Katarina Košmelj

UPORABNA STATISTIKA

Druga dopolnjena izdaja

Ljubljana, 2007

Recenzenta: prof. dr. Janez Stare in prof. dr. Anuška Ferligoj
Lektorica: prof. Mija Knop
Oblikovanje besedila in slik: mag. Matjaž Jeran

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

311(075.8)

KOŠMELJ, Katarina
Uporabna statistika [Elektronski vir] / Katarina Košmelj. - 2. dopolnjena izd. – Ljubljana: Biotehniška fakulteta, 2007

Način dostopa (URL): http://www.bf.uni-lj.si/agronomija/o-oddelku/katedre-in-druge-org-enote/za-statistiko/studijske-zadeve/Uporabna_statistika.pdf

ISBN 978-961-6275-26-2

235777024

PREDGOVOR

Make it as simple as possible but not simpler.
(Albert Einstein)

Delo je namenjeno začetnikom na področju uporabne statistike, ki imajo vsaj srednješolsko znanje matematike. Nastalo je na osnovi mojih predavanj na Biotehniški fakulteti na dodiplomskem študiju. Knjiga ima dve ravni, osnovno raven in nadgradnjo, ki je označena z zvezdicami.

Delo ima sedem poglavij: Uvod, Opisna statistika, Osnove verjetnostnega računa, Osnove statističnega sklepanja, Primerjava dveh populacij, Regresija in korelacija ter Hi-kvadrat preizkusi. Struktura poglavij je prikazana z miselnimi vzorci, ki so v prilogi. Vsako poglavje ima naloge, s katerimi študent preveri naučeno snov, rešitve so v prilogi.

Pobudo za nastanek statistike so dale različne stroke, matematika pa je dala orodje za njeno delovanje. Zato je statistika v bistvu matematična disciplina. Vendar to delo ni pisano s stališča matematične teorije, ki je podpora statistiki. Namerno je napisano tako, da je matematika podana v najmanjši možni meri. Matematično teorijo sem skušala nadomestiti z grafičnimi prikazi, ilustracijami in z velikim številom primerov. Vem, da za večino ljudi velja, da ustrezna slika nadomesti mnogo razlag ter da se največ naučimo iz primerov. Zato je v knjigi veliko primerov, večina iz realnega sveta, čeprav so zaradi razumljivih omejitev poenostavljeni. Rešitve primerov so podane v tolikšni meri, da je snov ustrezno ilustrirana.

Izdelava dela je bila tehnično zahtevna, pri izdelavi slik in pri končnem oblikovanju besedila mi je pomagal mag. Matjaž Jeran. Mnoge strokovne izboljšave sta predlagala recenzenta ter mag. Damijana Kastelec in mag. Matjaž Jeran. Lektorica prof. Mija Knop je izboljšala kakovost besedila. Vsem navedenim se za njihove predloge, trud in potrpežljivost najlepše zahvaljujem. Moja zahvala tudi vsem, ki so mi posredovali podatke. Statističnemu uradu R Slovenije pa se zahvaljujem za tiskanje učbenika.

Pri pisanju učbenika sem skušala slediti zgoraj navedenemu nasvetu Alberta Einsteina. Trudila sem se najti kompromis med teorijo in prakso, ki je primeren za začetnike, ter delo napisati tako, da bi se študent v čim krajšem času čim več naučil ter da bi mu bilo učenje v veselje. Kritičnim bralcem bom hvaležna za oceno, koliko sem pri teh ciljih uspela. Prav tako bodo dobrodošli vsi predlogi, popravki in izboljšave.

Katarina Košmelj

Ljubljana, januar 2001

Druga izdaja je dopolnjena in izdana v elektronski obliki.

Ljubljana, oktober 2007

Kazalo vsebine

1 UVOD

1.1 KAJ JE STATISTIKA

Začnimo to knjigo z opredelitvijo pojma statistika. Kratko in jedrnato bi povedali takole: **statistika** je veda, ki proučuje pojave, ki se kažejo v velikem številu v določenem času in prostoru. Takim pojavom rečemo množični pojavi. Namen statističnega proučevanja je globlje razumevanje množičnega pojava, odkrivanje njegovih zakonitosti in napovedovanje. Statistično proučevanje posredno pomaga drugim strokam pri upravljanju, vodenju in načrtovanju različnih procesov.

Pomemben poudarek v zgornji opredelitvi je na besedi množičen. Že na začetku jasno povejmo, da se statistika ukvarja z zakonitostmi, ki veljajo pri proučevanem množičnem pojavu, in ne s posameznimi enotami, ki sestavljajo ta pojav. Primer množičnega pojava je npr. študij na Univerzi v Ljubljani. Statistično proučevanje bi bilo lahko usmerjeno v proučevanje uspešnosti brucev v določenem obdobju. Poskušali bi ugotavljati zakonitosti, ki vplivajo na uspešnost brucev na Univerzi v Ljubljani. Te zakonitosti se nanašajo na vse obravnavane bruce, o posamezniku zvemo zelo malo.

Statistika proučuje zakonitosti množičnih pojavov na poseben način. Njeno delo sestavljajo tri faze, ki si sledijo druga za drugo:

- zbiranje in urejanje podatkov, ki opisujejo proučevani množični proces;
- analiza zbranih podatkov. Pri tem uporabljamo posebno statistiki lastno metodologijo, to so t. i. **statistične metode**. Nabor statističnih metod je velik, izbira metode je odvisna od namena analize ter od vrste podatkov. Analizo podatkov si lahko predstavljamo kot mlinček, v katerega damo podatke, izberemo ustrezno rezilo (statistične metode), iz mlinčka pa dobimo rezultate;
- razlaga rezultatov. Pri razlagi mora veljati poudarek vsebinski interpretaciji rezultatov, torej razlagi v smislu, kaj smo o proučevanem množičnem procesu zvedeli novega.

Vsaka izmed opisanih faz je ključna za uspešnost statističnega proučevanja in mora biti opravljena premišljeno, korektno in celovito.

V vsakdanjem življenju ima beseda statistika ožji pomen. Govorimo npr. o statistiki vpisa študentov, o statistiki prometnih nesreč, o statistiki brezposelnih ipd. V tem primeru ta beseda pomeni: zbiranje in urejanje podatkov, prikazovanje podatkov in izračun enostavnih kazalcev o pojavu.

V okviru statistične metodologije uporabljamo besedo statistika še v tretjem pomenu. Statistika predstavlja funkcijo, ki jo na določen način izračunamo iz podatkov, npr. z-statistika, t-statistika ipd. Te funkcije bomo spoznali kasneje.

S statistiko so se ukvarjale že zelo stare civilizacije. Ohranjeni so grafični prikazi števila oseb, živine, pridelka (npr. babilonska glinena plošča, 3800 pr. n. š.). Egipčani, Kitajci, Grki, Rimljani so izvajali popise prebivalstva. Rimljani so sistematično zbirali tudi druge podatke o osvojenih ozemljih. V srednjem veku so izvajali popise cerkvenih zemljišč (Pipin Mali, Karel Veliki, 8. stol.), nastala je prva zemljiška knjiga (Viljem Osvajalec, 11. stol.). Začne se sistematično proučevanje ekonomskih in demografskih pojavov. Pojavi se beseda statistika, ki izvira iz latinske besede *status* (stanje, pravni, državni položaj). 17. in 18. stoletje prineseta razvoj verjetnostnega računa, ki je temelj statistike (Blaise Pascal, Jacques Bernoulli, Siméon-Denis Poisson, Albert de Moivre, Carl Friedrich Gauss, Pierre Simon Laplace). Začetek 20. stoletja prinese razvoj regresijske in korelacijske analize (Francis Galton, Karl Pearson), okoli 1920 je začetek statističnega sklepanja (Ronald Fisher, Egon Pearson, Jerzy Neymann). Po letu 1960 računalniška tehnologija razširi dejavnost statistike in pospeši njen razvoj. Okoli

1980 se pojavijo novi načini in metode za analizo podatkov, npr. metode za raziskovanje podatkov (John Tukey).

Danes je statistika je interdisciplinarna veda. Je del mnogih naravoslovnih in družboslovnih ved. Npr. del bioloških in medicinskih ved je *biostatistika* in *biometrika*, del demografije je *demografska statistika*, del ekonomije je *ekonomska statistika* in *ekonometrika* itd.

Pomembna veja statistike je **državna statistika**, to je statistika za potrebe države. Vsaka država ima pooblaščeno institucijo, ki zbira in obdeluje za državo pomembne podatke. Za Slovenijo je pooblaščena institucija *Statistični urad R Slovenije* (SURS). Izdaja različne publikacije, najpomembnejša je *Statistični letopis* (SL), ki izhaja vsako leto.

Široka uporabnost statističnih metod izhaja iz dejstva, da imajo statistične metode matematične temelje. Z razvojem statističnih metod, ki so neodvisne od področja uporabe, se ukvarja posebna matematična disciplina, to je **matematična statistika**. Opira se na verjetnostni račun. V tem delu se z matematično statistiko ne bomo ukvarjali, bomo pa uporabljali njene rezultate.

1.2 OSNOVNI POJMI

1.2.1 Populacija, vzorec

Pri statističnem proučevanju množičnega pojava moramo najprej opredeliti **statistično populacijo.** Statistična populacija, krajše **populacija**, je podana s tremi opredelitvami:

- stvarna opredelitev: koga oz. kaj proučujemo?
- krajevna opredelitev: kje?
- časovna opredelitev: kdaj?

Primer populacije: študenti, vpisani v 1. letnik Biotehniške fakultete v šolskem letu 1998/99.

Populacije so stvarne in hipotetične. Stvarne populacije obstajajo v realnem svetu, hipotetične populacije si lahko le zamislimo.

Populacijo sestavljajo **enote**. Oznaka N označuje število enot v populaciji. Ker proučujemo množične pojave, je N navadno veliko število.

Del populacije imenujemo **vzorec**. Za statistiko je vzorec del populacije, katerega enote izbiramo z namenom, da ocenimo stanje v populaciji. Vzorec, ki dobro posreduje lastnosti celotne populacije, imenujemo reprezentativen vzorec. Oznaka n označuje število enot v vzorcu, $n < N$.

