VERJETNOSTNI RAČUN IN STATISTIKA

Aleksandar Jurišić, FRI

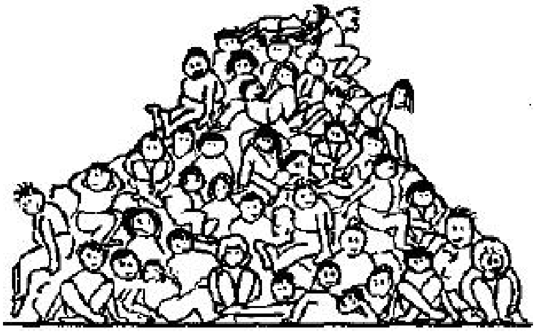
30. januar 2012

Seznam poglavij

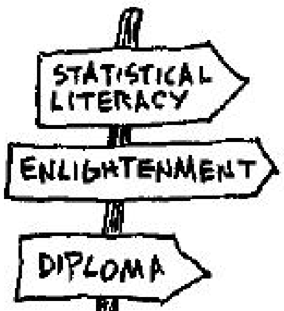
Kazalo

Poglavje 1

Uvod



Z veseljem bi predstavil učbenik, a ga nimamo (vsaj ne v slovenščini). Prav, pa predstavimo skripto in hkrati tudi naš predmet. Le-ta govori o dogodkih in podatkih.[1]



Iz učnega načrta: *"Predstavili bomo osnove teorije verjetnosti in njeno uporabo v statistiki, pa tudi nekaj osnov statistike. Bralec, ki bo predelal večji del naše snovi, bi moral znati opisati napovedljive vzorce, ki na dolgi rok vladajo slučajnim izidom ter doseči osnovno statistično pismenost, tj. sposobnost sledenja in razumevanja argumentov, ki izhajajo iz podatkov."* [2]

[1] Vsekakor ni samo prepreka na poti do diplome, pač pa se uporablja pri večini drugih predmetov na FRI. Če ne verjamete, me pridite negirat – moja vrata so Vam vedno odprta (Jadranska 21, vhod z južne strani, tj. Gradaščice, pritličje, soba 5), posebej v času govorilnih ur, ki so ta semester ob petkih 15:30-16:30.

[2] VAGABUND s foruma: huh, zveni preveč zapleteno in suhoparno, a saj to je pisano bolj za prfokse.

V nekem članku o računalniških virusih lahko preberemo debato o napadu črvov, ki so se razširili po Internetu ter upočasnili brskalnike in e-pošto širom po svetu. Koliko računalnikov je bilo okuženih? Strokovnjaki, na katere so se sklicali v članku, pravijo, da je bilo okužnih 39.000 računalnikov, ki so vplivali na stotine tisočev drugih sistemov. Kako so lahko prišli to te številke? Ali ne bi bilo težko priti do take številke? Ali so preverili vsak računalnik na Internetu, da bi se prepričali, če je okužen ali ne? Dejstvo, da je bil članek napisan v manj kot 24 urah od časa napada, sugerira, da je to število samo predpostavka. Vendar pa se lahko vprašamo, zakaj potem 39.000 in ne 40.000?

Statistika je znanost zbiranja, organiziranja in interpretiranja numeričnih dejstev, ki jih imenujemo *podatki*. Vsakodnevno smo s podatki takorekoč bombardirani. Večina ljudi povezuje "statistiko" z biti podatkov, ki izhajajo v dnevnem časopisju, novicah, reportažah: povprečna temperatura na današnji dan, procenti pri košarkaških prostih metih, procent tujih vlaganj na našem trgu, in anketa popularnosti politikov. Reklame pogosto trdijo, da podatki kažejo na superiornost njihovega produkta. Vse strani v javnih debatah o ekonomiji, izobraževanju in socialni politiki izhajajo iz podatkov. Kljub temu pa uporabnost statistike presega te vsakodnevne primere.

Podatki so navzoči pri delu mnogih, zato je izobraževanje na področju statistike izredno pomembno pri številnih poklicih. Ekonomisti, finančni svetovalci, vodstveni kader v politiki in gospodarstvu, vsi preučujejo najnovejše podatke o nezaposlenosti in inflaciji. Zdravniki morajo razumeti izvor in zanesljivost podatkov, ki so objavljeni v medicinskih revijah. Poslovne odločitve so običajno zasnovane na raziskavah tržišč, ki razkrijejo želje kupcev in njihovo obnašanje. Večina akademskih raziskav uporablja številke in tako hočeš nočeš izkorišča statistične metode.

Nič lažje se ni pobegniti podatkom kot se izogniti uporabi besed. Tako kot so besede na papirju brez pomena za nepismenega ali slabo izobraženega človeka, tako so lahko tudi podatki privlačni, zavajajoči ali enostavno nesmiselni. Statistična pismenost, tj. sposobnost sledenja in razumevanja argumentov, ki izhajajo iz podatkov, je pomembna za sleherno osebo.

Na statistiko in njene matematične temelje (verjetnost) lahko gledamo kot na učinkovito orodje, pa ne samo pri teoretičnem računalništvu (teoriji kompleksnosti, randomiziranih algoritmih, teoriji podatkovnih baz), pač pa tudi na praktičnih področjih. V vsakdanjem življenju ni pomembno da Vaš sistem obvlada čisto vse vhodne podatke, učinkovito pa naj opravi vsaj s tistimi, ki pokrijejo 99·99% primerov iz prakse.

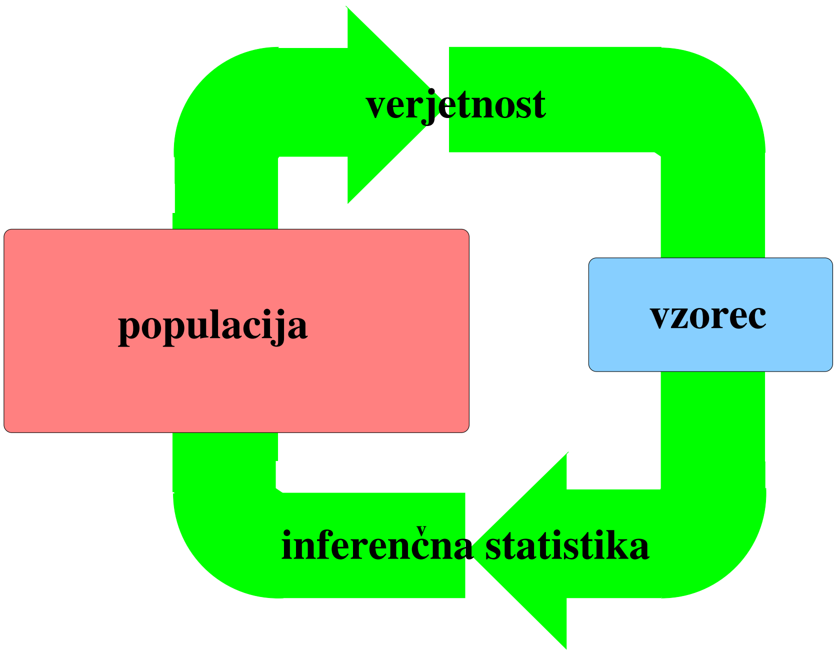
Zakaj verjetnost in statistika?

Statistika preučuje podatke, jih zbira, klasificira, povzema, organizira, analizira in interpretira. Glavni veji statistike sta **opisna statistika**, ki se ukvarja z organiziranjem, povzemanjem in opisovanjem zbirk podatkov (reduciranje podatkov na povzetke) ter **analitična statistika**, ki jemlje vzorce podatkov in na osnovi njih naredi zaključke (inferenčnost) o populaciji (ekstrapolacija). Slednjo bi po domače morda lahko poimenovali kar *napovedna statistika*, saj ne moremo nikoli biti popolnoma prepričani o zaključkih.

Populacija je podatkovna množica, ki ji je namenjena naša pozornost (vsi objekti, ki jih opazujemo).

Vzorec je podmnožica podatkov, ki so izbrani iz polulacije (po velikosti bistveno manjši od populacije).

Pravijo, da je slika vredna 1 000 besed, zato si še nazorno predočimo, kako nam verjetnost pomaga oceniti kakšen bo vzorec, ki ga bomo izbrali iz dane in dobro poznane populacije, medtem ko nam inferenčna statistika pomaga delati zaključke o celotni populaciji samo na osnovi vzorcev.



Za konec pa še tole. Matematika je dobra osnova, odličen temelj. Če pri konkretnem računalniškem ustvarjanju (pri tem mislim na razvoj programske opreme in reševanje logističnih problemov, ne pa prodajo in popravilo računalniške opreme) nimamo matematične izobrazbe/znanja, potem je skoraj tako kot če bi postavili hišo na blatu. Lahko izgleda lepo, vendar pa se bo začela pogrezati ob naslednjem nalivu.

Pogled od zunaj

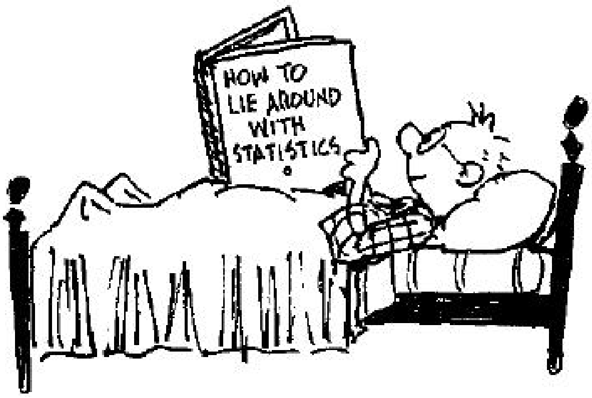
Števila so me pogosto begala,
še posebej, če sem imel pred seboj
neko njihovo razvrstitev,
tako da je tem primeru obveljala misel,
ki so jo pripisali Diaraeliju,
z vso pravico in močjo:

"Obstajajo tri vrste laži:

laži,
preklete laži in
statistika."

iz avtobiografije Marka Twaina

Vendar ne misliti, da vas bomo učili laganja – vse prej kot to.



Vaš cilj je lahko na primer, da se ne pustite drugim vleči za nos.

Del I

VERJETNOST

Ste se kdaj vprašali, zakaj so igre na srečo, ki so za nekatere rekreacija za druge droga, tako dober posel za igralnice?

Vsak uspešen posel mora iz uslug, ki jih ponuja, kovati napovedljive dobičke. To velja tudi v primeru, ko so te usluge igre na srečo. Posamezni hazarderji lahko zmagajo ali pa izgubijo. Nikoli ne morejo vedeti, če se bo njihov obisk igralnice končal z dobičkom ali z izgubo.

Igralnica pa ne kocka, pač pa dosledno dobiva in država lepo služi na račun loterij ter drugih oblik iger na srečo.

> *Presenetljivo je, da lahko skupni rezultat več tisoč naključnih izidov poznamo s skoraj popolno gotovostjo.*

Igralnici ni potrebno obtežiti kock, označiti kart ali spremeniti kolesa rulete. Ve, da ji bo na dolgi rok vsak stavljeni euro prinesel približno pet stotinov dobička.

Splača se ji torej osredotočiti na brezplačne predstave ali poceni avtobusne vozovnice, da bi privabili več gostov in tako povečali število stavljenega denarja. Posledica bo večji dobiček.

Igralnice niso edine, ki se okoriščajo z dejstvom, da so velikokratne ponovitve slučajnih izidov napovedljive.

Na primer, čeprav zavarovalnica ne ve, kateri od njenih zavarovancev bodo umrli v prihodnjem letu, lahko precej natančno napove, koliko jih bo umrlo. Premije življenjskih zavarovanj postavi v skladu s tem znanjem, ravno tako kot igralnica določi glavne dobitke.

Poglavje 2

Poskusi, dogodki in definicija verjetnosti

Ahil in Ajaks kockata, Amfora, okrog 530 pr.n.š, Eksekias, Vatikan

Naključnost so poznale že stare kulture: Egipčani, Grki, ... a je niso poskušale razumeti – razlagale so jo kot voljo bogov.

Na področju Francije je leta 1662 plemič Chevalier de Mere zastavil matematiku Blaise Pascalu vprašanje:

zakaj določene stave prinašajo dobiček druge pa ne.

Le-ta si je o tem začel dopisovati z matematikom Pierre Fermatom in iz tega so nastali začetki verjetnostnega računa.

Cardano

Prvo tovrstno razpravo je v resnici napisal italijanski kockar in matematik Cardano že leta 1545, a ni bila širše znana. Tudi leta 1662 je anglež John Graunt sestavil na osnovi podatkov prve zavarovalniške tabele.

Leta 1713 je švicarski matekatik Jakob Bernoulli objavil svojo knjigo *Umetnost ugibanja* s katero je verjetnostni račun postal resna in splošno uporabna veda. Njegov pomen je še utrdil Laplace, ko je pokazal njegov pomen pri analizi astronomskih podatkov (1812).

Leta 1865 je avstrijski menih Gregor Mendel uporabil verjetnostno analizo pri razlagi dednosti v genetiki. V 20. stoletju se je uporaba verjetnostnih pristopov razširila skoraj na vsa področja.

2.1 Poskusi in dogodki

Verjetnostni račun obravnava zakonitosti, ki se pokažejo v velikih množicah enakih ali vsaj zelo podobnih pojavov. Predmet verjetnostnega računa je torej iskustvene (empirične) narave in njegovi osnovni pojmi so povzeti iz izkušnje. Osnovni pojmi v verjetnostnem računu so: poskus, dogodek in verjetnost dogodka. **Poskus** je realizacija neke množice skupaj nastopajočih dejstev (kompleksa pogojev). Poskus je torej vsako dejanje, ki ga opravimo v natanko določenih pogojih.

Primeri: (1) met igralne kocke, (2) iz kupa 32 igralnih kart izberemo 5 kart, (3) met pikada v tarčo. ◇

Pojav, ki v množico skupaj nastopajočih dejstev ne spada in se lahko v posameznem poskusu zgodi ali pa ne, imenujemo **dogodek**. Za poskuse bomo privzeli, da jih lahko neomejeno velikokrat ponovimo. Dogodki se bodo nanašali na isti poskus. Poskuse označujemo z velikimi črkami iz konca abecede, npr. X, Y, X_1. Dogodke pa označujemo z velikimi črkami iz začetka abecede, npr. A, C, E_1.

Primeri: (1) v poskusu meta igralne kocke je na primer dogodek, da vržemo 6 pik; (2) v poskusu, da vlečemo igralno karto iz kupa 20 kart, je dogodek, da izvlečemo rdečo barvo. (3) v poskusu meta pikada zadanemo center (polje, ki označuje center). ◇

> Dogodek je **slučajen**, če so posamezni izidi negotovi, vendar pa je na dolgi rok vzorec velikega števila posameznih izidov napovedljiv.

Za statistika *slučajen* ne pomeni *neurejen.* Za slučajnostjo je neke vrste red, ki se pokaže šele na dolgi rok, po velikem številu ponovitev. Veliko pojavov, naravnih in tistih, ki so delo človeka, je slučajnih. Življenjska doba zavarovancev in barva las otrok sta primera naravne slučajnosti. Res, kvantna mehanika zagotavlja, da je na subatomskem nivoju v naravo vgrajena slučajnost. Teorija verjetnosti, matematični opis slučajnosti, je bistvenega pomena za prenekatero sodobno znanost.

Igre naključij so primeri slučajnosti, ki jo namenoma povzroči človek. Kocke v igralnicah so skrbno izdelane in izvrtane luknje, ki služijo označevanju pik, so zapolnjene z materialom, ki ima enako gostoto kot ostali del kocke. S tem je zagotovljeno, da ima stran s šestimi pikami enako težo kot nasprotna stran, na kateri je le ena pika. Tako je za vsako stran enako verjetno, da bo končala zgoraj. Vse verjetnosti in izplačila pri igrah s kockami temeljijo na tej skrbno načrtovani slučajnosti.

Izpostavimo nekatere posebne dogodke:

(a) **gotov** dogodek – oznaka $\boldsymbol{G}$: ob vsaki ponovitvi poskusa se zgodi.
Primer: dogodek, da vržemo 1, 2, 3, 4, 5, ali 6 pik pri metu igralne kocke. $\diamond$

(b) **nemogoč** dogodek – oznaka $\boldsymbol{N}$: nikoli se ne zgodi.
Primer: dogodek, da vržemo 7 pik pri metu igralne kocke. $\diamond$

(c) slučajen dogodek: včasih se zgodi, včasih ne.
Primer: dogodek, da vržemo 6 pik pri metu igralne kocke. $\diamond$

2.2 Računanje z dogodki

Dogodek A je **poddogodek** ali **način** dogodka B, kar zapišemo $\boldsymbol{A \subset B}$, če se vsakič, ko se zgodi dogodek A, zagotovo zgodi tudi dogodek B.

Primer: Pri metu kocke je dogodek A, da pade šest pik, način dogodka B, da pade sodo število pik. $\diamond$

Če je dogodek A način dogodka B in sočasno dogodek B način dogodka A, sta dogodka **enaka**: $(A \subset B) \wedge (B \subset A) \iff A = B$. **Vsota** dogodkov A in B, označimo jo z $\boldsymbol{A \cup B}$ ali $\boldsymbol{A + B}$, se zgodi, če se zgodi *vsaj* eden izmed dogodkov A in B.

Primer: Vsota dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je gotov dogodek. $\diamond$

V naslednji trditvi zberemo nekatere osnovne lastnosti operacij nad dogodki, ki smo jih vpeljali doslej (gotovo pa ne bi škodilo, če bi prej prebrali še Presekova članka Računala nove dobe 1. in 2., glej `http://lkrv.fri.uni-lj.si`, skrajšana verzija pa je v dodatku A.1 na strani 225).

Trditev 2.1. *Za poljubna dogodka A in B velja:*

- $A \cup B = B \cup A$ (tj. za vsoto velja pravilo o *zamenjavi*),
- $A \cup A = A$ (tj. vsi dogodki so *idempotenti* za vsoto),
- $A \cup N = A$ (tj. nemogoč dogodek N je za vsoto *nevtralen element*, tj. enota),
- $A \cup G = G$ (tj. gotov dogodek G za vsoto absorbira, je univerzalen),
- $B \subset A \iff A \cup B = A$
  (z vsoto znamo na drug način povedati kdaj je dogodek A način dogodka B),
- $A \cup (B \cup C) = (A \cup B) \cup C$ (tj. za vsoto velja pravilo o *združevanju*). □

Produkt dogodkov A in B, označimo ga z $\boldsymbol{A \cap B}$ ali $\boldsymbol{AB}$, se zgodi, če se zgodita A in B *hkrati*.

Primer: Produkt dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je nemogoč dogodek. ◇

Trditev 2.2. *Za poljubna dogodka A in B velja:*

- $A \cap B = B \cap A$ (tj. za produkt velja pravilo o *zamenjavi*),
- $A \cap A = A$ (vsi dogodki so *idempotenti* za produkt),
- $A \cap N = N$ (tj. nemogoč dogodek N absorbira, je univerzalen za produkt),
- $A \cap G = A$ (tj. gotov dogodek G je za produkt *nevtralen element*, tj. enota),
- $B \subset A \iff A \cap B = B$ (tudi s produktom znamo na drug način povedati kdaj je dogodek A način dogodka B),
- $A \cap (B \cap C) = (A \cap B) \cap C$ (tj. za produkt velja pravilo o *združevanju*),
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *in* $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  (tj. za produkt in vsoto veljata pravili o *porazdelitvi*). □

Dogodku A **nasproten** dogodek $\boldsymbol{\overline{A}}$, je tisti, ki se zgodi natanko takrat, ko se dogodek A ne zgodi, in ga imenujemo tudi *negacija* dogodka A.

Primer: Nasproten dogodek dogodku, da vržemo sodo število pik, je dogodek, da vržemo liho število pik. ◇

Trditev 2.3. *Za poljubna dogodka A in B velja:*

- $A \cap \overline{A} = N$,
- $A \cup \overline{A} = G$,
- $\overline{N} = G$, $\overline{G} = N$,
- $\overline{\overline{A}} = A$,
- $\overline{A \cup B} = \overline{A} \cap \overline{B}$ *in* $\overline{A \cap B} = \overline{A} \cup \overline{B}$ (*de Morganovi* pravili). □

Dogodka A in B sta **nezdružljiva**, če se ne moreta zgoditi hkrati. To se zgodi natanko tedaj, ko je njun produkt nemogoč dogodek, tj. $A\,B = N$.

Primer: Dogodka, A – da pri metu kocke pade sodo število pik in B – da pade liho število pik, sta nezdružljiva. ◇

Poljuben dogodek in njegov nasprotni dogodek sta vedno nezdružljiva. Ob vsaki ponovitvi poskusa se zagotovo zgodi eden od njiju, zato je njuna vsota gotov dogodek:

$$(A\,\overline{A} = N) \quad \wedge \quad (A + \overline{A} = G).$$

Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A **sestavljen** dogodek. Dogodek, ki ni sestavljen, imenujemo **osnoven** ali **elementaren** dogodek.

Primer: Pri metu kocke je šest osnovnih dogodkov: E_1, da pade 1 pika, E_2, da padeta 2 piki, ..., E_6, da pade 6 pik. Dogodek, da pade sodo število pik je sestavljen dogodek iz treh osnovnih dogodkov (E_2, E_4 in E_6). ◇

Množico dogodkov $S = \{A_1, A_2, \dots, A_n\}$ imenujemo **popoln sistem dogodkov**, če se v vsaki ponovitvi poskusa zgodi natanko eden od dogodkov iz množice S, tj.

- $A_i \neq N$, za $i = 1, 2, \dots, n$ (noben med njimi ni nemogoč),
- $A_i\,A_j = N$ za $i \neq j$ (so paroma nezdružljivi) in
- $A_1 + A_2 + \cdots + A_n = G$ (njihova vsota je gotov dogodek).

Primer: Popoln sistem dogodkov pri metu kocke sestavljajo na primer osnovni dogodki ali pa tudi dva dogodka: dogodek, da vržem sodo število pik, in dogodek, da vržem liho število pik. ◇

2.3 Definicija verjetnosti

Opišimo najpreprostejšo verjetnostno zakonitost. Denimo, da smo n–krat ponovili dan poskus in da se je k–krat zgodil dogodek A. Ponovitve poskusa, v katerih se A zgodi, imenujemo ugodne za dogodek A, število

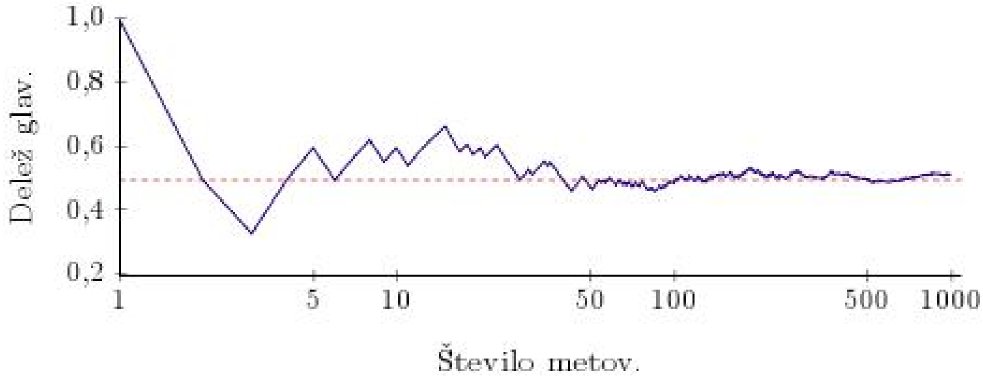
$$f(A) = \frac{k}{n}$$

pa je **relativna frekvenca** (pogostost) dogodka A v opravljenih poskusih. Statistični zakon, ki ga kaže izkušnja, je:

Če poskus X dolgo ponavljamo, se relativna frekvenca slučajnega dogodka ustali in sicer skoraj zmeraj toliko bolj, kolikor več ponovitev poskusa napravimo.

To temeljno zakonitost so empirično preverjali na več načinov.

Primer: (Metanje kovanca) Ko vržemo kovanec, sta le dva možna izida: glava ali cifra. Na sliki 2.1 je prikazan rezultat 1 000 metov kovanca. Za vsako število metov med 1 in 1 000 smo narisali delež tistih metov, katerih rezultat je bila glava. Prvi met je bila glava, zato je začetni delež glav enak 1. Pri drugem metu je padla cifra, zato se je delež glav po dveh metih zmanjšal na 0·5. Tretji met je bila spet cifra, sledili pa sta ji dve glavi, zato je bil delež glav po petih metih enak 3/5 ali 0·6.
Delež metov, pri katerih pade glava, se na začetku precej spreminja, vendar pa se ustali, ko število metov narašča. Nazadnje pride ta delež blizu 0·5 in tam obstane. Pogovorno rečemo, da se glava pojavi z *verjetnostjo* 0·5. Ta verjetnost je prikazana na grafu z vodoravno črto.
$\diamond$



Slika 2.1: Delež glav v odvisnosti od števila metov kovanca. Delež se sčasoma ustali pri verjetnosti za glavo.

Najbolj znani poskusi s kovanci, kjer so določali relativno frekvenco cifre ($f(A)$), pa so bili še bolj vstrajni:

- Buffon je v 4040 metih dobil $f(A) = 0{\cdot}5069$,
- Pearson je v 12000 metih dobil $f(A) = 0{\cdot}5016$,
- Pearson je v 24000 metih dobil $f(A) = 0{\cdot}5005$.

Ti in tudi drugi poskusi kažejo, da je ustalitev relativne frekvence v dovolj velikem številu ponovitev poskusa splošna zakonitost, je smiselna naslednja *statistična definicija verjetnosti*:

> *Verjetnost dogodka A v danem poskusu je število $P(A)$, pri katerem se navadno ustali relativna frekvenca dogodka A v velikem številu ponovitev tega poskusa.*

Ker je relativna frekvenca vedno nenegativna in kvečjemu enaka številu opravljenih poskusov, ni težko narediti naslednje zaključke.

Trditev 2.4. *Za poljubna dogodka A in B velja:*

1. $P(A) \geqslant 0$,
2. $P(G) = 1$, $P(N) = 0$ *in* $A \subset B \Rightarrow P(A) \leqslant P(B)$,
3. *če sta dogodka A in B nezdružljiva, potem je* $P(A + B) = P(A) + P(B)$. □

Klasični pristop k verjetnosti

Igralcem je že dolgo časa znano, da se meti kovanca, kart ali kock sčasoma ustalijo v točno določene vzorce. Verjetnostna matematika ima začetke v Franciji 17. stoletja, ko so hazarderji začeli prihajati k matematikom po nasvete (več v dodatku C.2, na strani 266). Ideja verjetnosti temelji na dejstvu, da lahko povprečni rezultat velikega števila slučajnih izidov poznamo z veliko gotovostjo. Vendar pa je definicija verjetnosti z izrazom "na dolgi rok" nejasna. Kdo ve, kaj “dolgi rok” je? Namesto tega matematično opišemo *kako se verjetnosti obnašajo*, pri čemer temeljimo na našem razumevanju deležev, ki se pojavljajo na dolgi rok. Da bi nadaljevali, si najprej zamislimo zelo preprost slučajni pojav, en sam met kovanca. Ko vržemo kovanec, ne vemo vnaprej, kakšen bo izid. Kaj pa vemo? Pripravljeni smo priznati, da bo izid bodisi glava bodisi cifra. Verjamemo, da se vsak od teh rezultatov pojavi z verjetnostjo 1/2. Opis meta kovanca sestoji iz dveh delov:

- seznama vseh možnih izidov in
- verjetnosti za vsakega od teh izidov.

Tak opis je osnova vseh verjetnostnih modelov. Tule je slovarček besed, ki jih pri tem uporabljamo:

Vzorčni prostor S slučajnega pojava je množica vseh možnih izidov.

Dogodek je katerikoli izid ali množica izidov slučajnega pojava.
Dogodek je torej podmnožica vzorčnega prostora.

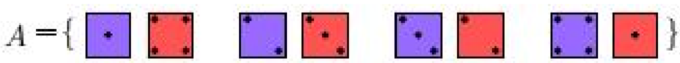
Verjetnostni model je matematični opis slučajnega pojava, sestavljen iz dveh delov: vzorčnega prostora S in predpisa, ki dogodkom priredi verjetnosti.

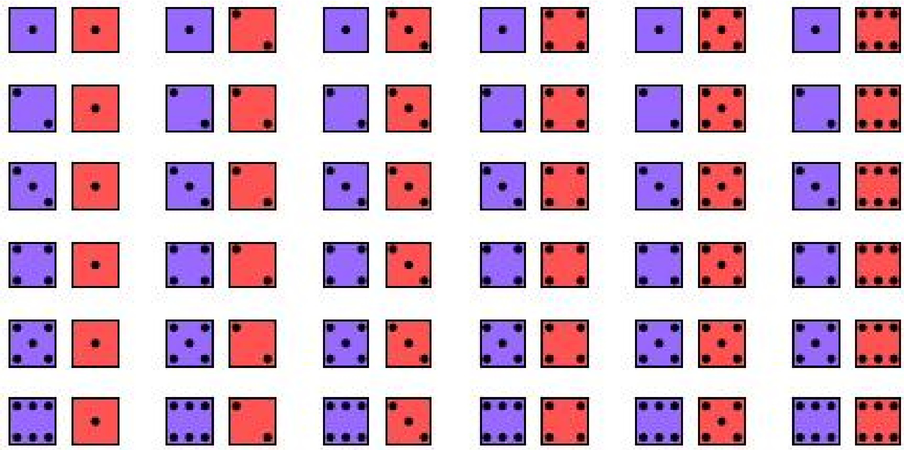
Vzorčni prostor S je lahko zelo preprost ali pa zelo zapleten. Ko vržemo kovanec enkrat, sta le dva možna izida, glava ali cifra. Vzorčni prostor je torej $S = \{G, C\}$. Če izbiramo slučajni vzorec 1500 polnoletnih Američanov kot v Gallupovih raziskavah, pa vzorčni prostor vsebuje vse možne izbire 1500 izmed več kot 200 milijonov odraslih prebivalcev. Ta S je hudo velik. Vsak element vzorčnega prostora S je možni vzorec za Gallupovo raziskavo, od koder ime *vzorčni prostor*.

Enako verjetni izidi in *klasična definicija verjetnosti*:

Pri enostavnem slučajnem vzorčenju imajo vsi vzorci enake možnosti, da so izbrani. Kadar je slučajnost produkt človekovega načrtovanja, se velikokrat zgodi, da so izidi iz vzorčnega prostora enako verjetni.

Primer: (Metanje kocke) Metanje dveh kock hkrati je običajni način za izgubljanje denarja v igralnicah. Ko vržemo dve kocki, je možnih 36 izidov, če med kockama razlikujemo. Ti možni izidi so prikazani na sliki 2.2. Tvorijo vzorčni prostor S.
“Pade 5” je nek dogodek, označimo ga z A. Vsebuje 4 od 36 možnih izidov:





Slika 2.2: Možni izidi pri metu dveh kock.

Če so kocke dobro izdelane, izkušnje kažejo, da se vsak od 36 izidov s slike 2.2 pojavi enako pogosto. Razumen verjetnostni model torej pripiše vsakemu od izidov verjetnost 1/36, verjetnost. ◇

Trditev 2.4 nam v primeru enako verjetnih izidov pove, kolikšne so te verjetnosti.

Vzemimo $s \in \mathbb{N}$ in predpostavimo, da so dogodki iz popolnega sistema dogodkov $\{E_1, E_2, \dots, E_s\}$ enako verjetni: $P(E_1) = P(E_2) = \dots = P(E_s) = p$. Tedaj je $P(E_i) = 1/s$, $1 \leqslant i \leqslant s$. Če je nek dogodek A sestavljen iz r dogodkov tega popolnega sistema dogodkov, potem je njegova verjetnost $P(A) = r/s$.

Z drugimi besedami:

Če je pri slučajnem pojavu možnih s izidov, ki so vsi enako verjetni, potem je verjetnost vsakega od izidov enaka $1/s$. Za poljubni dogodek A je njegova verjetnost enaka

$$P(A) = \frac{\text{število izidov v } A}{\text{število izidov v } S} = \frac{\text{število izidov v } A}{s}.$$

Primer: Izračunajmo verjetnost dogodka A, da pri metu kocke padejo manj kot 3 pike. Popolni sistem enako verjetnih dogodkov sestavlja 6 dogodkov. Od teh sta le dva ugodna za dogodek A (1 in 2 piki). Zato je verjetnost dogodka A enaka $2/6 = 1/3$. ◇

Primer: (Naključna števila) Z generatorjem naključnih števil generiramo števila iz množice $\{0, 1, \dots, 9\}$ za neko zaporedje. Če je generator dober, potem se vsak element iz zaporedja z enako verjetnostjo katerikoli od desetih možnih. Torej je verjetnost vsakega od desetih izidov enaka 1/10, verjetnostni model pa:

Izid	0	1	2	3	4	5	6	7	8	9
Verjetnost	0·1	0·1	0·1	0·1	0·1	0·1	0·1	0·1	0·1	0·1

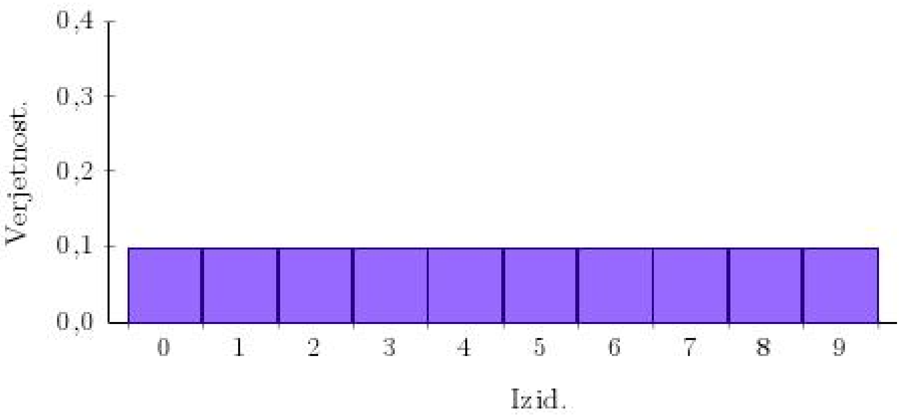
Na sliki 2.3 je ta model prikazan z verjetnostnim histogramom. Verjetnost poljubnega dogodka poiščemo tako, da preštejemo, koliko izidov vsebuje. Naj bo na primer

$$A = \{\text{lih izid}\} = \{1,3,5,7,9\} \quad \text{in} \quad B = \{\text{izid je manjši ali enak } 3\} = \{0,1,2,3\}.$$

Vidimo, da je $P(A) = 0,5$ in $P(B) = 0,4$. Dogodek $[A$ ali $B]$ vsebuje 7 izidov,

$$[A + B] = \{0,1,2,3,5,7,9\},$$

zato ima verjetnost 0·7. Ta verjetnost *ni* vsota verjetnosti $P(A)$ in $P(B)$, ker A in B *nista* disjunktna dogodka. Izida 1 in 3 pripadata tako A kot B. $\diamond$



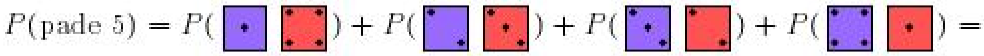
Slika 2.3: Verjetnostni histogram, ki prikazuje verjetnosti pri naključnem generiranju števke med 0 in 9.

Primer: (Verjetnost pri metanu dveh kock) V igrah kot je *craps*[1] je pomembna samo *vsota* pik, ki jih vržemo. Spremenimo torej izide, ki nas zanimajo, takole: vržemo dve kocki in preštejemo število pik. V tem primeru je možnih le 11 izidov, od vsote 2, ki jo dobimo, če vržemo dve enici, do vsote 12, ki jo dobimo z dvema šesticama. Vzorčni prostor je

$$S = \{2,3,4,5,6,7,8,9,10,11,12\}.$$

Če primerjamo ta S z vzorčnim prostorom na sliki 2.2, vidimo, da se S lahko spremeni, če spremenimo podrobnosti, ki jih pri nekem pojavu opazujemo. Izidi v tem novem vzorčnem prostoru *niso* več enako verjetni, ker lahko 7 dobimo na šest načinov, 2 ali 12 pa samo na enega. Za igralniške kocke je smiselno vsakemu od omenjenih 36 izidov predpisati isto verjetnost. Ker mora biti verjetnost vseh 36 dogodkov skupaj enaka 1, mora imeti vsak od izidov verjetnost 1/36. Kolikšna je verjetnost, da je vsota pik na obeh kockah enaka 5? Ker je ta dogodek sestavljen iz štirih možnih izidov, ki smo jih zapisali v prejšnjem primeru, nam pravilo vsote , tj. Trditev 2.4(3), pove, da je

[1] Ameriška igra z dvema kockama.

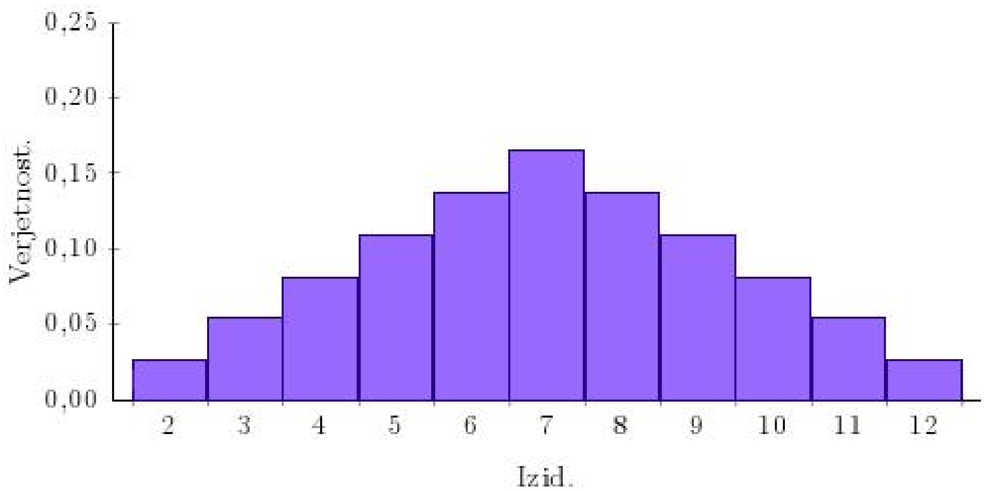


$$= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{4}{36} = 0{\cdot}111.$$

Kaj pa verjetnost, da dobimo 7? Na sliki 2.2 najdemo šest izidov, pri katerih je vsota pik enaka 7. Verjetnost za 7 je torej 6/36, kar je približno 0·167. Na ta način nadaljuj z računanjem, da dobiš celoten verjetnostni model (vzorčni prostor in predpise verjetnosti) za met dveh kock, pri katerem opazujemo vsoto pik. Tole je rezultat:

Izid	2	3	4	5	6	7	8	9	10	11	12
Verjetnost	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Na sliki 2.4 je *verjetnostni histogram* tega verjetnostnega modela. Višina vsakega stolpca prikazuje verjetnost dogodka, ki ga ta stolpec predstavlja. Ker so višine ravno verjetnosti, se seštejejo v 1. Slika 2.4 je idealizirana podoba rezultatov velikega števila metov kock. Kot idealizacija je popolnoma simetrična. ◇



Slika 2.4: Verjetnostni histogram za met dveh kock, ki prikazuje, kakšne so verjetnosti za posamezno vsoto pik.

Primer: V 17. stoletju je bila popularna igra, pri kateri se je stavilo na vsoto pik pri metu treh kock. Stavili so večinoma na 9 ali 10.

Kockarji so napisali vse možnosti na naslednji način:

vsota 9		
1	2	6
1	3	5
1	4	4
2	2	5
2	3	4
3	3	3

vsota 10		
1	3	6
1	4	5
2	2	6
2	3	5
2	4	4
3	3	4

Sklepali so, da sta stavi enakovredni. So imeli prav? Problem je kockarjem rešil sam Galileo Galilei (1564-1642). Na list papirja jim je napisal vse možne trojice padlih pik na kockah. Koliko jih je? ◇

Z zgornjimi primeri smo spoznali enega od načinov, kako dogodkom priredimo verjetnosti: predpišemo verjetnosti vsakemu od izidov, nato pa jih seštejemo, da dobimo verjetnost dogodka. Da bi tako prirejanje zadoščalo pravilom verjetnosti, se morajo verjetnosti posameznih izidov sešteti v 1.

Verjetnostni model za končen vzorčni prostor podamo tako, da predpišemo verjetnost vsakemu posameznemu izidu. Te verjetnosti morajo biti števila med 0 in 1 in njihova vsota mora biti enaka 1. Verjetnost poljubnega dogodka je vsota verjetnosti izidov, ki ga sestavljajo.

Primer: (Rangiranje srednješolcev) Naključno izbiramo študente in študentke prvih letnikov in jih vprašamo, kje so se v srednji šoli nahajali glede na učni uspeh. Tule so verjetnosti, ki jih dobimo iz deležev v velikem vzorcu študentov:

Razred	Zgornjih 20%	Drugih 20%	Tretjih 20%	Četrtih 20%	Spodnjih 20%
Verjetnost	0·41	0·23	0·29	0·06	0·01

Prepričaj se, da se te verjetnosti seštejejo v 1. Zdaj lahko izračunamo še naslednji dve verjetnosti:

$$\begin{aligned} P(\text{študent je v zgornjih } 40\%) &= P(\text{zgornjih } 20\%) + P(\text{drugih } 20\%) = \\ &= 0{\cdot}41 + 0{\cdot}23 = 0{\cdot}64, \end{aligned}$$

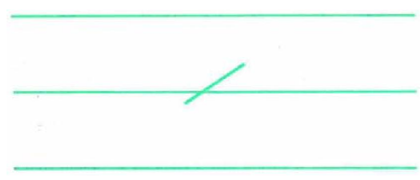
$$P(\text{študent ni v zgornjih } 40\%) = 0{\cdot}23 + 0{\cdot}29 + 0{\cdot}06 + 0{\cdot}01 = 0{\cdot}59.$$

Ali veš, zakaj je verjetnost, da študent ni v zgornjih 20%, enaka 1 minus verjetnost, da izbrani študent je v zgornjih 20%? ◇

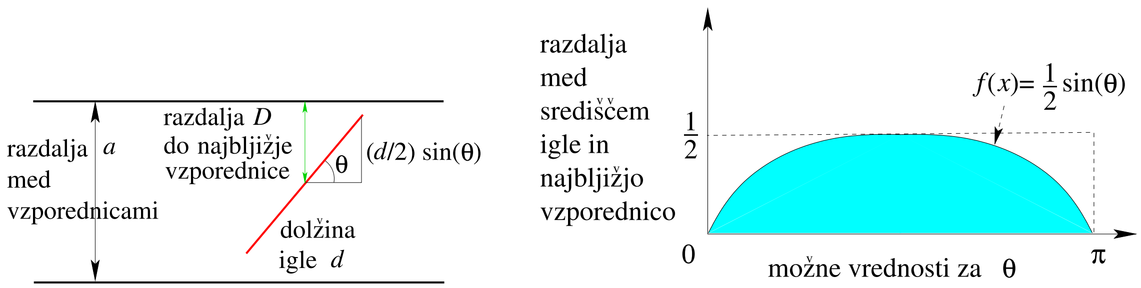
Geometrijska verjetnost

V primerih, ko lahko osnovne dogodke predstavimo kot 'enakovredne' točke na delu premice (ravnine ali prostora), določimo verjetnost sestavljenega dogodka kot razmerje dolžin (ploščin, prostornin) dela, ki ustreza ugodnim izidom, in dela, ki ustreza vsem možnim izidom. Verjetnost torej lahko računamo s pomočjo geometrije. Lahko pa se zgodi tudi obratno, da nam verjetnost pomaga določiti nekaj geometrijskega, npr. število π, ki ima neskončen neperiodičen zapis $3{\cdot}14159265\ldots$ S tem številom je povezana ena od znamenitih nalog klasičnega verjetnostnega računa iz leta 1777, ki jo je Buffon zastavil ter rešil z enostavno Monte Carlo metodo in jo predstavimo v naslednjem primeru.

Primer: Na večji papir narišemo na razdalji a vrsto vzporednih premic. Nato vzamemo iglo dolžine d (zaradi enostavnosti naj bo $d \leqslant a$) in jo na slepo večkrat vržemo na papir. Zanima nas verjetnost dogodka A, da igla seče eno od vzporednih premic. Igla seče najbljižjo vzporednico, če je $D \leqslant (d/2)\sin(\theta)$.



Slika: Vzporednice na razdalji a.



Slika 2.5: (a) Splošna a in d. (b) Za naše potrebe bi lahko vzeli kar $d = 1$ in $a = 1$.

Ploščino pod krivuljo funkcije $(d/2)\sin(\theta)$ izračunamo z določenim integralom te funkcije od 0 do π, in je enaka d. Ploščina pravokotnika $\pi \times (a/2)$ pa je $\pi a/2$. Kvocijent teh dveh ploščin pa nam da $P(A) = 2d/(\pi a)$. ◇

Pri zgornjem primeru je najbolj zanimivo, da v rezultatu nastopa število π, ki ga srečamo na primer pri obsegu ali pa ploščini kroga. Ravno ta prisotnost števila π je vzpodbudila vrsto ljudi k naslednjemu poskusu. Poskus so ponovili zelo velikokrat (n-krat) in prešteli, kolikokrat je sekala eno od črt (m-krat). Relativno frekvenco $f(A) = m/n$ so vzeli za približek zgornje verjetnosti in od tod ocenili približek za število π: $\pi \doteq 2dn/(am)$.

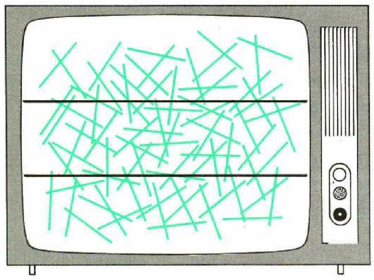
eksperimentator	*d*	*leto*	*število poskusov*	*ocena za* π
Wolf	0·8	1850	5000	3·1596
Smith	0·6	1855	3204	3·1553
deMorgan	1·0	1860	600	3·1370
Fox	0·75	1894	1120	3·1419

n	m	*ocena za* π
98	57	3·43860
234	150	3·14094
239	152	3·14474
357	220	3·24545

Tabela 2: Rezultati (a) iz raznih knjig (za $a = 1$), (b) našega programa (na 5 decimalnih mest).

Za računanje čim večjega števila decimalnih mest števila π je znanih vrsta drugih uspešnejših metod. Zgornji poskus z metanjem igle ne prinese velike natančnosti, prestavili smo ga bolj kot zanimivost. Pravzaprav je tu še en pomemben pogled. Metanje igle lahko zelo nazorno posnemamo z računalnikom. Zaradi enostavnosti vzemimo, da je $d = a$. E. Kramar je v Preseku 12/4 (1984/85), glej `http://www.presek.si/12/731-Kramar.pdf`, predstavil zgled programa v basicu za hišni računalnik ZX Spectrum. Vzel je $a = 50$ enot in na zaslonu narisal dve vzporednici na tej razdalji. Slučajnost pri metu je generiral s funkcijo `RND`, ki pomeni na slepo izbrano realno število med 0 in 1. Lego igle je določil s slučajnim kotom med $0°$ in $180°$ ($\texttt{fi} = \texttt{PI} * \texttt{RND}$) in slučajno razdaljo ene konice igle od zgornje vzporednice ($b = \texttt{INT}(a * \texttt{RND})$). Dodal je še slučajni pomik v levo ali desno (stavek 85), da igla pada nekako po celem ekranu, vendar to ne vpliva na rezultat.

```
  5 REM metanje igle
  6 REM in računanje približka za število pi
  7 REM
 10 INPUT "število metov"; n
 15 PRINT "število metov ";n
 20 LET a=50: LET m=0
 30 PLOT 0,70: DRAW 255,0
 40 PLOT 0,120: DRAW 255,0
 50 FOR i =1 TO n
 60 LET b=INT(a*RND): LET fi=PI*RND
 70 LET x1=a*COS(fi): LET y1=a*0.5*SIN(fi)
 80 LET z1=70+b-y1: LET z2=70+b+y1
 85 LET c=INT(150*RND)+50
 90 PLOT c,z1: DRAW x1,2*y1
100 IF z1 < 70 AND z2 > 120 THEN GO TO 140
110 LET m=m+1
120 PRINT AT 3,3; m;"   ";i
130 NEXT i
140 PRINT "približek za pi =    "; 2*n/m
150 STOP
```



Tule je še nekaj Kramerjevih komentarjev: “Bralec, ki ima dostop do ra čunalnika, bo gotovo tudi sam poskusil posnemati metanje igle (ali igel) na računalniku, pri tem bo spreminjal program po svoji želji. Če vzamemo večje število poskusov (n), lahko izključimo risanje, saj nas zanima samo rezultat na koncu. Velja opozoriti, da prevelikega števila n pravzaprav nima smisla vzeti, saj rezultati ne bodo bistveno boljši, zaradi vsakokratnega računanja nekaterih funkcij pa bo računalnik kar precej časa “mlel”. Prav tako nima smisla iti na lov za čim boljšim približkom števila π. Ker poznamo veliko njegovih decimalnih mest, bi lahko poskus priredili tako, da bi ga prekinili v trenutku za nas najbolj sprejemljivega približka. To pa ni poskus v zgornjem smislu, ko je število metov vnaprej izbrano. Drug način grobe potvorbe rezultatov pa bi bil, da bi vzeli ulomek, ki je dober približek za število π, kot je na primer star kitajski približek 355/113, in bi nekomu sporočili, da nam je pri 355-kratnem metu igle ta sekala vzporednico 226-krat. V resnici je prav malo verjetno (da se ugotoviti celo verjetnost tega dogodka, in sicer je manjša od 0·034), da se nam bo ravno to zgodilo. Omeniti velja še dejstvo, da smo zgoraj privzeli, da je generator slučajnih števil (RND) na računalniku dobro narejen, da dobro posnema slučajno izbiro realnega števila med 0 in 1, kar pa je zopet le približek.”

Za konec pa omenimo možnost eksperimentiranja s programi na internetu, npr.:
http://www.angelfire.com/wa/hurben/buff.html in
http://demonstrations.wolfram.com/BuffonsNeedleProblem/.

2.4 Osnovne lastnosti verjetnosti

Trditev 2.5. *Za poljubna dogodka A in B velja:*

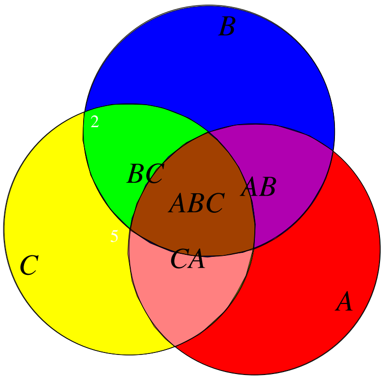
$$P(A+B) = P(A) + P(B) - P(AB).$$

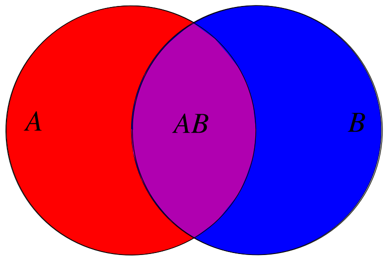
John Venn, angleški matematik 1834-1923.

Dokaz. Pri statističnem pristopu je dovolj, da se prepričamo o veljavnosti zveze

$$k_{A+B} + k_{AB} = k_A + k_B$$

(glej spodnjo sliko). □





Primer: Denimo, da je verjetnost, da študent naredi izpit iz Sociologije $P(S) = 2/3$. Verjetnost, da naredi izpit iz Politologije je $P(P) = 5/9$. Če je verjetnost, da naredi vsaj enega od obeh izpitov $P(S+P) = 4/5$, kolikšna je verjetnost, da naredi oba izpita?

$$P(SP) = P(S) + P(P) - P(S+P) = \frac{2}{3} + \frac{5}{9} - \frac{4}{5} = 0{\cdot}42. \qquad \diamond$$

Posledica 2.6. $P(\overline{A}) = 1 - P(A)$. □

Primer: Iz kupa 32 kart slučajno povlečemo 3 karte. Kolikšna je verjetnost, da je med tremi kartami vsaj en as (dogodek A)? Pomagamo si z nasprotnim dogodkom $\overline{A}$, da med tremi kartami ni asa. Njegova verjetnost po klasični definiciji verjetnosti je določena s kvocientom števila vseh ugodnih dogodkov v popolnem sistemu dogodkov s številom vseh dogodkov v tem sistemu dogodkov. Vseh dogodkov v popolnem sistemu dogodkov je $\binom{32}{3}$, ugodni pa so tisti, kjer zbiramo med ne–asi, tj. $\binom{28}{3}$. Torej je

$$P(\overline{A}) = \frac{\binom{28}{3}}{\binom{32}{3}} = 0{\cdot}66\ ;\ P(A) = 1 - P(\overline{A}) = 1 - 0{\cdot}66 = 0{\cdot}34. \qquad \diamond$$

Omenimo še dve posledici, ki prideta pogosto prav. Naslednjo trditev lahko dokažemo na enak način kot Trditev 2.5 ali pa kar z uporabo tega izreka.

Posledica 2.7. *Za dogodke A, B in C velja:*

$$P(A+B+C) = P(A)+P(B)+P(C) - \big(P(AB)+P(AC)+P(BC)\big) + P(ABC). \qquad \square$$

Kako lahko to pravilo posplošimo še na več dogodkov?

Namig: *Pravilo o vključitvi in izključitvi za množice $A_1, A_2, \ldots, A_n$:*

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{i=1}^{n} |A_i| - \sum_{1\leqslant i_1<i_2\leqslant n}^{n} |A_{i_1} \cap A_{i_2}|$$

$$+ \sum_{1\leqslant i_1<i_2<i_3\leqslant n}^{n} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \cdots + (-1)^{n-1}|A_1 \cap A_2 \cap \ldots \cap A_n|.$$

Posledica 2.8. *Če so dogodki A_i, $i \in I$ paroma nezdružljivi, velja*

$$P\Big(\sum_{i\in I} A_i\Big) = \sum_{i\in I} P(A_i).$$

Velja tudi za števno neskončne množice dogodkov. $\square$

Zgornjo relacijo si lažje zapomnimo, če jo na glas "odrecitiramo":

> "**verjetnost** *vsote* paroma nezdružljivih dogodkov
> je enaka
> *vsoti* **verjetnosti** teh dogodkov."

$$\begin{tikzcd}
(A,B) \arrow[r, "+"] \arrow[d] & A+B \arrow[d, "P(.)"] \\
\big(P(A),P(B)\big) \arrow[r, "+"] & P(A)+P(B) = P(A+B)
\end{tikzcd}$$

Torej gre za nekakšno pravilo o zamenjavi (med vrstnim redom računanja vsote in verjetnosti), ki velja za paroma nezdružljive dogodke.

S simulacijo lahko pridemo do števila $1/e$.

Primer: Za n različnih dopisov, namenjenih n osebam, so pripravljene že naslovljene ovojnice. Dopise bomo na slepo razdelili v ovojnice. Kolika je pri tem verjetnost, da niti en dopis ne bo prišel na pravi naslov?

Negacija dogodka A, ki mu iščemo verjetnost, je da pride vsaj eno pismo na pravi naslov. Pisma uredimo po nekem vrstnem redu in naj bo A_i $(1 \leqslant i \leqslant n)$ dogodek, da pride i-to pismo na pravi naslov. Potem je

$$\overline{A} = A_1 + \cdots + A_n,$$

dogodki na desni strani pa niso nezdružljivi. Torej lahko uporabimo pravilo o vključitvi in izključitvi, pri čemer označimo verjetnost i-te vsote z S_i $(1 \leqslant i \leqslant n)$. Potem ima vsota S_1 n členov, ki so vsi med seboj enaki, saj gre za izbiranje na slepo:

$$S_1 = n\,P(A_1).$$

Poskus je v našem primeru razdeljevanje n dopisov po ovojnicah, torej je možnih izidov $n!$ in so vsi med seboj enako verjetni. Med izidi so za A_1 ugodni tisti, pri katerih pride prvi dopis v prvo ovojnico, tj.

$$S_1 = \frac{n(n-1)!}{n!} = 1.$$

Nadalje je

$$S_2 = \binom{n}{2} P(A_1 A_2) = \binom{n}{2} \frac{(n-2)!}{n!} = \frac{1}{2!}.$$

in v splošnem $S_k = 1/k!$ $(1 \leqslant k \leqslant n)$. Torej je

$$P(\overline{A}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + \frac{(-1)^{n-1}}{n!} \quad \text{oziroma} \quad P(A) = \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}$$

in končno, če upoštevamo še Trditev A.5,

$$\lim_{n\to\infty} P(A) = \frac{1}{e},$$

tako da imamo že pri razmeroma majhnih n

$$P(A) \doteq \frac{1}{e} \doteq 0{\cdot}369.$$

$\diamond$

2.5 Aksiomi Kolmogorova

Poglejmo še, kako na verjetnost gledajo matematiki. Dogodek predstavimo z množico zanj ugodnih izidov; gotov dogodek G ustreza univerzalni množici; nemogoč dogodek pa prazni množici. Neprazna družina dogodkov $\mathcal{D}$ je **algebra dogodkov** (tudi σ-algebra), če velja:

- $A \in \mathcal{D} \Rightarrow \overline{A} \in \mathcal{D}$,
- $A, B \in \mathcal{D} \Rightarrow A + B \in \mathcal{D}$.

Pri neskončnih množicah dogodkov moramo drugo zahtevo posplošiti:

- $A_i \in \mathcal{D}, i \in I \Rightarrow \sum_{i \in I} A_i \in \mathcal{D}$.

Naj bo $\mathcal{D}$ algebra dogodkov v G. **Verjetnost na G** je preslikava $P : \mathcal{D} \to \mathbb{R}$ z lastnostmi:

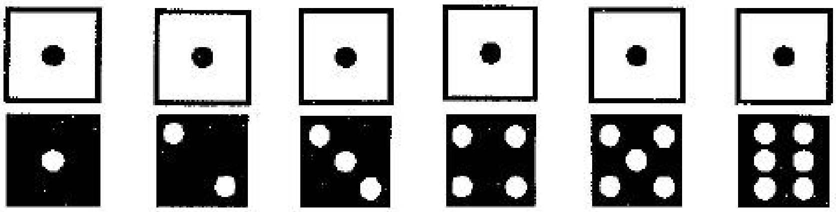
1. $P(A) \geqslant 0$ za vsak $A \in \mathcal{D}$.
2. $P(G) = 1$.
3. Če so dogodki A_i, $i \in I$ (I je množica indeksov), paroma nezdružljivi, je

$$P\left(\sum_{i \in I} A_i\right) = \sum_{i \in I} P(A_i).$$

Trojica $(G, \mathcal{D}, P)$ določa **verjetnostni prostor**.

Poglavje 3

Pogojna verjetnost



3.1 Intriga (po Kvarkadabri)

Ne podcenjujmo vpliva najrazličnejših rubrik v popularnih časopisnih prilogah, kjer nas domnevni strokovnjaki zasipajo z nasveti vseh vrst, rubrike krojijo mnenja ljudi in spreminjajo navade celotnih nacij, sprožajo obsežne polemike tako med širšimi množicami kot tudi v ozki strokovni javnosti. Na področju zdravja in prehrane tako burne odzive seveda pričakujemo, povsem nekaj drugega pa je, če jih sproži preprosto *matematično vprašanje*.

Revija *Parade* - kot prilogo jo vsako nedeljo dodajo več kot 400 ameriškim časopisom in doseže okoli 70 milijonov bralcev, že dolgo izhaja rubrika z imenom "Vprašajte Marilyn." Ureja jo **Marilyn vos Savant**. Sredi 80ih jo je *Guinnessova knjiga rekordov* razglasila za rekorderko z najvišjim inteligenčnim količnikom na planetu. V svoji rubriki zdaj že več kot 20 let odgovarja na najrazličnejša vprašanja bralcev in rešuje njihove težave. Med vsemi vprašanji, ki jih je kdaj obravnavala, ima prav posebno mesto na prvi pogled zelo preprost problem, ki ji ga je 9. septembra 1990 zastavil gospod Craig F. Whitaker:

Dve kozi in avtomobil

"Vzemimo, da sodelujete v nagradni igri, kjer vam ponudijo na izbiro troje vrat. Za enimi se skriva avto, za drugima dvema pa koza. Recimo, da izberete vrata številka 3, voditelj igre, ki ve, kaj se nahaja za posameznimi vrati, pa nato odpre vrata številka 1, za katerimi se pokaže koza. Nato vas vpraša: 'Bi se sedaj raje odločili za vrata številka 2?'

Zanima me, ali se tekmovalcu splača zamenjati izbor vrat?"

Poudariti je potrebno, da mora gostitelj nagradne igre vsakič postopati enako. Ne more enkrat ponuditi zamenjavo (npr. takrat, ko vidi, da nastopajoči kaže na vrata za katerimi se skriva avto), drugič pa ne (npr. takrat, ko nastopajoči kaže na vrata za katerimi je koza).

Vprašanja se je prijelo ime "*Monty Hall problem*", po imenu voditelja popularne ameriške televizijske oddaje Pogodimo se (Let's Make a Deal), v kateri je voditelj Monty Hall goste izzival, da so sprejemali ali zavračali najrazličnejše ponudbe, ki jim jih je zastavljal. Marilyn je bralcu v svoji rubriki odgovorila, da se nam vrata vsekakor splača zamenjati, saj se tako verjetnost, da bomo zadeli avto, poveča za dvakrat. Tole je njen odgovor: Seveda se splača zamenjati vrata. Prva vrata imajo le 1/3 verjetnosti za zmago, medtem ko imajo druga verjetnost 2/3.

Najlažje si vse skupaj predstavljate takole. Predpostavimo, da je na voljo milijon vrat in vi izberete prva. Nato voditelj, ki ve, kaj se nahaja za posameznimi vrati, odpre vsa vrata razen prvih vrat in vrat številka 777 777. V tem primeru bi zelo hitro zamenjali svoj izbor, kajne? Se najinteligentnejša ženska na planetu moti?

Sledila je ploha kritik (več kot 10 000 pisem jeznih bralcev, med katerimi je bilo ogromno učiteljev matematike). Skoraj 1 000 pisem je bilo podpisanih z imeni (dr. nazivi, napisana na papirju z glavo katere od ameriških univerz - www.marilynvossavant.com). Marylin bralce zavaja, saj se verjetnost za zadetek nikakor ne more spremeniti, če vmes zamenjamo izbor vrat. Neki profesor matematike je bil zelo neposreden: "Udarili ste mimo! ... Kot profesionalni matematik sem zelo zaskrbljen nad pomanjkanjem matematičnih veščin v širši javnosti. Prosim, da se opravičite in ste v prihodnosti bolj pazljivi." Drugi je Marylin celo obtožil, da je ona sama koza.

Polemika je pristala celo na naslovnici New York Timesa, v razpravo so se vključila tudi nekatera znana imena iz sveta matematike. O odgovoru vos Savantove, da naj tekmovalec zamenja vrata, so razpravljali tako na hodnikih Cie kot v oporiščih vojaških pilotov ob Perzijskem zalivu. Analizirali so ga matematiki z MIT in računalniški programerji laboratorijev Los Alamos v Novi Mehiki. Poleg žaljivih pisem, ki so njen odgovor kritizirala, je Marilyn vseeno prejela tudi nekaj pohval. Profesor s prestižnega MIT: "Seveda imate prav. S kolegi v službi smo se poigrali s problemom in moram priznati, da je bila večina, med njimi sem bil tudi sam, sprva prepričana, da se motite!"

Eksperimentalna ugotovitev

Marilyn se kritik ni ustrašila - navsezadnje je objektivno izmerljivo po inteligenčnem količniku pametnejša od vseh svojih kritikov, zato je v eni od svojih naslednjih kolumn vsem učiteljem v državi zadala nalogo, da to preprosto igrico igrajo s svojimi učenci v razredu (seveda ne s pravimi kozami in avtomobilom) in ji pošljejo svoje rezultate. Te je nato tudi objavila in seveda so se povsem skladali z njenim nasvetom, da se v tem konkretnem primeru bistveno bolj splača spremeniti izbiro vrat. Kdo ima prav?

Razprava o Monty Hall problemu spada na področje, ki mu matematiki pravijo **pogojna verjetnost**. Najbolj preprosto rečeno je to veda, ki se ukvarja s tem, kako prilagoditi verjetnost za posamezne dogodke, ko se pojavijo novi podatki. Bistvo zapleta, ki je izzval tako obsežno in čustveno nabito reakcijo bralcev, je v tem, da so bralci večinoma spregledali ključni podatek. Zelo pomembno je namreč dejstvo, da voditelj igre vnaprej ve, za katerimi vrati je avtomobil.

Ko v drugem delu odpre vrata, za katerimi se pokaže koza, vnaprej ve, da za temi vrati ni avtomobila. Če voditelj te informacije ne bi imel in bi vrata odpiral povsem naključno tako kot igralec, se verjetnost za zadetek ob spremembi vrat res ne bi povečala. Potem bi držale ugotovitve več 1 000 bralcev, ki so poslali jezna pisma na uredništvo revije, da Marilyn ne pozna osnov matematike. Matematična intuicija nam namreč pravi, da je verjetnost, da bo avto za enimi ali za drugimi vrati, ko so dvoja še zaprta, enaka. To je seveda res, če zraven ne bi bilo še voditelja, ki ve več kot mi.

Najlažje nejasnost pojasnimo, če analiziramo dogajanje izza kulis, od koder ves čas vidimo, za katerimi vrati je avto in kje sta kozi. Če tekmovalec že v prvo izbere vrata, za katerimi je avto, bo voditelj odprl katera koli od preostalih dveh vrat in zamenjava bo tekmovalcu v tem primeru le škodila. Ampak to velja le za primer, če v prvo izbere vrata, za katerimi je avto, verjetnost za to pa je 1/3. Če pa v prvo tekmovalec izbere vrata, za katerimi je koza, bo voditelj moral odpreti edina preostala vrata, za katerimi se nahaja koza. V tem primeru se bo tekmovalcu zamenjava vrat v vsakem primeru obrestovala in bo tako z gotovostjo zadel avto.

Če v prvo tekmovalec izbere kozo, se mu vedno splača zamenjati, če pa v prvo izbere avto, se mu zamenjava ne izplača. Verjetnost, da v prvo izbere kozo, je 2/3, medtem ko je verjetnost, da izbere avto, le 1/3. Če se tekmovalec odloči za strategijo zamenjave, je zato verjetnost, da zadane avtomobil, 2/3, če zamenjavo zavrne, pa je verjetnost pol manjša, tj. 1/3. Če se torej drži strategije zamenjave vrat, ko mu jo voditelj ponudi, bo tako vedno, ko v prvo izbere kozo, ob zamenjavi vrat dobil avto, kar ga do dobitka pripelje v 2× večjem številu primerov, kot sicer. Verjetnost za zadetek se mu tako s 33% poveča na 66%. Če vam ni takoj jasno, se ne sekirajte preveč. Tudi mnogi matematiki so potrebovali kar nekaj časa, da so si razjasnili ta problem.

3.2 Definicija pogojne verjetnosti

Opazujemo dogodek A ob poskusu X, ki je realizacija kompleksa pogojev K. Verjetnost dogodka A je tedaj $P(A)$. Kompleksu pogojev K pridružimo mogoč dogodek B, tj. $P(B) > 0$. Realizacija tega kompleksa pogojev $K' = K \cap B$ je poskus X' in verjetnost dogodka A v tem poskusu je $P_B(A)$, ki se z verjetnostjo $P(A)$ ujema ali pa ne. Pravimo, da je poskus X' poskus X s pogojem B in verjetnost $P_B(A)$ **pogojna verjetnost** dogodka A glede na dogodek B, kar zapišemo takole:

$$P_B(A) = P(A/B).$$

Pogojna verjetnost $P(A/B)$. (Opozorilo: A/B ne mešati z razliko množic $A\backslash B$.) v poskusu X' je verjetnost dogodka A v poskusu X s pogojem B. Je torej tudi verjetnost (podobno kot npr. je podgrupa grupe zopet grupa), le obravnavani kompleks pogojev, ki mora biti izpolnjen se je spremenil. Pogosto pogojno verjetnost pišejo tudi $P(A/B)$ (oziroma $P(A|B)$). Denimo, da smo n–krat ponovili poskus X in da se je ob tem k_B–krat zgodil dogodek B. To pomeni, da smo v n ponovitvah poskusa X napravili k_B–krat poskus X'. Dogodek A se je zgodil ob poskusu X' le, če se je zgodil tudi B, tj. AB. Denimo, da se je dogodek AB zgodil ob ponovitvi poskusa k_{AB}–krat. Potem je relativna frekvenca dogodka A v opravljenih ponovitvah poskusa X':

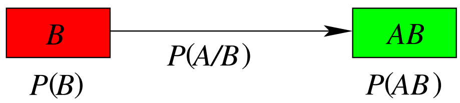
$$f_B(A) = f(A/B) = \frac{k_{AB}}{k_B} = \frac{k_{AB}/n}{k_B/n} = \frac{f(AB)}{f(B)}$$

in smo prišli do naslednje trditve.

Trditev 3.1. *Za dogodka A in B, kjer je $P(B) \neq 0$, velja*

$$P(A/B) = \frac{P(AB)}{P(B)}. \qquad \square$$

Pogojna verjetnost P_B ima prav take lastnosti kot 'brezpogojna'. Trojica $(B, \mathcal{D}_B, P_B)$, $\mathcal{D}_B = \{AB \mid A \in \mathcal{D}\}$ je zopet verjetnostni prostor.



Slika 3.3: Zvezo iz Trditve 3.1) lahko predstavimo grafično na naslednji način: verjetnost dogodka AB (desna škatla) je enaka produktu verjetnosti dogodka B (leva škatla) in pogojni verjetnosti dogodka A glede na dogodek B. Slednjo verjetnost zato postavimo ob povezavo med tema škatlama.

Primer: Denimo, da je v nekem naselju 900 polnoletnih prebivalcev. Zanima nas struktura prebivalcev po spolu (M – moški, Ž – ženski spol) in po zaposlenosti (Z – zaposlen(a), N – nezaposlen(a)). Podatke po obeh spremenljivkah uredimo v dvorazsežno frekvenčno porazdelitev, ki jo imenujemo tudi **kontingenčna tabela**:

spol \ *zap.*	Z	N	
M	460	40	500
Ž	240	160	400
	700	200	900

Poglejmo, kolikšna je verjetnost, da bo slučajno izbrana oseba moški pri pogoju, da je zaposlena.

$$P(Z) = \frac{700}{900}, \quad P(M\,Z) = \frac{460}{900}, \quad P(M/Z) = \frac{P(M\,Z)}{P(Z)} = \frac{460 \times 900}{900 \times 700} = \frac{460}{700}$$

ali neposredno iz kontingenčne tabele

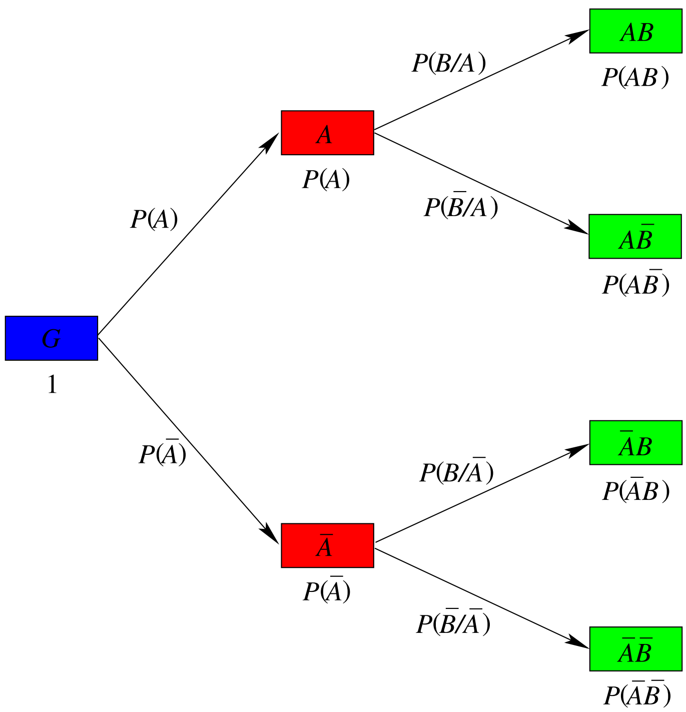
$$P(M/Z) = \frac{460}{700}. \qquad \diamond$$

Iz formule za pogojno verjetnost sledita naslednji zvezi:

$$P(A\,B) = P(B)\,P(A/B) \qquad \text{in} \qquad P(A\,B) = P(A)\,P(B/A). \tag{3.1}$$

(Drugo oz. 'dualno' enakost smo dobili iz prve tako, da smo v prvi zamenjali vlogi dogodkov A in B, seveda pa zanjo potrebujemo še pogoj $P(A) \neq 0$.) Torej velja:

$$P(A)\,P(B/A) = P(B)P(A/B). \tag{3.2}$$



Slika 3.4: Verjetnost vsakega izmed dogodkov $A\,B$, $A\,\overline{B}$, $\overline{A}\,B$ in $\overline{A}\,\overline{B}$ je enaka produktu verjetnosti na puščicah od začetka (koren na levi) pa do samega dogodka, kar nam zagotavlja desna identiteta iz (3.1). Primerjaj te dogodke s polji kontingenčne tabele 2×2. Le-ti sestavljajo popoln sistem dogodkov, vsota prvega in tretjega pa je ravno dogodek B.

Dogodka A in B sta **neodvisna**, če velja

$$P(A\,B) = P(A) \cdot P(B).$$

Zato za neodvisna dogodka A in B velja $P(A/B) = P(A)$. Za par nezdružljivih dogodkov A in B pa velja $P(A/B) = 0$.

Primer: Iz posode, v kateri imamo 8 belih in 2 rdeči krogli, $2\times$ na slepo izberemo po eno kroglo. Kolikšna je verjetnost dogodka, da je prva krogla bela (B_1) in druga rdeča (R_2)?

1. Če po prvem izbiranju izvlečene kroglo ne vrnemo v posodo (odvisnost), je:

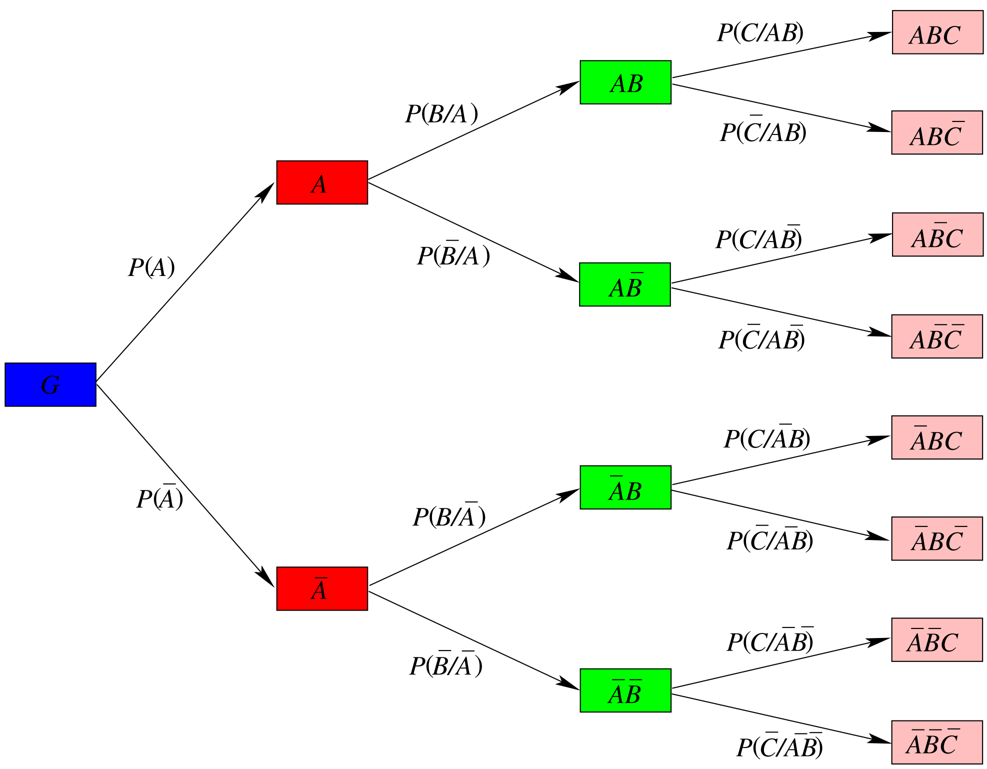
$$P(B_1\,R_2) = P(B_1) \cdot P(R_2/B_1) = \frac{8}{10} \times \frac{2}{9} = 0{\cdot}18.$$

2. Če po prvem izbiranju izvlečeno kroglo vrnemo v posodo (neodvisnost), je:

$$P(B_1\,R_2) = P(B_1) \cdot P(R_2/B_1) = P(B_1) \cdot P(R_2) = \frac{8}{10} \cdot \frac{2}{10} = 0{\cdot}16. \qquad \diamond$$

Trditev 3.2. *Dogodka A in B sta neodvisna, če je $P(A/B) = P(A/\overline{B})$. Nadalje velja*

$$P(A\,B\,C) = P(A) \cdot P(B/A) \cdot P(C/(A\,B)). \qquad \Box$$



Slika 3.5: Binarno drevo, ki nas pripelje do vseh osmih produktov med tremi dogodki A, B ter C in njihovimi nasprotnimi dogodki $\overline{A}$, $\overline{B}$ in $\overline{C}$, pri čemer mora nastopati v produktu natanko en izmed nasprotnih dogodkov X in $\overline{X}$ za vsak $X \in \{A, B, C\}$.

Dogodki A_i, $i \in I$ so **neodvisni**, če je

$$P\left(\prod_{i\in I} A_i\right) = \prod_{i\in I} P(A_i).$$

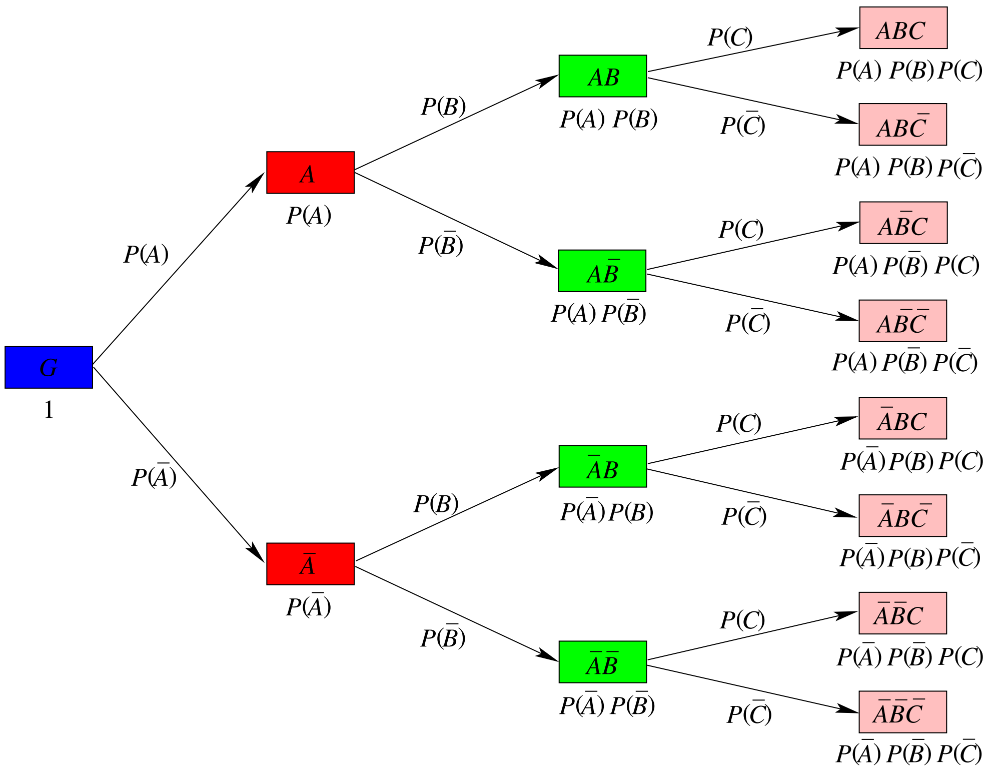
Pomembno je poudariti razliko med *nezdružljivostjo* in *neodvisnostjo*, zato za šalo zopet "recitirajmo":

> "**verjetnost** *produkta* paroma neodvisnih dogodkov
> je enaka
> *produktu* **verjetnosti** teh dogodkov."

$$\begin{tikzcd} (A,B) \arrow[r, "*"] \arrow[d, "\big(P(.),P(.)\big)"'] & A\,B \arrow[d, "P(.)"] \\ \big(P(A),P(B)\big) \arrow[r, "*"] & P(A)P(B) = P(AB) \end{tikzcd}$$

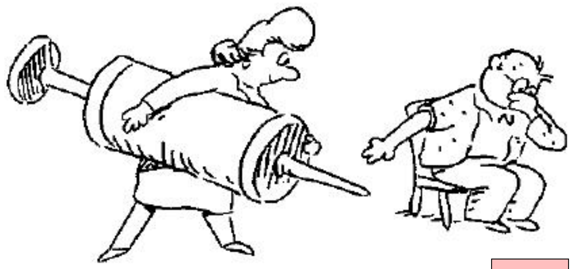
Torej gre za nekakšno pravilo o zamenjavi (med vrstnim redom računanja produkta in verjetnosti), ki velja za paroma neodvisne dogodke.

Za neodvisne dogodke A_i, $i \in I$ velja $P(A_j) = P(A_j / \prod_{i=1}^{j-1} A_i)$ za vsak $j \in I$.



Slika 3.6: Binarno drevo za tri neodvisne dogodke A, B in C.

Primer: Redko nalezljivo bolezen dobi ena oseba na 1 000.
Imamo dober, a ne popoln test za to bolezen:

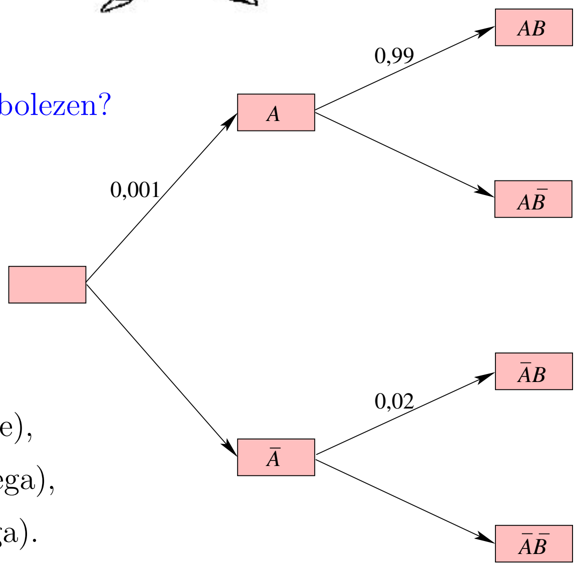


če ima neka oseba to bolezen,
potem test to pokaže v 99% primerih,

vendar pa test napačno označi tudi
2% zdravih pacientov za bolane.

V tvojem primeru je bil test pravkar **pozitiven**.
Kakšna je verjetnost, da si zares dobili nalezljivo bolezen?

Delamo z naslednjimi dogodki:



$\boldsymbol{A}$: pacient je dobil nalezljivo bolezen,

$\boldsymbol{B}$: pacientov test je bil pozitiven.

Izrazimo informacijo o učinkovitosti testov:

$\boldsymbol{P(A) = 0{\cdot}001}$ (en pacient na 1 000 se naleze),

$\boldsymbol{P(B/A) = 0{\cdot}99}$ (test pravilno označi okuženega),

$\boldsymbol{P(B/\overline{A}) = 0{\cdot}02}$ (test napačno označi zdravega).

Zanima nas $\boldsymbol{P(A/B)}$ (verjetnost, da smo se nalezli, če je test pozitiven).

1. način (algebraični pristop): Iz (3.2) dobimo:

$$P(A/B) = \frac{P(A)P(B/A)}{P(B)},$$

od koder je razvidno, da potrebujemo 'le še' $P(B)$. Spomnimo se popolnega sistema dogodkov, ki ga v našem primeru lahko predstavljata nasprotna dogodka A in $\overline{A}$. Potem velja:

$$P(B) = P(BG) = P(B(A + \overline{A})) = P(BA) + P(B\overline{A})$$

(zadnji enačaj valja zato, ker sta dogodka BA in $B\overline{A}$ nezdružljiva). Ali bi znali izračunati $P(BA)$? Pomagamo si lahko s formulo za pogojno verjetnost (ne morda prvo, ki nam pride na misel, pač pa tisto drugo, tj. 'dualno', glej (3.1)):

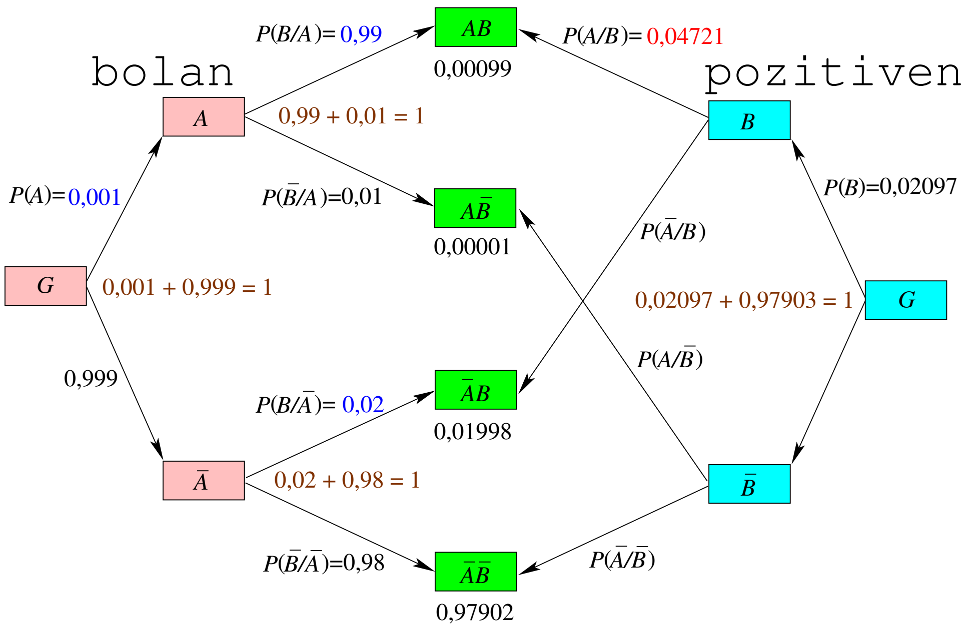
$$P(BA) = P(A)P(B/A) = 0{\cdot}001 \times 0{\cdot}99 = 0{\cdot}00099.$$

Spomnimo se tudi, da je $P(\overline{A}) = 1 - P(A) = 0{\cdot}999$ in že na enak način kot v prejšnjem primeru, ko smo računali $P(BA)$, dobimo:

$$P(B\overline{A}) = P(\overline{A})P(B/\overline{A}) = 0{\cdot}999 \times 0{\cdot}02 = 0{\cdot}01998,$$

kar nam da $P(B) = 0{\cdot}00099 + 0{\cdot}001998 = 0{\cdot}02097$ in $P(A/B) = 0{\cdot}00099/0{\cdot}02097 = 0{\cdot}04721$.

2. način (grafični pristop): Tokrat si pomagamo z binarnim drevesom na prejšnji strani. Na začetku imamo možnost, da se zgodi bodisi dogodek A ali pa njegov nasproten dogodek $\overline{A}$. Če poznamo verjetnost od enega od teh dveh dogodkov, potem po Posledici 2.6 poznamo tudi drugo verjetnost. Enako sklepamo tudi na naslednjem nivoju (z leve proti desni) in že imamo verjetnosti na vseh puščicah na naslednjem nivoju. Verjetnost vsakega izmed dogodkov v zelenih škatlah! je enaka produktu verjetnosti na puščicah od začetka (koren na levi) pa do samega dogodka, kar nam zagotavlja desna identiteta iz (3.1). Glej sliko 3.9.



Slika 3.9: Popoln sistem dogodkov $A\,B$, $A\,\overline{B}$, $\overline{A}\,B$ in $\overline{A}\,\overline{B}$ je označen z zeleno, njihove verjetnosti pa so $P(A\,B) = 0{\cdot}001 \times 0{\cdot}99 = 0{\cdot}00099$, $P(A\,\overline{B}) = 0{\cdot}001 \times 0{\cdot}01 = 0{\cdot}00001$, $P(\overline{A}\,B) = 0{\cdot}999 \times 0{\cdot}02 = 0{\cdot}01998$ in $P(\overline{A}\,\overline{B}) = 0{\cdot}999 \times 0{\cdot}98 = 0{\cdot}97902$. in so zapisane pod njihove škatle. Za vsak slučaj lahko preverimo, da je vsota pravkar izračunanih verjetnosti res enaka 1: $0{\cdot}00099 + 0{\cdot}00001 + 0{\cdot}01998 + 0{\cdot}97902 = 1$.

Verjetnost dogodka B je torej enaka $0{\cdot}00099 + 0{\cdot}01998 = 0{\cdot}02097$. Končno si pogledamo še pot, ki nas pripelje iz desnega korena do dogodka AB in upoštevamo še levo identiteto iz (3.1): $0{\cdot}00099/0{\cdot}02097 = 0{\cdot}04721$. Za konec pa naredimo še en preiskus:
$0{\cdot}02097 \times (1 - 0{\cdot}00099/0{\cdot}02097) = 0{\cdot}01998$, a premislite sami, kaj smo v resnici preverili. $\diamond$

3.3 Obrazec za popolno verjetnost in večstopenjski poskusi

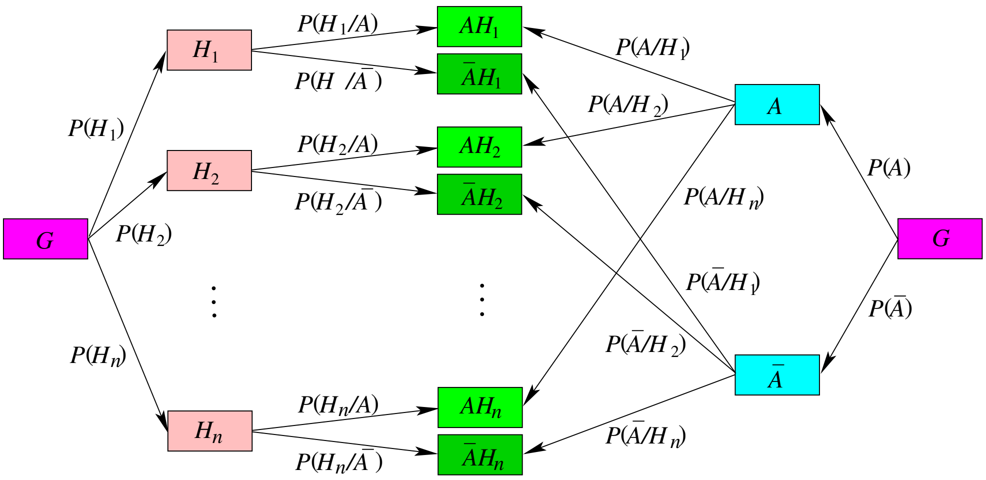
Naj bo H_i, $i \in I$ popoln sistem dogodkov, tj. **razbitje** gotovega dogodka: $\sum_{i\in I} H_i = G$, na paroma nezdružljive dogodke: $H_i\,H_j = N, i \neq j$. Gotov dogodek smo torej kot hlebec narezali s pomočjo domnev na posamezne kose, da jih bomo lažje obvladali. Zanima nas verjetnost dogodka A, če poznamo verjetnost $P(H_i)$, in pogojno verjetnost $P(A/H_i)$ za $i \in I$:

$$A = A\,(H_1 + H_2 + \cdots + H_n) = (A\,H_1) + \cdots + (A\,H_n).$$

Sedaj si bomo ogledali verjetnost dogodka A na posameznih 'kosih', skoraj tako kot pri marmornem kolaču. Ker so tudi dogodki $A\,H_i$ paroma nezdružljivi, velja:

Trditev 3.3. *Za popoln sistem dogodkov H_i, $i \in I$ in poljuben dogodek A velja*

$$P(A) = \sum_{i\in I} P(A\,H_i) = \sum_{i\in I} P(H_i)P(A/H_i). \qquad \square$$



Slika 3.10: Vsak stolpec škatel predstavlja popoln sistem dogodkov. V stolpcu zelenih škatel svetlejše predstavljajo popoln sistem dogodkov za A, preostale pa popoln sistem dogodkov za $\overline{A}$.

Zgornji trditvi pravimo tudi *izrek o popolni verjetnosti* ali pa obrazec za razbitje. Na to lahko pogledamo tudi kot na večstopenjski poskus: v prvem koraku se zgodi natanko eden od dogodkov H_i, ki ga imenujemo domneva (hipoteza) (domneve sestavljajo popoln sistem dogodkov). Šele izidi na prejšnjih stopnjah določajo, kako bo potekal poskus na naslednji stopnji. Omejimo se na poskus z dvema stopnjama. Naj bo A eden izmed mogočih dogodkov na drugi stopnji. Včasih nas zanima po uspešnem izhodu tudi druge stopnje, verjetnost tega, da se je na prvi stopnji zgodil dogodek H_i. Odgovor nam da naslednja trditev.

Thomas Bayes, angleški matematik (1702-1761).