Imamo osnovno populacijo velikosti N. To populacijo vzorčimo, da dobimo vzorec izbrane velikosti n. Število vzorcev, ki jih lahko dobimo iz osnovne populacije, je praviloma ogromno. Poglejmo bolj natančno to množico. Vsi vzorci velikosti n, ki jih lahko dobimo iz osnovne populacije velikosti N, tvorijo **populacijo vzorcev velikosti** n. To je hipotetična populacija, ki je izjemno pomembna za matematično statistiko. Enota populacije vzorcev velikosti n je en vzorec velikosti n.

Populacija vzorcev določene velikosti ima zelo veliko enot že pri majhnih osnovnih populacijah. Število vzorcev velikosti n, ki jih lahko dobimo iz osnovne populacije velikosti N, je odvisno od načina izbire enot: izbrano enoto vrnemo v populacijo, izbrane enote ne vrnemo v populacijo. Število vzorcev izračunamo z variacijami s ponavljanjem oz. brez ponavljanja takole:

- izbrano enoto vrnemo v populacijo:

  število vzorcev = $V_N^{n(p)} = N \cdot N \cdot \ldots \cdot N = N^n$

- izbrane enote ne vrnemo v populacijo:

število vzorcev = $V_N^n = N \cdot (N-1) \cdot (N-2) \cdot \ldots \cdot (N-n+1)$

Pri izračunu števila vzorcev smo upoštevali, da se vzorci, ki vsebujejo iste enote v različnem vrstnem redu, vsebinsko razlikujejo. V praksi navadno ta predpostavka ni utemeljena. Če je tako, je število vzorcev manjše in se izračuna s kombinacijami s ponavljanjem oz. brez ponavljanja takole:

- izbrano enoto vrnemo v populacijo:

  število vzorcev = $C_N^{n(p)} = \binom{N+n-1}{n}$

- izbrane enote ne vrnemo v populacijo:

  število vzorcev = $C_N^n = \binom{N}{n}$

Primer

Osnovna populacija so gospodinjstva v občini Dol na dan 1. 1. 1998. Teh je 1250. Za anketo potrebujemo 60 gospodinjstev. Vsi vzorci s po 60 gospodinjstvi iz občine Dol tvorijo populacijo vzorcev velikosti 60. Enota te populacije je en vzorec s 60 gospodinjstvi.

V vzorcu se gospodinjstva lahko ponavljajo, vzorci z istimi gospodinjstvi v različnem vrstnem redu se vsebinsko razlikujejo. Število vzorcev velikosti 60 izbranih iz osnovne populacije velikosti 1250 je

$1250^{60} = 6{,}5 \cdot 10^{185}$

V vzorcu naj bodo različna gospodinjstva. Število vzorcev velikosti 60 izbranih iz osnovne populacije velikosti 1250 je:

$1250 \cdot 1249 \cdot \ldots \cdot 1191 = 1{,}5 \cdot 10^{185}$

Če upoštevamo še, da so vzorci z istimi gospodinjstvi v različnem vrstnem redu ekvivalentni, je število vzorcev manjše:

$\binom{1250}{60} = 1{,}86 \cdot 10^{103}$

Primer

Osnovno populacijo predstavljajo volilni upravičenci v Sloveniji na dan 1. 10. 2000. Teh je približno 1,5 milijona. Raziskovalci so pred volitvami izvedli predvolilno anketo, v katero so vključili 300 volilnih upravičencev. V vzorcu se anketiranci ne ponavljajo, vzorci z istimi anketiranci v različnem vrstnem redu so ekvivalentni. Izračunajmo število možnih vzorcev:

$\binom{1500000}{300} = 2{,}13 \cdot 10^{1238}$

Izračunano število je nepredstavljivo veliko. Za njegov izračun smo uporabili dovolj zmogljiv računalnik.

1.2.2 Spremenljivke

Zanimajo nas lastnosti enot. Npr. pri proučevanju uspešnosti brucev nas zanimajo naslednje lastnosti brucev: leto rojstva, kraj rojstva, spol, opravljena srednja šola, smer študija itd. V

statistiki opisuje posamezno lastnost enote **statistična spremenljivka**, krajše jo bomo imenovali kar **spremenljivka**.

Vsaka spremenljivka ima določene vrednosti. Npr. vrednosti spremenljivke spol sta: moški, ženski; vrednosti spremenljivke leto rojstva so 1975, 1976, 1977 itd. Spremenljivke označujemo z velikimi črkami s konca abecede, njihove vrednosti pa s pripadajočimi malimi črkami. Npr. spremenljivka X ima vrednosti x_1, x_2 itd.

Glede na to, kako se vrednosti spremenljivke izražajo, ločimo:

- **opisne** (atributivne, kvalitativne) spremenljivke: vrednosti se izražajo z opisi (besede, znaki, številke);
- **številske** (numerične, kvantitativne) spremenljivke: vrednosti se izražajo s števili, s katerimi lahko računamo.

Primer

Opisne spremenljivke so: spol, kraj rojstva, izobrazba očeta, smer študija; številske spremenljivke so: leto rojstva, število otrok v družini.

Številske spremenljivke delimo v več skupin:

- *diskretne spremenljivke*. Za diskretno spremenljivko velja, da lahko zavzame le določene vrednosti. Njene vrednosti so navadno cela števila, pogosto jih dobimo s štetjem. Npr.: število otrok v družini, število opravljenih izpitov;
- *zvezne spremenljivke*. Spremenljivka je zvezna, če med poljubnima dvema vrednostma na njenem definicijskem območju vedno obstaja tretja vrednost. Vrednosti zvezne spremenljivke navadno dobimo z merjenjem ali izračunavanjem. Npr.: starost, višina, stopnja inflacije. Omeniti velja, da so dobljeni podatki za zvezne spremenljivke bolj ali manj diskretizirani zaradi omejene natančnosti merilnih naprav in zaradi zaokrožanja vrednosti. Tako npr. višino oseb merimo običajno v centimetrih, štejemo pa jo za zvezno spremenljivko.
- ostale, s katerimi se ne bomo ukvarjali.

Za statistično spremenljivko določimo njeno **mersko lestvico**. Le-ta je opredeljena glede na *urejenost vrednosti spremenljivke*: v vrednostih ni nobene urejenosti (npr. smer študija: ekonomija, biologija, zgodovina), vrednosti lahko uredimo v smiselni vrstni red (npr. izobrazba: osnovna, srednja, višja, visoka), vrednosti lahko odštevamo (npr. temperatura zraka: 10^0 C, -5^0 C), vrednosti lahko delimo (npr. masa: 10 kg, 120 kg). Glede na to, kolikšna je ta urejenost, ločimo štiri vrste spremenljivk:

- **imenska** (**nomimalna**) spremenljivka: v vrednostih spremenljivke ni nobene urejenosti. Npr.: kraj rojstva, smer študija, barva oči. Posebne vrste imenskih spremenljivk so **dvojiške** (**binarne**) spremenljivke, ki imajo samo po dve vrednosti. Npr.: spol (moški, ženski), štipendist (da, ne).
- **urejenostna** (**ordinalna**) spremenljivka: vrednosti lahko uredimo glede na nekakšen naravni vrstni red. Npr.: izobrazba, šolska ocena.
- **razmična** (**intervalna**) spremenljivka se izraža številsko. Poljubni dve vrednosti spremenljivke lahko odštevamo, ne pa delimo. Npr.: temperatura zraka, leto rojstva.
- **razmernostna** spremenljivka je številska spremenljivka, pri kateri je razmerje poljubnih dveh njenih vrednosti smiselno. Npr.: starost, višina štipendije, letna količina padavin.

Urejenost narašča od imenske do razmernostne spremenljivke. Opisne spremenljivke imajo imensko ali urejenostno mersko lestvico, številske pa razmično ali razmernostno. Izbira

ustrezne statistične metode temelji tudi na merski lestvici spremenljivke. Čim višja je merska lestvica spremenljivke, tem več je primernih statističnih metod.

NALOGE

1. Bruci

Populacija je opredeljena takole: vpisani v prvi letnik visokošolskega študija Univerze v Ljubljani v šolskem letu 1997/98. Ob vpisu študenti izpolnjujejo vprašalnik, na katerem so vprašanja, ki se nanašajo tudi na naslednje spremenljivke: priimek, spol, poštna številka kraja stalnega bivališča, izobrazba očeta, izobrazba matere, število otrok v družini, oddaljenost stalnega bivališča od fakultete (km), štipendist (Da/Ne), višina štipendije (SIT).

a) Kaj je stvarna, časovna in krajevna opredelitev populacije?
b) Kaj je enota te populacije?
c) Za vsako od spremenljivk navedite vsaj eno smiselno vrednost.
d) Ugotovite vrsto spremenljivke. Za številske spremenljivke ugotovite še vrsto številske spremenljivke.
e) Kakšna je spremenljivka glede na mersko lestvico?
f) Ali je katera od navedenih spremenljivk dvojiška?

2. Prometno dovoljenje

V Republiki Sloveniji vpišejo ob prvi registraciji v prometno dovoljenje za registrirano vozilo tudi naslednje podatke:

- podatki o prvi registraciji: datum, kraj, registrska označba;
- podatki o lastniku: priimek, ime, stalno prebivališče;
- podatki o vozilu: vrsta, znamka, tip, številka šasije, številka motorja, leto izdelave, moč motorja (kW), delovna prostornina (cm^3), masa praznega vozila (kg), število potniških mest.

Odgovorite na ista vprašanja kot pri nalogi Bruci.

3. Populacija vseh vzorcev

V namišljeni populaciji so tri enote, označimo jih A, B, C. Zapišite vse vzorce velikosti 2. Izračunajte, koliko jih je. Pogoji:

a) izbrane enote vračamo, vzorci z istimi elementi v različnem vrstnem redu se razlikujejo;
b) izbranih enot ne vračamo, vzorci z istimi elementi v različnem vrstnem redu se razlikujejo;
c) izbrane enote vračamo, vzorci z istimi elementi v različnem vrstnem redu so ekvivalentni;
d) izbranih enot ne vračamo, vzorci z istimi elementi v različnem vrstnem redu so ekvivalentni.

4. Gornji Dol

V kraju Gornji Dol je bilo dne 1. 1. 1998 60 gospodinjstev. Za anketo potrebujemo vzorec, v katerem je 5 različnih gospodinjstev, vzorci z istimi elementi v različnem vrstnem redu so ekvivalentni. Kolikšno je število vseh vzorcev?