Trditev 3.4. (**Bayesov obrazec**) *Za popoln sistem dogodkov H_i, $i \in I$ in poljuben dogodek A velja*

$$P(H_k/A) = \frac{P(H_k)\cdot P(A/H_k)}{\sum_{i\in I} P(H_i)\cdot P(A/H_i)}. \qquad \square$$

Primer: Trije lovci so hkrati ustrelili na divjega prašiča in ga ubili. Ko so prišli do njega, so našli v njem eno samo kroglo. Kolikšne so verjetnosti, da je vepra ubil (a) prvi, (b) drugi, (c) tretji lovec, če poznamo njihove verjetnosti, da zadanejo: 0·2; 0·4 in 0·6?

Na ta način jim namreč lahko pomagamo pri pošteni delitvi plena (kajti ne smemo pozabiti, da imajo vsi v rokah nevarno orožje). Sestavimo popoln sistem dogodkov in uporabimo dejstvo, da so lovci med seboj neodvisni, torej $P(A\,B\,C) = P(A)\,P(B)\,P(C)$. To nam zna pomagati pri računanju verjetnosti domnev (hipotez).

```
      .2  .4  .6
    prvi drugi tretji    P(H_i)     st.kr. P(E/H_i) P(E H_i)
H1    1    1    1    ,2*,4*,6 =0,048   3      0        0
H2    0    1    1    ,8*,4*,6 =0,192   2      0        0
H3    1    0    1    ,2*,6*,6 =0,072   2      0        0
H4    1    1    0    ,2*,4*,4 =0,032   2      0        0
H5    1    0    0    ,2*,6*,4 =0,048   1      1      0,048
H6    0    1    0    ,8*,4*,4 =0,128   1      1      0,128
H7    0    0    1    ,8*,6*,6 =0,288   1      1      0,288
H8    0    0    0    ,8*,6*,4 =0,192   0      0        0
vsota                         =1,000                  0,464
```

$$P(\text{ena krogla je zadela}) = 0,048 + 0,128 + 0,288 = 0,464 = P(E).$$

Ostale verjetnosti računamo za preiskus:

$$\begin{aligned}
&P(\text{nobena krogla ni zadela}) &=&\ 0,192 = P(N'),\\
&P(\text{dve krogli sta zadeli}) = 0,192 + 0,072 + 0,032 &=&\ 0,296 = P(D),\\
&P(\text{tri krogle so zadele}) &=&\ 0,048 = P(T).
\end{aligned}$$

Vsota teh verjetnosti je seveda enaka 1. Končno uporabimo Bayesov obrazec:

$$\begin{aligned}
P(H_5/E) &= \frac{P(H_5E)}{P(E)} = \frac{0{\cdot}048}{0{\cdot}464} = 0{\cdot}103 &=&\ P(\text{prvi je zadel/E}),\\
P(H_6/E) &= \frac{P(H_6E)}{P(E)} = \frac{0{\cdot}128}{0{\cdot}464} = 0{\cdot}276 &=&\ P(\text{drugi je zadel/E}),\\
P(H_7/E) &= \frac{P(H_7E)}{P(E)} = \frac{0{\cdot}288}{0{\cdot}464} = 0{\cdot}621 &=&\ P(\text{tretji je zadel/E}).
\end{aligned}$$

Tudi vsota teh verjetnosti pa je enaka 1. Delitev plena se opravi v razmerju $10{\cdot}3 : 27{\cdot}6 : 62{\cdot}1 = 3 : 8 : 18$ (in ne $2 : 4 : 6$ oziroma $16{\cdot}6 : 33{\cdot}3 : 50$, kot bi utegnili na hitro pomisliti). ◇

Bonus vprašanje: Kako bi si pravično razdelili plen, če bi v divjim prašiču našli dve krogli?

Za konec podajmo še presenetljivo kratko rešitev Monty Hall problema z uporabo pogojne verjetnosti:

Primer: (Rešitev problema Monty Hall) Z A označimo dogodek, da dobimo avto, z V pa da na začetku izberemo prava vrata. Potem obravnavamo naslednji možnosti:

1) Si ne premislimo: $P(A) = P(V) = 1/3$

2) Si premislimo: $P(A) = P(A/V)\cdot P(V) + P(A/\overline{V})\cdot P(\overline{V}) = 0 \times 1/3 + 1 \times 2/3 = 2/3$, kjer smo pri drugi možnosti uporabili obrazec za razbitje. Je pa tudi res, da se morata ti dve verjetnosti sešteti v 1 in da nam zato druge v resnici sploh ni treba računati s pomočjo obrazca za razbitje. Vendar smo tako vseeno bolj prepričani o pravilnosti našega sklepanja, kajti kot smo spoznali v sami zgodbi, se je kar veliko ljudi motilo pri tem problemu. ◇

Poglavje 4

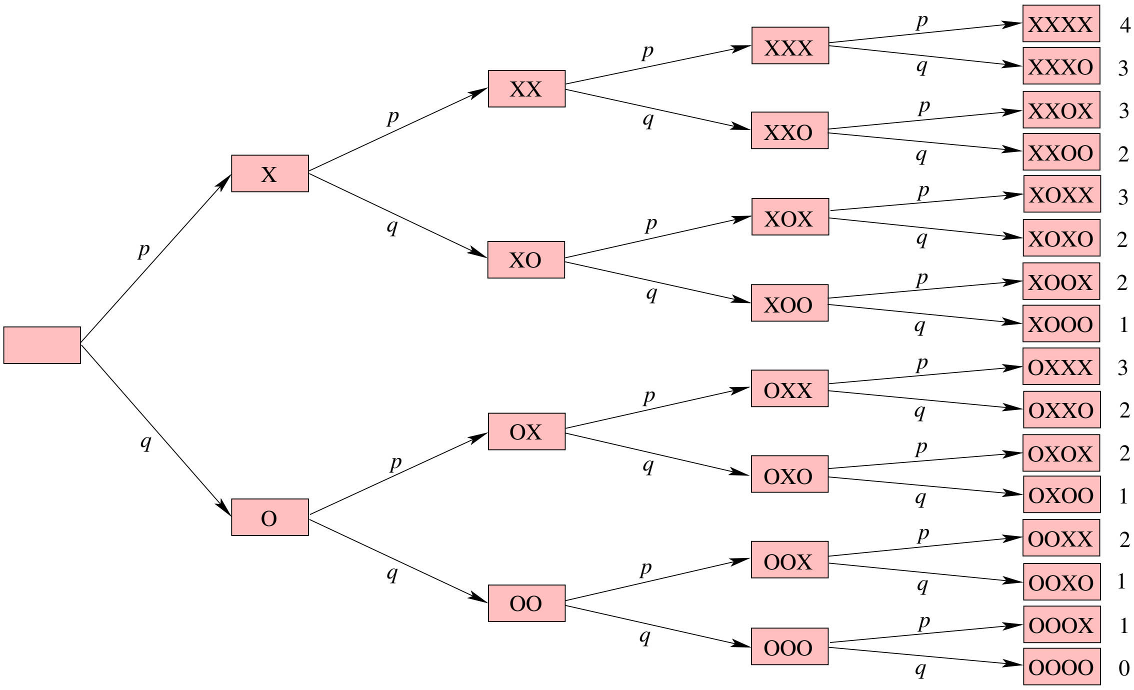
Bernoullijevo zaporedje neodvisnih poskusov

O zaporedju neodvisnih poskusov $X_1, X_2, \ldots, X_n, \ldots$ govorimo tedaj, ko so verjetnosti izidov v enem poskusu neodvisne od tega, kaj se zgodi v drugih poskusih.

Zaporedje neodvisnih poskusov se imenuje **Bernoullijevo zaporedje**, *če se more zgoditi v vsakem poskusu iz zaporedja neodvisnih poskusov le dogodek A z verjetnostjo $P(A) = p$ ali dogodek $\overline{A}$ z verjetnostjo $P(\overline{A}) = 1 - P(A) = 1 - p = q$.*

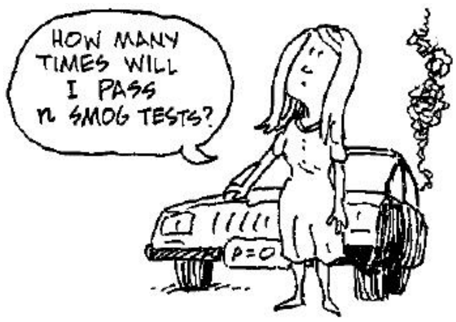
Jakob Bernoulli um 1687

Primer: Primer Bernoullijevega zaporedja poskusov je met kocke, kjer ob vsaki ponovitvi poskusa pade šestica (dogodek A) z verjetnostjo $P(A) = p = 1/6$ *ali* ne pade šestica (dogodek $\overline{A}$) z verjetnostjo $P(\overline{A}) = 1 - p = q = 5/6$. $\diamond$



Slika 4.3: Iz binarnega drevesa je očitno, da je število možnih izidov 2^n, če se seveda ustavimo po n zaporednih poskusih. Verjetnost, da pridemo do določenega vozla v drevesu je seveda enak produktu verjetnosti na povezavah od korena (začetka) pa vse do tega vozla.

V Bernoullijevem zaporedju neodvisnih poskusov nas zanima, kolikšna je verjetnost, da se v n zaporednih poskusih zgodi dogodek A natanko k–krat. To se lahko zgodi na primer tako, da se najprej zgodi k–krat dogodek A in nato v preostalih $(n-k)$ poskusih zgodi nasprotni dogodek $\overline{A}$:



$$P\left(\prod_{i=1}^{k}(X_i = A) \prod_{i=k+1}^{n}(X_i = \overline{A})\right) = \prod_{i=1}^{k} P(A) \cdot \prod_{i=k+1}^{n} P(\overline{A}) = p^k \cdot q^{n-k}.$$

Dogodek $P_n(k)$, da se dogodek A v n zaporednih poskusih zgodi natanko k–krat, se lahko zgodi tudi na druge načine in sicer je teh toliko, na kolikor načinov lahko izberemo k poskusov iz n poskusov. Teh je $\binom{n}{k}$. In ker je $\binom{n}{0} + \binom{n}{1} + \cdots \binom{n}{n} = 2^n$ nismo prav nobenega pozabili. Ker so ti načini nezdružljivi med seboj, je verjetnost dogodka $P_n(k)$ enaka

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.1}$$

Tej zvezi pravimo **Bernoullijev obrazec**.

Primer: Iz posode, v kateri imamo 8 belih in 2 rdeči krogli, na slepo izberemo po eno kroglo in po izbiranju izvlečeno kroglo vrnemo v posodo. Kolikšna je verjetnost, da v petih poskusih izberemo 3–krat belo kroglo? Dogodek A je, da izvlečem belo kroglo. Potem je

$$p = P(A) = \frac{8}{10} = 0{\cdot}8 \qquad \text{in} \qquad q = 1 - p = 1 - 0{\cdot}8 = 0{\cdot}2$$

Verjetnost, da v petih poskusih izberemo 3–krat belo kroglo, je:

$$P_5(3) = \binom{5}{3} 0{\cdot}8^3 (1 - 0{\cdot}8)^{5-3} = 0{\cdot}205. \qquad \diamond$$

4.1 Računanje verjetnosti $P_n(k)$

Ena možnost je uporaba rekurzije:

Trditev 4.1. *$P_n(0) = q^n$, $\qquad P_n(k) = \frac{(n-k+1)p}{kq} P_n(k-1) \qquad$ za $k = 1, \ldots n$.*

Dokaz.

$$\frac{P_n(k)}{P_n(k-1)} = \frac{\binom{n}{k} p^k q^{n-k}}{\binom{n}{k-1} p^{k-1} q^{n-k+1}} = \frac{n!\ (k-1)!(n-k+1)!\ p}{k!(n-k)!\ n!\ q} = \frac{(n-k+1)p}{kq}. \qquad \square$$

Vendar pa je takšno računanje izredno zamudno (eksponentno), tako kot rekurzivno računanje faktorjela, glej razdelek A.6.

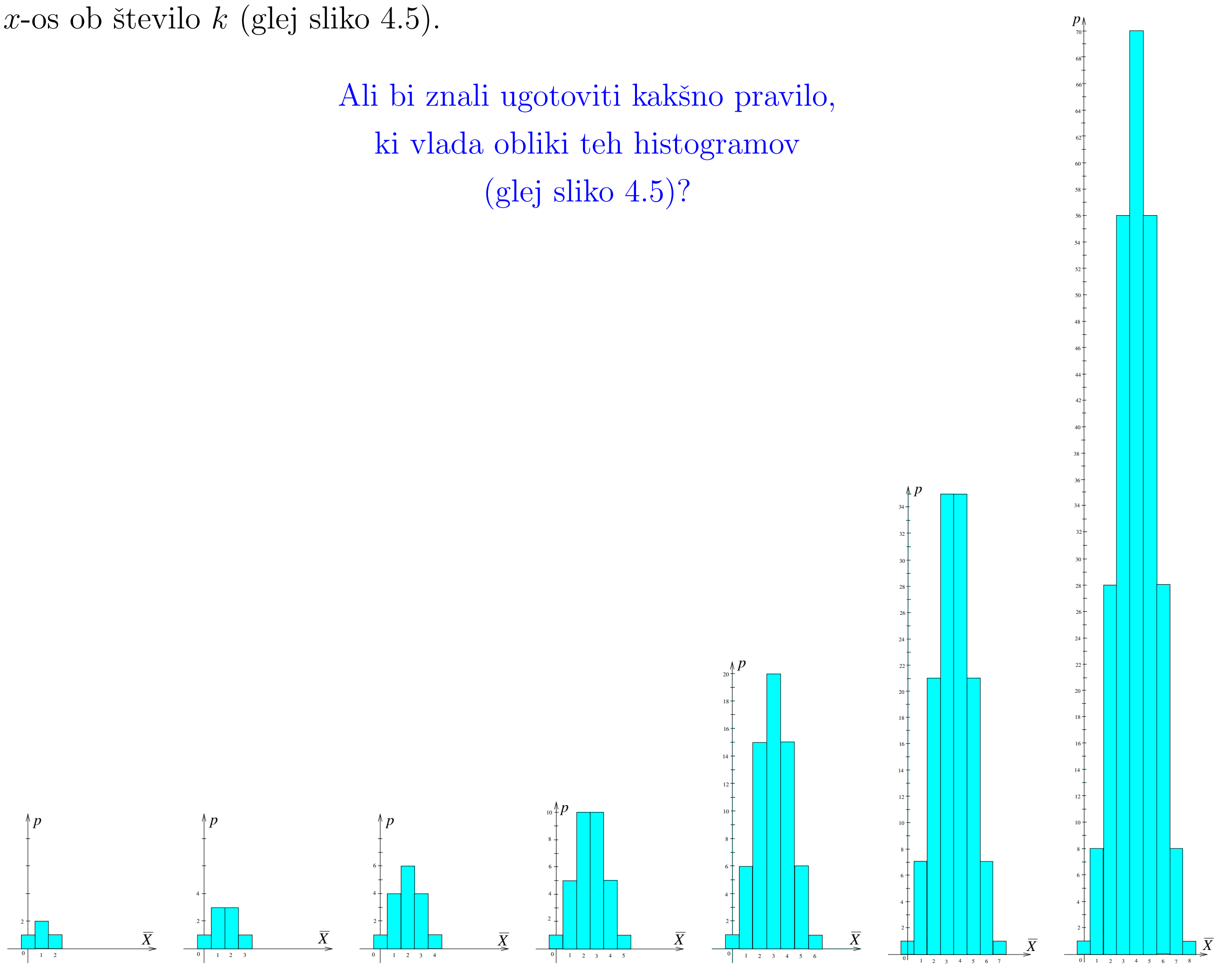
Program R: Vrednost $P_n(k)$ dobimo z ukazom `dbinom(`k`,size=`n`,prob=`p`)`

```
> dbinom(50,size=1000,prob=0.05)
[1] 0.05778798
```

4.2 Usojena krivulja v Pascalovemu trikotniku

Iz vsake vrstice Pascalovega trikotnika lahko narišemo histogram, tj. vsakemu binomskemu simbolu $\binom{n}{k}$ za $k = 0, 1, \ldots, n$, priredimo stolpec velikosti $\binom{n}{k} \times 1$ in ga postavimo (pokončno) na x-os ob število k (glej sliko 4.5).

Ali bi znali ugotoviti kakšno pravilo,
ki vlada obliki teh histogramov
(glej sliko 4.5)?



Slika 4.5: Predstavitev 2., 3., ..., 7. in 8. vrstice Pascalovega trikotnika v obliki histograma. Vsota višin stolpcev v posameznem histogramu je zaporedomo 4, 8, 16, 32, 64, 128 in 256, višina najvišjega stolpca pa je 2, 3, 6, ...

Eno je gotovo, višina histograma vrtoglavo raste in zato smo se ustavili že pri $n = 8$. Vsekakor gre za 'enogrbo kamelo', simetrično glede na $n/2$, tj. števila najprej rastejo, nato se pri lihih n največje število enkrat ponovi in nato (simetrično) padajo. Vsota vseh stolpcev je seveda enaka 2^n, saj po binomskem obrazcu velja:

$$2^n = (1+1)^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-2} + \binom{n}{n-1} + \binom{n}{n}.$$

Uvedemo nekaj dopolnil, ki bodo izboljšala naše slike, in bomo lahko z malo sreče v sliko spravili še kakšno vrstico Pascalovega trikotnika. Lahko se odločimo, da bo

- višina najvišjega stolpca v vsakem histogramu enaka,
- vsota ploščin vseh stolpcev v vsakem histogramu pa prav tako.

Drugo zahtevo hitro dosežemo tako, da za višino k-tega stolpca v histogramu uporabimo

$$\text{namesto} \quad \binom{n}{k} \qquad \text{raje} \quad \binom{n}{k}/2^n,$$

Slednje število je ravno $P_n(k)$ iz Bernoullijevega obrazca za $p = q = 1/2$. Vendar pa sedaj opazimo, da se višina najvišjega stolpca (ki ga dobimo za $k = \lfloor n/2 \rfloor$) z n počasi niža:

$$0{\cdot}500, 0{\cdot}500, 0{\cdot}375, 0{\cdot}375, 0{\cdot}313, 0{\cdot}313, 0{\cdot}273, 0{\cdot}273, 0{\cdot}246, 0{\cdot}246, \ldots$$

Prepričaj se, da je ponavljanje zgornjih vrednosti posledica lastnosti binomskega simbola. Zaradi tega opazujmo le še vrednosti, ki smo jih dobili za lihe n. Kako hitro pada višina najvišjega stolpca? Vsekakor ne linearno, saj se razlike med zaporednimi členi manjšajo. Potrebno bo torej najti funkcijo, ki pada počasneje. Poskusimo s funkcijo $\sqrt{n}$, ki pa jo zaradi njenega naraščanja obrnemo v padajočo funkcijo $1/\sqrt{n}$. Izkaže se, da smo imeli srečo, vsaj sodeč po zaporedju, ki smo ga dobimo iz prejšnjega z množenjem s $\sqrt{n}$:

$$0{\cdot}707, 0{\cdot}750, 0{\cdot}765, 0{\cdot}773, 0{\cdot}778, 0{\cdot}781, 0{\cdot}784, 0{\cdot}786, 0{\cdot}787, 0{\cdot}788, 0{\cdot}789, 0{\cdot}790, 0{\cdot}790, 0{\cdot}791, 0{\cdot}791, \ldots \tag{4.2}$$

Abraham de Moivre bi nam seveda pritrdil, da smo na pravi poti (saj je nekaj podobnega počel že leta 1733 – **The Doctrine of Chance**).

Limitno vrednost števila iz (4.2) poznamo samo na kakšno decimalko natančno, (če limita sploh obstaja), vendar pa bi ga radi natančno določili (če limita sploh obstaja seveda). Gre v resnici morda za $0{\cdot}8 = 4/5$? Predno se dodobra zamislimo, nam neutruden računalnik zaupa, da se po nekaj tisoč korakih pride do $0{\cdot}7979$ ali, če smo še bolj potrpežljivi do $0{\cdot}7978646\ldots$. Pa nadomestimo število z njegovim kvadratom oziroma s kvadratom recipročne vrednosti, da bo število večje od ena:

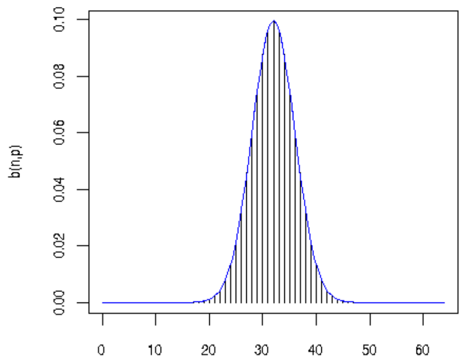
$$1{\cdot}57158.$$

To število je sumljivo blizu $\pi/2$. Tako blizu, da smo pripravljeni tvegati in postaviti domnevo:

$$\lim_{n\to\infty} \binom{n}{k} \frac{\sqrt{n\pi/2}}{2^n} = 1, \qquad \text{za } k = \lfloor n/2 \rfloor.$$

Sedaj pa bi nam pritrdil celo Stirling, kajti s pomočjo njegove (študentom dobro znane) ocene za faktoriel $n! \approx \sqrt{2\pi n}\,(n/e)^n$, kjer je e Eulerjeva konstanta ($e = 2{\cdot}71\ldots$), bi prišli do enakega zaključka (ki nam pravzaprav ponuja tudi hitro matematično utemeljitev zgornje domneve).

Toliko smo se ukvarjali z višino v histogramih (ki se sedaj v limiti približuje neki konstanti), da smo skoraj pozabili na ploščino, da o širini niti ne govorimo. V resnici lahko vse stolpce postavimo enega nad drugega in v tem primeru dobimo en sam stolpec, tj. pravokotnik. Želeli smo, da bi bila njegova ploščina enaka 1, vendar sedaj delimo vse višine stolpcev v histogramu namesto z 2^n le še s številom $2^n/(c\sqrt{n})$, kjer je c poljubna pozitivna konstanta, tj. po deljenju z 2^n še *množimo* s $c\sqrt{n}$. Zato je potrebno širino stolpcev *deliti* s $c\sqrt{n}$. Ker višine stolpcev v zadnjem koraku nismo spreminjali, smo končno izpolnili obe zahtevi (konstantna ploščina in konstantna maksimalna višina histograma).



Slika 4.7: Histogram za $n = 64$. Stolpce smo nadomestili kar z intervali.

Čeprav še nismo razmišljali o širini histograma, (le-ta vsebuje $n + 1$ stolpcev in je sedaj enaka $(n+1)/(c\sqrt{n})$, torej gre z rastočim n proti neskončno), zgornja slika kaže, da z njo ne bo težav, višine stolpcev so namreč na robu že zelo majhne (blizu 0) in ne bodo kaj dosti prispevale k obliki. Če želimo še bolje spoznati iskano obliko, je na vrsti študij *konveksnosti*, tj. kdaj moramo zavijati z avtom v desno ozirma v levo, če se vozimo po krivulji z leve proti desni. Pri tem si pomagamo z opazovanjem spreminjanja predznaka razlike dveh zaporednih binomskih simbolov. Naj bo

$$d_h := \binom{n}{h+1} - \binom{n}{h} \qquad \text{za} \quad h = 0, 1, \ldots, n-1.$$

Za $h = 0$ (oziroma $h = n - 1$) in $h = 1$ (oziroma $h = n - 2$) velja

$$d_0 = n = -d_{n-1} \qquad \text{in} \qquad d_1 = n(n-3)/2 = -d_{n-2}.$$

Zaradi simetrije binomskih simbolov, glede na $n/2$, tj. $\binom{n}{k} = \binom{n}{n-k}$, velja tudi

$$d_i = -d_{n-1-i}, \qquad 0 \leqslant i \leqslant n-1,$$

zato se bomo omejili le na območje za katerega velja $h \leqslant \lfloor (n-1)/2 \rfloor$. Naj bo $m \in \mathbb{N}$ in $n = 2m$ oziroma $n = 2m - 1$ glede na to ali je n sodo oziroma liho število. Potem je $\lfloor (n+1)/2 \rfloor = m$ in $d_m = 0$, če je n lih in $d_{m-1} = -d_m$, če je n sod. S slike oziroma iz pravkar navedenih lastnosti je očitno, da je potrebno na začetku zaviti v desno, potem pa bomo gotovo morali začeti zavijati še v levo, vsaj tik pred vrhom. Naslednja trditev nam zagotovi, da se na poti do vrha spremeni smer (zavijanja) le enkrat (iz leve v desno), zaradi simetrije pa potem enako velja tudi pri poti 'navzdol' (tokrat iz desne v levo).

Trditev 4.2. *Za* $h \in \{1, \dots, \lfloor (n-1)/2 \rfloor\}$ *velja*

$$d_{h-1} \leqslant d_h \quad \textit{natanko tedaj, ko velja} \quad h \leqslant \frac{n}{2} - \frac{\sqrt{n+2}}{2}$$

oziroma ekvivalentno

$$d_0 < d_1 < \cdots < d_{k-1} \leqslant d_k > d_{k+1} > d_{k+2} > \cdots > d_{\lfloor (n-1)/2 \rfloor},$$

kjer je $k := \lfloor (n - \sqrt{n+2})/2 \rfloor$. *Enačaj velja natanko tedaj, ko je* $n+2$ *popoln kvadrat in je* $h = k$.

Dokaz. Iz

$$d_h = \frac{n!}{(h+1)!(n-h-1)!} - \frac{n!}{h!(n-h)!} = \frac{n(n-2h-1)}{(h+1)!(n-h)!},$$

lahko zaključimo, da je predznak razlike

$$d_h - d_{h-1} = \frac{n!(n-2h-1)}{(h+1)!(n-h)!} - \frac{n!(n-2h+1)}{h!(n-h+1)!} = \frac{n!}{(h+1)!(n-h+1)!}\,\Delta$$

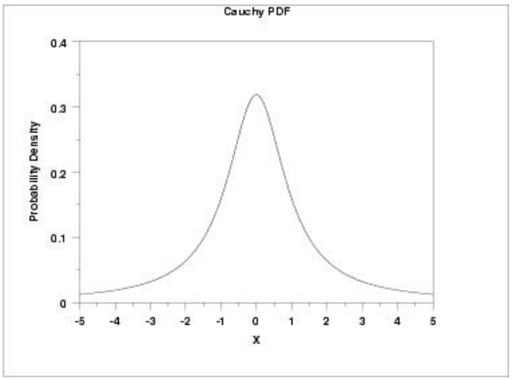
odvisna samo od predznaka razlike

$$\Delta := (n-2h-1)(n-h+1) - (n-2h+1)(h+1) = (n-2h)^2 - (n+2). \qquad \square$$

Zopet smo pred novim presenečenjem.

Vse popravljenje krivulje so si sedaj na moč podobne!

Za katero krivuljo pa gre? Če smo prej našli konstanto c, potem se moramo sedaj nekako podvizati in odkriti še skrivnostno krivuljo. Morda bi kdo najprej pomislil na **parabolo**, a ne bo prava, saj naša krivulja nikoli ne preseka x osi, temveč se ji samo asimptotično približuje. Iz zgornje trditve tudi sledi, da se oblika histogramov vsekakor bistveno razlikuje od oblike parabole, saj se na slednji sploh ne spremeni smer zavijanja. Kako pa je z obliko funkcije $\boldsymbol{\cos x}$ **(na intervalu** $[-\pi/2, \pi/2]$)? Na prvi pogled izgleda prava: tudi na njej moramo zaviti najprej v levo, nato v desno. V tem primeru pa je ključnega pomena to, da preidemo iz desnega zavijanja v levo zavijanje pri kosinusu ravno na sredi (od vznožja do vrha), v primeru binomskih simbolov pa se razdalja z večanjem n približuje $n/2$. Tretji poskus s funkcijo $\boldsymbol{1/(1+x^2)}$ prepustimo bralcu (tudi ta ni prava, a je matematična utemeljitev nekoliko bolj zapletena).



Slika 4.8: Graf funkcije $1/(\pi(1+x^2))$, kjer smo pod ulomkovo črto dodali še π, da bo ploščina pod krivuljo enaka 1.

V resnici naša funkcija ni niti racionalna (kvocient dveh polinomov). Morda pa je usoda hotela, da na ta način spoznamo povsem novo funkcijo. Zaenkrat jo poimenujmo kar **usojena** funkcija (da ne bomo pretiravali s tujkami - **fatalka**).

Sledimo Abrahamu de Moivru (The Doctrine of Chance, 1733), ki predlaga, da vpeljemo

$$x = \frac{k-(n/2)}{\sqrt{n}} \qquad \text{oziroma} \qquad k = x\sqrt{n} + (n/2)$$

(le ta zamenjava postavi vrh krivulje na y-os in normalizira razdaljo do prevoja) in izračunajmo

$$f(x) = \lim_{n\to\infty} \frac{\sqrt{n}}{2^n} \binom{n}{x\sqrt{n}+(n/2)} = \lim_{n\to\infty} \frac{\sqrt{n}}{2^n} \frac{n!}{\big((n/2)+x\sqrt{n}\big)!\ \big((n/2)-x\sqrt{n}\big)!}.$$

Z nekaj računske spretnosti in uporabo vrste za $\ln(1+x)$ (glej škatlo) pridemo do

$$f(x) = f(0)\ e^{-2x^2} = \sqrt{\frac{2}{\pi}}\ e^{-2x^2} \qquad \text{oziroma} \qquad N(t) = \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}}, \qquad \text{za}\ \ N(t) = f(t/2)/2.$$

Izpeljali smo naslednjo trditev.

Izrek 4.3. (**De Moivrov točkovni obrazec**)
Za velike n velja

$$P_n(k) \approx \frac{1}{\sqrt{\pi n/2}}\ e^{-\frac{(k-n/2)^2}{n/2}}.$$

De Moivrov točkovni obrazec je poseben primer **Laplaceovega točkovnega obrazca**. Slednjega smemo uporabljati, ko je n velik in p blizu $1/2$:

$$P_n(k) \approx \frac{1}{\sqrt{2\pi npq}}\ e^{-\frac{(k-np)^2}{2npq}}.$$

Skica dokaza De Moivrovega točkovnega obrazca

Vpeljimo zamenjavo $n = 4m^2$ (da se znebimo korenov in ulomkov) in si pobliže oglejmo kvocient

$$\frac{f(x)}{f(0)} = \lim_{n\to\infty} \frac{(n/2)!(n/2)!}{\big((n/2)+x\sqrt{n}\big)!\,\big((n/2)-x\sqrt{n}\big)!} = \lim_{m\to\infty} \frac{(2m^2)!\,(2m^2)!}{(2m^2+2mx)!\,(2m^2-2mx)!}.$$

Okrajšamo, kar se okrajšati da, in dobimo

$$\frac{f(x)}{f(0)} = \lim_{m\to\infty} \frac{(2m^2)(2m^2-1)\cdots(2m^2-2mx+2)(2m^2-2mx+1)}{(2m^2+1)(2m^2+2)\cdots(2m^2+2mx-1)(2m^2+2mx)}.$$

Opazimo, da se istoležeči faktorji, ki ležijo en nad drugim seštejejo v $4m^2+1$ in preoblikujmo dobljeni kvocient v naslednjo obliko:

$$\frac{f(x)}{f(0)} = \lim_{m\to\infty} \frac{\Big(2m^2+\frac{1}{2}-\frac{1}{2}\Big)\Big(2m^2+\frac{1}{2}-\frac{3}{2}\Big)\Big(2m^2+\frac{1}{2}-\frac{5}{2}\Big)\cdots\Big(2m^2+\frac{1}{2}-\frac{4mx-1}{2}\Big)}{\Big(2m^2+\frac{1}{2}+\frac{1}{2}\Big)\Big(2m^2+\frac{1}{2}+\frac{3}{2}\Big)\Big(2m^2+\frac{1}{2}+\frac{5}{2}\Big)\cdots\Big(2m^2+\frac{1}{2}+\frac{4mx-1}{2}\Big)}.$$

Sedaj pa delimo vsak faktor (nad in pod ulomkovo črto) z $2m^2+1/2$ in upoštevajmo, da lahko $1/2$ v limiti zanemarimo) in dobimo

$$\frac{f(x)}{f(0)} = \lim_{m\to\infty} \frac{\Big(1-\frac{1}{4m^2}\Big)\Big(1-\frac{3}{4m^2}\Big)\Big(1-\frac{5}{4m^2}\Big)\cdots\Big(1-\frac{4mx-1}{4m^2}\Big)}{\Big(1+\frac{1}{4m^2}\Big)\Big(1+\frac{3}{4m^2}\Big)\Big(1+\frac{5}{4m^2}\Big)\cdots\Big(1+\frac{4mx-1}{4m^2}\Big)}$$

oziroma

$$\begin{aligned}\ln\left(\frac{f(x)}{f(0)}\right) &= \lim_{m\to\infty} \ln\Big(1-\frac{1}{4m^2}\Big)+\ln\Big(1-\frac{3}{4m^2}\Big)+\ln\Big(1-\frac{5}{4m^2}\Big)+\cdots+\ln\Big(1-\frac{4mx-1}{4m^2}\Big)\\ &\quad -\ln\Big(1+\frac{1}{4m^2}\Big)-\ln\Big(1+\frac{3}{4m^2}\Big)-\ln\Big(1+\frac{5}{4m^2}\Big)-\cdots-\ln\Big(1+\frac{4mx-1}{4m^2}\Big).\end{aligned}$$

Uporabimo še vrsto $\ln(1+x) = x - x^2/2 + \cdots$ in opazimo, da gredo z izjemo prvega, vsi členi v tej vrsti s $m\to\infty$ proti nič:

$$\ln\left(\frac{f(x)}{f(0)}\right) = \lim_{m\to\infty} -2\frac{1+3+5+\cdots+(4mx-1)}{4m^2} = \lim_{m\to\infty} -\frac{(2mx)^2}{2m^2} = -2x^2.$$

Pri zadnjem enačaju smo uporabili še dejstvo, da je vsota prvih N lihih števil enaka kvadratu števila $(N+1)/2$. $\square$

Poglavje 5

Slučajne spremenljivke in porazdelitve

Denimo, da imamo poskus, katerega izidi so števila (npr. pri metu kocke so izidi števila pik). Se pravi, da je poskusom prirejena neka količina, ki more imeti različne vrednosti. Torej je spremenljivka. Katero od mogočih vrednosti zavzame v določeni ponovitvi poskusa, je odvisno od slučaja. Zato ji rečemo **slučajna spremenljivka**. Da je slučajna spremenljivka znana, je potrebno poznati

1. kakšne vrednosti more imeti (*zaloga vrednosti*) in
2. kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti.
   Predpis, ki določa te verjetnosti, imenujemo **porazdelitveni zakon**.

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami iz konca abecede, vrednosti spremenljivke pa z enakimi malimi črkami. Tako je npr. $(X = x_i)$ dogodek, da slučajna spremenljivka X zavzame vrednost x_i.

Primer: (i) Trikrat vržemo pošten kovanec. X je število padlih grbov.
(ii) Streljamo v tarčo, Y pa predstavlja razdaljo zadetka od središča tarče.
(iii) Z je število klikov na določeni spletni strani v časovni enoti. $\diamond$

Porazdelitveni zakon slučajne spremenljivke X je poznan, če je mogoče za vsako realno število x določiti verjetnost

$$F(x) = P(X \leqslant x).$$

$F(x)$ ali bolj natančno $F_X(x)$ imenujemo **porazdelitvena funkcija** (tudi kumulativna porazdelitvena funkcija). Najpogosteje uporabljamo naslednji vrsti slučajnih spremenljivk:

1. **diskretna** slučajna spremenljivka, pri kateri je zaloga vrednosti neka števna (diskretna) množica,
2. **zvezna** slučajna spremenljivka, ki lahko zavzame vsako realno število znotraj določenega intervala (bodisi končnega ali neskončnega).

V nekaterih primerih je slučajna spremenljivka lahko tudi kombinirana, tj. na nekem območju diskretna in drugje zvezna.

Izrek 5.1. **[Lastnosti porazdelitvene funkcije F]**

1. *Funkcija F je definirana na vsem $\mathbb{R}$ in zanjo velja $0 \leqslant F(x) \leqslant 1$ za vsak $x \in \mathbb{R}$.*
2. *Funkcija F je nepadajoča $x_1 < x_2 \Longrightarrow F(x_1) \leqslant F(x_2)$.*
3. $F(-\infty) := \lim_{x\to-\infty} F(x) = 0$ *in* $F(\infty) := \lim_{x\to\infty} F(x) = 1$.
4. *Funkcija je v vsaki točki zvezna od desne* $F(x+) := \lim_{0\leqslant h\to 0} F(x+h) = F(x)$.
5. *Funkcija ima lahko v nekaterih točkah skok. Vseh skokov je največ števno mnogo.*
6. $P(x_1 < X \leqslant x_2) = F(x_2) - F(x_1)$.
7. $P(x_1 \leqslant X \leqslant x_2) = F(x_2) - F(x_1+)$.
8. $P(X > x) = 1 - F(x)$.
9. $P(X = x) = F(x) - F(x-)$. □

5.1 Diskretne slučajne spremenljivke

Zaloga vrednosti diskretne slučajne spremenljivke X je števna množica

$$\{x_1, x_2, \dots, x_m, \dots\}.$$

Torej je lahko tudi števno neskončna, kot npr. množici naravnih ali celih števil: $\mathbb{N}$, $\mathbb{Z}$. Dogodki

$$X = x_k, \quad k = 1, 2, \dots, m, \dots$$

sestavljajo popoln sistem dogodkov. Označimo verjetnost posameznega dogodka s p_i, tj.

$$p_i = P(X = x_i).$$

Vsota verjetnosti vseh dogodkov je enaka 1:

$$p_1 + p_2 + \cdots + p_m + \cdots = 1.$$

Verjetnostna tabela prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti x_i, pod njimi pa so pripisane pripadajoče verjetnosti:

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_m & \cdots \\ p_1 & p_2 & \cdots & p_m & \cdots \end{pmatrix}.$$

Porazdelitveno funkcijo v diskretnem primeru lahko izrazimo na naslednji način

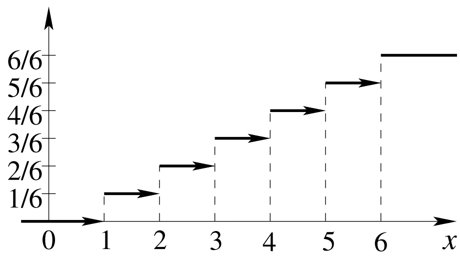
$$F(x_k) = P(X \leqslant x_k) = \sum_{i=1}^{k} p_i.$$

Je torej vsota vseh verjetnosti, ki pripadajo vrednostim, do neke določene vrednosti (kumulativa). Pare $(x_1, p_1), (x_2, p_2), \ldots, (x_m, p_m), \ldots$ lahko predstavimo v ravnini s točkami in se tako približamo histogramom ter razmišljamo kakšne oblike je katera porazdelitev.

5.1.1 Enakomerna diskretna porazdelitev

Končna diskretna slučajna spremenljivka se porazdeljuje **enakomerno**, če so vse njene vrednosti enako verjetne. Kakšna bi izgledala grafična predstavitev? Na desni je prestavljen graf porazdelitvene funkcije. Primer enakomerno porazdeljene slučajne spremenljivke je število pik pri metu kocke

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$



5.1.2 Binomska porazdelitev $B(n,p)$

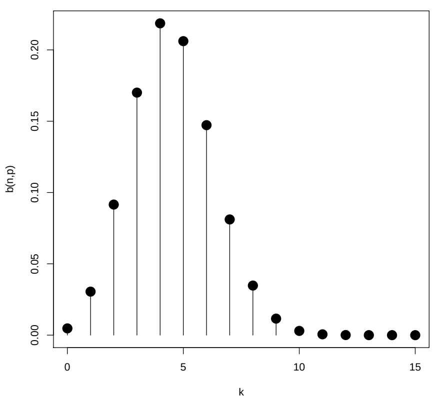
Kovanec vržemo 10 krat. Kolikšne so verjetnosti, da pade 'cifra' 0-krat, 1-krat, 2-krat, vsaj 3-krat itd.?

Binomska porazdelitev ima zalogo vrednosti $\{0, 1, 2, \ldots, n\}$ in verjetnosti, ki jih računamo po Bernoullijevem obrazcu (4.1):

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

$k = 0, 1, 2, \cdots, n$, glej sliko 4.3. Binomska porazdelitev je natanko določena z dvema podatkoma – parametroma: n in p. Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p, zapišemo:

$$X \sim B(n,p).$$




```
> h <- dbinom(0:15,size=15,prob=0.3)
> plot(0:15,h,type="h",xlab="k",ylab="b(n,p)")
> points(0:15,h,pch=16,cex=2)
```

Primer: Naj bo slučajna spremenljivka X določena s številom fantkov v družini s 4 otroki. Denimo, da je enako verjetno, da se v družini rodi fantek ali deklica:

$$P(F) = p = \frac{1}{2}, \qquad P(D) = q = \frac{1}{2}.$$

Spremenljivka X se tedaj porazdeljuje binomsko $B(4, \frac{1}{2})$ in njena verjetnostna shema je:

$$X \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1/16 & 4/16 & 6/16 & 4/16 & 1/16 \end{pmatrix}.$$

Na primer

$$P(X=2) = P_4(2) = \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{4-2} = \frac{6}{16} = \frac{3}{8}.$$

Porazdelitev obravnavane slučajne spremenljivke je simetrična. ◇

Pokazati se da, da je binomska porazdelitev simetrična, le če je $p = 0{\cdot}5$, sicer je asimetrična. Za $n = 1$ ji pravimo **Bernoullijeva** (oz. indikatorska) slučajna spremenljivka. Z njo najlažje izračunamo pričakovano vrednost in varianco binomske porazdelitve: $\mu = np$ in $\sigma^2 = npq$.

5.1.3 Poissonova porazdelitev $P(\lambda)$

Strokovnjaki za promet opazujejo neko križišče. Radi bi napovedali, kakšna je verjetnost, da v danem časovnem intervalu skozi križišče zapelje npr. 100 avtomobilov. Definirajo slučajno spremenljivko X = število avtomobilov, ki pridejo v križišče na uro in v modelu privzamejo še dve predpostavki:

- vsaka ura je enaka kot vsaka druga ura (čeprav vemo, da to ne drži, saj je v konici več prometa kot npr. ponoči).
- če v enem časovnem intervalu pride veliko avtomobilov, to še ne pomeni, da bo podobno tudi v naslednjem časovnem intervalu (takšni dogodki so med seboj neodvisni).

Za začetek si s štetjem prometa pridobijo grobo oceno za pričakovano število avtomobilov na časovno enoto: $\mathsf{E}X = \lambda$, npr. 9 avtomobilov na uro in poiskusijo v modelu uporabiti binomsko porazdelitev $B(n, p)$. Zanjo velja $\mathsf{E}X = np$, število n, ki predstavlja število ponovitev poskusa, je v našem primeru enako številu manjših časovnih enot, verjetnost posameznega dogodka p pa je enaka verjetnosti, da je v dani manjši časovni enoti prišel mimo vsaj en avto. Potem lahko sklepajo na naslednji način:

$$\lambda \,\text{avtomobilov/uro} = (60\,\text{min./h}) \times \big((\lambda/60)\,\text{avtomobilov/min.}\big),$$

formula za verjetnost v primeru binomske porazdelitve pa je enaka

$$P(X=k) = P_{60}(k) = \binom{60}{k} \left(\frac{\lambda}{60}\right)^k \left(\frac{1-\lambda}{60}\right)^{60-k}.$$

Vendar tu naletijo na problem, da gre v eni manjši časovni enoti mimo lahko več kot en avtomobil na minuto. V tem primeru štejejo, da se je dogodek zgodil, tj. četudi gre v tisti minuti mimo npr. 5 avtomobilov. Ta problem rešijo tako, da zmanjšajo časovno enoto:

$$P(X=k)=P_{3600}(k)=\binom{3600}{k}\left(\frac{\lambda}{3600}\right)^k\left(\frac{1-\lambda}{3600}\right)^{3600-k}.$$

Če pa tudi to ni dovolj, zamenjajo 3600 še z večjim številom. Pa poglejmo kaj se zgodi, če gre to število proti neskončno. Iz analize se spomnimo naslednje zveze

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=e^{-\lambda}.$$

Limito verjetnosti $P_n(k)$ lahko izračunamo za velike n in $\lambda=np$ na naslednji način:

$$\begin{aligned}\lim_{n\to\infty}P_n(k) &= \lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=\lim_{n\to\infty}\frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \left(\frac{\lambda^k}{k!}\right)\lim_{n\to\infty}\left[\underbrace{\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}}_{\to 1}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1}\right]=\left(\frac{\lambda^k}{k!}\right)e^{-\lambda}.\end{aligned}$$

Izpeljali[1] smo:

Izrek 5.2. (**Poissonov obrazec**) *Za velike n in majhne verjetnosti, tj. p blizu 0 velja*

$$P_n(k)\approx\frac{(np)^k e^{-np}}{k!}.$$ □

Dobljena limita predstavlja verjetnost pri porazdelitvi, ki jo želimo vpeljati v tem razdelku. **Poissonova porazdelitev** izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se ti dogodki pojavijo s poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek. Poissonovo porazdelitev lahko uporabimo tudi za število dogodkov v drugih intervalih, npr. razdalja, prostornina,... Ima zalogo vrednosti $\{0,1,2,\dots\}$, njena verjetnostna funkcija pa je

$$p_k=P(\texttt{\#dogodkov}=k)=\lambda^k\frac{e^{-\lambda}}{k!},$$

[1] Za dokaz $\lim_{n\to\infty}F=e^0=1$, kjer je $F:=n(n-1)\cdots(n-k+1)/n^k$, lahko uporabimo Stirlingovo aproksimacijo (A.4) na izrazu $\ln F=\ln(n!)-\ln[(n-k)!]-k\ln n$ in za velike n res dobimo:

$$\ln F\approx[n\ln n-n]-[(n-k)\ln(n-k)-(n-k)]-[k\ln n]=\underbrace{-\left(1-\frac{k}{n}\right)}_{\to -1}\underbrace{\ln\left(1-\frac{k}{n}\right)^n}_{\to -k}-k\approx k-k=0.$$

kjer je $\lambda > 0$ dani parameter – in predstavlja pričakovano pogostost nekega dogodka. Konstanta e je osnova naravnega logaritma, tj. $e = 2{\cdot}71828\ldots$.

$$p_{k+1} = \frac{\lambda}{k+1}\, p_k, \quad p_0 = e^{-\lambda}.$$

Vidimo, da zaloga vrednosti te slučajne spremenljivke ni omejena, saj je verjetnost, da se v nekem časovnem obdobju zgodi mnogo uspehov različna od nič. To je bistvena razlika v primerjavi z binomsko porazdelitvijo, kjer število uspehov seveda ne more presegati števila Bernoullijevih poskusov n. Poissonov obrazec lahko sedaj zapišemo tudi v naslednji obliki: $B(n, p) \approx P(np)$.

Primer: Posebno pomembna je ta porazdelitev v teoriji množične strežbe. Če se dogodek pojavi v povprečju 3-krat na minuto in nas zanima kolikokrat se bo zgodil v četrt ure, potem uporabimo za model Poissonovo porazdelitev z $\lambda = 15 \times 3 = 45$. $\diamond$

Naštejmo še nekaj primerov, ki jih dobro opišemo (modeliramo) s Poissonovo porazdelitvijo:

- število dostopov do omrežnega strežnika na minuto (pod predpostavko homogenosti),
- število telefonskih klicev na bazni postaji na minuto,
- število mutacij v danem intervalu RNK po določeni količini sprejete radiacije,
- število vojakov, ki so umrli vsako leto za posledicami konjske brce v vsaki diviziji Pruske konjenice, (iz knjige Ladislausa Josephovicha Bortkiewicza, 1868–1931).

5.1.4 Negativna binomska oziroma Pascalova porazdelitev $P(m, p)$

Kolikokrat moramo vreči kocko, da z verjetnostjo vsaj $0{\cdot}99$ pričakujemo, da bo padla vsaj ena šestica?

Negativna binomska $\mathrm{NegBin}(m, p)$, ki je znana tudi pod imenom **Pascalova porazdelitev** $P(m, p)$, ima zalogo vrednosti $\{m, m+1, m+2, \ldots\}$, verjetnostna funkcija pa je

$$p_k = \binom{k-1}{m-1} p^m q^{k-m},$$

kjer je $0 < p < 1$ dani parameter za verjetnost dogodka A v posameznem poskusu.

Opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi m–krat.

Za $m = 1$, porazdelitvi $G(p) = P(1, p)$ pravimo **geometrijska porazdelitev**. Opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi prvič.

Primer: Če mečemo kovanec toliko časa, da pade grb in z X označimo število potrebnih metov, vključno z zadnjim, potem je slučajna spremenljivka X geometrijsko porazdeljena. ◇

Primer: Marko zbira kartice svojih športnih junakov. Celotno zbirko sestavlja n različnih kartic. Poišči verjetnost, da bo moral kupiti vsaj t kartic, da bo sestavil celotno zbirko. Lahko se vprašamo tudi koliko je pričakovano število kartic, ki jih bo Marko kupil, da bo zbral vse kartice. Izkušnje kažejo, da je npr. za $n = 50$ potrebno kupiti 225 kartic, da zberemo vseh 50 različnih kartic. Naj bo T čas, da zberemo vseh n kartic in naj bo T_i čas da zberemo i-to kartico, potem ko smo že zbrali $i-1$ kartic. Potem sta T in T_i slučajni spremenljivki. Verjetnost, da bomo izbrali novo kartico, če smo jih izbrali že $i-1$ je enaka $p_i = (n-i+1)/n$. Torej je T_i slučajna spremenljivka z geometrijsko porazdelitvijo in pričakovano vrednostjo $1/p_i$. Zaradi linearnosti pričakovane vrednosti je

$$\begin{aligned}\mathrm{E}(T) &= \mathrm{E}(T_1) + \mathrm{E}(T_2) + \cdots + \mathrm{E}(T_n) = \frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n} \\ &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}\right) = n \cdot H_n,\end{aligned}$$

kjer je H_n *harmonično število*. Če upoštevamo še asimptotično vedenje harmoničnih števil, dobimo

$$\mathrm{E}(T) = n \cdot H_n = n \ln n + \gamma n + \frac{1}{2} + o(1), \quad \text{za } n \to \infty,$$

kjer je $\gamma \approx 0{\cdot}5772156649$ *Euler–Mascheronijeva konstanta*. ◇

5.1.5 Hipergeometrijska porazdelitev $H(n; M, N)$

Hipergeometrijska porazdelitev ima zalogo vrednosti $\{0, 1, 2, \ldots\}$, njena verjetnostna funkcija pa je

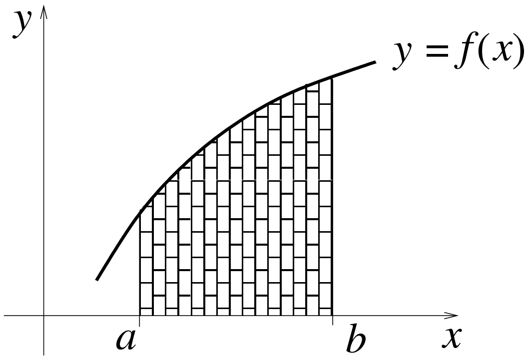
$$p_k = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

za $\max(0, n-(N-M)) \leqslant k \leqslant \min(M, n)$ in $n \leqslant N$. Opisuje verjetnost dogodka, da je med n izbranimi kroglicami natanko k belih, če je v posodi M belih in $N-M$ črnih kroglic in izbiramo n–krat brez vračanja. Zanima nas primer, ko je velikost n vzorca občutna glede na velikost populacije N, tj. $n/N > 0{\cdot}05$. Sicer bi za velike N, M in $N-M$ lahko uporabljali kar binomsko porazdelitev, pri čemer bi vzeli $p = M/N$. Slednje pomeni, da je pri veliki seriji praktično brez pomena ali izbiramo vzorec z vračanjem ali brez.

Primer: Mešalec dobro premeša 52 standardnih kart in nam jih dodeli pet.

- Kakšna je verjetnost, da dobimo med prejetimi kartami štiri ase?
- Kolikšno je pričakovano število asov, ki jih prejmemo? ◇

5.2 Ponovitev: integrali



Določeni integral predstavlja ploščino pod krivuljo. Naj bo funkcija $y = f(x)$ zvezna na $[a, b]$ in nenegativna. Ploščina lika med krivuljo $f(x) \geqslant 0$, in abscisno osjo na intervalu $[a, b]$ je enaka določenemu integralu

$$\int_a^b f(x)dx.$$

Trditev 5.3. **[Lastnosti določenega integrala]**

(1) *Za* $a, b \in \mathbb{R}$ *velja*

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

(2) *Če je* $f(x) \leqslant 0 \;\; \forall x \in [a, b]$, *je vrednost integrala negativna.*

(3) *Za* $c \in [a, b]$ *valja*

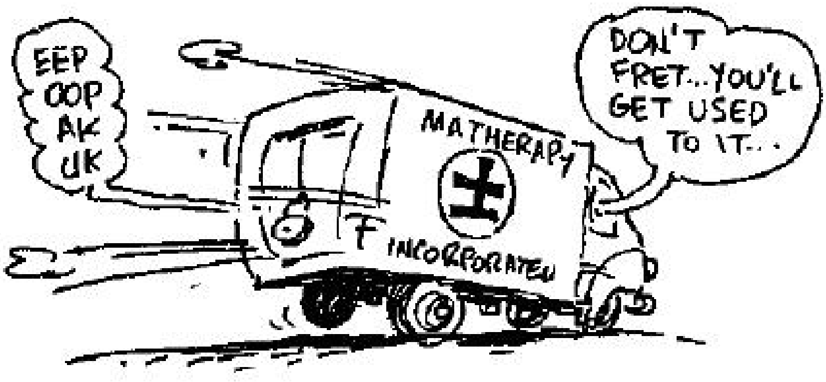
$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

(4) *Naj bo* $f(x) \geqslant g(x)$, $x \in [a, b]$, *potem velja*

$$\int_a^b f(x)\,dx \geqslant \int_a^b g(x)\,dx. \qquad \square$$

Saj vas razumem: potem pa uporabimo še ∞ za mejo pri integriranju. **A brez preplaha!**

Iščemo le celotno ploščino pod krivuljo, od enega konca do drugega, le da konca pravzaprav sploh ni.



5.3 Zvezne slučajne spremenljivke

Slučajna spremenljivka X je **zvezno porazdeljena**, če obstaja taka integrabilna funkcija p,[2] imenovana **gostota verjetnosti**, da za vsak $x \in \mathbb{R}$ velja:

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(t)\, dt,$$

kjer $p(x) \geqslant 0$. To verjetnost si lahko predstavimo tudi grafično v koordinatnem sistemu, kjer na abscisno os nanašamo vrednosti slučajne spremenljivke, na ordinatno pa gostoto verjetnosti $p(x)$. Verjetnost je tedaj predstavljena kot ploščina pod krivuljo, ki jo določa $p(x)$. Spomnimo se še Newton-Liebnitzove formule iz analize in zapišimo vse omenjene zveze:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1, \qquad P(x_1 < X \leqslant x_2) = \int_{x_1}^{x_2} p(t)\, dt = F(x_2) - F(x_1) \qquad \textit{ter} \qquad p(x) = F'(x).$$

5.3.1 Enakomerna porazdelitev zvezne slučajne spremenljivke

Verjetnostna gostota **enakomerno porazdeljene zvezne** slučajne spremenljivke je:

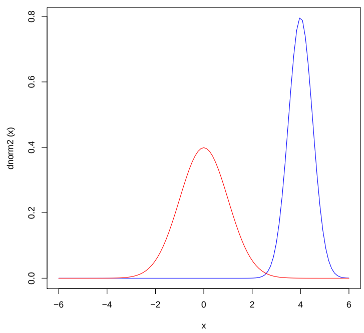
$$p(x) = \begin{cases} \dfrac{1}{b-a} & \text{za } a \leqslant x \leqslant b \\ 0 & \text{sicer.} \end{cases}$$

Grafično si jo predstavljamo kot pravokotnik nad intervalom (a, b) višine $\dfrac{1}{b-a}$.

[2] V zahodni literaturi se pogosto namesto črke p uporablja kar črka f, kar je v skladu s prejšnjim razdelkom.

5.3.2 Normalna ali Gaussova porazdelitev $N(\mu, \sigma)$

Leta 1738 je Abraham De Moivre (1667-1754) objavil aproksimacijo binomske porazdelitve, ki je normalna krivulja, glej prejšnje poglavje, konkretno De Moivrov točkovni obrazec in njegovo posplošitev, Laplaceov točkovni obrazec. Leta 1809 je Karl Frederic Gauss (1777-1855) raziskoval matematično ozadje planetarnih orbit, ko je prišel do normalne porazdelitvene funkcije.




```
> d2 <- function(x){dnorm(x,mean=4,sd=0.5)}
> curve(d2,-6,6,col="blue")
> curve(dnorm,-6,6,col="red",add=TRUE)
```

Zaloga vrednosti **normalno porazdeljene** slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Normalna porazdelitev je natanko določena s parametroma μ in σ. Če se slučajna spremenljivka X porazdeljuje normalno s parametroma μ in σ zapišemo:

$$X \sim N(\mu, \sigma).$$

Na sliki je vrh krivulje predstavljen s točko $\big(\mu, 1/(\sigma\sqrt{2\pi})\big)$. Če je krivulja, ki ustreza gostoti porazdelitve simetrična glede na premico $x = a$ za nek $a \in R$, potem velja $p(a-x) = p(a+x)$ za vsak $x \in \mathbb{R}$ in rečemo, da je porazdelitev *simetrična*. Ni se težko prepričati, da pričakovana vrednost poljubne simetrične porazdelitve enaka mediani:

$$(\mu - x)\, p(\mu - x) + (\mu + x)\, p(\mu + x) = (\mu - x + \mu + x)\, p(\mu + x) = 2\mu\, p(\mu + x),$$

če gre pa za zvonasto porazdelitev, potem je enaka tudi modusu:

Laplaceov intervalski obrazec

Zanima nas, kolikšna je verjetnost $P_n(k_1, k_2)$, da se v Bernoullijevem zaporedju neodvisnih poskusov v n zaporednih poskusih zgodi dogodek A vsaj k_1–krat in manj kot k_2–krat. Označimo

$$x_k = \frac{k - np}{\sqrt{npq}} \quad \text{in} \quad \Delta x_k = x_{k+1} - x_k = \frac{1}{\sqrt{npq}}.$$

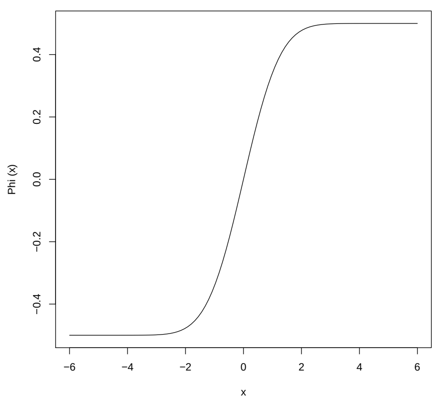
Tedaj je, če upoštevamo Laplaceov točkovni obrazec,

$$P_n(k_1,k_2) = \sum_{k=k_1}^{k_2-1} P_n(k) = \frac{1}{\sqrt{2\pi}} \sum_{k=k_1}^{k_2-1} e^{-\frac{1}{2}x_k^2} \Delta x_k.$$

Za (zelo) velike n lahko vsoto zamenjamo z integralom

$$P_n(k_1,k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx.$$

Funkcija napake $\Phi(x)$




```
> Phi <- function(x){pnorm(x)-0.5}
> curve(Phi,-6.6)
```

Funkcija napake imenujemo funkcijo

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt.$$

Funkcija napake je liha, zvezno odvedljiva, strogo naraščajoča funkcija, za katero velja $\Phi(0) = 0$, $P_n(k_1,k_2) \approx \Phi(x_{k_2}) - \Phi(x_{k_1})$, po izreku A.6 pa še $\Phi(\infty) = 1/2$ in $\Phi(-\infty) = -1/2$.

Vrednosti funkcije napake najdemo v tabelah ali pa je vgrajena v statističnih programih.

```
> x2 <- (50 - 1000*0.05)/sqrt(1000*0.05*0.95)
> x1 <- (0 - 1000*0.05)/sqrt(1000*0.05*0.95)
> pnorm(x2)-pnorm(x1)
[1] 0.5
```

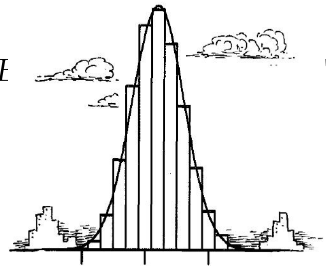
Porazdelitveno funkcijo slučajne spremenljivke $X \sim N(\mu,\sigma)$ izrazimo s funkcijo napake na naslednji način:

$$\begin{aligned} F(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{1}{2}s^2} ds \\ &= \frac{1}{2} + \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

oziroma velja

$$P(x_1 \leqslant X < x_2) = \Phi\left(\frac{x_2-\mu}{\sigma}\right) - \Phi\left(\frac{x_1-\mu}{\sigma}\right).$$

Porazdelitev $N(0,1)$ je **standardizirana normalna porazdelitev**.



Spremenljivko $X \sim N(\mu,\sigma)$ *pretvorimo s transformacijo* $z = \dfrac{x-\mu}{\sigma}$ *v standardizirano spremenljivko* $Z \sim N(0,1)$.

Iz Laplaceovega obrazca izhaja $B(n,p) \approx N(np,\sqrt{npq})$.

Izpeljimo še izrek, ki opravičuje statistično definicijo verjetnosti.

Izrek 5.4 (Bernoullijev zakon velikih števil – 1713). *Naj bo* k *frekvenca dogodka* A *v* n *neodvisnih ponovitvah danega poskusa, v katerem ima dogodek* A *verjetnost* p. *Tedaj za vsak* $\varepsilon > 0$ *velja*

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n}-p\right| \leqslant \varepsilon\right) = 1.$$

Bolj natančno, za dovolj velika naravna števila n *velja*

$$P\left(\left|\frac{k}{n}-p\right| \leqslant \varepsilon\right) \approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right).$$

Dokaz. Ker je n naravno število, lahko oba izraza v neenakost iz zgornje verjetnosti pomnožimo z n, nato pa še odpravimo absolutno vrednost in z upoštevanjem, da je tudi k celo število med 0 in n, dobimo oceno:

$$P\left(\left|\frac{k}{n}-p\right| le\varepsilon\right) = P(np-n\varepsilon \leqslant k \leqslant np+n\varepsilon) = P_n(k_1)+P_n(k_1+1)+\cdots+P_n(k_2),$$

kjer so $k_1 < k_1+1 < \cdots < k_2$ vsa cela števila na intervalu $[np-n\varepsilon, np+n\varepsilon]$, tj. intervalu s središčem v točki np in radijem $n\varepsilon$. Dobljeno vsoto smo označili s $P(k_1-1,k_2)$ in jo ocenili s funkcijo napake, kar nam da:

$$P(k_1-1,k_2) \approx \Phi\left(\frac{k_2-np}{\sqrt{npq}}\right) - \Phi\left(\frac{(k_1-1)-np}{\sqrt{npq}}\right) \approx 2\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right),$$

saj sta razliki k_2-np in $(np-1)-k_1$ za velike n približno enaki omenjenemu radiju, Φ pa je liha funkcija. Za konec upoštevamo še $\Phi(\infty)=1/2$ ter da so števila ε, p in $q=1-p$ vnaprej izbrane konstante, število n pa pošljemo proti neskončno in zgornja limita je izračunana. □

Primer: (a) Kolikšna je verjetnost, da se pri metu kovanca relativna frekvenca grba v 3600 metih ne razlikuje od 0·5 za več kot 0·01, se pravi, da grb pade med 1764 in 1836-krat? V tem primeru je $p=1/2$, $n=3600$, $\varepsilon=0{\cdot}01$, tako, da iz zgornje formule dobimo

$$2\,\Phi\left(0{\cdot}01\cdot\sqrt{\frac{3600}{0{\cdot}25}}\right) = 2\,\Phi(1{\cdot}2) = 2\cdot 0{\cdot}385 = 0{\cdot}77,$$

kar je presenetljivo veliko, kaj ne?

(b) Kolikokrat moramo vreči pošten kovanec, da bo verjetnost dogodka, da se relativna frekvenca grba razlikuje od 0·5 za manj kot 0·05, večja od 0·997? Iz tabele za Φ vidimo, da je $2\,\Phi(x) > 0{\cdot}997$ za $x=3$, zato moramo poiskati tak n, da bo

$$3 < \varepsilon\sqrt{\frac{n}{pq}} = 0{\cdot}05\sqrt{\frac{n}{0{\cdot}25}} \qquad \text{oziroma} \qquad n > \left(\frac{3}{0{\cdot}05}\right)^2 \times 0{\cdot}25 = 900. \qquad \diamond$$

5.3.3 Eksponentna porazdelitev

Gostota **eksponentne porazdelitve**, ki ji pravimo tudi porazdelitev Poissonovega toka, je enaka

$$p(x) = \lambda e^{-\lambda x}, \quad x \geqslant 0.$$

Potem za porazdelitveno funkcijo velja

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

5.3.4 Porazdelitev Gama

Prejšnjo porazdelitev lahko še precej posplošimo. Naj bosta $b, c > 0$. Tedaj ima **porazdelitev Gama** $\Gamma(b, c)$ gostoto:

$$p(x) = \frac{c^b}{\Gamma(b)} x^{b-1} e^{-cx}, \; 0 < x$$

in $p(x) = 0$ za $x \leqslant 0$. Za $b = 1$ seveda dobimo eksponentno porazdelitev. **Funkcijo Gama** lahko definiramo z določenim integralom za $\Re[z] > 0$ (Eulerjeva integralna forma)

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt = 2\int_0^\infty e^{-t^2} t^{2z-1}\, dt.$$

Leonhard Euler (1707-1783) [3]

Torej je $\Gamma(1) = 1$ in zaradi (A.5) tudi $\Gamma(1/2) = \sqrt{\pi}$. Z uvedbo nove spremenljivke dobimo še

$$\Gamma(z) = \int_0^1 \left[\ln \frac{1}{t}\right]^{z-1} dt.$$

(Je povsod analitična z izjemo $z = 0, -1, -2, \ldots$ in nima ničel. Glej npr. http://en.wikipedia.org/wiki/Gamma_function in http://en.wikipedia.org/wiki/Gamma_distribution.)

Integral poskusimo rešiti z integracijo po delih (po realnem argumentu). Za $v = t^x$ in $du = e^{-t}dt$ velja $dv = (x-1)t^{x-2}$ in $u = -e^{-t}$, od koder sledi

$$\begin{aligned}\Gamma(x) &= \int_0^\infty t^{x-1}e^{-t}\, dt = \Big[-t^{x-1}e^{-t}\Big]_0^\infty + \int_0^\infty (x-1)t^{x-2}e^{-t}\, dt \\ &= (x-1)\int_0^\infty t^{x-2}e^{-t}\, dt \quad = \quad (x-1)\Gamma(x-1).\end{aligned}$$

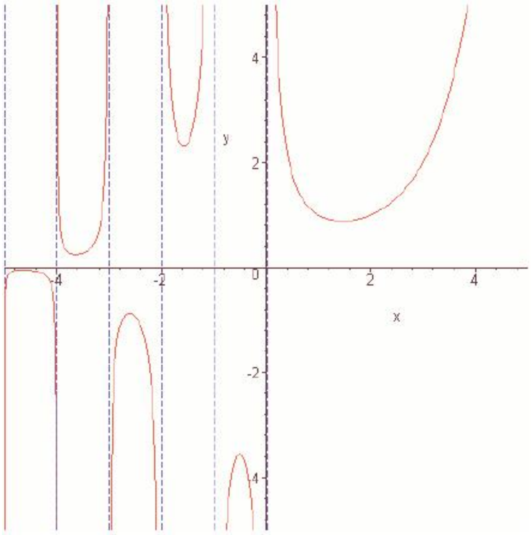
[3] Eulerju se lahko zahvalimo za zapis $f(x)$ za funkcijo (1734), e za osnovo naravnega logaritma (1727), i za kvadratni koren števila -1 (1777), π za konstanto pi, Σ za vsoto (1755), in mnoge druge oznake/koncepte, ki smo jih danes sprejeli za same po sebi umevne. Glej http://www.gap-system.org/~history/Biographies/Euler.html

Za naravno število $x = n \in \{1, 2, \dots\}$ torej dobimo

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)...1 = (n-1)!,$$

kar pomeni, da se v tem primeru funkcija Gama zreducira v 'faktorijel'. Tule je še majhna tabela za funkcijo Gama[4] ter njen graf:

x	Gama(x)
1.1	0.951351
1.2	0.918168
1.3	0.897471
1.4	0.887264
1.5	0.886225
1.6	0.893514
1.7	0.908642
1.8	0.931384
1.9	0.961767



5.3.5 Porazdelitev hi-kvadrat


Ernst Abbe
(1840-1905)

Porazdelitev **hi-kvadrat** je poseben primer porazdelitve Gama:

$$\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

($n \in \mathbb{N}$ je število prostostnih stopenj) in ima gostoto

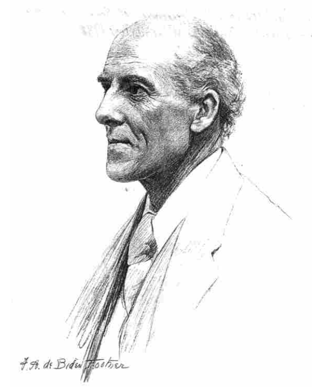
$$p(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x > 0.$$

Leta 1863 jo je prvi izpeljal nemški fizik Ernst Abbe, ko je preučeval porazdelitev vsote kvadratov napak. Če je $n = 2k$ za $k \in \mathbb{N}$, je $\Gamma(n/2) = (k-1)!$, če pa je $n = 2k+1$ za $k \in \mathbb{N}$, je

$$\Gamma(n/2) = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2k-1)\sqrt{\pi}/2^n.$$

Leta 1878 je Ludwig Boltzmann izpeljal hi-kvadrat porazdelitev z dvema in tremi prostostnimi stopnjami, ko je študiral kinetično energijo molekul.

Karl Pearson (1875-1937) je demonstriral uporabnost hi-kvadrat porazdelitve statistikom.



[4] http://functions.wolfram.com/webMathematica/FunctionEvaluation.jsp?name=Gamma

5.3.6 Cauchyeva porazdelitev

To je porazdelitev z gostoto

$$p(x) = \frac{a}{\pi} \, \frac{1}{1 + a^2(x-b)^2}, \qquad \text{za} \;\; x \in \mathbb{R} \quad \text{in} \quad a > 0.$$

Njena porazdelitvena funkcija je enaka

$$F(x) = \frac{a}{\pi} \int_{-\infty}^{x} \frac{1}{1 + a^2(x-b)^2} \mathrm{d}x = \frac{1}{\pi} \operatorname{arctg}\big(a(x-b)\big) + \frac{1}{2}.$$

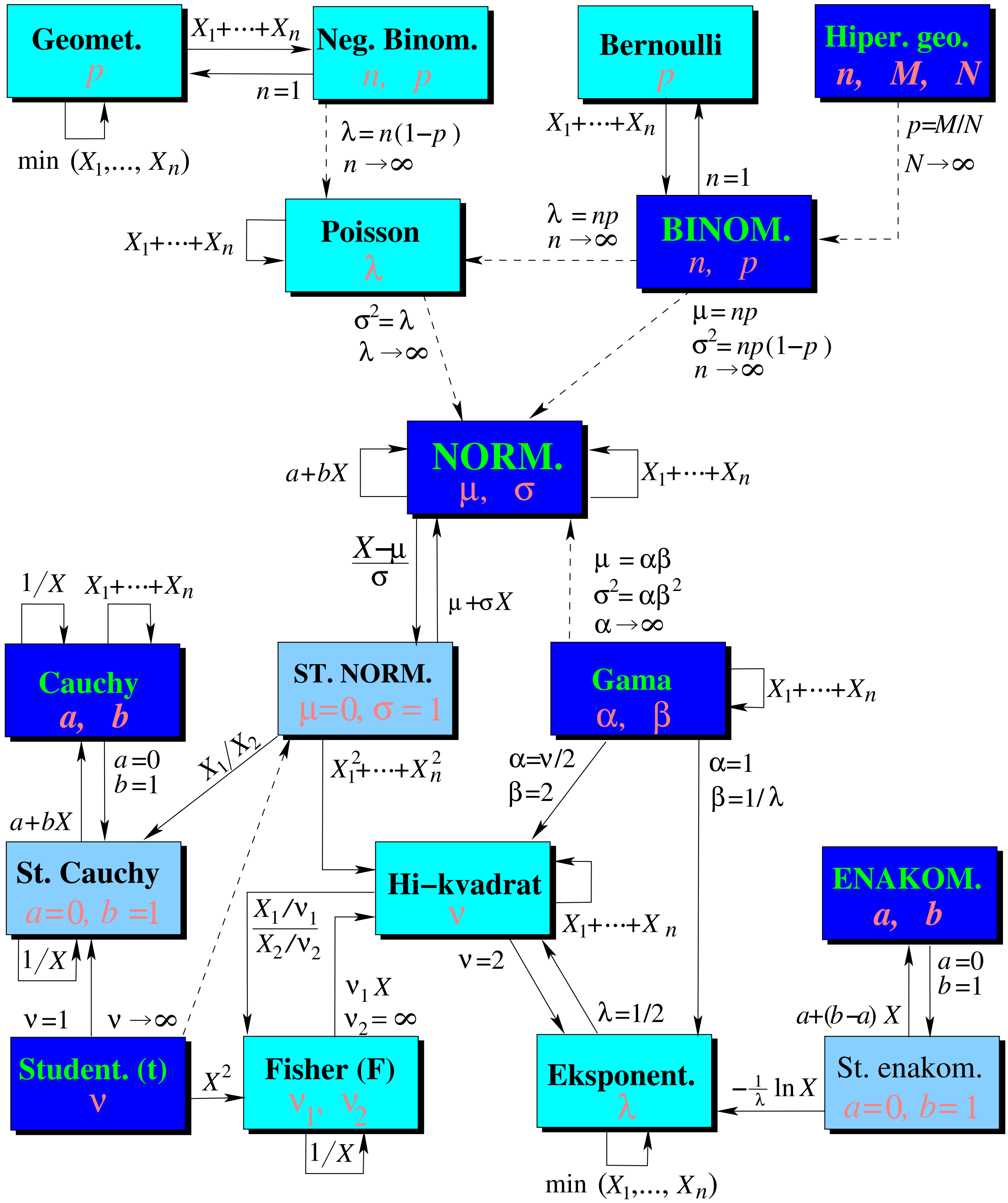
Porazdelitve v R-ju

V R-ju so za delo s pomembnejšimi porazdelitvami na voljo funkcije:
`d`*ime* – gostota porazdelitve *ime* $p_{ime}(x)$
`p`*ime* – porazdelitvena funkcija *ime* $F_{ime}(q)$
`q`*ime* – obratna funkcija: $q = F_{ime}(p)$
`r`*ime* – slučajno zaporedje iz dane porazdelitve

Za *ime* lahko postavimo:

- `unif` – zvezna enakomerna,
- `binom` – binomska,
- `norm` – normalna,
- `exp` – eksponentna,
- `lnorm` – logaritmičnonormalna,
- `chisq` – porazdelitev χ^2, ...

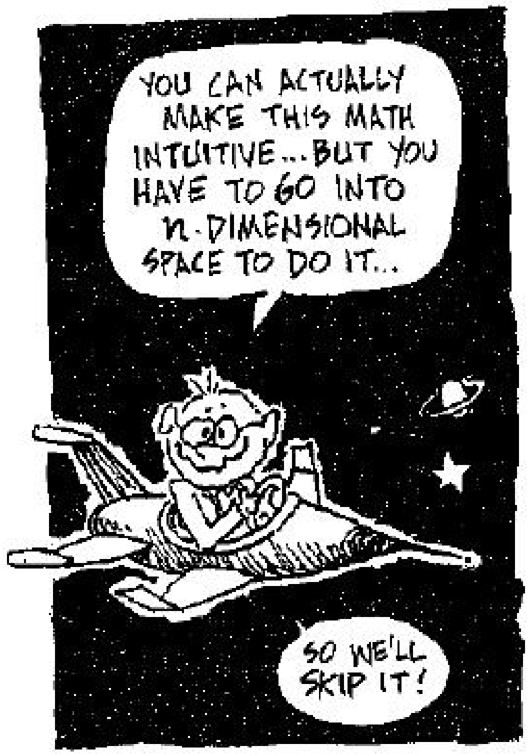
Opis posamezne funkcije in njenih parametrov dobimo z ukazom `help`. Na primer `help(rnorm)`.



Slika 5.16: Povezave med znanimi porazdelitvami. Iz normalne na standardno normalno pridemo s substitucijo $(X-\mu)/\sigma$. Iz binomske na normalno pridemo tako,da pošljemo $n \to \infty$ in dobimo, da je $\mu = np$ in $\sigma^2 = np(1-p)$, do Poissonove pa tako, da je $\lambda = np$...

Poglavje 6

Slučajni vektorji



Primer: Naj slučajna spremenljivka X predstavlja število naprav, ki so na voljo, slučajna spremenljivka Y pa število zaporednih operacij, ki jih moramo opraviti za procesiranje kosa materiala. Verjetnostna funkcija $P(X = x, Y = y) = p(x, y)$ je definirana z naslednjo tabelo:

$Y \setminus X$	1	2	3	4
0	0	0·10	0·20	0·10
1	0·03	0·07	0·10	0·05
2	0·05	0·10	0·05	0
3	0	0·10	0·05	0

Poišči verjetnostno tabelo spremenljivke X!

Seštejemo verjetnosti po stolpcih:

$Y \setminus X$	1	2	3	4	Y
0	0	0·10	0·20	0·10	0·40
1	0·03	0·07	0·10	0·05	0·25
2	0·05	0·10	0·05	0	0·20
3	0	0·10	0·05	0	0·15
X	0·08	0·37	0·40	0·15	1

in dobimo

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0{\cdot}08 & 0{\cdot}37 & 0{\cdot}40 & 0{\cdot}15 \end{pmatrix}.$$

Enako lahko storimo tudi za Y. *Ali sta slučajni spremenljivki X in Y neodvisni?*
Ne nista, saj velja npr.: $P(x = 4, y = 3) = 0 \neq 0{\cdot}15 \times 0{\cdot}15 = P(x = 4) \cdot P(y = 3)$. ◇

Slučajni vektor je n-terica slučajnih spremenljivk $\boldsymbol{X} = (X_1, \ldots, X_n)$. Opišemo ga s porazdelitveno funkcijo ($x_i \in \mathbb{R}$)

$$F(x_1, \ldots, x_n) = P(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n),$$

(pri čemer slednja oznaka pomeni $P\big(\{X_1 \leqslant x_1\} \cap \ldots \cap \{X_n \leqslant x_n\}\big)$) in za katero velja: $0 \leqslant F(x_1, \ldots, x_n) \leqslant 1$. Funkcija F je za vsako spremenljivko naraščajoča in od desne zvezna, veljati pa mora tudi

$$F(-\infty, \ldots, -\infty) = 0 \qquad \text{in} \qquad F(\infty, \ldots, \infty) = 1.$$

Funkciji $F_i(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$ pravimo **robna porazdelitvena funkcija** spremenljivke X_i.

Primer: Katere od naslednjih funkcij so lahko porazdelitvene funkcije nekega slučajnega vektorja (X, Y):

(a) $F(x, y)$ je enaka $1 - e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0,

(b) $F(x, y)$ je enaka $1 + e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0,

(c) $F(x, y) = x^2$,

(d) $F(x, y) = x^2 - y^2$.

(e) $F(x, y) = 0$.

Funkcija iz (a) nima vseh vrednosti na intervalu $[0, 1]$, npr. $F(0, 0) = 1 - 1 - 1 - 1 = -2 < 0$, zato ne more biti porazdelitvena funkcija. Podobno je tudi v primerih (c): $F(2, 0) = 4 \notin [0, 1]$ in (d): $F(0, 1) = -1 \notin [0, 1]$. V primeru (e) pa velja $F(\infty, \infty) = 0 \neq 1$, kar pomeni, da nam ostane le še možnost (b). V tem primeru lahko zapišemo $F(x, y) = (1 - e^{-x})(1 - e^{-y})$ od koder vidimo, da za $x \geqslant 0$ in $y \geqslant 0$ velja $F(x, y) \in [0, 1]$. Preverimo še $F(0, 0) = 0$ in $F(\infty, \infty) = 1$. ◇

Slučajni vektorji – primer

Naj bo

$$A(x,y) = \{(u,v) \in \mathbb{R}^2 : u \leqslant x \wedge v \leqslant y\}$$

(levi spodnji kvadrant glede na (x,y)). Naj porazdelitvena funkcija opisuje verjetnost, da je slučajna točka (X,Y) v množici $A(x,y)$

$$F(x,y) = P(X \leqslant x, Y \leqslant y) = P((X,Y) \in A(x,y)).$$

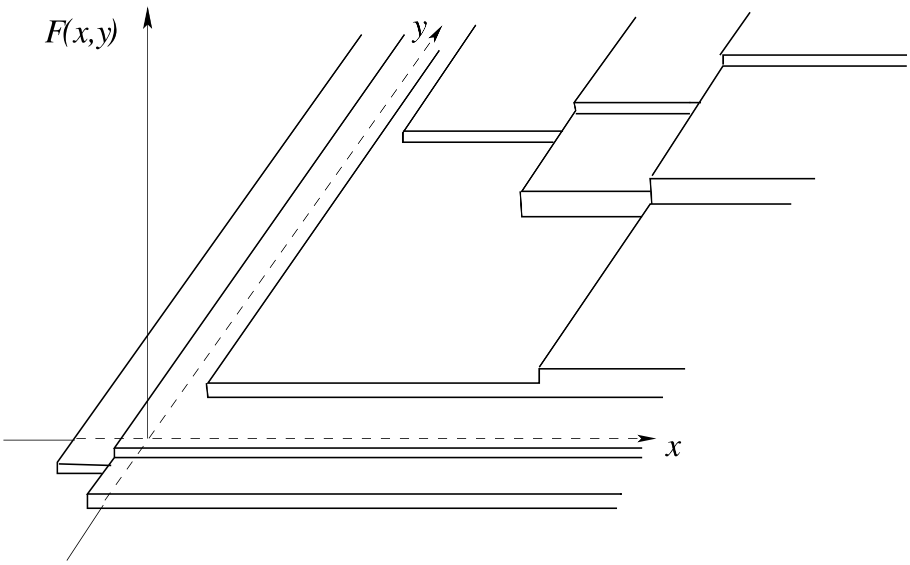
Tedaj je verjetnost, da je slučajna točka (X,Y) v pravokotniku $(a,b] \times (c,d]$ enaka

$$P\Big((X,Y) \in (a,b] \times (c,d]\Big) = F(b,d) - F(a,d) - F(b,c) + F(a,c) \qquad (6.1)$$

Zaloga vrednosti je kvečjemu števna množica. Opišemo jo z **verjetnostno funkcijo** $p_{k_1,\ldots,k_n} = P(X_1 = x_{k_1}, \ldots, X_n = x_{k_n})$. Za $n=2$, $X : \{x_1, x_2, \ldots, x_k\}$, $Y : \{y_1, \ldots, y_m\}$ in $P(X = x_i, Y = y_j)$, sestavimo **verjetnostno tabelo**:

$X \setminus Y$	y_1	y_2	$\ldots$	y_m	X
x_1	p_{11}	p_{12}	$\ldots$	p_{1m}	p_1
x_2	p_{21}	p_{22}	$\ldots$	p_{2m}	p_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	p_{k1}	p_{k2}	$\ldots$	p_{km}	p_k
Y	q_1	q_2	$\ldots$	q_m	1

$p_i = P(X = x_i) = \sum_{j=1}^{m} p_{ij}$ in $q_j = P(Y = y_j) = \sum_{i=1}^{k} p_{ij}$



Slika: Porazdelitvena funkcija $F(x,y)$, v primeru, ko sta spremenljivki X in Y diskretni.

Primer: Naj bosta X in Y diskretni slučajni spremenljivki z zalogami vrednosti $X : \{1,2\}$ in $Y : \{0,1\}$. Naj bo

$$P(X = x, Y = y) = p(x,y) = \frac{x - y + a}{5},$$

za neko konstanto a. Določi a!

Imamo štiri možne pare za (x, y): (1,0), (1,1), (2,0), (2,1), po zgornji formuli pa velja: $P\big((x, y) = (1, 0)\big) = (1 + a)/5$, $P\big((x, y) = (1, 1)\big) = a/5$, $P\big((x, y) = (2, 0)\big) = (2 + a)/5$, $P\big((x, y) = (2, 1)\big) = (1 + a)/5$. Vsota vseh verjetnosti je enaka 1, torej je $4 + 4a = 5$ ozirma $a = (5 - 4)/4 = 1/4$. ◇

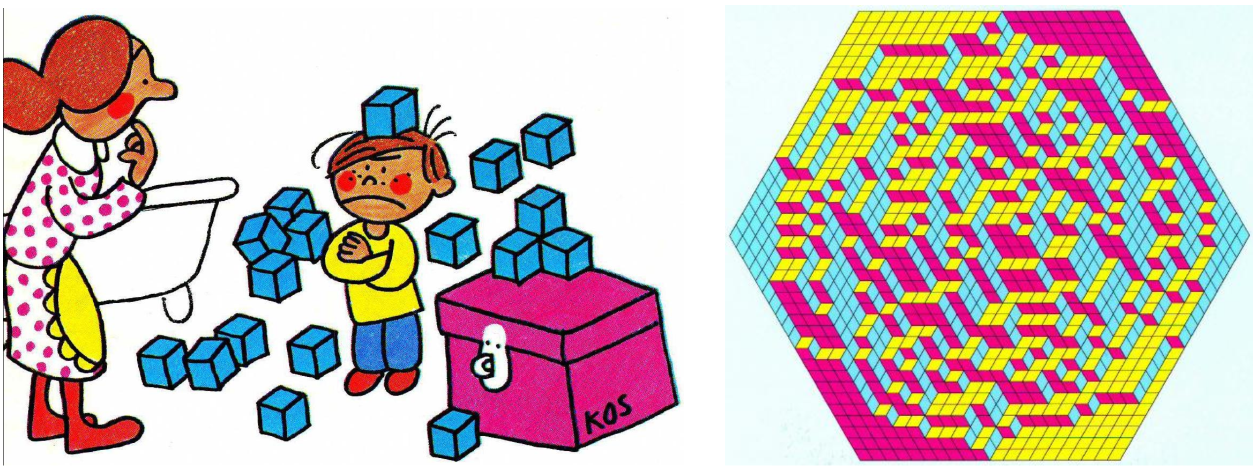
6.1 Diskretne večrazsežne porazdelitve – polinomska

Polinomska porazdelitev $P(n; p_1, p_2, \ldots, p_r)$, $\sum p_i = 1$, $\sum k_i = n$ je določena s predpisom

$$P(X_1 = k_1, \ldots, X_r = k_r) = \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r}.$$

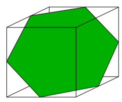
Koeficient šteje permutacije s ponavljanjem.[1] Za $r = 2$ dobimo binomsko porazdelitev, tj. $B(n, p) = P(n; p, q)$.

6.2 Ponovitev: dvojni integral



Slika: Kako bi najlažje prešteli vse kocke? Leva slika: 13 jih leži po tleh, ena je na fantovi glavi, na vijoličasti škatli pa so še 4 (čeprav četrte v resnici ne vidimo), torej jih je skupaj 18.

Sedaj pa poglejmo še desno sliko, kjer se kocke nahajajo znotraj večje kocke $20 \times 20 \times 20$, torej jih ne more biti več kot 8000, a izgleda kot da jih je ravno polovica. Po občutku jih je torej približno 4000.

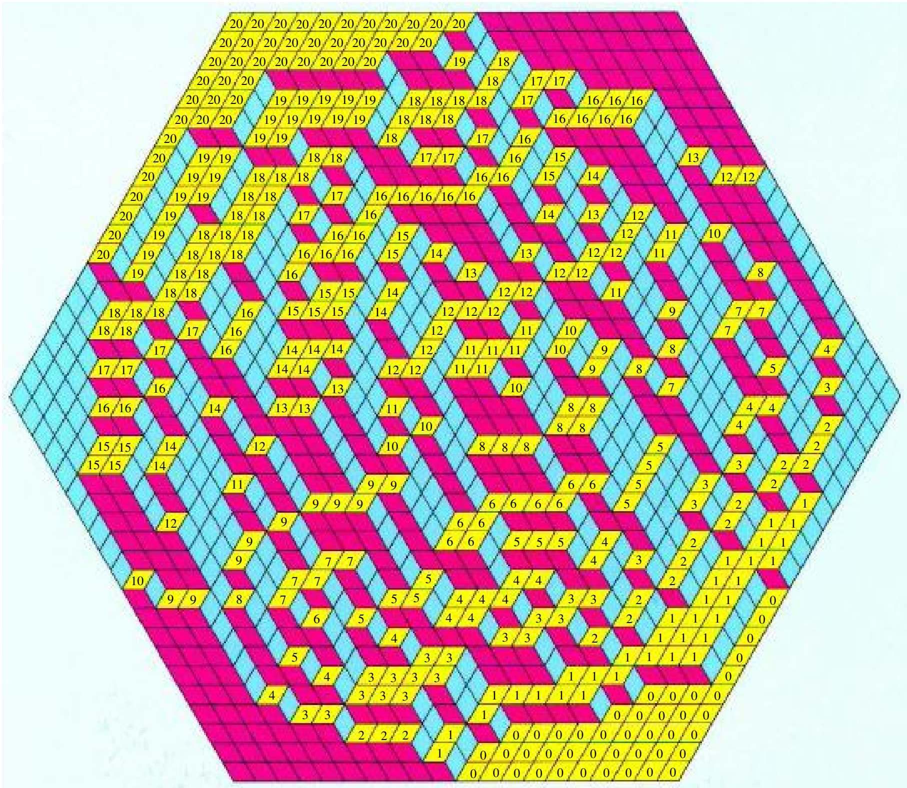


Če pa hočemo biti do kockice natančni, potem nam ne preostane drugega, kot da jih začnemo šteti. Problem pa ni samo v natančnosti, pač pa se moramo vprašati tudi katera 'polovica' nas zanima, kajti če sliko obrnemo na glavo, nam npr. zgornje ploskve doslej najvišjih kock postanejo tla.

Pa začnimo šteti po nivojih/slojih, ki so vzporedni s ploskvami kocke. Takoj vidimo, da imamo tri možnosti: v rumeni smeri (naj bo to z os), v modri smeri (naj bo to y os) in v vijolični smeri (ki naj bo x os). Mi se odločimo za rumeno in začnemo pri vrhu (te kocke namreč vse vidimo). V prvem stolpcu jih je 11, nato 10, pa 8, trikrat po 3 in sedemkrat po 1, skupaj torej $11 + 10 + 8 + 3 \times 3 + 7 \times 1 = 45$ (ki smo jih prešteli

[1]Glej npr. http://en.wikipedia.org/wiki/Multinomial_distribution.