1.3 NAČINI PROUČEVANJA MNOŽIČNIH POJAVOV

Najpomembnejša načina proučevanja množičnih pojavov sta **opazovanje** in **načrtovan poskus**.

1.3.1 Opazovanje

Raziskovalec opazuje oz. pregleduje enote in ugotavlja vrednosti zanj zanimivih spremenljivk pri teh enotah. Iz dobljenih vrednosti izračuna določene karakteristike spremenljivk. Npr.: kolikšen odstotek študentov Univerze v Ljubljani 1998/99 ima štipendijo in kolikšna je povprečna štipendija?

Glede na število proučevanih enot ločimo **popolno** in **delno opazovanje**. Pri popolnem opazovanju proučujemo vse enote v populaciji. To zahteva veliko denarja, časa, dobro organizacijo. Dva najpomembnejša načina popolnega opazovanja sta:

- **popis**: proučujemo vse enote populacije v določenem trenutku. Popisi se ponavljajo v fiksnih razmikih, da se ugotavlja dinamika pojava. Npr. popisi prebivalstva so v desetletnih razmikih;
- **tekoča registracija**: sprotno beležimo določene dogodke v populaciji, npr. registracija rojstev, smrti, opravljenih izpitov. Tekočo registracijo izvajajo pooblaščene službe.

Zelo pogosto popolno opazovanje ni izvedljivo zaradi pomanjkanja denarja, časa, osebja, včasih vseh podatkov ni mogoče dobiti. Zato opazujemo le določene enote iz celotne populacije. Takemu opazovanju rečemo delno opazovanje. Iz populacije izberemo vzorec, ki ga proučujemo, namesto da bi proučevali celotno populacijo. Ker sklepanje temelji na nepopolni informaciji, so statistični sklepi bolj ali manj verjetni.

Glede na način izbire enot ločimo dve vrsti vzorcev:

- **neslučajni vzorci**: izbira enot temelji na neslučajni izbiri enot, npr. na izbiri tipičnih enot, na izbiri najlažje dosegljivih enot;
- **slučajni vzorci**: izbira enot temelji na določenem verjetnostnem zakonu.

Preden začnemo z vzorčenjem, moramo razmisliti, kako naj izbiramo enote iz populacije, koliko enot naj bo v vzorcu ipd. Odgovore na ta vprašanja posreduje matematična statistika, vendar samo za slučajne vzorce. Zato se, če se le da, odločamo za slučajno vzorčenje.

Za izvedbo slučajnega vzorčenja potrebujemo **okvir vzorčenja**, to je seznam (skica) vseh enot populacije. V tem seznamu priredimo vsaki enoti njeno oznako: enote v okviru oštevilčimo s celimi števili od 1 do N.

Ločimo dve vrsti slučajnega vzorčenja: slučajno vzorčenje brez omejitev, slučajno vzorčenje z omejitvami. Če je populacija homogena, uporabimo slučajno vzorčenje brez omejitev, sicer so bolj primerne izvedenke slučajnega vzorčenja, ki upoštevajo različne omejitve.

1.3.1.1 Slučajno vzorčenje brez omejitev

Slučajno vzorčenje brez omejitev je **enostavno slučajno vzorčenje** in **sistematično vzorčenje** (pogojno).

1.3.1.1.1 Enostavno slučajno vzorčenje

Pri enostavnem slučajnem vzorčenju velja, da ima vsaka za vzorčenje razpoložljiva enota populacije na vsakem koraku vzorčenja enako verjetnost, da je izbrana v vzorec. Izbira enot za enostavni slučajni vzorec temelji na uporabi generatorja slučajnih števil, izvedbe pa so različne: loterija, tabele slučajnih števk, računalnik.

- *Loterijski način*: oštevilčene listke z oznakami od 1 do N damo v boben in jih slepo vlečemo;
- *Tabela slučajnih števk*: v Prilogi v Tabeli 1 je niz slučajnih števk. Zaradi lažje berljivosti so števke predstavljene v skupinicah. Ker so števke slučajne, jih lahko začnemo odčitavati kjerkoli v tabeli. Recimo, da je število enot v populaciji k-mestno število. Potem iz tabele slučajnih števk beremo po k slučajnih števk skupaj. Prebrani niz števk se nanaša na

oznako enote v okviru vzorčenja. Če te oznake enote v okviru ni, niz izpustimo. Postopek nadaljujemo toliko časa, da imamo izbranih potrebno število enot za vzorec.

Primer

V populaciji je 855 enot. Za enostavni slučajni vzorec potrebujemo 100 različnih enot. Število $N = 855$ je trimestno, torej je $k = 3$. V tabeli slučajnih števk npr. začnemo v prvi vrstici in prvem stolpcu. Beremo po 3 števke skupaj. Upoštevamo le tista števila, ki so na intervalu 1 do 855; ostale izpustimo:

034, 743, 738, 636, ~~964~~, 736, ...

Prečrtana števila izpustimo. Izbiro nadaljujemo tako dolgo, da izberemo 100 različnih enot.

- Funkcija RANDOM na kalkulatorju generira slučajna števila na intervalu [0, 1), števila imajo tri decimalna mesta. Naj s označuje slučajno število na kalkulatorju, S pa slučajno število na intervalu [1, N], ki ustreza celoštevilski oznaki enote v okviru vzorčenja. Transformacija slučajnega števila s v slučajno število S je:

  $$S = INT\ (N \cdot s + 1)$$

  Funkcija INT (angl. integer) decimalnemu številu odreže vse decimalke.

  Izbira s kalkulatorjem je bistveno hitrejša kot s tabelo slučajnih števk, vendar je postopek s kalkulatorjem primeren le za populacije, ki imajo največ 1000 enot.

Primer

$N = 855$, na kalkulatorju dobljeno slučajno število je $s = 0{,}876$. Potem je:

$$S = INT\ (855 \cdot 0{,}876) + 1 = 749$$

Izbrana je enota z oznako 749.

- *Generatorji psevdo-slučajnih števil v programski opremi*. Za večje primere je smiselno uporabiti generatorje psevdo-slučajnih števil v ustrezni programski opremi.

1.3.1.1.2 Sistematično vzorčenje

Izvedba enostavnega slučajnega vzorčenja na terenu je pogosto težka. Npr. enostavna slučajna izbira dreves v velikem sadovnjaku je tehnično zahtevno in zamudno delo. Zato pogosto enostavno slučajno vzorčenje nadomestimo s sistematičnim vzorčenjem, ki je v praksi lažje izvedljivo.

Najprej izračunamo *korak K* od ene izbrane enote do druge:

$$K = \mathrm{ROUND}\left(\frac{N}{n}\right)$$

Funkcija ROUND zaokroži rezultat na celo število.

Iz okvira vzorčenja sistematično izberemo vsako K- to enoto. Element slučajnosti vpeljemo s slučajno izbiro prve izbrane enote. Izbiramo med enotami na začetku okvira, njihove oznake so 1, 2, …, K. Pri tem uporabimo enostavno slučajno vzorčenje.

Sistematično vzorčenje lahko povzroči, da dobimo v vzorec nekaj enot preveč oz. nekaj enot premalo, odvisno od zaokrožanja pri izračunu koraka. V obeh primerih sistematično vzorčenje ponovimo: če smo dobili nekaj enot preveč, presežek enot izločimo, tako da ponovimo sistematično vzorčenje na izbranih enotah; če smo dobili nekaj enot premalo, izberemo manjkajoče enote s sistematičnim vzorčenjem na neizbranih enotah.

Sistematično vzorčenje ni slučajno vzorčenje, saj izbira temelji na sistemu. Izkaže pa se, da je navadno dober približek enostavnemu slučajnemu vzorčenju. Pogosto se uporablja v kmetijstvu in v gozdarstvu.

Primer

V okviru vzorčenja je 500 enot, ki imajo oznake od 1 do 500. Za vzorec potrebujemo 60 enot. Za izbiro bomo uporabili sistematično vzorčenje. Korak je:

$$K = \text{ROUND}\left(\frac{500}{60}\right) = 8$$

Izbirali bomo vsako osmo enoto iz okvira vzorčenja. S slučajno izbiro enot z oznakami 1, 2, 3, 4, 5, 6, 7, 8 določimo, kje začnemo. Recimo, da je slučajna izbira dala 2. Izbrane so enote z oznako: 2, 10, 18, 26, ..., torej enote: $2 + 8 \cdot k, \quad k = 0,\ 1,\ 2, ...,\ 500/8 = 62$. V vzorec smo dobili 63 enot, torej 3 preveč.

Naredimo nov okvir vzorčenja. Izbrane enote označimo 1 do 63 in ponovimo sistematično vzorčenje, s katerim bomo izmed 63 izločili 3 enote. Izračunamo korak:

$$K = \text{ROUND}\left(\frac{63}{3}\right) = 21$$

Slučajna izbira enot z oznakami 1 do 21 je dala 4. Torej iz okvira izločimo enote z oznakami 4, 25, 46. Preostalih 60 enot je namenjenih za vzorec.

1.3.1.2 Slučajno vzorčenje z omejitvami

Na kratko bomo predstavili dve vrsti slučajnega vzorčenja z omejitvami: **stratificirano vzorčenje** in **večstopenjsko vzorčenje**.

1.3.1.2.1 Stratificirano vzorčenje

V določenih primerih enostavno slučajno vzorčenje ni najbolj primerno. Recimo, da pri polnoletnih osebah proučujemo priljubljenost moderne glasbe. Pričakujemo lahko, da imajo mlajši drugačen odnos do moderne glasbe kot starejši, saj starost osebe vpliva na odnos do moderne glasbe; starost je *moteč dejavnik* pri proučevanju priljubljenosti moderne glasbe. Lahko bi se zgodilo, da bi s slučajno izbiro dobili vzorec, v katerem bi bil delež starejših bistveno večji kot v pripadajoči populaciji. Vzorec ne bi bil reprezentativen.V tem primeru je smiselno populacijo razdeliti na dva dela, na mlajše in na starejše, in iz vsakega dela izbrati enostavni slučajni vzorec ustrezne velikosti.

Če je populacija heterogena, jo poskusimo razdeliti na homogene delne populacije, ki jih imenujemo *stratumi*. Za stratume velja, da so znotraj stratumov podobne enote, med stratumi pa so razlike lahko zelo velike. Stratume je potrebno opredeliti na vsebinski osnovi, torej na osnovi poznavanja motečih dejavnikov, za kar je potrebno dobro poznavanje populacije. Npr. smiselni stratumi pri analizah javnega mnenja so starostne skupine, spol, včasih območje stalnega bivališča ipd.

Za vsak stratum potrebujemo okvir vzorčenja. Z enostavnim slučajnim vzorčenjem izberemo iz vsakega stratuma delni vzorec. Število enot, ki jih izberemo iz posameznega stratuma, je lahko določeno na različne načine; najpogosteje je proporcionalno velikosti stratuma.

Primer:

Televizijska hiša je proučevala gledanost določenih televizijskih oddaj. Populacija gospodinjstev s TV sprejemnikom je štela 70 892 gospodinjstev. Razdeljena je bila na tri

geografske stratume, prvi je štel 12 473 enot, drugi 35 241 in tretji 23 178. Spisek TV naročnikov po regijah je predstavljal okvir vzorčenja.

Odločili so se, da v vzorec vključijo 1% celotne populacije, torej 709 gospodinjstev, in da iz vsakega geografskega stratuma vzorčijo proporcionalni del. Z enostavnim slučajnim vzorčenjem je bilo iz prvega stratuma izbranih 125, iz drugega 352 in iz tretjega 232 gospodinjstev.

Neslučajna alternativa stratificiranemu vzorčenju s proporcionalno izbiro enot je **kvotno vzorčenje**. Uporabljamo ga takrat, kadar okvira vzorčenja za stratume nimamo, imamo pa znane t. i. kvote po stratumih, to je število enot po stratumih. Raziskovalec zavestno izbira enote v vzorec tako, da so izpolnjene kvote. Npr. če proučujemo populacijo polnoletnih oseb, ki je stratificirana po spolu, anketar na terenu izbira enote tako, da je odstotek žensk in moških v vzorcu enak kot v populaciji.

1.3.1.2.2 Večstopenjsko vzorčenje

Pri slučajnem vzorčenju je bistveno, da imamo ustrezen okvir vzorčenja. Le-ta pa pogosto ne obstaja oz. ni dosegljiv. Problem nedosegljivosti okvira vzorčenja lahko včasih rešimo z vzorčenjem v več stopnjah. Recimo, da proučujemo populacijo gimnazijcev v določenem šolskem letu. Spisek vseh gimnazijcev ni dostopen, dostopen pa je spisek gimnazij. Le-ta predstavlja *okvir vzorčenja v prvi stopnji*. Z enostavnim slučajnim vzorčenjem iz vseh gimnazij izberemo določeno število gimnazij. Za vsako izbrano gimnazijo dobimo spisek dijakov; ta služi za *okvir vzorčenja v drugi stopnji*. S slučajno izbiro izberemo določeno število dijakov v gimnazijah, ki so bile izbrane v prvi stopnji.

Pri opisanem primeru je bilo uporabljeno dvostopenjsko vzorčenje, stopenj je lahko tudi več.

1.3.2 Načrtovan poskus

Pri načrtovanem poskusu raziskovalec aktivno posega v proučevani pojav. Enote, ki so vključene v načrtovan poskus, imenujemo *poskusne enote*. Raziskovalec glede na namen raziskave poskusnim enotam priredi *obravnavanja*. V poskusu ugotavlja *izid poskusa*, to so vrednosti proučevane spremenljivke na poskusnih enotah.

Način dodelitve obravnavanj eksperimentalnim enotam sledi istemu konceptu kot pri opazovanju: zadoščeno mora biti principu slučajnosti. *Načrt poskusa* je skica, ki kaže, kako obravnavanja dodelimo poskusnim enotam.

Pri načrtovanih poskusih je zelo pomembna *zasnova poskusa*. Ta mora poleg proučevanih dejavnikov upoštevati tudi moteče dejavnike. Le-ti vplivajo na izid poskusa, vendar niso predmet proučevanja. Če poskuse izvajamo v laboratorijskih pogojih, lahko moteče dejavnike nevtraliziramo v bistveno večji meri kot pri poskusih, ki se izvajajo terenu.

Primer

V sadovnjaku je 30 jablan iste sorte in starosti, vse rastejo v enakih rastnih pogojih. Drevesa slabo rodijo, sadjarji domnevajo, da je vzrok temu pomanjkanje bora v zemlji. Odločili so se, da bodo izvedli poskus tako, da bodo polovici jablan dodali bor, polovici pa ne:

- poskusne enote: jablane v sadovnjaku;
- obravnavanja: kontrola, dodani bor;
- izid poskusa: pridelek na drevo.

Izdelajmo zasnovo poskusa. Drevesa oštevilčimo od 1 do 30 in z enostavnim slučajnim vzorčenjem izberemo 15 dreves. Tem bo dodan bor, preostalim pa ne. Slučajno izbiro smo izvedli s kalkulatorjem.

Tabela 1-1: Izbira dreves z enostavnim slučajnim vzorčenjem

Zaporedna izbira	*s*	*S*
1	0,740	23
2	0,282	9
3	0,778	24
4	0,564	17
5	0,501	16
6	0,967	30
7	0,590	18
8	0,042	2
9	0,824	25
10	0,102	4
11	0,608	19
12	0,913	28
13	0,832	~~25~~
14	0,386	12
15	0,266	8
16	0,540	~~17~~
17	0,098	3

Izbrana so bila drevesa z oznakami: 2, 3, 4, 8, 9, 12, 16, 17, 18, 19, 23, 24, 25, 28, 30.

NALOGE

1. Vaje

V razredu je 49 študentov, pri vajah bo pred tablo sodelovalo 5 študentov, vsak po enkrat. Izberite študente za vaje:

a) z enostavnim slučajnim vzorčenjem;

b) s sistematičnim vzorčenjem.

Uporabite tabelo slučajnih števk. Natančno opišite (vrstica, stolpec), kje ste odčitavali slučajne števke.

2. Anketa v Spodnjem Dolu

Imamo seznam gospodinjstev v kraju Spodnji Dol v letu 1998. Le-teh je 213. Za anketo potrebujemo 20 različnih gospodinjstev. Izberite gospodinjstva za anketo:

a) z enostavnim slučajnim vzorčenjem;

b) s sistematičnim vzorčenjem.

3. Cvetličarne

Da bi ugotovili, ali je v mestu dovolj cvetličarn, so se raziskovalci odločili izvesti telefonsko anketo, v katero bi vključili 150 telefonskih naročnikov. V mestu je 955 telefonskih naročnikov, njihove telefonske številke so v telefonskem imeniku.

a) Kaj je proučevana populacija?

b) Kaj predstavlja telefonski imenik?

c) Izvedite sistematično vzorčenje.

4. Gledanje TV

Raziskovalce zanima, koliko ur tedensko gleda študent Univerze v Ljubljani televizijo. V raziskavi bodo upoštevali le redno vpisane študente v tekočem šolskem letu. Raziskovalci so na osnovi predhodnih študij predpostavljali, da na gledanje TV vpliva način bivanja študentov: bivanje pri starših oz. sorodnikih, bivanje v študentskih domovih ter ostalo (najemniki, lastno gospodinjstvo). Razmislite:

a) Kaj v tej raziskavi predstavlja spremenljivka 'način bivanja'? Kaj določa?
b) Kakšen način vzorčenja je primeren, če je okvir vzorčenja po stratumih raziskovalcem dosegljiv?
c) Kakšen način vzorčenja je primeren, če okvir vzorčenja po stratumih raziskovalcem ni dosegljiv? Iz predhodnih raziskav je znano, da približno tretjina študentov biva pri starših ali sorodnikih, tretjina v študentskih domovih in tretjina drugje.
d) Ali bi bili rezultati reprezentativni za vse redno vpisane študente Univerze v Ljubljani v tekočem šolskem letu ali za vse študente Univerze v Ljubljani?

5. Travnik

Na travniku pravokotne oblike z merami 60 m × 40 m so želeli ugotoviti sestavo travne ruše. Na skici travnika so naredili mrežo 1 m × 1 m. Z enostavnim slučajnim vzorčenjem bodo izbrali 20 parcelic velikosti 1 m^2, ki ne smejo ležati na robu travnika. Izvedite postopek izbire.

Namig: vsaka parcelica naj bo opisana s koordinato v smeri osi x in v smeri osi y.

2 OPISNA STATISTIKA

2.1 RELATIVNA ŠTEVILA

Podatki postanejo zanimivi tedaj, ko jih med seboj primerjamo. Števila lahko primerjamo absolutno, npr. z razliko, ali pa relativno, npr. z razmerjem. Relativna števila primerjajo dva podatka z njunim razmerjem. Glede na to, v kakšni vsebinski povezavi sta podatka, ki ju primerjamo, imamo tri vrste relativnih števil: **strukture**, **koeficiente** in **indekse**. Pogledali si bomo njihov izračun in pripadajoče grafične prikaze.

2.1.1 Strukture

Celota je razdeljena v K skupin. Naj f_i določa število enot v i-ti skupini, $i = 1, 2, \ldots K$, to število imenujemo **frekvenca**. Pri izračunu strukture primerjamo število enot, ki jih ima posamični del celote, s številom enot, ki jih ima celota. Strukture izražamo v deležih ali odstotkih. Delež enot v i-ti skupini f_i^0 je:

$$f_i^0 = \frac{f_i}{\sum_{i=1}^{K} f_i}$$

Deleži so med 0 in 1. Vsota deležev je 1.

Odstotek enot v i-ti skupini $f_i\,\%$ je:

$$f_i\,\% = \frac{f_i}{\sum_{i=1}^{K} f_i} \cdot 100$$

Odstotki so med 0 in 100. Vsota odstotkov je 100. Izračun odstotkov je smiseln, če ima celota vsaj 100 enot, torej predstavlja 1 enota največ 1%. V rezultatih navadno navajamo odstotke zaokrožene na eno decimalno mesto.

Frekvence, izražene v deležih ali odstotkih imenujemo **relativne frekvence**.

Da pojasnimo, kaj določa skupine pri izračunu strukture, govorimo o *strukturi po …*; npr. struktura po spolu, struktura po smeri študija itd.

Za grafični prikaz strukture uporabljamo **strukturni stolpec** in **strukturni krog**.