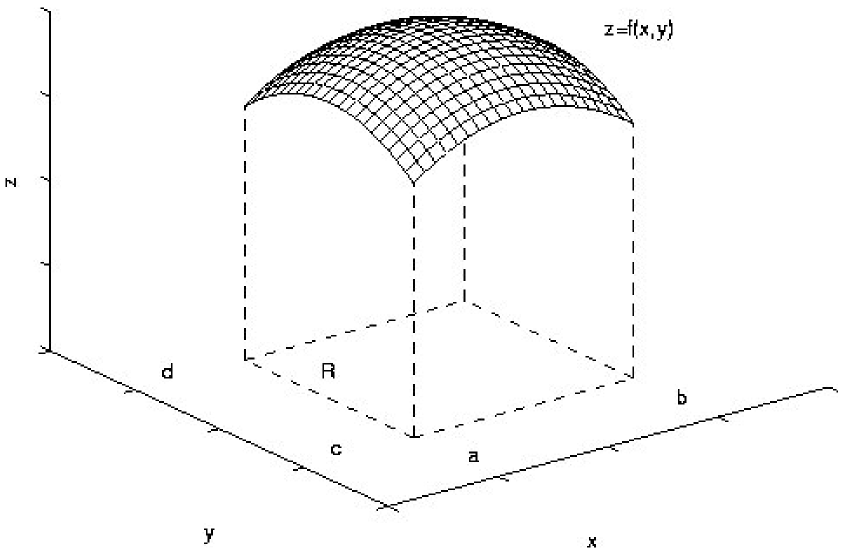
skoraj tako kot zidake pri integriranju). Ker pa so te kocke v 20 'nadstropju', pomnožimo njihovo število z 20 in dobimo 900. Nadaljujemo en nivo nižje. Kocke, ki jih ne vidimo, smo že šteli, tako da preštejemo še $0 + 1 + 0 + 2 \times 5 + 4 \times 2 + 4 \times 1 = 23$ kock in jih pomnožimo z 19, tako da dobimo 437. Tako nadaljujemo vse do konca:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
20 ·	(11 +	10 +	8 +	3 +	3 +	3 +	1 +	1 +	1 +	1 +	1 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	0 +	0)	= 20 · 45 = 900
19 ·	(0 +	1 +	0 +	5 +	5 +	2 +	2 +	2 +	2 +	1 +	1 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	0 +	0)	= 19 · 23 = 437
18 ·	(1 +	1 +	4 +	3 +	1 +	2 +	3 +	2 +	2 +	3 +	3 +	2 +	2 +	3 +	2 +	0 +	0 +	0 +	0 +	0)	= 18 · 34 = 612
17 ·	(2 +	1 +	0 +	1 +	2 +	0 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	1 +	1 +	2 +	0 +	0 +	0 +	0)	= 17 · 12 = 204
16 ·	(3 +	4 +	1 +	1 +	2 +	5 +	1 +	2 +	3 +	1 +	0 +	1 +	1 +	1 +	0 +	1 +	2 +	0 +	0 +	0)	= 16 · 29 = 612
15 ·	(0 +	0 +	1 +	1 +	0 +	0 +	1 +	1 +	0 +	2 +	3 +	0 +	0 +	0 +	0 +	0 +	0 +	2 +	2 +	0)	= 15 · 13 = 204
14 ·	(0 +	0 +	1 +	0 +	1 +	0 +	1 +	0 +	1 +	1 +	0 +	3 +	2 +	0 +	1 +	0 +	1 +	1 +	0 +	0)	= 14 · 13 = 464
13 ·	(1 +	0 +	0 +	1 +	0 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	1 +	2 +	0 +	0 +	0 +	0 +	0 +	0)	= 13 · 7 = 91
12 ·	(2 +	0 +	1 +	1 +	2 +	2 +	2 +	3 +	1 +	1 +	2 +	0 +	0 +	0 +	1 +	0 +	0 +	0 +	1 +	0)	= 12 · 19 = 228
11 ·	(0 +	0 +	1 +	1 +	0 +	1 +	0 +	1 +	3 +	2 +	0 +	1 +	0 +	0 +	0 +	1 +	0 +	0 +	0 +	0)	= 11 · 11 = 121
10 ·	(0 +	1 +	0 +	0 +	0 +	0 +	1 +	1 +	0 +	1 +	0 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	0 +	1)	= 10 · 7 = 70
9 ·	(0 +	0 +	0 +	0 +	1 +	0 +	1 +	1 +	0 +	0 +	0 +	0 +	0 +	2 +	3 +	1 +	1 +	1 +	0 +	2)	= 9 · 13 = 117
8 ·	(0 +	1 +	0 +	0 +	0 +	1 +	1 +	0 +	2 +	2 +	3 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	1 +	0)	= 8 · 11 = 88
7 ·	(0 +	0 +	2 +	1 +	0 +	0 +	1 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	2 +	2 +	1 +	0 +	0)	= 7 · 9 = 63
6 ·	(0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	2 +	4 +	2 +	2 +	0 +	0 +	0 +	1 +	0 +	0)	= 6 · 11 = 66
5 ·	(0 +	0 +	0 +	1 +	0 +	0 +	0 +	1 +	1 +	1 +	1 +	0 +	3 +	0 +	1 +	2 +	1 +	0 +	1 +	0)	= 5 · 13 = 65
4 ·	(0 +	1 +	0 +	0 +	2 +	1 +	0 +	0 +	0 +	0 +	0 +	1 +	1 +	2 +	3 +	2 +	1 +	0 +	1 +	1)	= 4 · 16 = 64
3 ·	(0 +	0 +	1 +	0 +	0 +	0 +	1 +	1 +	1 +	0 +	0 +	1 +	0 +	2 +	2 +	2 +	2 +	4 +	3 +	2)	= 3 · 22 = 66
2 ·	(0 +	0 +	0 +	1 +	1 +	2 +	1 +	1 +	1 +	1 +	1 +	1 +	1 +	1 +	1 +	0 +	0 +	0 +	0 +	3)	= 2 · 16 = 32
1 ·	(0 +	0 +	0 +	0 +	0 +	0 +	1 +	2 +	2 +	3 +	2 +	2 +	3 +	3 +	4 +	3 +	5 +	1 +	1 +	1)	= 1 · 33 = 33
0 ·	(0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	0 +	1 +	1 +	1 +	1 +	1 +	4 +	5 +	9 +	10 +	10)	= 0 · 43 = 0
vsota	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	400 4098

Vsota vseh števil v vseh oklepajih je ravno 400, kajti od zgoraj se vidjo vsa polja v mreži 20×20. Zadnje vrstice sicer ne potrebujemo, a je dobrodošla zaradi preiskusa. Načeloma so nas zanimale samo vsote števil v oklepajih, a tudi ta razčletnitev je v resnici koristna, saj se morajo števila v oklepajih po stolpcih sešteti v 20. Velja pa pripomniti še, da nam zgornja informacija zadostuje tudi za rekonstrukcijo celotne slike.

Končni odgovor je 4098. Če bi si sliko izpisali in bi jo ponesreči obrnili na glavo, bi dobili $8000 - 4098 = 3902 = 0 \cdot 45 + 1 \cdot 23 + 2 \cdot 34 + 3 \cdot 12 + 4 \cdot 29 + 5 \cdot 13 + 6 \cdot 13 + 7 \cdot 7 + 8 \cdot 19 + 9 \cdot 11 + 10 \cdot 7 + 11 \cdot 13 + 12 \cdot 11 + 13 \cdot 9 + 14 \cdot 11 + 15 \cdot 13 + 16 \cdot 16 + 17 \cdot 22 + 18 \cdot 16 + 19 \cdot 33 + 20 \cdot 44$. V resnici smo računali dvojni integral, to pa smo počeli s pomočjo običajnega integrala.



Dvojni integral predstavlja prostornino pod neko ploskvijo. Naj bo funkcija $z = f(x, y) \geqslant 0$ zvezna na nekem območju R v ravnini $\mathbb{R}^2$ (npr. kar $[a, b] \times [c, d]$). Prostornina telesa med ploskvijo, ki je podana z $z = f(x, y)$, in ravnino $z = 0$ je enaka dvojnemu integralu

$$\iint_R f(x,y)\,dxdy,$$

ki ga v primeru $R = [a, b] \times [c, d]$ izračunamo s pomočjo dvakratnega integrala

$$\int_c^d \left(\int_a^b f(x,y)\,dx \right) dy = \int_a^b \left(\int_c^d f(x,y)\,dy \right) dx.$$

Lastnosti dvojnega integrala

Trditev 6.1. (1) *Če je $f(x, y) \leqslant 0$ $\forall (x, y) \in R$, je vrednost dvojnega integrala negativna.*

(2) *Naj bo območje $R = R_1 \cup R_2$, kjer je $R_1 \cap R_2 = \emptyset$. Potem velja*

$$\iint_R f(x,y)\,dxdy = \iint_{R_1} f(x,y)\,dxdy + \iint_{R_2} f(x,y)\,dxdy.$$

(3) *Naj bo $f(x, y) \leqslant g(x, y)$, za vse točke $(x, y) \in R$, potem velja*

$$\iint_R f(x,y)\,dxdy \leqslant \iint_R g(x,y)\,dxdy. \qquad \square$$

Več o dvojnih integralih najdete npr. na:
http://www.math.oregonstate.edu/home/programs/undergrad/CalculusQuestStudyGuides/vcalc/255doub/255doub.html.

Računanje dvojnih integralov na pravokotnem območju se prevede na dva običajna (enkratna) integrala.

Kot bomo videli kasneje na primerih, pa je težje izračunati dvojni integral na območju, ki ni pravokotno, ampak je omejeno s poljubnimi krivuljami.

6.3 Zvezne večrazsežne porazdelitve

Slučajni vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ je **zvezno porazdeljen**, če obstaja integrabilna funkcija (**gostota verjetnosti**) $p(x_1, x_2, \ldots, x_n) \geqslant 0$ z lastnostjo

$$F(x_1, x_2, x_3, \ldots, x_n) = \int_{-\infty}^{x_1} \left(\int_{-\infty}^{x_2} \left(\cdots \left(\int_{-\infty}^{x_n} p(t_1, t_2, \ldots, t_n)\, dt_n \right) \cdots \right) dt_2 \right) dt_1$$

in $F(\infty, \infty, \infty, \ldots, \infty) = 1$.

Zvezne dvorazsežne porazdelitve

$$F(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} p(u, v)\, dv \right) du,$$

$$P\big((X, Y) \in (a, b] \times (c, d]\big) = \int_{a}^{b} \left(\int_{c}^{d} p(u, v)\, dv \right) du.$$

Velja

$$\frac{\partial F}{\partial x} = \int_{-\infty}^{y} p(x, v)\, dv \quad \text{in} \quad \frac{\partial^2 F}{\partial x \partial y} = p(x, y).$$

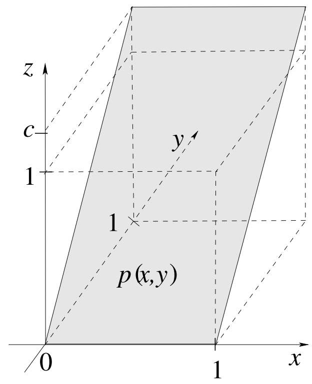
Robni verjetnostni gostoti sta

$$p_X(x) = F_X'(x) = \int_{-\infty}^{\infty} p(x, y)\, dy, \quad \text{in} \quad p_Y(y) = F_Y'(y) = \int_{-\infty}^{\infty} p(x, y)\, dx.$$

Primer: Naj bo gostota porazdelitve vektorja (X, Y) podana s

$$p(x, y) = \begin{cases} cy & \text{če je } 0 \leqslant x \leqslant 1 \text{ in } 0 \leqslant y \leqslant 1 \\ 0 & \text{sicer.} \end{cases}$$

Določi vrednost konstante c ter robni gostoti za slučajni spremenljivki X in Y!



Dvojni integral gostote verjetnosti je po eni strani enak 1, po drugi pa prostornini telesa, ki je pod osenčenim delom in nad ravnino xy, se pravi, da gre za polovico kvadra in znaša $1 \times 1 \times c \times 1/2$, od koder dobimo $c = 2$. Slučajna spremenljivka X je na intervalu $[0, 1]$ porazdeljena enakomerno, se pravi, da je $p(x) = 1$. To vidimo tako s slike (prečni prerez), kakor tudi iz definicije:

$$p_X(x) = \int_0^1 2y\, dy = y^2|_{y=0}^1 = 1.$$

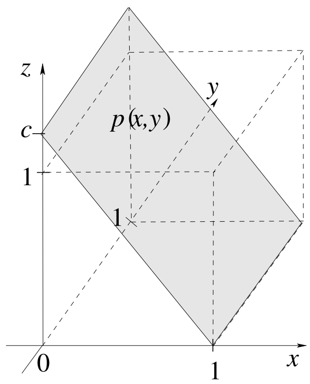
Za gostoto verjetnosti slučajne spremenljivke Y pa na intervalu $[0, 1]$ velja

$$p_Y(y) = \int_0^1 2y\, dx = 2xy|_{x=0}^1 = 2y.$$

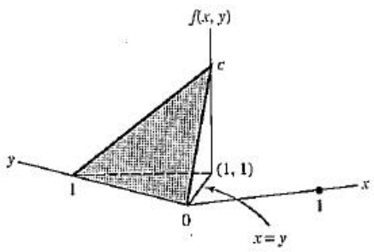
◇

Za vajo poskusi odgovoriti na ista vprašanja še za

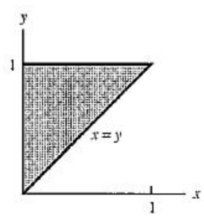
$$p(x, y) = \begin{cases} cx & \text{če je } 0 \leqslant x \leqslant 1 \text{ in } 0 \leqslant y \leqslant 1 \\ 0 & \text{sicer.} \end{cases}$$



Primer: Naj bo gostota porazdelitve slučajnega vektorja (X, Y) podana s $p(x, y) = f(x, y)$, kjer je



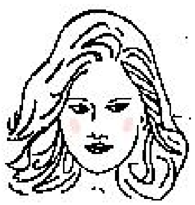
$$f(x, y) = \begin{cases} cx & \text{če je } 0 \leqslant x \leqslant y \text{ in } 0 \leqslant y \leqslant 1 \\ 0 & \text{sicer.} \end{cases}$$

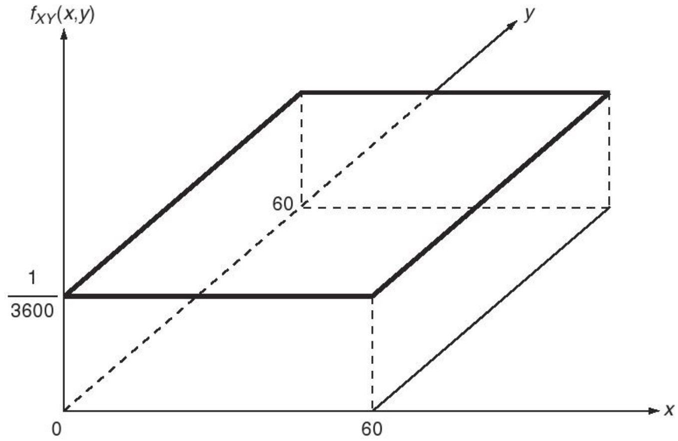


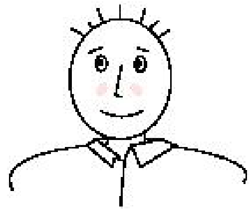
Izberi vrednost konstante c med:

(a) 6, **(b)** 2, **(c)** 3, **(d)** 1. ◇

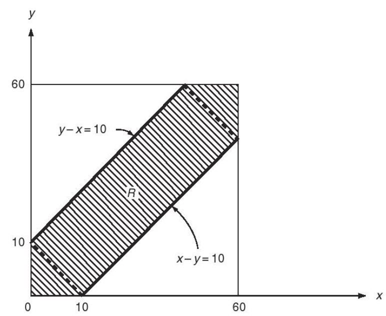
Primer: Dekle in fant se želita srečati na določenem mestu med 9-o in 10-o uro, pri čemer noben od njiju ne bo čakal drugega dlje od 10-ih minut. Če je vsak čas med 9-o in 10-o za vsakega od njiju enako verjeten, in sta njuna časa prihodov neodvisna, poišči verjetnost, da se bosta srečala. Naj bo čas prihoda fanta X minut po 9-i, pravtako pa naj bo čas prihoda dekleta Y minut po 9-i.







Ploskev, ki jo določa gostota porazdelitve, je ravnina, ker pa je prostornina pod njo enaka 1, je oddaljena od ravnine $z = 0$ za $1/3600$.



Prostornina, ki jo iščemo,
se nahaja nad področjem R,
ki je določeno z $|X - Y| \leqslant 10$,
torej je verjetnost srečanja enaka:

$$P(|X-Y| \leqslant 10) = \frac{(2 \times 5 \times 10 + 10\sqrt{2} \times 50\sqrt{2})}{3600} = \frac{11}{36}.$$

Pri bolj zapletenih gostotah verjetnosti, moramo dejansko izračunati integral

$$F(x,y) = \iint_R p(x,y)\, dydx.$$

Za vajo izračunajmo obe robni verjetnostni gostoti. Očitno velja:

$$F(x,y) = 0 \;\text{ za } (x,y) < (0,0) \qquad \text{in} \qquad F(x,y) = 1 \;\text{ za } (x,y) > (60,60).$$

Sedaj pa za $(0,0) \leqslant (x,y) \leqslant (60,60)$ velja

$$F(x,y) = \int_0^y \int_0^x \left(\frac{1}{3600}\right) dydx = \frac{xy}{3600}.$$

in

$$p_X(x) = F_X'(x) = \int_0^{60} \left(\frac{1}{3600}\right) dy = \frac{1}{60} \qquad \text{za} \quad 0 \leqslant y \leqslant 60,$$

$$p_Y(y) = F_Y'(y) = \int_0^{60} \left(\frac{1}{3600}\right) dx = \frac{1}{60} \qquad \text{za} \quad 0 \leqslant x \leqslant 60,$$

za vse ostale x in y pa je $p_X(x) = 0$ ter $p_Y(x) = 0$, torej sta X in Y obe enakomerno porazdeljeni slučajni spremenljivki na intervalu $[0,60]$. $\diamond$

Večrazsežna normalna porazdelitev

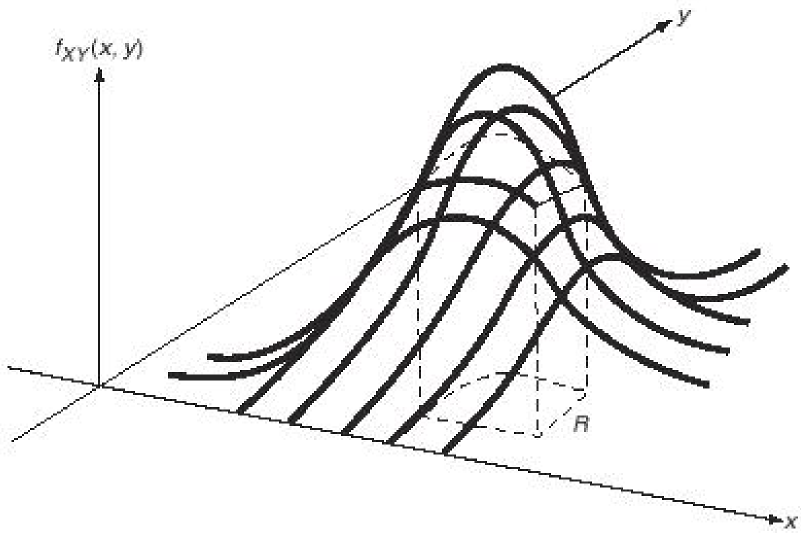
V dveh razsežnostih označimo normalno porazdelitev z $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ in ima gostoto

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\,\frac{x-\mu_x}{\sigma_x}\,\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)}.$$

V splošnem pa jo zapišemo v matrični obliki

$$p(\boldsymbol{x}) = \sqrt{\frac{\det A}{(2\pi)^n}}\, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T A(\boldsymbol{x}-\boldsymbol{\mu})},$$

kjer je A simetrična pozitivno definitna matrika, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, 'eksponent' T pa pomeni transponiranje (tj. zrcaljenje matrike preko glavne diagonale). Obe robni porazdelitvi sta normalni.



Primer: Pri študiju upora Y strukturnega elementa in sile X, ki deluje nanj, smatramo za slučajni spremenljivki. Verjetnost napake n_f je definirana z $P(Y \leqslant X)$. Predpostavimo, da je

$$p(x,y) = abe^{-(ax+by)} \qquad \text{za} \quad (x,y) > 0$$

in $p(x,y) = 0$ sicer, pri čemer sta a in b poznani pozitivni števili. Želimo izračunati n_f, tj.

$$F(x,y) = \iint_R p(x,y)\, dydx,$$

kjer je območje R določeno s pogojem $Y \leqslant X$. Ker slučajni spremenljivki X in Y zavzameta samo pozitivne vrednosti, velja

$$n_f = \int_0^\infty \int_y^\infty a\, b\, e^{-(ax+by)}\, dx\, dy = \int_0^\infty \int_0^x a\, b\, e^{-(ax+by)}\, dy\, dx.$$

Izračunajmo prvi integral. Upoštevamo $a\,dx = d(ax) = -d(-ax-by)$:

$$\int_0^\infty \int_y^\infty a\,b\,e^{-(ax+by)}\,dx\,dy = -b\int_0^\infty \left(\int_y^\infty e^{-(ax+by)}\,d(-ax-by)\right)dy$$

$$= -b\int_0^\infty \left(e^{-(ax+by)}\Big|_{x=y}^{\infty}\right)dy = b\int_0^\infty e^{-y(a+b)}\,dy$$

$$= \frac{-b}{a+b}\int_0^\infty e^{-y(a+b)}\,d\big(-y(a+b)\big) = \frac{-b}{a+b}\left(e^{-y(a+b)}\Big|_{y=0}^{\infty}\right) = \frac{b}{a+b}. \qquad \diamond$$

Za vajo izračunate tudi drugi dvojni integral.

6.4 Neodvisnost slučajnih spremenljivk

Podobno kot pri dogodkih pravimo za slučajne spremenljivke $X_1, X_2, \ldots, X_n$, da so med seboj **neodvisne**, če za poljubne vrednosti $x_1, x_2, \ldots, x_n \in \mathbb{R}$ velja

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)\cdot F_2(x_2)\cdots F_n(x_n),$$

kjer je F porazdelitvena funkcija vektorja, F_i pa so porazdelitvene funkcije njegovih komponent.

Trditev 6.2. *Diskretni slučajni spremenljivki X in Y z verjetnostnima tabelama*

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix} \quad \text{in} \quad Y \sim \begin{pmatrix} y_1 & y_2 & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}$$

ter verjetnostno funkcijo p_{ij} slučajnega vektorja (X,Y) sta X in Y neodvisni natanko takrat, ko je $p_{ij} = p_i\,q_j$ za vsak par naravnih števil i, j.

Dokaz. Zgornji pogoj za neodvisnost lahko zapišemo z verjetnostjo v naslednji obliki:

$$P(X \leqslant x_k,\ Y \leqslant y_\ell) = P(X \leqslant x_k)\cdot P(Y \leqslant y_\ell), \tag{6.2}$$

kjer sta k in ℓ poljubni naravni števili. Predpostavimo najprej, da je $p_{ij} = p_i\,q_j$ za vsak par naravnih števil i, j. Dokaz relacije (6.2) je sedaj precej direkten:

$$P(X \leqslant x_k,\ Y \leqslant y_\ell) = \sum_{i\leqslant k}\sum_{j\leqslant \ell} p_{ij} = \sum_{i\leqslant k}\sum_{j\leqslant \ell} p_i\cdot q_j = \sum_{i\leqslant k} p_i \cdot \sum_{j\leqslant \ell} q_j = P(X \leqslant x_k)\cdot P(Y \leqslant y_\ell).$$

Sedaj pa privzemimo pogoj (6.2). Zapišimo diskretno varianto relacije (6.1):

$$\begin{aligned} P(X = x_i,\ Y = y_j) = \;& P(X \leqslant x_{i+1}, Y \leqslant y_{j+1}) - P(X \leqslant x_{i+1},\ Y \leqslant y_j) \\ & - P(X \leqslant x_i,\ Y \leqslant y_{j+1}) + P(X \leqslant x_i,\ Y \leqslant y_j). \end{aligned}$$

Nariši si sliko, s katero se prepričaš o veljavnosti te relacije, nato pa uporabiš (6.2) na vsaki izmed verjetnosti na desni strani zgornje relacije:

$$\begin{aligned} p_{ij} &= P(X \leqslant x_{i+1}) \cdot P(Y \leqslant y_{j+1}) - P(X \leqslant x_i) \cdot P(Y \leqslant y_{j+1}) \\ &\quad - P(X \leqslant x_{i+1}) \cdot P(Y \leqslant y_j) + P(X \leqslant x_i) \cdot P(Y \leqslant y_j) \\ &= \big(P(X \leqslant x_{i+1}) - P(X \leqslant x_i)\big) \cdot \big(P(Y \leqslant y_{j+1}) - P(Y \leqslant y_j)\big) = p_i \cdot q_j. \end{aligned}$$

Sklicevanju na sliko pa se lahko izognemo na naslednji način. Najprej se ukvarjamo s spremenljivko X, tako da odštejemo naslednji enakosti:

$$\begin{aligned} P(X \leqslant x_i,\ Y \leqslant y_j) &= P(X \leqslant x_i) \cdot P(Y \leqslant y_j) \\ P(X \leqslant x_{i+1},\ Y \leqslant y_j) &= P(X \leqslant x_{i+1}) \cdot P(Y \leqslant y_j), \end{aligned}$$

kar nam da $P(X = x_i,\ Y \leqslant y_j) = p_i \cdot P(Y \leqslant y_j)$. Potem seveda velja tudi $P(X = x_i,\ Y \leqslant y_{j+1}) = p_i \cdot P(Y \leqslant y_{j+1})$, se pravi, da se lahko posvetimo še spremenljivki Y. Razlika zadnjih dveh relacij nam sedaj da $p_{ij} = p_i \cdot q_j$, kar smo želeli dokazati. □

Podobno dokažemo tudi naslednjo trditev za zvezno porazdeljeni slučajni spremenljivki (le da v prvemu delu namesto seštevanja integriramo, v drugem delu pa namesto odštevanja parcialno odvajamo).

Trditev 6.3. *Če sta X in Y zvezno porazdeljeni slučajni spremenljivki z gostotama p_X in p_Y ter je $p(x, y)$ gostota zvezno porazdeljenega slučajnega vektorja (X, Y), potem sta X in Y neodvisni natanko takrat, ko za vsak par x, y velja $p(x, y) = p_X(x) \cdot p_Y(y)$.* □

Primer: Naj bo dvorazsežni slučajni vektor (X, Y) z normalno porazdelitvijo $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$. Če je $\rho = 0$, je

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)} = p_X(x) \cdot p_Y(y).$$

Torej sta komponenti X in Y neodvisni. ◇

Brez dokaza pa omenimo še močnejšo trditev.

Izrek 6.4. *Zvezno porazdeljeni slučajni spremenljivki X in Y sta neodvisni natanko takrat, ko lahko gostoto $p(x, y)$ verjetnosti slučajnega vektorja (X, Y) zapišemo v obliki*

$$p(x, y) = f(x) \cdot g(y). \qquad \square$$

Naj bosta zvezno porazdeljeni slučajni spremenljivki X in Y tudi neodvisni ter A in B poljubni (Borelovi) podmnožici v $\mathbb{R}$. Potem sta neodvisna tudi dogodka $X \in A$ in $Y \in B$. Trditev velja tudi za diskretni slučajni spremenljivki X in Y. Pogosto pokažemo odvisnost spremenljivk X in Y tako, da najdemo množici A in B, za kateri je

$$P(X \in A, Y \in B) \neq P(X \in A) \cdot P(Y \in B).$$

Poglavje 7

Funkcije slučajnih spremenljivk in vektorjev

7.1 Funkcije slučajnih spremenljivk

Naj bo $X : G \to \mathbb{R}$ slučajna spremenljivka in $f : \mathbb{R} \to \mathbb{R}$ neka realna funkcija. Tedaj je njun kompozitum $Y = f \circ X$ določen s predpisom $Y(e) = f(X(e))$, za vsak $e \in G$, določa novo preslikavo $Y : G \to \mathbb{R}$. *Kdaj je tudi Y slučajna spremenljivka na $(G, \mathcal{D}, P)$?* V ta namen mora biti za vsak $y \in \mathbb{R}$ množica

$$(Y \leqslant y) = \{e \in G : Y(e) \leqslant y\} = \{e \in G : X(e) \in f^{-1}(-\infty, y]\}$$

dogodek – torej v $\mathcal{D}$. Če je ta pogoj izpolnjen, imenujemo Y **funkcija slučajne spremenljivke** X in jo zapišemo kar $Y = f(X)$. Njena porazdelitvena funkcija je v tem primeru

$$F_Y(y) = P(Y \leqslant y).$$

Če je funkcija f linearna, potem se porazdelitev verjetnosti ne spremeni, sicer pa se lahko, kot bomo videli v naslednjem primeru.

Primer: Naj bo diskretna slučajna spremenljivka X podana z

$$\begin{pmatrix} -1 & 0 & 1 \\ 1/2 & 1/3 & 1/6 \end{pmatrix}.$$

Potem je porazdelitev za slučajno spremenljivko $2X$ enaka

$$\begin{pmatrix} -2 & 0 & 2 \\ 1/2 & 1/3 & 1/6 \end{pmatrix}.$$

Sedaj pa izberimo še porazdelitev za slučajno spremenljivko X^2 med naslednjimi možnostmi:

(a) $\begin{pmatrix} -1 & 0 & 1 \\ 1/2 & 1/3 & 2/3 \end{pmatrix}$, **(b)** $\begin{pmatrix} -1 & 0 & 1 \\ 1/4 & 1/9 & 1/36 \end{pmatrix}$, **(c)** $\begin{pmatrix} -1 & 0 & 1 \\ 0 & 1/3 & 2/3 \end{pmatrix}$, **(d)** $\begin{pmatrix} -1 & 0 & 1 \\ 0 & 1/4 & 3/4 \end{pmatrix}$.

Hitro opazimo, da se v primeru (a) in (b) verjetnosti sploh ne seštejejo v 1, v primeru (d) pa se je spremenila verjetnost za vrednost 0, kar pa tudi ni mogoče. Ostane nam samo še možnost (c), ki pa je seveda prava, saj se verjetnost pri -1 spremenila v 0, verjetnost pri 0 pa je ostala nespremenjena. $\diamond$

Borelove množice


Emile Borel

Vprašanje: kakšna mora biti množica A, da je množica

$$X^{-1}(A) = \{e \in G : X(e) \in A\} \text{v } \mathcal{D}?$$

Zadoščajo množice A, ki so ali intervali, ali števne unije intervalov, ali števni preseki števnih unij intervalov – Borelove množice. **Kdaj je $f^{-1}(-\infty, y]$ Borelova množica?** Vsekakor je to res, ko je f zvezna funkcija. V nadaljevanju nas bodo zanimali samo taki primeri.

Primer: zvezne strogo naraščajoče funkcije

Naj bo $f : \mathbb{R} \to \mathbb{R}$ zvezna in strogo naraščajoča funkcija. Tedaj je taka tudi njena inverzna funkcija f^{-1} in velja

$$f^{-1}(-\infty, y] \quad = \quad \{x \in \mathbb{R} : f(x) \leqslant y\} = \{x \in \mathbb{R} : x \leqslant f^{-1}(y)\} = (-\infty, f^{-1}(y)]$$

ter potemtakem tudi $F_Y = F_X \circ f^{-1}$, o čemer se prepričamo takole

$$F_Y(y) = P(Y \leqslant y) = P(f(X) \leqslant y) = P(X \leqslant f^{-1}(y)) = F_X(f^{-1}(y)).$$

Če je X porazdeljena zvezno z gostoto $p(x)$, je

$$F_Y(y) = \int_{-\infty}^{f^{-1}(y)} p(x)\, dx$$

in, če je f odvedljiva, še

$$p_Y(y) = p(f^{-1}(y))f^{-1}(y)'.$$

Če funkcija ni monotona, lahko njeno definicijsko območje razdelimo na intervale monotonosti in obravnavamo vsak interval ločeno.

Primer: Obravnavajmo kvadrat normalno porazdeljene spremenljivke, tj. naj bo $X \sim N(0,1)$ in $Y = X^2$. Tedaj je $F_Y(y) = P(Y \leqslant y) = P(X^2 \leqslant y) = 0$ za $y \leqslant 0$, za $y > 0$ pa velja

$$F_Y(y) = P(|X| \leqslant \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

in ker je $p_X(x)$ soda funkcija

$$p_Y(y) = p_X(\sqrt{y})\frac{1}{2\sqrt{y}} + p_X(-\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{y}}p_X(\sqrt{y})$$

Vstavimo še standardizirano normalno porazdelitev

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}} & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

pa dobimo gostoto verjetnosti porazdelitve hi-kvadrat $\chi^2(1)$. $\diamond$

7.2 Funkcije in neodvisnost

Trditev 7.1. *Če sta X in Y neodvisni slučajni spremenljivki ter f in g zvezni funkciji na $\mathbb{R}$, sta tudi $U = f(X)$ in $V = g(Y)$ neodvisni slučajni spremenljivki.*

Dokaz. Za poljubna $u, v \in \mathbb{R}$ velja

$$\begin{aligned} P(U \leqslant u,\ V \leqslant v) &= P(f(X) \leqslant u,\ g(Y) \leqslant v) = P(X \in f^{-1}(-\infty, u],\ Y \in g^{-1}(-\infty, v]) \\ &\quad \text{(ker sta } X \text{ in } Y \text{ neodvisni, uporabimo Trditev 6.3)} \\ &= P(X \in f^{-1}(-\infty, u]) \cdot P(Y \in g^{-1}(-\infty, v]) \\ &= P(f(X) \leqslant u) \cdot P(g(Y) \leqslant v) = P(U \leqslant u) \cdot P(V \leqslant v). \quad \square \end{aligned}$$

Funkcije slučajnih vektorjev

Imejmo slučajni vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) : G \to \mathbb{R}^n$ in zvezno vektorsko preslikavo $f = (f_1, f_2, \ldots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$. Tedaj so $Y_j = f_j(X_1, X_2, \ldots, X_n)$, $j = 1, \ldots, m$ slučajne spremenljivke – komponente slučajnega vektorja $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$. Pravimo tudi, da je $\boldsymbol{Y}$ **funkcija slučajnega vektorja** $\boldsymbol{X}$, tj. $\boldsymbol{Y} = f(\boldsymbol{X})$. Porazdelitve komponent dobimo na običajen način

$$F_{Y_j}(y) = P(Y_j \leqslant y) = P\big(f_j(\boldsymbol{X}) \leqslant y\big) = P\big(\boldsymbol{X} \in f_j^{-1}(-\infty, y]\big)$$

in, če je $\boldsymbol{X}$ zvezno porazdeljen z gostoto $p(x_1, x_2, \ldots, x_n)$, potem je

$$F_{Y_j}(y) = \int\int \ldots \int_{f_j^{-1}(-\infty,y]} p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n.$$

Vsota slučajnih spremenljivk

Oglejmo si en enostaven primer. Naj bo $Z = X + Y$, kjer je (X, Y) zvezno porazdeljen slučajni vektor z gostoto $p(x, y)$. Tedaj je

$$F_Z(z) \;=\; P(Z \leqslant z) = P(X + Y \leqslant z) = \int\int_{x+y\leqslant z} p(x, y)\, dx\, dy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p(x, y)\, dy$$

in

$$p_Z(z) = F_Z'(z) = \int_{-\infty}^{\infty} p(x, z - x)\, dx = \int_{-\infty}^{\infty} p(z - y, y)\, dy.$$

Če sta spremenljivki X in Y neodvisni dobimo naprej zvezo

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)\, p_Y(z - x)\, dx.$$

Gostota $p_Z = p_X * p_Y$ je **konvolucija** funkcij p_X in p_Y.

Primer: Če je $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, je vsota $Z = X + Y$ zopet normalno porazdeljena: $Z \sim N(\mu_x + \mu_y, \sqrt{\sigma_x^2 + 2\rho\sigma_x\sigma_y + \sigma_y^2})$.

Če sta $X \sim \chi^2(n)$ in $Y \sim \chi^2(m)$ neodvisni slučajni spremenljivki, je tudi njuna vsota $Z = X + Y$ porazdeljena po tej porazdelitvi $Z \sim \chi^2(n + m)$. $\diamond$

Dosedanje ugotovitve lahko združimo v naslednjo trditev.

Trditev 7.2. *Če so $X_1, X_2, \ldots, X_n$ neodvisne standardizirano normalne slučajne spremenljivke, je slučajna spremenljivka $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ porazdeljena po $\chi^2(n)$.*

7.3 Funkcije slučajnih vektorjev

Naj bo sedaj $f : (x, y) \mapsto (u, v)$ transformacija slučajnega vektorja (X, Y) v slučajni vektor (U, V) določena z zvezama $u = u(x, y)$ in $v = v(x, y)$, torej je $U = u(X, Y)$ in $V = v(X, Y)$. Porazdelitveni zakon za nov slučajni vektor (U, V) je

$$F_{U,V}(u, v) \;=\; P(U < u, V < v) = P((U, V) \in A(u, v)) = P((X, Y) \in f^{-1}(A(u, v)).$$

Pri zvezno porazdeljenem slučajnem vektorju (X, Y) z gostoto $p(x, y)$ je

$$F_{U,V}(u, v) = \iint_{f^{-1}(A(u,v))} p(x, y)\, dx\, dy.$$

Če je f bijektivna z zveznimi parcialnimi odvodi, lahko nadaljujemo

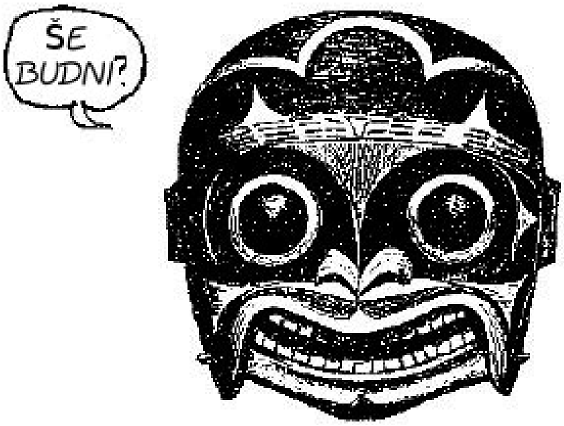
$$F_{U,V}(u,v) = \iint_{A(u,v)} p\big(x(u,v), y(u,v)\big)\, |J(u,v)|\, du\, dv,$$

kjer je (glej učbenik http://rkb.home.cern.ch/rkb/titleA.html)

$$J(u,v) = \frac{\partial(u,v)}{\partial(x,y)} = \det \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix}.$$

Jacobijeva determinanta (glej http://en.wikipedia.org/wiki/Jacobian za kakšen primer). Za gostoto $q(u,v)$ vektorja (U,V) dobimo od tu

$$q(u,v) = p\big(x(u,v), y(u,v)\big)\, |J(u,v)|.$$



Primer:

$$\Omega = \{(x,y)\,|\, 0 < x \leqslant 1,\ 0 < y \leqslant 1\}.$$

Naj bo

$$r = \sqrt{-2\log(x)}, \qquad \varphi = 2\pi y,$$
$$u = r\cos\varphi, \qquad v = r\sin\varphi.$$

Potem po pravilu za odvajanje posrednih funkcij in definiciji Jacobijeve matrike velja

$$\frac{\partial(u,v)}{\partial(x,y)} = \left(\frac{\partial(u,v)}{\partial(r,\varphi)}\right)\left(\frac{\partial(r,\varphi)}{\partial(x,y)}\right) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} \begin{pmatrix} \dfrac{-1}{rx} & 0 \\ 0 & 2\pi \end{pmatrix}.$$

Jacobijeva determinanta je

$$\det\left(\frac{d\boldsymbol{u}}{d\boldsymbol{x}}\right) = \det\left(\frac{\partial(u,v)}{\partial(x,y)}\right) = \det\left(\frac{\partial(u,v)}{\partial(r,\varphi)}\right)\det\left(\frac{\partial(r,\varphi)}{\partial(x,y)}\right) = r\frac{-2\pi}{rx} = \frac{-2\pi}{x}$$

in

$$d^2\boldsymbol{x} = \left|\det\left(\frac{d\boldsymbol{x}}{d\boldsymbol{u}}\right)\right| d^2\boldsymbol{u} = \left|\det\left(\frac{d\boldsymbol{u}}{d\boldsymbol{x}}\right)\right|^{-1} d^2\boldsymbol{u} = \frac{x}{2\pi}\ d^2\boldsymbol{u} = \frac{e^{-\dfrac{u^2+v^2}{2}}}{2\pi}\ d^2\boldsymbol{u}.$$

Od tod zaključimo, da za neodvisni slučajni spremenljivki x in y, ki sta enakomerno porazdeljeni med 0 in 1, zgoraj definirani slučajni spremenljivki u in v pravtako neodvisni in porazdeljeni normalno.

7.4 Pogojne porazdelitve

Naj bo B nek mogoč dogodek, tj. $P(B) > 0$. Potem lahko vpeljemo **pogojno porazdelitveno funkcijo**

$$F(x \mid B) = P(X \leqslant x \mid B) = \frac{P(X \leqslant x, B)}{P(B)}$$

V diskretnem primeru je:

$$p_{ik} = P(X = x_i, Y = y_k), \quad B = (Y = y_k) \quad \text{in} \quad P(B) = P(Y = y_k) = q_k.$$

Tedaj je pogojna porazdelitvena funkcija

$$F_X(x \mid y_k) \;=\; F_X(x \mid Y = y_k) = P(X \leqslant x \mid Y = y_k) = \frac{P(X < x, Y = y_k)}{P(Y = y_k)} = \frac{1}{q_k} \sum_{x_i \leqslant x} p_{ik}$$

Vpeljimo **pogojno verjetnostno funkcijo** s $p_{i|k} = \frac{p_{ik}}{q_k}$. Tedaj je $F_X(x \mid y_k) = \sum_{x_i \leqslant x} p_{i|k}$.

Primer: Nadaljujmo primer s katerim smo pričeli pogjavje 6. Zapiši pogojno verjetnostno porazdelitev slučajne spremenljivke X glede na pogoj $y = 2$!

$Y \setminus X$	1	2	3	4	Y
0	0	0·10	0·20	0·10	0·40
1	0·03	0·07	0·10	0·05	0·25
2	0·05	0·10	0·05	0	0·20
3	0	0·10	0·05	0	0·15
X	0·08	0·37	0·40	0·15	1

Verjetnosti v vrstici pri $y = 2$ moramo deliti s $P(Y = 2)$, ki je enaka 0·2:

$$X|y = 2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0{\cdot}25 & 0{\cdot}50 & 0{\cdot}25 & 0 \end{pmatrix}.$$ ◇

Primer: Dostavni tovornjak potuje od A do B in nazaj vsak dan. Na poti ima tri semaforje.

Naj bo X število rdečih semaforjev na katere naleti tovornjak na poti do dostavne točke B, in y število rdečih luči nazaj na poti do točke A. Inženir za promet je določil naslednjo verjetnostno porazdelitev:

$Y \setminus X$	0	1	2	3	Y
0	0·01	0·02	0·07	0·01	0·11
1	0·03	0·06	0·10	0·06	0·25
2	0·05	0·12	0·15	0·08	0·40
3	0·02	0·09	0·08	0·05	0·24
X	0·11	0·29	0·40	0·20	1

Poišči robno porazdelitev za Y.

$$Y \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{\cdot}11 & 0{\cdot}25 & 0{\cdot}40 & 0{\cdot}24 \end{pmatrix}.$$

Če vemo, da je tovornjak naletel na $x = 2$ luči do točke B, potem določi porazdelitev za Y.

$$Y|x = 2 \ \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 7/40 & 1/4 & 3/8 & 1/5 \end{pmatrix}. \qquad \diamond$$

Zvezne pogojne porazdelitve

Postavimo $B = (y < Y \leqslant y + h)$ za $h > 0$ in zahtevajmo $P(B) > 0$.

$$F_X(x \,|\, B) \ = \ P(X \leqslant x \,|\, B) = \frac{P(X \leqslant x,\ y < Y \leqslant y + h)}{P(y \leqslant Y < y + h)} = \frac{F(x, y + h) - F(x, y)}{F_Y(y + h) - F_Y(y)}.$$

Če obstaja limita (za $h \to 0$)

$$F_X(x|y) = F_X(x \,|\, Y = y) = \lim_{h\to 0} \frac{F(x, y + h) - F(x, y)}{F_Y(y + h) - F_Y(y)},$$

jo imenujemo **pogojna porazdelitvena funkcija** slučajne spremenljivke X glede na dogodek $(Y = y)$.

Gostota zvezne pogojne porazdelitve

Naj bosta gostoti $p(x, y)$ in $p_Y(y)$ zvezni ter $p_Y(y) > 0$. Tedaj je

$$F_X(x|y) = \lim_{h\to 0} \frac{\dfrac{F(x, y + h) - F(x, y)}{h}}{\dfrac{F_Y(y + h) - F_Y(y)}{h}} = \frac{\dfrac{\partial F}{\partial y}(x, y)}{F'_Y(y)} = \frac{1}{p_Y(y)} \int_{-\infty}^{x} p(u, y) du$$

oziroma, če vpeljemo **pogojno gostoto**

$$p_X(x|y) = \frac{p(x, y)}{p_Y(y)},$$

tudi $F_X(x|y) = \int_{-\infty}^{x} p_X(u|y)\, du$.

Primer: Za dvorazsežno normalno porazdelitev dobimo

$$p_X(x|y) \sim N\left(\mu_x + \rho\, \frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x \sqrt{1 - \rho^2}\right). \qquad \diamond$$

Poglavje 8

Momenti in kovarianca

8.1 Matematično upanje

Matematično upanje $\mathsf{E}X$ (oz. pričakovana vrednost) je posplošitev povprečne vrednosti diskretne spremenljivke $X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$, tj.

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{n} x_i k_i = \sum_{i=1}^{n} x_i f_i,$$

od koder izhaja

$$\mathsf{E}X = \sum_{i=1}^{n} x_i p_i.$$

Diskretna slučajna spremenljivka X z verjetnostno funkcijo p_k ima matematično upanje $\mathsf{E}X = \sum_{i=1}^{\infty} x_i p_i$, če je

$$\sum_{i=1}^{\infty} |x_i| p_i < \infty.$$

Zvezna slučajna spremenljivka X z gostoto $p(x)$ ima matematično upanje $\mathsf{E}X = \int_{-\infty}^{\infty} x p(x)\, dx$, če je

$$\int_{-\infty}^{\infty} |x| p(x)\, dx < \infty.$$

Primer: Poišči vsaj dve slučajni spremenljivki, za katere matematično upanje ne obstaja.

Diskretna: $x_k = (-1)^{k+1} 2^k / k$ in $p_k = 2^{-k}$.

Zvezna: $p_X(x) = \frac{1}{\pi(1+x^2)}$ – Cauchyeva porazdelitev.

V prvem primeru bi morala biti končna naslednja vsota:

$$S = \sum_{k=1}^{\infty} \frac{1}{k}.$$

Opazimo naslednje:

$$\frac{1}{3} + \frac{1}{4} \quad > \quad \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \qquad \text{in} \qquad \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \quad > \quad \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

ter v splošnem

$$\frac{1}{2^n + 1} + \cdots + \frac{1}{2^{n+1}} \quad > \quad \frac{2^n}{2^{n+1}} = \frac{1}{2}.$$

Na ta način pridemo do spodnje meje

$$\sum_{k=1}^{\infty} \frac{1}{k} \quad > \quad \sum_{n=1}^{\infty} \frac{1}{2},$$

od koder je očitno, da je vsota S neskončna. V drugem primeru pa

$$\int_{-\infty}^{\infty} \frac{|x|\,dx}{\pi(1+x^2)} = 2\int_{0}^{\infty} \frac{x\,dx}{\pi(1+x^2)} > \int_{0}^{\infty} \frac{d(x^2+1)}{\pi(1+x^2)} = \frac{1}{\pi}\int_{1}^{\infty} \frac{dz}{z} = \frac{1}{\pi}\ln z\Big|_{1}^{\infty} = \infty.$$

Velja omeniti, da je zadnji integral večji od S, tako da nam sploh ne bi bilo potrebno integrirati, da bi se prepričali, da integral ni končen. $\diamond$

Lastnosti matematičnega upanja

Naj bo a realna konstanta. Če je $P(X = a) = 1$, velja $\mathsf{E}X = a$.

Slučajna spremenljivka X ima matematično upanje natanko takrat, ko ga ima slučajna spremenljivka $|X|$. Očitno velja $|\mathsf{E}X| \leqslant \mathsf{E}|X|$. Za diskretno slučajno spremenljivko je $\mathsf{E}|X| = \sum_{i=1}^{\infty} |x_i| p_i$, za zvezno pa $\mathsf{E}|X| = \int_{-\infty}^{\infty} |x| p(x)\,dx$.

Velja splošno: matematično upanje funkcije $f(X)$ obstaja in je enako za diskretno slučajno spremenljivko $\mathsf{E}f(X) = \sum_{i=1}^{\infty} f(x_i) p_i$, za zvezno pa $\mathsf{E}f(X) = \int_{-\infty}^{\infty} f(x) p(x)\,dx$, če ustrezni izraz absolutno konvergira.

Primer: Za diskretno slučajno spremenljivko X je $E(X^2) = x_1^2 p_1 + x_2^2 p_1 + \cdots$ $\diamond$

Naj bo a realna konstanta. Če ima slučajna spremenljivka X matematično upanje, potem ga ima tudi spremenljivka aX in velja $\mathsf{E}(aX) = a\mathsf{E}X$. Če imata slučajni spremenljivki X in Y matematično upanje, ga ima tudi njuna vsota $X+Y$ in velja $\mathsf{E}(X+Y) = \mathsf{E}X + \mathsf{E}Y$. Za primer dokažimo zadnjo lastnost za zvezne slučajne spremenljivke. Naj bo p gostota slučajnega vektorja (X, Y) in $Z = X + Y$. Kot vemo, je $p_Z(z) = \int_{-\infty}^{\infty} p(x, z-x)\,dx$. Pokažimo najprej, da Z ima matematično upanje.

$$\begin{aligned}
\mathsf{E}|X+Y| &= \mathsf{E}|Z| = \int_{-\infty}^{\infty} |z|\,p_Z(z)\,dz \\
&= \int_{-\infty}^{\infty} |z| \left(\int_{-\infty}^{\infty} p(x, z-x)\,dx\right) dz = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |x+y|\,p(x,y)\,dx\right) dy \\
&\leqslant \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |x|\,p(x,y)\,dx\right) dy + \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |y|\,p(x,y)\,dx\right) dy \\
&= \int_{-\infty}^{\infty} |x|\,p_X(x)\,dx + \int_{-\infty}^{\infty} |y|\,p_Y(y)\,dy = \mathsf{E}|X| + \mathsf{E}|Y| < \infty.
\end{aligned}$$

Sedaj pa še zvezo

$$\begin{aligned}
\mathsf{E}(X+Y) &= \mathsf{E}Z = \int_{-\infty}^{\infty} z\,p_Z(z)\,dz \\
&= \int_{-\infty}^{\infty} z \left(\int_{-\infty}^{\infty} p(x, z-x)\,dx\right) dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)\,p(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x\,p(x,y)\,dx\right) dy + \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y\,p(x,y)\,dx\right) dy \\
&= \int_{-\infty}^{\infty} x\,p_X(x)\,dx + \int_{-\infty}^{\infty} y\,p_Y(y)\,dy = \mathsf{E}X + \mathsf{E}Y.
\end{aligned}$$

Torej je matematično upanje E **linearen funkcional**, tj.

$$\mathsf{E}(aX + bY) = a\mathsf{E}X + b\mathsf{E}Y.$$

Z indukcijo posplošimo to na poljubno končno število členov

$$\mathsf{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\mathsf{E}X_1 + a_2\mathsf{E}X_2 + \cdots + a_n\mathsf{E}X_n.$$

V Dodatku A.9 izpeljemo še naslednjo trditev.

Trditev 8.1. *Če obstajata matematični upanji* $\mathsf{E}X^2$ *in* $\mathsf{E}Y^2$, *obstaja tudi matematično upanje produkta* $\mathsf{E}XY$ *in velja ocena* $\mathsf{E}|XY| \leqslant \sqrt{\mathsf{E}X^2\mathsf{E}Y^2}$.

Enakost velja natanko takrat, ko velja $Y = \pm\sqrt{\mathsf{E}Y^2/\mathsf{E}X^2}\,X$ *z verjetnostjo* 1. □

Trditev 8.2. *Če sta slučajni spremenljivki, ki imata matematično upanje, neodvisni, obstaja tudi matematično upanje njunega produkta in velja* $\mathsf{E}XY = \mathsf{E}X \cdot \mathsf{E}Y$. □

Obstajajo tudi odvisne spremenljivke, za katere velja gornja zveza.

Primer: Naj bo slučajna spremenljivka X porazdeljena standardizirano normalno. Potem je $Y = X^2$ porazdeljena po $\chi^2(1)$. Velja tudi $\mathsf{E}X = 0$, $\mathsf{E}XY = \mathsf{E}(X^3) = 0$ in zato $\mathsf{E}XY = 0 = \mathsf{E}X \cdot \mathsf{E}Y$. Po drugi strani pa je $P(0 \leqslant X < 1,\ Y \geqslant 1) = 0$, $P(0 \leqslant X < 1) = \Phi(1) > 0$ in $P(Y \geqslant 1) = 1 - P(Y < 1) = 1 - P(-1 < X < 1) = 1 - 2\,\Phi(1) > 0$. ◇

Spremenljivki, za kateri velja $\mathsf{E}XY \neq \mathsf{E}X \cdot \mathsf{E}Y$ imenujemo **korelirani**.

Primer: Življenska doba varovalk (merjena v stotinah ur), ki jih uporabljamo pri računalniških monitorjih ima eksponentno porazdelitev s parametrom $\lambda = 5$. Vsak monitor ima dve varovalki, pri čemer ena deluje kot '*backup*' in prične delovati šele ko prva odpove.

- Če imata dve taki varovalki neodvisni življenski dobi X in Y, potem poišči gostoto porazdelitve $p(x, y)$.
- Efektivna skupna življenska doba dveh varovalk je $(X + Y)$. Poišči pričakovano skupno efektivno življensko dobo para dveh varovalk za monitor.

Odgovor: (a)

$$p(x, y) = \begin{cases} 25e^{-5(x+y)} & x, y > 0 \\ 0 & \text{sicer} \end{cases}$$

(b) 2/5. ◇

Primer: Koliko trčenj (rojstni dan na isti dan) lahko pričakujemo v skupini 100ih ljudi? ◇

8.2 Disperzija

Disperzija ali **varianca** $\mathsf{D}X$ slučajne spremenljivke, ki ima matematično upanje, je določena z izrazom

$$\mathsf{D}X = \mathsf{E}(X - \mathsf{E}X)^2.$$

Disperzija je vedno nenegativna, $\mathsf{D}X \geqslant 0$, je pa lahko tudi neskončna, tj. ne obstaja. Velja zveza

$$\mathsf{D}X = \mathsf{E}X^2 - (\mathsf{E}X)^2.$$

Naj bo a realna konstanta. Če je $P(X = a) = 1$, je $\mathsf{D}X = 0$. Iz linearnosti matematičnega upanja sledi tudi $\mathsf{D}(aX) = a^2\mathsf{D}X$ in $\mathsf{D}(X + a) = \mathsf{D}X$.

Trditev 8.3. *Če obstaja $\mathsf{D}X$ in je a realna konstanta, obstaja tudi $\mathsf{E}(X - a)^2$ in velja*

$$\mathsf{E}(X - a)^2 \geqslant \mathsf{D}X.$$

Enakost velja natanko za $a = \mathsf{E}X$.

Količino $\sigma X = \sqrt{\mathsf{D}X}$ imenujemo **standardna deviacija** ali **standardni odklon**. Za poljubno realno število a in slučajno spremenljivko X, za katero obstaja disperzija, velja $\sigma(aX) = a\,\sigma X$.

8.3 Standardizirane spremenljivke

Slučajno spremenljivko X **standardiziramo** s transformacijo

$$X_S = \frac{X-\mu}{\sigma},$$

kjer sta $\mu = \mathsf{E}X$ in $\sigma = \sqrt{\mathsf{D}X}$. Za X_S velja $\mathsf{E}X_S = 0$ in $\mathsf{D}X_S = 1$, saj je

$$\mathsf{E}X_S = \mathsf{E}\frac{X-\mu}{\sigma} = \frac{\mathsf{E}(X-\mu)}{\sigma} = \frac{\mu-\mu}{\sigma} = 0,$$

kjer smo upoštevali linearnost matematičnega upanja, ter

$$\mathsf{D}X_S = \mathsf{D}\frac{X-\mu}{\sigma} = \frac{\mathsf{D}(X-\mu)}{\sigma^2} = \frac{\sigma^2-0}{\sigma^2} = 1.$$

Matematična upanja in disperzije nekaterih porazdelitev

porazdelitev	$\mathsf{E}X$	$\mathsf{D}X$
binomska $B(n,p)$	np	npq
Poissonova $P(\lambda)$	λ	λ
Pascalova $P(m,p)$	m/p	mq/p^2
geometrijska $G(p)$	$1/p$	q/p^2
hipergeometrijska $H(n;M,N)$	nM/N	$\frac{M(N-M)n(N-n)}{N^2(N-1)}$
enakomerna zv. $E(a,b)$	$(a+b)/2$	$(b-a)^2/12$
normalna $N(\mu,\sigma)$	μ	σ^2
Gama $\Gamma(b,c)$	b/c	b/c^2
hi-kvadrat $\chi^2(n)$	n	$2n$

8.4 Kovarianca

Kovarianca $\mathsf{Cov}(X,Y)$ slučajnih spremenljivk X in Y je definirana z izrazom

$$\mathsf{Cov}(X,Y) = \mathsf{E}\big((X-\mathsf{E}X)(Y-\mathsf{E}Y)\big).$$

Zgornji izraz pa lahko poenostavimo na enak način kot pri varianci:

$$\mathsf{Cov}(X,Y) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y.$$

Velja tudi: $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$ (simetričnost) in

$\mathsf{Cov}(aX+bY,Z) = a\,\mathsf{Cov}(X,Z) + b\,\mathsf{Cov}(Y,Z)$ (bilinearnost).

V Dodatku A.9 izpeljemo še naslednjo trditev.

Trditev 8.4. *Če obstajata* $\mathsf{D}X$ *in* $\mathsf{D}Y$*, obstaja tudi* $\mathsf{Cov}(X,Y)$ *in velja*

$$|\mathsf{Cov}(X,Y)| \leqslant \sqrt{\mathsf{D}X\mathsf{D}Y} = \sigma X \sigma Y.$$

Enakost velja natanko takrat, ko je

$$Y - \mathsf{E}Y = \pm\frac{\sigma Y}{\sigma X}(X - \mathsf{E}X)$$

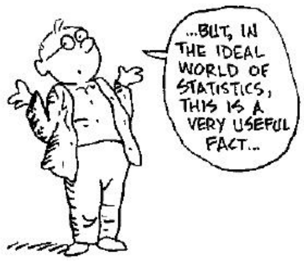
z verjetnostjo 1. □

Spremenljivki X in Y sta nekorelirani natanko takrat, ko je $\mathsf{Cov}(X,Y) = 0$. Če imata spremenljivki X in Y končni disperziji, jo ima tudi njuna vsota $X+Y$ in velja

$$\mathsf{D}(X+Y) = \mathsf{D}X + \mathsf{D}Y + 2\mathsf{Cov}(X,Y).$$

Če pa sta spremenljivki nekorelirani, je enostavno

$$\mathsf{D}(X+Y) = \mathsf{D}X + \mathsf{D}Y.$$



Zvezo lahko posplošimo na

$$\mathsf{D}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{D}X_i + \sum_{i\neq j} \mathsf{Cov}(X_i, X_j)$$

in za paroma nekorelirane spremenljivke

$$\mathsf{D}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{D}X.$$

Korelacijski koeficient

Korelacijski koeficient slučajnih spremenljivk X in Y je definiran z izrazom

$$r(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sigma X \sigma Y} = \frac{\mathsf{E}((X-\mathsf{E}X)(Y-\mathsf{E}Y))}{\sigma X \sigma Y}.$$

Primer: Za $(X,Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $r(X,Y) = \rho$. Torej sta normalno porazdeljeni slučajni spremenljivki X in Y neodvisni natanko takrat, ko sta nekorelirani. ◇

V splošnem velja:

$$-1 \leqslant r(X,Y) \leqslant 1.$$

$r(X,Y) = 0$ natanko takrat, ko sta X in Y nekorelirani.
$r(X,Y) = 1$ natanko takrat, ko je $Y = \frac{\sigma Y}{\sigma X}(X - \mathsf{E}X) + \mathsf{E}Y$ z verjetnostjo 1;

$r(X,Y) = -1$ natanko takrat, ko je $Y = -\frac{\sigma Y}{\sigma X}(X - \mathsf{E}X) + \mathsf{E}Y$ z verjetnostjo 1.

Torej, če je $|r(X,Y)| = 1$, obstaja med X in Y linearna zveza z verjetnostjo 1.

8.5 Pogojno matematično upanje

Pogojno matematično upanje je matematično upanje pogojne porazdelitve:

Diskretna slučajna spremenljivka X ima pri pogoju $Y = y_k$ pogojno verjetnostno funkcijo $p_{i|k} = p_{ik}/q_k$, $i = 1, 2, \dots$ in potemtakem pogojno matematično upanje

$$\mathsf{E}(X|y_k) = \sum_{i=1}^{\infty} x_i p_{i|k} = \frac{1}{q_k} \sum_{i=1}^{\infty} x_i p_{ik}.$$

Slučajna spremenljivka

$$\mathsf{E}(X|Y) \sim \begin{pmatrix} \mathsf{E}(X|y_1) & \mathsf{E}(X|y_2) & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}$$

ima enako matematično upanje kot spremenljivka X:

$$\mathsf{E}(\mathsf{E}(X|Y)) \;=\; \sum_{k=1}^{\infty} q_k \mathsf{E}(X|y_k) = \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} x_i p_{ik} = \sum_{i=1}^{\infty} x_i \sum_{k=1}^{\infty} p_{ik} = \sum_{i=1}^{\infty} x_i p_i = \mathsf{E}X.$$

Pogojno matematično upanje zvezne spremenljivke

Zvezna slučajna spremenljivka X ima pri pogoju $Y = y$ ima pogojno verjetnostno gostoto $p(x|y) = p(x, y)/p_Y(y)$, $x \in \mathbb{R}$ in potemtakem pogojno matematično upanje

$$\mathsf{E}(X|y) = \int_{-\infty}^{\infty} x p(x|y)\, dx = \frac{1}{p_Y(y)} \int_{-\infty}^{\infty} x p(x, y)\, dx.$$

Slučajna spremenljivka $\mathsf{E}(X|Y)$ z gostoto $p_Y(y)$ ima enako matematično upanje kot spremenljivka X

$$\mathsf{E}(\mathsf{E}(X|Y)) \;=\; \int_{-\infty}^{\infty} \mathsf{E}(X|y) p_Y(y)\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x p(x, y)\, dx\, dy = \int_{-\infty}^{\infty} x p_X(x) dx = \mathsf{E}X.$$

Regresijska funkcija

Preslikavo $x \mapsto \mathsf{E}(Y|x)$ imenujemo **regresija** slučajne spremenljivke Y glede na slučajno spremenljivko X.

Primer: Naj bo $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$. Tedaj je, kot vemo

$$p_X(x|y) \sim N\left(\mu_x + \rho \frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x \sqrt{1 - \rho^2}\right).$$

Torej je pogojno matematično upanje

$$\mathsf{E}(X|y) = \mu_x + \rho \frac{\sigma_x}{\sigma_y}(y - \mu_y)$$

in prirejena spremenljivka

$$\mathsf{E}(X|Y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y - \mu_y).$$

Na podoben način vpeljemo regresijo slučajne spremenljivke X glede na slučajno spremenljivko Y. Za dvorazsežno normalno porazdelitev dobimo

$$\mathsf{E}(Y|X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x).$$

Obe regresijski funkciji sta **linearni**.

Kovariančna matrika

Matematično upanje slučajnega vektorja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ je tudi vektor, in sicer $\mathsf{E}\boldsymbol{X} = (\mathsf{E}X_1, \mathsf{E}X_2, \ldots, \mathsf{E}X_n)$.

Primer: Za $(X,Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\mathsf{E}(X,Y) = (\mu_x, \mu_y)$. ◇

Naj slučajna spremenljivka Y prestavlja linearno kombinacijo spremenljivk $X_1, X_2, \ldots, X_n$. Spomnimo se, da je njeno matematično upanje

$$\mathsf{E}Y = \mathsf{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathsf{E}X_i.$$

Za disperzijo spremenljivke Y pa dobimo $\mathsf{D}Y = \mathsf{E}(Y - \mathsf{E}Y)^2 =$

$$\mathsf{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathsf{Cov}(X_i, X_j) = \boldsymbol{a}^T\boldsymbol{K}\boldsymbol{a},$$

kjer je $\mathsf{Cov}(X_i, X_j) = \mathsf{E}\big((X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j)\big)$ kovarianca spremenljivk X_i in X_j, $\boldsymbol{K} = [\mathsf{Cov}(X_i, X_j)]$ **kovariančna matrika** vektorja X, ter $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$.

Lastnosti kovariančne matrike

Kovariančna matrika $\boldsymbol{K} = [K_{ij}]$ je *simetrična*: $K_{ij} = K_{ji}$. Diagonalne vrednosti so disperzije spremenljivk: $K_{ii} = \mathsf{D}X_i$. Ker je $\boldsymbol{a}^T\boldsymbol{K}\boldsymbol{a} = \mathsf{D}Y \geqslant 0$, je pozitivno semidefinitna matrika. Naj bo $\boldsymbol{a}$, $||\boldsymbol{a}|| = 1$ lastni vektor, ki pripada lastni vrednosti λ kovariančne matrike $\boldsymbol{K}$, tj. $\boldsymbol{K}\boldsymbol{a} = \lambda\boldsymbol{a}$. Tedaj je $0 \leqslant \mathsf{D}Y = \boldsymbol{a}^T\boldsymbol{K}\boldsymbol{a} = \lambda$, kar pomeni, da so vse lastne vrednosti kovariančne matrike nenegativne. Če je kaka lastna vrednost enaka 0, je vsa verjetnost skoncentrirana na neki hiperravnini – porazdelitev je *izrojena*. To se zgodi natanko takrat, ko kovariančna matrika $\boldsymbol{K}$ ni obrnljiva, oziroma ko je $\det \boldsymbol{K} = 0$.

Primer: Za $(X,Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\boldsymbol{K} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$.

Ker je $|\rho| < 1$, je $\det \boldsymbol{K} = \sigma_x^2\sigma_y^2(1 - \rho^2) > 0$ in je potemtakem porazdelitev vedno neizrojena.

Za $N(\boldsymbol{\mu}, \boldsymbol{A})$ je $\boldsymbol{K} = \boldsymbol{A}^{-1}$. ◇

Poglejmo še, kako se spremeni kovariančna matrika pri linearni transformaciji vektorja $X' = \boldsymbol{A}X$, kjer je $\boldsymbol{A}$ poljubna matrika reda $n \times n$. Vemo, da je $\mathsf{D}(\boldsymbol{a}^T X) = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a}$. Tedaj je, če označimo kovariančno matriko vektorja X' s $\boldsymbol{K}'$,

$$\boldsymbol{a}^T \boldsymbol{K}' \boldsymbol{a} = \mathsf{D}(\boldsymbol{a}^T X') = \mathsf{D}(\boldsymbol{a}^T \boldsymbol{A} X) = \mathsf{D}((\boldsymbol{A}^T \boldsymbol{a})^T X) = (\boldsymbol{A}^T \boldsymbol{a})^T \boldsymbol{K} (\boldsymbol{A}^T \boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T \boldsymbol{a}$$

in potemtakem

$$\boldsymbol{K}' = \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T.$$

8.6 Višji momenti

Višji momenti so posplošitev pojmov matematičnega upanja in disperzije. **Moment reda** $k \in \mathbb{N}$ *glede na točko* $a \in \mathbb{R}$ imenujemo količino

$$m_k(a) = \mathsf{E}((X-a)^k).$$

Moment obstaja, če obstaja matematično upanje $\mathsf{E}(|X-a|^k) < \infty$. Za $a = 0$ dobimo **začetni moment** $z_k = m_k(0)$; za $a = \mathsf{E}X$ pa **centralni moment** $m_k = m_k(\mathsf{E}X)$.

Primer: $\mathsf{E}X = z_1$ in $\mathsf{D}X = m_2$. ◇

Če obstaja moment $m_n(a)$, potem obstajajo tudi vsi momenti $m_k(a)$ za $k < n$. Če obstaja moment z_n, obstaja tudi moment $m_n(a)$ za vse $a \in \mathbb{R}$.

$$m_n(a) = \mathsf{E}((X-a)^n) = \sum_{k=0}^{n} \binom{n}{k} (-a)^{n-k} z_k.$$

Posebej za centralni moment velja

$$m_n = m_n(z_1) = \sum_{k=0}^{n} \binom{n}{k} (-z_1)^k z_{n-k}$$

$m_0 = 1$, $m_1 = 0$, $m_2 = z_2 - z_1^2$, $m_3 = z_3 - 3z_2z_1 + 2z_1^3$, ...

Asimetrija spremenljivke X imenujemo količino $A(X) = \dfrac{m_3}{\sigma^3}$.

Sploščenost spremenljivke X imenujemo količino $K(X) = \dfrac{m_4}{\sigma^4} - 3$, kjer je $\sigma = \sqrt{m_2}$.

Za simetrično glede na $z_1 = \mathsf{E}X$ porazdeljene spremenljivke so vsi lihi centralni momenti enaki 0.

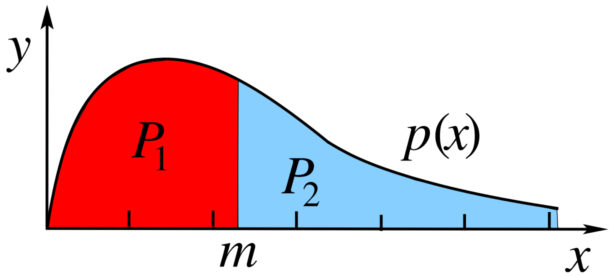
Primer: Za $X \sim N(\mu, \sigma)$ so $m_{2k+1} = 0$ in $m_{2k} = (2k-1)!!\sigma^{2k}$. Zato sta tudi $A(X) = 0$ in $K(X) = 0$. ◇

Če sta spremenljivki X in Y neodvisni, je $m_3(X+Y) = m_3(X) + m_3(Y)$.

Primer: Za binomsko porazdeljeno spremenljivko $X \sim B(n,p)$ pa je

$$m_3(X) = npq(q-p) \quad \text{in dalje} \qquad A(X) = \frac{q-p}{\sqrt{npq}}. \qquad \diamond$$

Kadar spremenljivka nima momentov, uporabljamo kvantile. **Kvantil reda** $p \in (0,1)$ je vsaka vrednost $x \in \mathbb{R}$, za katero velja $P(X \leqslant x) \geqslant p$ in $P(X \geqslant x) \geqslant 1-p$ oziroma $F(x) \leqslant p \leqslant F(x+)$. Kvantil reda p označimo z x_p. Za zvezno spremenljivko je $F(x_p) = p$. Kvantil $x_{\frac{1}{2}}$ imenujemo **mediana** (glej sliko);



Slika: Navpična črta pri vrednosti m razdeli ploščino med grafom gostote verjetnosti $p(x)$ in x-osjo na dva dela: P_1 in P_2. Ker je ploščina pod krivuljo enaka 1, velja $P_1 + P_2 = 1$. Vemo tudi, da je $P(X < m) = P_1$ in $P(X \geqslant m) = P_2$. Če je torej m mediana, potem je $P_1 = P_2$. Z drugimi besedami, *mediana* je tista vrednost oziroma tisto realno (predstavljeno na osi x), pri katerem ustrezna napičnica razdeli ploščino grafom $y = p(x)$ in x-osjo na dva enaka dela. Če je krivulja $y = p(x)$ simetrična, potem je mediana ravno na sredini.

Vrednosti $x_{\frac{i}{4}}$, $i = 0,1,2,3,4$ so **kvartili**. Kot nadomestek za standardni odklon uporabljamo **kvartilni razmik**

$$\frac{x_{\frac{3}{4}} - x_{\frac{1}{4}}}{2}$$

Poglavje 9

Karakteristične funkcije in limitni izreki

9.1 Karakteristična funkcija

Naj bo Z kompleksna slučajna spremenljivka, tj. $Z = X + iY$ za slučajni spremenljivki X in Y. Njeno upanje izračunamo z

$$E(Z) = E(X) + iE(Y),$$

disperzijo pa z

$$D(Z) = E(|Z - E(Z)|^2) = D(X) + D(Y).$$

Kompleksna funkcija realne slučajne spremenljivke je kompleksna slučajna spremenljivka, npr. e^{iX}.

Karakteristična funkcija realne slučajne spremenljivke X je kompleksna funkcija $\varphi_X(t)$ realne spremenljivke t določena z zvezo $\varphi_X(t) = \mathsf{E}e^{itX}$. Karakteristične funkcije vedno obstajajo in so močno računsko orodje. Posebej pomembni lastnosti sta:

Če obstaja začetni moment z_n, je karakteristična funkcija n-krat odvedljiva v vsaki točki in velja $\varphi_X^{(k)}(0) = i^k z_k$. Za neodvisni spremenljivki X in Y je $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$. Pojem karakteristične funkcije lahko posplošimo tudi na slučajne vektorje.

Reprodukcijska lastnost normalne porazdelitve

Vsaka linearna kombinacija *neodvisnih* in *normalno* porazdeljenih slučajnih spremenljivk je tudi sama **normalno** porazdeljena.

Če so slučajne spremenljivke $X_1, \ldots, X_n$ neodvisne in normalno porazdeljene $N(\mu_i, \sigma_i)$, potem je njihova vsota tudi normalno porazdeljena:

$$N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right).$$

Da ne bi vsota povprečij rastla z n, nadomestimo vsoto spremenljivk X_i z njihovim provprečjem $\overline{X}$ in dobimo

$$N\left(\overline{\mu}, \sqrt{\sum \left(\frac{\sigma_i}{n}\right)^2}\right).$$

Če privzamemo $\mu_i = \mu$ in $\sigma_i = \sigma$, dobimo $N\big(\mu, \sigma/\sqrt{n}\big)$.

9.2 Limitni izreki

Zaporedje slučajnih spremenljivk X_n **verjetnostno konvergira** k slučajni spremenljivki X, če za vsak $\varepsilon > 0$ velja

$$\lim_{n \to \infty} P(|X_n - X| \geqslant \varepsilon) = 0$$

ali enakovredno

$$\lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1.$$

Zaporedje slučajnih spremenljivk X_n **skoraj gotovo konvergira** k slučajni spremenljivki X, če velja

$$P(\lim_{n \to \infty} X_n = X) = 1.$$

Če zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X, potem za vsak $\varepsilon > 0$ velja

$$\lim_{m \to \infty} P(|X_n - X| < \varepsilon \quad \text{za vsak} \quad n \geqslant m) = 1.$$

Od tu izhaja:

če konvergira skoraj gotovo $X_n \to X$,
potem konvergira tudi verjetnostno $X_n \to X$.

Šibki in krepki zakon velikih števil

Naj bo $\boldsymbol{X_1, \ldots, X_n}$ zaporedje spremenljivk, ki imajo matematično upanje. Označimo $\boldsymbol{S_n} = \sum_{k=1}^n X_k$ in

$$\boldsymbol{Y_n} = \frac{S_n - \mathsf{E}S_n}{n} = \frac{1}{n}\sum_{k=1}^n (X_k - \mathsf{E}X_k) = \frac{1}{n}\sum_{k=1}^n X_k - \frac{1}{n}\sum_{k=1}^n \mathsf{E}X_k.$$

Pravimo, da za zaporedje slučajnih spremenljivk X_k velja:

- **šibki zakon velikih števil**, če gre verjetnostno $Y_n \to 0$, tj., če $\forall\varepsilon > 0$ velja
$$\lim_{n\to\infty} P\left(\left|\frac{S_n - \mathsf{E}S_n}{n}\right| < \varepsilon\right) = 1;$$
- **krepki zakon velikih števil**, če gre skoraj gotovo $Y_n \to 0$, tj., če velja
$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1.$$

Če za zaporedje $X_1, \ldots, X_n$ velja krepki zakon, velja tudi šibki.

Neenakost Čebiševa

Če ima slučajna spremenljivka X končno disperzijo, tj. $\mathsf{D}X < \infty$, velja za vsak $\varepsilon > 0$ **neenakost Čebiševa**

$$P(|X - \mathsf{E}X| \geqslant \varepsilon) \leqslant \frac{\mathsf{D}X}{\varepsilon^2}.$$

Dokaz: Pokažimo jo za zvezne spremenljivke

$$P(|X - \mathsf{E}X| \geqslant \varepsilon) = \int_{|x-\mathsf{E}X|\geqslant\varepsilon} p(x)\,dx = \frac{1}{\varepsilon^2}\int_{|x-\mathsf{E}X|\geqslant\varepsilon} \varepsilon^2 p(x)\,dx$$

$$\leqslant \frac{1}{\varepsilon^2}\int_{-\infty}^{\infty} (x - \mathsf{E}X)^2 p(x)\,dx = \frac{\mathsf{D}X}{\varepsilon^2}. \qquad \square$$

Posledice neenakosti Čebiševa

Izrek 9.1. *(**Markov**) Če gre za zaporedje slučajnih spremenljivk X_i izraz*

$$\frac{\mathsf{D}S_n}{n^2} \to 0,$$

ko gre $n \to \infty$, velja za zaporedje šibki zakon velikih števil.

Izrek 9.2. ***(Čebišev)*** *Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C, tj.*

$$\mathsf{D}X_i < C \qquad \text{za vsak } i,$$

velja za zaporedje šibki zakon velikih števil.

Za Bernoullijevo zaporedje X_i so spremenljivke paroma neodvisne, $\mathsf{D}X_i = pq$, $S_n = k$. Pogoji izreka Čebiševa so izpolnjeni in zopet smo prišli do Bernoullijevega zakona velikih števil, tj. izreka 5.4.

Še nekaj izrekov

Izrek 9.3. ***(Hinčin)*** *Če so neodvisne slučajne spremenljivke X_i enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = a$ za vsak i, potem velja zanje šibki zakon velikih števil, tj. za vsak $\varepsilon > 0$ je*

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - a\right| < \varepsilon\right) = 1.$$

Izrek 9.4. ***(Kolmogorov)*** *Če so slučajne spremenljivke X_i neodvisne, imajo končno disperzijo in velja $\sum_{n=1}^{\infty} \frac{\mathsf{D}S_n}{n^2} < \infty$, potem velja krepki zakon velikih števil:*

$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1.$$

Izrek 9.5. ***(Kolmogorov)*** *Če so slučajne spremenljivke X_i neodvisne,*
enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = \mu$,
potem velja krepki zakon velikih števil

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

Izrek 9.6. ***(Borel** 1909)* *Za Bernoullijevo zaporedje velja*

$$P\left(\lim_{n\to\infty} \frac{k}{n} = p\right) = 1.$$

9.3 Centralni limitni izrek (CLI)

Leta 1810 je Pierre Laplace (1749-1827) študiral anomalije orbit Jupitra in Saturna, ko je izpeljal razširitev De Moivrovega limitnega izreka,

"Vsaka vsota ali povprečje, če je število členov dovolj veliko, je približno normalno porazdeljena."

Centralni limitni zakon

Opazujmo sedaj zaporedje standardiziranih spremenljivk

$$Z_n = \frac{S_n - \mathsf{E}S_n}{\sigma(S_n)}.$$

Za zaporedje slučajnih spremenljivk X_i velja **centralni limitni zakon**, če porazdelitvene funkcije za Z_n gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve, to je, če za vsak $x \in \mathbb{R}$ velja

$$\lim_{n\to\infty} P\left(\frac{S_n - \mathsf{E}S_n}{\sigma(S_n)} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

(Osnovni CLI) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon.

Skica dokaza centralnega limitnega izreka

Naj bo $Z_i = \dfrac{X_i - \mu}{\sigma}$. Potem je

$$M_Z(t) = 1 - \frac{t^2}{2!} + \frac{t^3}{3!}E(Z_i^3) + \cdots$$

Za $\quad Y_n = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \dfrac{1}{\sqrt{n}} \, \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma} = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^n Z_i \quad$ velja

$$M_n(t) = \left[M_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\,k + \cdots\right)^n,$$

kjer je $k = E(Z_i^3)$.

$$\log M_n(t) = n \log\left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\,k + \cdots\right)$$

Za $x = \left(-\dfrac{t^2}{2\,n} + \dfrac{t^3}{3!\,n^{3/2}} + \cdots\right)$ velja

$$\log M_n(t) = n \log(1 + x) = n(x - \frac{x^2}{2} + \cdots) =$$

$$n\left[\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right) - \frac{1}{2}\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right)^2 + \cdots\right]$$

in od tod končno še

$$\lim_{n\to\infty} \log M_n(t) = -\frac{t^2}{2} \qquad \text{oziroma} \qquad \lim_{n\to\infty} M_n(t) = e^{-t^2/2}.$$

Iz konvergence karakterističnih funkcij φ_{Y_n} proti karakteristični funkciji standardizirano normalne porazdelitve lahko sklepamo po obratnem konvergenčnem izreku, da tudi porazdelitvene funkcije za Y_n konvergirajo proti porazdelitveni funkciji standardizirano normalne porazdelitve. Torej velja centralni limitni zakon. $\square$

Del II

STATISTIKA

Skozi življenje se prebijamo z odločitvami,
ki jih naredimo na osnovi nepopolnih informacij ...

Pridobivanje podatkov

Novice so polne številk. Televizijski napovedovalec pove, da se je stopnja nezaposlenosti zmanjšala na 4·7%. Raziskava trdi, da je 45% Američanov zaradi kriminala strah ponoči zapustiti domove. Od kod pridejo te številke? Ne vprašamo vseh ljudi, če so zaposleni ali ne. Raziskovalne agencije vprašajo le nekaj posameznikov, če zaradi strahu pred ropi ostajajo ponoči doma. Vsak dan se v novicah pojavi nov naslov. Eden od teh trdi: Aspirin preprečuje srčne infarkte. Nadaljnje branje razkrije, da je raziskava obravnavala 22 tisoč zdravnikov srednjih let. Polovica zdravnikov je vsak drugi dan vzela aspirin, druga polovica pa je dobila neaktivno tableto. V skupini, ki je jemala aspirin, je 139 zdravnikov doživelo srčni infarkt. V drugi skupini je bilo v enakem časovnem obdobju 239 infarktov. Ali je ta razlika dovolj velika, da lahko trdimo, da aspirin res preprečuje srčne infarkte?

Da bi ubežali neprijetnostim kot sta nezaposlenost in srčni infarkt, prižgimo televizijo. V pogovorni oddaji voditelj povabi gledalce, da sodelujejo v anketi. Tema pogovora je dobrodelnost in voditelja zanima, če gledalci redno prispevajo denar ali oblačila v dobrodelne namene. Med oddajo sprejmejo 50 tisoč klicev in 83% gledalcev trdi, da redno sodelujejo v tovrstnih akcijah. Ali je res, da smo tako zelo humanitarno osveščeni? Zanesljivost teh številk je v prvi vrsti odvisna od njihovega izvora. Podatkom o nezaposlenosti lahko zaupamo, v tistih 83% iz pogovorne oddaje pa najbrž lahko utemeljeno podvomimo. Naučili se bomo prepoznati dobre in slabe metode pridobivanja podatkov. Razumevanje metod, s katerimi lahko pridobimo zaupanja vredne podatke, je prvi (in najpomembnejši) korak k pridobivanju sposobnosti odločanja o pravilnosti sklepov, ki jih izpeljemo na osnovi danih podatkov. Izpeljava zaupanja vrednih metod za pridobivanje podatkov je področje, kjer vstopimo v svet statistike, znanosti o podatkih.

Obdelava podatkov

Za sodobno družbo je značilna poplava podatkov. Podatki, ali numerična dejstva, so bistveni pri odločanju na skoraj vseh področjih življenja in dela. Kot druge velike poplave nam poplava podatkov grozi, da nas bo pokopala pod sabo. Moramo jo kontrolirati s premišljeno organizacijo in interpretacijo podatkov. Baza podatkov kakšnega podjetja na primer vsebuje velikansko število podatkov: o zaposlenih, prodaji, inventarju, računih strank, opremi, davkih in drugem. Ti podatki so koristni le v primeru, ko jih lahko organiziramo in predstavimo tako, da je njihov pomen jasen. Posledice neupoštevanja podatkov so lahko hude. Veliko bank je izgubilo na milijarde dolarjev pri nedovoljenih špekulacijah njihovih zaposlenih, ki so ostale skrite med goro podatkov, ki jih odgovorni niso dovolj pozorno pregledali.

Statistično sklepanje

Sklepanje je proces, pri katerem pridemo do zaključkov na podlagi danih dokazov. Dokazi so lahko v mnogo različnih oblikah. V sojenju zaradi umora jih lahko predstavljajo izjave prič, posnetki telefonskih pogovorov, analize DNK iz vzorcev krvi in podobno. Pri statističnem sklepanju nam dokaze priskrbijo podatki. Po domače statistično sklepanje velikokrat temelji na grafični predstavitvi podatkov. Formalno sklepanje, tema tega predmeta, uporablja verjetnost, da pove, do kakšne mere smo lahko prepričani, da so naši zaključki pravilni.

Nekaj statističnih izzivov za začetnike

Trgovec je vašemu podjetju prodal 10.000 sodov rjavega fižola. Cena le-tega je na trgu za 10% višja od sivega (bolj obstojen in večja hranljiva vrednost). Še predno plačamo, odidemo do skladišča in odpremo naključno izban sod, ugotovimo, da je res napolnjen do vrha s fižolom, vendar pa so zrna rjava ali siva. Kako najhitreje ugotovimo, za koliko moramo znižati plačilo, če se odločimo, da bomo fižol vseeno prevzeli?

Dal bi vam toliko "odpustkov", kolikor las imam na glavi. Koliko las pa imamo na glavi?

Napisali smo diplomo, ki je dolga 100 strani, kolega pa ima za 20 strani daljšo diplomo. Če za trenutek pustimo ob strani samo vsebino (kvaliteto), je še vedno vprašanje ali je bil res boljši od nas v kvantiteti. Uporabljal je drugačen font, njegov rob je nekoliko večji,... Kako lahko na hitro ocenimo dejansko stanje (brez da bi primerjali sami datoteki)?

Nadaljujmo z branjem časopisov

Napoved vremena predstavlja naslednje področje statistike za množice, s svojimi napovedmi za dnevne najvišje in najnižje temperature (kako se lahko odločijo za 10 stopinj ne pa za 9 stopinj?). (Kako pridejo do teh številk? Z jemanjem vzorcev? Koliko vzorcev morajo zbrati in kje jih zbirajo? Najdete tudi napovedi za 3 dni naprej, morda celo teden, mesec in leto! Kako natančne so vremenske napovedi v današnjem času? Glede na to kolikokrat nas je ujel dež, ali kolikokrat so napovedali sonce, lahko zaključite, da morajo nadaljevati z raziskovanjem na tem področju. Verjetnost in računalniško modeliranja igrata pomembno vlogo pri napovedovanju vremena. Posebej uspešni so pri večjih dogodkih kot so orkani, potresi in vulkanski izbruhi. Seveda pa so računalniki le tako pametni kot ljudje, ki so napisali programsko opremo, ki jih poganja. Raziskovalci bodo imeli še veliko dela, predno bodo uspeli napovedati tornade še pred njihovim začetkom.

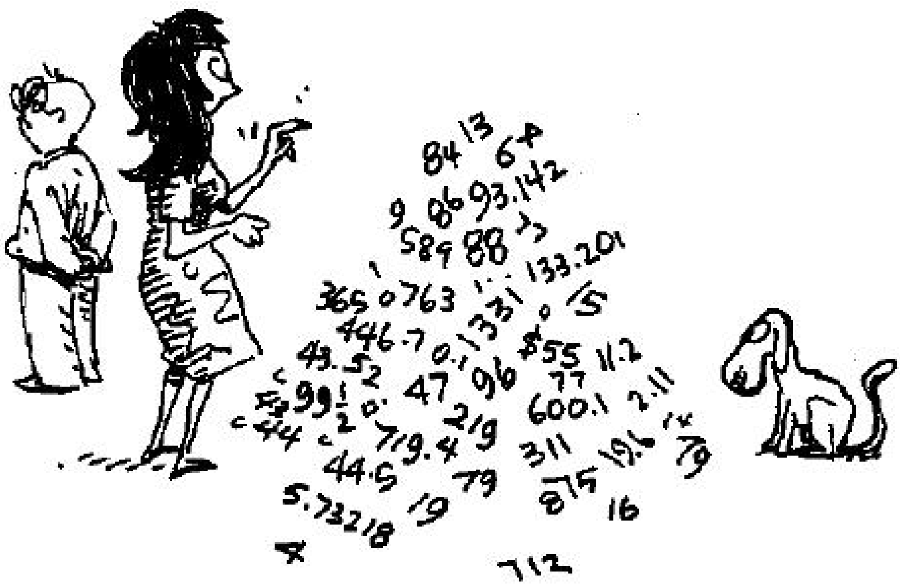
Poglejmo tisti del časopisa, ki se je posvečen filmom, ki jih trenutno vrtijo v kinematografih. Vsaka reklama vsebuje citate izbranih kritikov, npr. “Nepozabno!”, “Vrhunska predstava našega časa”, “Zares osupljivo”, ali “En izmed 10 najboljših filmov tega leta!” Ali Vam kritike kaj pomenijo? Kako se odločite katere filme si želite ogledati? Strokovnjaki so mnenja, da čeprav lahko vplivamo na popularnost filma s kritikami (dober ali slab) na samem začetku, pa je v celoti najbolj pomembno za film ustno izročilo. Študije so pokazale tudi, da bolj ko je dramatičen film, več kokic je prodanih. Res je, zabavna industrija beleži celo koliko hrustanja opravite med gledanjem. Kako v resnici zberejo vse te informacije in kako to vpliva na zvrsti filmov, ki jih delajo? Tudi to je del statistike: načrtovanje in izdelava študij, ki pomagajo določiti gledalce in ugotoviti kaj imajo radi, ter uporabiti informacijo za pomoč pri vodenju izdelave produkta/izdelka. Če Vas naslednjič nekdo ustavi z anketo in želi nekaj Vašega časa, si ga boste morda res vzeli v upanju, da bo upoštevana tudi Vaša volja.

Loterija in stave. Ko opazujemo zlorabo številk v vsakdanjem življenju, ne moremo mimo športnih stavnic, več milijardno industrijo (letno) ki prevzame tako občasnega stavca, kakor tudi profesionalnega igralca in impulzivnega zasvojenca z igrami na srečo. Na kaj lahko stavimo? Pravzaprav na takorekoč vse kar se konča na vsaj dva različna načina.

Številkam se ni mogoče izogniti niti s skokom v sekcijo potovanja. Tam najdemo tudi najbolj pogosto vprašanje naslovljeno na Urad za odzivni center transporta in varnosti, ki prejme tedensko povprečno 2.000 telefonskih klicev, 2.500 e-sporočil in 200 pisem (Bi želeli biti en izmed tistih, ki mora vse to prešteti?): “Ali lahko nesem to-in-to na letalo?”, pri čemer se “to-in-to” nanaša na takorekoč karkoli od živali do velikanske konzerve kokic (slednjega ne priporočam, saj je konzervo potrebno shraniti v vodoravni legi, med letom pa se stvari običajno premaknejo, pokrov se odpre in po pristanku vse skupaj pade na Vaše sopotnike - to se je enkrat celo v resnici zgodilo). To nas pripelje do zanimivega statističnega vprašanja: koliko telefonistov je potrebno v razičnih časovnih obdobjih tokom dneva, da obdelajo vse klice? Ocena števila klicev je samo prvi korak, in če nismo zadeli prave vrednosti, nas bo to bodisi drago stalo (v primeru, če je bila ocena prevelika) ali pa bomo prišli na slab glas (če je bila ocena prenizka).

Naslednja stvar, ki zbudi našo pozornost, je poročilo o povečanem številu mrtvih na naših cestah. Strokovnjaki nas opozarjajo, da se je število povečalo za več kot 50% od leta 1997 in nihče ne zna ugotoviti zakaj. Statistika nam pove zanimivo zgodbo. V letu 1997 je umrlo 2116 motoristov, v letu 2001 pa je statistični urad (National Highway Traffic Safety Administration - NHTSA) poročal o 3.181 žrtvah. V članku je obdelanih več možnih razlogov za povečanje števila žrtev, vključno z dejstvom, da so danes motoristi starejši (povprečna starost ponesrečenih motoristov se je povzpela z 29 let v letu 1990 na 36 let v letu 2001). Velikost dvokolesnikov je opisana kot druga možnost. Prostornina se je v povprečju povečala za skoraj 25% (iz 769 kubičnih centimeterov v letu 1990 na 959 kubičnih centimeters v letu 2001. Naslednja možnost je, da nekatere države ne izvajajo več tako strog nadzor nad zakonom o čeladah. V članku citirajo strokovnjake, da je potrebna veliko natančnejša študija, vendar pa najverjetneje ne bo opravljena, saj bi stala med 2 in 3 milijoni. En aspekt, ki v članku ni omenjen, je število motoristov v letu 2001 v primerjavi s številom v letu 1997. Večje število ljudi na cesti v glavnem pomeni tudi več žrtev, če vsi ostali faktorji ostanejo nespremenjeni. Kljub temu pa je v članku prikazan tudi graf, ki predstavi število smrtnih žrtev na 100 milijonov prepotovanih km od leta 1997 do 2001; ali ta podatek odgovori na vprašanje glede števila ljudi na cesti? Predstavljen je tudi stolpčni graf (diagram), ki primerja število smrtnih žrtev motoristov s številom nezgod s smrtnim izidom, ki so se pripetile z drugimi vozili. Le-ta prikaže 21 ponesrečenih motoristov na 100 milijonov prepotovanih km v primerjavi s samo 1·7 nezgodami s smrtnim izidom pri enakem številu prepotovanih km z avtom. Ta članek vsebuje veliko številk in statistike, toda kaj vse to sploh pomeni?

Ljudje običajno besedo *statistika* povezujejo z zbiranjem in urejanjem podatkov o nekem pojavu, izračunom raznih značilnosti iz teh podatkov, njih predstavitvijo in razlago. To je najstarejši del statistike in ima svoje začetke že v antiki – z nastankom večjih združb (držav) se je pojavila potreba po poznavanju stanja – “računovodstvo”, astronomija, ... Sama beseda *statistika* naj bi izvirala iz latinske besede *status* – v pomenu država. Kot smo že v uvodu omenili, pravimo Tej veji statistike *opisna statistika*. Druga veja, *inferenčna statistika*, poskuša spoznanja iz zbranih podatkov posplošiti (razširiti, podaljšati, napovedati, ...) in oceniti kakovost teh pospošitev. Statistiko lahko razdelimo tudi na *uporabno* in *teoretično* (računalniško in matematično) statistiko.



Statistika je veda, ki proučuje množične pojave.

(Statistična) enota – posamezna proučevana stvar ali pojav (npr. redni študent na Univerzi v Ljubljani v tekočem študijskem letu).

Populacija – množica vseh proučevanih enot; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko). Npr. vsi redni študentje na UL v tekočem študijskem letu.

Vzorec – podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih celotne populacije (npr. vzorec 300 slučajno izbranih rednih študentov).

Spremenljivka – lastnost enot; označujemo jih npr. z X, Y, X_1. Vrednost spremenljivke X na i–ti enoti označimo z x_i (npr. spol, uspeh iz matematike v zadnjem razredu srednje šole, izobrazba matere in višina mesečnih dohodkov staršev študenta itd.).

Posamezne spremenljivke in odnose med njimi opisujejo ustrezne porazdelitve.
Parameter je značilnost populacije, običajno jih označujemo z malimi grškimi črkami.
Statistika je značilnost vzorca; običajno jih označujemo z malimi latinskimi črkami.
Vrednost statistike je lahko za različne vzorce različna.