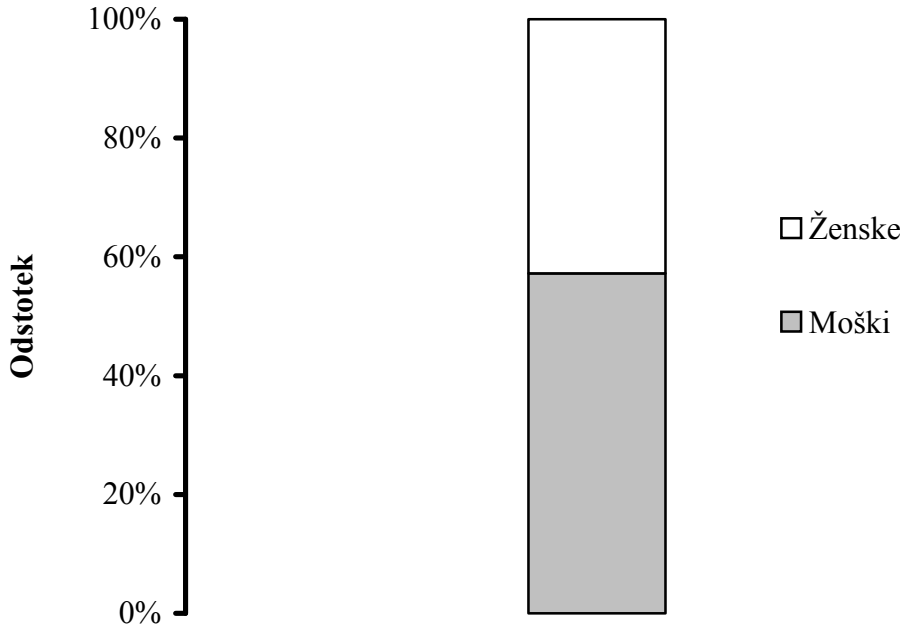
Primer

V tabeli so podatki za število študentov po smeri študija in po spolu za 1. letnik Biotehniške fakultete v 1997/98. Ogledali si bomo izračun različnih struktur in pripadajoče grafične prikaze.

Tabela 2-1: Število študentov 1. letnika BF v 1997/98 po smeri študija in po spolu (Vir: Arhiv Biotehniške fakultete)

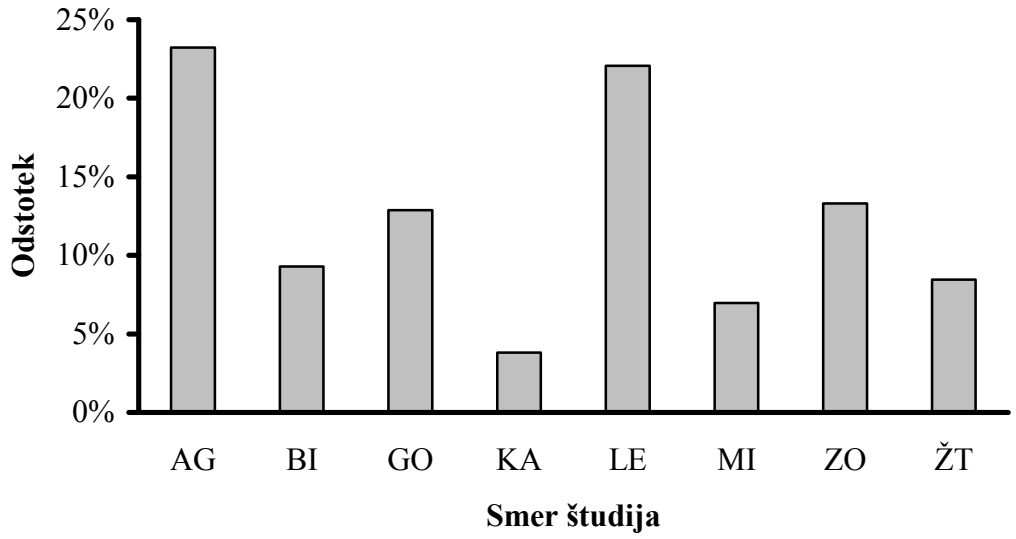
Smer študija	Moški	Ženske	Skupaj
Agronomija (AG)	103	117	220
Biologija (BI)	26	62	88
Gozdarstvo (GO)	98	24	122
Krajinska arhitektura (KA)	11	25	36
Lesarstvo (LE)	188	21	209
Mikrobiologija (MI)	19	47	66
Zootehnika (ZO)	63	63	126
Živilska tehnologija (ŽT)	34	46	80
Skupaj	542	405	947

- Grafični prikaz strukture po spolu s strukturnim stolpcem:



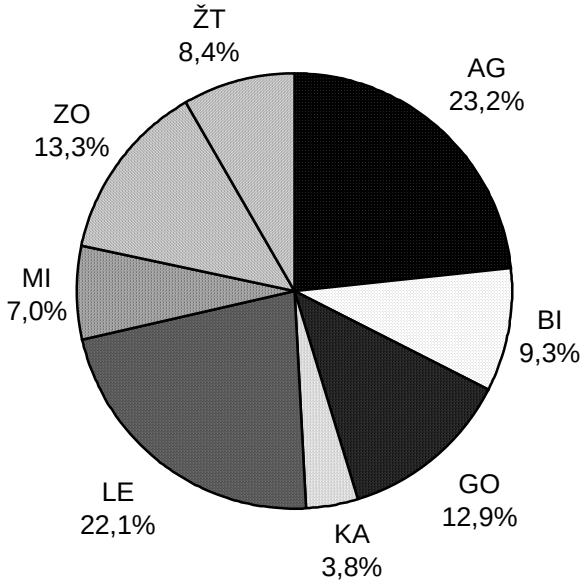
Slika 2-1: Struktura po spolu za študente 1. letnika Biotehniške fakultete 1997/98

- Grafični prikaz strukture po smeri študija z *razrezanim strukturnim stolpcem*. Ker je smeri študija osem, je bolj pregledno, če narišemo strukturni stolpec razrezan po delih, ki tvorijo 100%.



Slika 2-2: Struktura po smeri študija za študente 1. letnika Biotehniške fakultete 1997/98 prikazana z razrezanim strukturnim stolpcem

- Alternativni grafični prikaz strukture po smeri študija je s strukturnim krogom.



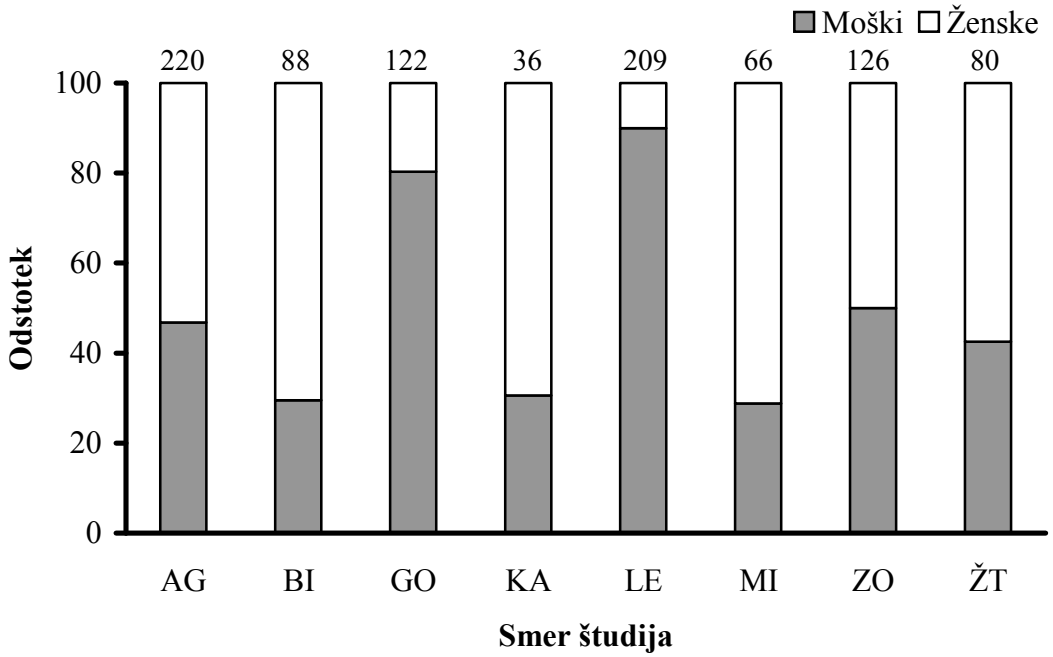
Slika 2-3: Struktura po smeri študija za študente 1. letnika Biotehniške fakultete 1997/98 prikazana s strukturnim krogom

Vizualna primerjava dveh krožnih izsekov je težja kot vizualna primerjava dveh stolpcev. Kadar je skupin veliko, postane krožni prikaz še posebej nepregleden. Zato je za grafičen prikaz strukture ponavadi bolje uporabiti strukturni stolpec kot strukturni krog. Krožni prikaz v treh razsežnostih zna biti še posebno zavajajoč, zato njegovo uporabo odsvetujemo.

- Izračun strukture po spolu za vsako smer študija je v tabeli. Najbolj ustrezen grafičen prikaz je prikaz s stolpci, kakor kaže slika.

Tabela 2-2: Struktura po spolu (%) za vsako smer študija v 1. letniku Biotehniške fakultete 1997/98

Smer študija	Moški	Ženske	Skupaj
AG	46,8	53,2	100,0
BI	29,5	70,5	100,0
GO	80,3	19,7	100,0
KA	30,6	69,4	100,0
LE	90,0	10,0	100,0
MI	28,8	71,2	100,0
ZO	50,0	50,0	100,0
ŽT	42,5	57,5	100,0



Slika 2-4: Struktura po spolu za vsako smer študija v 1. letniku BF v 1997/98. Nad stolpci je navedeno število študentov.

Slika kaže, da na študiju gozdarstva in lesarstva prevladujejo študenti, na študiju biologije, krajinske arhitekture in mikrobiologije pa študentke. Na ostalih treh smereh študija (agronomija, zootehnika in živilska tehnologija) je približno polovica študentov in polovica študentk. Iz prikaza strukture ni razvidno število študentov na vsaki smeri. Npr. na agronomiji 220, na krajinski arhitekturi 36. To informacijo napišemo na vrh pripadajočega stolpca.

Opomba: določene lastnosti izražamo v odstotkih, npr. relativna vlaga, obrestna mera, zasedenost turističnih ležišč. **Odstotna točka** izraža absolutno razliko dveh vrednosti spremenljivke, ki je izražena v odstotkih. Npr.: lahko mleko ima 1,5% maščob, navadno mleko 3,2%. Navadno mleko ima za 1,7 odstotne točke več maščob kot lahko mleko.

2.1.2 Koeficienti

Koeficient je razmerje dveh podatkov, ki sta vsebinsko povezana. Pogosto uporabljeni koeficienti so: gostota prebivalstva, število bolnikov na zdravnika, hitrost, poraba goriva na 100 km itd. Opozarjamo, da moramo poleg vrednosti koeficienta navesti tudi ustrezne merske enote. Včasih je vsebinsko smiselno izračunati tudi ustrezen *recipročni koeficient*. Npr. za Slovenijo imamo za leto 1991 podatke za število avtomobilov in za število prebivalcev: število avtomobilov = 594 289, število prebivalcev = 1 965 986. Smiselno je izračunati število prebivalcev na en avtomobil in število avtomobilov na 10 prebivalcev: 3,3 prebivalci/avto; 3,0 avta/10 prebivalcev.

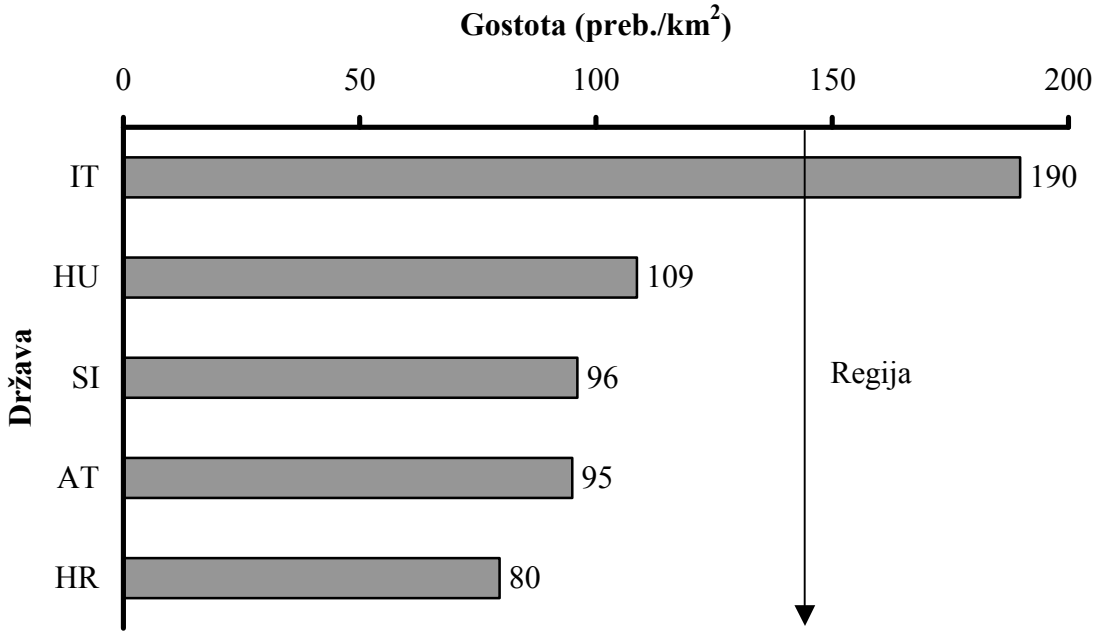
Grafično prikažemo izračunane koeficiente s stolpci. Lahko jih uredimo po velikosti in prikažemo navpično ali vodoravno.

Primer

Izračunali bomo gostoto prebivalstva za Slovenijo in za sosednje države ter za regijo, ki jo te države sestavljajo. Rezultate bomo grafično predstavili.

Tabela 2-3: Površina (km^2), število prebivalcev (1000) za Slovenijo in njene sosede v letu 1995 in izračunana gostota prebivalstva (Vir: Encarta Atlas, Microsoft, 1998)

Država	Površina (km^2)	Št. preb. (1000)	Gostota (preb./km^2)
Slovenija (SI)	20250	1946	96,1
Avstrija (AT)	83850	7968	95,0
Italija (IT)	301270	57187	189,8
Madžarska (HU)	93030	10115	108,7
Hrvaška (HR)	56540	4498	79,6
Regija	554940	81714	147,2



Slika 2-5: Gostota prebivalstva za Slovenijo in njene sosede v letu 1995. S poltrakom je predstavljena gostota prebivalstva za regijo, ki jo te države sestavljajo.

Posebne vrste koeficienti so **stopnje**, pogosto se uporabljajo v demografiji in medicini. Npr. letna stopnja rodnosti se izračuna z razmerjem: število živorojenih v vsem letu / število prebivalcev (v 1000).

Ker se število prebivalcev spreminja, običajno uporabimo oceno za število prebivalcev, ki se nanaša na sredino leta. Podobno računamo stopnjo smrtnosti, stopnjo naravnega prirasta.

Primer

Izračunali bomo stopnjo rodnosti in stopnjo smrtnosti po letih v obdobju 1980 - 1995 za Slovenijo.

Tabela 2-4: Število prebivalcev ocenjeno na dan 30. 6., število živorojenih in število umrlih po letih v obdobju 1980-2000 (Vir: SL-97, str. 85, SL-01) ter izračunana stopnja rodnosti in stopnja smrtnosti

Leto	Število prebivalcev	Število živorojenih	Število umrlih	Stopnja rodnosti (število živorojenih na 1000 prebivalcev)	Stopnja smrtnosti (število umrlih na 1000 prebivalcev)
1980	1901208	29902	18820	15,7	9,9
1985	1973151	25933	19854	13,1	10,1
1990	1998090	22368	18555	11,2	9,3
1995	1987505	18980	18968	9,5	9,5
2000	1990272	18180	18588	9,1	9,3

V poglavju o verižnih indeksih bomo spoznali še stopnjo rasti.

V statistični teoriji je več pomembnih mer, ki se izražajo s koeficientom, npr. koeficient variacije, koeficient korelacije, koeficient determinacije. Spoznali jih bomo kasneje.

2.1.3 Indeksi

Število x_i je vrednost številske spremenljivke X v i-ti skupini, $i = 1, 2, \ldots K$. Zaporedje vrednosti $x_1, x_2, \ldots x_K$ imenujemo **statistična vrsta**. Primeri statističnih vrst:

- vrednost potrošniške košarice (SIT) po krajih (krajevna vrsta);
- povprečna plača (SIT) po poklicih (stvarna vrsta);
- število brezposelnih po letih (časovna vrsta).

Izberemo en podatek v vrsti. Označimo ga x_0, imenujemo ga *osnova*. Indeksi primerjajo vsak podatek v vrsti z osnovo:

$$I_{i/0} = \frac{x_i}{x_0} \cdot 100, \; i = 1, 2, \ldots K$$

Indeks 120,5 pomeni, da je podatek za 20,5% večji od osnove; indeks 85,0 pomeni, da je podatek za 15,0% manjši od osnove. Indekse navadno predstavljamo na eno decimalno mesto. Izbira osnove je odvisna od tega, na katero skupino želimo primerjati druge skupine, torej je vsebinski problem.

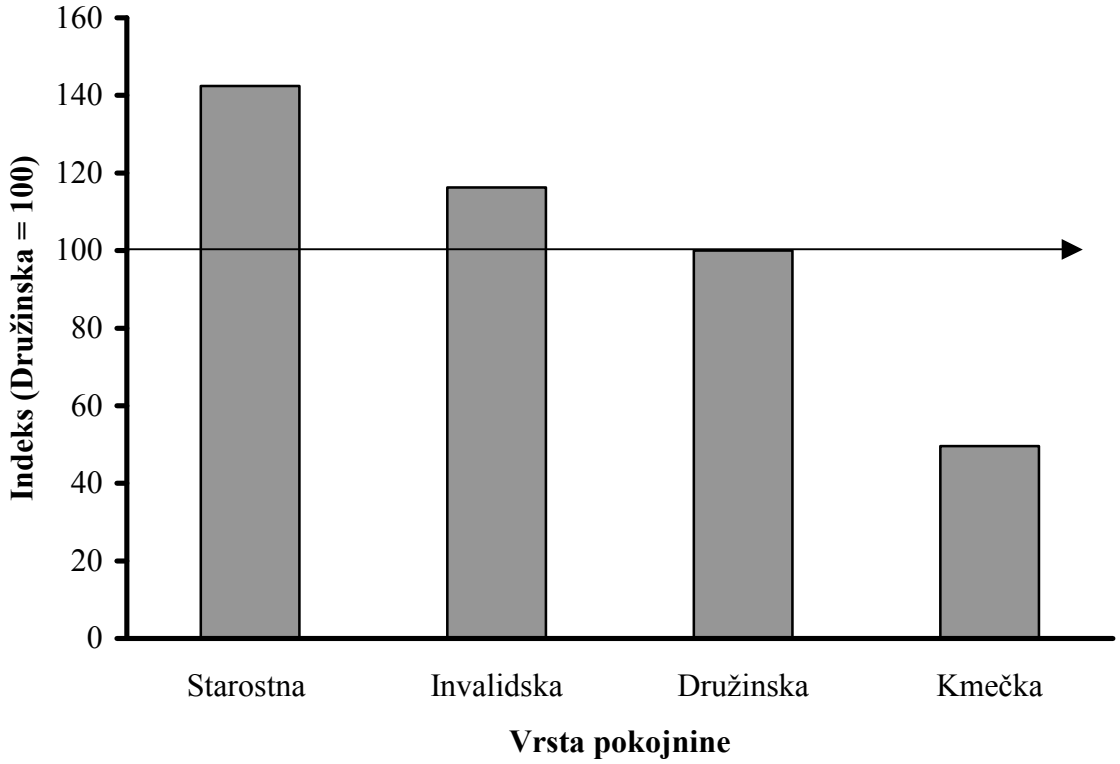
Za izračun indeksov mora biti merska lestvica spremenljivke razmernostna, vrednosti pozitivne. Za grafični prikaz indeksov uporabljamo stolpce.

Primer

V tabeli navajamo povprečni mesečni prejemek za starostne, invalidske, družinske in kmečke pokojnine za leto 1996 v Sloveniji. Za osnovo bomo vzeli družinske pokojnine in izračunali indekse ter rezultate grafično prikazali.