Eno izmed osnovnih vprašanj statistike je,
kako z uporabo ustreznih statistik oceniti
vrednosti izbranih parametrov.

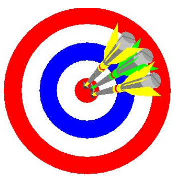
Poglavje 10

Opisna statistika

10.1 Vrste spremenljivk oziroma podatkov

Glede na vrsto vrednosti jih delimo na:

1. *številske* (ali numerične) spremenljivke – vrednosti lahko izrazimo s števili (npr. starost). Za podatke rečemo, da so *kvantitativni* kadar predstavljajo kvantiteto ali količino nečesa.



2. *opisne* (ali atributivne) spremenljivke – vrednosti lahko opišemo z imeni razredov (npr. poklic, uspeh, spol); Za podatke rečemo, da so *kvalitativni* (kategorični) kadar jih delimo v kategorije in zanje ni kvantitativnih interpretacij.

   ♂ ♀

   Če so kategorije brez odgovarjajočega vrstnega reda/urejenosti, rečemo, da so podatki *nominalni*. Kakor hitro imajo kategorije neko urejenost pa rečemo, da so podatki *ordinalni/številski*.

Glede na vrsto merske lestvice:

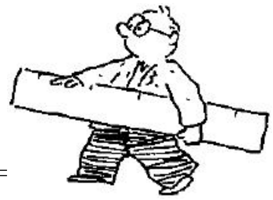
1. *imenske* (ali nominalne) spremenljivke – vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol);

2. *urejenostne* (ali ordinalne) spremenljivke – vrednosti lahko uredimo od najmanjše do največje (npr. uspeh);

3. *razmične* (ali intervalne) spremenljivke – lahko primerjamo razlike med vrednostima dvojic enot (npr. temperatura);

4. *razmernostne* spremenljivke – lahko primerjamo razmerja med vrednostima dvojic enot (npr. starost).

5. *absolutne* spremenljivke – štetja (npr. število prebivalcev).



dovoljene transformacije	*vrsta lestvice*	*primeri*
$f(x) = x$ (identiteta)	absolutna	štetje
$f(x) = a.x, a > 0$ podobnost	razmernostna	masa temperatura (K)
$f(x) = a \cdot x + b, \ a > 0$	razmična	temperatura (C,F) čas (koledar)
$x \geqslant y \Leftrightarrow f(x) \geqslant f(y)$ strogo naraščajoča	urejenostna	šolske ocene, kakovost zraka, trdost kamnin
f je povratno enolična	imenska	barva las, narodnost

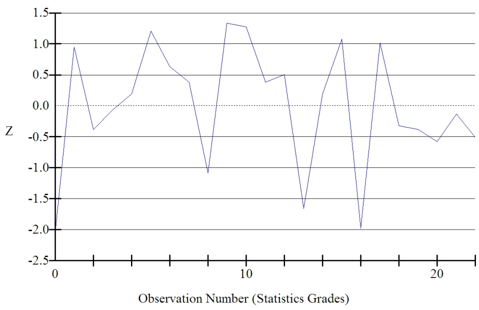
Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi. Urejenostne spremenljivke zadoščajo lastnostim, ki jih imajo imenske spremenljivke; in podobno razmernostne spremenljivke zadoščajo lastnostim, ki jih imajo razmične, urejenostne in imenske spremenljivke.

absolutna $\subset$ razmernostna $\subset$ razmična $\subset$ urejenostna $\subset$ imenska

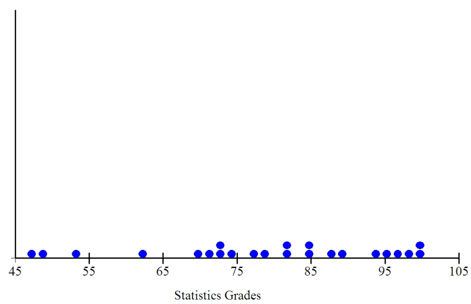
Posamezne statistične metode predpostavljajo določeno vrsto spremenljivk. Največ učinkovitih statističnih metod je razvitih za številske spremenljivke.

V teoriji merjenja pravimo, da je nek stavek *smiseln*, če ohranja resničnost/lažnost pri zamenjavi meritev z enakovrednimi (glede na dovoljene transformacije) meritvami.

10.2 Grafična predstavitev kvantitativnih podatkov



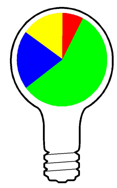
(a) grafični prikaz podatkov (angl. runs chart/plot),



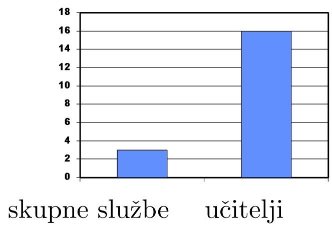
(b) diagram frnikul (angl. dot plot)

Primer: Oddelek sistemskih inženirjev

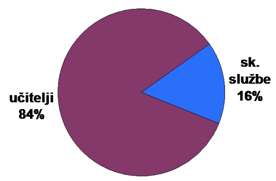
kategorija	frekvenca	relativna frekvenca
vrsta zaposlenih	število zaposlenih	delež
učitelji	16	0·8421
skupne službe	3	0·1579
skupaj	19	1·0000



Stolpčni prikaz (tudi stolpčni graf, poligonski diagram): Na eni osi prikažemo (urejene) razrede. Nad vsakim naredimo stolpec/črto višine sorazmerne frekvenci razreda.



Krožni prikaz (tudi strukturni krog, pogača, kolač): Vsakemu razredu priredimo krožni izsek s kotom $\alpha_i = \dfrac{k_i}{n} 360$ stopinj.



◇

Frekvenčna porazdelitev

Število vseh možnih vrednosti proučevane spremenljivke je lahko preveliko za pregledno prikazovanje podatkov. Zato sorodne vrednosti razvrstimo v skupine. Posamezni skupini priredimo ustrezno reprezentativno vrednost, ki je nova vrednost spremenljivke. Skupine vrednosti morajo biti določene *enolično*: vsaka enota s svojo vrednostjo je lahko uvrščena v natanko eno skupino vrednosti. *Frekvenčna porazdelitev* spremenljivke je *tabela*, ki jo določajo *vrednosti ali skupine vrednosti* in njihove *frekvence*. Če je spremenljivka vsaj urejenostna, vrednosti (ali skupine vrednosti) uredimo od najmanjše do največje. Skupine vrednosti številskih spremenljivk imenujemo *razredi*. Če zapišem podatke v vrsto po njihovi numerični velikosti pravimo, da gre za **urejeno zaporedje** oziroma *ranžirano vrsto*, ustreznemu mestu pa pravimo *rang*.

x_{min} in x_{max} – *najmanjša* in *največja* vrednost spremenljivke X.

$x_{i,min}$ in $x_{i,max}$ – *spodnja* in *zgornja meja* i-tega razreda.

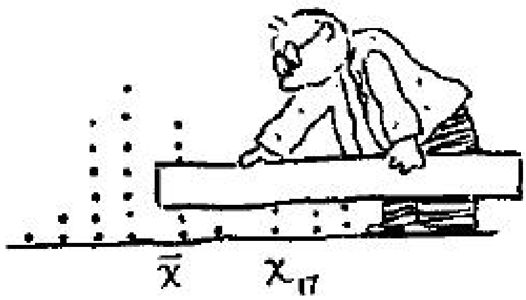
Meje razredov so določene tako, da velja $x_{i,max} = x_{i+1,min}$.

Širina i-tega razreda je $d_i = x_{i,max} - x_{i,min}$. Če je le mogoče, vrednosti razvrstimo v razrede enake širine.

Sredina i-tega razreda je $x_i = \frac{x_{i,min}+x_{i,max}}{2}$ in je značilna vrednost – predstavnik razreda.

Kumulativa (ali nakopičena frekvenca) je frekvenca do spodnje meje določenega razreda. Velja $F_{i+1} = F_i + f_i$, kjer je F_i kumulativa in f_i frekvenca v i-tem razredu. *Relativna frekvenca* rel.$f_i = f_i/N$ in *relativna kumulativa* rel.$F_i = F_i/N$, kjer je N število vseh podatkov.

Poligon: v koordinatnem sistemu zaznamujemo točke (x_i, f_i), kjer je x_i sredina i–tega razreda in f_i njegova frekvenca (tj. koliko podatkov je padlo v i-ti razred). K tem točkam dodamo še točki $(x_0, 0)$ in $(x_{k+1}, 0)$, če je v frekvenčni porazdelitvi k razredov. Točke zvežemo z daljicami.



Primer: Oglejmo si zaporedje podatkov

(a) Konstruiraj urejeno zaporedje.

(b) Nariši steblo-list diagram.

(c) Naredi histogram.

88	103	113	122	132
92	108	114	124	133
95	109	116	124	133
97	109	116	124	135
97	111	117	128	136
97	111	118	128	138
98	112	119	128	138
98	112	120	131	142
100	112	120	131	146
100	113	122	131	150

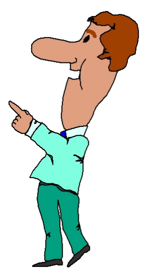
◇

Koraki za konstrukcijo steblo-list predstavitve

1. Razdeli vsako opazovanje-podatke na dva dela: **stebla** (angl. stem) in **listi** (angl. leaf).
2. Naštej stebla po vrsti v stolpec, tako da začneš pri najmanjšem in končaš pri največjem.
3. Upoštevaj vse podatke in postavi liste za vsak dogodek/meritev v ustrezno vrstico/steblo.
4. Preštej frekvence za vsako steblo.

Steblo-list diagram

stebla	listi	f_i	$rel.f_i$
08	8	1	2%
09	2 5 7 7 7 8 8	7	14%
10	0 0 3 8 9 9	6	12%
11	1 1 2 2 2 3 3 4 6 6 7 8 9	13	26%
12	0 0 2 2 4 4 4 8 8 8	10	20%
13	1 1 1 2 3 3 5 6 8 8	10	20%
14	2 6	2	4%
15	0	1	2%
		50	100%



Histogrami

Histogram: drug poleg drugega rišemo stolpce – pravokotnike, katerih ploščina je sorazmerna frekvenci v razredu. Če so razredi enako široki, je višina sorazmerna tudi frekvenci.

Ogiva: grafična predstavitev kumulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke $(x_{i,min}, F_i)$.

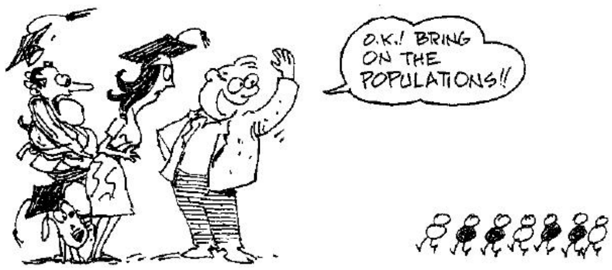
Kako zgradimo histogram

(a) Izračunaj **razpon** podatkov.

(b) Razdeli razpon na *5 do 20 razredov* enake širine.

(c) Za vsak razred preštej število vzorcev, ki spadajo v ta razred. To število imenujemo **frekvenca razreda**.

(d) Izračunaj vse **relativne frekvence razredov**.



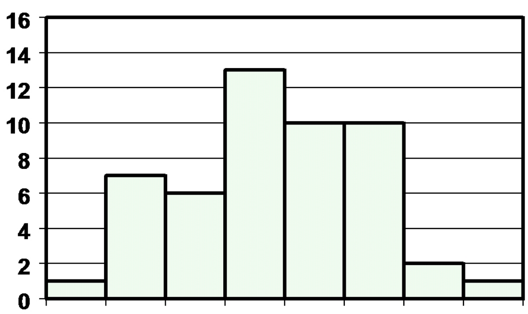
Pravilo za določanje števila razredov v histogramu

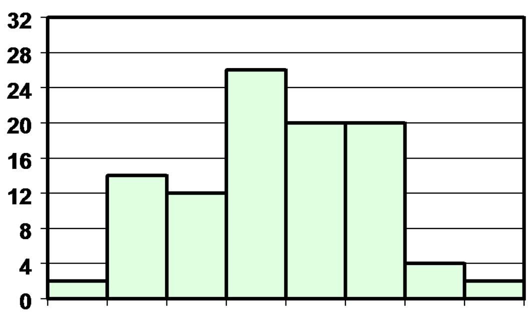
število vzorcev v množici podatkov	število razredov
manj kot 25	5 ali 6
25 – 50	7 – 14
več kot 50	15 – 20

Frekvenčna porazdelitev

razred	interval razreda	frekvenca	relativna frekvenca
1	80 – 90	1	2%
2	90 – 100	7	14%
3	100 – 110	6	12%
4	110 – 120	13	26%
5	120 – 130	10	20%
6	130 – 140	10	20%
7	140 – 150	2	4%
8	150 – 160	1	2%

Frekvenčni in procentni histogram

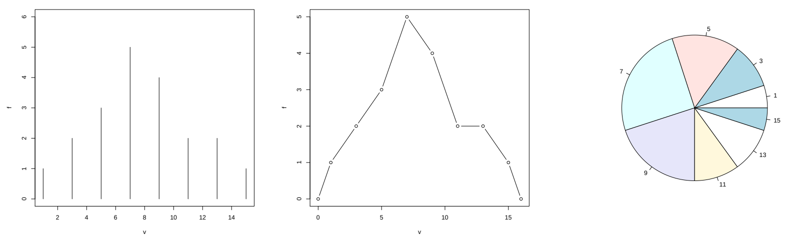




Nekaj ukazov v R-ju

```
> X <- c(5,11,3,7,5,7,15,1,13,11,9,9,3,13,9,7,7,5,9,7)
> n <- length(X)
> t <- tabulate(X)
> t
 [1] 1 0 2 0 3 0 5 0 4 0 2 0 2 0 1
> v <- (1:max(X))[t>0]
> f <- t[t>0]
> rbind(v,f)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
v    1    3    5    7    9   11   13   15
f    1    2    3    5    4    2    2    1
> plot(v,f,type="h")
> plot(c(0,v,16),c(0,f,0),type="b",xlab="v",ylab="f")
> pie(f,v)
> plot(c(0,v,16),c(0,cumsum(f)/n,1),col="red",type=š",
  xlab="v",ylab="f")
> x <- sort(rnorm(100,mean=175,sd=30))
> y <- (1:100)/100
> plot(x,y,main="Normalna porazdelitev, n=100",type=š")
> curve(pnorm(x,mean=175,sd=30),add=T,col="red")
```

stolpci poligon strukturni krog

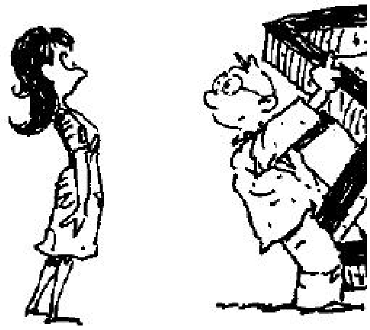


```
> x <- rnorm(1000,mean=175,sd=30)
> mean(x)
[1] 175.2683
> sd(x)
[1] 30.78941
> var(x)
[1] 947.9878
> median(x)
[1] 174.4802
> min(x)
[1] 92.09012
> max(x)
[1] 261.3666
> quantile(x,seq(0,1,0.1))
       0%       10%       20%       30%
 92.09012 135.83928 148.33908 158.53864
      40%       50%       60%       70%
166.96955 174.48018 182.08577 191.29261
      80%       90%      100%
200.86309 216.94009 261.36656
```

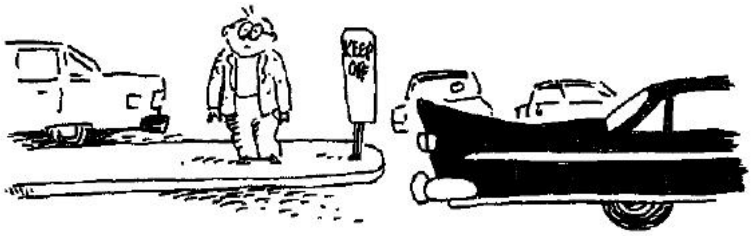
```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  92.09  154.20  174.50  175.30  195.50  261.40
> hist(x,freq=F)
> curve(dnorm(x,mean=175,sd=30),add=T,col="red")
```

10.3 Mere za lokacijo in razpršenost

- srednje vrednosti
- razpon (min/max)
- centili, kvartili
- varianca
- standardni odklon
- Z-vrednosti



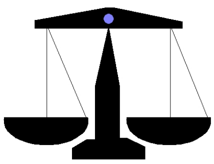
Modus (oznaka M_0) množice podatkov je tista vrednost, ki se pojavi z največjo frekvenco.

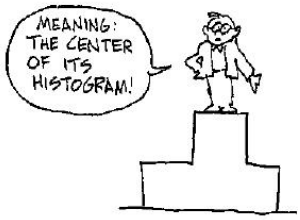


Da bi prišli do *mediane* (oznaka M_e) za neko množico podatkov, naredimo naslednje:

1. Podatke uredimo po velikosti v naraščujočem vrstnem redu,
2. Če je število podatkov liho, potem je mediana podatek na sredini,
3. Če je število podatkov sodo, je mediana enaka povprečju dveh podatkov na sredini.

Oznake: mediana populacije: μ mediana vzorca: m





Povrečje populacije: $$\mu = \frac{x_1 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

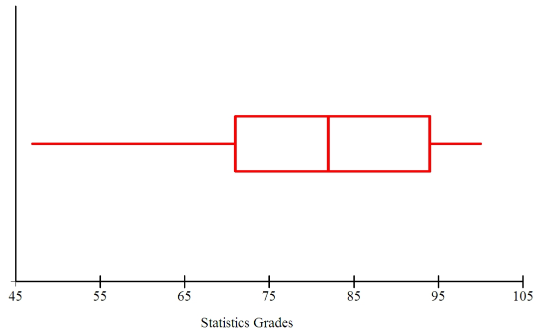
Povrečje vzorca: $$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Razpon je razlika med največjo in najmanjšo meritvijo v množici podatkov.

100p-ti centil ($p \in [0,1]$) je definiran kot število, od katerega ima 100p % meritev manjšo ali enako numerično vrednost. 100p-ti centil določimo tako, da izračunamo vrednost $p(n+1)$ in jo zaokrožimo na najbližje celo število. Naj bo to število enako i. Izmerjena vrednost z i-tim rangom je 100p-ti centil.

25. centil se imenuje tudi **1. kvartil**.
50. centil se imenuje **2. kvartil** ali **mediana**.
75. centil se imenuje tudi **3. kvartil**.

Škatla z brki (angl. box plot)



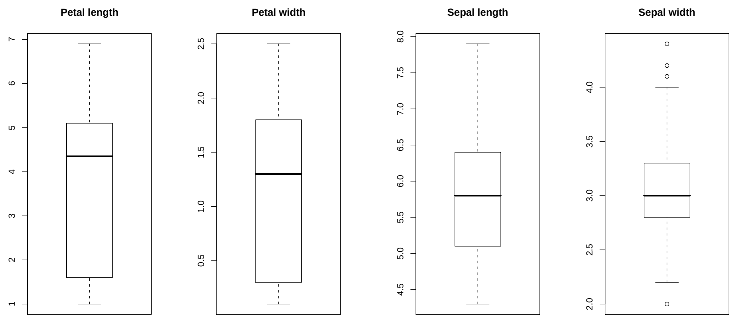
Še nekaj ukazov v R-ju: škatle in Q-Q-prikazi

Škatle (box-and-whiskers plot; grafikon kvantilov) `boxplot`:
škatla prikazuje notranja kvartila razdeljena z mediansko črto.
Daljici – brka vodita do robnih podatkov, ki sta največ za 1,5 dolžine
škatle oddaljena od nje. Ostali podatki so prikazani posamično.

Q-Q-prikaz `qqnorm` je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatka izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici. Premica `qqline` nariše premico skozi prvi in tretji kvartil.

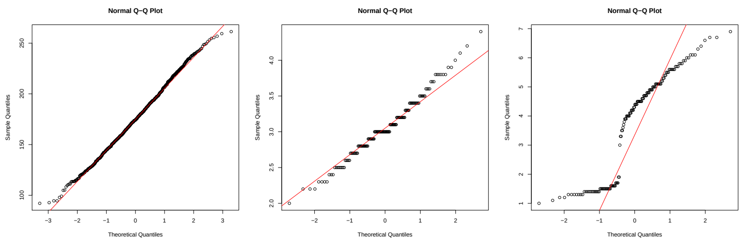
Obstaja tudi splošnejši ukaz `qqplot`, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametrom `datax=T` zamenjamo vlogo koordinatnih osi.

Škatle




```
> par(mfrow=c(1,2))
> boxplot(iris$Petal.Length,main="Petal length")
> boxplot(iris$Petal.Width,main="Petal width")
> boxplot(iris$Sepal.Length,main=Šepal length")
> boxplot(iris$Sepal.Width,main=Šepal width")
> par(mfrow=c(1,1))
```

Q-Q-prikaz




```
> qqnorm(x)
> qqline(x,col="red")
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width,col="red")
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length,col="red")
```

Mere razpršenosti

varianca

– kvadrat pričakovanega odklona (populacije)

– vsota kvadratov odklonov deljena s stopnjo prostosti (vzorca)

standardni odklon (deviacija)

– pozitivni kvadratni koren variance

koeficient variacije

– standardni odklon deljen s povprečjem

	populacija	vzorec
varianca	σ^2 $D,\ V$	$S^2,\ s^2$
standardni odklon	σ	$S,\ \ s$

Za vzorec smo vzeli osebje FRI in zabeležili naslednje število otrok: 1, 2, 2, 1, 2, 5, 1, 2.

Varianca in standardni odklon

Varianca populacije (končne populacije z N elementi): $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}.$$

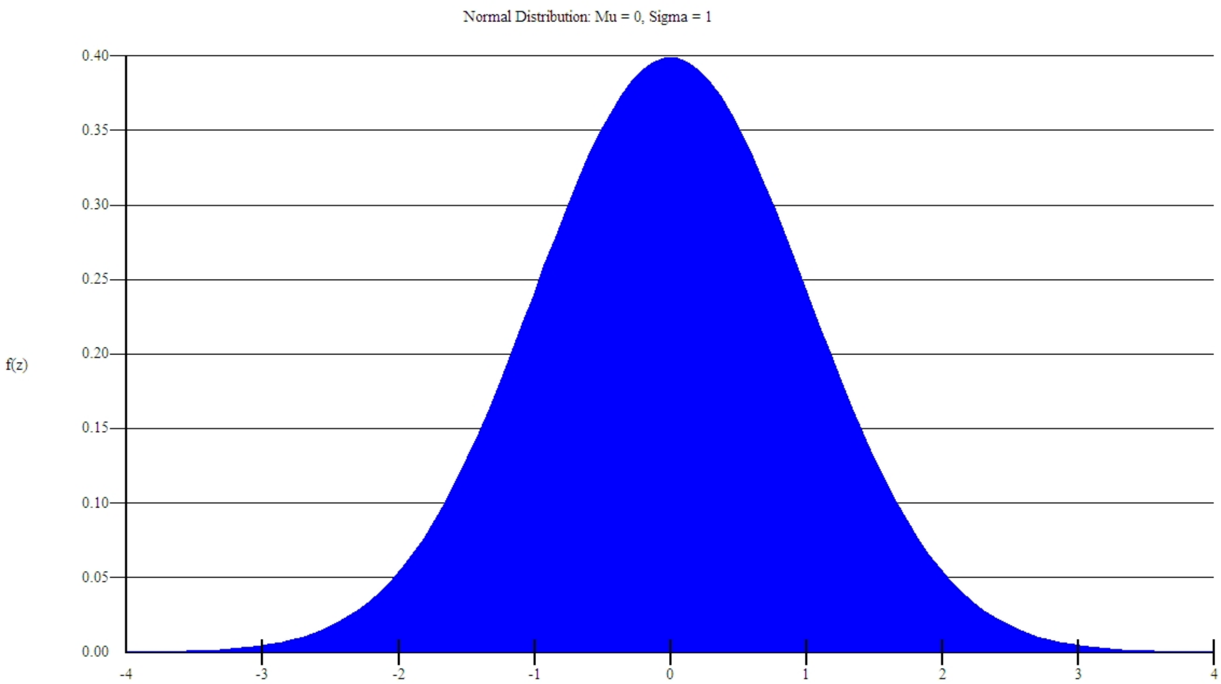
Varianca vzorca (n meritvami):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}$$

Standardni odklon je pozitivno predznačen kvadratni koren variance.

Normalna porazdelitev

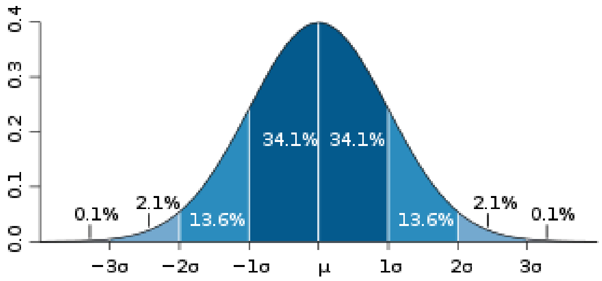
Veliko podatkovnih množic ima porazdelitev približno **zvonaste oblike** (unimodalna oblika - ima en sam vrh):



Empirična pravila

Če ima podatkovna množica porazdelitev približno zvonaste oblike, potem veljajo naslednja pravila (angl. rule of thumb), ki jih lahko uporabimo za opis podatkovne množice:

1. približno 68,3% vseh meritev leži na intervalu $[\mu - \sigma, \mu + \sigma]$,
2. približno 95,4% meritev leži na na intervalu $[\mu - 2\sigma, \mu + 2\sigma]$,
3. skoraj vse meritve (99,7%) ležijo na na intervalu $[\mu - 3\sigma, \mu + 3\sigma]$.



Mere oblike

Če je spremenljivka približno normalno porazdeljena, potem jo statistični karakteristiki povprečje in standardni odklon zelo dobro opisujeta. V primeru unimodalne porazdelitve spremenljivke, ki pa je bolj asimetrična in bolj ali manj sploščena (koničasta), pa je potrebno izračunati še stopnjo *asimetrije* in *sploščenosti (koničavosti)*.

Za $\ell \in \mathbb{N}$ je **ℓ-ti centralni moment** enak

$$m_\ell = \frac{(x_1 - \mu)^\ell + \cdots + (x_n - \mu)^\ell}{n}.$$

$m_1 = 0$, $m_2 = \sigma^2, \ldots$

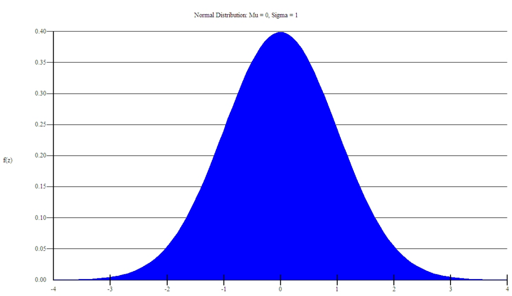
Koeficient asimetrije (s centralnimi momenti): $g_1 = m_3/m_2^{3/2}$. Mere asimetrije dobim tako, da opazujemo razlike med srednjimi vrednostimi. Le-te so tem večje čim bolj je porazdelitev asimetrična:

$$KA_{M_0} = (\mu - M_0)/\sigma, \qquad KA_{M_e} = 3(\mu - M_e)/\sigma.$$

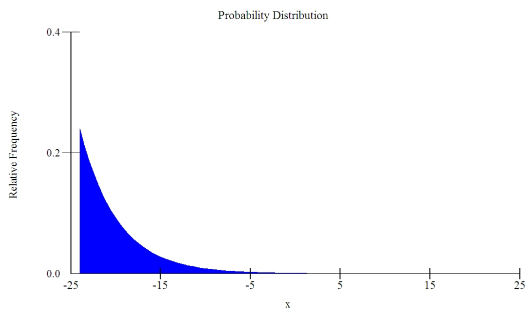
Koeficient sploščenosti (kurtosis) (s centralnimi momenti): $K = g_2 = m_4/m_2^2 - 3$

- $K = 3$ (ali 0) normalna porazdelitev zvonaste-oblike (*mesokurtic*),
- $K < 3$ (ali < 0) bolj kopasta kot normalna porazdelitev, s krajšimi repi (*platykurtic*),
- $K > 3$ (ali > 0) bolj špičasta kot normalna porazdelitev, z daljšimi repi (*leptokurtic*).

Normalna in asimetrična porazdelitev

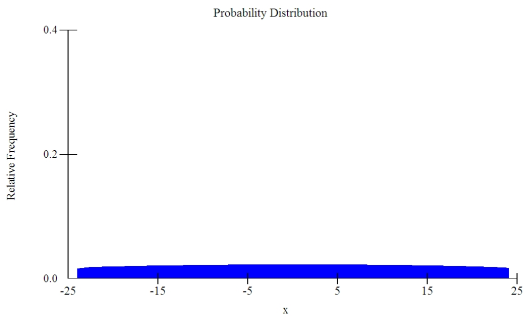


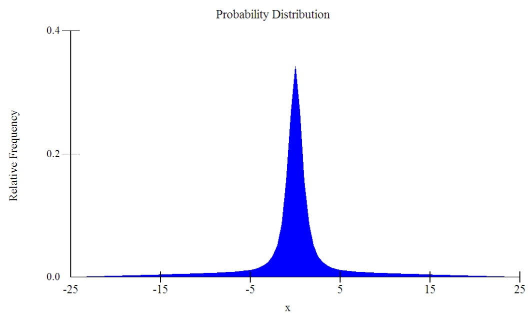
asim.= 0, sploščenost= 3 (mesokurtic).



asim.= 1·99, sploščenost= 8·85.

Kopasta in špičasta porazdelitev





asim.$= 0$, sploščenost$= 1{\cdot}86$ (platykurtic) asim.$= -1{\cdot}99$, sploščenost$= 8{\cdot}85$ (leptokurtic).

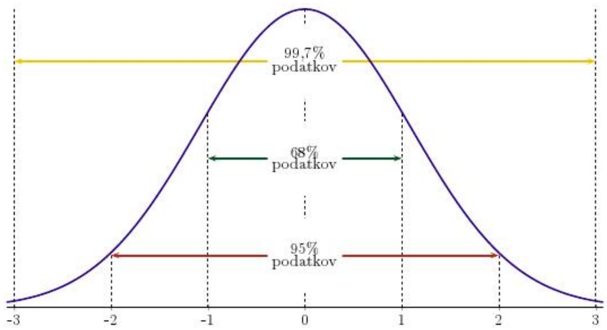
10.4 Standardizacija

Vsaki vrednosti x_i spremenljivke X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ:

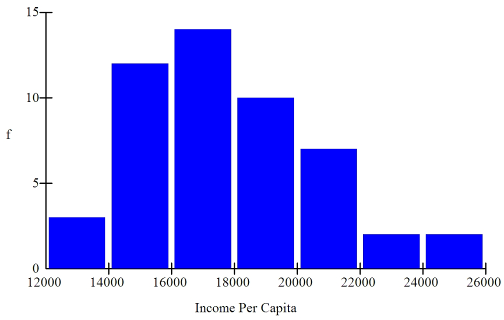
$$z_i = \frac{x_i - \mu}{\sigma}.$$

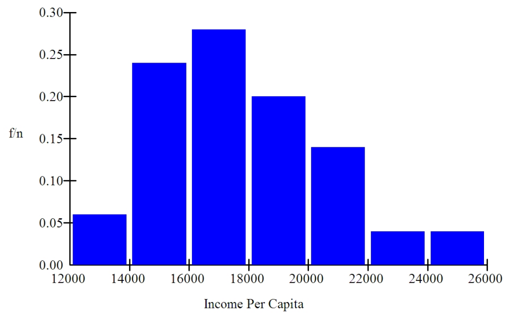
Za novo spremenljivko Z bomo rekli, da je **standardizirana**, z_i pa je **standardizirana vrednost**.

Potem je $\boldsymbol{\mu(Z) = 0}$ in $\boldsymbol{\sigma(Z) = 1}$.

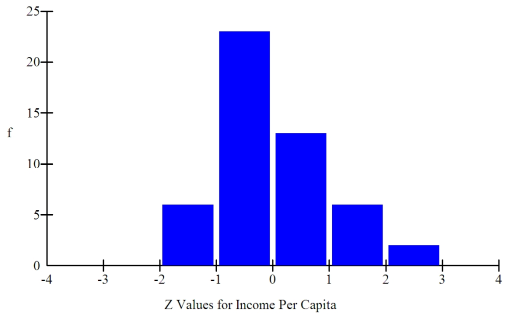


Frekvenčni in relativni frekvenčni histogram



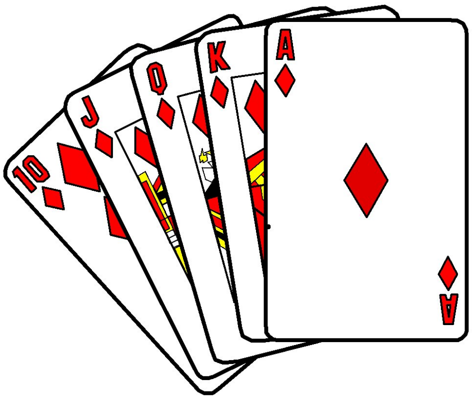


Histogram standardiziranih Z-vrednosti

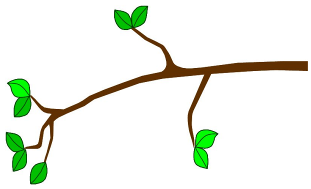


Poglavje 11

Vzorčenje



Analitična statistika je veja statistike, ki se ukvarja z uporabo vzorčnih podatkov, da bi z njimi naredili zaključek (inferenco) o populaciji.



Zakaj vzorčenje?

- cena
- čas
- destruktivno testiranje

Glavno vprašanje statistike je:

kakšen mora biti vzorec, da lahko iz podatkov zbranih na njem veljavno sklepamo o lastnostih celotne populacije.

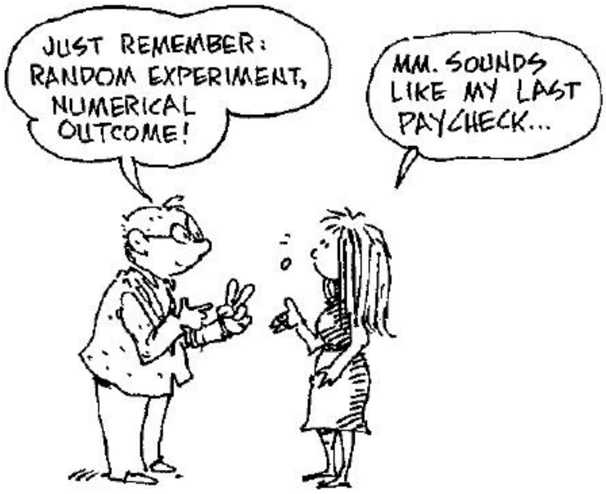
Kdaj vzorec dobro predstavlja celo populacijo? Preprost odgovor je:

- vzorec mora biti izbran *nepristransko*,
- vzorec mora biti *dovolj velik*.

Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor

$$(X_1, \dots, X_n),$$

ki mu rečemo *vzorec*. Število n je *velikost* vzorca.



Ker v vzorcu merimo isto spremenljivko in posamezna meritev ne sme vplivati na ostale, lahko predpostavimo:

1. vsi členi X_i vektorja imajo ***isto*** porazdelitev, kot spremenljivka X,
2. členi X_i so med seboj ***neodvisni***.

Takemu vzorcu rečemo *enostavni slučajni vzorec*. Večina statistične teorije temelji na predpostavki, da imamo opravka enostavnim slučajnim vzorcem. Če je populacija končna, lahko dobimo enostavni slučajni vzorec, tako da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo. Z vprašanjem, kako sestaviti dobre vzorce v praksi, se ukvarja posebno področje statistike – *teorija vzorčenja*. Načini vzorčenja

- ocena (priročnost).
- naključno
  - enostavno: pri enostavnem naključnem vzorčenju je vsak član populacije izbran/vključen z enako verjetnostjo.
  - deljeno: razdeljen naključni vzorec dobimo tako, da razdelimo populacijo na disjunktne množice oziroma dele (razrede) in nato izberemo enostavne naključne vzorce za vsak del posebej.
  - grozdno: takšno vzorčenje je enostavno naključno vzorčenje skupin ali klastrov/grozdov elementov.

11.1 Osnovni izrek statistike

Spremenljivka X ima na populaciji G porazdelitev $F(x) = P(X < x)$. Toda tudi vsakemu vzorcu ustreza neka porazdelitev. Za realizacijo vzorca $(x_1, \ldots, x_n)$ in $x \in \mathbb{R}$ postavimo

$$K(x) = \big|\{x_i : x_i < x, i = 1, \ldots, n\}\big| \quad \text{in} \quad V_n(x) = K(x)/n.$$

Slučajni spremenljivki $V_n(x)$ pravimo *vzorčna porazdelitvena funkcija*. Ker ima, tako kot tudi $K(x)$, $n+1$ možnih vrednosti k/n, $k = 0, \ldots, n$, je njena verjetnostna funkcija $B(n, F(x))$

$$P(V_n(x) = k/n) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

Če vzamemo n neodvisnih Bernoullijevih spremenljivk

$$Y_i(x) \sim \begin{pmatrix} 1 & 0 \\ F(x) & 1 - F(x) \end{pmatrix},$$

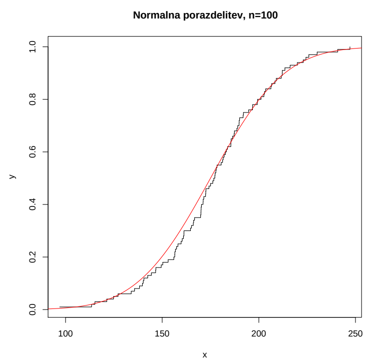
velja

$$V_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x).$$

Krepki zakon velikih števil tedaj zagotavlja, da za vsak x velja

$$P\Big(\lim_{n\to\infty} V_n(x) = F(x)\Big) = 1.$$

To je v bistvu Borelov zakon, da relativna frekvenca dogodka $(X < x)$ skoraj gotovo konvergira proti verjetnosti tega dogodka. Velja pa še več. $V_n(x)$ je stopničasta funkcija, ki se praviloma dobro prilega funkciji $F(x)$.



Odstopanje med $V_n(x)$ in $F(x)$ lahko izmerimo s slučajno spremenljivko

$$D_n = \sup_{x\in\mathbb{R}} \big|V_n(x) - F(x)\big|$$

za $n = 1, 2, \ldots$ Oznaka sup je okrajšava za supremum, ki predstavlja najmanjšo zgornjo mejo. V primeru končne množice razlik, gre za največjo razliko. Zanjo lahko pokažemo *osnovni izrek statistike*

$$P\Big(\lim_{n\to\infty} D_n = 0\Big) = 1.$$

Torej se z rastjo velikosti vzorca $V_n(x)$ enakomerno vse bolje prilega funkciji $F(x)$ – vse bolje povzema razmere na celotni populaciji.

11.2 Vzorčne ocene

Najpogostejša parametra, ki bi ju radi ocenili sta: *sredina populacije* μ glede na izbrano lastnost – matematično upanje spremenljivke X na populaciji; in *povprečni odklon* od sredine σ – standardni odklon spremenljivke X na populaciji. Statistike/ocene za te parametre so izračunane iz podatkov vzorca. Zato jim tudi rečemo *vzorčne ocene*.

Sredinske mere

Kot sredinske mere se pogosto uporabljajo:

Vzorčni modus – najpogostejša vrednost (smiselna tudi za imenske).
Vzorčna mediana – srednja vrednost, glede na urejenost, (smiselna tudi za urejenostne).
Vzorčno povprečje – povprečna vrednost (smiselna za vsaj razmične)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Vzorčna geometrijska sredina – (smiselna za vsaj razmernostne)

$$G(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Mere razpršenosti

Za oceno populacijskega odklona uporabljamo *mere razpršenosti*.

Vzorčni razmah $= \max_i x_i - \min_i x_i.$

Vzorčna disperzija $s_0^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Popravljena vzorčna disperzija $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

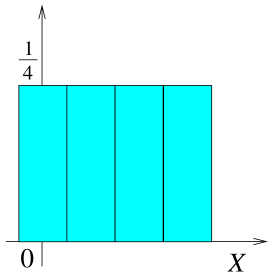
ter ustrezna *vzorčna odklona* s_0 in s.

11.3 Porazdelitve vzorčnih povprečij

Denimo, da je v populaciji N enot in da iz te populacije slučajno izbiramo n enot v enostavni slučajni vzorec ali na kratko slučajni vzorec (vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. $1/N$). Če hočemo dobiti slučajni vzorec, moramo izbrane enote pred ponovnim izbiranjem vrniti v populacijo (vzorec s ponavljanjem). Če je velikost vzorca v primerjavi s populacijo majhna, se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec, ki nastane s slučajnim izbiranjem brez vračanja.

Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce. Dobili smo populacijo vseh možnih vzorcev. Teh je v primeru enostavnih slučajnih vzorcev (s ponavljanjem) N^n. Število slučajnih vzorcev brez ponavljanja pa je $\binom{N}{n}$, če ne upoštevamo vrstnega reda izbranih enot v vzorcu, oziroma $\binom{N+n-1}{n}$, če upoštevamo vrstni red.

Primer: Vzemimo populacijo z $N = 4$ enotami, ki imajo naslednje vrednosti spremenljivke X: $0, 1, 2, 3$. in izračunamo populacijsko aritmetično sredino in varianco:



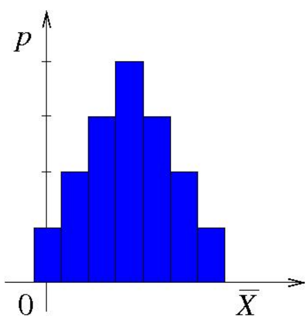
Slika 5. Histogram porazdelitve spremenljivke X.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{3}{2}, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{5}{4}.$$

Sedaj pa tvorimo vse možne vzorce velikosti $n = 2$ s ponavljanjem, in na vsakem izračunajmo vzorčno aritmetično sredino $\overline{X}$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{2}(X_1 + X_2).$$

vzorci	$\overline{X}$	vzorci	$\overline{X}$
0, 0	0	2, 0	1
0, 1	0·5	2, 1	1·5
0, 2	1	2, 2	2
0, 3	1·5	2, 3	2·5
1, 0	0·5	3, 0	1·5
1, 1	1	3, 1	2
1, 2	1·5	3, 2	2·5
1, 3	2	3, 3	3



Slika 6. Histogram porazdelitve spremenljivke $\overline{X}$.

Zapišimo verjetnostno shemo slučajne spremenljivke vzorčno povprečje $\overline{X}$ (glej sliko 6):

$$\overline{X} \; : \; \begin{pmatrix} 0 & 0{\cdot}5 & 1 & 1{\cdot}5 & 2 & 2{\cdot}5 & 3 \\ 1/16 & 2/16 & 3/16 & 4/16 & 3/16 & 2/16 & 1/16 \end{pmatrix}.$$

Izračunajmo matematično upanje ter disperzijo vzorčnega povprečja:

$$E(\overline{X}) = \sum_{i=1}^{m} \overline{X}_i\, p_i = \frac{0+1+3+6+6+5+3}{16} = \frac{3}{2},$$

$$D(\overline{X}) = \sum_{i=1}^{m} \Big(\overline{X}_i - E(\overline{X})\Big)^2 p_i = \frac{5}{8}.$$

◇

Na primeru smo spoznali, da je statistika 'vzorčna aritmetična sredina' slučajna spremenljivka s svojo porazdelitvijo. Sedaj pa izračunajmo matematično upanje in razpršenost za splošno **vzorčno povprečje**, določeno z zvezo

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

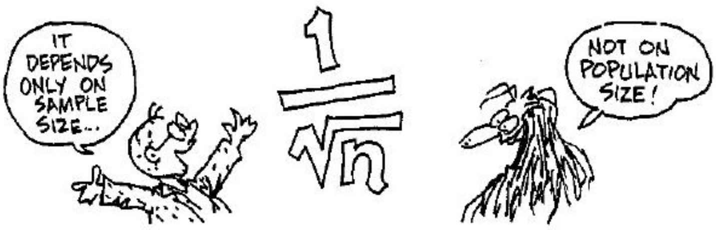
ki je tudi slučajna spremenljivka. Tedaj je

$$\mathsf{E}\overline{X} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i = \mu \qquad \text{in} \qquad \mathsf{D}\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{D}X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Zapišimo dobljeni ugotovitvi še v obliki trditve.

Trditev 11.1. *Naj bo $X_1, X_2, \ldots, X_n$ naključni vzorec, ki je sestavljen iz n meritev populacije s končnim povprečjem μ in končnim standardnim odklonom σ. Potem sta povprečje in standardni odklon vzorčnega povprečja $\overline{X}$ enaka*

$$\mu_{\overline{X}} = \mu \qquad \text{in} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$



Iz druge zveze vidimo, da standardna napaka $\sigma\overline{X} = \dfrac{\sigma}{\sqrt{n}}$ statistike $\overline{X}$ gre proti 0 z naraščanjem velikosti vzorca, tj. $\overline{X} \to \mu$; (enako nam zagotavlja tudi krepki zakon velikih števil).

Hitrost centralne tendence pri CLI

Spomnimo se, da nam **Centralni limitni izrek** v primeru povprečja pove:

Če je naključni vzorec velikosti n izbran iz populacije s končnim povprečjem μ in varianco σ^2, potem je lahko, če je n dovolj velik, vzorčna porazdelitev povprečja $\overline{X}$ aproksimirana z gostoto normalne porazdelitve.

Dokaz CLI je precej tehničen, kljub temu pa nam ne da občutka kako velik mora biti n, da se porazdelitev slučajne spremenljivke

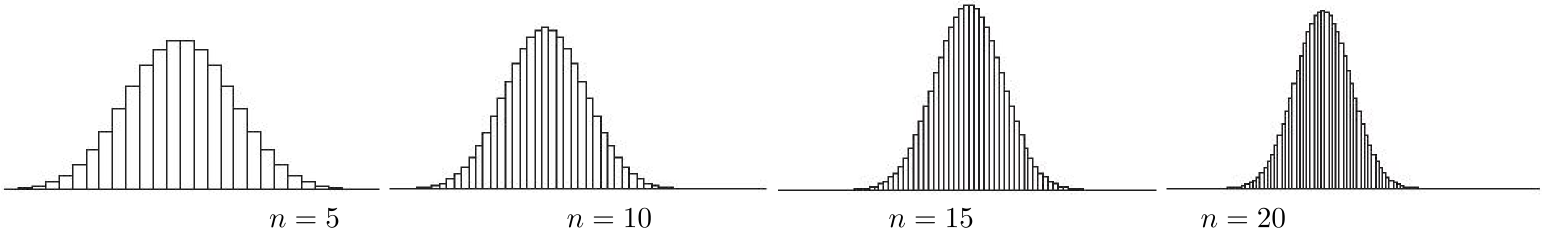
$$X_1 + \cdots + X_n$$

približa normalni porazdelitvi. Hitrost približevanja k normalni porazdelitvi je odvisna od tega kako simetrična je porazdelitev. To lahko potrdimo z eksperimentom: mečemo (ne)pošteno kocko, X_k naj bo vrednost, ki jo kocka pokaže pri k-tem metu.

Centralna tendenca za pošteno kocko

$$p_1 = 1/6, \quad p_2 = 1/6, \quad p_3 = 1/6, \quad p_4 = 1/6, \quad p_5 = 1/6, \quad p_6 = 1/6.$$

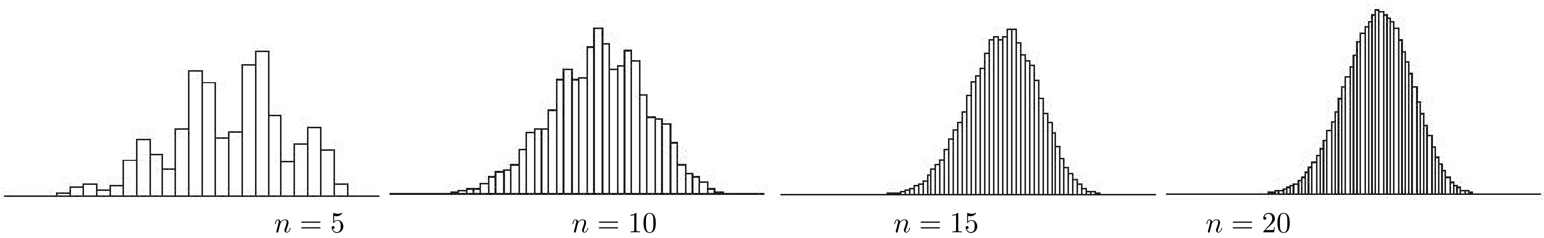
in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



Centralna tendenca za goljufivo kocko

$$p_1 = 0{\cdot}2, \quad p_2 = 0{\cdot}1, \quad p_3 = 0, \quad p_4 = 0, \quad p_5 = 0{\cdot}3, \quad p_6 = 0{\cdot}4.$$

in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



11.4 Vzorčna statistika

Vzorčna statistika je poljubna simetrična funkcija (tj. njena vrednost je neodvisna od permutacije argumentov) vzorca

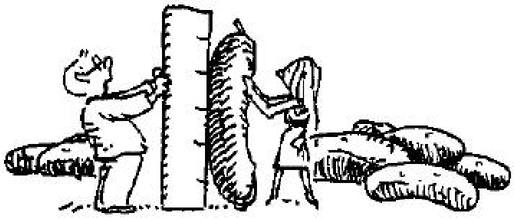
$$Y = g(X_1, X_2, \ldots, X_n).$$

Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca. Najzanimivejši sta značilni vrednosti

- njeno matematično upanje $\mathsf{E}Y$,
- standardni odklon σY, ki mu pravimo tudi *standardna napaka* statistike Y (angl. standard error – zato oznaka SE(Y)).

11.4.1 (A) Vzorčno povprečje

Proizvajalec embalaže za kumare bi rad ugotovil **povprečno dolžino** kumarice (da se odloči za velikost embalaže), ne da bi izmeril dolžino čisto vsake. Zato naključno izbere n kumar in izmeri njihove dolžine $X_1, \dots, X_n$. Sedaj nam je že blizu ideja, da je vsaka dolžina X_i **slučajna spremenljivka** (numerični rezultat naključnega eksperimenta).



Če je μ (iskano/neznano) povprečje dolžin, in je σ standardni odklon porazdelitve dolžin kumar, potem velja

$$\mathsf{E}X_i = \mu, \qquad \text{in} \qquad \mathsf{D}X_i = \sigma^2,$$

za vsak i, ker bi X_i bila lahko dolžina katerekoli kumare.

Denimo, da se spremenljivka X na populaciji porazdeljuje normalno $N(\mu, \sigma)$. Na vsakem vzorcu (s ponavljanjem) izračunamo vzorčno aritmetično sredino $\overline{X}$. Dokazati se da, da je **porazdelitev vzorčnih aritmetičnih sredin** normalna, kjer je

- matematično upanje vzorčnih aritmetičnih sredin enako aritmetični sredini spremenljivke na populaciji
$$E(\overline{X}) = \mu,$$
- standardni odklon vzorčnih aritmetičnih sredin
$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Za dovolj velike vzorce ($n > 30$) je porazdelitev vzorčnih aritmetičnih sredin približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se spremenljivka X porazdeljuje vsaj približno normalno s standardno napako $\mathrm{SE}(X)$, potem se

$$Z = \frac{X - E(X)}{\mathrm{SE}(X)}$$

porazdeljuje standardizirano normalno.

Vzorčno povprečje in normalna porazdelitev

Naj bo $X \sim N(\mu, \sigma)$. Tedaj je $\sum_{i=1}^{n} X_i \sim N(n\mu, \sigma\sqrt{n})$ in dalje $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$. Torej je vzorčna statistika

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1).$$

Kaj pa če porazdelitev X ni normalna? Izračun porazdelitve se lahko zelo zaplete. Toda pri večjih vzorcih ($n > 30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka Z porazdeljena skoraj standardizirano normalno. Vzorčno povprečje

$$\overline{X} = \frac{\sigma}{\sqrt{n}} Z + \mu$$

ima tedaj porazdelitev približno $N(0 + \mu, 1 \cdot \sigma/\sqrt{n}) = N(\mu, \sigma/\sqrt{n})$, kar je enako kot v primeru, ko smo začeli z normalno porazdeljenimi slučajnimi spremenljivkami.

Primer: Kolikšna je verjetnost, da bo pri 36 metih igralne kocke povprečno število pik večje ali enako 4? X je slučajna spremenljivka z vrednostmi $1, 2, 3, 4, 5, 6$ in verjetnostmi $1/6$. Zanjo je $\mu = 3{\cdot}5$ in standardni odklon $\sigma = 1{\cdot}7$. Vseh 36 ponovitev meta lahko obravnavamo kot slučajni vzorec velikost 36. Tedaj je

$$P(\overline{X} \geqslant 4) = P\Big(Z \geqslant (4 - \mu)\sqrt{n}/\sigma\Big) = P(Z \geqslant 1{\cdot}75) \approx 0{\cdot}04.$$

```
> x <- 1:6
> m <- mean(x)
> s <- sd(x)*sqrt(5/6)
> z <- (4-m)*6/s
> p <- 1-pnorm(z)
> cbind(m,s,z,p)
       m        s       z          p
[1,] 3.5 1.707825 1.75662 0.03949129
```

$\diamond$

11.4.2 (B) Vzorčna disperzija

Naj bo slučajna spremenljivka X na neki populaciji porazdeljena normalno, tj. $N(\mu, \sigma)$. Kako bi določili porazdelitev za vzorčno disperzijo ali popravljeno vzorčno disperzijo, tj.

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{oziroma} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2?$$

Raje izračunamo porazdelitev za naslednjo vzorčno statistiko

$$\chi^2 = \frac{nS_0^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Preoblikujemo jo lahko takole (pri čemer opozorimo, da vzorčno povprečje $\overline{X}$ in populacijsko povprečje μ nista enaka):

$$\begin{aligned}\chi^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{\sigma^2}(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) + \frac{n}{\sigma^2}(\mu - \overline{X})^2\end{aligned}$$

in, ker je $\sum_{i=1}^{n}(X_i - \mu) = n(\overline{X} - \mu) = -n(\mu - \overline{X})$, dalje

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2,$$

kjer so $Y_1, Y_2, \ldots, Y_n$ paroma neodvisne standardizirano normalno porazdeljene slučajne spremenljivke za katere velja $Y_i = (X_i - \mu)/\sigma$. Porazdelitvena funkcija za izbrano vzorčno statistiko χ^2 je

$$F_{\chi^2} = P(\chi^2 < z) = \int\!\!\int \cdots \int_{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 < z} e^{-(y_1^2 + y_2^2 + \ldots + y_n^2)/2}\, dy_n \ldots dy_1,$$

z ustrezno ortogonalno transformacijo v nove spremenljivke $z_1, z_2, \ldots, z_n$ dobimo po nekaj računanja (glej Hladnik)

$$F_{\chi^2} = \frac{1}{(2\pi)^{(n-1)/2}} \int\!\!\int \cdots \int_{\sum_{i=1}^{n-1} z_i^2 < z} e^{-(z_1^2 + z_2^2 + \ldots + z_{n-1}^2)/2} dz_{n-1} \ldots dz_1.$$

Pod integralom je gostota vektorja $(Z_1, Z_2, \ldots, Z_{n-1})$ z neodvisnimi standardizirano normalnimi členi. Integral sam pa ustreza porazdelitveni funkciji vsote kvadratov

$$Z_1^2 + Z_2^2 + \ldots + Z_{n-1}^2.$$

Tako je porazdeljena tudi statistika χ^2. Kakšna pa je ta porazdelitev? Ker so tudi kvadrati Z_1^2, Z_2^2, $\ldots$,Z_{n-1}^2 med seboj neodvisni in porazdeljeni po zakonu $\chi^2(1)$, je njihova vsota porazdeljena po zakonu $\chi^2(n-1)$. Tako je torej porazdeljena tudi statistika χ^2.

Ker vemo, da je $\mathsf{E}\chi^2(n) = n$ in $\mathsf{D}\chi^2(n) = 2n$, lahko takoj izračunamo

$$\mathsf{E}S_0^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n} = \frac{(n-1)\sigma^2}{n}, \qquad \mathsf{E}S^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n-1} = \sigma^2$$

in

$$\mathsf{D}S_0^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n} = \frac{2(n-1)\sigma^4}{n^2} \qquad \mathsf{D}S^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n-1} = \frac{2\sigma^4}{n-1}$$

Če je n zelo velik, je po centralnem limitnem izreku statistika χ^2 porazdeljena približno normalno in sicer po zakonu

$$N\Big(n-1, \sqrt{2(n-1)}\Big),$$

vzorčna disperzija S_0^2 približno po

$$N\left(\frac{(n-1)\sigma^2}{n}, \frac{\sqrt{2(n-1)}\,\sigma^2}{n}\right)$$

in popravljena vzorčna disperzija S^2 približno po

$$N\left(\sigma^2, \sqrt{\frac{2}{n-1}}\,\sigma^2\right).$$

11.5 Nove porazdelitve

Pri normalno porazdeljeni slučajni spremenljivki X je tudi porazdelitev $\overline{X}$ normalna, in sicer $N(\mu, \sigma/\sqrt{n})$. Statistika

$$Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$$

je potem porazdeljena standardizirano normalno.

Pri ocenjevanju parametra μ z vzorčnim povprečjem $\overline{X}$ to lahko uporabimo le, če poznamo σ; sicer ne moremo oceniti standardne napake – ne vemo, kako dobra je ocena za μ. Kaj lahko naredimo, če σ ne poznamo?

Parameter σ lahko ocenimo s S_0 ali S. *Toda* S je slučajna spremenljivka in porazdelitev statistike

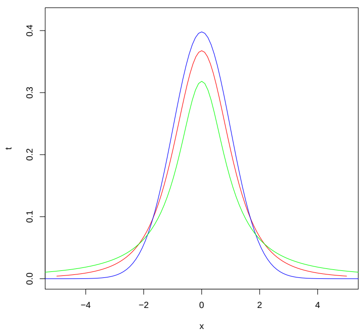
$$\frac{\overline{X}-\mu}{S}\sqrt{n}$$

ni več normalna $N(0,1)$ (razen, če je n zelo velik in S skoraj enak σ). Kakšna je porazdelitev nove vzorčne statistike

$$T = \frac{\overline{X}-\mu}{S}\sqrt{n} \quad ?$$

'Student' in 1908

11.5.1 Studentova porazdelitev



Leta 1908 je W.S. Gosset (1876-1937) pod psevdonimom 'Student' objavil članek, v katerem je pokazal, da ima statistika T porazdelitev $\boldsymbol{S(n-1)}$ z gostoto

$$p(t) = \frac{\left(1 + \dfrac{t^2}{n-1}\right)^{-n/2}}{\sqrt{n-1}\, B\left(\dfrac{n-1}{2}, \dfrac{1}{2}\right)}.$$

Tej porazdelitvi pravimo **Studentova porazdelitev** *z $n-1$ prostostnimi stopnjami*.

Pri tem smo uporabili Beta funkcijo, ki jo lahko vpeljemo z Gama funkcijo na naslednji način:

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = B(y,x).$$

Posebne vrednosti:

$$B\left(\frac{1}{2}, \frac{1}{2}\right) = \pi, \qquad B(m,n) = \frac{(m-1)!(n-1)!}{(m+n-1)!} \quad \text{za } m,n \in \mathbb{N}.$$

```
> plot(function(x) dt(x,df=3),-5,5,ylim=c(0,0.42),ylab="t",
  col="red")
> curve(dt(x,df=100),col="blue",add=T)
> curve(dt(x,df=1),col="green",add=T)
```

Za $S(1)$ dobimo Cauchyevo porazdelitev z gostoto

$$p(t) = \frac{1}{\pi(1+t^2)}$$

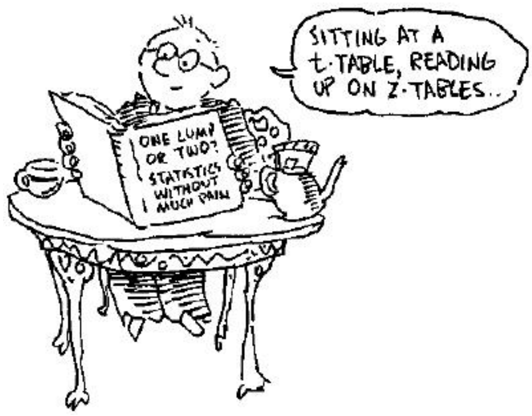
Za $n \to \infty$ pa gre

$$\frac{1}{\sqrt{n-1}\, B\left(\frac{n-1}{2}, \frac{1}{2}\right)} \to \sqrt{2\pi} \quad \text{in} \quad \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \to e^{-t^2/2}.$$

Torej ima limitna porazdelitev gostoto

$$p(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

standardizirane normalne porazdelitve.



Če zadnji sliki dodamo

```
> curve(dnorm(x),col="magenta",add=T)
```

ta pokrije modro krivuljo.

11.5.2 Fisherjeva porazdelitev

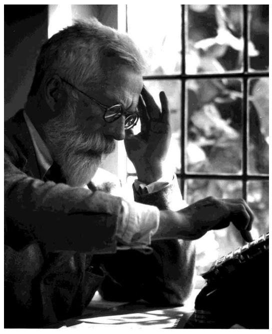
Poskusimo najti še porazdelitev kvocienta $Z = \dfrac{U}{V}$,

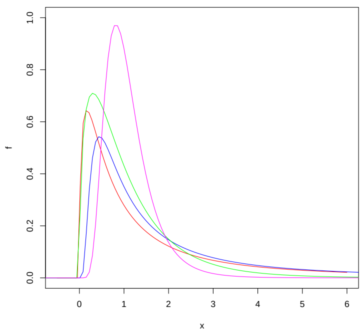
kjer sta $U \sim \chi^2(m)$ in $V \sim \chi^2(n)$ ter sta U in V neodvisni.

Z nekaj računanja (glej Hladnik) je mogoče pokazati, da je za $x > 0$ gostota ustrezne porazdelitve $F(m,n)$ enaka

$$p(x) = \frac{m^{m/2}\, n^{n/2}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{x^{(m-2)/2}}{(n+mx)^{(m+n)/2}}$$

in je enaka 0 drugje.





Porazdelitvi $F(m,n)$ pravimo **Fisherjeva** ali tudi **Snedecorjeva porazdelitev** ***F*** *z* (m,n) *prostostnimi stopnjami*.

```
> plot(function(x) df(x,df1=3,df2=2),-0.5,6,ylim=c(0,1),ylab="f",
  col="red")
> curve(df(x,df1=20,df2=2),col="blue",add=T)
> curve(df(x,df1=3,df2=20),col="green",add=T)
> curve(df(x,df1=20,df2=20),col="magenta",add=T)
```

Po zakonu $F(m-1, n-1)$ je na primer porazdeljena statistika

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

saj vemo, da sta spremenljivki

$$U = (m-1)S_X^2/\sigma_X^2 \quad \text{in} \quad V = (n-1)S_Y^2/\sigma_Y^2$$

porazdeljeni po χ^2 z $m-1$ oziroma $n-1$ prostostnimi stopnjami in sta neodvisni. Velja še:

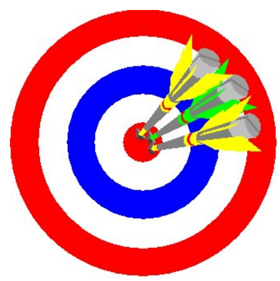
če je $U \sim F(m,n)$, je $1/U \sim F(n,m)$,

če je $U \sim S(n)$, je $U^2 \sim F(1,n)$.

Poglavje 12

Cenilke

12.1 Osnovni pojmi



Točkovna cenilka je pravilo ali formula, ki nam pove, kako izračunati numerično oceno parametra populacije na osnovi merjenj vzorca.
Število, ki je rezultat izračuna, se imenuje **točkovna ocena** (in mu ne moremo zaupati v smislu verjetnosti).

Cenilka parametra ζ je vzorčna statistika $C = C(X_1, \ldots, X_n)$, katere porazdelitveni zakon je odvisen le od parametra ζ, njene vrednosti pa ležijo v prostoru parametrov. Seveda je odvisna tudi od velikosti vzorca n .

Primer: Vzorčna mediana $\tilde{X}$ in vzorčno povprečje $\overline{X}$ sta cenilki za populacijsko povprečje μ; popravljena vzorčna disperzija S^2 pa je cenilka za populacijsko disperzijo σ^2. ◇

Cenilka C parametra ζ (grška črka zeta) je **dosledna**, če z rastočim n zaporedje C_n verjetnostno konvergira k parametru ζ, tj. za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P(|C_n - \zeta| < \varepsilon) = 1.$$

Primeri: vzorčno povprečje $\overline{X}$ je dosledna cenilka za populacijsko povprečje μ. Tudi vsi **vzorčni začetni momenti**

$$Z_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

so dosledne cenilke ustreznih začetnih populacijskih momentov $z_k = \mathsf{E}X^k$, če le-ti obstajajo.

Primer: Vzorčna mediana $\tilde{X}$ je dosledna cenilka za populacijsko mediano. $\diamond$

Trditev 12.1. *Če pri pogoju $n \to \infty$ velja $\mathsf{E}C_n \to \zeta$ in $\mathsf{D}C_n \to 0$, je C_n dosledna cenilka parametra ζ.*

Dokaz. To spremidimo takole:

$$1 - P(|C_n - \zeta| < \varepsilon) = P(|C_n - \zeta| \geqslant \varepsilon) \leqslant P(|C_n - \mathsf{E}C_n| + |\mathsf{E}C_n - \zeta| \geqslant \varepsilon),$$

upoštevajmo še, da za dovolj velike n velja $|\mathsf{E}C_n - \zeta| < \varepsilon/2$, in uporabimo neenakost Čebiševa

$$P(|C_n - \mathsf{E}C_n| \geqslant \varepsilon/2) \leqslant \frac{4\mathsf{D}C_n}{\varepsilon^2} \to 0. \qquad \square$$

Primer: Naj bo $X \sim N(\mu, \sigma)$. Ker za $n \to \infty$ velja

$$\mathsf{E}S_0^2 = \frac{(n-1)\sigma^2}{n} \to \sigma^2 \quad \text{in} \quad \mathsf{D}S_0^2 = \frac{2(n-1)\sigma^4}{n^2} \to 0,$$

je vzorčna disperzija S_0^2 dosledna cenilka za σ^2. $\diamond$

Nepristrana cenilka z najmanjšo varianco

Cenilka C_n parametra ζ je **nepristranska**, če je $\mathsf{E}C_n = \zeta$ (za vsak n); in je **asimptotično nepristranska**, če je $\lim_{n\to\infty} \mathsf{E}C_n = \zeta$. Količino $B(C_n) = \mathsf{E}C_n - \zeta$ imenujemo **pristranost** (angl. *bias*) cenilke C_n.

Primer: Vzorčno povprečje $\overline{X}$ je nepristranska cenilka za populacijsko povprečje μ; vzorčna disperzija S_0^2 je samo asimptotično nepristranska cenilka za σ^2, popravljena vzorčna disperzija S^2 pa je nepristranska cenilka za σ^2. $\diamond$

Disperzija nepristranskih cenilk

Izmed nepristranskih cenilk istega parametra ζ je boljša tista, ki ima manjšo disperzijo – v povprečju daje bolj točne ocene. Če je razred cenilk parametra ζ *konveksen* (vsebuje tudi njihove konveksne kombinacije), obstaja v bistvu ena sama cenilka z najmanjšo disperzijo: Naj bo razred nepristranskih cenilk parametra ζ konveksen. Če sta C in C' nepristranski cenilki, obe z najmanjšo disperzijo σ^2, je $C = C'$ z verjetnostjo 1. Za to poglejmo

$$\mathsf{D}((C + C')/2) = \frac{1}{4}(\mathsf{D}C + \mathsf{D}C' + 2\mathsf{Cov}(C, C')) \leqslant \left(\frac{1}{2}(\sqrt{\mathsf{D}C} + \sqrt{\mathsf{D}C'})\right)^2 = \sigma^2.$$

Ker sta cenilki minimalni, mora biti tudi $\mathsf{D}((C+C')/2) = \sigma^2$ in dalje $\mathsf{Cov}(C, C') = \sigma^2$ oziroma $r(C, C') = 1$. Torej je $C' = aC + b$, $a > 0$ z verjetnostjo 1. Iz $\mathsf{D}C = \mathsf{D}C'$ izhaja $a = 1$, iz $\mathsf{E}C = \mathsf{E}C'$ pa še $b = 0$.

Srednja kvadratična napaka

Včasih je celo bolje vzeti pristransko cenilko z manjšo disperzijo, kot jo ima druga, sicer nepristranska, cenilka z veliko disperzijo. Mera *učinkovitosti* cenilk parametra ζ je *srednja kvadratična napaka*

$$q(C) = \mathsf{E}(C - \zeta)^2.$$

Ker velja

$$q(C) = \mathsf{E}(C - \mathsf{E}C + \mathsf{E}C - \zeta)^2 = \mathsf{E}(C - \mathsf{E}C)^2 + (\mathsf{E}C - \zeta)^2,$$

jo lahko zapišemo tudi v obliki

$$q(C) = \mathsf{D}C + B(C)^2.$$

Za nepristranske cenilke je $B(C) = 0$ in zato $q(C) = \mathsf{D}C$. Če pa je disperzija cenilke skoraj 0, je $q(C) \approx B(C)^2$.

12.2 Rao-Cramérjeva ocena

Naj bo f gostotna ali verjetnostna funkcija slučajne spremenljivke X in naj bo odvisna še od parametra ζ, tako da je $f(x; \zeta)$ njena vrednost v točki x. Združeno gostotno ali verjetnostno funkcijo slučajnega vzorca $(X_1, \ldots, X_n)$ označimo z L in ji pravimo *funkcija verjetja* (tudi *zanesljivosti*, angl. *likelihood*)

$$L(x_1, \ldots, x_n; \zeta) = f(x_1; \zeta) \cdots f(x_n; \zeta).$$

Velja

$$\int\int \ldots \int L(x_1, \ldots, x_n; \zeta)\, dx_1 \ldots dx_n = 1. \tag{12.1}$$

$L(X_1, \ldots, X_n)$ je funkcija vzorca – torej slučajna spremenljivka. Privzemimo, da je funkcija L vsaj dvakrat zvezno odvedljiva po ζ na nekem intervalu I in naj na tem intervalu tudi integral odvoda L po ζ enakomerno konvergira. Odvajajmo enakost (12.1) po ζ in upoštevajmo $\frac{\partial \ln L}{\partial \zeta} = \frac{1}{L}\frac{\partial L}{\partial \zeta}$ pa dobimo

$$\int\int \ldots \int \frac{\partial \ln L}{\partial \zeta} L\, dx_1 dx_2 \ldots dx_n = 0,$$

kar lahko tolmačimo kot

$$\mathsf{E}\frac{\partial \ln L}{\partial \zeta} = 0.$$

Naj bo sedaj C nepristranska cenilka parametra ζ, torej $\mathsf{E}C = \zeta$, oziroma zapisano z integrali $\int\int \ldots \int C L dx_1 dx_2 \ldots dx_n = \zeta$. Ker C ni odvisna od ζ, dobimo z odvajanjem po parametru ζ:

$$\int\int \ldots \int C \frac{\partial \ln L}{\partial \zeta} L\, dx_1 dx_2 \ldots dx_n = 1$$

kar z drugimi besedami pomeni
$$\mathsf{E}\left(C\frac{\partial \ln L}{\partial \zeta}\right) = 1.$$
Če to enakost združimo s prejšnjo (pomnoženo s ζ), dobimo:
$$\mathsf{E}\left((C-\zeta)\,\frac{\partial \ln L}{\partial \zeta}\right) = 1.$$
Od tu po $(\mathsf{E}XY)^2 \leqslant \mathsf{E}X^2\mathsf{E}Y^2$ izhaja naprej
$$1 = \left(\mathsf{E}\left((C-\zeta)\,\frac{\partial \ln L}{\partial \zeta}\right)\right)^2 \leqslant \mathsf{E}(C-\zeta)^2\,\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2 = \mathsf{D}C\;\,\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2,$$
kar da *Rao-Cramérjevo oceno*
$$\mathsf{D}C \geqslant \left(\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2\right)^{-1} = \left(-\mathsf{E}\frac{\partial^2 \ln L}{\partial \zeta^2}\right)^{-1} = \left(n\mathsf{E}\left(\frac{\partial \ln f}{\partial \zeta}\right)^2\right)^{-1}.$$

12.3 Učinkovitost cenilk

Rao-Cramérjeva ocena da absolutno spodnjo mejo disperzije za vse nepristranske cenilke parametra ζ (v dovolj gladkih porazdelitvah). Ta meja ni nujno dosežena. Cenilka, ki jo doseže, se imenuje *najučinkivitejša cenilka* parametra ζ in je ena sama (z verjetnostjo 1).

Kdaj pa je ta spodnja meja dosežena?

V neenakosti $(\mathsf{E}XY)^2 \leqslant \mathsf{E}X^2\mathsf{E}Y^2$, ki je uporabljena v izpeljavi Rao-Cramérjeve ocene, velja enakost natanko takrat, ko je $Y = cX$ z verjetnostjo 1. Torej velja v Rao-Cramérjevi oceni enakost natanko takrat, ko je
$$\frac{\partial \ln L}{\partial \zeta} = A(\zeta)(C-\zeta),$$
kjer je $A(\zeta)$ konstanta, odvisna od ζ in neodvisna od vzorca. Zato je tudi
$$(\mathsf{D}C)^{-1} = \mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2 = A(\zeta)^2\mathsf{E}(C-\zeta)^2 = A(\zeta)^2\mathsf{D}C$$
oziroma končno
$$\mathsf{D}C = |A(\zeta)|^{-1}.$$

Najučinkovitejše cenilke za parametre normalne porazdelitve

Naj bo $X \sim N(\mu,\sigma)$. Tedaj je
$$L = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-((\frac{X_1-\mu}{\sigma})^2+\cdots+(\frac{X_n-\mu}{\sigma})^2)/2}$$

in

$$\ln L = \ln \frac{1}{(2\pi)^{n/2}\sigma^n} - \frac{1}{2}\left(\left(\frac{X_1-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_n-\mu}{\sigma}\right)^2\right)$$

ter dalje

$$\frac{\partial \ln L}{\partial \mu} = \frac{X_1-\mu}{\sigma^2} + \cdots + \frac{X_n-\mu}{\sigma^2} = \frac{n}{\sigma^2}(\overline{X}-\mu).$$

Torej je vzorčno povprečje $\overline{X}$ najučinkovitejša cenilka za μ z disperzijo $\mathsf{D}\overline{X} = \sigma^2/n$. Prvi člen v izrazu za $\ln L$ lahko zapišemo tudi $-\frac{n}{2}(\ln 2\pi + \ln \sigma^2)$. Tedaj je, če privzamemo, da je μ znano število

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}((X_1-\mu)^2 + \cdots + (X_n-\mu)^2) = \frac{n}{2\sigma^4}(S_\mu^2 - \sigma^2).$$

To pomeni, da je $S_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ najučinkovitejša cenilka za parameter σ^2 z disperzijo $\mathsf{D}S_\mu^2 = 2\sigma^4/n$.

Primer: Za Poissonovo porazdelitev $P(\lambda)$ s parametrom λ, tj. $p_k = \lambda^k e^{-\lambda}/k!$, je

$$L = e^{-n\lambda}\,\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}$$

in dalje

$$\ln L = -n\lambda + (x_1+\cdots+x_n)\ln\lambda - \ln(x_1!\cdots x_n!)$$

ter končno

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1+\cdots+x_n}{\lambda} = \frac{n}{\lambda}(\overline{X}-\lambda). \tag{12.2}$$

Najučinkovitejša cenilka za parameter λ je $\overline{X}$ z disperzijo $\mathsf{D}\overline{X} = \lambda/n$. ◇

Naj bo C_0 najučinkovitejša cenilka parametra ζ in C kaka druga nepristranska cenilka. Tedaj je *učinkovitost* cenilke C določena s predpisom

$$e(C) = \frac{\mathsf{D}C_0}{\mathsf{D}C}.$$

Učinkovitost najučinkovitejše cenilke je $e(C_0) = 1$. Če najučinkovitejša cenilka ne obstaja, vzamemo za vrednost $\mathsf{D}C_0$ desno stran v Rao-Cramérjevi oceni.

Primer: Naj bo $X \sim N(\mu,\sigma)$. Pri velikih n-jih je vzorčna mediana $\tilde{X}$ – ocena za μ, porazdeljena približno po $N(\mu, \sigma\sqrt{\pi/2n})$. Torej je

$$e(\tilde{X}) = \frac{\mathsf{D}\overline{X}}{\mathsf{D}\tilde{X}} = \frac{\dfrac{\sigma^2}{n}}{\dfrac{\pi\sigma^2}{2n}} = \frac{2}{\pi} \approx 0{\cdot}64.$$

◇

Primer: Naj bo $X \sim N(\mu, \sigma)$. Če poznamo μ, je najučinkovitejša cenilka za σ^2 statistika S_μ^2 z disperzijo $\mathsf{D}S_\mu^2 = 2\sigma^4/n$. Popravljena vzorčna disperzija S^2 pa je nepristranska cenilka istega parametra z disperzijo $\mathsf{D}S^2 = 2\sigma^4/(n-1)$. Tokrat je

$$e(S^2) = \frac{\mathsf{D}S_\mu^2}{\mathsf{D}S^2} = \frac{\dfrac{2\sigma^4}{n}}{\dfrac{2\sigma^4}{n-1}} = \frac{n-1}{n}.$$

Iz tega vidimo, da $e(S^2) \to 1$, ko $n \to \infty$. Pravimo, da je cenilka S^2 *asimptotično najučinkovitejša cenilka* za σ^2. ◇

12.4 Metoda momentov

Recimo, da je za zvezno slučajno spremenljivko X njena gostota f odvisna od m parametrov $f(x; \zeta_1, \ldots, \zeta_m)$ in naj obstajajo momenti

$$z_k = z_k(\zeta_1, \ldots, \zeta_m) = \int_{-\infty}^{\infty} x^k f(x; \zeta_1, \ldots, \zeta_m)\, dx \qquad \text{za } k = 1, \ldots, m.$$

Če se dajo iz teh enačb enolično izračunati parametri $\zeta_1, \ldots, \zeta_m$ kot funkcije momentov $z_1, \ldots, z_m$:

$$\zeta_k = \varphi_k(z_1, \ldots, z_m),$$

potem so

$$C_k = \varphi_k(Z_1, \ldots, Z_m)$$

cenilke parametrov ζ_k po *metodi momentov*. k-ti vzorčni začetni moment $Z_k = \frac{1}{n}\sum_{i=1}^n X_i^k$ je cenilka za ustrezni populacijski moment z_k. Cenilke, ki jih dobimo po metodi momentov so dosledne.

Primer: Naj bo $X \sim N(\mu, \sigma)$. Tedaj je $z_1 = \mu$ in $z_2 = \sigma^2 + \mu^2$. Od tu dobimo $\mu = z_1$ in $\sigma^2 = z_2 - z_1^2$. Ustrezni cenilki sta

$$Z_1 = \overline{X} \quad \text{za } \mu \qquad \text{in} \qquad Z_2 - Z_1^2 = \overline{X^2} - \overline{X}^2 = S_0^2 \quad \text{za } \sigma^2,$$

torej vzorčno povprečje in disperzija. ◇

Metoda največjega verjetja

Funkcija verjetja

$$L(x_1, \ldots, x_n; \zeta) = f(x_1; \zeta) \cdots f(x_n; \zeta)$$

je pri danih $x_1, \ldots, x_n$ odvisna še od parametra ζ. Izberemo tak ζ, da bo funkcija L dosegla največjo vrednost. Če je L vsaj dvakrat zvezno odvedljiva, mora veljati

$$\frac{\partial L}{\partial \zeta} = 0 \qquad \text{in} \qquad \frac{\partial^2 L}{\partial \zeta^2} < 0.$$

Največja vrednost parametra je še odvisna od $x_1, \ldots, x_n$: $\zeta_{\max} = \varphi(x_1, \ldots, x_n)$. Tedaj je cenilka za parameter ζ enaka

$$C = \varphi(X_1, \ldots, X_n).$$

Metodo lahko posplošimo na večje število parametrov. Pogosto raje iščemo maksimum funkcije $\ln L$. Če najučinkovitejša cenilka obstaja, jo dobimo s to metodo.

Primer: Naj bo $X \sim B(1, p)$. Tedaj je $f(x; p) = p^x(1-p)^{1-x}$, kjer je $x = 0$ ali $x = 1$. Ocenjujemo parameter p. Funkcija verjetja ima obliko $L = p^x(1-p)^{n-x}$, kjer je sedaj $x \in \{0, \ldots, n\}$. Ker je $\ln L = x \ln p + (n-x)\ln(1-p)$, dobimo

$$\frac{\partial \ln L}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p},$$

ki je enak 0 pri $p = x/n$. Ker je v tem primeru

$$\frac{\partial^2 \ln L}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0,$$

je v tej točki maksimum. Cenilka po metodi največjega verjetja je torej $P = X/n$, kjer je X binomsko porazdeljena spremenljivka – frekvenca v n ponovitvah. Cenilka P je nepristranska, saj je $\mathsf{E}P = \mathsf{E}X/n = p$. Ker gre $\mathsf{D}P = \mathsf{D}X/n^2 = p(1-p)/n \to 0$ za $n \to \infty$, je P dosledna cenilka. Je pa tudi najučinkovitejša

$$\frac{\partial \ln L}{\partial p} = \frac{X}{p} - \frac{n-X}{1-p} = \frac{n}{p(1-p)}\left(\frac{X}{n} - p\right) = \frac{n}{p(1-p)}(P - p).$$ ◇

Primer: Nadaljujmo primer Poissonove porazdelitve. Odvod (12.2) je enak 0 za $\lambda = \overline{X}$. Drugi odvod v tej točki je

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{x_1 + \cdots + x_n}{\lambda^2} < 0,$$

kar pomeni, da je v tej točki maksimum. Cenilka za λ je po metodi največjega verjetja vzorčno povprečje $\overline{X}$. Je tudi najučinkovitejša cenilka za λ z disperzijo $\mathsf{D}\overline{X} = \lambda/n$. ◇

12.5 Vzorčna statistika (nadaljevanje)

12.5.1 (C) Vzorčne aritmetične sredine

Primer: Denimo, da se spremenljivka inteligenčni kvocient na populaciji porazdeljuje normalno z aritmetično sredino $\mu = 100$ in standardnim odklonom $\sigma = 15$, tj.

$$X \sim N(100, 15)$$

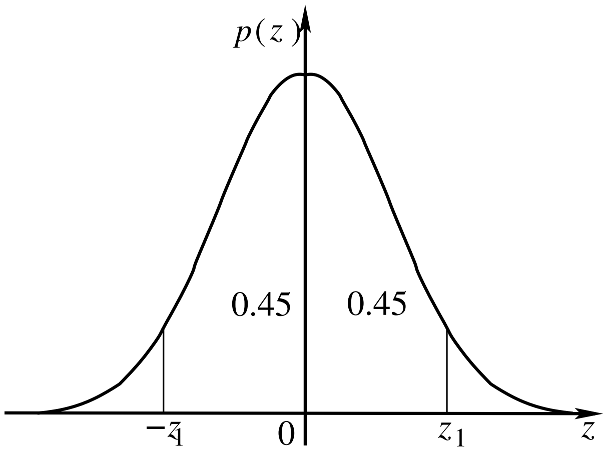
Denimo, da imamo vzorec velikosti $n = 225$. Tedaj se vzorčne aritmetične sredine porazdeljujejo normalno $\overline{X} \sim N(100, 15/\sqrt{225}) = N(100, 1)$. Izračunajmo, kolikšne vzorčne aritmetične sredine ima 90% vzorcev (simetrično na povprečje).

90% vzorčnih aritmetičnih sredin se nahaja na intervalu:

$$P(\overline{X}_1 < \overline{X} < \overline{X}_2) = 0{\cdot}90$$

$$P(-z_1 < z < z_1) = 0{\cdot}90 \quad \Longrightarrow \quad 2\Phi(z_1) = 0{\cdot}90$$

$$\Phi(z_1) = 0{\cdot}45 \quad \Longrightarrow \quad z_1 = 1{\cdot}65.$$



Potem se vzorčne aritmetične sredine nahajajo v intervalu

$$P\left(\mu - z_1\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_1\frac{\sigma}{\sqrt{n}}\right) = 0{\cdot}90 \quad \text{oziroma} \quad P\Big(100 - 1{\cdot}65 < \overline{X} < 100 + 1{\cdot}65\Big) = 0{\cdot}90.$$

90% vseh slučajnih vzorcev velikosti 225 enot bo imelo povprečja za inteligenčni kvocient na intervalu $(98{\cdot}35, 101{\cdot}65)$. Lahko preverimo, da bi bil ta interval v primeru večjega vzorca ožji. Npr. v primeru vzorcev velikosti $n = 2500$ je ta interval

$$P\left(100 - 1{\cdot}65\frac{15}{\sqrt{2500}} < \overline{X} < 100 + 1{\cdot}65\frac{15}{\sqrt{2500}}\right) = 0{\cdot}90 \quad \text{oziroma} \quad (99{\cdot}5, 100{\cdot}5). \quad \diamond$$

12.5.2 (D) Vzorčni deleži

Denimo, da želimo na populaciji oceniti delež enot π z določeno lastnostjo.

Zato na vsakem vzorcu poiščemo vzorčni delež p. Pokazati se da, da se za dovolj velike slučajne vzorce s ponavljanjem (za deleže okoli 0,5 je dovolj 20 enot ali več) vzorčni deleži porazdeljujejo približno normalno z

- aritmetično sredino vzorčnih deležev, ki je enaka deležu na populaciji
$$\mathsf{E}p = \pi,$$
- standardnim odklonom vzorčnih deležev
$$\mathsf{SE}(p) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Za manjše vzorce se vzorčni deleži porazdeljujejo binomsko. Cenilka populacijskega deleža je nepristranska cenilka, ker velja $\mathsf{E}p = \pi$.

Primer: V izbrani populaciji prebivalcev je polovica žensk $\pi = 0{\cdot}5$. Če tvorimo vzorce po $n = 25$ enot, nas zanima, kolikšna je verjetnost, da je v vzorcu več kot 55% žensk? To pomeni, da iščemo verjetnost $P(p > 0{\cdot}55)$. Vzorčni deleži p se porazdeljujejo približno normalno:

$$p \,:\, N\left(0{\cdot}5\,,\ \sqrt{\frac{\pi(1-\pi)}{n}}\right) = N\left(0{\cdot}5\,,\ \sqrt{\frac{0{\cdot}5\times 0{\cdot}5}{25}}\right) = N(0{\cdot}5\,,\ 0{\cdot}1).$$

Zato je

$$\begin{aligned} P(p > 0{\cdot}55) &= P\left(Z > \frac{0{\cdot}55 - 0{\cdot}5}{0{\cdot}1}\right) = P(Z > 0{\cdot}5) \\ &= 0{\cdot}5 - \Phi(0{\cdot}5) = 0{\cdot}5 - 0{\cdot}1915 = 0{\cdot}3085. \end{aligned}$$

Rezultat pomeni, da lahko pričakujemo, da bo pri približno 31% vzorcev delež žensk večji od $0{\cdot}55$. Poglejmo, kolikšna je ta verjetnost, če bi tvorili vzorce velikosti $n = 2500$ enot:

$$P(p > 0{\cdot}55) = P\left(Z > \frac{0{\cdot}55 - 0{\cdot}5}{\sqrt{\frac{0{\cdot}5(1-0{\cdot}5)}{2500}}}\right) = P(Z > 5) = 0{\cdot}5 - \Phi(5) = 0{\cdot}5 - 0{\cdot}5 = 0.$$

V 10-krat večjih vzorcih kot prej ne moremo pričakovati več kot 55% žensk. $\diamond$

12.5.3 (E) Razlika vzorčnih aritmetičnih sredin

Denimo, da imamo dve populaciji velikosti N_1 in N_2 in se spremenljivka X na prvi populaciji porazdeljuje normalno $N(\mu_1, \sigma)$, na drugi populaciji pa $N(\mu_2, \sigma)$ (standardna odklona sta na obeh populacijah enaka!). V vsaki od obeh populacij tvorimo neodvisno slučajne vzorce velikosti n_1 in n_2. Na vsakem vzorcu (s ponavljanjem) prve populacije izračunamo vzorčno aritmetično sredino $\overline{X}_1$ in podobno na vsakem vzorcu druge populacije $\overline{X}_2$. Dokazati se da, da je porazdelitev razlik vzorčnih aritmetičnih sredin normalna, kjer je

- matematično upanje razlik vzorčnih aritmetičnih sredin enako

$$\mathsf{E}(\overline{X}_1 - \overline{X}_2) = \mathsf{E}\overline{X}_1 - \mathsf{E}\overline{X}_2 = \mu_1 - \mu_2,$$

- disperzija razlik vzorčnih aritmetičnih sredin enaka

$$\mathsf{D}(\overline{X}_1 - \overline{X}_2) = \mathsf{D}\overline{X}_1 + \mathsf{D}\overline{X}_2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \cdot \frac{n_1 + n_2}{n_1 n_2}.$$

Primer: Dvema populacijama študentov na neki univerzi (tehnikom in družboslovcem) so izmerili neko sposobnost s povprečjema $\mu_t = 70$ in $\mu_d = 80$ točk in standardnim odklonom, ki je na obeh populacijah enak, $\sigma = 7$ točk.

Kolikšna je verjetnost, da je aritmetična sredina slučajnega vzorca družboslovcev ($n_d = 36$) večja za več kot 12 točk od aritmetične sredine vzorca tehnikov ($n_t = 64$)? Zanima nas torej verjetnost:

$$\begin{aligned} P(\overline{X}_d - \overline{X}_t > 12) &= P\left(Z > \frac{12-10}{7\sqrt{\frac{36+64}{36\cdot 64}}}\right) = P(Z > 1{\cdot}37) = 0{\cdot}5 - \Phi(1{\cdot}37) \\ &= 0{\cdot}5 - 0{\cdot}4147 = 0{\cdot}0853. \end{aligned}$$

Torej, približno 8·5% parov vzorcev je takih, da je povprečje družboslovcev večje od povprečja tehnikov za 12 točk. ◇

12.5.4 (F) Razlika vzorčnih deležev

Podobno kot pri porazdelitvi razlik vzorčnih aritmetičnih sredin naj bosta dani dve populaciji velikosti N_1 in N_2 z deležema enot z neko lastnostjo π_1 in π_2. Iz prve populacije tvorimo slučajne vzorce velikosti n_1 in na vsakem izračunamo delež enot s to lastnostjo p_1.

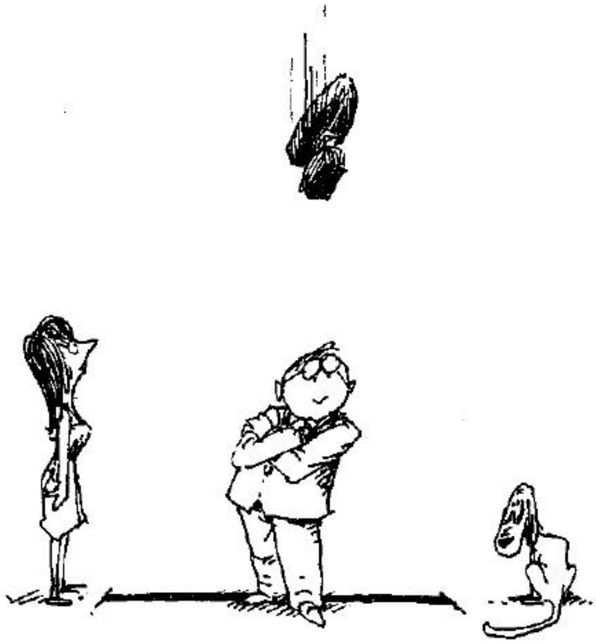
Podobno naredimo tudi na drugi populaciji: tvorimo slučajne vzorce velikosti n_2 in na njih dolocimo deleže p_2. Pokazati se da, da se za dovolj velike vzorce razlike vzorčnih deležev porazdeljujejo približno normalno z

- matematičnim upanjem razlik vzorčnih deležev

$$\mathsf{E}(p_1 - p_2) = \mathsf{E}p_1 - \mathsf{E}p_2 = \pi_1 - \pi_2,$$

- disperzijo razlik vzorčnih deležev

$$\mathsf{D}(p_1 - p_2) = \mathsf{D}p_1 + \mathsf{D}p_2 = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}.$$



Poglavje 13

Intervali zaupanja

Denimo, da s slučajnim vzorcem ocenjujemo parameter γ. Poskušamo najti statistiko g, ki je nepristranska, tj. $\mathsf{E}g = \gamma$ in se na vseh možnih vzorcih vsaj približno normalno porazdeljuje s standardno napako $\mathsf{SE}(g)$. Nato poskušamo najti interval, v katerem se bo z dano gotovostjo $(1-\alpha)$ nahajal ocenjevani parameter:

$$P(a < \gamma < b) = 1 - \alpha,$$

kjer je a je spodnja meja zaupanja, b je zgornja meja zaupanja, α verjetnost tveganja oziroma $1-\alpha$ verjetnost gotovosti. Ta interval imenujemo **interval zaupanja** in ga interpretiramo takole: z verjetnostjo tveganja α se parameter γ nahaja v tem intervalu.