Tabela 2-5: Povprečni mesečni prejemek (SIT) v letu 1996 za razne vrste pokojnin (Vir: Slovenija v številkah 1997) in izračunani indeksi z osnovo 'Družinska'

Vrsta pokojnine	Mesečni prejemek (SIT)	Družinska = 100
Starostna	61586	142,4
Invalidska	50278	116,3
Družinska	43241	100,0
Kmečka	21440	49,6



Slika 2-6: Indeksi z osnovo 'Družinska' za povprečne mesečne prejemke po vrstah pokojnine za leto 1996 za Slovenijo

2.1.3.1 Časovni indeksi

Časovni indeksi so osnovno orodje za analizo časovne vrste. Poglejmo najprej opredelitev časovne vrste.

Številsko spremenljivko Y opazujemo v času, torej je $Y = Y(t)$. Podatki se nanašajo na zaporedna časovna obdobja: $t_1, t_2, \ldots, t_T$. Statistično vrsto $y_1, y_2, \ldots y_T$ imenujemo **časovna vrsta**, T je dolžina časovne vrste.

Grafični prikaz časovne vrste je **linijski grafikon**, na abscisni osi je časovna skala. Pri grafičnem prikazu moramo ustrezno prikazati časovno zaporedje vrednosti. Če časovna vrsta ni ekvidistantna (med opazovanji ni enak časovni razmik), moramo na sliki to upoštevati.

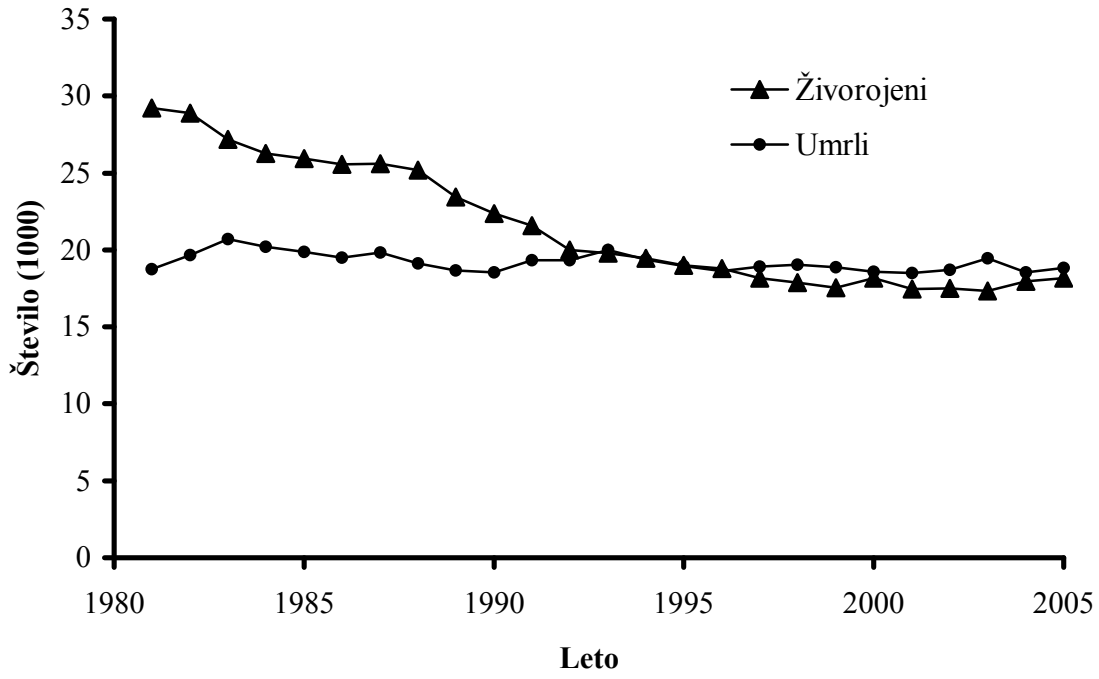
Primer

V tabeli so navedeni podatki o številu živorojenih in o številu umrlih po letih v Sloveniji v obdobju 1981 - 2005.

Tabela 2-6: Število živorojenih in število umrlih v Sloveniji po letih v obdobju 1981 – 2005 (Vir: SL - 94, str. 74, SL – 99, str. 84, SL-06)

Leto	Živorojeni	Umrli
1981	29220	18733
1982	28894	19647
1983	27200	20703
1984	26274	20214
1985	25933	19854
1986	25570	19499
1987	25592	19837
1988	25209	19126
1989	23447	18669
1990	22368	18555
1991	21583	19324
1992	19982	19333
1993	19793	20012
1994	19463	19359
1995	18980	18968
1996	18788	18620
1997	18165	18928
1998	17856	19039
1999	17533	18885
2000	18180	18588
2001	17477	18508
2002	17501	18701
2003	17321	19451
2004	17961	18523
2005	18157	18825

Grafično prikažimo podatke za število živorojenih in za število umrlih.



Slika 2-7: Število živorojenih in število umrlih v Sloveniji po letih v obdobju 1981 - 2005

Število živorojenih v opazovanem obdobju pada, rečemo, da ima časovna vrsta negativen trend. Število umrlih je skoraj konstantno. Naravni prirast, to je razlika med številom živorojenih in številom umrlih, je od leta 1993 zelo majhen ali celo negativen.

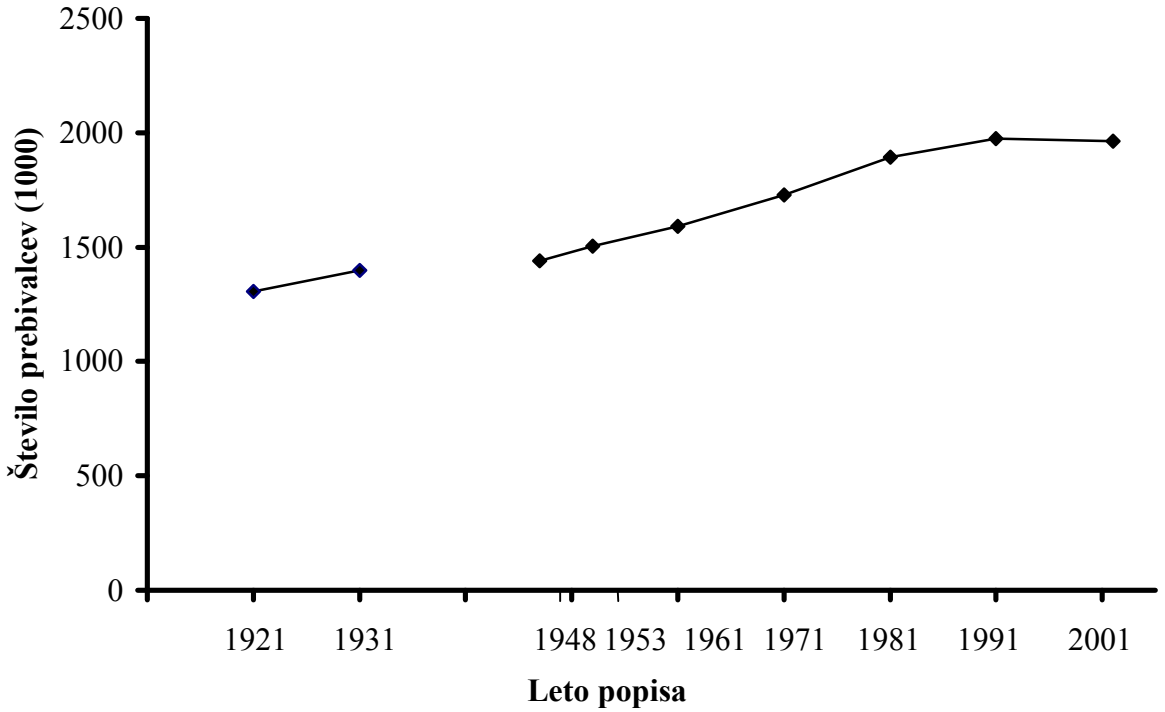
Primer

Popisi prebivalstva po državah se praviloma izvajajo v desetletnih intervalih. Na ozemlju Slovenije je bil popis prebivalstva leta 1921, leta 1941 popisa prebivalstva zaradi vojne ni bilo. Po II. svetovni vojni je bil popis leta 1948 in 1953. Podatki o številu prebivalcev v tisočih so v tabeli.

Tabela 2-7: Število prebivalcev po zaporednih popisih prebivalstva v Sloveniji v obdobju 1921 - 2002 (Vir: SL - 91, str. 71, Popisni atlas Slovenije 2002)

Leto popisa	Število prebivalcev (1000)
1921	1305
1931	1398
1948	1440
1953	1504
1961	1592
1971	1727
1981	1892
1991	1975
2002	1964

Grafično prikažimo podatke.



Slika 2-8: Število prebivalcev po zaporednih popisih prebivalstva v Sloveniji

Točk za leti 1931 in 1948 nismo povezali, ker je bila vmes vojna, ki je bistveno vplivala na število prebivalcev.

Osnovno orodje za analizo časovnih vrst so indeksi in sicer:

- **indeksi s stalno osnovo**,
- **indeksi s premično osnovo**.