Konstruirajmo interval zaupanja. Na osnovi omenjenih predpostavk o porazdelitvi statistike g lahko zapišemo, da se statistika

$$Z = \frac{g - \mathsf{E}g}{\mathsf{SE}(g)} = \frac{g-\gamma}{\mathsf{SE}(g)}$$

porazdeljuje standardizirano normalno $N(0,1)$. Če tveganje α porazdelimo polovico na levo in polovico na desno na konce normalne porazdelitve, lahko zapišemo

$$P\left(-z_{\alpha/2} < \frac{g-\gamma}{\mathsf{SE}(g)} < z_{\alpha/2}\right) = 1 - \alpha.$$

Po ustrezni preureditvi lahko izpeljemo naslednji interval zaupanja za parameter γ

$$P\Big(g - z_{\alpha/2}\,\mathsf{SE}(g) < \ \gamma \ < g + z_{\alpha/2}\,\mathsf{SE}(g)\Big) = 1 - \alpha,$$

kjer je $z_{\alpha/2}$ določen le s stopnjo tveganja α. Vrednosti $z_{\alpha/2}$ lahko razberemo iz tabele za verjetnosti za standardizirano normalno porazdelitev, ker velja

$$\Phi(z_{\alpha/2}) = 0{\cdot}5 - \frac{\alpha}{2}.$$

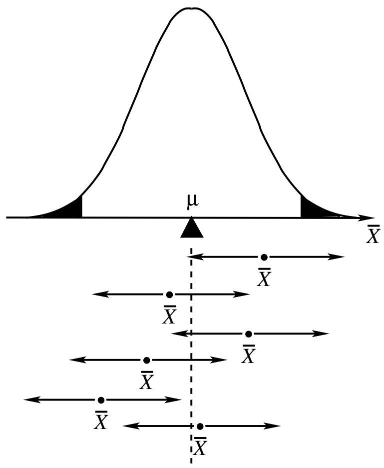
Podajmo vrednost $z_{\alpha/2}$ za nekaj najbolj standardnih tveganj:

- $\alpha = 0{\cdot}10, \quad z_{\alpha/2} = 1{\cdot}65,$
- $\alpha = 0{\cdot}05, \quad z_{\alpha/2} = 1{\cdot}96,$
- $\alpha = 0{\cdot}01, \quad z_{\alpha/2} = 2{\cdot}58.$

13.1 Pomen stopnje tveganja

Za vsak slučajni vzorec lahko ob omenjenih predpostavkah izračunamo ob izbrani stopnji tveganja α interval zaupanja za parameter γ. Ker se podatki vzorcev razlikujejo, se razlikujejo vzorčne ocene parametrov in zato tudi izračunani intervali zaupanja za parameter γ. To pomeni, da se intervali zaupanja od vzorca do vzorca razlikujejo. Meji intervala sta slučajni spremenljivki.

Primer: Vzemimo stopnjo tveganja $\alpha = 0{\cdot}05$. Denimo, da smo izbrali 100 slučajnih vzorcev in za vsakega izračunali interval zaupanja za parameter γ. Tedaj lahko pričakujemo, da 5 intervalov zaupanja od 100 ne bo pokrilo iskanega parametra γ. Povedano je lepo predstavljeno tudi grafično (glej sliko). V tem primeru ocenjujemo parameter aritmetično sredino inteligenčnega kvocienta. Kot vemo, se vzorčne aritmetične sredine $\overline{X}$ za dovolj velike vzorce porazdeljujejo nomalno. $\diamond$



Denimo, da v tem primeru poznamo vrednost parametra ($\mu = 100$). Za več slučajnih vzorcev smo izračunali in prikazali interval zaupanja za μ ob stopnji tveganja $\alpha = 0{\cdot}05$. Predstavitev več intervalov zaupanja za aritmetično sredino μ pri 5% stopnji tveganja: približno 95% intervalov pokrije parameter μ.

13.2 Intervalsko ocenjevanje parametrov

Naj bo X slučajna spremenljivka na populaciji G z gostoto verjetnosti odvisno od parametra ζ. Slučajna množica $M \subset \mathbb{R}$, ki je odvisna le od slučajnega vzorca, ne pa od parametra ζ, se imenuje *množica zaupanja* za parameter ζ, če obstaja tako število α, $0 < \alpha < 1$, da velja $P(\zeta \in M) = 1 - \alpha$. Število $1 - \alpha$ imenujemo tedaj *stopnja zaupanja*; število α pa *stopnja tveganja*. Stopnja zaupanja je običajno 95% ali 99% – $\alpha = 0{\cdot}05$ ali $\alpha = 0{\cdot}01$. Pove nam, kakšna je verjetnost, da M vsebuje vrednost parametra ζ ne glede na to, kakšna je njegova dejanska vrednost. Če je množica M interval $M = [A, B]$, ji rečemo *interval zaupanja* (za parameter ζ). Njegovi krajišči sta funkciji slučajnega vzorca – torej statistiki.

Naj bo $X \sim N(\mu, \sigma)$ in recimo, da poznamo parameter σ in ocenjujemo parameter μ. Izberimo konstanti a in b, $b > a$, tako da bo $P(a \leqslant Z \leqslant b) = 1 - \alpha$, kjer je $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$. Tedaj je

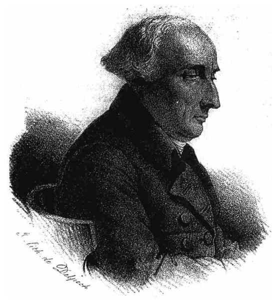
$$P\left(\overline{X} - \frac{b\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} - \frac{a\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Označimo $A = \overline{X} - b\sigma/\sqrt{n}$ in $B = \overline{X} - a\sigma/\sqrt{n}$.

Za katera a in b je interval $[A, B]$ najkrajši?

Pokazati je mogoče (Lagrangeova funkcija), da mora biti $a = -b$ in $\Phi(b) = (1-\alpha)/2$; oziroma, če označimo $b = z_{\alpha/2}$, velja $P(Z > z_{\alpha/2}) = \alpha/2$. Iskani interval je torej

$$A = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ in } B = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$



tj., z verjetnostjo $1-\alpha$ je $|\overline{X} - \mu| < z_{\alpha/2}\sigma/\sqrt{n}$.

Od tu dobimo, da mora za to, da bo napaka manjša od ε z verjetnostjo $1-\alpha$, veljati $n > (z_{\alpha/2}\sigma/\varepsilon)^2$.

Če pri porazdelitvi $X \sim N(\mu, \sigma)$ tudi parameter σ ni znan, ga nadomestimo s cenilko S in moramo zato uporabiti Studentovo statistiko $T = \frac{\overline{X}-\mu}{S}\sqrt{n}$. Ustrezni interval je tedaj

$$A = \overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \qquad B = \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}},$$

kjer je $P(T > t_{\alpha/2}) = \alpha/2$.

Če pa bi ocenjevali parameter σ^2, uporabimo statistiko $\chi^2 = (n-1)S^2/\sigma^2$, ki je porazdeljena po $\chi^2(n-1)$. Tedaj je

$$A = \frac{(n-1)S^2}{b} \qquad \text{in} \qquad B = \frac{(n-1)S^2}{a}.$$

Konstanti a in b včasih določimo iz pogojev

$$P(\chi^2 < a) = \alpha/2 \qquad \text{in} \qquad P(\chi^2 > b) = \alpha/2,$$

najkrajši interval pa dobimo, ko velja zveza $a^2 p(a) = b^2 p(b)$ in seveda $\int_a^b p(t)dt = 1-\alpha$.

Teoretična interpretacija koeficienta zaupanja $(1-\alpha)$

Če zaporedoma izbiramo vzorce velikosti n iz dane populacije in konstruiramo $[(1-a)100]\%$ interval zaupanja za vsak vzorec, potem lahko pričakujemo, da bo $[(1-a)100]\%$ intervalov dalo prvo vrednost parametra.

stopnja tveganja = 1 − stopnja zaupanja

13.2.1 Povprečje μ s poznanim σ

Točki

$$\overline{y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

prestavljata krajišči intervala zaupanja, pri čemer je:

$z_{\alpha/2}$ vrednost spremenljivke, ki zavzame površino $\alpha/2$ na svoji desni;

σ je standardni odklon za populacijo;

n je velikost vzorca;

$\overline{y}$ je vrednost vzorčnega povprečja.

13.2.2 Velik vzorec za povprečje μ

$$\overline{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right),$$

kjer je s standardni odklon vzorca.

Primer: Na vzorcu velikosti $n = 151$ podjetnikov v majhnih podjetjih v Sloveniji, ki je bil izveden v okviru ankete 'Drobno gospodarstvo v Sloveniji' (Prašnikar, 1993), so izračunali, da je povprečna starost anketiranih podjetnikov $\overline{X} = 40{\cdot}4$ let in standardni odklon $s = 10{\cdot}2$ let. Pri 5% tveganju želimo z intervalom zaupauja oceniti povprečno starost podjetnikov v majhnih podjetjih v Sloveniji:

$$P\left(\overline{y} - z_{\alpha/2}\frac{s}{\sqrt{n}} \;<\; \mu \;<\; \overline{y} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) \;=\; 1-\alpha.$$

oziroma

$$40{\cdot}4 - \frac{1{\cdot}96 \times 10{\cdot}2}{\sqrt{151}} \;<\; \mu \;<\; 40{\cdot}4 + \frac{1{\cdot}96 \times 10{\cdot}2}{\sqrt{151}}$$

in končno $40{\cdot}4 - 1{\cdot}6 < \mu < 40{\cdot}4 + 1{\cdot}6$. Torej je 95% interval zaupanja za povprečno starost podjetnikov v majhnih podjetjih v Sloveniji med $38{\cdot}8$ in $42{\cdot}0$ leti. ◇

13.2.3 Majhen vzorec za povprečje μ

$$\overline{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right),$$

kjer je porazdelitev spremenljivke y vzeta na osnovi $(n-1)$ prostostnih stopenj.
Privzeli smo, da je vzorec iz približno **normalno porazdeljene** populacije.

Primer: Vzemimo, da se spremenljivka X - število ur branja dnevnih časopisov na teden - porazdeljuje normalno $N(\mu, \sigma)$. Na osnovi podatkov za 7 slučajno izbranih oseb ocenimo interval zaupanja za aritmetično sredino pri 10% tveganju. Podatki in ustrezni izračuni so:

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
5	-2	4
7	0	0
9	2	4
7	0	0
6	-1	1
10	3	9
5	-2	4
49	0	22

$$\overline{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{49}{7} = 7,$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \overline{X})^2}{n-1} = \frac{22}{6} = 3{\cdot}67.$$

Iz tabele za t–porazdelitev preberemo, da je $t_{\alpha/2}(n-1) = t_{0{\cdot}05}(6) = 1{\cdot}943$ in interval zaupanja je

$$7 - 1{\cdot}943 \times \frac{1{\cdot}9}{\sqrt{7}} < \mu < 7 + 1{\cdot}943 \times \frac{1{\cdot}9}{\sqrt{7}} \quad \text{oziroma} \quad 7 - 1{\cdot}4 < \mu < 7 + 1{\cdot}4. \qquad \diamond$$

13.2.4 Razlika povprečij $\mu_1 - \mu_2$ s poznanima σ_1 in σ_2

$$\overline{y}_1 - \overline{y}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

13.2.5 Veliki vzorec za razliko povprečij $\mu_1 - \mu_2$ z neznanima σ_1 in σ_2

$$\overline{y}_1 - \overline{y}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

13.2.6 Majhen vzorec za razliko povprečij $\mu_1 - \mu_2$ z neznanima $\sigma_1 = \sigma_2$

$$\overline{y}_1 - \overline{y}_2 \pm t_{\alpha/2, n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad \text{kjer je} \qquad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

Privzeli smo, da sta obe populaciji porazdeljeni **približno normalni**, da sta varianci **enaki** in da so naključni vzorci izbrani **neodvisno**.

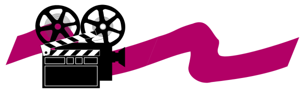
13.2.7 Majhen vzorec za razliko povprečij $\mu_1 - \mu_2$ z neznanima σ_1 in σ_2

$$\overline{y}_1 - \overline{y}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \text{kjer je} \quad \nu = \left\lfloor \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \right\rfloor.$$

Če ν ni naravno število, zaokroži ν navzdol do najbližjega naravnega števila za uporabo t-tabele.

Primer: Naslednji podatki predstavljajo dolžine filmov, ki sta jih naredila dva filmska studija. Izračunaj 90%-ni interval zaupanja za razliko med povprečnim časom filmov, ki sta jih producirala ta dva studija. Predpostavimo, da so dolžine filmov porazdeljene **približno normalno**. Čas (v minutah)

Studio 1:	103	94	110	87	98		
Studio 2:	97	82	123	92	175	88	118



Podatke vnesemo v Minitab. Dva vzorca T-Test in interval zaupanja

```
Dva vzorca T za C1 : C2
      N    povpr.   St.odk.   SE povpr.
C1    5    98.40     8.73       3.9
C2    7    110.7     32.2       12

90\%-ni interval zaupanja za mu C1-mu C2: (-36.5, 12)
T-TEST mu C1=mu C2 (vs ni=):
T = -0.96   P = 0.37   DF = 7
```

◇

13.2.8 Velik vzorec za razliko $\mu_d = \mu_1 - \mu_2$

$$\overline{d} \pm z_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right), \quad \text{kjer je } n \text{ število parov.}$$

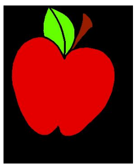
13.2.9 Majhen vzorec za razliko $\mu_d = \mu_1 - \mu_2$

$$\overline{d} \pm t_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right), \quad \text{kjer je } n \text{ število parov.}$$

Privzeli smo, da je populacija razlik parov normalno porazdeljena.

Primer: Špricanje jabolk lahko pozroči kontaminacijo zraka. Zato so v času najbolj intenzivnega špricanja zbrali in analizirali vzorce zraka za vsak od 11ih dni. Raziskovalci želijo vedeti ali se povprečje ostankov škropiv (diazinon) razlikuje med dnevom in nočjo. Analiziraj podatke za 90% interval zaupanja.

```
            Diazinon Residue   razlika
Datum        dan        noč    dan-noč
Jan. 11      5.4       24.3     -18.9
     12      2.7       16.5     -13.8
     13     34.2       47.2     -13.0
     14     19.9       12.4       7.5
     15      2.4       24.0     -21.6
     16      7.0       21.6     -14.6
     17      6.1      104.3     -98.2
     18      7.7       96.9     -89.2
     19     18.4      105.3     -86.9
     20     27.1       78.7     -51.6
     21     16.9       44.6     -27.7
```



Podatke vnesemo v Minitab, pri čemer sta drugi in tretji stolpec zgoraj `C1` in `C2`.

```
MTB > Let C3=C1-C2.

T interval zaupanja

Spremen. N   povpr. Stdev SEpovpr.
C3       11   -38.9   36.6    11.0
```

Torej je 90·0% interval zaupanja (58·9, 18·9). ◇

Za p – delež populacije, $\hat{p}$ – delež vzorca, kjer je $\hat{p} = y/n$ in je y število uspehov v n poskusih.

13.2.10 Delež π s poznanim $\sigma_{\hat{\pi}}$

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}}.$$

13.2.11 Veliki vzorec za delež populacije

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Privzeli smo, da je velikost vzorca n dovolj velika, da je aproksimacija veljavna.
Izkušnje kažejo, da je za izpolnitev pogoja "dovolj velik vzorec" priporočljivo privzeti (angl. rule of thumb): $n\hat{p} \geqslant 4$ in $n\hat{q} \geqslant 4$.

Primer: Na vzorcu ($n = 151$), ki je bil izveden v okviru ankete 'Drobno gospodarstvo v Sloveniji', so izračunali. da je delež obrtnih podjetij $p = 0{\cdot}50$. Pri 5% tveganju želimo z intervalom zaupanja oceniti delež obrtnih majhnih podjetij v Sloveniji. Po zgornji formuli dobimo

$$0{\cdot}50 - 1{\cdot}96\,\sqrt{\frac{0{\cdot}50 \times 0{\cdot}50}{151}} \quad < \quad \pi \quad < \quad 0{\cdot}50 + 1{\cdot}96\,\sqrt{\frac{0{\cdot}50 \times 0{\cdot}50}{151}}$$

oziroma $0{\cdot}50 - 0{\cdot}08 < \pi < 0{\cdot}50 + 0{\cdot}08$. S 5% stopnjo tveganja trdimo, da je delež obrtnih majhnih podjetij v Sloveniji glede na vsa majhna podjetja med 0·42 in 0·58. ◇

13.2.12 Razlika deležev $\pi_1 - \pi_2$ s poznanim $\sigma_{\hat{p}_1 - \hat{p}_2}$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{\hat{p}_1-\hat{p}_2}.$$

13.2.13 Veliki vzorec za razliko deležev $\pi_1 - \pi_2$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

Privzeli smo, da je velikost vzorca n dovolj velika, da je aproksimacija veljavna.

Izkušnje kažejo, da je za izpolnitev pogoja "dovolj velik vzorec" priporočljivo privzeti: $n_1\hat{p}_1 \geqslant 4$, $n_1\hat{q}_1 \geqslant 4$, $n_2\hat{p}_2 \geqslant 4$ in $n_2\hat{q}_2 \geqslant 4$.

13.2.14 Veliki vzorec za varianco σ^2

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}.$$

Privzeli smo, da je vzorec iz približno **normalno porazdeljene** populacije.

Primer: Vzemimo prejšnji primer spremenljivke o številu ur branja dnevnih časopisov na teden. Za omenjene podatke iz vzorca ocenimo z intervalom zaupanja varianco pri 10% tveganju. Iz tabele za χ^2–porazdelitev preberemo, da je

$$\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0{\cdot}95}(6) = 12{\cdot}6, \qquad \chi^2_{\alpha/2}(n-1) = \chi^2_{0{\cdot}05}(6) = 1{\cdot}64.$$

90% interval zaupanja za varianco je tedaj

$$\frac{6 \times 3{\cdot}67}{12{\cdot}6} < \sigma^2 < \frac{6 \times 3{\cdot}67}{1{\cdot}64} \qquad \text{oziroma} \qquad 1{\cdot}75 < \sigma^2 < 13{\cdot}43. \qquad \diamond$$

13.2.15 Kvocient varianc σ_1^2/σ_2^2

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} \leqslant \frac{\sigma_1^2}{\sigma_2^2} \leqslant \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,n_2-1,n_1-1}}.$$

Privzeli smo, da sta obe populaciji iz katerih izbiramo vzorce, imata **približno normalni porazdelitvi** relativnih frekvenc in da so naključni vzorci izbrani **neodvisno** iz obeh populacij.

13.3 Izbira velikosti vzorca

V tem razdelku bomo izbirali velikosti vzorcev za oceno različnih parametrov, ki je pravilna znotraj H enot z verjetnostjo $(1-\alpha)$.

- Populacijsko povprečje μ: $n = \left(\dfrac{z_{\alpha/2}\,\sigma}{H}\right)^2$. Populacijski odklon je običajno aproksimiran.
- Razlika med parom populacijskih povprečij, tj. $\mu_1 - \mu_2$: $n_1 = n_2 = \left(\dfrac{z_{\alpha/2}}{H}\right)^2 (\sigma_1^2 + \sigma_2^2)$.
- Populacijski delež π: $n = \left(\dfrac{z_{\alpha/2}}{H}\right)^2 pq$.

  Opozorilo: v tem primeru potrebujemo oceni za p in q. Če nimamo nobene na voljo, potem uporabimo $p = q = 0{\cdot}5$ za konzervativno izbiro števila n.
- Razlika med parom populacijskih deležev, tj. $p_1 - p_2$: $n_1 = n_2 = \left(\dfrac{z_{\alpha/2}}{H}\right)^2 (p_1q_1 + p_2q_2)$.

Poglavje 14

Preverjanje statističnih domnev

Načrt

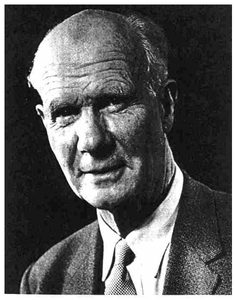
- postopek
- elementi
  - napake 1. in 2. vrste
  - značilno razlikovanje
  - moč statističnega testa
- testi
  - centralna tendenca
  - delež
  - varianca

Cilj

- postaviti domnevo (trditev) o populaciji,
- izberati vzorec, s katerim bomo preverili domnevo,
- zavrniti ali sprejeti domnevo.

Domneva je testirana z določanjem verjetja,
da dobimo določen rezultat, kadar jemljemo vzorce iz
populacije s predpostavljenimi vrednostimi parametrov.

Teorijo preverjanja domnev sta v 20. in 30. letih
prejšnjega stoletja razvila J. Neyman in E.S. Pearson.



EGON SHARPE PEARSON

Statistična domneva (ali hipoteza) je vsaka domneva o porazdelitvi slučajne spremenljivke X na populaciji. Če poznamo vrsto (obliko) porazdelitve $f(x;\zeta)$ in postavljamo/raziskujemo domnevo o parametru ζ, govorimo o *parametrični domnevi*. Če pa je vprašljiva tudi sama vrsta porazdelitve, je domneva *neparametrična*. Domneva je *enostavna*, če natačno določa porazdelitev (njeno vrsto in točno vrednost parametra); sicer je *sestavljena*.

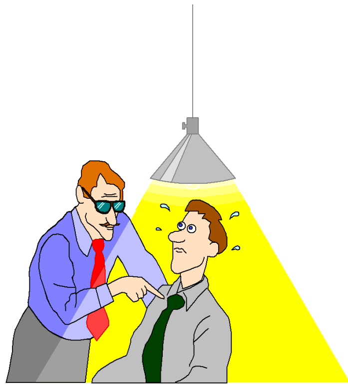
Primer: Naj bo $X \sim N(\mu, \sigma)$. Če poznamo σ, je domneva $H : \mu = 0$ enostavna; Če pa parametra σ ne poznamo, je sestavljena. Primer sestavljene domneve je tudi $H : \mu > 0$. $\diamond$

Statistična domneva je lahko pravilna ali napačna. Želimo seveda sprejeti pravilno domnevo in zavrniti napačno. Težava je v tem, da o pravilnosti/napačnosti domneve ne moremo biti gotovi, če jo ne preverimo na celotni populaciji. Ponavadi se odločamo le na podlagi vzorca. Če vzorčni podatki preveč odstopajo od domneve, rečemo, da niso *skladni* z domnevo, oziroma, da so *razlike značilne*, in domnevo zavrnemo. Če pa podatki domnevo podpirajo, jo ne zavrnemo – včasih jo celo sprejmemo. To ne pomeni, da je domneva pravilna, temveč da ni zadostnega razloga za zavrnitev.

- **Ničelna domneva (H_0)**
  – je trditev o lastnosti populacije za katero predpostavimo, da drži
  (oziroma za katero verjamemo, da je resnična),
  – je trditev, ki jo test skuša ovreči.
- **Alternativna (nasprotna) domneva (H_a)**
  – je trditev nasprotna ničelni domnevi,
  – je trditev, ki jo s testiranjem skušamo dokazati.

Okvirni postopek testiranja domneve

- postavimo ničelno in alternativno domnevo,
- izberemo testno statistiko,
- določimo zavrnitveni kriterij,
- izberemo naključni vzorec,
- izračunamo vrednost na osnovi testne statistike,
- sprejmemo odločitev,
- naredimo ustrezen zaključek.



14.1 Ilustrativni primer (ameriški sodni sistem)

- H_0: obtoženec je nedolžen (ničelna domneva),
- H_a: obtoženec je kriv (alternativna domneva).

Odločitev in zaključek

- Porota je spoznala obtoženca za **krivega**. Zaključimo, da je bilo dovolj dokazov, ki nas prepričajo, da je obtoženec storil kaznivo dejanje.
- Porota je spoznala obtoženca za **nedolžnega**. Zaključimo, da je ni bilo dovolj dokazov, ki bi nas prepričali, da je obtoženec storil kaznivo dejanje.

Elementi preverjanja domneve

		odločitev	
		nedolžen	**kriv**
dejansko stanje	**nedolžen**	pravilna odločitev	napaka 1. vrste (α)
	kriv	napaka 2. vrste (β)	moč testa ($1-\beta$)

- Verjetnost napake 1. vrste (α) je verjetnost za obtožbo nedolžnega obtoženca.
- Značilno razlikovanje (signifikantno) oziroma **stopnja značilnosti**.
- Količina dvoma (α), ki ga bo porota še sprejela:
  - kriminalna tožba: Beyond a reasonable doubt...
  - civilna tožba: The preponderance of evidence must suggest...
- Verjetnost napake 2. vrste (β) je verjetnost,
  da spoznamo krivega obtoženca za nedolžnega.
- Moč testa ($1-\beta$) je verjetnost, da obtožimo krivega obtoženca.

Sodba

- breme dokazov,
- potrebno je prepričati poroto, da je obtoženi kriv (alternativna domneva) preko določene stopnje značilnosti:
  - kriminalna tožba: Reasonable Doubt,
  - civilna tožba: Preponderance of evidence.

Obramba

- Ni bremena dokazovanja.
- Povzročiti morajo dovolj dvoma pri poroti, če je obtoženi resnično kriv.

14.2 Alternativna domneva in definicije

Za začetek si oglejmo parametrične statistične domneve:

- ničelna domneva $H_0 : q = q_0$
- alternativna domneva
  - $H_a : q \neq q_0$
  - $H_a : q > q_0$
  - $H_a : q < q_0$

Nekaj konkretnih primerov ničelnih domnev:

$$H_0 : \mu = 9\text{mm (premer 9 milimetrskega kroga)},$$
$$H_0 : \mu = 600\text{km (doseg novih vozil)},$$
$$H_0 : \mu = 3 \text{ dnevi}$$

Neusmerjena H_a
Merjenje 9 mm kroga:

$$H_0 : \mu = 9mm,$$
$$H_a : \mu \neq 9mm.$$

"Manj kot" H_a
Proizvalajec trdi, da je to doseg novih vozil:

$$H_0 : \mu = 600 \text{ km},$$
$$H_a : \mu < 600 \text{ km}.$$

"Več kot" H_a
Čas odsotnosti določenega artikla pri neposredni podpori:

$$H_0 : \mu = 3 \text{ dnevi},$$
$$H_a : \mu > 3 \text{ dnevi}.$$

Definicije

1. **Napaka 1. vrste** je zavrnitev ničelne domneve, če je le-ta pravilna. Verjetnost, da naredimo napako 1. vrste, označimo s simbolom α in ji pravimo **stopnja tveganja**, $(1 - \alpha)$ pa je **stopnja zaupanja**.

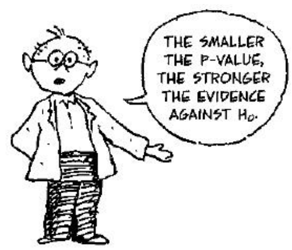
2. ***P*-vrednost** oziroma **(bistvena) stopnja značilnosti testa** (tudi **signifikantnosti**) je največja vrednost parametra α, ki jo je vodja eksperimenta pripravljen sprejeti (zgornja meja za napako 1. vrste) glede na dan vzorec.
3. Če ne zavrnemo ničelno domnevo, v primeru, da je napačna, pravimo, da gre za **napako 2. vrste**. Verjetnost, da naredimo napako 2. vrste, označimo s simbolom β.
4. **Moč statističnega testa**, $(1-\beta)$, je verjetnost zavrnitve ničelne domneve v primeru, ko je le-ta v resnici napačna.

		odločitev	
		FTR H_0	**zavrni** H_0
dejansko stanje	H_0 **je pravilna**		(α)
	H_0 **je napačna**	(β)	$(1-\beta)$

velikost vzorca n	napaka 1.vrste α	napaka 2.vrste β	moč $1-\beta$
konst.	$\uparrow$	$\downarrow$	$\uparrow$
konst.	$\downarrow$	$\uparrow$	$\downarrow$
povečanje	$\downarrow$	$\downarrow$	$\uparrow$
zamnjšanje	$\uparrow$	$\uparrow$	$\downarrow$

Še enkrat z drugimi besedami: P-vrednost (ali ugotovljena bistvena stopnja značilnosti za določen statistični test) je verjetnost (ob predpostavki, da drži H_0), da ugotovimo vrednost testne statistike, ki je vsaj toliko v protislovju s H_0 in podpira H_a kot tisto, ki je izračunana iz vzorčnih podatkov.

- Sprejemljivost domneve H_0 na osnovi vzorca
  - Verjetnost, da je opazovani vzorec (ali podatki) bolj ekstremni, če je domneva H_0 pravilna.
- Najmanjši α pri katerem zavrnemo domnevo H_0:
  - če je P-vrednost $> \alpha$, potem FTR H_0,
  - če je P-vrednost $< \alpha$, potem zavrni H_0.



14.3 Primeri statističnih domnev

Primer: Pascal je visoko-nivojski programski jezik, ki smo ga nekoč pogosto uporabljali na miniračunalnikih in microprocesorjih. Narejen je bil eksperiment, da bi ugotovili delež Pascalovih spremenljivk, ki so tabelarične spremenljivke (v kontrast skalarim spremenljivkam, ki so manj učinkovite, glede na čas izvajanja). 20 spremenljivk je bilo naključno izbranih iz množice Pascalovih programov, pri tem pa je bilo zabeleženo število tabelaričnih spremenljivk Y. Predpostavimo, da želimo testirati domnevo, da je Pascal bolj učinkovit jezik kot Agol, pri katerem je 20% spremenljivk tabelaričnih.

Postavitev statistične domneve

- $H_0 \,:\, p = 0{\cdot}20$ (ničelna domneva),
- $H_a \,:\, p > 0{\cdot}20$ (alternativna domneva),

kjer je p verjetnost, da imamo tabelarično spremenljivko na vsakem poskusu od 20ih neodvisnih poskusov.

(a) Določi α za območje zavrnitve $y > 8$. Izračunati želimo verjetnost, da se bo zgodila napaka 1. vrste, torej da bomo zavrnili pravilno domnevo. Predpostavimo, da je domneva H_0 pravilna, tj. $Y \,:\, \mathrm{B}(20, 0{\cdot}2)$. Če se bo zgodilo, da bo Y pri izbranem vzorcu večji ali enak 8, bom domnevo zavrnili, čeprav je pravilna. Torej velja:

$$\begin{aligned}\alpha &= \mathrm{P}(Y \geqslant 8) = 1 - \mathrm{P}(Y \leqslant 7) = 1 - \sum_{i=0}^{7} \mathrm{P}(Y = i) \\ &= 1 - \sum_{i=0}^{7} \binom{20}{i}\, 0{\cdot}2^i\; 0{\cdot}2^{20-i} = 1 - 0{\cdot}9679 = 0{\cdot}0321 = 3{\cdot}21\%.\end{aligned}$$

(b) Določi α za območje zavrnitve $y \geqslant 5$. Do rezultata pridemo na enak način kot v prejšnji točki:

$$\begin{aligned}\alpha &= \mathrm{P}(Y \geqslant 5) = 1 - \mathrm{P}(Y \leqslant 4) = 1 - \sum_{i=0}^{4} \mathrm{P}(Y = i) \\ &= 1 - \sum_{i=0}^{4} \binom{20}{i} 0{\cdot}2^i\; 0{\cdot}2^{20-i} = 1 - 0{\cdot}6296 = 0{\cdot}3704 = 37{\cdot}04\%.\end{aligned}$$

(c) Določi β za območje zavrnitve $Y \geqslant 8$, če je $p = 0{\cdot}5$. Izračunati želimo verjetnost, da se bo zgodila napaka 2. vrste, torej da bomo sprejeli napačno domnevo. Ker vemo, da je $p = 0{\cdot}5$, velja $Y \sim \mathrm{B}(20, 0{\cdot}5)$. Napačno domnevo bomo sprejeli, če bo y pri izbranem vzorcu manjši od 8.

$$\beta = \mathrm{P}(y \leqslant 7) = \sum_{i=0}^{7} \binom{20}{i} 0{\cdot}5^i\; 0{\cdot}5^{20-i} = 0{\cdot}1316 = 13{\cdot}16\%.$$

(d) Določi β za območje zavrnitve $y \geqslant 5$, če je $p = 0{\cdot}5$. Do rezultata pridemo na enak način kot v prejšnji točki:

$$\beta = \mathrm{P}(y \leqslant 4) = \sum_{i=0}^{4} \binom{20}{i} 0{\cdot}5^i\; 0{\cdot}5^{20-i} = 0{\cdot}0059 = 0{\cdot}59\%.$$

(e) Katero območje zavrnitve $y \geqslant 8$ ali $y \geqslant 5$ je bolj zaželeno, če želimo minimizirati verjetnost napake 1. stopnje oziroma če želimo minimizirati verjetnost napake 2. stopnje. Napako 1. stopnje minimiziramo z izbiro območja $y \geqslant 8$, napako 2. stopnje pa z izbiro območja $y \geqslant 5$.

(f) Določi območje zavrnitve $y \geqslant a$ tako, da je α približno 0·01. Na osnovi točke (e) zaključimo, da se z večanjem števila a manjša verjetnost α in s poskušanjem (ki ga pričnemo na osnovi izkušenje iz točke (a) pri 9 pridemo do $a = 9$.

(g) Za območje zavrnitve določeno v točki (f) določi moč testa, če je v resnici $p = 0{\cdot}4$. Moč testa je $1 - \beta$. Verjetnost β izračunamo enako kot v točkah (c) in (d). Velja $Y \sim \mathrm{B}(20, 0{\cdot}4)$ in

$$\beta = \mathrm{P}(y \leqslant 8) = \sum_{i=0}^{8} \binom{20}{i} 0{\cdot}4^i \; 0{\cdot}6^{20-i} = 0{\cdot}5956 = 59{\cdot}56\%.$$

Moč testa znaša 0·4044.

(h) Za območje zavrnitve določeno v točki (f) določi moč testa, če je v resnici $p = 0{\cdot}7$. Tokrat velja $Y \sim \mathrm{B}(20, 0{\cdot}7)$ in

$$\beta = \mathrm{P}(y \leqslant 8) = \sum_{i=0}^{8} \binom{20}{i} 0{\cdot}7^i \; 0{\cdot}3^{20-i} = 0{\cdot}0051 = 0{\cdot}51\%.$$

Moč testa znaša 0·995. ◇

Primer: Ali je dejanska mediana (τ) pH iz določene regije 6·0? Da bi odgovorili na to vprašanje bomo izbrali 10 vzorcev zemlje iz te regije, da ugotovimo, če empirični vzorci močno podpirajo, da je dejanska mediana manšja ali enaka 6·0?

Predpostavke

- naključni vzorec
  - neodvisen
  - enako porazdeljen (kot celotna populacija),
- vzorčenje iz zvezne porazdelitve,
- verjetnostna porazdelitev ima mediano.

Postavitev statistične domneve

- $H_0 \,:\, \tau = 6{\cdot}0$ (ničelna domneva),
- $H_a \,:\, \tau < 6{\cdot}0$ (alternativna domneva).

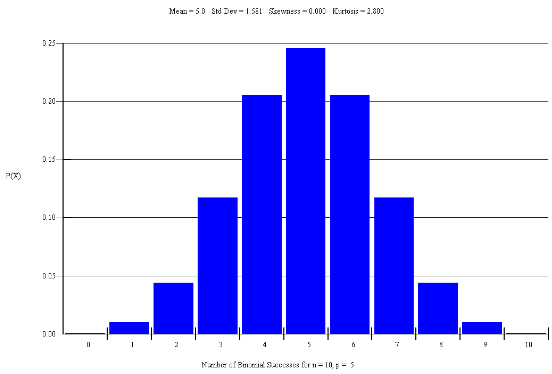
Izbira testne statistike (TS)

- S_+ = število vzorcev, ki so večji od mediane τ_0 iz domneve.
- S_- = število vzorcev, ki so manšji od mediane τ_0 iz domneve.

Porazdelitev testne statistike

- vsak poskus je bodisi uspeh ali neuspeh,
- fiksen vzorec, velikosti n,
- naključni vzorci
  - neodvisni poskusi,
  - konstantna verjetnost uspeha.

Gre za binomsko porazdelitev:
$S_+ \approx B(10, 0{\cdot}5)$,
torej je $E(S_+) = np = 5$.



x	$P(X=x)$	$F(x)$
0	0·000977	0·00098
1	0·009766	0·01074
2	0·043945	0·05469
3	0·117188	0·17188
4	0·205078	0·37695
5	0·246094	0·62305
6	0·205078	0·82813
7	0·117188	0·94531
8	0·043945	0·98926
9	0·009766	0·99902
10	0·000977	1·00000

Določimo zavrnitveni kriterij

- Stopnja značilnosti testa $\alpha = 0{\cdot}01074$,
- Kritična vrednost: $S_+ = 1$,
- Območje zavrnitve: $S_+ \leqslant 1$,

Izberemo naključni vzorec

Predpostavimo, da je dejanska mediana (τ) pH iz določene regije 6·0. Da bi preverili to trditev, smo izbrali 10 vzorcev zemlje iz te regije in jih podvrgli kemični analizi in na ta način določili pH vrednost za vsak vzorec. Ali empirični podatki podpirajo trditev, da je dejanska mediana manšja ali enaka 6·0?

pH	predznak
5·93	−
6·08	+
5·86	−
5·91	−
6·12	+
5·90	−
5·95	−
5·89	−
5·98	−
5·96	−

Izračunajmo P-vrednost iz testne statistike

$S_+ = 2$, P-vrednost $= P(S_+ \geqslant 2 \,|\, \tau = 6{\cdot}0) = 0{\cdot}05469$.

$S_- = 8$, P-vrednost $= P(S_- \geqslant 8 \,|\, \tau = 6{\cdot}0) = 0{\cdot}05469$.

Odločitev in zaključek

- P-vrednost $> \alpha = 0{\cdot}01074$.
- Ni osnove za zavrnitev domneve H_0.

- Zavrni ničelno domnevo.
  - Zaključimo, da empirični podatki sugerirajo, da velja alternativna trditev.
- Ni osnove za zavrnitev ničelne domneve (angl. fail to reject, kratica FTR).
  - Zaključimo, da nimamo dovolj osnov, da bi dokazali, da velja alternativna trditev.
- Privzemimo, da je pH enaka 6·0 v tej konkretni regiji, saj imamo premalo podatkov, da bi pokazali, da je dejanska mediana pH manjša od 6·0. ◇

14.4 Preverjanje predznaka

- test: $H_0 : \tau = \tau_0$ (mediana populacije)
- predpostavke
  - naključno vzorčenje
  - vzorčenje iz zvezne porazdelitve

Preverjanje domneve z enim vzorcem

Testi za mero centralne tendence???

SASA diagram....

14.5 Wilcoxonov predznačen-rang test

- test
  - $H_0 : \tau = \tau_0$ (mediana populacije)
  - $H_0 : \mu = \mu_0$ (povprečje populacije)
- Predpostavke
  - naključni vzorec iz zvezne porazdelitve.
  - porazdelitev populacije ima simetrično obliko.
  - verjetnostna porazdelitev ima povprečje (mediano).

Testna statistika

- S_+ = vsota rangov, ki ustrezajo pozitivnim številom.
- S_- = vsota rangov, ki ustrezajo negativnim številom.

Primer: Naj bo $H_0 : \tau = 500$ in $H_a : \tau > 500$.

Postopek:

- izračunaj odstopanje od τ_0
- razvrsti odstopanja glede na velikost absolutne vrednosti (tj., brez upoštevanja predznaka).

- seštej range, ki ustrezajo bodisi pozitivnemu ali negativnemu predznaku.

meritve	odstopanje	abs. vrednost	rang	+	–
499,2	-0,8	0,8	1		1
498,5	-1,5	1,5	2		2
502,6	2,6	2,6	3	3	
497,3	-2,7	2,7	4		4
496,9	-3,1	3,1	5		5
				S_+= 3	**S_- = 12**

Porazdelitev testne statistike

- 2^n enako verjetnih zaporedij,
- največji rang $= n(n+1)/2$.

S+	1	2	**S–**	p	F	
0	–	–	3	0,25	0,25	
1	+	–	2	0,25	0,5	
2	–	+	1	0,25	0,75	
3	+	+	0	0,25	1	

S+	1	2	3	**S–**	p	F
0	–	–	–	6	0,125	0,125
1	+	–	–	5	0,125	0,25
2	–	+	–	4	0,125	0,375
3	–	–	+	3	0,125	0,5
3	+	+	–	3	0,125	0,625
4	+	–	+	2	0,125	0,75
5	–	+	+	1	0,125	0,875
6	+	+	+	0	0,125	1

S+	1	2	3	4	**S–**	p	F
0	–	–	–	–	10	0,0625	0,0625
1	+	–	–	–	9	0,0625	0,125
2	–	+	–	–	8	0,0625	0,1875
3	+	+	–	–	7	0,0625	0,25
3	–	–	+	–	7	0,0625	0,3125
4	+	–	+	–	6	0,0625	0,375
4	–	–	–	+	6	0,0625	0,4375
5	+	–	–	+	5	0,0625	0,5
5	–	+	+	–	5	0,0625	0,5625
6	–	+	–	+	4	0,0625	0,625
6	+	+	+	–	4	0,0625	0,6875
7	+	+	–	+	3	0,0625	0,75
7	–	–	+	+	3	0,0625	0,8125
8	+	–	+	+	2	0,0625	0,875
9	–	+	+	+	1	0,0625	0,9375
10	+	+	+	+	0	0,0625	1

S+	1	2	3	4	5	**S–**	p	F
0	–	–	–	–	–	15	0,03125	0,03125
1	+	–	–	–	–	14	0,03125	0,0625
2	–	+	–	–	–	13	0,03125	0,09375
3	–	–	+	–	–	12	0,03125	0,125
3	+	+	–	–	–	12	0,03125	0,15625
4	–	–	–	+	–	11	0,03125	0,1875
4	+	–	+	–	–	11	0,03125	0,21875
5	–	–	–	–	+	10	0,03125	0,25
5	+	–	–	+	–	10	0,03125	0,28125
5	–	+	+	–	–	10	0,03125	0,3125
6	+	–	–	–	+	9	0,03125	0,34375
6	–	+	–	+	–	9	0,03125	0,375
6	+	+	+	–	–	9	0,03125	0,40625
7	–	+	–	–	+	8	0,03125	0,40625
7	–	–	+	+	–	8	0,03125	0,4375
7	+	+	–	+	–	8	0,03125	0,46875
8	–	–	+	–	+	7	0,03125	0,5

P-vrednost

- Sprejemljivost domneve H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je domneva H_0 pravilna.
- Najmanjši α pri katerem zavrnemo domnevo H_0
  - Če je P-vrednost $> \alpha$, potem FTR H_0.
  - Če je P-vrednost $< \alpha$, potem zavrni H_0.
  - Če je P-vrednost $= 2\,P(Z > 1{\cdot}278) = 2 \times 0{\cdot}1003 = 0{\cdot}2006$.

Odločitev in zaključek

- odločitev
  - P-vrednost $= P(S_+ = 3)$ ali $P(S_- = 12)$

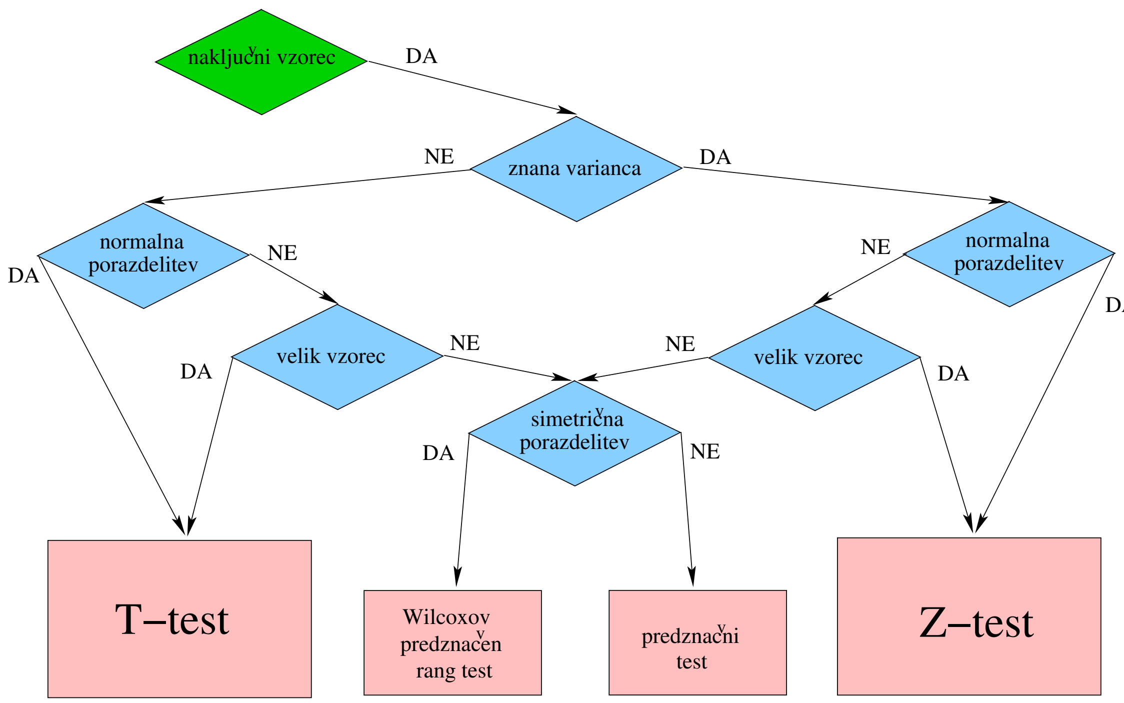
  – P-vrednost $= 0{\cdot}15625 > 0{\cdot}1 = \alpha \implies$ FTR H_0.

- zaključek: privzemimo $\tau = 500$, saj ni osnov, da bi pokazali $\tau > 500$ $\diamond$

14.6 Formalen postopek za preverjanje domnev

1. Postavi domnevi o parametrih (ničelno H_0 in alternativno H_1).
2. Za parameter poiščemo kar se da dobro cenilko (npr. nepristransko) in njeno porazdelitev ali porazdelitev ustrezne statistike (izraz, v katerem nastopa cenilka).
3. Določi odločitveno pravilo. Izberemo stopnjo značilnosti (α). Na osnovi stopnje značilnosti in porazdelitve statistike določimo kritično območje;
4. Zberi/manipuliraj podatke ter na vzorčnih podatkih izračunaj (eksperimentalno) vrednost testne statistike.
5. Primerjaj in naredi zaključek.

- če eksperimentalna vrednost pade v kritično območje, ničelno domnevo zavrni in sprejmi osnovno domnevo ob stopnji značilnosti α.
- če eksperimentalna vrednost ne pade v kritično območje, pa pravimo da vzorčni podatki kažejo na statistično neznačilne razilike med parametrom in vzorčno oceno.



14.7 Preverjanje domneve za povprečje $H_0 : \mu = \mu_0$

Delovne *predpostavke* v tem rezdelku so (ne pozabite navesti vse Vaše predpostavke pri vsakem premeru):

– naključno vzorčenje
– izbiramo vzorce iz normalne porazdelitve in/ali imamo vzorec pri katerem je n velik.

14.7.1 Znan odklon σ

Če poznamo odklon populacije σ, potem na osnovi predpostavk na začetku tega razdelka in diagrama iz razdelka 14.6

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

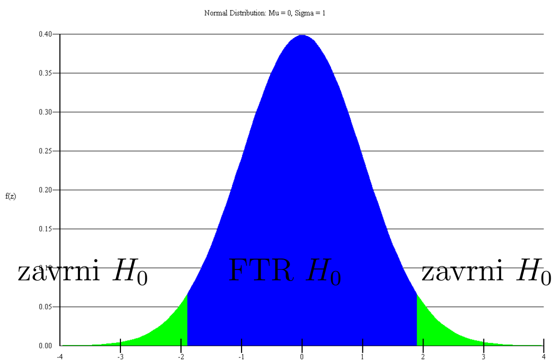
Primer: Proizvajalec omake za špagete da v vsako posodo 28 unče omake za špagete. Količina omake, ki je v vsaki posodi, je porazdeljena normalno s standardnim odklonom $0,005$ unče. Podjetje ustavi proizvodni trak in popravi napravo za polnenje, če so posode bodisi premalo napolnjene (to razjezi kupce), ali preveč napolnjene (kar seveda pomeni manjši profit). **Ali naj na osnovi vzorca iz 15ih posod ustavijo proizvodno linijo?** Uporabi stopnjo značilnosti 0·05. Postavimo domnevo

- $H_0 \,:\, \mu = 28$ (ničelna domneva),
- $H_a \,:\, \mu \neq 28$ (alternativna domneva).

Dodatna predpostavka: poznamo varianco populacije.

Izberemo testno statistiko: Z-Test: $H_0 \,:\, \mu = \mu_0$ (povprečje populacije)

Določimo zavrnitveni kriterij



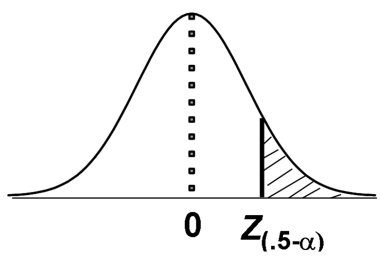
Rezultati testiranja

- vzami naključni vzorec – zanj dobimo vzorčno povprečje: 28·0165
- izračunaj vrednost testne statistike: $Z = (28{\cdot}0165 - 28)/0{\cdot}0129 = 1{\cdot}278$.
- naredi odločitev: FTR H_0.
- zaključek: privzemi $\mu = 28$

P-vrednost

- Sprejemljivost domneve H_0 na osnovi vzorca (možnost za opazovanje vzorca ali bolj ekstremno podatkov, če je domneva H_0 pravilna):
  – P-vrednost $= 2\,P(Z > 1{\cdot}278) = 2 \times 0{\cdot}1003 = 0{\cdot}2006$.
- Najmanjši α pri katerem zavrnemo domnevo H_0
  – P-vrednost $> \alpha$, zato FTR H_0.

Za $H_a \,:\, \mu > \mu_0$ je **odločitveno pravilo**: zavrni H_0, če je **T.S.** $\geqslant z_{(0\cdot5-\alpha)}$



Za $H_a : \mu < \mu_0$ **odločitveno pravilo**: zavrni H_0, če je **T.S.** $\leqslant \boldsymbol{z_{(0{\cdot}5-\alpha)}}$.
Za $H_a : \mu \neq \mu_0$ **odločitveno pravilo**: zavrni H_0 **če je T.S.** $\leqslant -\boldsymbol{z_{(0{\cdot}5-\alpha)}}$ ali **če je T.S.** $\geqslant \boldsymbol{z_{(0{\cdot}5-\alpha)}}$. ◇

14.7.2 Neznan odklon σ in velik vzorec

Če ne poznamo odklona σ in je $n \geqslant 30$, potem

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{s}\sqrt{n} \quad \textbf{sledi } t\textbf{-porazdelitev z } n-1 \textbf{ prostostnimi stopnjami.}$$

(Velja omeniti še, da se pri tako velikem n z- in t-porazdelitev tako ne razlikujeta kaj dosti.)

14.7.3 Neznan odklon σ, normalna populacija in majhen vzorec

Če ne poznamo odklona σ, populacija je normalna in je $n < 30$, potem

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{\dfrac{s}{\sqrt{n}}} \quad \textbf{sledi } t\textbf{-porazdelitev z } n-1 \textbf{ prostostnimi stopnjami.}$$

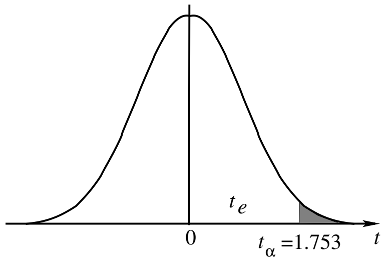
Primer: Za slučajni vzorec: 16-ih odraslih Slovencev smo izračunali povprečno število in variance priznanih let šolanja: $\overline{X} = 9$ in $s^2 = 9$. Predpostavljamo, da se spremenljivka na populaciji porazdeljuje normalno. **Ali lahko sprejmemo domnevo, da imajo odrasli Slovenci v povprečju več kot osemletko pri 5% stopnji značilnosti?** Postavimo najprej ničelno in osnovno domnevo:

$$H_0 : \mu = 8 \qquad \text{in} \qquad H_1 : \mu > 8.$$

Ustrezna testna statistika

$$t = \frac{\overline{X} - \mu_H}{s}\sqrt{n}$$

je porazdeljena po t-porazdelitvi s 15imi prostostnimi stopnjami. Ker gre za enostranski test, je glede na osnovno domnevo kritično območje na desni strani porazdelitve in kritična vrednost $t_{0{\cdot}05}(15) = 1{\cdot}753$. Izračunajmo eksperimentalno vrednost statistike:



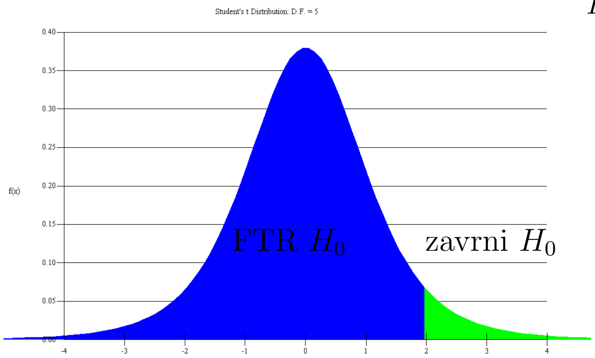
$$t_e = \frac{9-8}{3}\sqrt{16} = 1{\cdot}3.$$

Eksperimentalna vrednost ne pade v kritično območje. Zato ničelne domneve ne moremo zavrniti in sprejeti osnovne domneve, da imajo odrasli Slovenci več kot osemletko. ◇

Primer: Ravnatelj bežigrajske gimnazije trdi, da imajo najboljši PT program v Sloveniji s povprečjem APFT 240. Predpostavi, da je porazdelitev rezultatov testov približno normalna.

Uporabi $\alpha = 0{\cdot}05$ za določitev **ali je povprečje APFT rezultatov šestih naključno izbranih dijakov iz bežigrajske gimnazije statistično večje od 240**. *Dodatna predpostavka:* ne poznamo varianco populacije. Postavimo domnevi: $H_0 : \mu = 240$ in $H_a : \mu > 240$ in zaradi dodatne predpostavke izberemo T-test za testno statistiko.

Določimo zavrnitveni kriterij



Rezultati testiranja

- naredi naključni vzorec:
  - vzorčno povprečje: $255{\cdot}4$
  - vzorčni standardni odklon: $40{\cdot}07$
- izračunaj vrednost testne statistike: $T = (255{\cdot}4 - 240)/16{\cdot}36 = 0{\cdot}9413$.
- sprejmi odločitev: FTR H_0

Zaključek: Bežigrajska gimnazija ne more pokazati, da imajo višje povprečje APFT rezultatov, kot slovensko povprečje.

P-vrednost

- Sprejemljivost domneve H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je domneva H_0 pravilna
  - P-vrednost $= P(T > 0{\cdot}9413) = 0{\cdot}1949$.
- Najmanjši α pri katerem zavrnemo domnevo H_0
  - P-vrednost $> \alpha$, zato FTR H_0 .

Vstavimo podatke v Minitab:

`C1` : 2610, 2750, 2420, 2510, 2540, 2490, 2680.

T-test povprečja

```
Test of mu = 2500.0 vs mu > 2500.0
   N      MEAN   STDEV    SE MEAN      T     p-VALUE
C1 7    2571.4   115.1       43.5   1.64       0.076
```

Razlaga P-vrednosti

1. Izberi največjo vrednost za α, ki smo jo pripravljeni tolerirati.
2. Če je P-vrednost testa manjša kot maksimalna vrednost parametra α, potem zavrni ničelno domnevo.

$\diamond$

14.8 Preverjanje domneve za razliko povprečij $H_0 : \mu_1 - \mu_2 = D_0$

14.8.1 Znana odklona σ_1 in σ_2

Vzorce jemljemo neodvisno, zato

$$\text{T.S.} = \Big((\overline{y}_1 - \overline{y}_2) - D_0\Big)\Big/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

Primer: Preveriti želimo domnevo, da so dekleta na izpitu boljša od fantov. To domnevo preverimo tako, da izberemo slučajni vzorec 36 deklet in slučajni vzorec 36 fantov, za katere imamo izpitne rezultate, na katerih izračunamo naslednje statistične karakteristike:

$$\overline{X}_F = 7{\cdot}0, \quad s_F = 1 \quad \text{in} \quad \overline{X}_D = 7{\cdot}2, \quad s_D = 1$$

Domnevo preverimo pri 5% stopnji značilnosti. Postavimo ničelno in osnovno domnevo:

$$H_0 : \mu_D = \mu_F \quad \text{oziroma} \quad \mu_D - \mu_F = 0,$$

$$H_1 : \mu_D > \mu_F \quad \text{oziroma} \quad \mu_D - \mu_F > 0.$$

Za popularijsko razliko aritmetičnih sredin na vzorcih računamo vzorčno razliko aritmetičnih sredin, ki se za dovolj velike vzorce porazdeljuje normalno

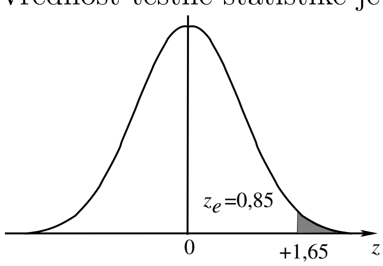
$$\overline{X}_D - \overline{X}_F : N\left(\mu_D - \mu_F, \sqrt{\frac{s_D^2}{n_D} + \frac{s_F^2}{n_F}}\right).$$

oziroma statistika

$$z = \Big(\overline{X}_D - \overline{X}_F - (\mu_D - \mu_F)_H\Big)\Big/\sqrt{\frac{s_D^2}{n_D} + \frac{s_F^2}{n_F}}$$

standardizirano normalno $N(0,1)$. Osnovna domneva kaže enostranski test: možnost napake 1. vrste je le na desni strani normalne porazdelitve, kjer zavračamo ničelno domnevo. Zato je kritično območje določeno z vrednostmi večjimi od $1{\cdot}65$. Vrednost testne statistike je

$$z_e = (7{\cdot}2 - 7 - 0)\Big/\sqrt{\frac{1}{36} + \frac{1}{36}} = 0{\cdot}852.$$



Torej ne pade v kritično območje. Ničelne domneve ne moremo zavrniti. Povprečna uspešnost deklet in fantov ni statistično značilno različna.

◇

14.8.2 Neznana odklona σ_1 in/ali σ_2, $n_1 \geqslant 30$ in/ali $n_2 \geqslant 30$

Ker vzorce jemljemo neodvisno, velja

$$\text{T.S.} = ((\overline{y}_1 - \overline{y}_2) - D_0)\Big/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

14.8.3 Neznana σ_1 in/ali σ_2, norm. pop., $\sigma_1 = \sigma_2$, $n_1 < 30$ ali $n_2 < 30$

Ker vzorce jemljemo neodvisno, velja:

$$\text{T.S.} = \frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

sledi t-porazdelitev z $n_1 + n_2 - 2$ prostostnimi stopnjami. Privzeli smo:

1. Populaciji iz katerih jemljemo vzorce imata obe približno **normalno** relativno porazdelitev frekvenc.
2. Varianci obeh populacij sta **enaki**.
3. Naključni vzorci so izbrani **neodvisno** iz obeh populacij.

14.8.4 Neznana σ_1 in/ali σ_2, norm. pop., $\sigma_1 \neq \sigma_2$, $n_1 < 30$ ali $n_2 < 30$

Ker jemljemo vzorce neodvisno, velja

$$\text{T.S.} = \frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

sledi t-porazdelitev z ν prostostnimi stopnjami, kjer je

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1-1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

(Če ν ni naravno število, zaokroži ν navzdol do najbližjega naravnega števila za uporabo t-tabele.)

14.9 Preverjanje domneve za povprečje $H_0 : \mu_d = D_0$

14.9.1 Velik vzorec

Ker vzorce jemljemo neodvisno velja

$$\text{T.S.} = \frac{(\overline{d} - D_0)}{s_d}\sqrt{n} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

14.9.2 Normalna populacija razlik in majhen vzorec

Če je populacija razlik normalno porazdeljena in če je $n \leqslant 30$, potem velja

$$\text{T.S.} = \frac{\overline{d} - D_0}{s_d}\sqrt{n} \quad \textbf{sledi } t\textbf{-porazdelitev z } n-1 \textbf{ prostostnimi stopnjami.}$$

naloga	clovek. urnik	avtomatizirana metoda	razlika
1	185.4	180.4	5.0
2	146.3	248.5	-102.2
3	174.4	185.5	-11.1
4	184.9	216.4	-31.5
5	240.0	269.3	-29.3
6	253.8	249.6	-4.2
7	238.8	282.0	-43.2
8	263.5	315.9	-52.4

Vstavimo podatke v Minitab

C1: 185·4 146·3 174·4 184·9 240·0 253·8 238·8 263·5

C2: 180·4 248·5 185·5 216·4 269·3 249·6 282·0 315·9

Test za parjenje in interval zaupanja. Parjenj T za C1-C2:

	N	povpr.	StDev	SE povpr.
C1	8	210.9	43.2	15.3
C2	8	243.4	47.1	16.7
Razlika	8	032.6	35.0	12.4

95% interval zaupanja za razliko povprečja: $(-61{\cdot}9, -3{\cdot}3)$.

T-test za razliko povpr. $= 0$ (proti $\neq 0$): T-vrednost$=-2{\cdot}63$, P-vrednost$=0{\cdot}034$.

14.10 Preverjanje domneve za delež $H_0 : p = p_0$

14.10.1 Velik vzorec

$$\text{T.S.} = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

Kot splošno pravilo bomo zahtevali, da velja $\quad n\hat{p} \geqslant 4 \quad$ in $\quad n\hat{q} \geqslant 4$.

Primer: Postavimo domnevo o vrednosti parametra, npr. π – delež enot z določeno lastnostjo na populaciji. Denimo, da je domneva $H : \pi_H = 0{\cdot}36$. Tvorimo slučajne vzorce npr. velikosti $n = 900$ in na vsakem vzorcu določimo vzorčni delež p (delež enot z določeno lastnostjo na vzorcu). Ob predpostavki, da je domneva pravilna, vemo, da se vzorčni deleži porazdeljujejo približno normalno

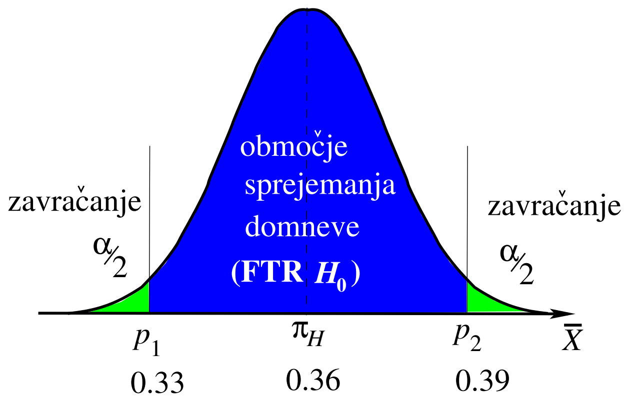
$$N\left(\pi_H, \sqrt{\frac{\pi_H(1-\pi_H)}{n}}\right).$$

Vzemimo en slučajni vzorec z vzorčnim deležem p. Ta se lahko bolj ali manj razlikuje od π_H. Če se zelo razlikuje, lahko podvomimo o resničnosti domneve π_H. Zato okoli π_H naredimo območje sprejemanja domneve in izven tega območja območje zavračanja domneve. Denimo, da je območje zavračanja določeno s 5% vzorcev, ki imajo ekstremne vrednosti deležev (2·5% levo in 2·5% desno). Deleža, ki ločita območje sprejemanja od območja zavračanja lahko izračunamo takole:

$$p_{1,2} = \pi_H \pm z_{\alpha/2}\sqrt{\frac{\pi_H(1-\pi_H)}{n}},$$

$$p_{1,2} = 0{\cdot}36 \pm 1{\cdot}96\sqrt{\frac{0{\cdot}36 \times 0{\cdot}64}{900}}$$

$$= 0{\cdot}36 \pm 0{\cdot}03.$$



Kot smo že omenili, je sprejemanje ali zavračanje domnev po opisanem postopku lahko napačno v dveh smislih:

Napaka 1. vrste (α): Če vzorčna vrednost deleža pade v območje zavračanja, domnevo π_H zavrnemo. Pri tem pa vemo, da ob resnični domnevi π_H obstajajo vzorci, ki imajo vrednosti v območju zavračanja. Število α je verjetnost, da vzorčna vrednost pade v območje zavračanja ob prodpostavki, da je domneva resnična. Zato je α verjetnost, da zavrnemo pravilno domnevo – **napaka 1. vrste**. Ta napaka je merljiva in jo lahko poljubno manjšamo.

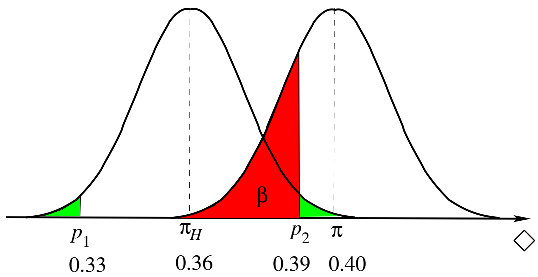
Napaka 2. vrste (β): Vzorčna vrednost lahko pade v območje sprejemanja, čeprav je domnevna vrednost parametra napačna. V primeru, ki ga obravnavamo, naj bo prava vrednost deleža na populaciji $\pi = 0{\cdot}40$. Tedaj je porazdelitev vzorčnih deležev

$$N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right) = N(0{\cdot}40, 0{\cdot}0163).$$

Ker je območje sprejemanja, domneve v intervalu $0{\cdot}33 < p < 0{\cdot}39$, lahko izračunamo verjetnost, da bomo sprejeli napačno domnevo takole:

$$\beta = P(0{\cdot}33 < p < 0{\cdot}39) = 0{\cdot}27$$

Napako 2. vrste lahko izračunamo le, če imamo znano resnično vrednost parametra π. Ker ga ponavadi ne poznamo, tudi ne poznamo napake 2. vrste. Zato ne moremo sprejemati domnev.



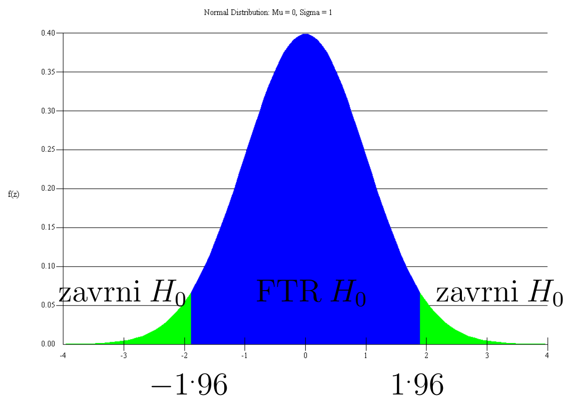
Primer: Državni zapisi indicirajo, da je od vseh vozil, ki gredo skozi testiranje izpušnih plinov v preteklem letu, 70% uspešno opravilo testiranje v prvem poskusu. Naključni vzorec 200ih avtomobilov testiranih v določeni pokrajni v tekočem letu je pokazalo, da jih je 156 šlo čez prvi test. Ali to nakazuje, da je dejanski delež populacije za to pokrajno v tekočem letu različno od preteklega državnega deleža? Pri testiranju domneve uporabi $\alpha = 0{\cdot}05$. $\diamond$

14.10.2 Majhen vzorec

- Ničelna domneva $H_0 \,:\, p = 0{\cdot}7$
- Alternativna domneva $H_a \,:\, p \neq 0{\cdot}7$
- Test: $H_0 \,:\, p = p_0$ (delež populacije)
- Predpostavke: naključni vzorec in izbiranje vzorca iz binomske porazdelitve
- T.S.: $\quad Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}},$ $\quad$ upoštevamo pa tudi $\quad n\hat{p} \geqslant 4 \quad$ in $\quad n\hat{q} \geqslant 4.$

Primer:

Določimo zavrnitveni kriterij



Rezultati testiranja

- Naredi naključni vzorec: dobimo delež vzorca: $156/200 = 0{\cdot}78$.
- Izračunaj vrednost testne statistike: $Z = (0{\cdot}78 - 0{\cdot}7)/0{\cdot}0324 = 2{\cdot}4688$.
- Naredi odločitev: zavrni domnevo H_0
- Zaključek: pokrajna ima drugačen kriterij.

P-vrednost

- Sprejemljivost domneve H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je domneva H_0 pravilna
  - P-vrednost $= 2\,P(Z > 2{\cdot}469) = 2 \times 0{\cdot}0068 = 0{\cdot}0136$
- Najmanjši α pri katerem zavrnemo domnevo H_0
  - P-vrednost $< \alpha$, zato zavrni domnevo H_0 $\diamond$

14.11 Razlika deležev dveh populaciji $H_0 : p_1 - p_2 = D_0$

Če želimo, da je vzorec za testiranje domneve o $p_1 - p_2$ velik,

bomo kot splošno pravilo zahtevali

$$n_1\hat{p}_1 \geqslant 4, \qquad n_1\hat{q}_1 \geqslant 4,$$

$$n_2\hat{p}_2 \geqslant 4 \quad \text{in} \quad n_2\hat{q}_2 \geqslant 4.$$

14.11.1 Velik vzorec in $D_0 = 0$

$$\text{T.S.} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \textbf{sledi } z\textbf{-porazdelitev, kjer je } \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}.$$

Primer: Želimo preveriti, ali je predsedniški kandidat različno priljubljen med mestnimi in vaškimi prebivalci. Zato smo slučajni vzorec mestnih prebivalcev povprašali, ali bi glasovali za predsedniškega kandidata. Od 300 vprašanih (n_1) jih je 90 glasovalo za kandidata (k_1). Od 200 slučajno izbranih vaških prebivalcev (n_2) pa je za kandidata glasovalo 50 prebivalcev (k_2). Domnevo, da je kandidat različno priljubljen v teh dveh območjih preverimo pri 10% stopnji značinosti.

$$H_0 \,:\, \pi_1 = \pi_2 \quad \text{oziroma} \quad \pi_1 - \pi_2 = 0,$$

$$H_1 \,:\, \pi_1 \neq \pi_2 \quad \text{oziroma} \quad \pi_1 - \pi_2 \neq 0.$$

Vemo, da se razlika vzorčnih deležev porazdeljuje približno normalno:

$$p_1 - p_2 \,:\, N\left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}\right).$$

Seveda π_1 in π_2 nista znana. Ob predpostavki, da je ničelna domneva pravilna, je matematično upanje razlike vzorčnih deležev hipotetična vrednost razlike deležev, ki je v našem primeru enaka 0. Problem pa je, kako oceniti standardni odklon. Ker velja domneva $\pi_1 = \pi_2 = \pi$, je disperzija razlike vzorčnih deležev

$$\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} = \frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2} = \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Populacijski delež π ocenimo z uteženim povprečjem vzorčnih deležev p_1 in p_2

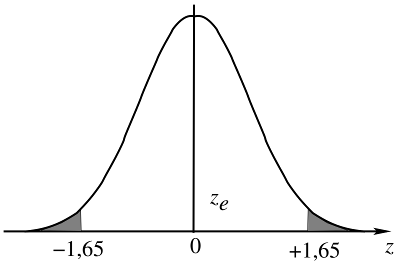
$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}.$$

Vrnimo se na primer. Vzorčna deleža sta: $p_1 = 90/300 = 0{\cdot}30$ in $p_2 = 50/200 = 0{\cdot}25$. Ocena populacijskega deleža je $p = (50 + 90)/(200 + 300) = 0{\cdot}28$. Kot smo že omenili, se testna statistika

$$z = \Big(p_1 - p_2 - (\pi_1 - \pi_2)_H\Big) \Big/ \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

porazdeljuje približno standardizirano normalno $N(0,1)$. Ker gre za dvostranski test, sta kritični vrednosti $\pm z_{\alpha/2} = \pm 1{\cdot}65$. Eksperimentalna vrednost statistike pa je

$z_e = (0{\cdot}30 - 0{\cdot}025 - 0) \Big/ \sqrt{0{\cdot}28(1 - 0{\cdot}28)\left(\frac{1}{300} + \frac{1}{200}\right)} = 1{\cdot}22.$



Eksperimentalna vrednost ne pade v kritično območje. Zato ničelne domneve ne moremo zavrniti. Priljubljenost predsedniškega kandidata ni statistično značilno različna med mestnimi in vaškimi prebivalci. ◇

14.11.2 Velik vzorec in $D_0 \neq 0$.

$$\text{T.S.} = \Big((\hat{p}_1 - \hat{p}_2) - D_0\Big) \Big/ \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \qquad \textbf{sledi } z\textbf{-porazdelitev.}$$

Primer: Neka tovarna cigaret proizvaja dve znamki cigaret. Ugotovljeno je, da ima 56 od 200 kadilcev raje znamko A in da ima 29 od 150 kadilcev raje znamko B. Preveri domnevo pri $0,06$ stopnji zaupanja, da bo prodaja znamke A boljša od prodaje znamke B za 10% proti alternativni domnevi, da bo razlika manj kot 10%. ◇

14.12 Analiza variance

Če opravljamo isti poskus v nespremenjenih pogojih, kljub temu v rezultatu poskusa opažamo spremembe (variacije) ali odstopanja. Ker vzrokov ne poznamo in jih ne moremo kontrolirati, spremembe pripisujemo *slučajnim vplivom* in jih imenujemo *slučajna odstopanja*. Če pa enega ali več pogojev v poskusu spreminjamo, seveda dobimo dodatna odstopanja od povprečja. Analiza tega, ali so odstopanja zaradi sprememb različnih faktorjev ali pa zgolj slučajna, in kateri faktorji vplivajo na variacijo, se imenuje *analiza variance*.

Zgleda:

(a) Namesto dveh zdravil proti nespečnosti kot v Studentovem primeru lahko preskušamo učinkovitost več različnih zdravil A, B, C, D,... in s preskušanjem domneve $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ raziskujemo, ali katero od zdravil sploh vpliva na rezultat. Torej je to posplošitev testa za $H_0 : \mu_1 = \mu_2$

(b) Raziskujemo hektarski donos pšenice. Nanj vplivajo različni faktorji: različne sorte pšenice, različni načini gnojenja, obdelave zemlje itd., nadalje klima, čas sejanja itd.

Analiza variance je nastala prav v zvezi z raziskovanjem v kmetijstvu. Glede na število faktorjev, ki jih spreminjamo, ločimo t.i. *enojno klasifikacijo* ali *enofaktorski eksperiment*, *dvojno klasifikacijo* ali *dvofaktorski eksperiment*, itd. V izrazu

$$Q_v^2 = \sum_{i=1}^{r} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_i)^2 = \sum_{i=1}^{r} (n_i - 1) S_i^2 \quad \text{je} \quad S_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_i)^2$$

nepristranska cenilka za disperzijo v i-ti skupini; neodvisna od S_j^2, za $i \neq j$. Zato ima

$$\frac{Q_v^2}{\sigma^2} = \sum_{i=1}^{r} (n_i - 1) \frac{S_i^2}{\sigma^2}$$

porazdelitev $\chi^2(n-r)$, saj je ravno $\sum_{i=1}^{r}(n_i - 1) = n - r$ prostostnih stopenj. Ker je $\mathsf{E}\frac{Q_v^2}{\sigma^2} = n - r$, je tudi $S_v^2 = \frac{1}{n-r} Q_v^2$ nepristranska cenilka za σ^2. Izračunajmo še Q_m^2 pri predpostavki o veljavnosti osnovne domneve H_0. Dobimo

$$Q_m^2 = \sum_{i=1}^{r} n_i (\overline{X}_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Torej je

$$\frac{Q_m^2}{\sigma^2} = \sum_{i=1}^{r} n_i \Big(\frac{\overline{X}_i - \mu}{\sigma/\sqrt{n_i}} \Big)^2 - n \Big(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \Big)^2,$$

od tu pa spremidimo, da je testna statistika $\boldsymbol{Q_m^2/\sigma^2}$ porazdeljena po $\chi^2(r-1)$. Poleg tega je $S_m^2 = Q_m^2/(r-1)$ nepristranska cenilka za σ^2, neodvisna od S_v^2. Ker sta obe cenilki za varianco σ^2, pri domnevi H_0, njuno razmerje $F = S_m^2/S_v^2$ ne more biti zelo veliko.

Iz

$$F = \frac{S_m^2}{S_v^2} = \frac{Q_m^2/(r-1)}{Q_v^2/(n-r)} = \frac{\frac{Q_m^2}{\sigma^2}/(r-1)}{\frac{Q_v^2}{\sigma^2}/(n-r)}$$

vidimo, da gre za Fisherjevo (Snedecorjevo) porazdelitev $F(r-1, n-r)$, glej *tabelo analize variance*.

VV	VK	PS	PK	F
faktor	Q_m^2	$r-1$	S_m^2	F
slučaj	Q_v^2	$n-r$	S_v^2	
	Q^2	$n-1$		

Analiza variance v R-ju

Zgled: Petnajst enako velikih njiv je bilo posejanih z isto vrsto pšenice, vendar gnojeno na tri različne načine – z vsakim po pet njiv.

```
> ena <- c(47,47,40,32,40)
> dva <- c(76,68,71,46,54)
> tri <- c(49,40,34,36,44)
> d <- stack(list(e=ena,d=dva,t=tri))
> names(d)
[1] "values" "ind"
> oneway.test(values ~ ind, data=d, var.equal=TRUE)

        One-way analysis of means

data:  values and ind
F = 10.5092, num df = 2, denom df = 12, p-value = 0.002304
> av <- aov(values ~ ind, data=d)
> summary(av)
            Df  Sum Sq Mean Sq F value   Pr(>F)
ind          2 1628.93  814.47  10.509 0.002304 **
Residuals   12  930.00   77.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Domnevo H_0 zavrnemo.

14.12.1 Preverjanje domneve o varianci $\sigma^2 = \sigma_0^2$

$$\text{T.S.} = \frac{(n-1)s^2}{\sigma_0^2} \quad \textbf{sledi} \quad \chi^2\textbf{–porazd.}$$

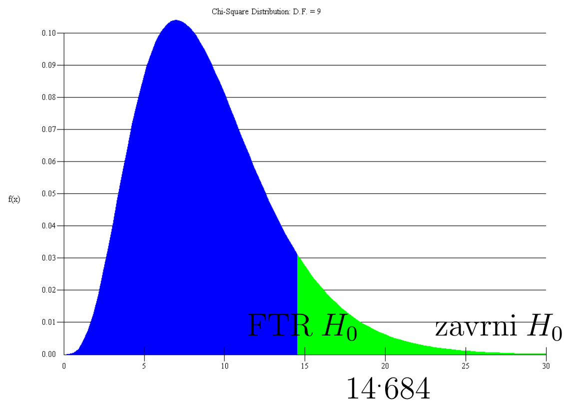
Če je

- $H_a : \sigma^2 > \sigma_0^2$, potem je **odločitveno pravilo**:
  zavrni ničelno domnevo, če je testna statistika večja ali enaka $\chi^2_{(\alpha,n-1)}$.
- $H_a : \sigma^2 < \sigma_0^2$ potem je **odločitveno pravilo**:
  zavrni ničelno domnevo, če je testna statistika manjša ali enaka $\chi^2_{(\alpha,n-1)}$.
- $H_a : \sigma^2 \neq \sigma_0^2$, potem je **odločitveno pravilo**:
  zavrni ničelno domnevo, če je testna statistika manjša ali enaka $\chi^2_{(\alpha,n-1)}$
  ali če je testna statistika večja ali enaka $\chi^2_{(\alpha,n-1)}$.

Primer: Količina pijače, ki jo naprava za mrzle napitke zavrže je normalno porazdeljena s povprečjem 12 unčev in standardnim odklonom 0,1 unče. Vsakič, ko servisirajo napravo, si izberejo 10 vzorcev in izmerijo zavrženo tekočino. Če je razpršenost zavržene količine prevelika, potem mora naprava na servis. Ali naj jo odpeljejo na servis? Uporabi $\alpha = 0{\cdot}1$.

- Ničelna domneva $H_0 : \sigma^2 = 0{\cdot}01$,
- Alternativna domneva $H_a : \sigma^2 > 0{\cdot}01$,
- Predpostavke: naključni vzorec in vzorčenje iz normalne porazdelitve.
- Testna statistika $\chi^2_{\nu=n-1} = S^2(n-1)/\sigma_0^2$.

Določimo zavrnitveni kriterij



Rezultati testiranja

- naredi naključni vzorec izračunamo naslednjo varianco vzorca: $0{\cdot}02041$,
- izračunaj vrednost testne statistike $\chi^2 = (0{\cdot}02041)(9)/(0{\cdot}01) = 18{\cdot}369$,
- naredi odločitev: zavrni H_0,
- zaključek popravi napravo.

P-vrednost

- Sprejemljivost domneve H_0 na osnovi vzorca
  – možnost za opazovanje vzorca/podatkov, če je domneva H_0 pravilna
  – P-vrednost $= P(\chi^2 > 18{\cdot}369) = 0{\cdot}0311$
- Najmanjši α pri katerem zavrnemo domnevo H_0
  – P-vrednost $< \alpha$, zato zavrni domnevo H_0.

$\diamond$

14.12.2 Preverjanje domneve o kvocientu varianc $\sigma_1^2/\sigma_2^2 = 1$

Če velja $H_a : \sigma_1^2/\sigma_2^2 > 1$, potem je **testna statistika** enak s_1^2/s_2^2, **odločitveno pravilo** pa je: zavrni ničelno domnevo, če velja T.S. $\geqslant F_{\alpha,n_1-1,n_2-1}$.

Če velja $H_a : \sigma_1^2/\sigma_2^2 < 1$, potem je **testna statistika** enaka

$$\frac{\text{varianca večjega vzorca}}{\text{varianca manjšega vzorca}},$$

odločitveno pravilo pa je: zavrni ničelno domnevo, če velja $s_1^2 > s_2^2$ in T.S. $\geqslant F_{\alpha,n_1-1,n_2-1}$ oziroma zavrni ničelno domnevo, če velja $s_1^2 < s_2^2$ in T.S. $\geqslant F_{\alpha,n_2-1,n_1-1}$.

14.13 Preverjanje domnev o porazdelitvi spremenljivke

Do sedaj smo ocenjevali in preverjali domnevo o parametrih populacije kot μ, σ in π. Sedaj pa bomo preverjali, če se spremenljivka porazdeljuje po določeni porazdelitvi. Test je zasnovan na dejstvu, kako dobro se prilegajo empirične (eksperimentalne) frekvence vrednosti spremenljivke hipotetičnim (teoretičnim) frekvencam, ki so določene s predpostavljeno porazdelitvijo.

14.13.1 Preverjanje domneve o enakomerni porazdelitvi

Za primer vzemimo met kocke in za spremenljivko število pik pri metu kocke. Preverimo domnevo, da je kocka poštena, kar je enakovredno domnevi, da je porazdelitev spremenljivke enakomerna. Tedaj sta ničelna in osnovna domneva

H_0 : spremenljivka se porazdeljuje enakomerno,
H_1 : spremenljivka se ne porazdeljuje enakomerno.

Denimo, da smo 120-krat vrgli kocko ($n = 120$) in štejemo kolikokrat smo vrgli posamezno število pik. To so empirične ali opazovane frekvence, ki jih označimo s f_i. Teoretično, če je kocka poštena, pričakujemo, da bomo dobili vsako vrednost z verjetnostjo 1/6 oziroma 20 krat. To so teoretične ali pričakovane frekvence, ki jih označimo s f_i'. Podatke zapišimo v naslednji tabeli

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6
f_i'	20	20	20	20	20	20
f_i	20	22	17	18	19	24

S primerjavo empiričnih frekvenc z ustreznimi teoretičnim frekvencami se moramo odločiti, če so razlike posledica le vzorčnih učinkov in je kocka poštena ali pa so razlike prevelike, kar kaže, da je kocka nepoštena.

Testna statistika, ki meri prilagojenost empiričnih frekvenc teoretičnim je

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

ki se porazdeljuje po χ^2 porazdelitvi z $m = k - 1$ prostostnimi stopnjami, ki so enake številu vrednosti spremenijivke ali celic (k) minus število količin dobljenih iz podatkov, ki so uporabljene za izračun teoretičnih frekvenc.

V našem primeru smo uporabili le eno količino in sicer skupno število metov kocke ($n = 120$). Torej število prostostnih stopenj je $m = k-1 = 6-1 = 5$. Ničelna in osnovna domneva sta tedaj

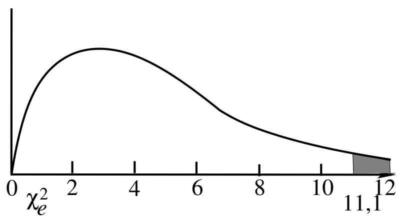
$$H_0 \,:\, \chi^2 = 0 \qquad \text{in} \qquad H_1 \,:\, \chi^2 > 0.$$

Doomnevo preverimo pri stopnji značilnosti $\alpha = 5\%$. Ker gre za enostranski test, je kritična vrednost enaka

$$\chi^2_{1-\alpha}(k-1) = \chi^2_{0{\cdot}95}(5) = 11,1.$$

Eksperimentalna vrednost statistike pa je

$$\chi_e^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20}$$
$$= \frac{4+9+4+1+16}{20} = \frac{34}{20} = 1{\cdot}7.$$



Ker eksperimentalna vrednost statistike ne pade v kritično območje, ničelne domneve ne moremo zavrniti. Empirične in teoretične frekvence niso statistično značilno različne med seboj.

14.13.2 Preverjanje domneve o normalni porazdelitvi

Omenjeni test najpogosteje uporabljamo za preverjanje ali se spremenjivka porazdeljuje normalno. V tem primeru je izračun teoretičnih frekvenc potrebno vložiti malo več truda.

Primer: Preverimo domnevo, da se spremenljivka telesna višina porazdeljuje normalno $N(177, 10)$. Domnevo preverimo pri 5% stopnji značilnosti. Podatki za 100 slučajno izbranih oseb so urejeni v frekvenčni porazdelitvi takole:

	f_i
nad 150-160	2
nad 160-170	20
nad 170-180	40
nad 180-190	30
nad 190-200	8
	100

Ničelna in osnovna domneva sta tedaj

$$H_0 \,:\, \chi^2 = 0 \qquad \text{in} \qquad H_1 \,:\, \chi^2 \neq 0.$$

Za test uporabimo statistiko

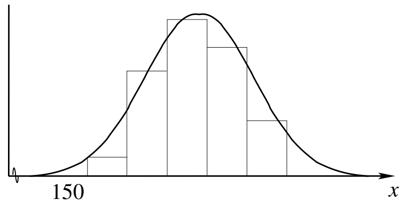
$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

ki se porazdeljuje po χ^2 porazdelitvi z $m = 5 - 1$ prostostnimi stopnjami. Kritična vrednost je

$$\chi^2_{0{\cdot}95}(4) = 9{\cdot}49.$$

V naslednjem koraku je potrebno izračunati teoretične frekvence.

Najprej je potrebno za vsak razred izračunati verjetnost p_i, da spremenljivka zavzame vrednosti določenega intervala, če se porazdeljuje normalno (glej sliko).



Tako je na primer verjetnost, da je višina med 150 in 160 cm:

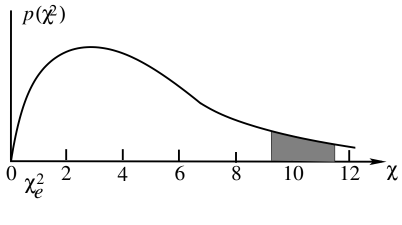
$$\begin{aligned} P(150 < X < 160) &= P\left(\frac{150-177}{10} < Z < \frac{160-177}{10}\right) \\ &= P(-2{\cdot}7 < Z < -1{\cdot}7) = H(2{\cdot}7) - H(1{\cdot}7) = 0{\cdot}4965 - 0{\cdot}4554 \\ &= 0{\cdot}0411. \end{aligned}$$

Podobno lahko izračunamo ostale verjetnosti. Teoretične frekvence so $f_i' = n \times p_i$. Izračunane verjetnosti p_i in teoretične frekvence f_i' postavimo v tabelo.

	f_i	p_i	f_i'
nad 150-160	2	0·0411	4·11
nad 160-170	20	0·1974	19·74
nad 170-180	40	0·3759	37·59
nad 180-190	30	0·2853	28·53
nad 190-200	8	0·0861	8·61
	100		98·58

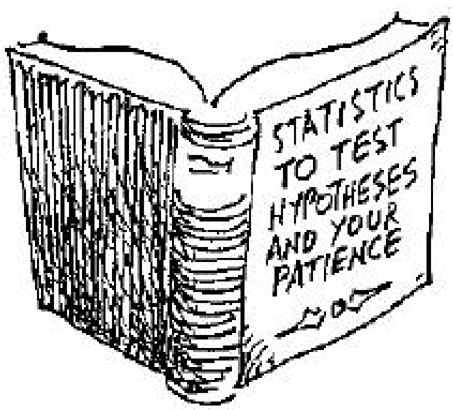
Eksperimentalna vrednost statistike je tedaj

$$\chi_e^2 = \frac{(2-4.11)^2}{4{\cdot}11} + \frac{(20-19{\cdot}74)^2}{19{\cdot}74} + \frac{(40-37{\cdot}59)^2}{37{\cdot}59} + \frac{(30-28{\cdot}53)^2}{28{\cdot}53} + \frac{(8-8{\cdot}61)^2}{8{\cdot}61} \approx 1$$



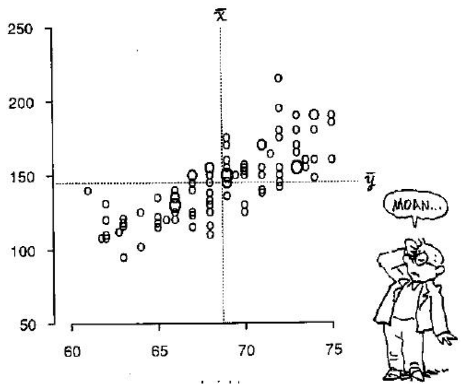
Ker eksperimentalna vrednost ne pade v kritično območje, ne moremo zavrniti ničelne domneve, da je spremenljivka normalno porazdeljena. ◇

Obstajajo tudi drugi testi za preverjanje porazdelitve spremenljivke, npr. Kolmogorov-Smirnov test.



Poglavje 15

Bivariatna analiza in regresija



Bivariatna analiza

$$X \longleftrightarrow Y \quad \text{povezanost}$$

$$X \longrightarrow Y \quad \text{odvisnost}$$

Mere povezanosti ločimo glede na tip spremenljivk:

1. NOMINALNI tip para spremenljivk (ena od spremenljivk je nominalna): χ^2, kontingenčni koeficienti, koeficienti asociacije;

2. ORDINALNI tip para spremenljivk (ena spremenljivka je ordinalna druga ordinalna ali boljša) koeficient korelacije rangov;

3. ŠTEVILSKI tip para spremenljivk (obe spremenljivki sta številski): koeficient korelacije.

15.1 Povezanost dveh nominalnih spremenljivk

Primer:

- ENOTA: dodiplomski študent neke fakultete v letu 1993/94;
- VZOREC: slučajni vzorec 200 študentov;
- 1. SPREMENLJIVKA: spol;
- 2. SPREMENLJIVKA: stanovanje v času študija.

Zanima nas ali študentke drugače stanujejo kot študentje oziroma: ali sta spol in stanovanje v času študija povezana. V ta namen podatke študentov po obeh spremenljivkah uredimo v dvorazsežno frekvenčno porazdelitev. To tabelo imenujemo kontingenčna tabela.

Denimo, da so podatki za vzorec urejeni v naslednji kontingenčni tabeli:

	starši	št. dom	zasebno	skupaj
moški	16	40	24	80
ženske	48	36	36	120
skupaj	64	76	60	200

Ker nas zanima ali študentke drugače stanujejo v času študija kot študentje, moramo porazdelitev stanovanja študentk primerjati s porazdelitvijo študentov.

Ker je število študentk različno od števila študentov, moramo zaradi primerjave izračunati relativne frekvence:

	starši	št. dom	zasebno	skupaj
moški	20	50	30	100
ženske	40	30	30	100
skupaj	32	38	30	100

Če med spoloma ne bi bilo razlik, bi bili obe porazedelitvi (za moške in ženske) enaki porazdelitvi pod "skupaj". Naš primer kaže, da se odstotki razlikujejo: npr. le 20% študentov in kar 40% študentk živi med študijem pri starših. Odstotki v študentskih domovih pa so ravno obratni. Zasebno pa stanuje enak odstotek deklet in fantov. Že pregled relativnih frekvenc (po vrsticah) kaže, da sta spremenljivki povezani med seboj.

Relativne frekvence lahko računamo tudi po stolpcih:

	starši	št. dom	zasebno	skupaj
moški	25	56·6	40	40
ženske	75	43·4	60	60
skupaj	100	100	100	100

Relativno frekvenco lahko prikažemo s stolpci ali krogi. Kontingenčna tabela kaže podatke za slučajni vzorec. Zato nas zanima, ali so razlike v porazdelitvi tipa stanovanja v času študija po spolu statistčno značilne in ne le učinek vzorca:

H_0: spremenljivki nista povezani

H_1: spremenljivki sta povezani

Za preverjanje domneve o povezanosti med dvema nominalnima spremenljivkama na osnovi vzorčnih podatkov, podanih v dvo-razsežni frekvenčni porazdelitvi, lahko uporabimo χ^2 test.

Ta test sloni na primerjavi empiričnih (dejanskih) frekvenc s teoretičnimi frekvencami, ki so v tem primeru frekvence, ki bi bile v kontingenčni tabeli, če spremenljivki ne bi bili povezani med seboj. To pomeni, da bi bili porazdelitvi stanovanja v času študija deklet in fantov enaki. Če spremenljivki nista povezani med seboj, so verjetnosti hkratne zgoditve posameznih vrednosti prve in druge spremenljivke enake produktu verjetnosti posameznih vrednosti. Npr., če označimo moške z M in stanovanje pri starših s S, je:

$$P(M) = \frac{80}{200} = 0{\cdot}40, \qquad P(S) = \frac{64}{200} = 0{\cdot}32,$$

$$P(M\,S) = P(M) \cdot P(S) = \frac{80}{200} \times \frac{64}{200} = 0{\cdot}128.$$

Teoretična frekvenca je verjetnost $P(M\,S)$ pomnožena s številom enot v vzorcu:

$$f'(M\,S) = n \cdot P(M\,S) = 200 \times \frac{80}{200} \times \frac{64}{200} = 25{\cdot}6.$$

Podobno izračunamo teoretične frekvence tudi za druge celice kontingenčne tabele.

Če teoretične frekvence zaokrožimo na cela števila, je tabela izračunanih teoretičnih frekvenc f_i' naslednja:

	starši	št. dom	zasebno	skupaj
moški	26	30	24	80
ženske	38	46	36	120
skupaj	64	76	60	200

Spomnimo se tabel empiričnih (dejanskih) frekvenc f_i: Testna statistika χ^2, ki primerja dejanske in teoretične frekvence je

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

kjer je k število celic v kontingenčni tabeli. Testna statistika χ^2 se porazdeljuje po χ^2 porazdelitvi s $(s-1)(v-1)$ prostostnimi stopnjami, kjer je s število vrstic v kontingenčni tabeli in s število stolpcev. Ničelna in osnovna domneva sta v primeru tega testa

H_0: $\chi^2 = 0$ (spremenljivki nista povezani)
H_1: $\chi^2 > 0$ (spremenljivki sta povezani)

Iz tabele za porazdelitev χ^2 lahko razberemo kritične vrednost te statistike pri 5% stopnji značilnosti:

$$\chi^2_{1-\alpha}[(s-1)(v-1)] = \chi^2_{0{\cdot}95}(2) = 5{\cdot}99.$$

Eksperimentalna vrednost statistike χ^2 pa je:

$$\chi^2_e = \frac{(16-26)^2}{26} + \frac{(40-30)^2}{30} + \frac{(24-24)^2}{24} + \frac{(48-38)^2}{38} + \frac{(36-46)^2}{46} + \frac{(36-36)^2}{36} = 12.$$

Ker je ekperimentalna vrednost večja od kritične vrednosti, pomeni, da pade v kritično območje. To pomeni, da ničelno domnevo zavrnemo. Pri 5% stopnji značilnosti lahko sprejmemo osnovno domnevo, da sta spremenljivki statistično značilno povezani med seboj. ◇

Testna statistika χ^2 je lahko le pozitivna. Zavzame lahko vrednosti v intervalu $[0, \chi^2_{\max}]$, kjer je $\chi^2_{\max} = n(k-1)$, če je $k = \min(v, s)$. Statistika χ^2 v splošnem ni primerljiva. Zato je definiranih več **kontingenčnih koeficientov**, ki so bolj ali manj primerni. Omenimo naslednje:

1. **Pearsonov koeficient**: $\Phi = \chi^2/n$, ki ima zgornjo mejo $\Phi^2_{\max} = k-1$.

2. **Cramerjev koeficient**: $\alpha = \sqrt{\dfrac{\Phi^2}{k-1}} = \sqrt{\dfrac{\chi^2}{n(k-1)}}$, ki je definiran na intervalu $[0,1]$.

3. **Kontingenčni koeficient**: $C = \sqrt{\chi^2/(\chi^2+n)}$, ki je definiran na intervalu $[0, C_{\max}]$, kjer je $C_{\max} = \sqrt{k/(k-1)}$.

15.2 Koeficienti asociacije

Denimo, da imamo dve nominalni spremenljivki, ki imata le po dve vrednosti (sta dihotomni). Povezanost med njima lahko računamo poleg kontingenčnih koeficientov s **koeficienti asociacije** na osnovi frekvenc iz kontingenčne tabele 2×2:

$Y \backslash X$	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
	$a+c$	$b+d$	N

kjer je $N = a+b+c+d$. Na osnovi štirih frekvenc v tabeli je definirnih več koeficientov asociacije. Omenimo najpomembnejše:

- **Yulov koeficient asociacije**:
$$Q = \frac{ad-bc}{ad+bc} \in [-1,1].$$
- **Sokal Michenerjev koeficient**:
$$S = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \in [0,1].$$
- **Pearsonov koeficient**:
$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1,1].$$
  Velja $\chi^2 = N \cdot \phi^2$.
- **Jaccardov koeficient**:
$$J = \frac{a}{a+b+c} \in [0,1],$$

Primer: povezanost mod kaznivimi dejanji in alkoholizmom. Tabela kaže podatke za $N =$ 10.750 ljudi

alk. \ kaz. d.	DA	NE	skupaj
DA	50	500	550
NE	200	10.000	10.200
skupaj	250	10.500	10.750

Izračunajmo koeficiente asociacije:

$$Q = \frac{50 \times 10000 - 200 \times 500}{50 \times 10000 + 200 \times 500} = 0{\cdot}67,$$

$$S = \frac{10050}{10750} = 0{\cdot}93, \quad \text{in} \quad J = \frac{50}{50 + 500 + 200} = 0{\cdot}066.$$

Izračunani koeficienti so precej različni. Yulov in Sokal Michenerjev koeficient kažeta na zelo močno povezanost med kaznjivimi dejanji in alkoholizmom, medtem kot Jaccardov koeficient kaže, da med spremenljivkama ni povezanosti. Pri prvih dveh koeficientih povezanost povzroča dejstvo, da večina alkoholiziranih oseb ni naredila kaznivih dejanj in niso alkoholiki (frekvenca d). ◇

Ker Jaccardov koeficient upošteva le DA DA ujemanje, je lažji za interpretacijo. V našem primeru pomeni, da oseba, ki je naredila kaznivo dejanje, sploh ni nujno alkoholik.

15.3 Povezanost dveh ordinalnih spremenljivk

V tem primeru gre za študij povezanosti med dvema spremenljivkama, ki sta vsaj ordinalnega značaja.

Primer: Vzemimo slučajni vzorec šestih poklicev in ocenimo, koliko so odgovorni (O) in koliko fizično naporni (N).

V tem primeru smo poklice uredili od najmanj odgovornega do najbolj odgovornega in podobno od najmanj fizično napornega do najbolj napornega. Poklicem smo torej priredili range po odgovornosti (R_0) in po napornosti (R_N) od 1 do 6 (glej tabelo).

poklic	R_0	R_N
A	1	6
B	2	4
C	3	5
D	4	2
E	5	3
F	6	1

Povezanost med spremenljivkama lahko merimo s koeficientom korelacije rangov r_s (Spearman), ki je definiran takole:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}.$$

kjer je d_i razlika med **rangoma** v i-ti enoti.[1] Koeficient korelacije rangov lahko zavzame vrednosti na intervalu $[-1, 1]$. Če se z večanjem rangov po prvi spremenjivki večajo rangi

[1] Mimogrede: ni se težko prepričati o naslednji identiteti: $1^2 + 2^2 + 3^2 + \cdots + n^2 = n(n+1)(2n+1)/6$ (npr. s popolno indukcijo, ali pa neposredno iz vsote enačb $(1+i)^3 = 1 + 3i + 3i^2 + i^3$ za $i = 0, \ldots, n$).

tudi po drugi spremenljivki, gre za pozitivno povezanost. Tedaj je koeficient pozitiven in blizu 1. Če pa se z večanjem rangov po prvi spremenljivki rangi po drugi spremenljivki manjšajo, gre za negativno povezanost. Koeficient je tedaj negativen in blizu -1. V našem (preprostem) primeru gre negativno povezanost. Če ne gre za pozitivno in ne za negativno povezanost, rečemo, da spremenljivki nista povezani.

Izračunajmo koeficient korelacije rangov za primer šestih poklicev:

poklic	R_0	R_N	d_i	d_i^2
A	1	6	-5	25
B	2	4	-2	4
C	3	5	-2	4
D	4	2	2	4
E	5	3	2	4
F	6	1	5	25
vsota			0	66

$$r_s = 1 - \frac{6 \times 66}{6 \times 35} = 1 - 1,88 = -0{\cdot}88.$$

Res je koeficient blizu -1, kar kaže na močno negativno povezanost teh 6-ih poklicev.

Omenili smo, da obravnavamo 6 slučajno izbranih poklicev. Zanima nas, ali lahko na osnovi tega vzorca posplošimo na vse poklice, da sta odgovornost in fizična napornost poklicev (negativno) povezana med seboj. Upoštevajmo 5% stopnjo značilnosti. Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho_s = 0$ (spremenljivki nista povezani)
H_1: $\rho_s \neq 0$ (spremenljivki sta povezani)

kjer populacijski koeficient označimo s ρ_s. Pokaže se, da se testna statistika

$$t = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

porazdeljuje približno po t-porazdelitvi z $m = (n-2)$ prostostnimi stopnjami. Ker gre za dvostranski test, sta kritični vrednosti enaki

$$\pm t_{\alpha/2} = \pm t_{0{\cdot}025}(4) = \pm 2{\cdot}776.$$

Eksperimentalna vrednost statistike je za naš primer

$$t_e = \frac{-0{\cdot}88 \times 2}{\sqrt{1-(-0{\cdot}88)^2}} = \frac{-1{\cdot}76}{0{\cdot}475} = -3{\cdot}71.$$

Eksperimentalna vrednost pade v kritično območje. Pri 5% stopnji značilnosti lahko rečemo, da sta odgovornost in fizična napornost (negativno) povezani med seboj. ◇

Če je ena od obeh spremenljivk številska, moramo vrednosti pred izračunom d_i rangirati. Če so kakšne vrednosti enake, zanje izračunamo povprečne pripadajoče range.

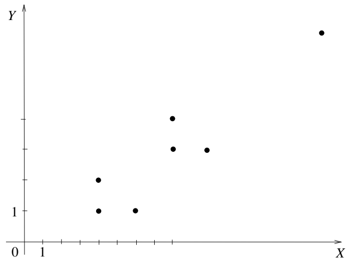
15.4 Povezanost dveh številskih spremenljivk

Vzemimo primer dveh številskih spremenljivk:

X - izobrazba (število priznanih let šole)
Y - število ur branja dnevnih časopisov na teden

Podatki za 8 slučajno izbranih oseb so:

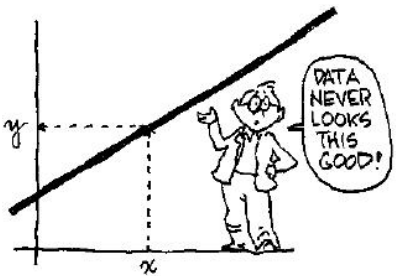
X	10	8	16	8	6	4	8	4
Y	3	4	7	3	1	2	3	1

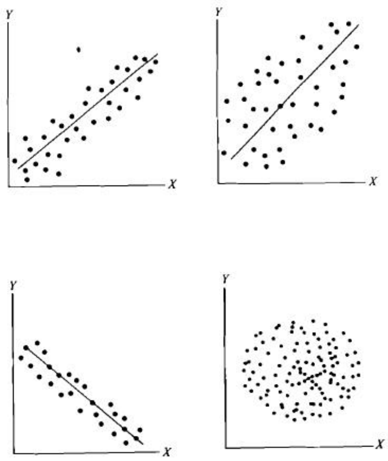


Grafično lahko ponazorimo povezanost med dvema številskima spremenljivkama z *razsevnim grafikonom.* To je, da v koordinatni sistem, kjer sta koordinati obe spremenljivki, vrišemo enote s pari vrednosti.

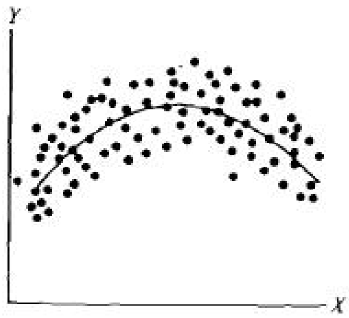
Tipi povezanosti:

- **funkcijska** povezanost: vse točke ležijo na krivulji:
- **korelacijska** (stohastična] povezanost: točke so od krivulje bolj ali manj odklanjajo (manjša ali večja povezanost).





Tipični primeri linearne povezanosti spremenljivk



Primer nelinearne povezanosti spremenljivk.

Kovarianca

$$\mathsf{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_X)\cdot(y_i-\mu_Y)$$

meri linearno povezanost med spremenljivkama.
$\mathsf{Cov}(X,Y) > 0$ pomeni pozitivno linearno povezanost,
$\mathsf{Cov}(X,Y) = 0$ pomeni da ni linearne povezanosti,
$\mathsf{Cov}(X,Y) < 0$ pomeni negativno linearno povezanost.

(Pearsonov) koeficient korelacije je

$$\rho_{XY} = \frac{\mathsf{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)\cdot(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_Y)^2}}.$$

Koeficient korelacije lahko zavzame vrednosti v intervalu $[-1,1]$. Če se z večanjem vrednosti prve spremenljivke večajo tudi vrednosti druge spremenljivke, gre za **pozitivno povezanost**. Tedaj je koeficient povezanosti blizu 1. Če pa se z večanjem vrednosti prve spremenljivke vrednosti druge spremenljivke manjšajo, gre za **negativno povezanost**. Koeficient je tedaj negativen in blizu -1. Če ne gre za pozitivno in ne za negativno povezanost, rečemo da spremenljivki nista povezani in koeficient je blizu 0.

Statistično sklepanje o korelacijski povezanosti:

Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho = 0$ (spremenljivki nista linearno povezani)
H_1: $\rho \neq 0$ (spremenljivki sta linearno povezani)

Pokaže se, da se testna statistika

$$t = \frac{r\cdot\sqrt{n-2}}{\sqrt{1-r^2}}$$

porazdeljuje po t porazdelitvi z $m = (n-2)$ prostostnimi stopnjami. Z r označujemo koeficient korclacije na vzorcu in z ρ koeficient korelacije na populaciji.

Primer: Preverimo domnevo, da sta izobrazba (število priznanih let šole) in število ur branja dnevnih časopisov na teden povezana med seboj pri 5% stopnji značilnosti. Najprej izračunajmo vzorčni koeficient korelacije:

x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x)\cdot$ $(y_i - \mu_y)$
10	3	2	0	4	0	0
8	4	0	1	0	1	0
16	7	8	4	64	16	32
8	3	0	0	0	0	0
6	1	-2	-2	4	4	4
4	2	-4	-1	16	1	4
8	3	0	0	0	0	0
4	1	-4	-2	16	4	8
64	24	0	0	104	26	48

$$r = \frac{48}{\sqrt{104\times 26}} = 0{\cdot}92.$$

Ker gre za dvostranski test, je kritično območje določeno s kritičnima vrednostima

$$\pm t_{\alpha/2}(n-2) = \pm t_{0{\cdot}025}(6) = \pm 2{\cdot}447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \frac{0{\cdot}92\sqrt{8-2}}{\sqrt{1-0{\cdot}92^2}} = 2{\cdot}66.$$

Eksperimentalna vrednost pade v kritično območje. *Zaključek:* ob 5% stopnji značilnosti lahko rečemo, da je izobrazba linearno povezana z branjem dnevnih časopisov.

15.5 Parcialna korelacija

Včasih je potrebno meriti zvezo med dvema spremenljivkama in odstraniti vpliv vseh ostalih spremenljivk. To zvezo dobimo s pomočjo koeficienta parcialne korclacije. Pri tem seveda predpostavljamo, da so vse spremenljivke med seboj linearno povezane. Če hočemo iz zveze med spremenljivkama X in Y odstraniti vpliv tretje spremenljivke Z, je **koeficient parcialne korelacije**:

$$\boldsymbol{r_{XY,Z} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{1-r_{XZ}^2}\,\sqrt{1-r_{YZ}^2}}.}$$

Tudi ta koeficient, ki zavzema vrednosti v intervalu $[-1, 1]$, interpretiramo podobno kot običajni koeficient korelacije. S pomočjo tega obrazca lahko razmišljamo naprej, kako bi izločili vpliv naslednjih spremenljivk.

Primer: V neki ameriški raziskavi, v kateri so proučevali vzroke za kriminal v mestih, so upoštevali naslednje spremenljivke: X : % nebelih prebivalcev, Y : % kaznivih dejanj, Z : % revnih prebivalcev, U : velikost mesta. Izračunali so naslednje koeficiente korelacije:

	X	Z	U	Y
X	1	0·51	0·41	0·36
Z		1	0·29	0·60
U			1	0·49
Y				1

Zveza med nebelim prebivalstvom in kriminalom je $r_{XY} = 0{\cdot}36$. Zveza je kar močna in lahko bi mislili, da nebeli prebivalci povzročajo več kaznivih dejanj. Vidimo pa še, da je zveza med revščino in kriminalom tudi precejšna $r_{YZ} = 0{\cdot}60$. Lahko bi predpostavili, da revščina vpliva na zvezo med nebelci in kriminalom, saj je tudi zveza med revnimi in nebelimi precejšna $r_{XZ} = 0{\cdot}51$. Zato poskusim odstraniti vpliv revščine iz zveze: "nebelo prebivalstvo : kazniva dejanja":

$$r_{XY,Z} = \frac{0{\cdot}36 - 0{\cdot}51 \times 0{\cdot}60}{\sqrt{1-0{\cdot}51^2}\sqrt{1-0{\cdot}60^2}}.$$

Vidimo, da se je linearna zveza zelo zmanjšala. Če pa odstranimo še vpliv velikosti mesta, dobimo parcialno korelacijo $-0{\cdot}02$ oziroma zveze praktično ni več. ◇

15.6 Regresijska analiza

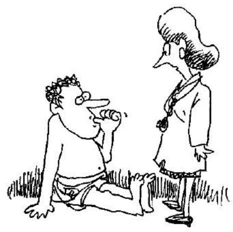
Regresijska funkcija $Y' = f(X)$ kaže, kakšen bi bil vpliv spremenljivke X na Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y. Ker pa so ponavadi še drugi vplivi na proučevano spremenljivko Y, se točke, ki predstavljajo enote v razsevnem grafikonu, odklanjajo od idealne regresijske krivulje

$$Y = Y' + E = f(X) + E,$$

kjer X imenujemo neodvisna spremenljivka, Y odvisna spremenljivka in E člen napake (ali motnja, disturbanca). Če je regresijska fukcija linearna:

$$Y' = f(X) = a + bX,$$

je regresijska odvisnost

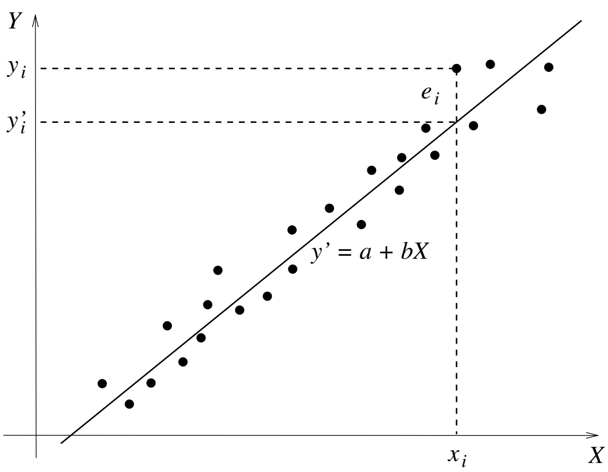


$$Y = Y' + E = a + bX + E$$

oziroma za i to enoto

$$y_i = y'_i + e_i = a + bx_i + e_i.$$

Regresijsko odvisnost si lahko zelo nazorno predstavimo v razsevnem grafikonu:



Regresijsko funkcijo lahko v splošnem zapišemo

$$Y' = f(X, a, b, \cdots),$$

kjer so a, b, ... parametri funkcije. Ponavadi se moramo na osnovi pregleda razsevnega grafikona odločiti za tip regresijske funkcije in nato oceniti parametre funkcije, tako da se regresijska krivulja kar se da dobro prilega točkam v razsevnem grafikonu.

Pri dvorazsežno normalno porazdeljenem slučajnem vektorju $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je, kot vemo

$$\mathsf{E}(Y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

Pogojna porazdelitev Y glede na X je tudi normalna:

$$N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y\sqrt{1 - \rho^2}).$$

Regresija je linearna in regresijska krivulja premica, ki gre skozi točko (σ_x, σ_y). Med Y in X ni linearne zveze, sta le 'v povprečju' linearno odvisni. Če označimo z $\boldsymbol{\beta} = \rho\frac{\sigma_y}{\sigma_x}$ *regresijski koeficient*, $\boldsymbol{\alpha} = \mu_y - \beta\mu_x$ in $\boldsymbol{\sigma^2} = \sigma_y\sqrt{1 - \rho^2}$, lahko zapišemo zvezo v obliki

$$y = \alpha + \beta x.$$

Preverjanje regresijskih koeficientov

Po metodi momentov dobimo cenilki za α in β:

$$B = R\frac{C_y}{C_x} = \frac{C_{xy}}{C_x^2} \qquad \text{in} \qquad A = \overline{Y} - B\overline{X},$$

kjer so $C_x^2 = \sum_{i=1}^n (X_i - \overline{X})^2$, $C_y^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2$ in $C_{xy} = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$.

Kako sta cenilki B in A porazdeljeni?

$$B = \frac{C_{xy}}{C_x^2} = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}(Y_i - \overline{Y}) = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2} Y_i.$$

Ker proučujemo pogojno porazdelitev Y glede na X (torej so vrednost X poznane), obravnavamo spremenljivke $X_1, \ldots, X_n$ kot konstante. Ker je B linearna funkcija spremenljivk $Y_1, \ldots, Y_n$, ki so normalno porazdeljene $Y_i \sim N(\alpha + \beta X_i, \sigma)$, je tudi B normalno porazdeljena. Določimo parametra te porazdelitve:

$$\mathsf{E}B = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2} \mathsf{E}Y_i = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}(\alpha + \beta X_i) = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2} \beta (X_i - \overline{X}) = \beta.$$

Pri tem upoštevamo, da je $\sum_{i=1}^n (X_i - \overline{X}) = 0$ in da sta α ter $\overline{X}$ konstanti.

$$\mathsf{D}B = \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{C_x^4} \mathsf{D}Y_i = \frac{\sigma^2}{C_x^2}.$$

Torej je $\quad B \sim N\Big(\beta, \dfrac{\sigma}{C_x}\Big)$, $\quad$ oziroma $\quad \dfrac{B - \beta}{\sigma} C_x \sim N(0,1)$.

Podobno dobimo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2 \Big(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\Big).$$

Težje se je dokopati do cenilke za parameter σ^2. Označimo $Q^2 = \sum_{i=1}^n (Y_i - A - BX_i)^2$. Po nekaj računanja se izkaže, da velja $\mathsf{E}\dfrac{Q^2}{\sigma^2} = n - 2$.

Torej je $S^2 = \dfrac{Q^2}{n-2} = \dfrac{\sigma^2}{n-2}\,\dfrac{Q^2}{\sigma^2}$ nepristranska cenilka za σ^2.

S^2 je neodvisna od A in B. Testni statistiki za A in B sta tedaj

$$T_A = \frac{A - \mathsf{E}A}{\sqrt{\mathsf{D}A}} = \frac{A - \alpha}{S}\sqrt{\frac{nC_x^2}{C_x^2 + n\overline{X}^2}} = \frac{A - \alpha}{S} C_x \sqrt{\frac{n}{\sum_{i=1}^n X_i^2}},$$

$$T_B = \frac{B - \mathsf{E}B}{\sqrt{\mathsf{D}B}} = \frac{B - \beta}{S} C_x,$$

ki sta obe porazdeljeni po Studentu $S(n-2)$. Statistika za σ^2 pa je spremenljivka $\dfrac{Q^2}{\sigma^2} = (n-2)\dfrac{S^2}{\sigma^2}$, ki je porazdeljena po $\chi^2(n-2)$. Pokazati je mogoče tudi, da velja

$$Q^2 = C_y^2 - B^2 C_x^2 = C_y^2 (1 - R^2).$$

To nam omogoča S v statistikah zapisati z C_y in R. Te statistike uporabimo tudi za določitev intervalov zaupanja za parametre α, β in σ^2.

15.7 Linearni model

Pri proučevanju pojavov pogosto teorija postavi določeno funkcijsko zvezo med obravnavanimi spremenljivkami. Oglejmo si primer *linearnega modela*, ko je med spremenljivkama x in y linearna zveza

$$y = \alpha + \beta x.$$

Za dejanske meritve se pogosto izkaže, da zaradi različnih vplivov, ki jih ne poznamo, razlika $u = y - \alpha - \beta x$ v splošnem ni enaka 0, čeprav je model točen. Zato je ustreznejši *verjetnostni linearni model*

$$Y = \alpha + \beta X + U,$$

kjer so X, Y in U slučajne spremenljivke in $\mathsf{E}U = 0$ – model je vsaj v povprečju linearen.

Slučajni vzorec (meritve) $(X_1, Y_1), \dots, (X_n, Y_n)$ je realizacija slučajnega vektorja. Vpeljimo spremenljivke

$$U_i = Y_i - \alpha - \beta X_i$$

in predpostavimo, da so spremenljivke U_i med seboj neodvisne in enako porazdeljene z matematičnim upanjem 0 in disperzijo σ^2. Torej je:

$$\mathsf{E}U_i = 0, \quad \mathsf{D}U_i = \sigma^2 \quad \text{in} \quad \mathsf{E}(U_iU_j) = 0, \quad \text{za } i \neq j.$$

Običajno privzamemo še, da lahko vrednosti X_i točno določamo – X_i ima vedno isto vrednost. Poleg tega naj bosta vsaj dve vrednosti X različni. Težava je, da (koeficientov) premice $y = \alpha + \beta x$ ne poznamo. Recimo, da je približek zanjo premica $y = a + bx$. Določimo jo po *načelu najmanjših kvadratov* z minimizacijo funkcije

$$f(a, b) = \sum_{i=1}^{n} (y_i - (bx_i + a))^2.$$

Naloga zadošča pogojem izreka. Iz pogoja $\nabla P = 0$ dobimo enačbi

$$\frac{\partial f}{\partial a} = \sum_{i=1}^{n} 2(y_i - (bx_i + a)) = 0, \qquad \frac{\partial f}{\partial b} = \sum_{i=1}^{n} 2(y_i - (bx_i + a))x_i = 0,$$

z rešitvijo

$$b = \frac{n\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}, \qquad a = \frac{1}{n}\Big(\sum y - b\sum x\Big).$$

oziroma, če vpeljemo oznako $\overline{z} = \frac{1}{n}\sum z$:

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \qquad a = \overline{y} - b\overline{x}.$$

Poglejmo še Hessovo matriko

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial a^2} & \frac{\partial^2 f}{\partial a \partial b} \\ \frac{\partial^2 f}{\partial b \partial a} & \frac{\partial^2 f}{\partial b^2} \end{bmatrix} = 2 \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}.$$

Ker je $\Delta_1 = 2\sum x^2 > 0$ in

$$\Delta_2 = 4\Big(n\sum x^2 - \Big(\sum x\Big)^2\Big) = 2\sum\sum(x_i - x_j)^2 > 0,$$

je matrika H pozitivno definitna in zato funkcija P strogo konveksna. Torej je *regresijska premica* enolično določena. Seveda sta parametra a in b odvisna od slučajnega vzorca – torej slučajni spremenljivki. Iz dobljenih zvez za a in b dobimo že znani cenilki za koeficienta α in β

$$B = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \overline{Y} - B\overline{X}.$$

Iz prej omenjenih predpostavk lahko (brez poznavanja porazdelitve Y in U) pokažemo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2\Big(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\Big), \ \mathsf{E}B = \beta \quad \text{in} \quad \mathsf{D}B = \frac{\sigma^2}{C_x^2}, \quad K(A,B) = -\sigma^2\frac{\overline{X}}{C_x^2}.$$

Cenilki za A in B sta najboljši linearni nepristranski cenilki za α in β.

To metodo ocenjevanja parametrov regresijske funkcije imenujemo **metoda najmanjših kvadratov**. Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

$$Y = \mu_Y + \frac{\mathsf{Cov}(X,Y)}{\sigma_X^2}(X - \mu_X).$$

To funkcijo imenujemo tudi **prva** regresijska funkcija. Podobno bi lahko ocenili linearno regresijsko funkcijo

$$X = a^* + b^*Y.$$

Če z metodo najmanjših kvadratov podobno ocenimo parametra a^* in b^*, dobimo:

$$X = \mu_X + \frac{\mathsf{Cov}(X,Y)}{\sigma_Y^2}(Y - \mu_Y).$$

To funkcijo imenujemo **druga** regresijska funkcija,

Primer: Vzemimo primer 8 osob, ki smo ga obravnavali v poglavju o povezanosti dveh številskih spremenljivk. Spremenjivki sta bili:
X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.
Spomnimo se podatkov za teh 8 slučajno izbranh osob:

X	10	8	16	8	6	4	8	4
Y	3	4	7	3	1	2	3	1

Zanje izračunajmo obe regresijski premici in ju vrišimo v razsevni grafikon. Ko smo računali koeficient korelacije smo že izračunali aritmetični sredini

$$\mu_X = \frac{64}{8} = 8, \qquad \mu_Y = \frac{24}{8} = 3,$$

vsoti kvadratov odklonov od aritmetične sredine za obe spremenljivki

$$\sum_{i=1}^{n}(x_i-\mu_X)^2 = 104, \qquad \sum_{i=1}^{n}(y_i-\mu_Y)^2 = 26$$

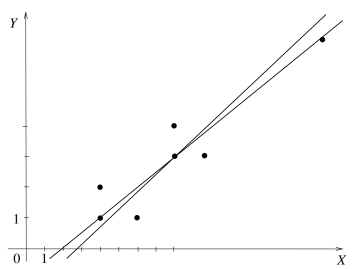
in vsoto produktov odklonov od obeh aritmetičnih sredin

$$\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y) = 48.$$

Potem sta regresijski premici

$$\begin{aligned} Y &= \mu_Y + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\displaystyle\sum_{i=1}^{N}(x_i-\mu_X)^2}(X-\mu_X), \\ X &= \mu_X + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\displaystyle\sum_{i=1}^{N}(y_i-\mu_Y)^2}(Y-\mu_Y), \end{aligned}$$

oziroma $Y = 0{\cdot}46X - 0{\cdot}69$ in $X = 1{\cdot}85Y + 2{\cdot}46$. Obe regresijski premici lahko vrišemo v razsevni grafikon in preverimo, če se res najbolje prilegata točkam v grafikonu: $\diamond$



Regresijski premici se sečeta v točki, določeni z aritmetičnima sredinama spremenljivk X in Y. Dokažite, da se premici vedno sečeta v tej točki.

15.7.1 Statistično sklepanje o regresijskem koeficientu

Vpeljimo naslednje oznake:

$Y = \alpha + \beta X$ regresijska premica na populaciji,
$Y = a + bX$ regresijska premica na vzorcu.

Denimo, da želimo preveriti domnevo o regresijskem koeficientu β. Postavimo ničelno in osnovno domnevo: H_0: $\beta = \beta_0$, H_1: $\beta \neq \beta_0$.

Nepristranska cenilka za regresijski koeficient β je $b = \mathsf{Cov}(X,Y)/s_X^2$, ki ima matematično upanje in standardno napako:

$$\mathsf{E}b = \beta; \qquad \mathsf{SE}(b) = \frac{s_Y\sqrt{1-r^2}}{s_X\sqrt{n-2}}.$$

Testna statistika za zgornjo ničelno domnevo je:

$$t = \frac{s_Y\sqrt{n-2}}{s_X\sqrt{1-r^2}}\,(b-\beta_0),$$

ki se porazdeljuje po t-porazdelitvi z $m = (n-2)$ prostostnimi stopnjami.

Primer: Vzemimo primer, ki smo ga že obravnavali. Spremenljivki sta

X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.

Preverimo domnevo, da je regresijski koeficient različen od 0 pri $\alpha = 5\%$.

Postavimo najprej ničelno in osnovno domnevo: H_0: $\beta = 0$, H_1: $\beta \neq 0$. Gre za dvostranski test. Zato je ob 5% stopnji značilnosti kritično območje določeno s kritičnimia vrednostima:

$$\pm t_{\alpha/2}(n-2) = \pm t_{0\cdot 025}(6) = \pm 2\cdot 447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \sqrt{\frac{104\times(8-2)}{26\times(1-0\cdot 92^2)}}\cdot(0\cdot 46-0) = 5\cdot 8.$$

Regresijski koeficient je statistično značilno različen od 0. ◇

15.7.2 Pojasnjena varianca (ang. ANOVA)

Vrednost odvisne spremenljivke Y_i lahko razstavimo na tri komponente:

$$y_i = \mu_Y + (y_i' - \mu_Y) + (y_i - y_i'),$$

kjer so pomeni posameznih komponent

μ_Y : rezultat splošnih vplivov,

$(y_i' - \mu_Y)$: rezultat vpliva spremenljivke X (regresija),

$(y_i - y_i')$: rezultat vpliva drugih dejavnikov (napake/motnje).

Če zgornjo enakost najprej na obeh straneh kvadriramo, nato seštejemo po vseh enotah in končno delimo s številom enot (N), dobimo:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_Y)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i'-\mu_Y)^2 + \frac{1}{N}\sum_{i=1}^{N}(y_i-y_i')^2.$$

To lahko zapišemo takole: $\sigma_Y^2 = \sigma_{Y'}^2 + \sigma_e^2$, kjer posamezni členi pomenijo:

σ_Y^2 : celotna varianca spremenljivke Y,

$\sigma_{Y'}^2$: pojasnjena varianca spremenljvke Y,

σ_e^2 : nepojasnjena varianca spremnenljivke Y.

Delež pojasnjene variance spremenljivke Y s spremenljivko X je $R = \sigma_{Y'}^2/\sigma_Y^2$. Imenujemo ga **determinacijski koeficient** in je definiran na intervalu $[0, 1]$. Pokazati se da, *da je v primeru linearne regresijske odvisnosti determinacijski koeficient enak* $R = \rho^2$, *kjer je* ρ *koeficient korelacije.* Kvadratni koren iz nepojasnjene variance σ_e imenujemo **standardna napaka regresijske ocene**, ki meri razpršenost točk okoli regresijske krivulje. Standarna napaka ocene meri kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo. V primeru linearne regresijske odvisnosti je standardna napaka enaka: $\sigma_e = \sigma_Y \sqrt{1-\rho^2}$.

Primer: Vzemimo spremenljivki

X - število ur gledanja televizije na teden

Y - število obiskov kino predstav na mesec

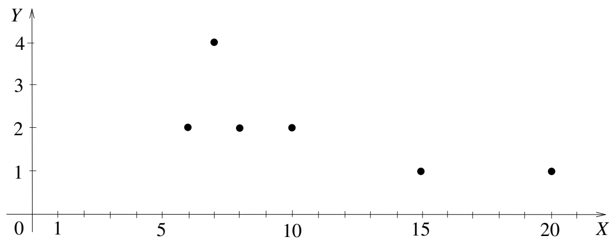
Podatki za 6 oseb so:

X	10	15	6	7	20	8
Y	2	1	2	4	1	2

Z linearno regresijsko funkcijo ocenimo, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden televizijo. Kolikšna je standardna napaka?

Kolikšen delež variance obiska kinopredstav lahko pojasnimo z gledanjem televizije?

Najprej si podatke predstavimo v razsevnem grafikonu:



Za odgovore potrebujemo naslednje izračune:

x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x) \cdot (y_i - \mu_y)$
10	2	-1	0	1	0	0
15	1	4	-1	16	1	-4
6	2	-5	0	25	0	0
7	4	-4	2	16	4	-8
20	1	9	-1	81	1	-9
8	2	-3	0	9	0	0
66	12	0	0	148	6	-21

$$Y' = 2 - \frac{21}{148}(X - 11) = 3{\cdot}54 - 0{\cdot}14X,$$

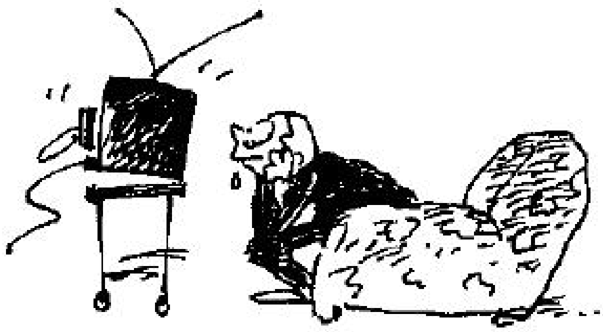
$$y'(18) = 3{\cdot}54 - 0{\cdot}14 \times 18 = 1{\cdot}02,$$

$$\rho = \frac{21}{146 \times 6} = -0{\cdot}70.$$

Torej je $\sigma_e^2 = \frac{6}{6}\sqrt{1-(-0{\cdot}70)^2} = 0{\cdot}71$ in $R = (0{\cdot}70)^2 = 0{\cdot}49$. Če oseba gleda 18 ur na teden televizijo, lahko pričakujemo, da bo 1-krat na mesec šla v kino, pri čemer je standardna napaka 0·7. 49% variance obiska kino predstav lahko pojasnimo z gledanjem televizije. ◇

Poglavje 16

Časovne vrste in trendi



Družbeno-ekonomski pojavi so časovno spremenljivi. Spremembe so rezultat delovanja najrazličnejših dejavnikov, ki tako ali drugače vplivajo na pojave. Sliko dinamike pojavov dobimo s časovimi vrstami. *Časovna vrsta* je niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke. Osnovni namen analize časovnih vrst je

- opazovati časovni razvoj pojavov,
- iskati njihove zakonitosti in
- predvidevati nadaljni razvoj.

Seveda to predvidevanje ne more biti popolnoma zanesljivo, ker je skoraj nemogoče vnaprej napovedati in upoštevati vse faktorje, ki vplivajo na proučevani pojav. Napoved bi veljala strogo le v primeru, če bi bile izpolnjene predpostavke, pod katerimi je napoved izdelana. Časovne vrste prikazujejo individualne vrednosti neke spremenljivke v času. Čas lahko interpretiramo kot trenutek ali razdobje; skladno s tem so časovne vrste

- trenutne, npr. število zaposlenih v določenem trenutku:
- intervalne, npr. družbeni proizvod v letu 1993.

Časovne vrste analiziramo tako, da opazujemo spreminjanje vrednosti členov v časovih vrstah in iščemo zakonitosti tega spreminjanja. Naloga enostavne analize časovnih vrst je primerjava med členi v isti časovni vrsti. Z metodami, ki so specifične za analizo časovnih vrst, analiziramo zakonitosti dinamike ene same vrste, s korelacijsko analizo pa zakonitosti odvisnosti v dinamiki več pojavov, ki so med seboj v zvezi.

Primer: Vzemimo število nezaposlenih v Sloveniji v letih od 1981 do 1990.

leto	X_k
1981	12 315
1982	13 700
1983	15 781
1984	15 300
1985	11 657
1986	14 102
1987	15 184
1988	21 311
1989	28 218
1990	44 227

V metodoloških pojasnilih v Statističnem letopisu Republike Slovenije 1991, so nezaposlni (spremenljivka X) opredeljeni takole:

Brezposelna oseba je oseba, ki je sposobna in voljna delati ter je pripravljena sprejeti zaposlitev, ki ustreza njeni strokovni izobrazbi oz. z delom pridobljeni delovni zmožnosti, vendar brez svoje krivde nima dela in možnosti, da si z delom zagotavlja sredstva za preživetje in se zaradi zaposlitve prijavi pri območni enoti Zavoda za zaposlovanje (do leta 1989 skupnosti za zaposlovanje). ◇

16.1 Primerljivost členov v časovni vrsti

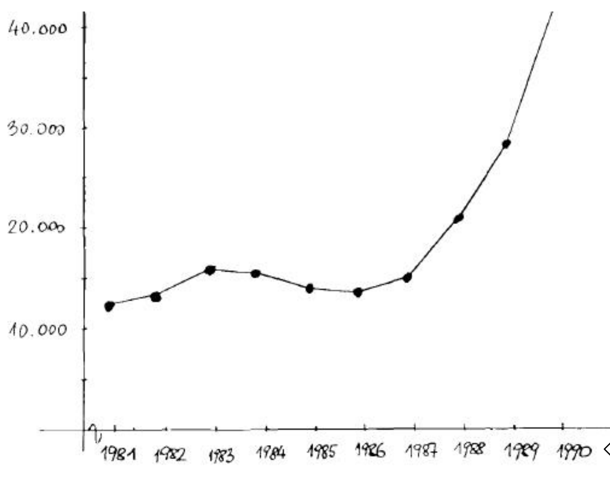
Kljub temu, da so členi v isti časovni vrsti istovrstne količine, dostikrat niso med seboj neposredno primerljivi. Osnovni pogoj za primerljivost členov v isti časovni vrst je pravilna in nedvoumna opredelitev pojava, ki ga časovna vrsta prikazuje. Ta opredelitev mora biti vso dobo opazovanja enaka in se ne sme spreminjati. Ker so spremembe pojava, ki ga časovna vrsta prikazuje bistveno odvisne od časa, je zelo koristno, če so **časovni razmiki med posameznimi členi enaki**. Na velikost pojavov dostikrat vplivajo tudi **administrativni** ukrepi, ki z vsebino proučevanja nimajo neposredne zveze.

En izmed običajnih vzrokov so upravnoteritorialne spremembe, s katerimi se spremeni geografska opredelitev pojava, ki onemogoča primerljivost podatkov v časovni vrsti. V tem primeru je potrebno podatke časovne vrste za nazaj preračunati za novo območje.

16.2 Grafični prikaz časovne vrste

Kompleksen vpogled v dinamiko pojavov dobimo z grafičnim prikazom časovnih vrst v koordinatnem sistemu, kjer nanašamo na abscisno os čas in na ordinatno vrednosti dane spremenljivke. V isti koordinatni sistem smemo vnašati in primerjati le istovrstne časovne vrste.

Primer: Grafično prikažimo število brezposelnih v Sloveniji v letih od 1981 do 1990.



◇

16.3 Indeksi

Denimo, da je časovna vrsta dana z vrednostmi neke spremenljivke v časovnih točkah takole:

$$X_1, X_2, \ldots, X_n.$$

O indeksih govorimo, kadar z relativnimi števili primerjamo istovrstne podatke. Glede na to, kako določimo osnovo, s katero primerjamo člene v časovni vrsti, ločimo dve vrsti indeksov:

- **Indeksi s stalno osnovo.** Člene časovnih vrst primerjamo z nekim stailnim členom v časovni vrsti, ki ga imenujemo osnova X_0: $I_{k/0} = (X_k/X_0) \cdot 100$.
- **Verižni indeksi.** Za dano časovno vrsto računamo vrsto verižnih indeksov tako, da za vsak člen vzamemo za osnovo predhodni člen: $I_k = (X_k/X_{k-1}) \cdot 100$.

Člene časovne vrste lahko primerjamo tudi z absolutno in relativno razliko med členi:

- **Absolutna razlika:** $D_k = X_k - X_{k-1}$.
- **Stopnja rasti** (relativna razlika med členi): $T_k = \big((X_k - X_{k-1})/X_{k-1}\big) \cdot 100 = I_k - 100$.

Interpretacija indeksov

	pojav		
indeks	raste	stagnira	pada
s stalno osnovo	$I_{k+1/0} > I_{k/0}$	$I_{k+1/0} = I_{k/0}$	$I_{k+1/0} < I_{k/0}$
verižni indeks	$I_k > 100$	$I_k = 100$	$I_k < 100$
stopnja rasti	$T_k > 0$	$T_k = 0$	$T_k < 0$

Primer: Izračunajmo omenjene indekse za primer brezposelnih v Sloveniji:

leto	X_k	$I_{k/0}$	I_k	T_k
1981	12 315	100	–	–
1982	13 700	111	111	11
1983	15 781	128	115	15
1984	15 300	124	97	−3
1985	11 657	119	96	−4
1986	14 102	115	97	−3
1987	15 184	124	107	7
1988	21 311	173	141	41
1989	28 218	229	132	32
1990	44 227	359	157	57

Rezultati kažejo, da je bila brezposenost v letu 1990 kar 3·5-krat večja kot v letu 1981 (glej indeks s stalno osnovo). Iz leta 1989 na leto 1990 je bil prirast nezposlenih 57% (glej stopnjo rasti).

16.4 Sestavine dinamike v časovnih vrstah

Posamezne vrednosti časovnih vrst so rezultat številnih dejavnikov, ki na pojav vplivajo. Iz časovne vrste je moč razbrati skupen učinek dejavnikov, ki imajo širok vpliv na pojav, ki ga proučujemo. Na časovni vrsti opazujemo naslednje vrste sprememb:

1. Dolgoročno gibanje ali trend - X_T podaja dolgoročno smer razvoja. Običajno ga je mogoče izraziti s preprostimi rahlo ukrivljenimi krivuljami.

2. Ciklična gibanja - X_C, so oscilarijo okoli trenda. Poriode so ponavadi daljše od enega leta in so lahko različno dolge.

3. Sezonske oscilacije - X_S so posledice vzrokov, ki se pojavljajo na stalno razdobje. Periode so krajše od enega leta, ponavadi sezonskega značaja.

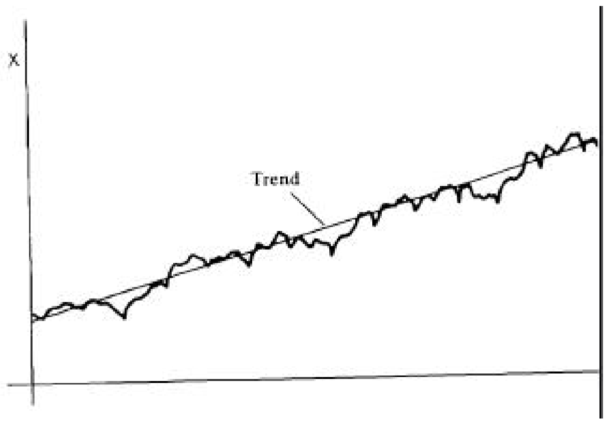
4. Naključne spremembe - X_E so spremembe, ki jih ne moremo razložiti s sistematičnimi gibanji (1, 2 in 3).

Časovna vrsta ne vsebuje nujno vseh sestavin. Zvezo med sestavinami je mogoče prikazati z nekaj osnovnim modeli. Npr.:

$$X = X_T + X_C + X_S + X_E$$

$$\text{ali} \quad X = X_T \cdot X_C \cdot X_S \cdot X_F;$$

$$\text{ali} \quad X = X_T \cdot X_C \cdot X_S + X_E.$$



Primer časovne vrste z vsemi štirimi sestavinami

Ali je v časovni vrsti trend?

Obstaja statistični test, s katerim preverjamo ali trend obstaja v časovni vrsti. Med časom in spremenljivko izračunamo koeficient korelacije rangov

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

kjer je d_i, razlika med rangoma i tega časa in pripadajoče vrednosti spremenljivke. Ničelna in osnovna domneva sta:

H_0: $\rho_e = 0$ trend ne obstaja
H_1: $\rho_e \neq 0$ trend obstaja

Ustrezna testna statistika

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

se porazdeluje približno po t porazdelitvi z $(n-2)$ prostostnimi stopnjami.

Metode določanja trenda

- Prostoročno
- Metoda drsečih sredin
- Metoda najmanjših kvadratov
- Druge analitične metode

Drseče sredine

Metoda drsečih sredin lahko pomaga pri določitvi ustreznega tipa krivulje trenda. V tem primeru namesto člena časovne vrste zapišemo povprečje določenega števila sosednjih členov. Če se odločimo za povprečje treh členov, govorimo o tričlenski vrsti drsečih sredin. Tedaj namesto članov v osnovni časovni vrsti X_k: tvorimo tričlenske drseče sredine X :

$$X_k' = \frac{X_{k-1} + X_k + X_{k+1}}{3}.$$

V tem primeru prvega in zadnjega člena časovne vrste ne moremo izračunati.

- Včasih se uporablja utežena aritmetična sredina, včasih celo geometrijska za izračun drsečih sredin.
- Če so v časovni vrsti le naključni vplivi, dobimo po uporabi drsečih sredin ciklična gibanja (učinek Slutskega).
- Če so v časovni vrsti stalne periode, lahko drseče sredine zabrišejo oscilacije v celoti.
- V splošnem so drseče sredine lahko dober približek pravemu trendu.

Primer: Kot primer drsečih sredin vzemimo zopet brezposelne v Sloveniji. Izračunajmo tričlensko drsečo sredino:

T	X_k	tričl. drs. sred.
1981	12 315	–
1982	13 700	13 032
1983	15 781	14 030
1984	15 240	15 249
1985	15 300	14 710
1986	14 657	14 678
1987	14 102	15 184
1988	21 341	21 581
1989	28 218	31 262
1990	44 227	–

◇

Analitično določanje trenda

Trend lahko obravnavamo kot posebni primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend $X_T = f(T)$, lahko parametre trenda določimo z metoda najmanjših kvadratov $\sum_{i=1}^{n}(X_i - X_{iT})^2 = \min$. V primeru linearnega trenda

$$X_T = a + bT, \qquad \sum_{i=1}^{n}(X_i - a - bT_i)^2 = \min .$$

dobimo naslednjo oceno trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X})(T_i - \overline{T})}{\sum_{i=1}^{n}(T_i - \overline{T})^2}\,(T - \overline{T}).$$

Ponavadi je čas T transformiran tako, da je $t = 0$. Tedaj je ocena trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X})\cdot t_i}{\sum_{i=1}^{n} t_i^2}\, t.$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je

$$\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{iT})^2}.$$

Primer: Kot primer ocenimo število doktoratov znanosti v Sloveniji v razdobju od leta 1986 do 1990. Z linearnim trendom ocenimo koliko doktorjev znanosti je v letu 1991. Izračunajmo tudi standardno napako ocene. Izračunajmo najprej trend:

T	Y_i	t_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})t_i$	t_i^2
1986	89	-2	$-19{\cdot}8$	$39{\cdot}6$	4
1987	100	-1	$-8{\cdot}8$	$8{\cdot}8$	1
1988	118	0	$9{\cdot}2$	0	0
1989	116	1	$7{\cdot}2$	$7{\cdot}2$	1
1990	121	2	$12{\cdot}2$	$24{\cdot}4$	4
	544	0		80	10

$$\overline{Y} = \frac{544}{4} = 1088,$$

$$Y_T = 108{\cdot}8 + \frac{80}{10}\,t = 108{\cdot}8 + 8t,$$

$$Y_T(1991) = 108{\cdot}8 + 8 \times 3 = 132{\cdot}8.$$

Ocena za leto 1991 je približno 133 doktorjev znanosti.

Zdaj pa izračunajmo standardno napako ocene. Za vsako leto je potrebno najprej izračunat

T	Y_i	Y_{iT}	$Y_i - Y_{iT}$	$(Y_i - Y_{iT})^2$
1986	89	$92{\cdot}8$	$-3{\cdot}8$	$14{\cdot}14$
1987	100	$100{\cdot}8$	$-0{\cdot}8$	$0{\cdot}64$
1988	118	$108{\cdot}8$	$9{\cdot}2$	$84{\cdot}64$
1989	116	$116{\cdot}8$	$-0{\cdot}8$	$0{\cdot}64$
1990	121	$124{\cdot}8$	$-3{\cdot}8$	$14{\cdot}44$
	544	544	0	$114{\cdot}8$

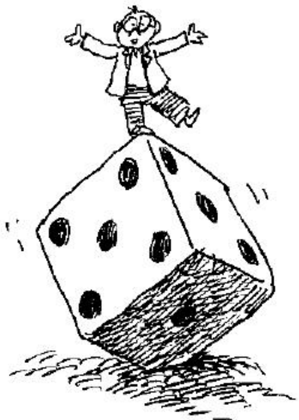
$$\sigma_e = \sqrt{\frac{114{\cdot}8}{5}} = 4{\cdot}8.$$

$\diamond$

Del III

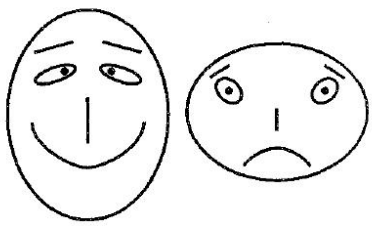
KAM NAPREJ

Osnovne principe in orodja, ki smo jih spoznali pri VIS, lahko posplošimo in razširimo do te mere, da se dajo z njimi rešiti tudi bolj kompleksni problemi.



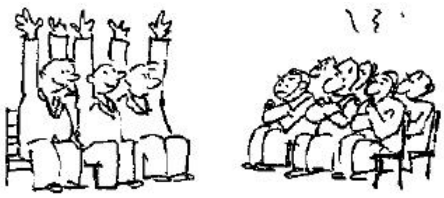
Spoznali smo kako predstaviti **eno** spremenljivko (dot-plot, histogrami,...) in **dve** spremenljivki (razsevni diagram). **Kako pa predstavimo več kot dve spremenljivki na ravnem listu papirja?** Med številnimi možnostmi moramo omeniti idejo **Hermana Chernoffa**, ki je uporabil človeški obraz, pri čemer je vsako lastnost povezal z eno spremenljivko. Oglejmo si Chernoffov obraz:

- X =naklon obrvi,
- Y =velikost oči,
- Z =dolžina nosu,
- T =dolžina ust,
- U =višino obraza, itd.



Multivariantna analiza

Širok izbor multivariantnih modelov nam omogoča analizo in ponazoritev n-razsežnih podatkov.
Združevalna/grozdna tehnika (ang. cluster technique): iskanje delitve populacije na homogene podskupine, npr. z analizo vzorcev senatorskih glasovanj v ZDA zaključimo, da *jug* in *zahod* tvorita dva različna grozda.

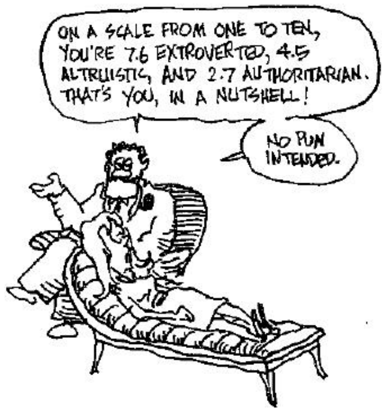


Diskriminacijska analiza

je obraten proces. Npr. odbor/komisija za sprejem novih študentov bi rad našel podatke, ki bi že vnaprej opozorili ali bodo prijavljeni kandidati nekega dne uspešno zaključili program (in finančno pomagali šoli - npr. z dobrodelnimi prispevki) ali pa ne bodo uspešni (gre delati dobro po svetu in šola nikoli več ne sliši zanj(o)).

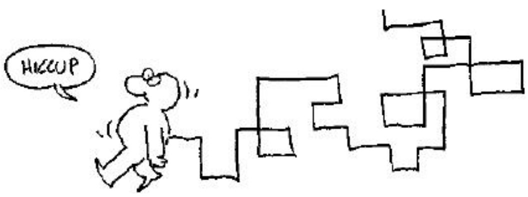
Analiza faktorjev

išče poenostavljeno razlago večrazsežnih podatkov z manjšo skupino spremenljivk. Npr. Psihiater lahko postavi 100 vprašanj, skrivoma pa pričakuje, da so odgovori odvisni samo od nekaterih faktorjev: ekstravertiranost, avtoritativnost, alutarizem, itd. Rezultate testa lahko potem povzamemo le z nekaterimi sestavljenimi rezultati v ustreznih dimenzijah.



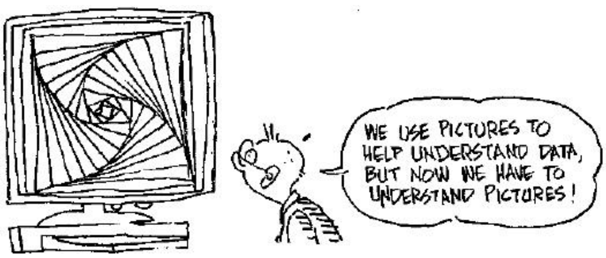
Naključni sprehodi

Pričnejo se z metom kovanca, recimo, da se pomaknemo korak nazaj, če pade grb, in korak naprej, če pade cifra. (z dvema kovancema se lahko gibljemo v 2-razsežnemu prostoru - tj. ravnini). Če postopek ponavljamo, pridemo do *stohastičnega procesa*, ki ga imenujemo naključni sprehod (ang. random walk). Modeli na osnovi naključnih sprehodov se uporabljajo za nakup/prodajo delnic in portfolio management.



Vizualizacija in analiza slik

Sliko lahko sestavlja 1 000 × 10 00 pikslov, ki so predstavljeni z eno izmed 16·7 milijonov barv. Statistična analiza slik želi najti nek pomen iz "informacije" kot je ta.



Ponovno vzorčenje

Pogosto ne moremo izračunati standardne napake in limite zaupanja. Takrat uporabimo tehniko ponovnega vzorčenja, ki tretira vzorec, kot bi bila celotna populacija. Za takšne tehnike uporabljamo pod imeni: randomization Jackknife, in Bootstrapping.

Kvaliteta podatkov

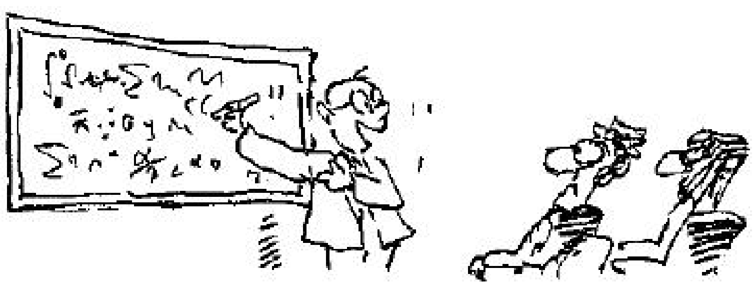
Navidezno majhne napake pri vzorčenju, merjenju, zapisovanju podatkov, lahko povzročijo katastrofalne učinke na vsako analizo. R. A. Fisher, genetik in ustanovitelj moderne statistike ni samo načrtoval in analiziral eksperimentalno rejo, pač pa je tudi čistil kletke in pazil na živali. Zavedal se je namreč, da bi izguba živali vplivala na rezultat. Moderni statistiki, z njihovimi računalniki in podatkovnimi bazami ter vladnimi projekti (beri denarjem) si pogosto ne umažejo rok.

Inovacija

Najboljše rešitve niso vedno v knjigah (no vsaj najti jih ni kar tako). Npr. mestni odpad je najel strokovnjake, da ocenijo kaj sestavljajo odpadki, le-ti pa so se znašli pred zanimivimi problemi, ki se jih ni dalo najti v standardnih učbenikih.

Komunikacija

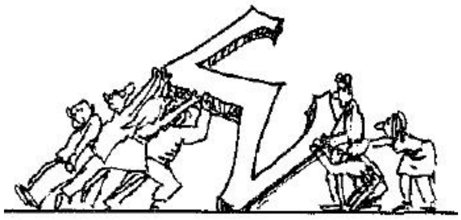
Še tako uspešna in bistroumna analiza je zelo malo vredna, če je ne znamo jasno predstaviti, vključujoč stopnjo statistične značilnosti.



Npr. v medijih danes veliko bolj natančno poročajo o velikosti napake pri svojih anketah.

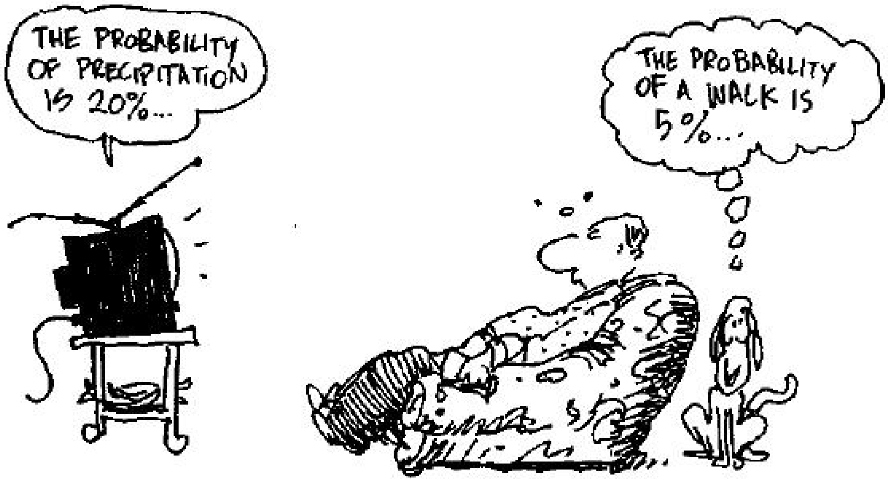
Timsko delo

V današnji kompleksni družbi. Reševanje številnih problemov zahteva *timsko delo*. Inženirji, statistiki in delavci sodelujejo, da bi izboljšali kvaliteto produktov. Biostatistiki, zdravniki, in AIDS-aktivisti združeno sestavljajo klinične poiskuse, ki bolj učinkovito ocenijo terapije.



Poglavje 17

Nekaj primerov uporabe verjetnosti



17.1 Zamenjalna šifra

Tomaž Pisanski, Skrivnostno sporočilo, *Presek* **V**/1, 1977/78, str. 40-42.

```
YHW?HD+CVODHVTHVO-!JVG:CDCYJ(JV/-V?HV(-T?HVW-4YC4(?-DJV/-(?S-VO3CWC%J(-V4-DC
V!CW-?CVNJDJVD-?+-VO3CWC%J(-VQW-DQ-VJ+V?HVDWHN-V3C:CODCV!H+?-DJVD-?+CV3JO-YC
```

(črko Č smo zamenjali s C, črko Ć pa z D). Imamo $26! = 403291461126605635584000000$ možnosti z direktnim preverjanjem, zato v članku dobimo naslednje nasvete:

(0) Relativna frekvenca črk in presledkov v slovenščini: presledek 173,

E	A	I	O	N	R	S	L	J	T	V	D	K	M	P	U	Z	B	G	Č	H	Š	C	Ž	F
89	84	74	73	57	44	43	39	37	37	33	30	29	27	26	18	17	15	12	12	9	9	6	6	1

(1) Na začetku besed so najpogostejše črke N, S, K, T, J, L.

(2) Najpogostejše končnice pa so E, A, I, O, U, R, N.

(3) Ugotovi, kateri znaki zagotovo predstavljajo samoglasnike in kateri soglasnike.

(4) V vsaki besedi je vsaj en samoglasnik ali samoglasniški R.

(5) V vsaki besedi z dvema črkama je ena črka samoglasnik, druga pa soglasnik.

(6) detektivska sreča

Pa začnimo z reševanem (oziroma kakor pravijo kriptografi: z razbijanjem):

```
(0)  V  -  C  D  J  ?  H  W  O  (  +  3  Y  4  !  /  Q  :  %  T  N  S  G
    23 19 16 12 11 10  9  7  6  6  5  4  4  3  3  2  2  2  2  2  2  1  1
```

Zaključek `V --> ' '` (drugi znaki z visoko frekvenco ne morejo biti). Dve besedi se ponovita: `O3CWC%J(-`, opazimo pa tudi eno sklanjatev: `D-?+-` ter `D-?+C`. Torej nadaljujemo z naslednjim tekstom:

```
YHW?HD+C ODH TH O-!J G:CDCYJ(J /- ?H (-T?H W-4YD4(?-DJ /-(?S- O3CWC%J(- 4-DC
!CW-?C NJDJ D-?+- O3CWC%J(- QW-DQ- J+ ?H DWHN- 3C:CODC !H+?-DJ D-?+C 3JO-YC
```

(3) Kanditati za samoglasnike e,a,i,o so znaki z visokimi frekvancami. Vzamemo:

$$\{e,a,i,o\} = \{-,C,J,H\}$$

(saj D izključi -,H,J,C in ? izključi -,H,C, znaki -,C,J,H pa se ne izključujejo)

Razporeditev teh znakov kot samoglasnikov izgleda prav verjetna. To potrdi tudi gostota končnic, gostota parov je namreč:

```
AV CV HV JV VO ?H -D DC JM W- DJ UC CW -? VD
7  5  5  5  4  4  4  3  3  3  3  3  3  3  3
```

(5) Preučimo besede z dvema črkama:

```
Samoglasnik na koncu                          Samoglasnik na začetku

1) da ga na pa ta za (ha ja la)               1) ar as             (ah aj au)
2) če je le me ne se še te ve že (he)      2) en ep             (ej eh)
3) bi ji ki mi ni si ti vi                    3) in iz ig
4) bo do (ho) jo ko no po so to               4) on ob od os on     (oh oj)
5) ju mu tu (bu)                              5) uk up uš ud um ur (uh ut)
6) rž rt
```

in opazujemo besedi: /- ?H ter besedi: J+ ?H. J+ ima najmanj možnosti, + pa verjetno ni črka n, zato nam ostane samo še:

```
J+ ?H        DWHN-
/- ?H
iz te       (ne gre zaradi: D-?+C)
ob ta(e,o)  (ne gre zaradi: D-?+C)
od te       (ne gre zaradi: D-?+C)
```

tako da bo potrebno nekaj spremeniti in preveriti še naslednje: on bo; on jo; in so; in se; in je; in ta; en je; od tu ...

(6) Če nam po dolgem premisleku ne uspe najti rdeče niti, bo morda potrebno iskati napako s prijatelji (tudi računalniški program z metodo lokalne optimizacije ni zmogel problema zaradi premajhne dolžine tajnopisa, vsekakor pa bi bilo problem mogoče rešiti s pomočjo elektronskega slovarja). Tudi psihološki pristop pomaga, je svetoval Martin Juvan in naloga je bila rešena (poskusite sami!).

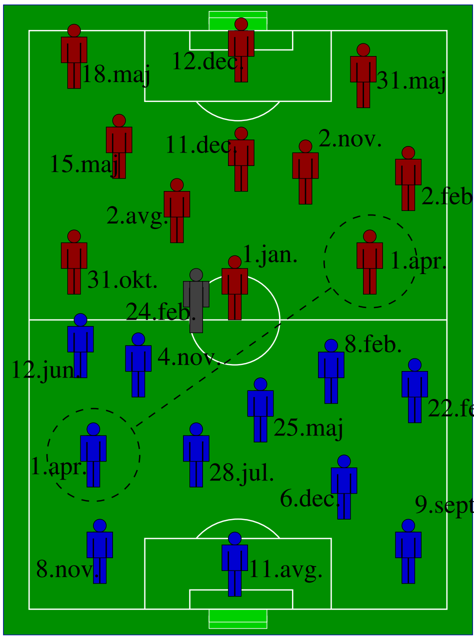
Kaj pa tuji jeziki

Podobna naloga je v angleščini dosti lažja, saj je v tem jeziku veliko členov THE, A in AN, vendar pa zato običajno najprej izpustimo presledke iz teksta, ki ga želimo spraviti v tajnopis. V angleščini imajo seveda črke drugačno gostoto kot v slovenščini. Razdelimo jih v naslednjih pet skupin:

1. E, z verjetnostjo okoli 0·120,
2. T, A, O, I, N, S, H, R, vse z verjetnostjo med 0·06 in 0·09,
3. D, L, obe z verjetnostjo okoli 0·04,
4. C, U, M, W, F, G, Y, P, B, vse z verjetnostjo med 0·015 in 0·028,
5. V, K, J, X, Q, Z, vse z verjetnostjo manjšo od 0·01.

Najbolj pogosti pari so (v padajočem zaporedju): TH, HE, IN, ER, AN, RE, ED, ON, ES, ST, EN, AT, TO, NT, HA, ND, OU, EA, NG, AS, OR, TI, IS, ET, IT, AR, TE, SE, HI in OF. Najbolj pogoste trojice pa so (v padajočem zaporedju): THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR in DTH.

17.2 Kakšno naključje!!! Mar res?

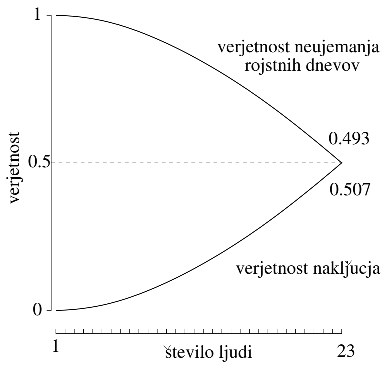


Na nogometni tekmi sta na igrišču dve enajsterici in sodnik, skupaj **23 oseb**.

Kakšna je verjetnost, da imata **dve osebi** isti rojstni dan?

Ali je ta verjetnost lahko večja od $0{\cdot}5$?

Ko vstopi v sobo k-ta oseba, je verjetnost, da je vseh k rojstnih dnevov različnih enaka:



$$\frac{365}{365}\times\frac{364}{365}\times\frac{363}{365}\times\cdots\times\frac{365-k+1}{365} = \begin{cases} 0{\cdot}493; & \text{če je } k=22 \\ 0{\cdot}507; & \text{če je } k=23 \end{cases}$$

V poljubni skupini 23-ih ljudi je verjetnost,
da imata vsaj dva skupni rojstni dan $> 1/2$.

Čeprav je 23 majhno število, je med 23 osebami 253 različnih parov. To število je veliko bolj povezano z iskano verjetnostjo. Testirajte to na zabavah z več kot 23 osebami. Organizirajte stave in dolgoročno boste gotovo na boljšem, na velikih zabavah pa boste zlahka zmagovali.

Napad s pomočjo paradoksa rojstnih dnevov (angl. *Birthday Attack*)

To seveda ni paradoks, a vseeno ponavadi zavede naš občutek.

Ocenimo še splošno verjetnost. Mečemo k žogic v n posod in gledamo, ali sta v kakšni posodi vsaj dve žogici. Poiščimo spodnjo mejo za verjetnost zgoraj opisanega dogodka:

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) = \prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right)$$

Iz Taylorjeve vrste

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$

ocenimo $1 - x \approx e^{-x}$ in dobimo

$$\prod_{i=1}^{k-1}\left(1 - \frac{i}{n}\right) \approx \prod_{i=1}^{k-1} e^{\frac{-i}{n}} = e^{\frac{-k(k-1)}{2n}}.$$

Torej je verjetnost trčenja $1 - e^{-k(k-1)/(2n)}$. Potem velja $e^{-k(k-1)/(2n)} \approx 1 - \varepsilon$ oziroma $-k(k-1)/(2n) \approx \log(1-\varepsilon)$, tj. $k^2 - k \approx -2n\log(1-\varepsilon)$ in če ignoriramo $-k$, dobimo končno

$$\boxed{k \approx \sqrt{2n \log \frac{1}{1-\varepsilon}}\ .}$$

Za $\varepsilon = 0,5$ je

$$k \approx 1{\cdot}17\sqrt{n},$$

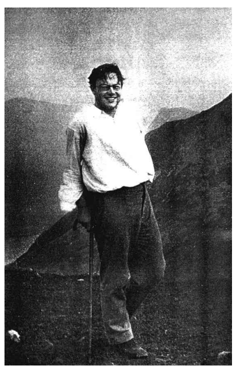
kar pomeni, da, če zgostimo nekaj več kot $\sqrt{n}$ elementov, je bolj verjetno, da pride do trčenja kot da ne pride do trčenja. V splošnem je k proporcionalen s $\sqrt{n}$.

Raba v kriptografiji

Napad s pomočjo paradoksa rojstnih dnevov s tem določi spodnjo mejo za velikost zaloge vrednosti zgoščevalnih funkcij, ki jih uporabljamo v kriptografiji in računalniški varnosti. 40-bitna zgostitev ne bi bila varna, saj bi prišli do trčenja z nekaj več kot 2^{20} (se pravi milijon) naključnimi zgostitvami z verjetnostjo vsaj 1/2. V praksi je priporočena najmanj 128-bitna zgostitev in standard za shema digitalnega podpisa (160 bitov) to vsekakor upošteva. Podobno si lahko pomagamo tudi pri napadih na DLP in še kje.

17.3 Ramseyjeva teorija

- intuitivna ideja
- Ramseyjev izrek
- Erdősev izrek
- primeri uporabe



Po 3,500 let starem zapisu je antični sumerski učenjak pogledal v nebo in zagledal leva, bika in škorpijona. **Ali gre za kozmične sile?** Astronom bi rekel: kolekcija zvezd, tj. začasna konfiguracija zvezd, ki jo gledamo z roba navadne galaksije.

1928 Frank Plumpton Ramsey (26 let, angleški matematik, filozof in ekonomist)

Popoln nered je nemogoč.

Ramseyjeva teorija: **Vsaka dovolj velika struktura vsebuje urejeno podstrukturo.**
Konkretna naloga: **Koliko objektov nam zagotavlja željeno podstrukturo?**

Izrek (SIM). *V družbi šestih ljudi obstaja trojica v kateri se vsaka dva poznata ali pa vsaka dva ne poznata.*

– naivni prostop: preverimo $2^{15} = 32.768$ možnosti,

– barvanje povezav polnega grafa K_6 in Dirichletov princip.

Nekaj težja naloga: V družbi 17ih znanstvenikov se vsaka dva dopisujeta o eni izmed treh tem. Dokaži, da obstajajo trije, ki se dopisujejo o isti temi!

Ramseyjevo število $r(k,\ell)$ je najmanjše število za katerega vsak graf na $r(k,\ell)$ vozliščih vsebuje bodisi k-kliko bodisi ℓ-antikliko. Prepričaj se, da je $r(k,\ell) = r(\ell,k)$.

Primeri: $r(k,1) = 1 = r(1,\ell)$, $r(2,\ell) = \ell$, $r(k,2) = k$, SIM: $r(3,3) \leqslant 6$.

Ramseyjev izrek. $\forall k,\ell \in \mathbb{N}$

$$r(k,\ell) \leqslant r(k,\ell-1) + r(k-1,\ell).$$

Če sta obe števili na desni strani neenakosti sodi, potem velja stroga neenakost.

Zgled uporabe: $r(3,3) \leqslant r(3,2) + r(2,3) = 3 + 3 = 6$.

Dokaz: (1935 Erdős & Szekeres, 1955 Greenwood & Gleason) Naj bo G graf na $r(k,\ell-1) + r(k-1,\ell)$ vozliščih. Potem velja ena izmed naslednjih možnosti:

(a) Vozlišče v ni sosednje množici S z vsaj $r(k,\ell-1)$ vozlišči.

kar pomeni, da $G[S]$ vsebuje ali k-kliko ali $(\ell-1)$-antikliko.

(b) Vozlišče v je sosednje množici T z vsaj $r(k-1,\ell)$ vozlišči.

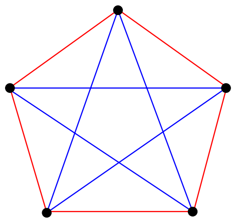
kar pomeni, da $G[T]$ vsebuje ali $(k-1)$-kliko ali ℓ-antikliko.

Od tod sledi, da G vsebuje bodisi k-kliko bodisi ℓ-antikliko. Naj bosta $r(k,\ell-1)$ in $r(k-1,\ell)$ sodi števili in $|G| = r(k,\ell-1) + r(k-1,\ell) - 1$. Potem obstaja vozlišče $v \in V(G)$, katerega stopnja je sodo število. Torej v ni soseden točno $r(k-1,\ell) - 1$ vozliščem in velja bodisi (a) bodisi (b). $\square$

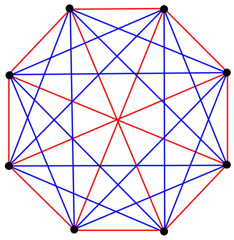
Pokaži:

$$r(3,4) \leqslant 9, \quad r(3,5) \leqslant 14, \quad r(4,4) \leqslant 18, \quad r(k,\ell) \leqslant \binom{k+\ell-2}{k-1}.$$

To je bila zgornja meja. Kaj pa spodnja meja?



$5 < r(3,3) = 6$



$8 < r(3,4) = 9$

Podobno dobimo tudi $13 < r(3,5) = 14$, $17 < r(3,6) = 18$,

$22 < r(3,7) = 23$, $27 < r(3,8) \leqslant 29$ in $35 < r(3,9) = 36$.

Erdősev Izrek. $\forall k \in \mathbb{N} \quad r(k,k) \geqslant 2^{k/2}$.

Zgled uporabe: $r(3,3) \geqslant 3$ and $r(4,4) \geqslant 4$.

Če Marsovci napadejo Zemljo nam morda uspe izračunati $r(5,5) \in [43,49]$ (Exoo 1989, McKay and Radziszowski 1995), nikakor pa ne moremo izračunati $r(6,6) \in [102,165]$ (Kalbfleisch 1965, Mackey 1994).

Znana Ramseyeva števila:

$k\backslash\ell$	3	4	5	6	7	8	9	10			
3	6	9	14	18	23	28	36	?			
4	9	18	25	?	?	?	?	?	?	?	?
6	18	?	?	?	?	?	?	?			

Ester Klein je leta 1933 predstavil naslednjo geometrijsko nalogo:

Med petimi točkami v ravnini, od katerih nobene tri niso kolinearne (ležijo na premici), lahko vedno izberemo štiri, ki določajo konveksen četverokotnik.

Rešitev: Vpeljemo pojem konveksne ogrinjače ...

Če je konveksna ogrinjača teh petih točk

(a) petkotnik, potem vsake 4 točke med njimi sestavljajo konveksen četverokotnik,

(b) štirikotnik, potem so njegovi vrhovi tiste 4 točke, ki smo jih iskali,

(c) trikotnik, potem ga lahko označimo z A, B in C, preostali točki pa z D in E, tako da sta točki A in B na isti s strani premice DE. V tem primeru je četverokotnik $ABCD$ konveksen. $\square$

Nalogo lahko posplošimo na 9 točk in iskanje konveksnega petkotnika ter počasi pridemo do Erdőseve domneve, da za konveksen k-kotnik potrebujemo v ravnini vsaj

$$n = 1 + 2^{k-2}$$

točk od katerih nobene 3 niso kolinearne. Pravzaprav se je najprej Szekeres prepričal, da za dovolj velik n vedno obstaja konveksen k-kotnik, potem pa je Erdős postavil svojo domnevo.

Erdőseva probabilistična metoda (1947)

34 točk določa 561 premic. Da se to zgodi v eni barvi, je verjetnost

$$2^{-561} \approx 2{\cdot}6 \times 10^{-169}.$$

Velja tudi $\binom{1.000.000}{34} = 3{\cdot}4 \times 10^{165}$. Torej lahko pričakujemo $\binom{10^6}{34} = 3{\cdot}4 \times 10^{165} \approx 0{\cdot}01$ oziroma $0{\cdot}01\%$ enobarvnih. To pomeni, da v $99{\cdot}9\%$ ne dobimo enobarvnega K_{34}.

Slednjo idejo pretvorimo v Erdősev dokaz.

Dokaz Erdősevega izreka: Probabilistična metoda (ni konstruktivna) in štetje. Naj bo $\mathcal{G}_n$ množica grafov z vozlišči $v_1, v_2, \dots, v_n$. Naj bo $\mathcal{G}_n^k$ množica grafov iz $\mathcal{G}_n$, ki vsebujejo k-kliko. Potem je $|\mathcal{G}_n| = 2^{\binom{n}{2}}$, $\quad |\mathcal{G}_n^k| = 2^{\binom{n}{2}-\binom{k}{2}}\binom{n}{k}$ in

$$q \quad = \quad |\mathcal{G}_n^k|/|\mathcal{G}_n| \quad \leqslant \quad \frac{n^k\, 2^{-\binom{k}{2}}}{k!}.$$

Če je $n < 2^{k/2}$, $\quad$ velja $\quad q \leqslant \dfrac{2^{\frac{k^2}{2}-\binom{k}{2}}}{k!} < \dfrac{1}{2}$.

Se pravi, da manj kot polovica grafov iz $\mathcal{G}_n$ vsebuje k-klike.

Iz $\mathcal{G}_n = \{G \,|\, \overline{G} \in \mathcal{G}_n\}$ pa sledi, da manj kot polovica grafov iz $\mathcal{G}_n$ vsebuje k-antiklike. $\square$

Posledica. *Za* $m := \min(k, \ell)$ *velja* $\quad r(k, \ell) \geqslant 2^{m/2}$.

Uporaba: Pobarvaj z modro in rdečo števila 1 2 3 4 5 6 7 8 9.

Posledica Ramseyjevega izreka (Waerden 1926):
3 rdeča ali 3 modra števila tvorijo aritmetično zaporedje.

$\texttt{PODVOJ}(x) = 2x$, $\quad \texttt{EKSPONENT}(x) = 2^x$, $\quad \texttt{STOLP}(x) = 2^{2^{2^{\cdots^{2}}}}$ $\quad$ (x dvojk)

$\texttt{UAU}(1) = \texttt{STOLP}(1)=2$. $\quad \texttt{UAU}(2) = \texttt{STOLP}(2)=4$. $\quad \texttt{UAU}(3) = \texttt{STOLP}(4)=65{,}536$

$\texttt{UAU}(4)$ = prevelik za vse knjige, za vse računalnike ..., $\quad \texttt{UAU}(x) = \dots$.

Zaporedje $1, 2, \dots, \texttt{ACKERMANN}(k)$ pobarvamo z dvema barvama.
Potem obstaja monokromatično (enobarvno) aritmetično podzaporedje s k členi.

17.4 Teorija kodiranja

Claude Shannon je postavil teoretične osnove **teorije informacij** in zanesljivega prenosa digitalnih podatkov kmalu po koncu druge svetovne vojne.

Glavni mejniki teorija kodiranja

1947-48: začetki teorije informacij: znamenita izreka o "**Source Coding**" in pa "**Channel Capacity**" (C. Shannon)

1949-50: odkritje *prvih kod* za odpravljanje napak (M. Golay, R. Hamming).

1959-60: odkritje **BCH-kod** (R. Bose, D. Ray-Chaudhuri, A. Hochquenghem).

1967: Viterby algoritm za odkodiranje **konvolucijskih kod**, (ki sta jih predlagala Elias 1955, Hagelbarger 1959).

1993: razvoj **turbo kod** (C. Berrou, A. Glavieux, P. Titimajshima).

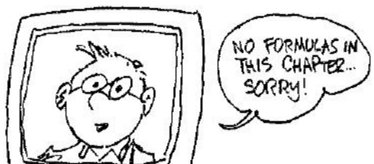
Poglavje 18

Uporaba statistike

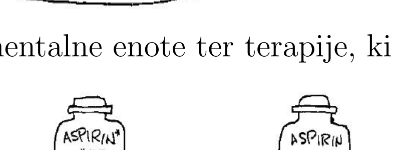
(Testiranje PRNG, Teorija informacij in entropija)

18.1 Načrtovanje eksperimentov

Načrtovanje eksperimentov se pogosto neposredno prevede v uspeh oziroma neuspeh. V primeru parjenja lahko statistik spremeni svojo vlogo iz pasivne v aktivno. Predstavimo samo osnovne ideje, podrobno numerično analizo pa prepustimo statistični programski opremi.



Elementi načrta so eksperimentalne enote ter terapije, ki jih želimo uporabiti na enotah.



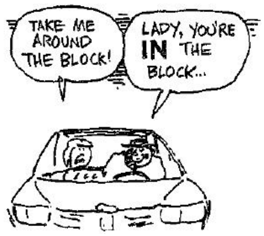
- medicina: bolniki (enote) in zdravila (terapije),
- optimizacija porabe: taxi-ji (enote) in različne vrste goriva (terapije),
- agronomija: območja na polju in različne vrste kulture, gnojiva, špricanja,...

Danes uporabljamo ideje načrtovanja eksperimentov na številnih področjih:

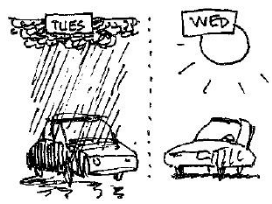
- optimizacija industrijskih procesov,
- medicina,
- sociologija.

Na primeru bomo predstavili tri osnovne principe načrtovanja eksperimentov:

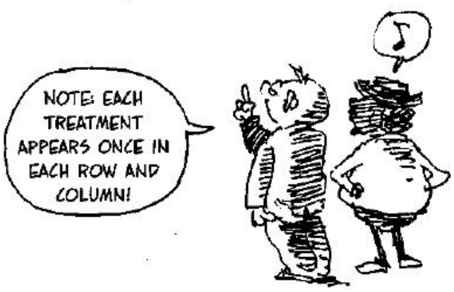
1. Ponavljanje: enake terapije pridružimo različnim enotam, saj ni mogoče oceniti naravno spremenljivost (ang. natural variability) in napake pri merjenju.
2. Lokalna kontrola pomeni vsako metodo, ki zmanjša naravno spremenljivost.
En od načinov grupira podobne enote eksperimentov v **bloke**. V primeru taxijev uporabimo obe vrsti goriva na vsakem avtomobilu in rečemo, da je avto blok.



3. Naključna izbira je bistven korak povsod v statistiki! Terapije za enote izbiramo naključno. Za vsak taksi izberemo vrsto goriva za torek oziroma sredo z metom kovanca. Če tega ne bi storili, bi lahko razlika med torkom in sredo vplivala na rezultate.



		DAY 1	2	3	4
CAB	1	a	b	c	d
	2	b	c	d	a
	3	c	d	a	b
	4	d	a	b	c

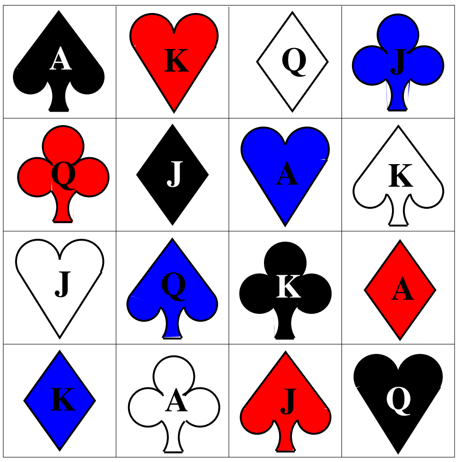


Latinski kvadrati

Latinski kvadrat reda v je $v \times v$-razsežna matrika, v kateri vsi simboli iz množice

$$\{1, \dots, v\}$$

nastopajo v vsaki vrstici in vsakem stolpcu.



Trije paroma ortogonalni latinski kvadrati reda 4, tj. vsak par znak-črka ali črka-barva ali barva-znak se pojavi natanko enkrat.

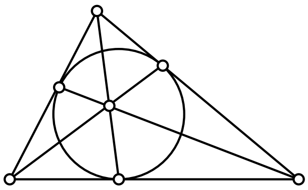
Projektivni prostor $PG(d, q)$ (razsešnosti d nad q) dobimo iz vektorskega prostora $[GF(q)]^{d+1}$, tako da naredimo kvocient po 1-razsežnih podprostorih.

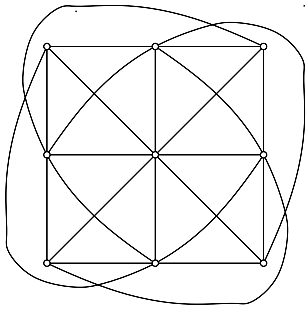
Projektivna ravnina $PG(2, q)$ je incidenčna struktura z 1- in 2-dim. podprostori prostora $[GF(q)]^3$ kot **točkami** in **premicami**, kjer je "$\subset$" incidenčna relacija. To je 2-$(q^2+q+1, q+1, 1)$-design, tj.,

- $v = q^2 + q + 1$ je število točk (in število premic b),
- vsaka premica ima $k = q + 1$ točk (in skozi vsako točko gre $r = q + 1$ premic),
- vsak par točk leži na $\lambda = 1$ primicah (in vsaki premici se sekata v natanko eno točki).

Primeri:

1. Projektivno ravnino $PG(2, 2)$ imenujemo **Fano ravnina** (7 točk in 7 premic).





2. $PG(2,3)$ lahko skonstruiramo iz 3×3 mreže oziroma afine ravnine $AG(2,3)$.

Bose in Shrikhande

Prva sta konec tridesetih let prejšnjega stoletja vpeljala **asociativne sheme** **Bose** in **Nair** a potrebe statistike.

Toda **Delsarte** je pokazal, da nam lahko služijo kot povezava med številnimi področji matematike, naprimer teorijo kodiranja in teorijo načrtov.

Literatura

[1] A. Ferligoj: *Osnove statistike na prosojnicah.* Samozaložba, Ljubljana 1995.

[2] L. Gonick in W. Smith, *The Cartoon guide to Statistics*, 1993.

[3] M. Hladnik: *Verjetnost in statistika.* Založba FE in FRI, Ljubljana 2002.

[4] W. Mendenhall in T. Sincich, *Statistics for engineering and the sciences*, 4th edition, Prentice Hall, 1995.

[5] D. S. Moore (Purdue University), *Statistika: znanost o podatkih*
(5. izdaja prevedena v slovenščino leta 2007).

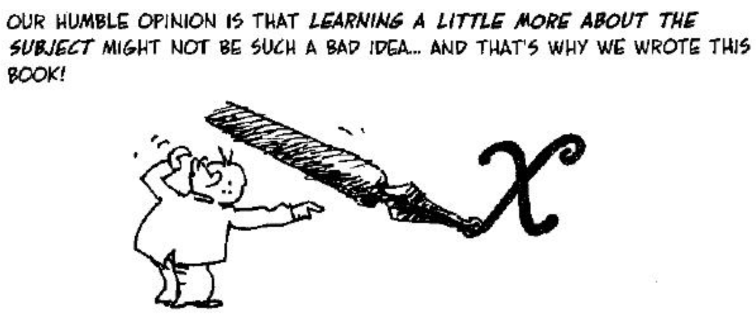
Obstaja obilna literatura na spletu in v knjižnicah.

Gradiva bodo dosegljiva preko internetne učilnice (moodle).

Pri delu z dejanskimi podatki se bomo v glavnem naslonili na prosti statistični program `R`.

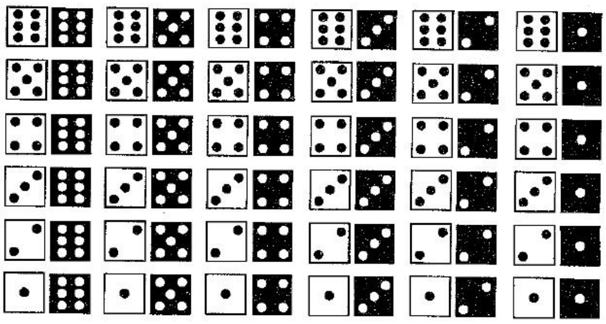
Program je prosto dostopen na: http://www.r-project.org/

Proti koncu semestra pa morda tudi Minitab.



Dodatek A

MATEMATIČNE OSNOVE (ponovitev)



A.1 Računala nove dobe

Ste že kdaj razmišljali o računanju (aritmetiki), ki ga uporabljamo v vsakdanjem življenju? Večina ljudi jo zamenjuje kar za celotno matematiko. Na kakšen način računajo računalniki ter ostale digitalne naprave (digit je angl. beseda za število), ki nas obkrožajo v času informacijske dobe? Nekateri se sicer skušajo prilagajati našemu načinu računanja, vse več pa je takih, ki so jim časovna in prostorska učinkovitost ter preciznost ključnega pomena. Take naprave računajo na malce drugačen način. V tem razdelku se bomo poskusili s pomočjo osnovnošolskega računanja približati računalom, ki jih preko številnih naprav, kot so osebni računalniki, diskmani in pametne kartice, uporabljamo v vsakdanji praksi.

Poleg seštevanja in množenja pa se v prvih razredih osnovne šole naučimo tudi odštevati in deliti. Seveda začnemo najprej odštevati manjša števila od večjih. Če želimo izračunati $a - b$ in je $a \geqslant b$, se lahko vprašamo

$$b \text{ plus koliko je } a?$$

Šele nekoliko kasneje se naučimo, da moramo v primeru, ko želimo odšteti večje število od manjšega, števili najprej zamenjati, na koncu pa dobljeni razliki spremeniti predznak. Zaradi tega smo povečali množico naravnih števil $\mathbb{N} = \{1, 2, \dots\}$ do množice celih števil $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$.

Deljenje ni tako preprosto. Če želimo a deliti z b, se lahko prav tako kot prej vprašamo "b krat koliko je a?" Vendar se pogosto zgodi, da število a sploh ni deljivo s številom b. Množico števil lahko sicer povečamo do množice ulomkov $\mathbb{Q}$, kjer se da deliti s poljubnim od nič različnim številom, a potem nastopijo druge težave. Najdemo lahko različne ulomke, ki so si poljubno blizu, tudi tako blizu, da jih računalnik ne more več ločiti. Ker pa si želimo, da bi se računalniki čim manj motili, se vprašajmo po množicah, v katerih bi lahko brez problemov tudi delili, po možnosti na enak način kot znamo odštevati. Pravzaprav se je potrebno vprašati, na katera pravila se želimo pri računanju opreti. Naštejmo jih nekaj.

1. Običajno je prvo pravilo *zaprtost*, rezultat, ki ga dobimo po opravljeni operaciji med dvema številoma, je tudi v množici, iz katere smo izbrali števili. Množica naravnih števil je zaprta za seštevanje in množenje, saj v tabelah 1a in 1b nastopajo samo naravna števila. Ni pa množica naravnih števil zaprta za odštevanje. To lastnost ima na primer množica celih števil.

2. V množici celih števil igra pomembno vlogo število 0; pa ne samo zato, ker loči pozitivna števila od negativnih, pač pa tudi zato, ker se nobeno število s prištevanjem števila 0, ne spremeni. Tudi pri množenju najdemo nekaj podobnega. Če pomnožimo katerokoli od nič različno število z 1, dobimo zopet isto število. Takemu številu pravimo *nevtralni element* ali pa tudi *enota* za ustrezno operacijo.

3. V množici celih števil sta poljubni števili $-a$ in a povezani z enoto za seštevanje na naslednji način: $a + (-a) = 0$. Pravimo, da je $-a$ *nasprotni* element števila a. Celo število b je *obratni element* celega števila a, če je $ab = 1$. Od tod sledi $a = b = 1$, tj. v množici celih števil imata le števili 1 in -1 obratni element.

4. Če si izberemo poljubna števila a, b in c, potem velja $a+(b+c) = (a+b)+c$ in $a(bc) = (ab)c$. O drugi enakosti se lahko prepričamo z računanjem prostornine kvadra s stranicami a, b in c. Tem lastnostim pravimo *zakon o združevanju* za seštevanje oziroma za množenje (ali tudi *asociativnost*). Le-ta nam pove, da je vseeno, ali začnemo računati z leve ali z desne. To seveda ne drži za odštevanje ali deljenje.

Če v neki množici G z binarno (dvočleno) operacijo $\circ$, tj. operacijo, ki vsakemu urejenemu paru elementov iz G priredi natanko določen element, veljajo naslednja pravila:

(G1) za vsaka $a, b \in G$ je $\boxed{a \circ b \in G}$,

(G2) obstaja tak element $e \in G$, da za vsak $g \in G$ velja $\boxed{e \circ g = g \circ e = g}$,

(G3) za vsak element $g \in G$ obstaja tak $f \in G$, da je $\boxed{g \circ f = f \circ g = e}$,

(G4) za vse $a, b, c \in G$ velja $\boxed{(a \circ b) \circ c = a \circ (b \circ c)}$,

potem pravimo, da je par $(G, \circ)$ **grupa**. Elementu e pravimo **enota** grupe, elementu f pa **inverz** elementa g. Množica celih števil je grupa za seštevanje, ni pa grupa za množenje, saj ni izpolnjeno pravilo (G3) (le 1 in -1 imata inverzni element za množenje).

Morda bo kdo pomislil, da je prišla definicija grupe iz glave enega samega matematika, pa temu sploh ni tako. Matematiki so potrebovali več kot 100 let trdega dela, da so končno (eksplicitno) zapisali zgornja pravila (*aksiome*). *Joseph Louis Lagrange* (1736-1813) je leta 1771 postavil prvi pomembnejši izrek. *Augustin Louis Cauchy* (1789-1857) je študiral grupe permutacij, medtem, ko je *Niels Henrik Abel* (1802-1829) s teorijo grup pokazal, da enačba 5. stopnje ni rešljiva z radikali (tj. rešitve ne znamo zapisati s formulami kot v primeru enačb nižjih stopenj). Po njem pravimo grupam, v katerih velja pravilo zamenjave, tudi *Abelove grupe* (ali komutativne grupe). Pravi pionir abstraktnega pristopa pa je bil *Evariste Galois* (1811-1832), ki je leta 1823 prvi uporabil besedo "grupa". Proces poudarka na strukturi se je nadaljeval vse do leta 1854, ko je *Arthur Cayley* (1821-1895) pokazal, da je grupo moč definirati ne glede na konkretno naravo njenih elementov.

Galois je vpeljal tudi naslednji pojem. Če za neko množico $\mathcal{O}$ z binarnima operacijama, ki ju bomo označili s $+$ in $*$ (četudi ne predstavljata nujno običajnega seštevanja in množenja), velja

(O1) par $(\mathcal{O}, +)$ je grupa z enoto 0,

(O2) par $(\mathcal{O}\backslash\{0\}, *)$ je grupa z enoto 1,

(O3) za vse $a, b, c \in \mathcal{O}$ je $\boxed{a*(b+c) = a*b+b*c}$ in $\boxed{(b+c)*a = b*a+c*a}$,

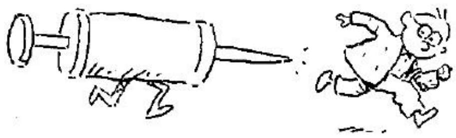
potem imenujemo trojico $(\mathcal{O}, +, *)$ **obseg**. Množica ulomkov z običajnim seštevanjem in množenjem je primer obsega. O lastnosti (O3), ki jo imenujemo *zakon o razčlenjevanju* oziroma *distributivnost*, se lahko prepričamo z računanjem površine pravokotnika s stranicama a in $b + c$.

Primer: Za cilj postavimo iskanje obsega s končno mnogo elementi, v katerem bo računanje v nekem smislu še udobnejše kot v obsegih, ki jih srečamo v osnovni ali srednji šoli (racionalna števila $\mathbb{Q}$, realna števila $\mathbb{R}$ ali celo kompleksna števila $\mathbb{C}$).

Gotovo ste hitro ugotovili, da mora imeti grupa zaradi aksioma (G2) vsaj en element, enoto e namreč, obseg pa vsaj dva, enoto za operacijo "+" in enoto za operacijo "$*$". Potem se ni več težko prepričati, da je en element v primeru grupe že dovolj, saj nam $e \circ e = e$ zadovolji vse aksiome (G1)-(G4). V primeru obsega z dvema elementoma je enako z multiplikativno grupo: $1 * 1 = 1$ zadovolji aksiom (O2). Ne pozabite, da operaciji "+" in "$*$" ne predstavljata (nujno) običajnega seštevanja in množenja. V tem sestavku bomo spoznali kar nekaj takih obsegov. V vsakem obsegu je produkt poljubnega elementa a z aditivno enoto 0 enak 0, saj je $0*a = (0+0)*a = 0*a+0*a$ (upoštevali smo (G2) in (O3)) in po krajšanju z $0*a$ res dobimo $0*a = 0$. Opozoriti je treba, da *pravilo krajšanja* velja v poljubni grupi, saj v resnici na obeh straneh "dodamo" inverzni element in nato upoštevamo zakon o združevanju (G4) ter (G2) (do tega ste se gotovo dokopali že sami pri reševanju 3. naloge iz prejšnjega razdelka). Seveda velja enako tudi, kadar vrstni red zamenjamo: $a * 0 = 0$. Torej tudi v primeru najmanjšega obsega velja $0 * 0 = 0$ in $0 * 1 = 0 = 1 * 0$, kjer je 1 multiplikativna enota. Kako pa je z grupo, ki ima dva elementa, npr. enoto e in a? Poleg $e \circ e = e$ in $e \circ a = a = a \circ e$ mora zaradi aksioma (G3), pravila krajšanja in $e \neq a$ veljati še $a \circ a = e$ in že so izpolnjeni vsi aksiomi (G1)-(G4). Torej velja za obseg z dvema elementoma in pravkar odkrito aditivno grupo tudi aksiom (O1). Zlahka preverimo še (O3) in že smo ugnali tudi najmanjši obseg. ◇

A.2 Funkcije/preslikave

Funkcija f iz množice A v množico B je predpis, ki vsakemu elementu iz množice A priredi natanko določen element iz množice B, oznaka $f : A \longrightarrow B$.



Funkcija $f : A \longrightarrow B$ je:

- **injektivna** (angl. one to one) če za $\forall\, x, y \in A \quad x \neq y \ \Rightarrow\ f(x) \neq f(y)$,
- **surjektivna** (angl. on to), če za $\forall\, b \in B \quad \exists\, a \in A$, tako da je $f(a) = b$.

Injektivni in surjektivni funkciji pravimo **bijekcija**. Množicama med katerima obstaja bijekcija pravimo **bijektivni** množici. Bijektivni množici imata enako število elementov (npr. končno, števno neskončno, itd).

Trditev A.1. *Če sta množici A in B končni ter je $f : A \longrightarrow B$ funkcija, iz injektivnosti funkcije f sledi surjektivnost, in obratno, iz surjektivnosti funkcije f sledi injektivnost.* □

A.3 Permutacije

Permutacija elementov $1, \ldots, n$ je bijekcija, ki slika iz množice $\{1, \ldots, n\}$ v množico $\{1, \ldots, n\}$. Npr. permutacija kart je običajno premešanje kart (spremeni se vrstni red, karte pa ostanejo iste). Število permutacij n elementov, tj. razvrstitev n-tih različnih elementov, je enako $n! := 1 \cdot 2 \cdot \ldots \cdot n$ (oziroma definirano rekurzivno $n! = (n-1)!n$ in $0! = 1$). Permutacijo lahko opišemo z zapisom:

$$\pi = \begin{pmatrix} 1 & 2 & \ldots & n \\ a_1 & a_2 & \ldots & a_n \end{pmatrix},$$

kjer je $\{1, 2, \ldots, n\} = \{a_1, a_2, \ldots, a_n\}$. To pomeni $\pi(1) = a_1$, $\pi(2) = a_2, \ldots,$ $\pi(n) = a_n$.

Primer: $n = 11$,

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix}$$ ◇

Naj bo A neka množica. Permutacije množice A med seboj množimo po naslednjem pravilu: $\pi = \pi_2 \circ \pi_1$ je permutacija množice A, ki preslika $a \in A$ v $\pi_2(\pi_1(a))$.

Primer:

$$\begin{aligned} \pi_1 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix} \\ \pi_2 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 8 & 2 & 1 & 3 & 10 & 9 & 4 & 5 & 7 & 6 & 11 \end{pmatrix} \end{aligned}$$

Potem je

$$\pi = \pi_2 \circ \pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 1 & 3 & 10 & 6 & 2 & 8 & 4 & 7 & 11 & 9 & 5 \end{pmatrix}$$ ◇

Cikel je permutacija, za katero je

$$\pi(a_1) = a_2,\ \pi(a_2) = a_3, \quad \ldots, \pi(a_r) = a_1,$$

ostale elementi pa so fiksni (tj. $\pi(a) = a$). Na kratko jo zapišemo z $(a_1\, a_2 \ldots\, a_r)$.

Trditev A.2. *Vsako permutacijo lahko zapišemo kot produkt disjunktnih ciklov.* □

Primer:

$$\begin{aligned} \pi_1 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix} \\ \pi_2 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 8 & 2 & 1 & 3 & 10 & 9 & 4 & 5 & 7 & 6 & 11 \end{pmatrix} \\ \pi &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 1 & 3 & 10 & 6 & 2 & 8 & 4 & 7 & 11 & 9 & 5 \end{pmatrix} \end{aligned}$$

Potem je

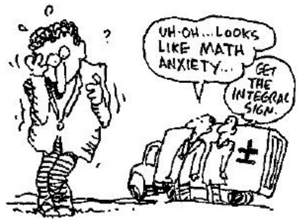
$$\pi_1 = (1\,3\,5\,2\,4\,10\,6)\,(8\,9\,11), \qquad \pi_2 = (1\,8\,5\,10\,6\,9\,7\,4\,3), \qquad \pi = (2\,3\,10\,9\,11\,5\,2)\,(4\,6\,8\,7)$$ ◇

Transpozicija je cikel dolžine 2. Vsak cikel pa je produkt transpozicij:

$$(a_1\, a_2\, a_3\, \ldots\, a_r) = (a_{r-1}\, a_r) \circ \cdots \circ (a_2\, a_3) \circ (a_1\, a_2),$$

torej je tudi vsaka permutacija produkt transpozicij. Seveda ta produkt ni nujno enolično določen, vseeno pa velja:

Trditev A.3. *Nobena permutacija se ne da zapisati kot produkt sodega števila in kot produkt lihega števila permutacij.*



Dokaz. Naj bodo $x_1, x_2, \dots, x_n$ različna realna števila. Poglejmo si produkt:

$$P = \prod_{i<j} (x_i - x_j).$$

Izberimo indeksa a in b, $a<b$, in poglejmo v katerih razlikah se pojavita:

$x_1 - x_a, \dots, x_{a-1} - x_a,$		$x_a - x_{a+1}, \dots, x_a - x_{b-1},$	$x_a - x_b,$	$x_a - x_{b+1}, \dots, x_a - x_n,$
$x_1 - x_b, \dots, x_{a-1} - x_b,$	$x_a - x_b,$	$x_{a+1} - x_b, \dots, x_{b-1} - x_b,$		$x_b - x_{b+1}, \dots, x_b - x_n.$

Razliko $x_a - x_b$ smo navedli dvakrat, a se v produktu P pojavi samo enkrat. Če na množici indeksov opravimo transpozicijo $(a\, b)$, razlika $x_a - x_b$ preide v razliko $x_b - x_a$, torej zamenja predznak, razlike iz prvega in zadnjega stolpca se med seboj zamenjajo, razlike iz srednjega stolpca pa tudi zamenjajo predznake (vendar je le-teh sodo mnogo in zato ne vplivajo na produkt P). Sedaj pa napravimo na množici indeksov permutacijo π. V tem primeru je produkt

$$P_\pi = \prod_{i<j} (x_{\pi(i)} - x_{\pi(j)}).$$

enak $\pm P$. Če uporabimo sodo število transpozicij, potem je $P_\pi = P$, sicer pa $P_\pi = -P$. $\square$

Glede na sodo oziroma liho število transpozicij imenujemo permutacijo **soda** oziroma **liha** permutacija.

Permutacije s ponavljanjem

Permutacije s ponavljanjem so nekakšne permutacije, pri katerih pa ne ločimo elementov v skupinah s $k_1, k_2, \dots, k_r$ elementi, torej imamo $n = k_1 + k_2 + \dots + k_r$ elementov - zato delimo število vseh permutacij n elementov s številom njihovih vrstnih redov, tj. permutacij:

$$P_n^{k_1, k_2, \dots, k_r} = \frac{n!}{k_1! k_2! \cdots k_r!}.$$

Primer: 8 vojakov je potrebno poslati na stražo v štiri postojanke. Recimo, da želimo vsak dan izbrati drugačno razporeditev. Na koliko načinov lahko to storimo?
Odgovor: $P_8^{2,2,2,2} = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 / 2^4 = 7 \times 6 \times 5 \times 4 \times 3 = 42 \times 60 = 2520$. Torej je načinov vsekakor preveč, da bi vojaki odšli na stražo na vse možne načine, četudi bi služili vojaški rok celo leto in šli na stražo prav vsak dan po šestkrat. $\diamond$

Če je $r = 1$, je število permutacij s ponavljanjem enako 1, če je $r = n$, pa gre za čisto navadne permutacije. Če je $r = 2$, ločimo elemente v dve skupini. Če je $k = k_1$, je $n - k = k_2$ in pišemo

$$\binom{n}{k} := P_n^{k,n-k}.$$

Ta primer bomo obravnavali posebej v naslednjem razdelku.

Primer: Na koliko načinov lahko med sošolke Aleksandro, Evo in Rebeko razdelimo pet knjig, če dobi vsaka vsaj eno knjigo?
Naj bo m število sošolk in n število knjig, iskano število pa označimo s $S(n, m)$. Potem je očitno $S(n, 1) = 1$. Poglejmo si sedaj primer $m = 2$, tj. primer, ko knjige dobita le dve sošolki. Za vsako knjigo se odločimo, kateri sošolkii jo damo in hitro vidimo, da imamo 2^n različnih možnosti, vendar pa dve možnosti nista pravi (tisti pri katerih bi vse knjige dali prvi oziroma drugi sošolki), tj. iskano število je enako

$$S(n, 2) = 2^n - 2.$$

Morda bo sedaj kaj lažje ugnati primer $m = 3$. Za $n = 5$ bi morda lahko izpisali vsa petmestna števila v trojiškem sistemu (le teh je natanko $3^5 = 3^2 \times 3^2 \times 3 = 9 \times 9 \times 3 = 81 \times 3 = 243$), nato pa označili z $*$ tista števila, ki imajo vse števke enake (teh je ravno 3), z x, y in z pa zaporedoma še tista preostala števila, ki ne vsebujejo nobeno dvojko, enico oziroma ničlo, vendar iz prejšnjega primera že vemo, da je število oznak x (oziroma y oziroma z) je ravno $S(n, 2) = 2^5 - 2$. Torej je iskano število enako:

$$S(n, 3) = 3^n - \binom{3}{2}(2^n - 2) - \binom{3}{1} = 3^n - 3(2^n - 2) - 3.$$

Za $n = 5$ pa dobimo $S(5, 3) = 3(3^4 - 2^5 + 2 - 1) = 3(81 - 32 + 2 - 1) = 3 \times 50 = 150$.

Preverimo dobljeno formulo za $S(n, 3)$ še s formulo za permutacije s ponavljanjem:

n	$S(n,3)$			
3	$3^3 - 3(2^3 - 2) - 3$	$= 6$	$= 3!$	$= P_3^{111}$
4	$3^4 - 3(2^4 - 2) - 3$	$= 36$	$= 3 \times 4!/2!$	$= 3 \cdot P_4^{112}$
5	$3^5 - 3(2^5 - 2) - 3$	$= 150$	$= 3 \cdot \left(\frac{5!}{3!} + \frac{5!}{2!2!}\right)$	$= 3 \cdot (P_5^{113} + P_5^{122})$
6	$3^6 - 3(2^6 - 2) - 3$	$= 540$	$= \frac{3 \cdot 6!}{4!} + \frac{6 \cdot 6!}{2!3!} + \frac{6!}{2!2!2!}$	$= 3 \cdot P_6^{114} + 6 \cdot P_6^{123} + 1 \cdot P_6^{222}$
7	$3^7 - 3(2^7 - 2) - 3$	$= 1806$	$= \frac{3 \times 7!}{5!} + \frac{6 \times 7!}{2!4!} + \frac{3 \times 7!}{3!3!} + \frac{3 \times 7!}{2!2!3!}$	$= 3P_7^{115} + 6P_7^{124} + 3P_7^{133} + 3P_7^{223}$

V primeru $n = 3$ je bil izračun otročje lahek (pa tudi za zelo majhno število gre), že v naslednjem primeru pa smo morali upoštevati tri možnosti, tj. katera od deklet dobi dve knjigi (tudi to število v resnici šteje permutacije s ponavljanjem: $P_3^{21} = 3$). Primer $n = 5$ je zelo podoben prejšnjemu primeru: če dobi eno dekle tri knjige, ostali dve morata dobiti po eno, če pa damo eni dve knjigi, bo dobila še ena dve, za tretjo pa ostane ena sama knjiga. Za $n = 6$ in $n = 7$ pa omenimo še $P_3^{111} = 6$ in $P_3^3 = 1$.

00000$*$	01000x	02000y	10000x	11000x	12000	20000y	21000	22000y
00001x	01001x	02001	10001x	11001x	12001	20001	21001	22001
00002y	01002	02002y	10002	11002	12002	20002y	21002	22002y
00010x	01010x	02010	10010x	11010x	12010	20010	21010	22010
00011x	01011x	02011	10011x	11011x	12011	20011	21011	22011
00012y	01012	02012	10012	11012	12012	20012	21012	22012
00020	01020	02020y	10020	11020	12020	20020y	21020	22020y
00021	01021	02021	10021	11021	12021	20021	21021	22021
00022y	01022	02022y	10022	11022	12022	20022y	21022	22022y
00100x	01100x	02100	10100x	11100x	12100	20100	21100	22100
00101x	01101x	02101	10101x	11101x	12101	20101	21101	22101
00102	01102	02102	10102	11102	12102	20102	21102	22102
00110x	01110x	02110	10110x	11110x	12110	20110	21110	22110
00111x	01111x	02111	10111x	11111$*$	12111z	20111	21111z	22111z
00112	01112	02112	10112	11112z	12112z	20112	21112z	22112z
00120	01120	02120	10120	11120	12120	20120	21120	22120
00121	01121	02121	10121	11121z	12121z	20121	21121z	22121z
00122	01122	02122	10122	11122z	12122z	20122	21122z	22122z
00200y	01200	02200y	10200	11200	12200	20200y	21200	22200y
00201	01201	02201	10201	11201	12201	20201	21201	22201
00202y	01202	02202y	10202	11202	12202	20202y	21202	22202y
00210	01210	02210	10210	11210	12210	20210	21210	22210
00211	01211	02211	10211	11211z	12211z	20211	21211z	22211z
00212	01212	02212	10212	11212z	12212z	20212	21212z	22212z
00220y	01220	02220y	10220	11220	12220	20220y	21220	22220y
00221	01221	02221	10221	11221z	12221z	20221	21221z	22221z
00222y	01222	02222y	10222	11222z	12222z	20222y	21222z	22222$*$
$7x$ $7y$	$8x$	$8y$	$8x$	$7x$ $7z$	$8z$	$8y$	$8z$	$7y$ $7z$
$27-1-14$	$27-8$	$27-8$	$27-8$	$27-1-14$	$27-8$	$27-8$	$27-8$	$27-1-14$
12	19	19	19	12	19	19	19	12
	50			50			50	

Tabela. Nejeverni Tomaži si lahko res izpišejo vseh 3^5 možnosti delitve knjig in prečrtajo (označijo) napačne, pa bodo zopet prišli do $S(5, 3) = 150$. Naj pa bodo pozorni, da bi bilo dovolj izpisati prve tri stolpce, saj drugi trije in zadnji trije izgledajo precej podobno (v resnici jih dobimo iz prvega s permutacijo oznak: (012) oziroma (021)). To pa pomeni, da bi lahko z enako truda izračunali tudi $S(6, 3)$. V resnici lahko v ta namen uporabimo kar zgornjo tabelo - le ničlo si moramo prestavljati na začetku vsake peterice. To pa pomeni, da šesteric označenih z z ni potrebno več odševati, dve šesterici, ki sta označeni z zvezdico pa bi morali označiti z x oziroma y. Torej je $S(6, 3) = (150 + 30)3 = 540$. V splošnem pa dobimo na ta način rekurzivno zvezo $S(n+1, 3) = 3\big(S(n, 3) + S(n, 2)\big)$.

Če nadaljujemo izračun števila $S(n, 3)$ s uporabo formule za permutacije s ponavljanjem, pa stvar postane že skoraj rutinirano dolgočasna:

$$n = 8: \quad 3^8 - 3(2^8 - 2) - 3 = 5796 = 3P_8^{116} + 6P_8^{125} + 6P_8^{134} + 3P_8^{224} + 3P_8^{233},$$

$$n = 9: \quad 3^9 - 3(2^9 - 2) - 3 = 18150 = 3P_9^{117} + 6P_9^{126} + 6P_9^{135} + 3P_9^{144} + 3P_9^{225} + 6P_9^{234} + P_9^{333}.$$

Vse več je sumandov na desni strani, kar pomeni, da postaja za večje m prvi način bolj praktičen/učinkovit. Da pa se ne bi preveč dolgočasili, je morda čas, da rešimo še kakšen primer, npr. ko je $m = 4$: Zapišimo zvezo iz katerih smo izračunali $S(n, 2)$ in $S(n, 3)$ za splošen m:

$$m^n = S(n, m)\binom{m}{m} + S(n, m-1)\binom{m}{m-1} + \cdots + S(n, 2)\binom{m}{2} + S(n, 1)\binom{m}{1}.$$

Le-ta nam da rekurzivno formulo za $S(n,m)$. V primeru $m = 4$ dobimo

$$S(n,4) = 4^n - 4 \cdot S(n,3) - 6 \cdot S(n,2) - 4 \cdot S(n,1)$$

oziroma, če upoštevamo še formule za $S(n,3)$, $S(n,2)$ in $S(n,1)$:

$$S(n,4) = 4^n - 4\Big(3^n - 3(2^n - 2) - 3\Big) - 6(2^n - 2) - 4 = 4^n - 4 \cdot 3^n + 6 \cdot 2^n - 4.$$

Bralcu prepuščamo, da pravkar dobljeno formulo testira (npr. bodisi s permutacijami s ponavljanjem ali pa kar štetjem izračuna $S(5,4)$ (glej sp. tabelo) ter $S(6,4)$ in $S(7,4)$). $\diamond$

0000	0200	1000	1200	2000	2200	3000	3200
0001	0201	1001	1201	2001	2201	3001	3201
0002	0202	1002	1202	2002	2202	3002	3202
0003	0203	1003	1203	2003	2203	3003	3203
0010	0210	1010	1210	2010	2210	3010	3210
0011	0211	1011	1211	2011	2211	3011	3211
0012	0212	1012	1212	2012	2212	3012	3212
0013	0213	1013	1213	2013	2213	3013	3213
0020	0220	1020	1220	2020	2220	3020	3220
0021	0221	1021	1221	2021	2221	3021	3221
0022	0222	1022	1222	2022	2222	3022	3222
0023	0223	1023	1223	2023	2223	3023	3223
0030	0230	1030	1230	2030	2230	3030	3230
0031	0231	1031	1231	2031	2231	3031	3231
0032	0232	1032	1232	2032	2232	3032	3232
0033	0233	1033	1233	2033	2233	3033	3233
0100	0300	1100	1300	2100	2300	3100	3300
0101	0301	1101	1301	2101	2301	3101	3301
0102	0302	1102	1302	2102	2302	3102	3302
0103	0303	1103	1303	2103	2303	3103	3303
0110	0310	1110	1310	2110	2310	3110	3310
0111	0311	1111	1311	2111	2311	3111	3311
0112	0312	1112	1312	2112	2312	3112	3312
0113	0313	1113	1313	2113	2313	3113	3313
0120	0320	1120	1320	2120	2320	3120	3320
0121	0321	1121	1321	2121	2321	3121	3321
0122	0322	1122	1322	2122	2322	3122	3322
0123	0323	1123	1323	2123	2323	3123	3323
0130	0330	1130	1330	2130	2330	3130	3330
0131	0331	1131	1331	2131	2331	3131	3331
0132	0332	1132	1332	2132	2332	3132	3332
0133	0333	1133	1333	2133	2333	3133	3333

Tabela. Vsa štirimestna števila s števkami $0, 1, 2, 3$.

A.4 Kombinacije

Binomski simbol oz. število **kombinacij**, tj. število m-elementnih podmnožic množice moči n, je

$$\binom{n}{m} = \frac{n \cdot (n-1) \cdots (n-m+1)}{1 \cdot 2 \cdots m} = \frac{n!}{m!(n-m)!},$$

saj lahko prvi element izberemo na n načinov, drugi na $n-1$ načinov, ..., zadnji na $n-m+1$ načinov, ker pa vrstni red izbranih elementov ni pomemben, dobljeno število še delimo s številom permutacij.

Trditev A.4. *Za binomske simbole velja* $\binom{n}{0} = 1$, $\binom{n}{m} = \binom{n}{n-m}$,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \quad \textit{in} \quad \binom{n}{m} + \binom{n}{m+1} = \binom{n+1}{m+1}.$$

Zgornji enakosti na levi pravimo **binomski obrazec**, na desni pa **pravilo Pascalovega trikotnika** (glej sliko 1).

Dokaz. Prvi dve relaciji sta očitni že iz same definicije binomskega sibola, tretja pa sledi neposredno iz dejstva, da binomski simbol enak številu vseh kombinacij ustreznega reda. Po definiciji binomskega simbola in množenju obeh strani z $(m+1)!(n-m)!/n!$ je zadnja relacija ekvivalentna $m + 1 + n - m = n + 1$, kar očitno drži. $\square$

Pascalov trikotnik

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

1 10 45 120 210 252 210 120 45 10 1

1 11 55 165 330 462 462 330 165 55 11 1

1 12 66 220 495 792 924 792 495 220 66 12 1

1 13 78 286 715 1287 1716 1716 1287 715 286 78 13 1

1 14 91 364 1001 2002 3003 3432 3003 2002 1001 364 91 14 1

1 15 105 455 1365 3003 5005 6435 6435 5005 3003 1365 455 105 15 1

1 16 120 560 1820 4368 8008 11440 12870 11440 8008 4368 1820 560 120 16 1

1 17 136 680 2380 6188 12376 19448 24310 24310 19448 12376 6188 2380 680 136 17 1

1 18 153 816 3060 8568 18564 31824 43758 48620 43758 31824 18564 8568 3060 816 153 18 1

Slika 1: To so v obliki trikotnika zapisani binomski simboli, vsaka vrstica pa ustreza enemu binoskemu obrazcu.

Primer: Na kvizu je danih 6 vprašanj. Pri vsakem so možni 4 odgovori (npr. (a), (b), (c) in (d)), od katerih je natanko en pravilen. Kakšna je verjetnost, da bomo na vsaj polovico vprašanj odgovorili pravilno, če odgovore izbiramo naključno? Vseh možnih odgovorov je $4^6 = 2^{12} = 4096$, kar je vsekakor preveč, da bi jih vse držali v glavi, tudi na papir bi jih bilo zoprno pisati, s `kopiraj` in `prilepi` pa že ni več tako naporno. Privzameš lahko, da je pravilna rešitev kar `aaaaaa`. Ni pa nič narobe, če nalogo najprej nekoliko poenostavimo in si najprej pogledamo kviz z dvemi, tremi ali celo štirimi vprašanji:

```
aa DA   ba DA   ca DA   da DA   | aaaa DA  acaa DA  baaa DA  bcaa DA  caaa DA  ccaa DA  daaa DA  dcaa DA
ab DA   bb      cb      db      | aaab DA  acab DA  baab DA  bcab     caab DA  ccab     daab DA  dcab
ac DA   bc      cc      dc      | aaac DA  acac DA  baac DA  bcac     caac DA  ccac     daac DA  dcac
ad DA   bd      cd      dd      | aaad DA  acad DA  baad DA  bcad     caad DA  ccad     daad DA  dcad
           n=2                  | aaba DA  acba DA  baba DA  bcba     caba DA  ccba     daba DA  dcba
                                | aabb DA  acbb     babb     bcbb     cabb     ccbb     dabb     dcbb
--------------------------------| aabc DA  acbc     babc     bcbc     cabc     ccbc     dabc     dcbc
                                | aabd DA  acbd     babd     bcbd     cabd     ccbd     dabd     dcbd
aaa DA  baa DA  caa DA  daa DA  | aaca DA  acca DA  baca DA  bcca     caca DA  ccca     daca DA  dcca
aab DA  bab     cab     dab     | aacb DA  accb     bacb     bccb     cacb     cccb     dacb     dccb
aac DA  bac     cac     dac     | aacc DA  accc     bacc     bccc     cacc     cccc     dacc     dccc
aad DA  bad     cad     dad     | aacd DA  accd     bacd     bccd     cacd     cccd     dacd     dccd
aba DA  bba     cba     dba     | aada DA  acda DA  bada DA  bcda     cada DA  ccda     dada DA  dcda
abb     bbb     cbb     dbb     | aadb DA  acdb     badb     bcdb     cadb     ccdb     dadb     dcdb
abc     bbc     cbc     dbc     | aadc DA  acdc     badc     bcdc     cadc     ccdc     dadc     dcdc
abd     bbd     cbd     dbd     | aadd DA  acdd     badd     bcdd     cadd     ccdd     dadd     dcdd
aca DA  bca     cca     dca     | abaa DA  adaa DA  bbaa DA  bdaa DA  cbaa DA  cdaa DA  dbaa DA  ddaa DA
acb     bcb     ccb     dcb     | abab DA  adab DA  bbab     bdab     cbab     cdab     dbab     ddab
acc     bcc     ccc     dcc     | abac DA  adac DA  bbac     bdac     cbac     cdac     dbac     ddac
acd     bcd     ccd     dcd     | abad DA  adad DA  bbad     bdad     cbad     cdad     dbad     ddad
ada DA  bda     cda     dda     | abba DA  adba DA  bbba     bdba     cbba     cdba     dbba     ddba
adb     bdb     cdb     ddb     | abbb     adbb     bbbb     bdbb     cbbb     cdbb     dbbb     ddbb
adc     bdc     cdc     ddc     | abbc     adbc     bbbc     bdbc     cbbc     cdbc     dbbc     ddbc
add     bdd     cdd     ddd     | abbd     adbd     bbbd     bdbd     cbbd     cdbd     dbbd     ddbd
            n=3                 | abca DA  adca DA  bbca     bdca     cbca     cdca     dbca     ddca
                                | abcb     adcb     bbcb     bdcb     cbcb     cdcb     dbcb     ddcb
                                | abcc     adcc     bbcc     bdcc     cbcc     cdcc     dbcc     ddcc
                                | abcd     adcd     bbcd     bdcd     cbcd     cdcd     dbcd     ddcd
                                | abda DA  adda DA  bbda     bdda     cbda     cdda     dbda     ddda
                                | abdb     addb     bbdb     bddb     cbdb     cddb     dbdb     dddb
                                | abdc     addc     bbdc     bddc     cbdc     cddc     dbdc     dddc
                                | abdd     addd     bbdd     bddd     cbdd     cddd     dbdd     dddd
                                |                                      n=4
```

in dobimo zaporedoma 7/16, 10/64 in 67/256. Študent računalništva zna seveda napisati kratek programček, ki opravi delo namesto njega. Nalogo lahko tudi malo posplošimo. Namesto 6ih vprašanj bi lahko vzeli, da jih kviz vsebuje n, kjer je $n \in \mathbb{N}$. Tudi število možnih odgovorov je lahko namesto 4 kar $m \in \mathbb{N}$, $m \geqslant 2$. Vpeljimo še en parameter, označimo ga s k, in si postavimo dodatno vprašanje. Kakšna je verjetnost, da smo s poskušanjem odgovorili pravilno na natanko k vprašanj, pri čemer je $0 \leqslant k \leqslant n$. Če to verjetnost označimo s $P_{n,m,k}$, potem je iskana verjetnost enaka $P_{6,4,3} + P_{6,4,4} + P_{6,4,5} + P_{6,4,6}$ oziroma $1 - P_{6,4,0} - P_{6,4,1} - P_{6,4,2}$. Z uporabo binomskih koeficientov dobimo

$$P_{n,m,0} = \frac{(m-1)^n}{m^n}, \quad P_{n,m,1} = \frac{n\,(m-1)^{n-1}}{m^n}, \ldots, P_{n,m,k} = \binom{n}{k}\frac{(m-1)^{n-k}}{m^n}, \ldots,$$

glede na to, da najprej med n nalogami izberemo k tistih s pravilnimi odgovori, ostale naloge pa nimajo pravilnih odgovorov, zato imamo pri vsakem odgovoru le $m-1$ možnosti. Iskana verjetnost je torej enaka

$$1 - \frac{1}{m^n}\sum_{i=0}^{\lfloor (n-1)/2 \rfloor}\binom{n}{i}(m-1)^{n-i} \tag{A.1}$$

(če bi seštevali po vseh $i \in \{0, 1, \ldots, n\}$ bi po binomskem obrazcu dobili 0) oziroma za

- $n = 2$ in $m = 4$: $(4^2 - 3^2)/4^2 = 7/16 = 0{\cdot}438$,
- $n = 3$ in $m = 4$: $(4^3 - 3^3 - 3 \times 3^2)/4^3 = 10/64 = 0{\cdot}156$,
- $n = 4$ in $m = 4$: $(4^4 - 3^4 - 4 \times 3^3)/4^4 = 67/256 = 0{\cdot}262$,
- $n = 5$ in $m = 4$: $(4^5 - 3^5 - 5 \times 3^4 - 10 \times 3^3)/4^5 = 106/1024 = 0{\cdot}104$, in končno
- $n = 6$ in $m = 4$: $(4^6 - 3^6 - 6 \times 3^5 - 15 \times 3^4)/4^6 = 694/4096 = 0{\cdot}169$.

Rešitvi (A.1) nekoliko bolj zaupamo, ker se odgovori za $n = 2, 3, 4$ ujemajo z verjetnostmi, ki smo jih dobili s preštevanjem. Zaupanje pa bi se še povečalo, če bi preverili še $n = 6$ in $m = 3$ ali vsaj $n = 6$ in $m = 2$. $\diamond$

Primer: Na polico bi radi postavili 4 matematične, 6 fizikalnih in 2 kemijski knjigi. Na koliko načinov lahko to storimo:

(a) če naj knjige iste stroke stojijo skupaj,

(b) če kemijski knjigi ne smeta stati skupaj[1],

(c) če morajo matematične knjige stati na začetku. $\diamond$

Primer: Na koliko načinov lahko sestavimo iz 7ih soglasnikov in 5ih samoglasnikov besedo, ki ima 4 soglasnike in 3 samoglasnike?[2] $\diamond$

[1] Namig: če želimo ponagajati prijateljici, dajmo tisti dve kemijski knjigi namenoma skupaj.

[2] Namig: nalogo lahko rešimo tako s kombinacijami in permutacijami, kot tudi z variacijami in permutacijami s ponavljanjem.

A.5 Vrsta za e

Število e, ki predstavlja osnovo za naravni logaritem, pogosto definiramo s formulo

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n. \tag{A.2}$$

Leta 1683 je Jakob Bernoulli poskušal izračunati limito $(1+1/n)^n$, ko gre n proti neskončno. Uporabil je binomski obrazec:

$$\left(1+\frac{1}{n}\right)^n = 1+\binom{n}{1}\frac{1}{n}+\binom{n}{2}\frac{1}{n^2}+\binom{n}{3}\frac{1}{n^3}+\cdots+\binom{n}{n}\frac{1}{n^n}.$$

k-ti sumand na desni strani zgornje relacije je enak

$$\binom{n}{k}\frac{1}{n^k} = \frac{1}{k!}\cdot\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}.$$

Za $n\to\infty$, gre slednji ulomek na desni proti 1, tj.

$$\lim_{n\to\infty}\binom{n}{k}\frac{1}{n^k} = \frac{1}{k!},$$

kar pomeni, da lahko e zapišemo kot vrsto.

$$e = 1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots. \tag{A.3}$$

O tem se prepričamo zato, ker je vsak člen v binomski razširitvi naraščujoča funkcija od n, sledi iz izreka o monotoni konvergenci za vrste, da je vsota te neskončne vrste enaka e. Bernoulli se je na ta način prepričal, da število e leži med 2 in 3. To sta torej v nekem smislu prva približka za e, vendar pa Bernoulli nikoli ni povezal tega števila z logaritmom[3] Za kaj takega je bilo potrebno izkristalizirati pojem funkcije in dognati, da sta eksponentna in logaritemska funkcija inverzni. Euler je v resnici prvi dokazal zgornjo zvezo (A.3), hkrati pa izračunal prvih 18 decimalk števila e: $e \approx 2{\cdot}718281828459045235$

Še splošnejšo vrsto pa dobimo z uporabo Taylorjeve vrste:

Trditev A.5.

$$e^x = 1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots.$$

[3]Glej http://www.gap-system.org/~history/HistTopics/e.html#s19.

A.6 Stirlingov obrazec

Stirlingovo aproksimacijo (ozriroma formulo ali obrazec) uporabljamo za učinkovito računanje/ocenjevanje velikih faktorjelov in je poimenovana po škotskem matematiku Jamesu Stirlingu (1692–1770):

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \tag{A.4}$$

ko gre $n \to \infty$. Zavedati se moramo, da je naivno računanje zgornjega faktorjela $1 \cdot 2 \cdot 3 \dots (n-1) \cdot n$ eksponentne časovne zahtevnosti v odvisnosti od dolžina zapisa števila n (le-ta je seveda enaka naravnemu številu k, ki je zelo blizu $\log n$, pri čemer za logaritemsko osnovo vzamemo številsko osnovo, v kateri zapišemo število n). V čem smo torej na boljšem, ko računamo izraz na desni strani (A.4)? Računanje potence lahko izvedemo tako, da najprej izračunamo naslednje potence $(n/e)^{2^0}, (n/e)^{2^1}, (n/e)^{2^2}, \dots, (n/e)^{2^k}$ z zaporednim kvadriranjem, nato pa zmnožimo med seboj tiste katerih eksponenti ustrezajo mestom enic v binarni predstavitvi števila n, kar pomeni da smo opravili največ $2k$ množenj.

Primer: Namesto, da bi izračunali $P = a^{21}$ z dvajsetimi množenji ($P = a$, dvajsetkrat ponavljaj $P := P * a$), raje izračunamo potence a, a^2, a^4, a^8, a^{16}, nato pa zaradi $21 = 2^4 + 2^2 + 2^1$ še $P = a^{16} \cdot a^4 \cdot a^1$, kar znese samo 4 kvadriranja in 2 množenji. ◇

Formulo (A.4) je prvi odkril Abraham de Moivre v naslednji obliki

$$n! \approx [\text{konstanta}] \cdot n^{n+1/2} e^{-n},$$

ko gre $n \to \infty$, pri čemer je konstanto izrazil s hyperboličnim logaritmom. James Stirlingov prispevek pa je bil, da je konstanta v resnici enaka $\sqrt{2\pi}$. Formulo tipično uporabljamo v aplikacijah v obliki

$$\ln n! \approx n \ln n - n,$$

ko gre $n \to \infty$. V zgornji verziji manjka faktor $\frac{1}{2}\ln(2\pi n)$, ki ga lahko za velike n zanemarimo v primerjavi z drugimi faktorji. Zapišimo $\ln(n!) = \ln 1 + \ln 2 + \cdots + \ln n$, pri čemer lahko na desno stran zgornje relacije gledamo kot na približek za integral

$$\int_1^n \ln x \, dx = n \ln n - n + 1.$$

Od tu naprej pa si lahko pomagamo z Euler–Maclaurinovo formulo in uporabo Bernoullijevih števil. Glej npr. `http://en.wikipedia.org/wiki/Stirling's_approximation`.

A.7 Normalna krivulja v prostoru

Želimo pokazati naslednjo identiteto.[4]

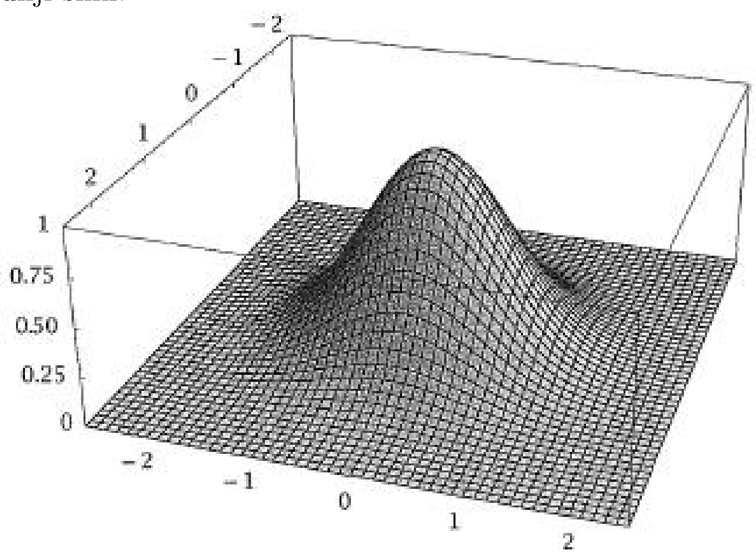
Izrek A.6.

$$\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}. \tag{A.5}$$

Dokaz. Označimo vrednost integrala na levi strani (A.5) z I. Funkcija

$$g(s,t) = e^{-(s^2+t^2)} = e^{-s^2}e^{-t^2}$$

je narisana na spodnji sliki.



Slika: Normalna gora.

Sedaj pa prerežimo zgornjo ploskev z ravnino $s = 1$. Prerez seveda izgleda kot normalna krivulja, ploščina pod dobljeno krivuljo pa je enaka ploščini pod normalno krivuljo, ki je pomnožena z e^{-1}:

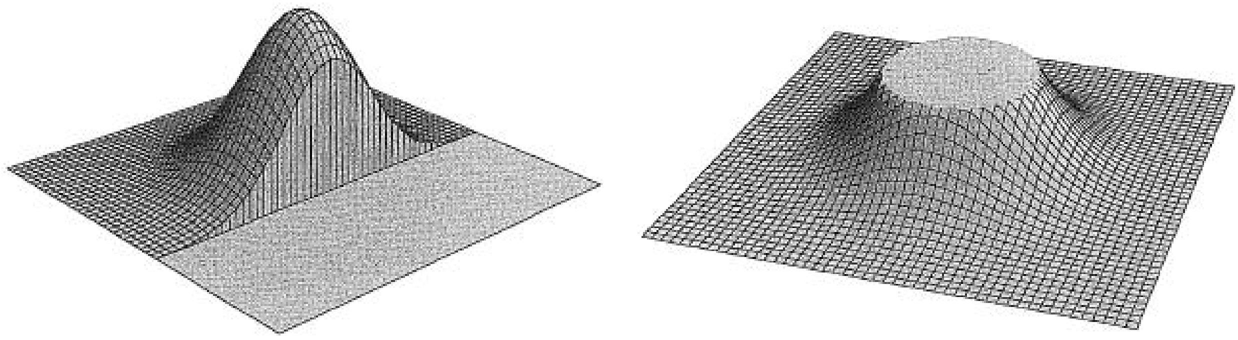
$$\int_{-\infty}^{\infty} e^{-1}e^{-t^2}\,dt = e^{-1}\,I.$$

Podobno je za katerokoli drugo vrednost števila s ploščina ustrezne krivulje enaka $e^{-s^2}\,I$. Sedaj lahko izračunamo prostornino normalne gore z naslednjim integralom

$$V = \int_{-\infty}^{\infty} e^{-s^2}\,I\,ds = I^2.$$

Preostane nam le še, da dokažemo, da je $V = \pi$. Tokrat presekajmo normalno ploskev z ravnino $z = h$.

[4]Stara zgodba pravi, da je Lord Kelvin nekoč dejal, da je matematik nekdo, za katerega je ta identiteta očitna.



Slika: Vertikalni in horizontalni prerez normalne ploskve.

Potem za točke preseka velja

$$e^{-s^2-t^2} \geqslant h \qquad \text{oziroma} \qquad s^2+t^2 \leqslant -\ln h.$$

To pa je ravno krog s središčem $(0,0)$ in polmerom $r = \sqrt{-\ln h}$. Ploščina tega kroga je $\pi r^2 = \pi(-\ln h)$, prostornino V pa dobimo z integriranjem od $h=0$ do $h=1$:

$$V = \int_0^1 \pi(-\ln h)\,dh = -\pi \int_0^1 \ln h\,dh.$$

Postavimo $u = \ln x$, $dv = dx$, $du = dx/x$ in $v = \displaystyle\int dx = x$. Z integriranjem po delih dobimo

$$\int_0^1 \ln x\,dx = x\ln x\Big|_0^1 - \int_0^1 \frac{x\,dx}{x} = -1$$

in od tod $V = \pi$. $\square$

A.8 Sredine nenegativnih števil $a_1, \ldots, a_n$

Aritmetična: $$A_n = \frac{a_1 + \cdots + a_n}{n}$$

Geometrična: $$G_n = \sqrt[n]{a_1 \cdot \ldots \cdot a_n}$$

Harmonična: $$H_n = \frac{n}{\frac{1}{a_1} + \cdots + \frac{1}{a_n}}$$

Kvadratna: $$K_n = \sqrt{\frac{a_1^2 + \cdots + a_n^2}{n}}$$

Sredine dveh števil: $H_2 \leqslant G_2 \leqslant A_2 \leqslant K_2$

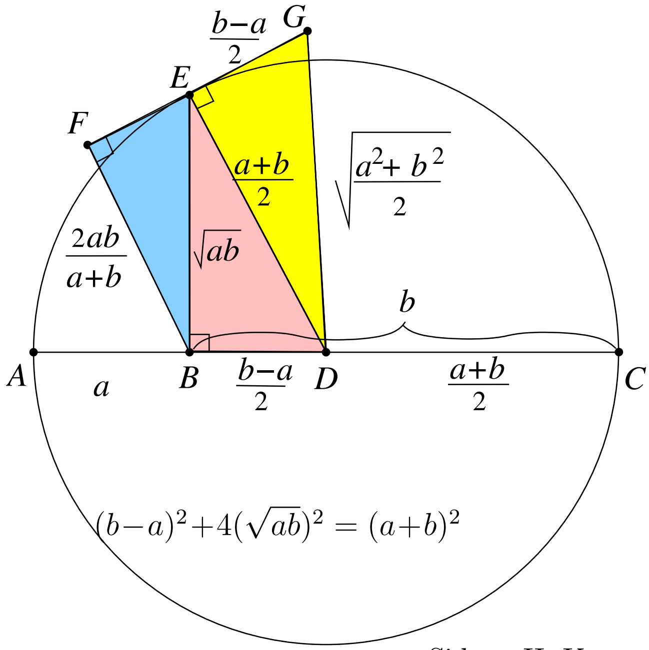
$a, b \geqslant 0$

$$\boxed{H_2 = \frac{2}{\frac{1}{a} + \frac{1}{b}}}$$

$$\boxed{G_2 = \sqrt{ab}}$$

$$\boxed{A_2 = \frac{a+b}{2}}$$

$$\boxed{K_2 = \sqrt{\frac{a^2+b^2}{2}}}$$



Sidney H. Kung
(iz R.B. Nelsenove knjige "Dokazi brez besed")

Potenčna (stopnje k): $$P_{n,k} = \sqrt[k]{\frac{a_1^k + \cdots + a_n^k}{n}}$$

Velja:

$$H_n = P_{n,-1}, \quad G_n = \lim_{k \to 0} P_{n,k} \quad A_n = P_{n,1} \ \text{ in } \ K_n = P_{n,2}$$

ter

$$H_n \leqslant G_n \leqslant A_n \leqslant K_n \qquad \text{oziroma za} \quad k \leqslant m \quad P_{n,k} \leqslant P_{n,m}$$

A.9 Cauchyjeva neenakost

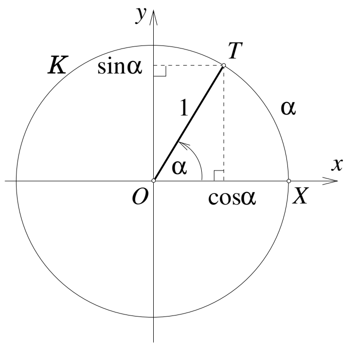
Skalarni produkt vektorjev $\vec{u}$ in $\vec{v}$ iz $\mathbb{R}^n$ je po definiciji enak

$$\vec{u} \cdot \vec{v} = |\vec{u}| \operatorname{proj}_{\vec{u}} \vec{v},$$

kjer je $\operatorname{proj}_{\vec{u}} \vec{v}$ pravokotna projekcija vektorja $\vec{v}$ na vektor $\vec{u}$. Od tod sledi $\vec{u} \cdot \vec{v} = |\vec{u}||\vec{v}| \cos\varphi$ (glej sliko). Naj bo $\vec{u} = (u_1, ..., u_n)$, $\vec{v} = (v_1, ..., v_n)$, potem lahko skalarni produkt izračunamo z naslednjo vsoto

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + ... + u_n v_n.$$

Potem je dolžina vektorja $\vec{v}$ enaka $|\vec{v}\,| = \sqrt{\vec{v} \cdot \vec{v}}$. Neposredno iz $|\cos\varphi\,| \leqslant 1$ sledi za realne vektorje naslednja neenakost.



Slika A.4.1: Definiciji *sinusa* in *kosinusa*. V ravnini narišemo enotsko krožnico $\mathcal{K}$ s središčem O v izhodišču koordinatnega sistema. Iz točke $X = (1, 0)$ se v nasprotni smeri od urinega kazalca poda na pot po krožnici $\mathcal{K}$ točka T. Ko ima za seboj "prehojen" lok dolžine α (takrat je kot $\angle XOT$ enak α radianov), ima točka T koordinati $(\cos\alpha, \sin\alpha)$. Funkcija sinus je pozitivna v prvem in drugem kvadrantu, funkcija cosinus pa v prvem in četrtem. Obe funkciji sta periodični s periodo 2π (tj. $360°$).

Trditev A.7. (**Cauchyjeva neenakost**) *Za poljubna vektorja $\vec{u}$ in $\vec{v}$ velja*

$$|\vec{u} \cdot \vec{v}\,| \leqslant |\vec{u}\,||\vec{v}\,|.$$

Enakost velja natanko tedaj, ko je kota med vektorjema $\vec{u}$ in $\vec{v}$ enak $k\pi$, za $k \in \mathbb{N}$, to je natanko tedaj, ko sta vektorja $\vec{u}$ in $\vec{v}$ kolinearna.

Pravimo ji tudi Cauchy-Schwarzova neenakost ali neenakost Bunjakovskega, glej `http://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality`. To je ena izmed najpomembnejših neenakosti. Omogoča nam, da preverimo kolinearnost dveh vektorjev tako, da izračunamo vrednosti na levi in desni strani zgornje neenakosti in preverimo, če sta enaki.

Za $n = 2$ Cauchyjeva neenakost sledi tudi iz identitete

$$(a^2 + b^2)(c^2 + d^2) = (ad - bc)^2 + (ac + bd)^2, \qquad a, b, c, d \in \mathbb{R},$$

ki je že naša stara znanka in pove, da za kompleksni števili $z = a + ib$ in $w = c + id$ *produkt absolutnih vrednosti dveh kompleksnih števil enak absolutni vrednosti ustreznega produkta*, tj. $|z| \cdot |w| = |zw|$.

Za $n = 3$ lahko dokažemo Cauchyjevo neenakost s pomočjo vektorskega produkta in Lagrangeove identitete. **Vektorski produkt** vektorjev $\vec{u} = (u_1, u_2, u_3)$ in $\vec{v} = (v_1, v_2, v_3)$ iz $\mathbb{R}^3$ je vektor v $\mathbb{R}^3$ podan s formulo: $\vec{u} \times \vec{v} = (u_2 v_3 - u_3 v_2, -u_1 v_3 + u_3 v_1, u_1 v_2 - u_2 v_1)$.

Dolžina vektorja $\vec{u}\times\vec{v}$ je enaka $|\vec{u}\times\vec{v}\,| = |\vec{u}\,||\vec{v}\,||\sin\varphi\,|$, geometrično pa to pomeni, da je dolžina vektorskega produkta enaka ploščini paralelograma, ki ga razpenjata vektorja $\vec{u}$ in $\vec{v}$. **Lagrangeova identiteta** $|\vec{u}|^2|\vec{v}|^2 = (\vec{u}\vec{v})^2 + |\vec{u}\times\vec{v}|^2$, ni nič drugega kot na drugačen način zapisana relacija $\sin^2\varphi + \cos^2\varphi = 1$ (oziroma Pitagorjev izrek).

Za splošen $n\in\mathbb{N}$ lahko Cauchyjevo neenakost zapišemo tudi v naslednji obliki:

$$\big(a_1^2+\cdots+a_n^2\big)\big(b_1^2+\cdots+b_n^2\big) \geqslant \big(a_1b_1+\cdots+a_nb_n\big)^2,$$

kjer so $a_1,\dots,a_n,b_1,\dots,b_n$ poljubna realna števila. Lagrangeova identiteta za splošen n pa izgleda takole:

$$\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) = \left(\sum_{i=1}^n a_ib_i\right)^2 + \sum_{i<j}(a_ib_j-a_jb_i)^2.$$

Zadnja vsota ima $(n-1)+(n-2)+\cdots+2+1 = (n-1)n/2$ členov.

Raba v verjetnosti

Za slučajni spremenljivki X in Y je matematično upanje njunega produkta skalarni produkt, tj.

$$\langle X,Y\rangle := \mathsf{E}(XY)$$

zadovoljuje tri aksiome iz naslednje škatle. (V tem primeru velja $\langle X,X\rangle = 0$ natanko tedaj, ko je $P(X=0)=1$.) Potem iz Cauchyjeve neenakosti sledi

$$|\mathsf{E}(XY)|^2 \leqslant \mathsf{E}(X^2)\,\mathsf{E}(Y^2).$$

Naj bo $\mu = E(X)$ in $\nu = E(Y)$. Potem po Cauchyjevi neenakosti velja

$$\begin{aligned}|\mathsf{Cov}(X,Y)|^2 &= |\mathsf{E}\big((X-\mu)(Y-\nu)\big)|^2 = |\langle X-\mu, Y-\nu\rangle|^2\\ &\leqslant \langle X-\mu, X-\mu\rangle\langle Y-\nu, Y-\nu\rangle = \mathsf{E}\big((X-\mu)^2\big)\,\mathsf{E}\big((Y-\nu)^2\big) = \mathsf{D}(X)\,\mathsf{D}(Y),\end{aligned}$$

kjer je D disperzija, Cov pa kovarianca.

Formalno je **vektorski prostor s skalarnim produktom** (rečemo tudi *unitarni vektorski prostor*) vektorski prostor V nad poljubnim obsegom $\mathbb{F}$ s skalarnim produktom, tj. s preslikavo

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$$

ki zadovoljuje naslednje tri aksiome za poljubne vektorje $x, y, z \in V$ in skalarje $a \in \mathbb{F}$:

- *Konjugirana simetrija*: $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
- *Linearnost na prvi koordinati*: $\langle ax, y \rangle = a\langle x, y \rangle . \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.
- *Positivna-definitnost*: $\langle x, x \rangle \geqslant 0$, kjer velja enakost, če in samo če je $x = 0$.

(zgoraj smo opustili vektorske oznake). Glej http://en.wikipedia.org/wiki/Inner_product_space.

Predstavimo še dokaz Cauchyjeve neenakosti za vektorski prostor s skalarnim produktom.

Dokaz. Naj bosta u in v poljubna vektorja vektorskega prostora V nad obsegom $\mathbb{F}$. Neenakost je očitna za $v = 0$, zato predpostavimo, da je $\langle v, v \rangle \neq 0$. Naj bo $\delta \in \mathbb{F}$. Potem velja

$$0 \leqslant |u - \delta v|^2 = \langle u - \delta v, u - \delta v \rangle = \langle u, u \rangle - \bar{\delta}\langle u, v \rangle - \delta\langle u, v \rangle + |\delta|^2 \langle v, v \rangle.$$

Sedaj pa izberimo $\delta = \langle u, v \rangle \cdot \langle v, v \rangle^{-1}$, in dobimo

$$0 \leqslant \langle u, u \rangle - |\langle u, v \rangle|^2 \cdot \langle v, v \rangle^{-1} \qquad \text{oziroma} \qquad |\langle u, v \rangle|^2 \leqslant \langle u, u \rangle \cdot \langle v, v \rangle,$$

in končno po korenjenju neenakost, ki smo jo želeli pokazati. $\square$

Trikotniško neenakost za vektorske prostore s skalarnim produktom pogosto pokažemo kot posledico Cauchyjeve neenakosti na naslednji način: za vektorja x in y velja

$$|x + y|^2 = \langle x + y, x + y \rangle = |x|^2 + \langle x, y \rangle + \langle y, x \rangle + |y|^2 \leqslant |x|^2 + 2|x|\,|y| + |y|^2 = (|x| + |y|)^2 .$$

Po korenjenju dobimo trikotniško neenakost.

Dodatek B

Vadnica

Stari pregovor pravi: "Brez muje se še čevelj ne obuje." Ni kaj, snov je najbolje utrjevati z reševanjem nalog. Sprva ni lahko, potem pa lahko postane celo zabavno.

B.1 Vaje za uvod

Za tiste, ki ste že nekoliko pozabili snov iz srednje šole[1] priporočam ponavljanje. Možnosti je veliko, npr.

- Nives Mihelič Erbežnik et al., Priprave na maturo, matematika, 1. izd., 1. natis. – Ljubljana: DZS, 2001. (Zbirka nalog za srednje šole) 399, str.

Za ogrevanje smo si izposodili nekaj nalog iz verjetnosti:

1. Kaj je poskus in kaj slučajni dogodek?

- **(a)** Iz kompleta 32 kart izberemo 3 karte.
- **(b)** Izvlečena karta iz komleta 32 kart je srčni kralj.

Obkroži poskus.

2. Zapisa $A \cap B \subset A$ in $A \cap B \subset B$ govorita o:

- **(a)** vsoti dogodkov,
- **(b)** produktu dogodkov.

3. Produkt nezdružljivih dogodkov A in B je:

- **(a)** nemogoč dogodek,
- **(b)** gotov dogodek,
- **(c)** sestavljen dogodek.

4. Poljuben dogodek in njegova negacija sta:

- **(a)** nezdružljiva in nasprotna dogodka,
- **(b)** združljiva in nasprotna dogodka,
- **(c)** združljiva in neodvisna dogodka.

5. Za dogodek A, da padejo pri metu igralne kocke več kot 3 pike, je nasprotni dogodek (negacija):

- **(a)** da ne pade več pik kot 3,
- **(b)** da ne padejo kvečjemu štiri pike.

6. Pošteno igralno kocko vržemo enkrat. Popolni sistem elementarnih dogodkov tega poskusa so dogodki, da:

- **(a)** padejo šestice $\{e_6\}$,
- **(b)** pade katerokoli število pik $\{e_1, e_2, e_3, e_4, e_5, e_6\}$.

[1] Pogosto se zgodi, da snov kot je verjetnost (ali celo statistika) učitelji v srednjih šolah sploh ne obdelajo, večinoma pa jo le na hitro preletijo.

7. Vržemo igralno kocko. Sestavljeni dogodek je:
   (a) pade šestica,
   (b) pade sodo število pik, manjše od 5,
   (c) pade liho število pik, večje od 4.

8. Klasična in statistična verjetnost se razlikujeta po tem, da je ocena za verjetnost dogodkov:
   (a) pri prvi dobljena računsko, pri drugi empirično,
   (b) pri prvi dobljena empirično, pri drugi računsko

9. Elementarni dogodki, ki jih obravnavamo s klasično definicijo verjetnosti:
   (a) morajo sestavljati popoln simetričen sistem dogodkov,
   (b) ne smejo sestavljati popolnega simetričnega sistema dogodkov.

10. V posodi so 3 rdeče kroglice in 2 modri. Na slepo izvlečemo kroglico, jo vrnemo v posodo, jih premešamo in ponovno na slepo eno izvlečemo. Izid poskusa je elementarni dogodek, ki ga določata barvi dveh kroglic. Vseh elementarnih dogodkov v tem poskusu je:

(a) V_5^2, **(b)** C_2^5.

11. V posodi sta 2 beli kroglici in 5 rdečih. Na slepo izvlečemo iz nje eno kroglico. Verjetnost, da je izvlečena kroglica rdeča, je:

(a) 2/7, **(b)** 5/7.

12. Iz kupa 32 igralnih kart izvlečemo 3 karte. Verjetnost, da je med njimi vsaj en as, je:

(a) 0·34, **(b)** 0·66.

13. Dva moška in pet žensk naključno razporedimo v vrsto. Verjetnost, da bosta moška sedela skupaj na začetku ali na koncu vrste, je:

(a) 2/7, **(b)** 2/21.

14. Hkrati vržemo 3 poštene igralne kocke. Verjetnost, da pade ena šestica, hkrati pa vsaka kocka pokaže različno število pik, je:

(a) $20/6^3$, **(b)** 5/18.

15. V veslaškem klubu so 3 krmarji in 10 veslačev, od tega 5 članov in 5 mladincev. Trener mora v naglici izbrati posadko za četverec s krmarjem, zato jo sestavi kar na slepo. Verjetnost, da bo izbral najboljšega krmarja, najboljšega člana in najboljšega mladinca, je:

(a) 2/45, **(b)** 3/45.

16. Problem rešujeta neodvisno drug od drugega dva učenca. Učenec A rešuje probleme tako, da je verjetnost za posamezno rešitev 0·9, učenec B pa z verjetnostjo 0·6. Verjetnost, da bo problem rešen, je:

(a) 0·04, **(b)** 0·96.

17. Probleme rešujeta neodvisno drug od drugega dva učenca. Učenec A rešuje probleme tako, da je verjetnost za posamezno rešitev 0·9, učenec B pa z verjetnostjo 0·6. Problem naj reši le en učenec. Verjetnost, da bo problem rešil samo učenec A, je:

(a) 0·857, **(b)** 0·143.

18. Pri poskusu sodelujejo štiri osebe. Vsaka si povsem naključno in neodvisno od drugih izbere neko naravno število, manjše od 100. Verjetnost, da bo vsaj eno od teh štirih števil deljivo s 3, je

(a) 1/3, **(b)** 2/3, **(c)** 65/81.

19. Študent obvlada 8 izpitnih vprašanj od 12-ih. Na izpitu mora na slepo izbrati 4 vprašanja. Pozitivno oceno doseže, če pravilno odgovori vsaj na 2 vprašanji. Verjetnost, da bo študent izpit opravil, je:

(a) 2/3, **(b)** 14/15.

20. V prvi vrečki so 3 bele kroglice, v drugi pa 2 rdeči. Na slepo izberemo eno od obeh vrečk, v njej izberemo kroglico in jo prenesemo v sosednjo vrečko. Naposled še enkrat sežemo na slepo v eno od obeh vrečk in v njej vnovič na slepo izberemo kroglico. Verjetnost, da bo nazadnje izbrana kroglica bele barve, je:

(a) 3/4, **(b)** 25/48.

21. V športnem oddelku gimnazije, ki ga obiskuje 24 dijakov, se jih 15 ukvarja z nogometom, 15 jih kolesari in 10 šahira. Nogomet in šah igra 6 dijakov. Z nogometom in kolesarjenjem se jih ukvarja 10, 7 pa jih kolesari in igra šah. Vse tri športe gojijo 3 dijaki. Izračunajte verjetnost dogodkov:

(a) da se slučajno izbran dijak ne ukvarja z nobenim športom,

(b) da slučajno izbran dijak kolesari in šahira, a ne igra nogometa,

(c) da se med dvema slučajno izbranima dijakoma eden ukvarja samo z nogometom, drugi pa samo kolesari.

22. Iz kompleta 52 igralnih kart na slepo izberemo tri.

(a) Kolikšna je verjetnost, da so vse enake barve?

(b) Kolikšna je verjetnost, da so vse tri različnih barv?

23. V n-kotniku na slepo izberemo dve oglišči. Kolikšno mora biti najmanj število n, da bo verjetnost, da sta izbrani točki krajišči diagonale mnogokotnika, vsaj $0{\cdot}95$?

24. Črke besede **HIPERBOLA** napišemo na 9 listkov in jih premešamo. Nato na slepo izberemo 5 listkov in jih položimo na mizo v naključnem zaporedju. Izračunajte verjetnost naslednjih dogodkov:

(a) sestavili smo besedo **LOPAR**,

(b) sestavljena beseda vsebuje 3 soglasnike in 2 samoglasnika,

(c) sestavljena beseda vsebuje črko **H**,

(d) sestavljena beseda vsebuje 4 soglasnike, se začne s črko **H** in konča s črko **A**.

25. Sedem različnih učbenikov na slepo pospravimo v dva predala: v večjega štiri, v manjšega tri učbenike. Kolikšna je verjetnost, da bosta določena dva učbenika znašla v istem predalu?

26. Na zabavi se je zbralo 6 družin. Vsako družino prestavljajo oče, mati in trije otroki. Za igro slepo izberemo dve odrasli osebi in enega otroka. Izračunajte verjetnost dogodkov:

(a) da so osebe izbrane iz iste družine,

(b) da so osebe izbrane iz treh različnih družin in sta odrasli osebi različnega spola.

27. Trener razdeli na slepo pet različnih štartnih številk petim atletom: Andreju, Borisu, Cenetu, Vidu in Žigi. Kolikšna je verjetnost, da dobi Andrej nižjo startno številko kot Vid in Žiga?

28. Štiri različna pisma za štiri različne naslovnike bomo na slepo zalepili v štiri ovojnice z njihovimi naslovi. Izračunaj verjetnost, da bosta natanko dve pismi prišli na pravi naslov.

29. Na treh kroglicah so številke 1, 3 in 5, na treh ploščicah pa 2, 4 in 6. Kroglice in ploščice postavimo v raven niz.

- **(a)** V koliko razporeditvah stojijo kroglice skupaj?
- **(b)** V koliko razporeditvah se niz začne in konča s ploščico?
- **(c)** Koliko štirimestnih števil lahko sestavimo iz vseh števk, ki so na kroglicah in ploščicah?
- **(č)** Kolikšna je verjetnost dogodka, da je štirimestno število večje od 5000?

30. V stolpnici stanuje 5 družin z enim otrokom, 3 družine s 3 otroki in 2 družini s 5 otroki. Zaradi anketiranja izberemo na slepo 3 družine. Izračunajte verjetnosti, da

- **(a)** imata dve izmed izbranih družin isto število otrok,
- **(b)** imajo vse tri izbrane družine skupaj 7 otrok.

31. Kolikšna je verjetnost, da pri naključni permutaciji črk besede **ŠALA** enaki črki ne bosta stali druga zraven druge?

32. Naključen izbor treh črk v besedi **RAČUN** naj bo elementaren dogodek v poskusu.

- **(a)** Koliko je vseh različnih dogodkov v tem poskusu?
- **(b)** Kolikšna je verjetnost, da sta med izbranimi tremi črkami vsaj dva soglasnika?
- **(c)** Kolikšna je verjetnost, da se pri naključni permutaciji črk besede **RAČUN** sestavi nova beseda, v kateri oba samoglasnika nista sosednji črki?

33. Micka ima v vrečki pet kroglic, ki se navzven razlikujejo le po barvi: 3 so bele, 2 pa sta črni. Na slepo vlečemo kroglice eno za drugo iz vrečke. Kolikšna je verjetnost, da ji bo šele v tretjem poskusu uspelo prvič izvleči črno kroglico?

34. Na naši fakulteti so si študentje izbirne predmete izbrali takole: 55 študentov statistiko, 80 študentov kriptografijo, 75 študentov verjetnost, 25 študentov kriptografijo in verjetnost, 20 študenti statistiko in verjetnost, 5 študentov vse tri predmete.

- **(a)** Koliko študentov je v tem letniku?
- **(b)** Koliko študentov je izbralo dva in koliko samo en predmet od vseh treh naštetih?
- **(c)** Koliko študentov je izbralo kriptografijo in ne verjetnosti?
- **(č)** Izračunaj verjetnost dogodka, da je slučajno izbran študent izbral statistiko.
- **(d)** Izračunaj verjetnost dogodka, da je slučajno izbran študent izbral verjetnost, pri pogoju, da je izbral statistiko.

35. Pri poskusu sodelujejo štiri osebe. Vsaka si povsem naključno in neodvisno od drugih izbere neko naravno število, manjše od 100. Kolikšna je verjetnost, da bo vsaj eno od teh štirih števil deljivo s 3?

36. Lokostrelec cilja v mirujočo tarčo. Na voljo ima 4 puščice, vrednost zadetka pri vsakem strelu je 0·6 (po zadetku streljanje prekine). Izračunaj verjetnost dogodkov:

- **(a)** za zadetek porabi natanko 3 puščice,
- **(b)** za zadetek porabi največ 2 puščici,
- **(c)** tarča je zadeta,
- **(d)** lokostrelec porabi vse 4 puščice.

37. Iz škatle, v kateri so 3 bele in 2 črni kroglici, vlečemo na slepo po eno kroglico brez vračanja, dokler ni število izvlečenih belih kroglic enako številu izvlečenih črnih kroglic, ali dokler v škatli ne zmanjka kroglic. Izračunajte verjetnost, da bo število izvlečenih belih kroglic enako številu izvlečenih črnih.

38. V podjetju z velikim številom zaposlenih je 60% delavcev moških. 35% moških in 25% žensk ima visoko izobrazbo.

- **(a)** Izračunajte verjetnost, da ima naključno izbrani delavec visoko izobrazbo.
- **(b)** Naključno izbrani delavec ima visoko izobrazbo. Kolikšna je verjetnost, da je to ženska?
- **(c)** Vsaj koliko delavcev moramo izbrati, da je med njimi z verjetnostjo večjo od 0·95, vsaj eden z visoko izobrazbo.

39. Lastnik stojnice na zabavišču je ugotovil, da je verjetnost, da nalkjučni gost s pikadom zadene v polno in s tem osvoji za nagrado plišastega medvedka, približno 1/10. Najmanj kolikokrat mora tedaj oče Hinko vreči pikado, da bo verjetnost, da osvoji vsaj enega medvedka za svojo hčer, vsaj 1/2? Računajte, da namerava oče vnaprej plačati število metov in ne glede na vmesne izide vreči vsa plačana pikada.

40. Robotek stoji v spodnjem levem polju šahovnice velikosti 3×3 (ki jo narišemo na vrhu visoke stolpnice). Vsak njegov premik je naključen: z enako verjetnostjo se premakne zmeraj za eno polje bodisi v desno ali pa navzgor. Na ta način lahko robotek torej tudi zdrsne čez rob šahovnice, s čimer je njegove poti seveda konec. Kolikšna je verjetnost, da robotku uspe priti na zgornje desno polje?

41. Vsak od šesterice prijateljev je pred kasaško dirko na slepo stavil na enega od treh konjev A, B ali C.

- **(a)** Na koliko načinov je malhko v tem primeru stavilo teh šest prijateljev?
- **(b)** Kolikšna je verjetnost, da sta Albert in Bruno, da iz omenjene šesterice prijateljve, stavila na istega konja?
- **(c)** Kolikšna je verjetnost, da je vsaj eden od teh šestih prijateljev stavil na konja A?
- **(č)** Kolikšna je verjetnost, da jih je med njimi na konja A stavilo toliko, kot na oba ostala konja?

42. Igralno kocko vržemo petkrat zapored.

- **(a)** Kolikšna je verjetnost, da bo v prvem in zadnjem metu padlo enako število pik?
- **(b)** Kolikšna je verjetnost, da bo padlo pri tem sodo število pik vsaj enkrat?
- **(c)** Kolikšna je verjetnost, da bo šele

v tretjem metu prvič padlo manj kakor tri pike?

(č) Kolikšna je verjetnost, da bo padlo v vsakem naslednjem metu več pik kakor v prejšnjem?

43. Štirje igralci drug za drugim v krogu mečejo igralno kocko. Zmaga igralec, ki prvi vrže šestico. Izračunajte verjetnost, da zmaga igralec, ki igro začne?

44. Igralna kocka je prirejena tako, da so nekateri izidi bolj verjetno od drugih. Izidi enica, dvojka in trojka so enako verjetni, izida štirica in petica sta dvakrat bolj verjetna kot enica, šestica je trikrat verjetnejša od enice.

(a) Izračunajte verjetnosti elementarnih dogodkov.

(b) Izračunajte verjetnosti, da pade sodo število pik.

(c) Izračunajte verjetnosti, da pade v treh poskusih vsaj enkrat šestica.

45. Pokaži, da iz $A \subset B$ sledi $AB = A$.

46. Pokaži, da iz $A \subset B$ sledi $A \cup B = B$.

47. Kocko vržemo sedemkrat. Kolikšna je verjetnost, da padejo več kot štiri šestice?

48. Katero število grbov je nabolj verjetno, če vržemo pošten kovanec

(a) 70-krat, **(b)** 75-krat.

49. Tovarna izdeluje žarnice; med njimi je 4% takih, ki po kakovosti ne ustrezajo normi. Oceni verjetnost, da so med 50 kupljenimi žarnicami 4 neustrezne.

50. Trije lovci so hkrati ustrelili na divjega prašiča, ki je ubit z eno samo kroglo. Kolikšne so verjetnosti, da je vepra ubil

(a) prvi,

(b) drugi,

(b) tretji

lovec, če poznamo njihove verjetnosti, da zadanejo: 0·2, 0·4, 0·6?

B.2 Poskusi, dogodki in definicija verjetnosti

1. Standardno kocko (s pikami od 1 do 6) želimo obtežiti tako, da ko jo bomo dvakrat (neodvisno) vrgli, bo vsota pik obeh metov zavzela vrednost od 2 do 12, vsako z enako verjetnostjo. Poišči tako obtežitev ali dokaži, da ne obstaja.

2. Skupina sedmih moških in petih žensk se odpravlja na taborjenje. Imajo dva šotora za tri osebe in tri šotore za dve osebi. Na koliko načinov se lahko razdelijo v šotore? Kaj pa, če naj bo v vsakem šotoru vsaj en moški in vsaj ena ženska? Kaj pa, če morajo biti v istem šotoru le osebki istega spola? (Ljudi in šotore med seboj ločimo.)

3. Trije gusarji, Kuki, Luki in Muki, najdejo zaklad, štiri zlatnike. Razdelijo si jih na naslednji način: Kuki in Luki vržeta pošten kovanec. Če pade grb, dobi prvi zlatnik Kuki, sicer ga dobi Luki. Tako en gusar dobi prvi kovanec, ostala dva gusarja pa nato mečeta kovanec za drugi zlatnik. Za tretji zlatnik se na isti način potegujeta tista dva gusarja, ki nista dobila drugega zlatnika, ter za četrti zlatnik tista dva, ki nista dobila

tretjega zlatnika. Kaj je bolj verjetno, da ima Kuki isto število zlatnikov kot Luki ali kot Muki?

4. Janez gre z avtobusom v službo. Možno se je peljati z dvema progama, prva vozi na 5, druga pa na 7 minut in sta neodvisni. Če gre s prvo progo, potrebuje od trenutka, ko stopi na avtobus, pa do službe 15 minut, če gre z drugo, pa 13 minut. Recimo, da gre Janez na prvi avtobus, ki pride. Kolikšna je verjetnost, da bo do službe (skupaj s čakanjem) potreboval manj kot 18 minut?

5. V družbi je n ljudi. Vsak od njih da svojo vizitko v posodo, nato pa iz nje vsak na slepo izbere po eno vizitko. Dokaži, da je verjetnost, da bo natanko r ljudi ($0 \leqslant r \leqslant n$) dobilo svojo vizitko, enaka

$$p_r(n) = \frac{1}{r!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-r}}{(n-r)!}\right).$$

6. Na FRI nekoč niso hoteli vzeti v službo enega dobrega raziskovalca. Kot kaže, se ta s tem ni mogel sprijazniti in je po mnogih letih prišel na Fakulteto ter ugrabil 100 študentov - med njimi tudi Tebe. Zaprl vas je v veliko predavalnico in jo obdal z eksplozivom. Prosili ste ga, da vas izpusti, vendar se ga ni dalo omehčati saj tudi njega, kjub njegovi veliki želji, da bi delal na FRI, daleč nazaj ali pa sedaj niso uslišali. Ker pa je imel tako zelo rad računalništvo (posebej verjetnost in statistiko, kjer se je naučil tudi kakšno novo strategijo), se je odločil, da vam vseeno ponudi naslednjo možnost:

   - v drugi predavalnici je pripravil 100 omaric (oštevilčenih od 1 do 100) in zbral vseh 100 vaših študentskih izkaznic;
   - v vsako omarico je dal po eno izkaznico in vam - študentom povedal, da boste lahko šli en po en v to predavalnico z omaricami.
   - vsak izmed vas bo lahko odprl največ 50 omaric (lahko tudi manj), jih nato zaprl in odšel ven;
   - vsak študent, ki pride ven iz sobe z omaricami, se ne more pogovarjati (ali kako drugače komunicirati) z drugimi, saj dobi v usta veliko nogavico, čez glavo pa žakelj;
   - če se bo zgodilo, da je vsak od vas videl med drugim tudi svojo izkaznico, potem vas bo izpustil, sicer pa bo šla cela stavba v zrak (pa čeprav ni dolgo tega kar je bila renovirana).

   Študente je za trenutek zgrabila panika, saj so pomislili, da se jim ne piše dobro, potem pa so se spomnili, da si med njimi tudi Ti zvit(a) študent(ka), ki jim lahko razložiš kakšno strategijo lahko uberete, da bo verjetnost, da se rešite bistveno večja od $(1/2)^{100}$, recimo blizu 1/3. Gotovo misliš, da gre samo za sanje, pa temu žal ni tako. Ali se še spomniš, kaj si jim povedal(a)?

B.3 Pogojna verjetnost

1. Pepe deli karte pri taroku. Ko deli talon (6 kart), pogleda, ali je v njem škis (ena izmed 54 kart, kolikor jih je vseh skupaj). Če je škis v talonu, mu ga s pogojno verjetnostjo 50% uspe neopaženo vtihotapiti med svojih 12 kart, z verjetnostjo 30% mu to ne uspe (a tudi nihče nič ne opazi), z verjetnostjo 20% pa ga razkrinkajo. Recimo, da soigralci niso opazili nič sumljivega. Kolikšna je pogojna verjetnost, da ima Pepe škisa?
2. Osebi A in B mečeta kovanec, pri katerem grb pade z verjetnostjo p. Zmaga igralec, ki prej vrže grb. Najprej kovanec vrže oseba A in če pade grb zmaga, sicer poda kovanec osebi B. Če oseba B vrže grb zmaga, sicer poda kovanec nazaj osebi A in igra se nadaljuje, dokler nekdo ne zmaga.
   (a) Denimo, da je kovanec pravičen ($p = 1/2$). Kolikšna je verjetnost, da zmaga oseba A? Kolikšna, da zmaga oseba B?
   (b) Posplošimo igro na k igralcev. Torej najprej vrže kovanec oseba A_1, nato $A_2, \ldots, A_k, A_1, \ldots$ dokler ena oseba ne vrže grba. Kolikšna je verjetnost, da zmaga oseba A_i?
   (c) Denimo, da je zmagala oseba A_i. Kolikšna je pogojna verjetnost, da je zmagala v drugem krogu (oseba A_i je v drugem poskusu vrga grb)?
3. Med dvajsetimi kovanci sta dva kovanca z dvema grboma ter dva kovanca z dvema ciframa. Ostali kovanci imajo en grb in eno cifro. Naključno si izberemo en kovanec in ga vržemo 5-krat.
   (a) Kolikšna je verjetnost, da vržemo natanko 5 grbov?
   (b) Recimo, da smo vrgli 5 grbov. Kolikšna je pogojna verjetnost, da smo povlekli kovanec z dvema grboma?
4. Vsak izmed 10 študentov si izbere bodisi belo bodisi črno kroglico, vsako enako verjetno. Nato 8-krat seže v to posodo in vsakič izvleče kroglico, ki je ne vrne. Izkaže se, da je izžrebal 5 belih in 3 črne.
   (a) Kolikšna je verjetnost, da sta v posodi ostali še dve črni kroglici?
   (b) Študent Anže ve, da je v posodo prispeval belo kroglico. Koliko je zanj pogojna verjetnost, da sta v posodi ostali še dve črni kroglici?
   (c) Študent Blaž ve, da je v posodo vrgel črno kroglico (ne ve pa, kaj ve Anže). Koliko je zanj pogojna verjetnost, da sta v posodi še dve črni kroglici?
5. V prvi posodi sta dve beli in ena črna kroglica, v drugi posodi pa dve črni in ena bela kroglica. Najprej Marjetica vrže goljufivi kovanec, na katerem grb pade z verjetnostjo

40%. Če pade grb, z izvleče eno kroglico iz prve posode, sicer izvleče eno kroglico in druge posode. Izvlečene kroglice ne vrne nazaj v posodo. Za njo pride Trdoglav, ki izvleče eno kroglico iz iste posode iz katere je c kroglico izvlekla Marjetica.

(a) Kolikšna je verjetnost, da je Trdoglav izvlekel črno kroglico?

(b) Kolikšna je pogojna verjetnost, da je Trdoglav izvlekel črno kroglico, če je Marjetica izvlekla črno kroglico?

(c) Kolikšna je pogojna verjetnost, da je Marjetica izvlekla črno kroglico, če je Trdoglav izvlekel črno kroglico?

6. Gusar želi poiskati zaklad, ki se nahaja na enem izmed n otokov. Otoke začne raziskovati enega za drugim, tako da za naslednji otok izbere enega od preostalih, vse z enako verjetnostjo. Ker želi gusar v čim krajšem času poiskati zaklad je malce površen in zato najde zaklad na posameznem otoku le z verjetnostjo p, če se zaklad na tem otoku tudi nahaja. Naj bo k celo število, $0 \leqslant k \leqslant n$. Kolikšna je verjetnost, (a) da gusar ne najde zaklada na enem izmed prvih k otokov, in (b) da se s zaklad ne nahaja na nobenem izmed prvih k otokov, če vemo, da gusar ni našel zaklada na prvih k otokih.

B.4 Bernoullijevo zaporedje neodvisnih poskusov

B.5 Slučajne spremenljivke in porazdelitve

B.6 Slučajni vektorji

B.7 Funkcije slučajnih spremenljivke in vektorjev

1. Hkrati vržemo tri kovance in za tem še enkrat vržemo tiste kovance, na katerih je padel grb. Meti so med seboj neodvisni, verjetnosti, da pade grb je na vseh kovancih enaka 50%. Naj bo X število kovancev, na katerih je padla cifra v prvem metu in naj bo Y število kovancev, na katerih je padla s cifra v drugem metu.

(a) Zapišite porazdelitev slučajnega vektorja (X, Y).

(b) Poiščite robni porazdelitvi in porazdelitev produkta XY.

(c) Sta slučajni spremenljivki X in Y neodvisni?

(d) Izračunajte matematično upanje $E(X - 2Y)$.

B.8 Momenti in kovarianca

1. Fakulteta za računalništvo in informatiko ima nad pritličjem 8 različnih nadstropij $(M_1, 1, M_2, \ldots, M_4, 4)$, ki so bila nekoč označena s številkami od 1 do 8. Denimo, da so

tudi sedaj označena tako. V dvigalo vstopi 6 profesorjev, med katerimi si vsak izbere nadstropje naključno, vsakega z isto verjetnostjo in neodvisno od ostalih profesorjev.

(a) Kolikšno je pričakovano število postankov dvigala?

(b) Naj bo H najvišje nadstropje, v katerem se dvigalo ustavi. Poiščite tako število (mediano) m, za katerega velja $P(H \leqslant m) \leqslant 1/2 \leqslant P(H \leqslant m)$.

2. Dragić, Nachbar in Lakovič tekmujejo v metanju trojk. Vsak izmed njih vrže na koš enkrat. Prvi na koš vrže Dragić, za njim Nachbar in na koncu se Lakovič. Preden začnejo metati na koš, da vsak izmed njih v kapo 100EUR. Kdor koš zgreši, mora znesek v kapi podvojiti. Kdor koš zadane, dobi ves denar iz kape, nato pa vsi prispevajo začetnih 100EUR. Kar je v kapi po koncu metov, si razdelijo na enake dele. Recimo, da Dragić zadane trojko z verjetnostjo 25%, Nachbar z verjetnostjo 50% in Lakovič z verjetnostjo 75%. Koliko ima Dragić v povprečju dobička oz. izgube?

3. Kako sta povezani mediana zaporedja in mediana slučajne spremenljivke?

4. Predpostavi, da je graf gostote verjetnosti $p(x)$ zvezne slučajne spremenljivke simetričen. Kaj lahko v tem primeru poveš o mediani in pričakovani vrednosti (mat. upanje) te slučajne spremenljivke?

B.9 Karakteristične funkcije in limitni izreki

B.10 Opisna statistika

1. Množico, ki jo statistično opazujemo, imenujemo:

(a) populacija,

(b) vzorec,

(c) statistična spremenljivka (znak).

2. Značilnosti populacije kot celote imenujemo:

(a) statistične enote,

(b) statistične spremenljivke,

(c) statistične parametri.

3. Frekvenca je število, ki pove:

(a) koliko je v opazovani populaciji vseh možnih vrednosti statistične spremenljivke;

(b) kolikokrat je v opazovani populaciji nastopila določena vrednost statistične spremenljivke.

4. Frekvenčna distribucija je prikaz:

(a) zbranih vrednosti statistične spremenljivke v ustrezni tabeli;

(b) zbranih vrednosti statistične spremenljivke v ustreznem grafikonu.

5. Frekvenca posameznega razreda je enaka:

(a) številu enot opazovane populacije;

(b) številu vrednosti številske spremenljivke, ki spada v tisti razred.

(c) količniku med številom vrednosti številske spremenljivke, ki spada v tisti razred, in številom vseh enot populacije.

6. Širina razreda je enaka:
   (a) aritmetični sredini obeh mej razreda,
   (b) razliki med zgornjo in spodnjo mejo razreda.

7. Porazdelitev absolutnih in porazdelitev relativnih frekvenc lahko prikažemo s:
   (a) frekvenčnimi poligoni,
   (b) histogrami,
   (c) frekvenčnimi kolači.

8. Frekvenčni kolač (strukturni krog) sestavljajo krožni izseki. Ti kažejo
   (a) deleže enot, ki sodijo v posamezne razrede;
   (b) število enot, ki sodijo v posamezne razrede.

9. Uteženo aritmetično sredino koristno uporabljamo, če:
   (a) statistična spremenljivka zavzame različne vrednosti na več enotah populacije;
   (b) statistična spremenljivka zavzame isto vrednost na več enotah populacije;
   (c) so vrednosti statistične spremenljivke razdeljene v razrede.

10. Varianca, mera za razpršenost posameznih vrednosti, je:
   (a) število, ki je kvadratni koren iz povprečja kvadratov odklonov posameznih vrednosti statistične spremenljivke od aritmetične sredine;
   (b) število, ki je enako povprečju kvadratov odklonov posameznih vrednosti statistične spremenljivke od aritmetične sredine;

11. V razredu s 25 učenci je 8% odličnih, 28% pravdobrih, trije so nezadostni, sedem je zadostnih. Ostali so dobri. Izračunajte povprečno oceno razreda in narišite histogram frekvenc ocen.

12. V razredu je 25 učencev. Učenka Andreja je računala povprečno število točk pri šolski nalogi. Pri prvem računanju se je zmotila: ni upošteva svojega dosežka in dobila povprečje 74·5 točk. Ko je napako popravila, je dobila povprečje 75 točk. Koliko točk je dosegla Andreja pri šolski nalogi?

13. Za izhodišče vzemimo stavek:

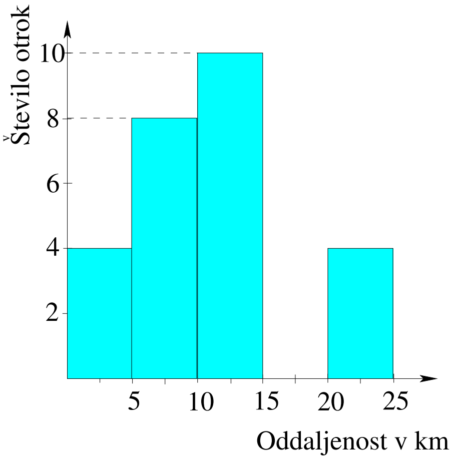
Je zvito kakor kozji rog.
   (a) V zgornjem stavku opazujte števila črk v posameznih besedah. Izračunajte povprečno število črk v besedah tega stavka in standardno deviacijo števila črk v besedah.
   (b) Opazujte besede v zgornjem stavku še enkrat: tokrat štejte, koliko samoglasnikov vsebuje ta in ona beseda. Ali je razpršenost teh podatkov večja ali manjša od razpršenosti podatkov o dolžini posameznih besed iz prejšnjega vprašanja?

14. V nekem razredu je 30 otrok. Njihovi domovi so od šole oddaljeni največ 25 km. Oddaljenosti so prikazane s frekvenčnim histogramom, v katerem manjka podatek za oddaljenost od 15km do 20 km. Mejne oddaljenosti 5 km, 10 km, ... štejemo v histogramu k manjšim vrednostim, na primer 15 km$\in (10\text{km}, 15\text{km}]$.

(a) Koliko domov otrok je od šole oddaljenih od 5 km do 10 km?

(b) Koliko domov otrok je od šole oddaljenih od 15 km do 20 km?

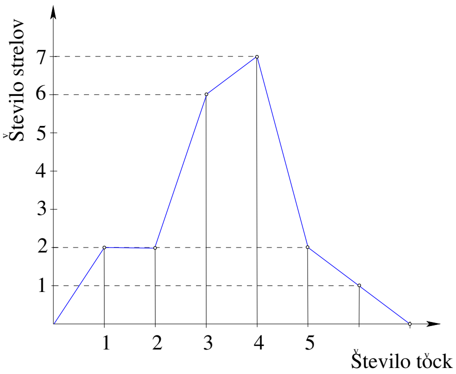
(c) Koliko so domovi otrok povprečno oddaljeni od šole? Izračunajte povprečno vrednost, ki jo lahko dobite iz podatkov v histogramu.



15. Dva strelca sta vsak dan dvajsetkrat ustrelila v tarčo. Pri vsakem strelu sta lahko dosegla 0 točk, če sta tarčo zgrešila, 1 točko, 2 točki, ..., 5 točk, če sta zadela v sredino tarče. Točke prvega so prikazane s tabelo, točke drugega pa s frekvenčnim poligonom.

(a) Kateri strelec ima večje povprečje točk?

(b) Kateri strelec ima manjši standardni odklon?



B.11 Vzorčenje

B.12 Cenilke

B.13 Intervali zaupanja

B.14 Preverjanje statističnih domnev

1. Proizvajalec merilcev električne moči, ki se uporabljajo za uravnavanje pragov energije pri podatkovno–komunikacijskih sistemih, trdi, da ob normalno delujoči proizvodni liniji proizvede največ 10% nedelujočih merilcev.

   Trgovec je pravkar dobil pošiljko 25 omenjenih merilcev. Recimo, da želi testirati domnevo $H_0 : p = 0{\cdot}10$ napram $H_1 : p > 0{\cdot}10$, kjer je p pravi delež pokvarjenih merilcev. Pri testu uporabi $y \geqslant 6$ za prag zavrnitve (y je seveda pravo število pokvarjenih merilcev v pošiljki).

(a) Določite vrednost α za ta test.

(b) Določite β, če je $p = 0{\cdot}2$. Koliko je moč testa za to vrednost p?

(c) Določite β, če je $p = 0{\cdot}4$. Koliko je moč testa za to vrednost p?

2. S pomočjo računalniške simulacije so raziskovali učinek napak na strojih na performanse proizvodnega sistema (Industrial Engineering, Aug. 1990). Študija se je osredotočila na sistem z enim strojem. Povprečni čas med začetkom dveh procesiranj je $1{\cdot}25$ minute pri konstantnem času procesiranja 1 minuta. Naprava je pokvarjena 10% časa. Po $n = 5$ neodvisnih simulacijah dolžine 160 ur je povprečna produktivnost na teden (40-urni delavnik) $\overline{y} = 1908{\cdot}8$ izdelkov. Za sistem brez napak je povprečna produktivnost 1920 izdelkov. Če predvidevamo, da je standardna deviacija 5 simulacij $s = 18$ izdelkov, testiraj domnevo, da je resnična povprečna produktivnost manjša od 1920 izdelkov. Testirajte pri značilnosti $\alpha = 0{\cdot}05$.

3. Rezultati druge raziskave o nacionalnem zdravju in prehranjevanju v ZDA so pokazali, da imajo ljudje v starosti med pol leta in 74 let v krvi povprečno koncentracijo svinca 14μg/dl (Analytical Chemistry, feb. 1986). Poleg tega so ugotovili, da imajo črnski otroci v starosti do pet let znatno višje koncentracije od ostalih.

   V naključnem vzorcu 200 črnskih otrok v starosti pod pet let je bila ugotovljena povprečna koncentracija svinca v krvi 21μg/dl s standardnim odklonom 10μg/dl. Je to dovolj, da lahko trdimo, da je pravo povprečje pri črnski populaciji pod pet let res večje od 14μg/dl? Testirajte z vrednostjo $\alpha = 0{\cdot}01$.

4. Dovoljena koncentracija PCB-ja (nevarna substanca) v vodi po standardu, ki ga je postavila EPA, je 5 delcev na milijon. Večji proizvajalec PCB-ja, ki se uporablja za električno izolacijo, spušča manjše količine PCB-ja skupaj z odpadno vodo. Uprava tovarne je izdala navodila, da je treba ustaviti proizvodnjo, če povprečna koncentracija PCB-ja v odplakah preseže 3 delce/milijon. Analiza 50 naključnih vzorcev odpadne vode je pokazala naslednje rezultate:

$$\overline{y} = 3{\cdot}1 \text{ delcev na milijon} \qquad \text{in} \qquad s = 0{\cdot}5 \text{ delcev na milijon}$$

   (a) Ali so zgornji statistični rezultati analize dovoljšen dokaz, da je proizvodnjo potrebno ustaviti? Uporabite $\alpha = 0{\cdot}01$.

   (b) Če bi bili vi menedžer tovarne, ali bi uporabili večjo ali manjšo vrednost za α v testu pod točko (a)? Pojasnite!

5. Vrtanje "globokih lukenj" je družina procesov, ki omogočajo vrtanje lukenj, ki so vsaj desetkrat globlje kot je premer svedra. Eden bistvenih problemov pri globokem vrtanju je zastajanje izvrtanih okruškov v utorih svedra. Izveden je bil eksperiment, s katerim

so preučili uspešnost globokega vrtanja v primeru, ko se okruški sprijemajo (Journal of Engineering for Industry, maj 1993). Globina izvrtanih lukenj pri 50 vrtanjih je v povprečju znašala $y = 81{\cdot}2\,\mathrm{mm}$ s standardno deviacijo vzorca $s = 50{\cdot}2\,\mathrm{mm}$. Testirajte domnevo, da se pravo povprečje izvrtanih lukenj μ razlikuje od 75 mm. Uporabite stopnjo značilnosti $\alpha = 0{\cdot}01$.

6. Inštitut za okolje in tehnologijo je izdal študijo o onesnaženosti zemlje na Nizozemskem. Skupno so zbrali, posušili in analizirali za prisotnost cianida 72.400 gramov vzorcev zemlje. Z infrardečo mikroskopsko metodo so merili koncentracijo cianida v miligramih na kilogram zemlje v vzorcih. Povprečna koncentracija cianida v vzorcih je bila $y = 84\,\mathrm{mg/kg}$, standardna deviacija pa $s = 80\,\mathrm{mg/kg}$.

   Uporabite to informacijo za testiranje domneve, da je resnično povprečje koncentracije cianida v zemlji na Nizozemskem manjše od $100\,\mathrm{mg/kg}$ pri stopnji značilnosti $\alpha = 0{\cdot}10$.

7. Ali tekmovanje med različnimi odseki za raziskovanje in razvoj (R&R) v sklopu ameriškega ministrstva za obrambo, ki delajo na istem projektu, izboljša produktivnost? Za odgovor na to vprašanje so raziskali produktivnost pri 58 projektih, dodeljenih večim oddelkom, ter pri 63 projektih, dodeljenih enemu samemu oddelku (IEEE Transactionson Engineering Management, feb. 1990). Glede na rezultate so dobili povprečen faktor uspešnosti pri prvih (tekmovalnih) projektih enak $7{\cdot}62$, pri drugih (netekmovalnih) pa $6{\cdot}95$.

   (a) Postavite ničelno in alternativno domnevo za odločanje, ali povprečje uspešnosti pri tekmovalnih projektov presega povprečje uspešnosti pri netekmovalnih projektih.

   (b) Določite območje zavrnitve za test pri vrednosti $\alpha = 0{\cdot}05$.

   (c) Izkazalo se je, da pri zgornjem testu p–vrednost leži v intervalu med $0{\cdot}02$ in $0{\cdot}03$. Kakšen je pravilen zaključek?

8. Inštitut za okolje in prostor je narediš študijo o insekticidih, ki se uporabljajo v nasadu orhidej v dolini San Joaquin v Kaliforniji. Zbrali so vzorce zraka v nasadu in jih testirali vsak dan v obdobju najbolj intenzivnega škropljenja. V spodnji tabeli so prikazane količine oksonov in thionov v zraku (v $\mathrm{ng/m^3}$) in razmerje med količino thionov in oksonov. Primerjaj povprečje razmerij med oksoni in thioni v meglenih in jasnih/oblačnih pogojih v nasadu orhidej z uporabo domneve. Uporabi stopnjo značilnosti $\alpha = 0{\cdot}05$.

Datum	Vremenske razmere	Thioni	Oksoni	Razmerje oksoni/thioni
15. jan	megla	38·2	10·3	0·270
17.	megla	28·6	6·9	0·241
18.	megla	30·2	6·2	0·205
19.	megla	23·7	12·4	0·523
20.	megla	62·3	-	vzorec izgubljen
20.	jasno	74·1	45·8	0·618
21.	megla	88·2	9·9	0·112
21.	jasno	46·4	27·4	0·591
22.	megla	135·9	44·8	0·330
23.	megla	102·9	27·8	0·270
23.	oblačno	28·9	6·5	0·225
25.	megla	46·9	11·2	0·239
25.	jasno	44·3	16·6	0·375

9. Tovarna bi rada ugotovila kateri od naslednjih dveh virov energije – plin in elektrika – bo dal cenejšo energijo. Ena mera za ekonomično proizvodnjo energije (v angl. se imenuje plant investment perdelivered quad) je izračunana tako, da vzamemo vsoto vloženega denarja (dolarji) v določen vir s strani neke tovarne in jo delimo s količino energije (v kvadrilionih britanskih termalnih enot). Manjši kot je kvocient manj plača tovarna za dobavljeno elektriko. Izbran je bil naključni vzorec 11ih tovarn, ki uporabljata električno energijo

1	2	3	4	5	6	7	8	9	10	11
204·15	0·57	62·76	89·72	0·35	85·46	0·78	0·65	44·38	9·28	78·60

in 16ih tovarn, ki uporabljajo plin:

1	2	3	4	5	6	7	8
0·78	16·66	74·94	0·01	0·54	23·59	88·79	0·64
9	10	11	12	13	14	15	16
0·82	91·84	7·20	66·64	0·74	64·67	165·60	0·36

Nato je bil izračunan zgoraj omenjeni kvocient za vsako tovarno in podatki vnešeni v tabele, sledi pa še izpis iz MINITABa za analizo podatkov.

```
---------------------------------------------------------------
TWOSAMPLE T FOR electric VS gas
          N   MEAN  STDEV     SE MEAN
electric 11   52,4  62,4      19
gas      16   37,7  49,0      12

95 PCT CI FOR MU electric - MU gas: (-30, 59)

TTEST MU electric = MU gas (VS NE) : T=0,68    P=0,50   DF= 25
```

```
POOLED STDEV = 54,8 (standardna e)
------------------------------------------------------------
```

(a) Ali ti podatki predstavljajo dovolj dober pokazatelj za stopnjo značilnosti $\alpha = 0{\cdot}05$, ki kaže na razliko med povprečjem kvocientov, ki uporabljajo plin in tistimi, ki uporabljajo elektriko?

(b) Kakšne predpostavke so potrebne, da bo postopek veljaven?

(c) Preveri, če so predpostavke iz točke (b) smiselno izpolnjene. Kako to vpliva na upravičenost rezultata v točki (a)?

10. Raziskovalci so naredili eksperiment za določitev vpliva puščavskih granivorjev (živali, ki jedo semena) na gostoto in porazdelitev semen v prsti (Ecology, Dec. 1979). Določene vrste puščavskih glodalcev delajo zaloge semen na površju zemlje, zato je bil eksperiment zasnovan tako, da so raziskovali ali zaradi takšnih zalog semen v povprečju na tistem območju iz semen zraste več mladih rastlinic kot na sosednjih kontrolnih območjih. Locirali so 40 majhnih območij, kjer so glodalci kopičili semena in jih pokrili z mrežo, da so glodalcem preprečili ponoven dostop. Zamrežili so tudi sosednja območja za kontrolo. Potem so opazovali število semen, ki je vzklilo na zamreženih območjih. V spodnji tabeli je povzetek zbranih podatkov. Ali so rezultati eksperimenta zadosten dokaz (pri stopnji značilnosti $\alpha = 0{\cdot}05$), da je povprečno število vzkaljenih semen na območjih, kjer so jih kopičili glodalci bistveno večje kot na kontrolnih območjih?

Območja z zalogami	Kontrolna območja
$n_1 = 40$	$n_2 = 40$
$y_1 = 5{\cdot}3$	$s_2 = 2{\cdot}7$
$s_1 = 1{\cdot}3$	$s_2 = 0{\cdot}7$

11. Percepcija govora pretežno gluhih oseb se zanaša predvsem na branje z ust, tj., zaznavanje pogovornega jezika z opazovanjem artikuliranih gibov, izrazov na obrazu ter gest sogovornika. Ali se da percepcijo govora izboljšati tako, da bralcu glasu vizulano predstavimo podatek o poudarku zlogov. Da bi raziskali ta fenomen je 10 oseb z normalnim sluhom sodelovalo v eksperimentu pri katerem so morali ustno ponoviti zvočno predvajane stavke, katerih informacijo niso videli na video-monitorju. (Journal of the Acoustical Society of America, Feb. 1986). Stavki so bili predstavljeni osebam pod naslednjima pogojema:

(1) branje govora z informacijo o frekvenci in amplitudi govornega signala (oznaka $S + F + A$) ter

(2) samo branje govora (oznaka S).

Za vsakega od 10-ih oseb, je bila izračunana razlika med procentom pravilno reproducirane vsebine pod pogojem $S + F + A$ in pod pogojem S. Povprečje in standardni odklon razlik sta naslednja:

$$\overline{d} = 20{\cdot}4 \quad \text{in} \quad s_d = 17{\cdot}44.$$

Testiraj domnevo, da je povprečje procentov pravilnih vsebin pri pogoju $S + F + A$ presega ustrezno povprečje pod pogoju S. Privzami $\alpha = 0{\cdot}05$.

12. Tetraklorodibenzo-p-dioksin (TCDD) je visoko-toksična substanca, ki jo najdemo v industrijskih odpadkih. Znanstveniki so naredili študijo s katero so določili količino TCDD-ja v tkivih volovskih žab, ki živijo na območju Rocky Branch Creek v Arkansasu, za katerega se ve, da je kontaminirano z TCDD (Chemosphere, Feb. 1986). Merili so količino TCDD-ja (v delcih na trilijon) v različnih tkivih štirih samic volovskih žab. Za vsako žabo so izmerili razmerje med količino TCDD-ja v tkivu in v nožni mišici. Relativno razmerje med koncentracijo TCDD-ja v jetrih in jajčnikih s je podana v spremljajoči preglednici. Raziskovalci po zbranih podatkih sklepajo, da je povprečna koncentracija TCDD-ja v jajčnikih samic volovskih žab večja kot povprečna koncentracija v jetrih. Testirajte to trditev z uporabo $\alpha = 0{\cdot}05$.

Žaba	Jetra	Jajčniki
A	11·0	34·2
B	14·6	41·2
C	14·3	32·5
D	12·2	26·2

13. V Merckovih raziskovalnih laboratorijih so izvedli poskus ocene učinka novega zdravila, pri čemer so uporabili prirejen plavalni labirint. Devetnajstim oplojenim jezovnim podganam so dali 12·5 mg zdravila. Iz vsakega legla so za plavanje v labirintu naključno izbrali enega moškega ter enega ženskega mladiča. Vsak podganji mladič je postavljen v vodo na enem koncu labirinta, plava pa, dokler uspešno ne najde izhoda na nasprotnem koncu. Če mladiču ne uspe najti izhoda v določenem časovnem intervalu, ga spet postavijo na začetek in mu dajo še eno priložnost za pobeg. Poskus ponavljamo vse dokler vsak mladič uspešno ne pobegne trikrat. Število plavanj, potrebnih za tri uspešne pobege posameznega mladiča, je predstavljeno v spodnji razpredelnici. Ali lahko sklepamo, da je povprečno število potrebnih pobegov različno pri moških oz. ženskih mladičih?

Leglo	Moški	Ženski	Leglo	Moški	Ženski
1	8	5	11	6	5
2	8	4	12	6	3
3	6	7	13	12	5
4	6	3	14	3	8
5	6	5	15	3	4
6	6	3	16	8	12
7	3	8	17	3	6
8	5	10	18	6	4
9	4	4	19	9	5
10	4	4			

```
TEST  MU = 0.000 proti MU $\not =$ 0.000

            N   E       STDEV     SE MEAN      T      P VALUE
Swimoiff   19   0.368     3.515     0.806      0.46      0.65
```

15. Raziskovalci z Univerze v Rochestru so študirali trenje, ki nastane v procesu zajemanja papirja iz kasete v fotokopirnem stroju (Journal of Engineering for Industry, May 1993). Poskus je vseboval opazovanje zamika posameznih listov papirja v kupu, s katerega je fotokopirni stroj zajemal. Posamezno zajemanje je označeno kot uspešno, če se noben list z izjemo vrhnjega ni zamaknil za več kot 25% celotne predvidene dolžine premika. V kupu stotih listov je bilo zajemanje uspešno 94-krat. Predviden delež uspešnosti postopka je 0·9. Pri stopnji značilnosti $\alpha = 0{\cdot}1$ oceni, ali zanesljivost presega 0·9.

16. Ameriška agencija za znanost, NSF, je v študiji 2237-ih doktorandov s področij tehniških ved na ameriških univerzah ugotovila, da je bilo med njimi 607 ameriških državljanov (revija Science, 24. september 1993). S stopnjo značilnosti $\alpha = 0{\cdot}01$ testiraj domnevo, da dejanski odstotek doktorskih nazivov, podeljenih tujim (neameriškim) državljanom na ameriških univerzah, presega 0·5.

17. Zaradi dvomov o zadostni varnosti v letalskem prometu je ameriška zvezna agencija za letalstvo (FAA) uvedla sankcije zoper letalske družbe, ki na preizkusih varnosti dobijo nezadostno oceno. Ena izmed serij poskusov, izpeljanih na mednarodnem letališču v Los Angelesu (LAX), je pokazala, da so varnostniki odkrili zgolj 72 izmed 100 lažnih orožij, ki so jih inspektorji FAA bodisi nosili pri sebi bodisi spravili v osebno prtljago. Kot so zatrdili z FAA, je bila stopnja odkritih primerov občutno pod državnim povprečjem, ki znaša 0·8. Testiraj to domnevo, pri čemer za stopnjo značilnosti vzemi $\alpha = 0{\cdot}1$.

18. Oddelek za mehaniko organizacije ASEE vsakih 10 let opravi nacionalno raziskavo o poddiplomskem poučevanju mehanike na kolidžih in univerzah. Leta 1985 so na 66 izmed 100 kolidžev poučevali statiko tekočin v poddiplomskem programu strojništva. Leta 1975 pa je ta predmet poučevalo 43% kolidžev (Engineering Education, april 1986).

Izvedite test, s katerim boste določili ali se je odstotek kolidžev, ki poučujejo statiko tekočin, povečal med leti 1975 in 1985, če sklepamo da je bilo leta 1975 v raziskavo pravtako vključenih 100 kolidžev. Uporabite stopnjo značilnosti $\alpha = 0{\cdot}01$.

19. V študiji, katere namen je bilo ugotoviti vpliv t.i. multifunkcijske delovne postaje (MFDP) na način dela zaposlenih (Datamation, 15 februar 1986), sta sodelovali dve skupini svetovalcev iz varnostne agencije s sedežem v St. Louisu, in sicer skupina 12-ih svetovalcev, ki trenutno uporabljajo programsko opremo MFDP, ter kontrolna skupina 25-ih, ki je ne uporabljajo. Eno izmed vprašanj, zastavljenih udeležencem, se je dotikalo virov informacij. V testni skupini uporabnikov MFDP so štirje izjavili, da je računalnik njihov glavni vir informacij, medtem ko sta v kontrolni skupini isto izjavila dva njena člana.
    (a) Ali lahko sklepamo o razliki med deležem uporabnikov MFDP, ki se med vsemi viri informacij najbolj zanašajo na računalnik, ter istim deležem med neuporabniki MFDP? Test izvedi z $\alpha = 0{\cdot}1$.
    (b) Ali sta vzorca dovolj velika, da lahko uporabimo aproksimacijski postopek iz točke (a)?

20. Sisteme za ogrevanje hiš, ki izkoriščanjo sončno energijo lahko razdelimo na dve skupini - pasivne in aktivne. Pri pasivnih sistemih hiša deluje kot kolektor sončne energije, pri aktivnih sistemih pa se uporablja napredna mehanska oprema, ki pretvarja sončne žarke v toploto. V raziskavi je bilo vključenih 50 domov z pasivnimi sistemi in 50 domov z aktivnimi sistemi. Raziskovalci so zabeležili število domov, ki so porabili manj kot 200 galon kurilnega olja za ogrevanje v zadnjem letu. Ali rezultati raziskave pri stopnji značilnosti $\alpha = 0{\cdot}02$ nakazujejo na to, da se pasivni in aktivni sistemi razlikujejo po učinkovitosti?

	št. domov	št. domov, ki so porabili manj kot 200 galon kurilnega olja za ogrevanje
pasivni sončni sistemi	50	37
aktivni sončni sistemi	50	46

21. Najpogostejša metoda za dezinfekcijo vode za pitje je prosta rezidualna klorinacija. Pred kratkim pa je kot alternativa precej pozornosti pritegnil postopek preamoniacije (tj. dodajanja amoniaka vodi pred dodajanjem prostega klora). V eni študiji je bila preamoniacija izvedena na 44-ih vzorcih vode in se je izkazalo, da je imel indeks odtočne umazanosti povpreče $1{\cdot}8$ ter standardni odklon $0{\cdot}16$ (American Water Works Journal, januar 1986).

    Ali lahko sklepamo, da varianca indeksa odtočne umazanosti vodnih vzorcev, dezinficiranih s preamoniacijo, presega $0{\cdot}0016$? (Vrednost $0{\cdot}0016$ predstavlja doslej znano varianco omenjenega indeksa.) Pri testu vzemi $\alpha = 0{\cdot}01$.

22. Poliklorinirani bifenili (PCB-ji), ki se uporabljajo v proizvodnji velikih električnih transformatorjev in kondenzatorjev so izjemno nevarni onesnaževalci okolja. Agencija za zaščito okolja (EPA) preizkuša novo napravo za merjenje koncentracije PCB-jev v ribah. Za testiranje natančnosti nove naprave so naredili 7 merjenj koncentracije PCB-ja na isti ribi. Podatki so zabeleženi v spodnji tabeli (v delcih na milijon):

1	2	3	4	5	6	7
6·2	5·8	5·7	6·3	5·9	5·8	6·0

Ali nova naprava ustreza specifikacijam Agencije za zaščito okolja, če le ta zahteva, da naprave za merjenje koncentracije PCB-ja merijo z varianco manjšo od 0·1. Testirajte pri stopnji značilnosti $\alpha = 0{\cdot}05$.

23. (Nadaljevanje 8. naloge.) Spomnite se, da je bila izvedena študija za primerjavo povprečnih razmerij med oksoni in thioni v Kalifornijskem nasadu orhidej v dveh vremenskih pogojih - megla in jasno/oblačno vreme. Testirajte predpostavko, da so variance v različnih vremenskih pogojih enake, ker je to pogoj da je primerjava povprečij veljavna. Testirajte pri stopnji značilnosti $\alpha = 0{\cdot}05$.

24. (Nadaljevanje 11. naloge - o branju govora z ust.) Za osebe z normalnim sluhom je bil izveden še en eksperiment za primerjavo razpršenosti v zaznavanju pogovornega jezika v skupini tistih brez izkušenj v branju z ust in skupino tistih z izkušnjami v branju z ust. Vzorec je sestavljalo 24 neizkušenih in 12 izkušenih oseb. Vsaka oseba je morala ustno ponoviti stavke pod različnimi pogoji, med katerim je bil tudi dodatek za branje govora s podatkom o poudarku zlogov. Povzetek rezultatov (procent pravilnih zlogov) za obe skupini je podan v tabeli. Testirajte ali obstaja razlika med variancama v procentu pravilno reproduciranih zlogov med obema skupinama. Pri testu porabite $\alpha = 0{\cdot}10$.

neizkušeni	izkušeni
$n_1 = 24$	$n_2 = 12$
$\overline{y}_1 = 87{\cdot}1$	$\overline{y}_2 = 86{\cdot}1$
$s_1 = 8{\cdot}7$	$s_2 = 12{\cdot}4$

B.15 Bivariatna analiza in regresija

B.16 Časovne vrste in trendi

Dodatek C

Zgodovina in biografije iz verjetnosti in statistike

Kombinatoriko je definiral Jakob Bernoulli v svoji knjigi *Ars Conjectandi*, ki je izšla po njegovi smrti v Baslu leta 1713. Osrednja ideja kombinatorike je preštevanje:

1. koliko bo treba prebrati za izpit?[1]
2. diagonale n-kotnika
3. število 7. mestnih dvojiških števil: $2^7 = 128$
4. L. Euler in N. Bernoulli sta si dopisovala o *problemu zamenjave pisem* in ga tudi rešila.
5. Več faz, kombinatorično drevo ...

Zgodovina: Rimski patricij Boetius v 5. stoletju, indijski matematik Bhaskara v 12. stoletju, Židovski učenjak iz Avignona Levi ben Gerson v 14. stoletju (ki je znal izračunati permutacije, kombinacije in variacije brez ponavljanja). Leta 1494 Luca Pacioli v knjigi *Suma de aritmeticae* na koliko načinov se lahko 10 (ali je to pravo število) vsede v prvo vrsto. 16. stoletje: Cardan, Tartaglio, Leibniz, Buteom 17. stoletje: Pascal Fermat, Jakob Bernoulli

ime		dela
1. Blaise Pascal	(1623-1662)	pisma (Fermatu)
2. Jacob Bernoulli	(1654-1705)	Umetnost ugibanja 1713 (Ars Conjectandi)
3. Abraham de Moivre	(1667-1754)	The Doctrine of Chances 1718, 1733
4. Thomas Bayes	(1702-1761)	
5. Leonhard Euler	(1707-1783)	
6. Pierre-Simon Laplace	(1749-1827)	
7. Frederic Gauss	(1777-1855)	Disquisitiones Arithmeticae, 1801
8. Simeon Poisson	(1781{1840)	
9. Augustin Louis Cauchy	(1789-1857)	

[1]Pa poskusimo na hitro oceniti, kaj imamo pred seboj: 36 vrstic $\times$ 86 črk je $\approx 3K$, 100 strani je torej 300K, 200 strani 600K in čez 300 strani skoraj 1M.

```
10. William Sealy Gosset  (1876-1937)
11. Emile Borel           (1871-1956)
12. Sir Ronald A. Fisher  (1890-1962)
13. Frank Plumpton Ramsey (1903-1930)
14. Andrei Kolmogorov     (1903-1987)
15. Paul Erd\" os         (1913-1996)
```

C.1 Blaise Pascal (1623-1662)

French mathematician, philosopher, and religious figure. He studied the region above the mercury in a barometer, maintaining that it was a vacuum. In his investigations of the barometer, he found that the height to which the mercury rose was the same regardless of shape. Based on his double vacuum experiment, he formulated Pascal's principle, which states that the pressure is constant throughout a static fluid. He performed an experiment in which he convinced his brother-in-law to climb the Puy-de-Dôme Mountain in France. He found the height of the mercury dropped with altitude, indicating pressure decreases with altitude.

Pascal also designed and built mechanical adding machines, and incorporated a company in 1649 to produce and market them. Unfortunately, the machines were rather highly priced and had reliability problems. Only seven of Pascal's devices survive today.

Pascal suffered from serious health problems, and spent most of his final years writing on religious philosophy. http://scienceworld.wolfram.com/biography/Pascal.html

C.2 Matematični Bernoulliji

Redke družine so prispevale k matematiki toliko, kot Bernoullijevi iz švicarskega Basla. Kar sedem Bernoullijev iz treh generacij med leti 1680 in 1800 je bilo odličnih matematikov. Pet izmed njih je pomagalo zgraditi novo matematično teorijo verjetnosti.

Jacob (1654–1705) (glej tudi naslednji razdelek), in *Johann* (1667–1748) sta bila otroka uspešnega švicarskega trgovca, vendar sta matematiko študirala proti volji svojega praktičnega očeta. Oba sta bila med najboljšimi matematiki svojega časa, vendar je bil Jacob tisti, ki se je osredotočil na verjetnost. Bil je prvi, ki je jasno videl idejo z dolgoročnimi povprečji kot način za merjenje slučajnosti.

Johannov sin *Daniel* (1700–1782) ter Jacobov in Johannov nečak *Nicholas* (1687–1759) sta se prav tako ukvarjala z verjetnostjo. Nicholas je opazil, da lahko z verjetnostjo pojasnimo vzorec v rojstvu dečkov in deklic. Kljub temu da se je tudi sam upiral očetovim željam, je Johann želel, da bi njegov sin Daniel postal trgovec ali zdravnik, a Daniela to ni odvrnilo in je vseeno postal še en matematik iz družine Bernoullijev. Na področju verjetnosti se je ukvarjal s pravičnim določanjem cen v igrah naključij, poleg tega pa je dokazal učinkovitost cepiva proti kozam.

Matematična družina Bernoullijev je, tako kot njihovi glasbeni sodobniki iz družine Bach, nenavaden primer nadarjenosti za določeno področje, ki se pokaže v zaporednih generacijah. Delo Bernoullijev je pomagalo verjetnosti, da je od svojega rojstva v svetu hazarda zrasla do spoštovanega orodja svetovne uporabnosti.

C.3 Jacob Bernoulli (1654-1705)

Jacob Bernoulli (also referred to as Jacques or James) was born in Basel Switzerland on December 27, 1654, the first of the famous Bernoulli family of Swiss mathematicians. He first studied theology, but then against his father's wishes, turned to mathematics, astronomy, and physics. Bernoulli was appointed professor of mathematics at the University of Basel in 1687.

Bernoulli's mathematical work mostly involved the new theory of calculus. He worked on infinite series, conics, cycloids, transcendental curves, isoperimetry, isochronous curves, catenaries, and the logarithmic spiral, publishing his results widely in Acta Eruditorum. Bernoulli also wrote what is considered the second book devoted to probability, Ars Conjectandi, which was published after his death in 1713. Bernoulli formulated the version of the law of large numbers for independent trials, now called Bernoulli trials in his honor, and studied the binomial distribution. Jacques Bernoulli died on August 16, 1705 in Basel. The logarithmic spiral is carved into his tombstone.

`http://www.math.uah.edu/STAT/biographies/Bernoulli.xhtml`

C.4 Abraham de Moivre (1667-1754)

A French Huguenot, de Moivre was jailed as a Protestant upon the revocation of the Edict of Nantes in 1685. When he was released shortly thereafter, he fled to England. In London he became a close friend of Sir Isaac Newton and the astronomer Edmond Halley. De Moivre was elected to the Royal Society of London in 1697 and later to the Berlin and Paris academies. Despite his distinction as a mathematician, he never succeeded in securing a permanent position but eked out a precarious living by working as a tutor and a consultant on gambling and insurance.

De Moivre expanded his paper "De mensura sortis" (written in 1711), which appeared in Philosophical Transactions, into The Doctrine of Chances (1718). Although the modern theory of probability had begun with the unpublished correspondence (1654) between Blaise Pascal and Pierre de Fermat and the treatise De Ratiociniis in Ludo Aleae (1657; "On Ratiocination in Dice Games") by Christiaan Huygens of Holland, de Moivre's book greatly advanced probability study. The definition of statistical independence—namely, that the probability of a compound event composed of the intersection of statistically independent events is the product of the probabilities of its components—was first stated in de Moivre's Doctrine. Many problems in dice and other games were included, some of which appeared in the Swiss mathematician Jakob (Jacques) Bernoulli's Ars conjectandi (1713; "The Conjectural Arts"), which was published before de Moivre's Doctrine but after his "De mensura." He derived the principles of probability from the mathematical expectation of events, just the reverse of present-day practice.

De Moivre's second important work on probability was Miscellanea Analytica (1730; "Analytical Miscellany"). He was the first to use the probability integral in which the integrand is the exponential of a negative quadratic. He originated Stirling's formula. In 1733 he used Stirling's formula to derive the normal frequency curve as an approximation of the binomial law. De Moivre was one of the first mathematicians to use complex numbers in trigonometry. The formula known by his name, $(\cos x + i \sin x)^n = \cos nx + i \sin nx$, was

instrumental in bringing trigonometry out of the realm of geometry and into that of analysis.
`http://www.britannica.com/EBchecked/topic/387796/Abraham-de-Moivre`

C.5 Thomas Bayes (1702-1761)

was an English clergyman who set out his theory of probability in 1764. His conclusions were accepted by Laplace in 1781, rediscovered by Condorcet, and remained unchallenged until Boole questioned them. Since then Bayes' techniques have been subject to controversy.
`http://www.york.ac.uk/depts/maths/histstat/bayesbiog.pdf`
`http://www.britannica.com/EBchecked/topic/56807/Thomas-Bayes`

English Nonconformist theologian and mathematician who was the first to use probability inductively and who established a mathematical basis for probability inference (a means of calculating, from the frequency with which an event has occurred in prior trials, the probability that it will occur in future trials. See probability theory: Bayes's theorem.

Bayes set down his findings on probability in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the Philosophical Transactions of the Royal Society. That work became the basis of a statistical technique, now called Bayesian estimation, for calculating the probability of the validity of a proposition on the basis of a prior estimate of its probability and new relevant evidence. Disadvantages of the method—pointed out by later statisticians—include the different ways of assigning prior distributions of parameters and the possible sensitivity of conclusions to the choice of distributions.

The only works that Bayes is known to have published in his lifetime are Divine Benevolence; or, An Attempt to Prove That the Principal End of the Divine Providence and Government Is the Happiness of His Creatures (1731) and An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst (1736), which was published anonymously and which countered the attacks by Bishop George Berkeley on the logical foundations of Sir Isaac Newton's calculus. Bayes was elected a fellow of the Royal Society in 1742.

C.6 Leonhard Euler (1707-1783)

Swiss mathematician who was tutored by Johann Bernoulli. He worked at the Petersburg Academy and Berlin Academy of Science. He had a phenomenal memory, and once did a calculation in his head to settle an argument between students whose computations differed in the fiftieth decimal place. Euler lost sight in his right eye in 1735, and in his left eye in 1766. Nevertheless, aided by his phenomenal memory (and having practiced writing on a large slate when his sight was failing him), he continued to publish his results by dictating them. Euler was the most prolific mathematical writer of all times finding time (even with his 13 children) to publish over 800 papers in his lifetime. He won the Paris Academy Prize 12 times. When asked for an explanation why his memoirs flowed so easily in such huge quantities, Euler is reported to have replied that his pencil seemed to surpass him in intelligence. François Arago said of him "He calculated just as men breathe, as eagles sustain themselves in the air"(Beckmann 1971, p. 143; Boyer 1968, p. 482).

Euler systematized mathematics by introducing the symbols e, i, and $f(x)$ for f a function of x. He also made major contributions in optics, mechanics, electricity, and magnetism.

He made significant contributions to the study of differential equations. His Introducio in analysin infinitorum (1748) provided the foundations of analysis. He showed that any complex number to a complex power can be written as a complex number, and investigated the beta and gamma functions. He computed the Riemann zeta function to for even numbers.

He also did important work in number theory, proving that that the divergence of the harmonic series implied an infinite number of Primes, factoring the fifth Fermat number (thus disproving Fermat's conjecture), proving Fermat's lesser theorem, and showing that e was irrational. In 1772, he introduced a synodic coordinates (rotating) coordinate system to the study of the three-body problem (especially the Moon). Had Euler pursued the matter, he would have discovered the constant of motion later found in a different form by Jacobi and known as the Jacobi integral.

Euler also found the solution to the two fixed center of force problem for a third body. Finally, he proved the binomial theorem was valid for any rational exponent. In a testament to Euler's proficiency in all branches of mathematics, the great French mathematician and celestial mechanic Laplace told his students, "Liesez Euler, Liesez Euler, c'est notre maître à tous"("Read Euler, read Euler, he is our master in everything"(Beckmann 1971, p. 153).

http://scienceworld.wolfram.com/biography/Euler.html

C.7 Pierre-Simon Laplace (1749–1827)

Born into the French middle class but died a Marquis having prospered as well during the French Revolution as during the monarchy. His name is well known to mathematicians both for his work on transformations and for his work on planetary motion. He had a deterministic view of the universe. Reputedly, his reply, when asked by Napoleon where God fitted in, was 'I have no need of that hypothesis'. Napoleon appointed him as Minister of the Interior — but removed him six weeks later for trying 'to carry the spirit of the infinitesimal into administration'. In Statistics he worked on many probability problems including the Laplace distribution: he is credited with independently discovering Bayes's Theorem. He was elected FRS in 1789 and FRSE in 1813. A street in Paris is named after him, as is the Promontorium Laplace on the Moon.

http://www.answers.com/topic/pierre-simon-laplace

C.8 Johann Carl Friedrich Gauss (1777–1855)

German mathematician who is sometimes called the "prince of mathematics." He was a prodigious child, at the age of three informing his father of an arithmetical error in a complicated payroll calculation and stating the correct answer. In school, when his teacher gave the problem of summing the integers from 1 to 100 (an arithmetic series) to his students to keep them busy, Gauss immediately wrote down the correct answer 5050 on his slate. At age 19, Gauss demonstrated a method for constructing a heptadecagon using only a straightedge and compass which had eluded the Greeks. (The explicit construction of the heptadecagon was accomplished around 1800 by Erchinger.) Gauss also showed that only regular polygons of a certain number of sides could be in that manner (a heptagon, for example, could not be constructed.)

Gauss proved the fundamental theorem of algebra, which states that every polynomial has a root of the form $a+bi$. In fact, he gave four different proofs, the first of which appeared in his dissertation. In 1801, he proved the fundamental theorem of arithmetic, which states that every natural number can be represented as the product of primes in only one way.

At age 24, Gauss published one of the most brilliant achievements in mathematics, Disquisitiones Arithmeticae (1801). In it, Gauss systematized the study of number theory (properties of the integers). Gauss proved that every number is the sum of at most three triangular numbers and developed the algebra of congruences.

In 1801, Gauss developed the method of least squares fitting, 10 years before Legendre, but did not publish it. The method enabled him to calculate the orbit of the asteroid Ceres, which had been discovered by Piazzi from only three observations. However, after his independent discovery, Legendre accused Gauss of plagiarism. Gauss published his monumental treatise on celestial mechanics Theoria Motus in 1806. He became interested in the compass through surveying and developed the magnetometer and, with Wilhelm Weber measured the intensity of magnetic forces. With Weber, he also built the first successful telegraph.

Gauss is reported to have said "There have been only three epoch-making mathematicians: Archimedes, Newton and Eisenstein" (Boyer 1968, p. 553). Most historians are puzzled by the inclusion of Eisenstein in the same class as the other two. There is also a story that in 1807 he was interrupted in the middle of a problem and told that his wife was dying. He is purported to have said, "Tell her to wait a moment 'till I'm through" (Asimov 1972, p. 280).

Gauss arrived at important results on the parallel postulate, but failed to publish them. Credit for the discovery of non-Euclidean geometry therefore went to Janos Bolyai and Lobachevsky. However, he did publish his seminal work on differential geometry in Disquisitiones circa superticies curvas. The Gaussian curvature (or "second" curvature) is named for him. He also discovered the Cauchy integral theorem for analytic functions, but did not publish it. Gauss solved the general problem of making a conformal map of one surface onto another.

Unfortunately for mathematics, Gauss reworked and improved papers incessantly, therefore publishing only a fraction of his work, in keeping with his motto "pauca sed matura" (few but ripe). Many of his results were subsequently repeated by others, since his terse diary remained unpublished for years after his death. This diary was only 19 pages long, but later confirmed his priority on many results he had not published. Gauss wanted a heptadecagon placed on his gravestone, but the carver refused, saying it would be indistinguishable from a circle. The heptadecagon appears, however, as the shape of a pedestal with a statue erected in his honor in his home town of Braunschweig.

C.9 Simeon Poisson (1781–1840)

"Life is good for two things, learning mathematics and teaching mathematics."

(b. Pithviers, d. Paris). Simeon Poisson developed many novel applications of mathematics for statistics and physics. His father had been a private soldier, and on his retirement was given a small administrative post in his native village. When the French revolution broke out, his father assumed the government of the village, and soon became a local dignitary.

He was educated by his father who prodded him to be a doctor. His uncle offered to teach him medicine, and began by making him prick the veins of cabbage-leaves with a lancet. When he had perfected this, he was allowed to practice on humans, but in the first case that he did this by himself, the patient died within a few hours. Although the other physicians assured him that this was not an uncommon occurance, he vowed he would have nothing more to do with the medical profession.

Upon returning home, he discovered a copy of a question set from the Polytechnic school among the official papers sent to his father. This chance event determined his career. At the age of seventeen he entered the Polytechic. A memoir on finite differences which he wrote when only eighteen was so impressive that it was rapidly published in a prestigious journal. As soon as he had finished his studies he was appointed as a lecturer. Throughout his life he held various scientific posts and professorships. He made the study of mathematics his hobby as well as his business.

Over his life Simeon Poisson wrote between 300-400 manuscripts and books on a variety of mathematical topics, including pure mathematics, the application of mathematics to physical problems, the probability of random events, the theory of electrostatics and magnetism (which led the forefront of the new field of quantum mechanics), physical astronomy, and wave theory.

One of Simeon Poisson's contributions was the development of equations to analyze random events, later dubbed the Poisson Distribution. The fame of this distribution is often attributed to the following story. Many soldiers in the Prussian Army died due to kicks from horses. To determine whether this was due to a random occurance or the wrath of god, the Czar commissioned the Russian mathematician Ladislaus Bortkiewicz to determine the statistical significance of the events. Fourteen corps were examined,

C.10 William Sealy Gosset (1876-1937)

Leta 1908 je pod psevdonimom 'Student' objavil članek, v katerem je vpeljal Studentovo t-porazdelitev. He is famous as a statistician, born in Canterbury, England. attended Winchester College After graduating chemistry and mathematics at New College, Oxford in 1899, he joined the Dublin brewery of Arthur Guinness & Son. Guinness was a progressive agrochemical business and Gosset would apply his statistical knowledge both in the brewery and on the farm—to the selection of the best yielding varieties of barley. Gosset acquired that knowledge by study, trial and error and by spending two terms in 1906–7 in the biometric laboratory of Karl Pearson. Gosset and Pearson had a good relationship and Pearson helped Gosset with the mathematics of his papers. Pearson helped with the 1908 papers but he had little appreciation of their importance. The papers addressed the brewer's concern with small samples, while the biometrician typically had hundreds of observations and saw no urgency in developing small-sample methods. Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. This meant that Gosset was unable to publish his works under his own name. He therefore used the pseudonym Student for his publications to avoid their detection by his employer. Thus his most famous achievement is now referred to as Student's t-distribution, which might otherwise have been Gosset's t-distribution.

Gosset had almost all of his papers including The probable error of a mean published in Pearson's journal Biometrika using the pseudonym Student. However, it was R. A. Fisher who appreciated the importance of Gosset's small-sample work, after Gosset had written to him to say I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them! Fisher believed that Gosset had effected a "logical revolution". Ironically the t-statistic for which Gosset is famous was actually Fisher's creation. Gosset's statistic was $z = t/\sqrt{n-1}$. Fisher introduced the t-form because it fit in with his theory of degrees of freedom. Fisher was also responsible for the applications of the t-distribution to regression.

Although introduced by others, Studentized residuals are named in Student's honor because, like the problem that led to Student's t-distribution, the idea of adjusting for estimated standard deviations is central to that concept.

Gosset's interest in barley cultivation led him to speculate that design of experiments should aim, not only at improving the average yield, but also at breeding varieties whose yield was insensitive (robust) to variation in soil and climate. This principle only occurs in the later thought of Fisher and then in the work of Genichi Taguchi in the 1950s.

In 1935, he left Dublin to take up the position of Head Brewer, in charge of the scientific side of production, at a new Guinness brewery at Park Royal in North West London. He died in Beaconsfield, England of a heart attack.

Gosset was a friend of both Pearson and Fisher, an achievement, for each had a massive ego and a loathing for the other. Gosset was a modest man who cut short an admirer with the comment that "Fisher would have discovered it all anyway."

C.11 Emile Borel (1871–1956)

French mathematician educated at the École Normale Supérieure (1889-1892) who received is Docteur es sciences (1894). He was appointed "Maitre de conférence" at the University of Lille (1893), and subsequently at the École Normale Supérieure in 1897. He was professor of the theory of functions at the Sorbonne (1909-1920) and professor of Calcul des Probabilités et de Physique mathématiques (1920-1941). He was scientific director of the École Normale Supérieure (1911), and became a member of the Académie des sciences in 1921. He was also a member of French Chamber of Deputies (1924-1936), and Minister of the Navy for a few months in 1925 (not the fifteen years claimed in many biographies).

Borel worked on divergent series, the theory of functions, probability, game theory, and was the first to define games of strategy. In particular, he found an elementary proof of Picard's theorem in 1896. Borel founded measure theory, which is the application of the theory of sets to the theory of functions, thus becoming founding with Lebesgue and Baire of modern theory of functions of real variables. Borel's work culminated in the Heine-Borel theorem. He showed that a "sum" could be defined for some divergent series. Borel wrote more thirty books and near 300 papers, including a number of popular works of high scientific quality, and devoted more fifty papers to history and philosophy of sciences.
`http://scienceworld.wolfram.com/biography/Borel.html`

C.12 Augustin Louis Cauchy (1789-1857)

`http://www-history.mcs.st-and.ac.uk/Mathematicians/Cauchy.html`
Cauchy pioneered the study of analysis, both real and complex, and the theory of permutation groups. He also researched in convergence and divergence of infinite series, differential equations, determinants, probability and mathematical physics.

`http://scienceworld.wolfram.com/biography/Cauchy.html`
French mathematician who wrote 789 papers, a quantity exceeded only by Euler and Cayley,

which brought precision and rigor to mathematics. He invented the name for the determinant and systematized its study and gave nearly modern definitions of limit, continuity, and convergence. Cauchy founded complex analysis by discovering the Cauchy-Riemann equations (although these had been previously discovered by d'Alembert).

Cauchy also presented a mathematical treatment of optics, hypothesized that ether had the mechanical properties of an elasticity medium, and published classical papers on wave propagation in liquids and elastic media. After generalizing Navier's equations for isotropic media, he formulated one for anisotropic media. Cauchy published his first elasticity theory in 1830 and his second in 1836. Both were rather ad hoc and were riddled with problems, and Cauchy proposed a third theory in 1839. Cauchy also studied the reflection from metals and dispersion relationships.

Cauchy extended the polyhedral formula in a paper which was criticized by Malus. His theory of substitutions led to the theory of finite groups. He proved that the order of any subgroup is a divisor of the order of the group. He also proved Fermat's three triangle theorem. He refereed a long paper by Le Verrier on the asteroid Pallas and invented techniques which allowed him to redo Le Verrier's calculations at record speed. He was a man of strong convictions, and a devout Catholic. He refused to take an oath of loyalty, but also refused to leave the French Academy of Science.

C.13 Sir Ronald A. Fisher(1890-1962)

Ideje in metode, ki jih danes poznamo pod imenom *statistika*, so v devetnajstem in dvajsetem stoletju izumili ljudje, ki so se ukvarjali s problemi, pri katerih je bilo potrebno analizirati velike količine podatkov. Astronomija, biologija, družboslovne vede in celo geodezija lahko trdijo, da so igrale pomembno vlogo pri rojstvu statistike. Če pa si kdo zasluži naziv "oče statistike", je to Sir Ronald A. Fisher.

Fisherjevi zapiski so opredelili statistiko kot posebno področje proučevanja, katerega metode lahko uporabimo v številnih panogah. Sistematiziral je matematično teorijo statistike in izumil številne nove metode. Slučajeni primerjalni eksperiment je najbrž njegov največji dosežek.

Kot druge statistične pionirje so tudi Fisherja gnale zahteve praktičnih problemov. Od leta 1919 naprej je delal na področju kmetijstva na terenu v Rothamstedu v Angliji. Kako naj razporedimo sajenje različnih vrst pridelka ali pa uporabo različnih gnojil, da jih lahko primerjamo? Ker se rodovitnost in druge lastnosti spreminjajo, ko se premikamo po polju, so eksperimentatorji uporabljali zamotane vzorce, ki so posnemali šahovnico. Fisher je imel boljšo idejo: namenoma uredimo parcele naključno.

C.14 Frank Plumpton Ramsey (1903-1930)

The Cambridge philosopher Ramsey, a wunderkind of first order, wrote three important contributions to economics.

The first, "Truth and Probability" (written in 1926, published 1931), was the first paper to lay out the theory of subjective probability and begin to axiomatize choice under (subjective) uncertainty, a task completed decades later by Bruno de Finetti and Leonard Savage. (This was written in opposition to John Maynard Keynes's own information-theoretic Treatise on Probability.) Ramsey's second contribution was his theory of taxation (1927), generating the famous "Boiteux-Ramsey"pricing rule. Ramsey's third contribution was his exercise in determining optimal savings (1928), the famous "optimal growth"model - what has since become known as the "Ramsey model- one of the earliest applications of the calculus of variations to economics.

Frank Ramsey died on January 27, 1930, just before his 27th birthday. In his tragically short life he produced an extraordinary amount of profound and original work in economics, mathematics and logic as well as in philosophy: work which in all these fields is still extremely influential.

C.15 Andrei Kolmogorov (1903-1987)

`http://www.exploratorium.edu/complexity/CompLexicon/kolmogorov.html`

Kolmogorov was one of the broadest of this century's mathematicians. He laid the mathematical foundations of probability theory and the algorithmic theory of randomness and made crucial contributions to the foundations of statistical mechanics, stochastic processes, information theory, fluid mechanics, and nonlinear dynamics. All of these areas, and their interrelationships, underlie complex systems, as they are studied today. Kolmogorov graduated from Moscow State University in 1925 and then became a professor there in 1931. In 1939 he was elected to the Soviet Academy of Sciences, receiving the Lenin Prize in 1965 and the Order of Lenin on seven separate occasions.

His work on reformulating probability started with a 1933 paper in which he built up probability theory in a rigorous way from fundamental axioms, similar to Euclid's treatment of geometry. Kolmogorov went on to study the motion of the planets and turbulent fluid flows, later publishing two papers in 1941 on turbulence that even today are of fundamental importance. In 1954 he developed his work on dynamical systems in relation to planetary motion, thus demonstrating the vital role of probability theory in physics and re-opening the study of apparent randomness in deterministic systems, much along the lines originally conceived by Henri Poincare.

In 1965 he introduced the algorithmic theory of randomness via a measure of complexity, now referred to Kolmogorov Complexity. According to Kolmogorov, *the complexity of an object is the length of the shortest computer program that can reproduce the object.* Random objects, in his view, were their own shortest description. Whereas, periodic sequences have low Kolmogorov complexity, given by the length of the smallest repeating "template" sequence they contain. Kolmogorov's notion of complexity is a measure of randomness, one that is closely related to Claude Shannon's entropy rate of an information source.

Kolmogorov had many interests outside mathematics research, notable examples being the quantitative analysis of structure in the poetry of the Russian author Pushkin, studies of agrarian development in 16th and 17th century Novgorod, and mathematics education.

C.16 Paul Erdős (1913-1996)

`http://www.maa.org/mathland/mathland_10_7.html`
`http://www.britannica.com/EBchecked/topic/191138/Paul-Erdos: Paul Hoffman`
Hungarian "freelance" mathematician (known for his work in number theory and combinatorics) and legendary eccentric who was arguably the most prolific mathematician of the 20th century, in terms of both the number of problems he solved and the number of problems he

convinced others to tackle. Erdös did mathematics with a missionary zeal, often 20 hours a day, turning out some 1,500 papers, an order of magnitude higher than his most prolific colleagues produced. He was active to the last days of his life. At least 50 papers on which he is listed as a coauthor are yet to appear, representing the results of various recent collaborative efforts. His interests were mainly in number theory and combinatorics, though they ranged into topology and other areas of mathematics. He was fascinated by relationships among numbers, and numbers served as the raw materials for many of his conjectures, questions, and proofs.

In 1930, at age 17, Erdös entered the Péter Pázmány University in Budapest, where in four years he completed his undergraduate work and earned a Ph.D. in mathematics. As a college freshman (18), he made a name for himself in mathematical circles with a stunningly simple proof of Chebyshev's theorem, which says that a prime can always be found between any integer n (greater than 1) and its double $2n$. A little later, he proved his own theorem that there is always a prime of the form $4k+1$ and $4k+3$ between n and $2n$. For example, the interval between 100 and 200 contains the prime-number pair 101 and 103 ($k = 25$). Paul Erdös has the theory that God has a book containing all the theorems of mathematics with their absolutely most beautiful proofs, and when [Erdös] wants to express particular appreciation of a proof, he exclaims, 'This is one from the book!' During his university years he and other young Jewish mathematicians, who called themselves the Anonymous group, championed a fledgling branch of mathematics called Ramsey theory.

In 1934 Erdös, disturbed by the rise of anti-Semitism in Hungary, left the country for a four-year postdoctoral fellowship at the University of Manchester in England. In September 1938 he emigrated to the United States, accepting a one-year appointment at the Institute for Advanced Study in Princeton, New Jersey, where he cofounded the field of probabilistic number theory. During the 1940s he wandered around the United States from one university to the next—Purdue, Stanford, Notre Dame, Johns Hopkins—spurning full-time job offers so that he would have the freedom to work with anyone at any time on any problem of his choice. Thus began half a century of nomadic existence that would make him a legend in the mathematics community. With no home, no wife, and no job to tie him down, his wanderlust took him to Israel, China, Australia, and 22 other countries (although sometimes he was turned away at the border—during the Cold War, Hungary feared he was an American spy, and the United States feared he was a communist spy). Erdös would show up—often unannounced—on the doorstep of a fellow mathematician, declare "My brain is open!" and stay as long as his colleague served up interesting mathematical challenges.

In 1949 Erdös had his most satisfying victory over the prime numbers when he and Atle Selberg gave The Book proof of the prime number theorem (which is a statement about the frequency of primes at larger and larger numbers). In 1951 John von Neumann presented the Cole Prize to Erdös for his work in prime number theory. In 1959 Erdös attended the first International Conference on Graph Theory, a field he helped found. During the next three decades he continued to do important work in combinatorics, partition theory, set theory, number theory, and geometry—the diversity of the fields he worked in was unusual. In 1984 he won the most lucrative award in mathematics, the Wolf Prize, and used all but \$720 of the \$50,000 prize money to establish a scholarship in his parents' memory in Israel. He was elected to many of the world's most prestigious scientific societies, including the Hungarian Academy of Science (1956), the U.S. National Academy of Sciences (1979), and the British Royal Society (1989). Defying the conventional wisdom that mathematics was a young man's game, Erdös went on proving and conjecturing until the age of 83, succumbing to a heart attack only hours after disposing of a nettlesome problem in geometry at a conference in Warsaw.

Erdös had once remarked that mathematics is eternal because it has an infinity of problems. In the same spirit, his own contributions have enriched mathematics. Erdös problems – solved and unsolved – abound in the mathematical literature, lying in wait to provoke thought and elicit surprise.

Erdös loved problems that people could understand without learning a mass of definitions. His hallmark was the deceptively simple, precisely stated problem and the succinct and ingenious argument to settle the issue. Though simply stated, however, his problems were often notoriously difficult to solve. Here's a sample, not-so-difficult Erdös problem that concerns sequences of $+1$'s and -1's. Suppose there are equal numbers of $+1$'s and -1's lined up in a row. If there are two $+1$'s and two -1's, for example, a row could consist of $+1 +1 -1 -1$. Because these terms can be listed in any order, there are in fact six different ways to write such a row. Of course, the sum of all the numbers in a row is zero. However, it's interesting to look at the partial sums in each row. In the example above, the partial sums are $+1$ (after one term), $+2$ (after two terms), $+1$ (after three terms), and 0 (after four terms). The problem is to determine how many rows out of all the possibilities yield no partial sum that is negative. Of the six different rows for $n = 2$, only two escape a negative partial sum. Of the 20 rows for $n = 3$, just five have exclusively nonnegative partial sums; for $n = 4$, 14 out of 70 rows have this particular characteristic; and so on. The answer turns out to be a sequence called the Catalan numbers: $1/(n + 1)$ times the number of different rows for n $+1$'s and n -1's. One can liken these rows to patrons lined up at a theater box office. The price of admission is 50 cents, and half the people have the exact change while the other half have one-dollar bills.Thus, each person provides one unit of change for the cashier's later use or uses up one unit of change. In how many ways can the patrons be lined up so that a cashier, who begins with no money of her own, is never stuck for change?

He turned mathematics into a social activity, encouraging his most hermetic colleagues to work together. The collective goal, he said, was to reveal the pages in the Book. Erdös himself published papers with 507 coauthors. In the mathematics community those 507 people gained the coveted distinction of having an "Erdös number of 1," meaning that they wrote a paper with Erdös himself. Someone who published a paper with one of Erdös's coauthors was said to have an Erdös number of 2, and an Erdös number of 3 meant that someone wrote a paper with someone who wrote a paper with someone who worked with Erdös. Albert Einstein's Erdös number, for instance, was 2. The highest known Erdös number is 15; this excludes nonmathematicians, who all have an Erdös number of infinity.

Erdös enjoyed offering monetary rewards for solving particular problems, ranging from \$10,000 for what he called "a hopeless problem" in number theory to \$25 for something that he considered not particularly difficult but still tricky, proposed in the middle of a lecture. One problem worth a \$3,000 reward concerns an infinite sequence of integers, the sum of whose reciprocals diverges. The conjecture is that such a sequence contains arbitrarily long arithmetic progressions. "This would imply that the primes contain arbitrarily long arithmetic progressions," Erdös remarked. "This would be really nice. And I don't expect to have to pay this money, but I should leave some money for it in case I leave."

Dodatek D

PROGRAM R (Martin Raič)

Predstavili bomo osnovne ukaze v programu R, ki so povezani z našim predmetom.

Informacije

Dokumentacija na Linuxu: `/usr/share/doc/r-doc-html/manual` .

Pomoč v R-ovem pozivniku:

- `help(točno določena funkcija)`
- `help.search("nekaj približnega")`

D.1 Izvajanje programa

Iz pozivnika našega operacijskega sistema program zaženemo z ukazom `R`. Če ni določeno drugače, se znajdemo v R-ovem pozivniku. Program lahko zaženemo z obilo opcijami. Njihov seznam izpiše ukaz `R -h` ali `R --help` (in nato takoj konča). R ne zažene pozivnika in le izpiše rezultat, nakar konča, če mu na vhodu predpišemo ukazni niz. To pa je lahko:

- Standardni vhod, če R zaženemo kot cevovod ali pa z dostavkom `< datoteka`.
  V tem primeru mu moramo predpisati še opcijo `--save`, `--no-save` ali `--vanilla`.
- Vsebina datoteke z imenom, ki sledi opciji `-f` oziroma `--file`.
- Ukazni niz, ki sledi opciji `-e`.

Pri izvajanju R-a na ta način pride prav opcija `--slave`, ki izključi ves nenujni izhod.

Zadnjo vrnjeno vrednost dobimo z ukazom `.Last.value`.

Izhod iz programa dosežemo z ukazom `q()`.

D.2 Aritmetika

Elementarne binarne operacije: `+`, `−`, `*`, in `**`

Elementarne funkcije: `sqrt`, `exp`, `log`, `sin`, `cos`, `tan`, `asin`, `acos`, `atanr`

Konstanta: `pi`

Izpis na določeno število (n) signifikantnih decimalk: `print(x, digits=n)`
(Več o izpisovanju kasneje).

Zaokrožitvene funkcije: `trunc`, `round`, `floor`, `ceiling`
Fakulteta: `factorial`
Funkcija gama: `gamma`
Binomski simbol: `choose(n, k)` vrne n nad k, tj. $\binom{n}{k}$.
Naključna števila: `runif(1)` vrne psevdonaključno število med 0 in 1 z enakomerno porazdelitvijo. Več o naključnih številih kasneje.

D.3 Najosnovnejše o spremenljivkah

Prireditev: `x <- nekaj` ali `nekaj -> x`

Izbris: `rm(x)` ali tudi `rm(x, y)`.

Ukaz `ls` vrne seznam vseh simbolov, ki so trenutno definirani, razen tistih, katerih imena se začenjajo s piko. Če želimo vključiti še te, vnesemo `ls(all=TRUE)`.
Ukaz `rm(list=ls(all=TRUE))` izbriše vse trenutno definirane simbole.

Osnovni podatkovni tipi:

- števila: cela (npr. `integer(-42)`), realna (npr. `1.23`) in kompleksna (npr. `2 + 1i`);
- nizi znakov (npr. `"Zemlja"`);
- logične vrednosti: `TRUE`, `FALSE`;
- prazna vrednost: `NULL`.

D.4 Uporabnikove funkcije

Anonimna funkcija: `function(parametri) telo`
Funkcija z imenom: `ime <- function(parametri) telo`

Ukaz `plot(funkcija, sp. meja, zg. meja)` nariše graf funkcije.

D.5 Numerično računanje

`uniroot(f, c(a, b))` vrne ničlo zvezne funkcije f na intervalu $[a, b]$. Vrednosti na krajiščih morata imeti nasproten predznak. Vselej vrne le eno ničlo. Pravzaprav vrne celotno poročilo o ničli. Če želimo le ničlo, ukažemo: `uniroot(f, c(a, b))$root`.

`integrate(f, a, b)` numerično integrira funkcijo f od a do b. Spet vrne celo poročilo, če želimo le vrednost, ukažemo: `integrate(f, a, b)$value`.

D.6 Podatkovne strukture

D.6.1 Vektorji

Primer: Konstrukcija vektorja: `c(7, 8, 9)` da isto kot `7:9` ali `seq(7, 9)`. ◇

Pri ukazu `seq` lahko predpišemo tudi korak, recimo `seq(70, 90, 10)`. `runif(n)` vrne naključni vektor dolžine n. Več o tem kasneje. Vektorje lahko tvorimo tudi iz nizov:

```
x <- c("Merkur", "Venera", "Zemlja", "Mars", "Jupiter", "Saturn", "Uran", "Neptun").
```

Ukaz `c` tudi združuje vektorje.
POZOR! Vsi elementi v vektorju morajo biti istega tipa. Če so tipi različni, se nižji tipi pretvorijo v višje.

Primeri:

- `c(1, 2, "Zemlja")` se pretvori v `c("1", "2", "Zemlja")`.
- `c(1, 2, TRUE)` se pretvori v `c(1, 2, 1)`.
- `c(1, 2, TRUE,"Zemlja")` se pretvori v `c("1", "2", "TRUE", "Zemlja")`. ◇

Če ni določeno drugače, ukaz `c` izpusti vrednosti `NULL`. Z ukazom `[ ... ]` dobimo ven posamezne komponente vektorja. Natančneje, če je v vektor, ukaz `v[i]` deluje odvisno od narave objekta i na naslednji način:

- Če je i naravno število, vrne element z indeksom i. Indeksi se štejejo od 1 naprej.
- Če je i negativno celo število, vrne vektor brez elementa z ustreznim indeksom.
- Če je i vektor iz naravnih števil, vrne vektor iz elementov z ustreznimi indeksi (primerno za komponiranje preslikav).
  POZOR! Primer tovrstne kode je `v[c(2, 4, 3, 2)]` in ne recimo `v[2, 4, 3, 2]`: slednje pomeni večrazsežni indeks v tabeli – glej kasneje.

- Če je i vektor iz negativnih celih števil, vrne vektor, ki ima ustrezne elemente izpuščene.
- Če je i vektor iz logičnih vrednosti, vrne vektor iz komponent vektorja v, pri katerih je na odgovarjajočem položaju v i vrednost `TRUE`.

Ponovitev elementa ali vektorja: `rep(vrednost ali vektor, kolikokrat)`.

Obrat vektorja: `rev(x)` vrne vektor `x` od zadaj naprej.

`plot(x)` nariše točke, ki pripadajo koordinatam vektorja x.
To je isto kot `plot(x, type="p")`.

Druge opcije za ukaz `type`:

- `"l"`: nariše lomljenko;
- `"b"`: točke poveže z daljicami;
- `"h"`: nariše histogram iz navpičnih črt;
- `"s"`: nariše stopnice, pri čemer gre posamezna stopnica desno od točke;
- `"S"`: nariše stopnice, pri čemer gre posamezna stopnica levo od točke;

Ukaz `barplot(x)` nariše vektor x v obliki lepega histograma.

D.6.2 Matrike

Matriko:

```
 1   2   3
 4   5   6
 7   8   9
10  11  12
```

vnesemo z enim izmed naslednjih ukazov:

- `matrix(c(1, 4, 7, 10, 2, 5, 8, 11, 3, 6, 9, 12), nrow=4, ncol=3)`
- `matrix(1:12, nrow=4, ncol=3, byrow=TRUE)`
- `array(c(1, 4, 7, 10, 2, 5, 8, 11, 3, 6, 9, 12), dim=c(4, 3))`
- `rbind(1:3, 4:6, 7:9, 10:12)`
- `cbind(c(1, 4, 7, 10), c(2, 5, 8, 11), c(3, 6, 9, 12))`

Preprost je tudi vnos diagonalnih matrik, npr.

`diag(3, nrow=5)` ali `diag(c(3, 2, 4, 5, 1))`.

Priklic elementov matrike:

- `A[i, j]` vrne element v i-ti vrstici in j-tem stolpcu matrike A.

- `A[i,]` vrne i-to vrstico matrike A.
- `A[,j]` vrne j-ti stolpec matrike A.
- `A[i]` vrne i-ti element matrike, pri čemer so elementi urejeni po stolpcih. Tako pri zgornji matriki `A[8]` vrne 11.

Ukaz `c(A)` ali `as.vector(A)` iz matrike A naredi dolg vektor, pri čemer združuje po stolpcih.

Vse aritmetične operacije delujejo tudi na vektorjih in matrikah – po komponentah. Tako `A*B` zmnoži matriki A in B po komponentah. Ukaz `A + 2` pa vrne matriko A, ki ima vse elemente povečane za 2. Seveda lahko matrike in vektorje posredujemo tudi funkcijam. Recimo `(function(x)  x ** 3 + x )(0:3)` vrne isto kot `c(0, 2, 10, 30)`.

Matrične operacije:

- `%*%`: matrično množenje,
- `%o%`: tenzorski produkt – množenje vsake komponente z vsako,
- `t`: transponiranje,
- `det`: determinanta,
- `solve(A, B)`: reši sistem $Ax = b$,
- `solve(A)`: poišče $A - 1$,
- `eigen(A)`: poišče lastne vrednosti in lastne vektorje (dobro deluje le, če se da matrika diagonalizirati).

Osnovna obdelava podatkov na vektorjih in matrikah:

- `sum(x)`: vsota elementov vektorja ali matrike
- `prod(x)`: produkt elementov vektorja ali matrike
- `mean(x)`: povprečje elementov vektorja ali matrike
- `sd(x)`: popravljeni standardni odklon elementov vektorja ali matrike
- `cumsum(x)`: vektor iz kumulativnih vsot
- `cumprod(x)`: vektor iz kumulativnih produktov
- `min(x)`: minimum elementov
- `max(x)`: maksimum elementov

Posredovanje funkcij po komponentah:

- Ukaz `sapply(vektor ali matrika, funkcija)` posreduje funkcijo posameznim komponentam vektorja ali matrike. Tako npr. ukaz `sapply(c(x1, ...  xn), f)` vrne `c(f(x1), ...  f(xn))`.

- Ukaz `apply(matrika, 1, funkcija)` posreduje funkcijo posameznim vrsticam.
- Ukaz `apply(matrika, 2, funkcija)` posreduje funkcijo posameznim stolpcem.
- Nasploh ukaz `array` deluje na poljubnorazsežnih tabelah. V drugem argumentu določimo številke razsežnosti, ki ostanejo.
- Ukaz `mapply(f, c(x11, ... x1n), ... c(xm1, ... xmn))` vrne `c(f(x11, ... xm1), ... f(x1n, ... xmn))`.
- Ukaz `outer(c(x1, ... xm), c(y1, ... yn), FUN=f)` vrne matriko z elementi `f(xi, yj)`. Recimo, operacija `%o%` se da definirati z ukazom `outer`, kjer za funkcijo izberemo množenje. Toda pozor: ukaz je implementiran tako, da funkcijo kliče na ustreznih zgeneriranih matrikah. Rešitev:
  `outer(c(x1, ... xm), c(y1, ... yn), FUN=function(v1, v2) mapply(f, v1, v2) )`.

D.6.3 Tabele

Tabele imajo lahko poljubno mnogo razsežnosti. Dobimo jih lahko iz vektorjev z ukazom: `array(vektor, dim=c(razsežnosti))`. Z ukazom `[ ... ]` lahko podobno kot pri matrikah izluščimo poljubno razsežne komponente tabele.

D.6.4 Vektorji, matrike in tabele z označenimi indeksi

Vektor z označenimi indeksi lahko dobimo z ukazom

`c(oznaka1 = vrednost1, oznaka2 = vrednost2, ... )`.

Lahko pa konstruiramo tudi enorazsežno tabelo z ukazom `array`.

Primer: `array(c(20, 50, 30), dimnames=list(c("prvi", "drugi", "tretji")))` ◇

Seveda nastavitev `dimnames` deluje za poljubno razsežno tabelo.
Deluje tudi pri ukazu `matrix`.

Primer: `matrix(1:12, nrow=4, ncol=3, byrow=TRUE, dimnames=list(c("prva", "druga", "tretja", "cetrta"), c("prvi", "drugi", "tretji")))`

Vrsticam in stolpcem lahko imena prirejamo ali spreminjamo tudi naknadno, recimo: `dimnames(x) <- list(c("prvi", "drugi", "tretji"))`. ◇

POZOR! Vektor ni eno-razsežna tabela, zato mu ne moremo prirejati nastavitve `dimnames`. Če želimo to, ga moramo najprej z ukazom `array` pretvoriti v enorazsežno tabelo.

Če ima tabela označene indekse, jih lahko uporabimo tudi znotraj oklepajev `[ ... ]`. Seveda pa lahko še vedno uporabimo tudi številke.

Oznake komponent upoštevata tudi ukaza `plot` in `boxplot`.

D.6.5 Zapisi

Zapisi so objekti, sestavljeni iz več komponent, ki so lahko različnih tipov. Komponentam bomo rekli rubrike. Zapise konstruiramo z ukazom list.

Primer: `nas_clovek <- list(ime="Janez", starost=35,`
`zakonec="Marija", starosti_otrok=c(15, 13, 2))`. ◇

Posamezne rubrike dobimo z ukazom `$`, npr. `nas_clovek$ime`.
Lahko uporabimo tudi `nas_clovek[["ime"]]`.
Lahko priklicujemo nadaljnje komponente, npr. `nas_clovek$starosti_otrok[2]`.
Z oglatimi oklepaji izluščimo del zapisa, npr.

`nas_clovek[ime]` ali `nas_clovek[c("ime", "starost")]`.

Imena rubrik dobimo in nastavljamo z ukazom `names`.

D.6.6 Kontingenčne tabele in vektorji s predpisanimi vrednostmi

Kontingenčne tabele so 1- ali 2-razsežne tabele z označenimi indeksi, elementi pa so števila. Dobimo jih lahko iz vektorjev z označenimi indeksi z ukazom `as.table`. Ukaz `table(vektor)` pa iz vektorja naredi tabelo zastopanosti njegovih vrednosti. Le-te sortira po abecedi. Če želimo vrednosti ali njihov vrstni red predpisati, uporabimo vektor s predpisanimi vrednostmi, ki ga dobimo z ukazom `factor` in nastavitvijo `levels`:
`table(factor(vektor, levels=vrednosti))`.

Pri vektorju s predpisanimi vrednosti so le-te vedno nizi. Navaden vektor dobimo nazaj z ukazom `as.vector`. POZOR! Ukaz `c` spremeni vrednosti v naravna števila!

Ukaz `table(vektor1, vektor2)` naredi 2-razsežno kontingenčno tabelo. Tabele prav tako rišemo z ukazoma `plot` in `barplot`. Če je t kontingenčna tabela, ukaza
`t[ ... ]` in `t[[ ... ]]` delujeta podobno kot pri zapisih.

D.6.7 Preglednice

Preglednice so podobne dvo-razsežnimi tabelam – z razliko, da so lahko stolpci različnih tipov, zato jim bomo tudi tu rekli rubrike. V posamezni rubriki so bodisi sama števila bodisi sami nizi bodisi same logične vrednosti – tako kot pri vektorjih. Preglednice konstruiramo z ukazom `data.frame`.

Primer:

```
nasi_mozje <- data.frame(
    ime=c("Janez", "Franc", "Joze"),
    starost=c(35, 42, 55), zena=c("Marija", "Stefka", "Lojzka"), st_otrok=c(3, 2, 4)
).
```

◇

Dostop do posameznih komponent je podoben kot pri matrikah – s tem, da:

- če priklicujemo vrstice, je rezultat spet preglednica;
- če priklicujemo stolpce, je rezultat vektor, ki ima predpisane vrednosti, če gre za nize;
- če priklicujemo posamezen element v stolpcu z nizi, je rezultat še vedno vektor s predpisanimi vrednostmi.

Posamezne rubrike lahko podobno kot pri zapisih dobimo tudi z ukazoma `$` in `[[ ... ]]`. Imena rubrik tudi tu dobimo in nastavljamo z ukazom `names`.

Primer:

```
nase_zene < nasi_mozje[,c(žena", "starost", "ime", "st_otrok")]
names(nase_zene) <- c("ime", "starost", "moz", "st_otrok")
```

◇

D.7 Osnove programiranja

R je močan programski jezik, ki podpira tudi objektno orientirano programiranje. Pišemo lahko kratke programčke v pozivniku, daljše programe pa lahko shranimo v datoteke.

D.7.1 Izvajanje programov, shranjenih v datotekah

Samostojne programe, shranjene v datotekah, lahko izvajamo z naslednjimi klici:

- `R < datoteka` – v tem primeru mu moramo predpisati še opcijo `--save`, `--no-save` ali `--vanilla`;
- `R -f datoteka` ali `R --file datoteka`;
- `Rscript datoteka`;
- na Linuxu lahko izvedemo tudi samo datoteko, če se le-ta začne z `#!/usr/bin/Rscript`.

Pri prvih dveh načinih R izpiše celoten potek izvajanja programa. To mu lahko preprečimo z opcijo `--slave`.

Morebitne parametre iz ukazne vrstice dobimo z ukazom `commandArgs(TRUE)`, ki vrne vektor iz pripadajočih nizov. Vključevanje datotek iz pozivnika ali med izvajanjem programa

Pomožne programe lahko vključimo v R-ov pozivnik ali drugo programsko kodo z ukazom `source`.

Pregled nad delovnim imenikom (direktorijem):

- Ukaz `getwd()` vrne lokacijo imenika, kjer R bere in piše.
- Ukaz `setwd(imenik)` nastavi lokacijo delovnega.
- Ukaz `list.files()` vrne vsebino delovnega imenika – kot vektor.
  Lahko mu predpišemo obilo nastavitev.

D.7.2 Najosnovnejši programski ukazi

Posamezne stavke ločimo s podpičjem ali prehodom v novo vrstico. Slednje lahko storimo tudi znotraj oklepajev ali narekovajev.

Znak `#` označuje komentar: vsebina od tega znaka do konca vrstice se ignorira.

Za splošno izpisovanje uporabimo ukaz `cat`. Če mu podamo več argumentov ali vektor, jih loči s presledkom. To lahko spremenimo z nastavitvijo sep. V nasprotju z ukazom print ukaz `sep` ne zaključi vrstice. To dosežemo z nizom `"\n"`.

Primer: `cat(1:5, 10, sep=", "); cat("\n")` ◇

Zaviti oklepaji `{ ... }` označujejo blok. Blok je ukaz, ki zaporedoma izvaja ukaze znotraj zavitih oklepajev in vrne vrednost zadnjega.

D.7.3 Krmilni stavki

Ukaz `if(pogoj) izraz1 else izraz2` ali `ifelse(pogoj, izraz1, izraz2)` vrne `izraz1`, če je pogoj pravilen, sicer pa `izraz2`. Pri prvi konstrukciji lahko del z `else` izpustimo. V tem primeru, če je pogoj napačen, dobimo vrednost `NULL`.

Zanke:

- Ukaz `for(spremenljivka in vektor) ukaz` zaporedoma izvaja ukaz,
  pri čemer spremenljivki prireja vrednosti v vektorju.
- Ukaz `while(pogoj) ukaz` izvaja ukaz, dokler je pogoj pravilen.
- Ukaz `repeat ukaz` ponavlja izvajanje ukaza.
- Ukaz `break` prekine izvajanje zanke.
- Ukaz `next` prekine izvajanje tekočega cikla zanke.

Ukazi za zanke vrnejo izhod zadnjega izvedenega ukaza.

D.7.4 Nekaj več o funkcijah

Funkcije lahko sprejemajo izbirne argumente (opcije), ki jim predpišemo privzete vrednosti. To storimo v deklaraciji:

```
function(parametri, opcija1=vrednost1, opcija2=vrednost2 ...)
```

Primer: Če deklariramo: `f <- function(x, pristej=0) { x*x + pristej }`, ukaz `f(2)` vrne 4, ukaz `f(2, pristej=3)` pa vrne 7. ◇

Ukaz `return(vrednost)` prekine izvajanje funkcije in vrne predpisano vrednost.

D.8 Knjižnice z dodatnimi funkcijami

Ukaz `library()` vrne seznam knjižnic, ki so na voljo. Ukaz `library(knjižnica)` naloži ustrezno knjižnico.

D.9 Vhod in izhod

D.9.1 Pisanje

Včasih želimo kaj izpisati še kako drugače, kot to stori R. Poleg tega R ne izpisuje avtomatično znotraj zank, ker pač le-te ne vračajo argumentov. Če želimo to, uporabimo ukaz `print`, ki izpiše vrednost, tako kot bi jo sicer izpisal R (z nekaj dodatnimi nastavitvami, kot je npr. `digits`). Izpisovanje v osnovni obliki dosežemo z ukazom `cat`. Sprejme več argumentov, ki jih loči s presledkom. Ločitveni `niz` lahko spremenimo z nastavitvijo `sep`.

Primeri:

- `cat("bla", "ble", "bli")`
- `cat("bla", "ble", "bli", sep="*")`
- `cat("bla", "ble", "bli", "\n")`
- `cat("bla", "ble", "bli", sep="\n")`
- `cat("bla", "ble", "bli", sep="*\n")` ◇

POZOR! Če ločitveni niz vsebuje znak za novo vrstico, R slednjega avtomatično doda tudi na konec.

Ukaz `paste` deluje podobno kot `cat`, le da vrednost vrne, namesto da bi jo izpisal. Če mu kot argument podamo več nizov, vrne isto, kot bi izpisal `cat`, le da ne upošteva pravila o novi vrstici.

Ukazu `paste` pa lahko posredujemo tudi več vektorjev. V tem primeru vrne vektor, čigar i-to komponento sestavljajo vse i-te komponente podanih vektorjev, ločene z nizom, podanim s `sep`.

Če ukazu `sep` predpišemo nastavitev `collapse=niz`, iz ustvarjenega vektorja naredi enoten niz, pri čemer komponente loči s predpisanim nizom.

Primer: ukaz `paste(c(1, 2, 3), c(4, 5, 6, 7), sep=, collapse="\n")` vrne "1 4\n2 5\n3 6\n1 7". ◇

Ukaz `format` je namenjen izpisovanju števil v predpisanem formatu (denimo na določeno število decimalk), deluje pa tudi za nize. Podobno kot paste tudi ukaz `format` ne izpisuje, temveč vrne niz. Ena od mnogih nastavitev je `justify`, ki deluje za poravnavo nizov (možne vrednosti so `"left"`, `"right"`, `čentre"` in `"none"`). Števila pa so vedno poravnana desno.

Primer: ukaz `format(2/3, trim = TRUE, digits = 3, width = 6, justify = "left")` vrne "0.667", ukaz: `format(format(2/3, digits = 3), width = 6, justify = "left")` pa vrne "0.667". ◇

POZOR! Pri obratni poševnici je hrošč – šteje jo kot dva znaka.

Še nekaj ukazov za pisanje:

- `writeLines` izpiše komponente podanega vektorja kot vrstice.
- `write.table` izpiše preglednico.
- `write.csv` izpiše preglednico v formatu csv.

D.9.2 Delo z datotekami

Pregled nad delovnim imenikom (direktorijem):

- Ukaz `getwd()` vrne lokacijo imenika, kjer R bere in piše.
- Ukaz `setwd(imenik)` nastavi lokacijo delovnega.
- Ukaz `list.files()` vrne vsebino delovnega imenika – kot vektor.
  Lahko mu predpišemo obilo nastavitev.

Branje in pisanje navadno izvedemo tako, da ustreznemu ukazu predpišemo opcijo `file`. Če jo nastavimo na `niz`, to pomeni enkratno branje oz. pisanje na datoteko z danim imenom. R pa podpira tudi postopno delo z datotekami. V tem primeru datoteko:

- Najprej odpremo z ukazom `file(datoteka, open = način)`.
  Tipične vrednosti načina so `"r"` (branje), `"w"` (pisanje) in `"a"` (dodajanje).
- Funkcija `file` vrne kazalec na datoteko, ki ga podajamo pri opciji file.

- Nazadnje datoteko zapremo z ukazom `close(kazalec)`.

Primer: ukaz `cat("blabla", file = "blabla.txt")` naredi isto kot zaporedje ukazov:

```
bla <- file("blabla.txt", "w")
cat("blabla", file = bla)
close(bla).
```

◇

D.9.3 Branje

Osnovni ukazi:

- `readChar(datoteka, n)` prebere n znakov z datoteke in vrne prebrani niz.
- `readLines(datoteka)` prebere datoteko in vrne vektor iz njenih vrstic.
- `scan(datoteka, what = tip)` prebere datoteko in njeno vsebino interpretira kot seznam elementov predpisanega tipa, recimo `"logical"`, `"numeric"`, `"character"` ali `"list"`. Rezultat je vektor.

Ukaz `read.table` prebere preglednico. Prva vrstica so glave, elementi v posamezni vrstici so ločeni s presledki, če imajo nizi presledke, jih damo v narekovaje. Posamezne glave morajo biti brez presledkov. Če ima npr. datoteka Preglednica.txt vsebino:

```
Ime Prva Druga
"Novak Janez" 5 35
Jurak 3 37
```

ima ukaz: `read.table("Preglednica.txt", header=TRUE)` isti rezultat kot:

```
data.frame(
  Ime=factor(c("Novak Janez", "Jurak")),
  Prva=factor(c(5, 3))
  Druga=factor(c(35,37))
).
```

Ukaz `read.csv` je namenjen branju preglednic v formatu `csv`.

D.9.4 Izvoz grafike

Datoteko najprej odpremo s klicem ustreznega gonilnika, npr. `postscript(datoteka)` ali `pdf(datoteka)`. Nato kličemo ustrezne ukaze, npr. `plot`. Nazadnje datoteko zapremo z ukazom `dev.off()`.

D.10 Verjetnostne porazdelitve

Porazdelitve v R-u računamo z ukazom: `predpona+porazdelitev(vrednost, parametri)`. Parametri so opcije ustrezne funkcije, tj. oblike `parameter=vrednost` (glej spodaj). Možne predpone so:

- `'d'` za točkasto verjetnost $P(X = x)$ diskretnih oz. gostoto $p_X(x)$ zveznih porazdelitev;
- `'p'` za kumulativno porazdelitveno funkcijo $P(X \leqslant x)$;
- `'q'` za kvantilno funkcijo;
- `'r'` za simulacijo.

Primer: `dbinom(3, size=10, prob=0.4)` vrne $P(X = 3)$, kjer je slučajna spremenljivka X porazdeljena biniomsko `Bi(10, 0.4)`. $\diamond$

Varianta za `'d'`:

- `pporazdelitev(x, parametri, log=TRUE)` vrne $\ln P(X = x)$ oz. $\ln p_X(x)$.

Variante za `'p'`:

- `pporazdelitev(x, parametri, lower.tail=FALSE)` vrne $P(X > x)$.
- `pporazdelitev(x, parametri, log.p=TRUE)` vrne $\ln P(X \leqslant x)$.
- `pporazdelitev(x, parametri, lower.tail=FALSE, log.p=TRUE)` vrne $\ln P(X > x)$.

Najpogostejše diskretne porazdelitve:

porazdelitev	R-ovo ime	parametri
binomska	`binom`	`size`, `prob`
geometrijska	`geom`	`prob`
neg. binomska	`nbinom`	`size`, `prob`
Poissonova	`pois`	`lambda`
hipergeometrijska	`hyper`	`m`, `n`, `k`

Najpogostejše zvezne porazdelitve:

porazdelitev	R-ovo ime	parametri
enakomerna	`unif`	`min`, `max`
normalna	`norm`	`mean`, `sd`
eksponentna	`exp`	`rate`
Cauchyjeva	`cauchy`	`location`, `scale`
Studentova	`t`	`df`, `ncp`
Fisherjeva	`f`	`df1`, `df2`, `ncp`

Določeni parametri imajo privzete vrednosti, npr. `mean=0` in `sd=1` pri normalni porazdelitvi.

D.11 Simulacije

Ukaz `rporazdelitev(n)` naredi vektor iz n realizacij slučajne spremenljivke z dano porazdelitvijo. Recimo funkcija `runif` ustreza funkciji `rnd` ali `random` iz mnogih programskih jezikov.

Vzorčenju je namenjen ukaz `sample`:

- `sample(x)` vrne slučajno permutacijo vektorja x.
- `sample(x, velikost)` vrne vzorec brez ponavljanja ustrezne velikosti iz x.
- `sample(x, velikost, replace = TRUE)` vrne vzorec iz x s ponavljanjem.

Dodatna možna nastavitev: `prob` za vektor verjetnosti, recimo:
`sample(c("a", "e", "i"), 20, replace=TRUE, prob=c(0.2, 0.5, 0.4))`

Isto naredi ukaz:
`sample(c("a", "e", "i"), 20, replace=TRUE, prob=c(20, 50, 40))`.

D.12 Statistično sklepanje

D.12.1 Verjetnost uspeha poskusa/delež v populaciji

Zanima nas verjetnost, da poskus uspe oziroma delež enot v populaciji z dano lastnostjo. To označimo s p. Izvedemo n poskusov, k jih uspe (oziroma vzamemo vzorec n enot in k jih ima iskano lastnost).

Interval zaupanja za p pri stopnji zaupanja β dobimo z ukazom:

`binom.test(k, n, conf.level =`β`)`.

ali tudi:

`binom.test(c(k, n - k), conf.level =`β`)`.

Hipotezo, da je $p = p_0$, testiramo z ukazom:

`binom.test(k, n, p = p0, alternative = "two.sided"ali "less"ali "greater")`.

Kot rezultat dobimo p-vrednost. Privzeta vrednost za določilo p je 0.5.

Določilo `alternative` pove, kaj je alternativna domneva:

izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi $p \neq p_0$.

Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi $p < p_0$. Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi $p > p_0$.

D.12.2 Primerjava verjetnosti dveh poskusov/deležev v dveh populacijah

Testiramo ničelno domnevo, da sta verjetnosti dveh različnih poskusov oziroma deleža enot z določeno lastnostjo v dveh različnih populacijah enaka. Označimo ju s p_1 in p_2. Izvedemo n_1 poskusov prve vrste, od katerih jih uspe k_1, in n_2 poskusov druge vrste, od katerih jih uspe k_2. Ekvivalentno, vzamemo vzorec n_1 enot iz prve populacije, izmed katerih jih ima k_1 iskano lastnost, in vzorec n_2 enot iz druge populacije, izmed katerih jih ima k_2 našo lastnost. Test ničelne domneve $p_1 = p_2$ izvedemo z ukazom:

`prop.test(c(k1, k2), c(n1,n2), alternative = "two.sided" ali "less" ali "greater")`.

Določilo `alternative` pove, kaj je alternativna domneva:

- Izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi $p_1 \neq p_2$.
- Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi $p_1 < p_2$.
- Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi $p_1 > p_2$.

D.12.3 Primerjava verjetnosti več poskusov/deležev več populacijah

Testiramo ničelno domnevo, da so vse verjetnosti več različnih poskusov oziroma vsi deleži enot z določeno lastnostjo v več različnih populacijah enaki. Izvedemo poskuse oziroma iz vsake populacije vzamemo vzorec. Števila vseh poskusov posamezne vrste oziroma velikosti vzorcev iz posamezne populacije naj tvorijo vektor n, števila vseh uspelih poskusov posamezne vrste oziroma števila enot iz posamezne populacije z določeno lastnostjo pa vektor k. Test izvedemo z ukazom: `prop.test(k, n)`.

D.12.4 Populacijsko povprečje – T-test

Zanima nas povprečje spremenljivke na populaciji, kjer privzamemo, da ima (vsaj približno) normalno porazdelitev. Označimo ga z μ. Vzamemo vzorec, vrednosti spremenljivke na vzorcu naj tvorijo vektor v. Parameter v pa je lahko tudi kontingenčna tabela (recimo dobljena z ukazom `as.table`.

Interval zaupanja za μ pri stopnji zaupanja β dobimo z ukazom:

`t.test(v, conf.level` $= \beta$`)`.

Hipotezo, da je $\mu = \mu_0$, testiramo z ukazom:

`t.test(v, alternative = "two.sided"ali "less"ali "greater",` $\mu = \mu_0$`)`.

Privzeta vrednost določila μ je 0.

Določilo `alternative` pove, kaj je alternativna domneva:

- Izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi $\mu \neq \mu_0$.
- Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi $\mu < \mu_0$.
- Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi $\mu > \mu_0$.

D.12.5 Test mediane

Testiramo domnevo, da ima dana urejenostna (ordinalna) spremenljivka na populaciji mediano μ_0. To lahko izvedemo tudi na številskih spremenljivkah namesto T-testa, še posebej, če so njihove porazdelitve daleč od normalne. Vzamemo vzorec, vrednosti naše spremenljivke na vzorcu naj tvorijo vektor v. Obravnavali bomo dva možna testa.

Test z znaki. Pri tem testu je pomembno le, ali je dana vrednost večja, manjša ali enaka μ_0, zato je njegova uporaba smiselna pri čistih urejenostnih spremenljivkah, ko ni določeno, kateri dve vrednosti imata isti odmik od μ_0 navzgor in navzdol. Test se tako prevede na testiranje deleža in ga izvedemo z ukazom:

`binom.test(sum(sapply(v, function(x) {` $x > \mu_0$ `})), sum(sapply(v, function(x) { x !=` μ_0 `})),`
`alternative = "two.sided" ali "less" ali "greater")`.

Wilcoxonov test z rangi. Ta test izvedemo, če je določeno, kateri dve vrednosti imata isti odmik od μ_0. Test izvedemo z ukazom:

`wilcox.test(v, alternative = "two.sided"ali "less"ali "greater")`.

D.12.6 Primerjava porazdelitev dveh spremenljivk

1. Dve (približno) normalni spremenljivk na isti populaciji – T-test

Ta primer se prevede na testiranja povprečja ene same spremenljivke, če obe spremenljivki odštejemo: če je količina tvori vektor v, druga pa vektor w, ukažemo:

```
t.test(v - w, alternative = "two.sided"ali "less"ali "greater").
```

Isto pa dosežemo tudi z ukazom:

```
t.test(v, w, paired = TRUE, alternative = "two.sided"ali "less"ali "greater").
```

2. Dve urejenostni spremenljivki na isti populaciji

Tukaj primerjamo spremenljivki, ki sta bodisi urejenostni ali pa ne moremo privzeti, da sta njuni porazdelitvi blizu normalne. Spet lahko uporabimo bodisi test z znaki bodisi test z rangi. Če sta v in w vektorja podatkov, test z znaki izvedemo z ukazom:

```
binom.test(sum(mapply(quote(≫"), x, y)),
sum(mapply(quote("!="), x, y)),
alternative = "two.sided"ali "less"ali "greater").
```

test z rangi pa izvedemo z ukazom:

```
wilcox.test(v, w, paired = TRUE, alternative = "two.sided"ali "less"ali "greater").
```

3. Povprečji dveh spremenljivk na dveh populacijah – T-test

Testiramo domnevo, da imata spremenljivki, definirani na različnih populacijah, enako povprečje, pri čemer privzamemo, da sta vsaj približno normalno porazdeljeni. Z drugimi besedami, če povprečji označimo z μ_1 in μ_2, ničelna domneva trdi, da je $\mu_1 = \mu_2$. Iz vsake populacije vzamemo vzorec, vrednosti na prvem naj tvorijo vektor v, vrednosti na drugem pa vektor w. Test izvedemo z ukazom:

```
t.test(v, w, alternative = "two.sided"ali "less"ali "greater").
```

Natančneje, ta ukaz izvede heteroskedastični T-test, ki ne privzame enakosti varianc. Če smemo privzeti, da sta varianci obeh spremenljivk enaki, izvedemo homoskedastični test:

```
t.test(v, w, var.equal = TRUE, alternative = "two.sided"ali "less"ali "greater").
```

Pomen določila `alternative` je tak kot ponavadi:

- Izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi $\mu_1 \neq \mu_2$.
- Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi $\mu_1 < \mu_2$.
- Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi $\mu_1 > \mu_2$.

4. Dve urejenostni spremenljivki na dveh populacijah – Wilcoxon-Mann-Whitneyev test

Spet primerjamo spremenljivki, ki sta bodisi urejenostni ali pa ne moremo privzeti, da sta njuni porazdelitvi blizu normalne. Iz vsake populacije vzamemo vzorec, vrednosti na prvem naj tvorijo vektor v, vrednosti na drugem pa vektor w. Hipotezo, da sta spremenljivki enako porazdeljeni, testiramo z ukazom:

`wilcox.test(v, w, paired = TRUE, alternative = "two.sided"ali "less"ali "greater")`.

Določilo `alternative` pove, kaj je alternativna domneva:

- izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi, da sta porazdelitvi različni.
- Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi, da je prva spremenljivka stohastično večja od druge.
- Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi, da je prva spremenljivka stohastično manjša od druge.

Če je eden od vzorcev velik, je računanje zahtevno. Lahko ga poenostavimo z določilom `exact = FALSE` – v tem primeru bo R računal približek.

5. Povprečji spremenljivk na več populacijah – analiza variance (ANOVA) z enojno klasifikacijo

Testiramo domnevo, da imajo spremenljivke, za katere privzamemo, da so porazdeljene normalno, enaka povprečja. Podatke uvrstimo v dva vektorja, denimo v in s. V vektorju v so vse vrednosti, v vektorju s pa skupine (ta vektor torej pove, iz katere populacije je dani podatek). Elementi vektorja s morajo biti nizi, ne števila! Test izvedemo z ukazom:

`anova(lm(vr   sk, data.frame(vr = v, sk = s)))`.

p-vrednost odčitamo v rubriki $Pr(< F)$ na desni strani R-ovega poročila.

6. Urejenostni spremenljivki na več populacijah – Kruskal-Wallisov test

Ravnamo podobno kot pri analizi variance: podatke uvrstimo v dva vektorja, denimo v in s, pri čemer so v vektorju v vrednosti, v vektorju s pa skupine. Toda pozor: v nasprotju z analizo variance morajo biti tu v vektorju s števila, ne nizi!

Test izvedemo z ukazom: `kruskal.test(v, s)`.

D.12.7 Koreliranost

Zanima nas koreliranost dveh spremenljivk na isti populaciji. Vzamemo vzorec, vrednosti prve spremenljivke naj tvorijo vektor v, vrednosti druga pa vektor w.

Interval zaupanja za korelacijski koeficient pri stopnji zaupanja β poiščemo z ukazom:

`cor.test(v, w, method = "pearson"ali "kendall"ali "spearman", conf.level =`β`)`.

Določilo method pove, kateri korelacijski koeficient nas zanima.

- Izbira `"pearson"` ali `"p"` ali opustitev določila pomeni običajni Pearsonov korelacijski koeficient, ki je primeren za številske spremenljivke, porazdeljene (približno) normalno.
- Izbira `"spearman"` ali `š"` pomeni Spearmanov,
- izbira `"kendall"` ali `"k"` pa Kendallov korelacijski koeficient;

Slednja testa sta primerna za urejenostne spremenljivke. Spearmanov korelacijski koeficient (ρ) je lažji za računanje, o Kendallovem koeficientu (τ) pa dosti statistikov meni, da je verodostojnejši.

Hipotezo, da sta spremenljivki nekorelirani (oziroma neodvisni) pa testiramo z ukazom:

`cor.test(v, w, method = "pearson"ali "kendall"ali "spearman", alternative = "two.sided" ali "less" ali "greater")`.

Določilo `alternative` pove, kaj je alternativna domneva:

- Izbira `"two.sided"` ali `"t"` ali izpustitev določila pomeni, da alternativna domneva trdi, da sta spremenljivki odvisni (oziroma korelirani).
- Izbira `"less"` ali `"l"` pomeni, da alternativna domneva trdi, da sta spremenljivki pozitivno asociirani (oziroma korelirani).
- Izbira `"greater"` ali `"g"` pa pomeni, da alternativna domneva trdi, da sta spremenljivki negativno asociirani (oziroma korelirani).

Računanje p-vrednosti za Kendallov korelacijski koeficient je zahtevno in R sam izbere, kdaj bo računal natančno vrednost in kdaj približek. To lahko spremenimo z določilom `exact = TRUE` ali `exact = FALSE`.

Stvarno kazalo