2.1.3.1.1 Indeksi s stalno osnovo

Neki podatek v časovni vrsti izberemo za osnovo, označimo ga y_0. Vse podatke primerjamo na ta podatek, osnova je stalna:

$$I_{t/0} = \frac{y_t}{y_0} \cdot 100, \;\; t = 1, 2, \ldots T$$

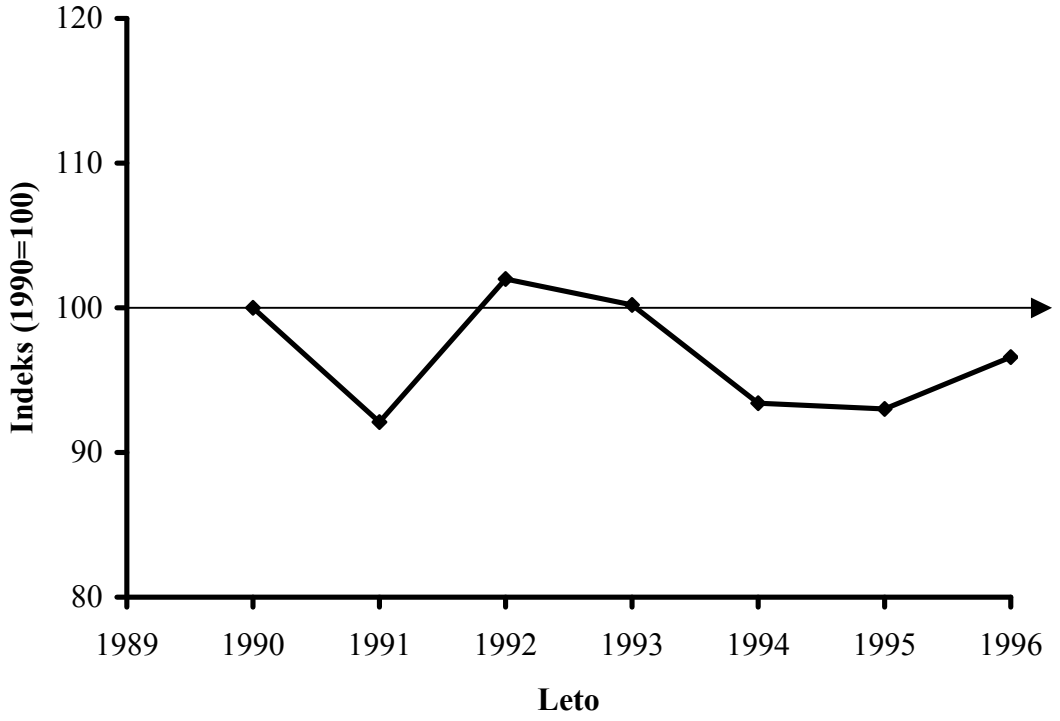
Rezultat je nova časovna vrsta, vse vrednosti osnovne časovne vrste so deljene z vrednostjo y_0. Za osnovo ne izbiramo neobičajnih obdobij, npr. obdobij s sušo, vojno, epidemijo.
Grafični prikaz indeksov s stalno osnovo je enak kot prikaz izhodiščne časovne vrste.

Primer

Podajamo podatke za število štipendistov v Sloveniji po letih v obdobju 1990 – 1996. Izračunali bomo indekse z osnovo 1990 in jih grafično prikazali.

Tabela 2-8: Število štipendistov (v 1000) v Sloveniji v obdobju 1990-1996 (Vir: SL-97, str. 131) in izračunani indeksi s stalno osnovo 1990

Leto	Št. štipendistov (1000)	1990=100
1990	65,6	100,0
1991	60,4	92,1
1992	66,9	102,0
1993	65,7	100,2
1994	61,3	93,4
1995	61,0	93,0
1996	63,4	96,6



Slika 2-9: Indeksi z osnovo 1990 za število štipendistov v Sloveniji po letih v obdobju 1990-1996

2.1.3.1.2 Indeksi s premično osnovo

Osnova ni stalna, ampak se spreminja. Če za osnovo vedno vzamemo predhodni podatek v časovni vrsti, dobimo **verižne indekse**:

$$I_{t/t-1} = \frac{y_t}{y_{t-1}} \cdot 100, \ t = 2, 3, \ldots T$$

Verižni indeksi so smiselni samo, če je časovna vrsta ekvidistantna. Prvi verižni indeks ne obstaja; standardni znak za neobstoječo vrednost je —.

Pri interpretaciji uporabljamo **stopnjo rasti**. Dobimo jo tako, da od verižnega indeksa odštejemo 100:

$S_{t/t-1} = I_{t/t-1} - 100, \ t = 2, 3, \ldots T$.

Stopnja rasti je relativna sprememba, ki v odstotkih izraža spremembo pojava:

$$S_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100$$

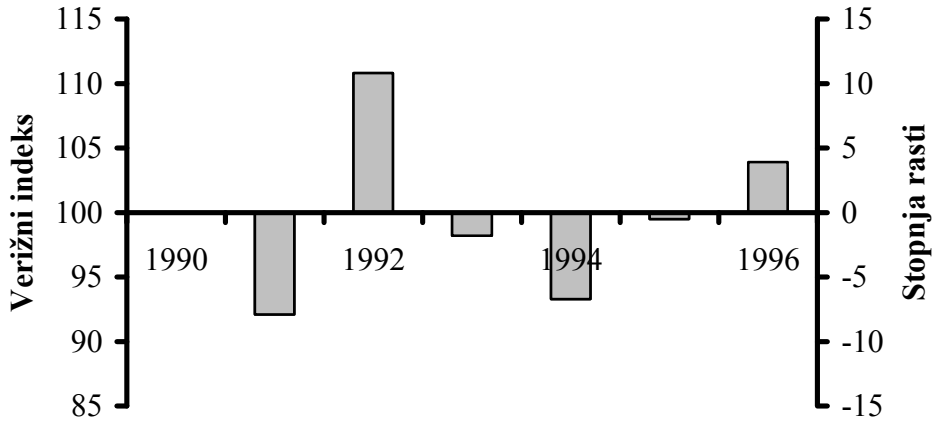
Stopnja rasti je lahko pozitivna, negativna ali ničelna. Za grafičen prikaz verižnih indeksov in stopenj rasti uporabljamo prikaz s stolpci.

Primer

Izračunajmo verižne indekse in stopnje rasti za število štipendistov in rezultate grafično prikažimo.

Tabela 2-9: Število štipendistov (v 1000) v Sloveniji v obdobju 1990-1996 (Vir: SL-97, str. 131) in izračunani verižni indeksi in stopnje rasti

Leto	Št. štipendistov (1000)	Verižni indeks	Stopnja rasti
1990	65,6	—	—
1991	60,4	92,1	-7,9
1992	66,9	110,8	10,8
1993	65,7	98,2	-1,8
1994	61,3	93,3	-6,7
1995	61,0	99,5	-0,5
1996	63,4	103,4	3,9



Slika 2-10: Verižni indeksi in stopnje rasti za število štipendistov po letih v obdobju 1990-1996

NALOGE

1. Zemljišča v Sloveniji

Skupna površina zemljišč se deli na nerodovitna zemljišča, na gozdna zemljišča ter na kmetijska zemljišča. Kmetijska zemljišča se delijo na travinje, to so travniki in pašniki, na njive in vrtove, in na sadovnjake in vinograde. V tabeli so podatki za Slovenijo za leto 1990.

Tabela 2-10: Površina zemljišč (ha) po vrsti zemljišč v Sloveniji leta 1990 (Vir: SL - 91, str. 214)

Vrsta zemljišč	Površina zemljišč (ha)	Površina kmet. zemljišč (ha)
Nerodovitna	136754	
Gozdna	1024535	
Kmetijska	864184	
-travinje		559264
-njive, vrtovi		247083
-sadov., vinog.		57837
Skupaj	2025473	864184

Izračunajte strukturo površin in strukturo kmetijskih površin.

2. Radio

V tabeli je podan čas predvajanja glasbenega programa po programih Radia Slovenija ter po vrsti glasbe za leto 1992.

Tabela 2-11: Čas (ure) predvajanja glasbenega programa po programih Radia Slovenija in po vrstah glasbe

Vrsta glasbe	Prvi program	Drugi program	Tretji program
Resna	106	0	3723
Zabavna in popularna	4320	3491	113
Narodna glasba	45	0	36
Glasba za otroke	16	0	145
Druga glasba	107	177	191

a) Izračunajte strukturo po vrsti glasbe za vsak program.

b) Rezultate grafično prikažite in obrazložite.

3. Zasedenost ležišč

V tabeli so podatki o številu ležišč in o številu nočitev gostov po vrstah krajev v Sloveniji v letu 1993.

Tabela 2-12: Število ležišč po vrstah krajev, stanje 31. 08. 1993 v Sloveniji in število nočitev gostov v letu 1993 po vrstah krajev v Sloveniji (Vir: SL - 94, str. 388 in str. 397)

Vrste krajev	Ležišča	Nočitve gostov
Ljubljana	3285	322883
Zdraviliški kraji	12111	1819398
Obmorski kraji	21867	1454500
Gorski kraji	27475	1259751
Drugi turistični kraji	8998	493560
Drugi kraji	1089	34534
Skupaj	74825	5384626

a) Izračunajte zasedenost ležišč (%) v letu 1993 po vrstah krajev in za Slovenijo.

b) Rezultate grafično predstavite in obrazložite.

c) Primerjajte zasedenost v Ljubljani z zasedenostjo v obmorskih krajih.

4. Queteletovo razmerje

Za sebe izračunajte Queteletovo razmerje: Q = telesna masa (kg) / (telesna višina (m))^2

in ugotovite, v katero skupino ste uvrščeni:

do pod 18:	shujšanost,
18 do pod 20:	premajhna teža,
20 do pod 27:	normalna teža,
27 do pod 30:	prekomerna teža,
30 in več:	debelost.

5. Gostota prebivalstva v Franciji in Monaku

Tabela 2-13: Površina (km^2) in število prebivalcev (1000) za Francijo in Monako za leto 1995 (Vir: Encarta Atlas, Microsoft, 1998)

Država	Površina (km^2)	Štev. preb. (1000)
Francija	551 500	57 981
Monaco	2	32

a) Izračunajte gostoto prebivalstva za Francijo in za Monako.

b) Izračunajte gostoto prebivalstva za regijo 'Francija in Monako'. Primerjajte dobljeno gostoto z gostotama, izračunanima pod a). Kaj ugotovite?

6. Prometne nesreče

V tabeli je navedeno število prometnih nesreč s smrtnim izidom ali s telesno poškodbo na cestah v Sloveniji in število smrtno ponesrečenih v teh nesrečah.

Tabela 2-14: Število prometnih nesreč in število smrtno ponesrečenih po letih v obdobju 1988 - 1998 (Vir: SL - 93, str. 274, SL - 99, str. 364)

Leto	Prometne nesreče	Število smrtno ponesrečenih
1988	6085	551
1989	5825	554
1990	5177	517
1991	5479	462
1992	5882	492
1993	6290	493
1994	6586	505
1995	6540	415
1996	6348	389
1997	6951	358
1998	5864	309

a) Za število prometnih nesreč izračunajte: indekse z osnovo 1988, verižne indekse in stopnje rasti. Rezultate grafično prikažite in obrazložite.

b) Za število smrtno ponesrečenih izračunajte: indekse z osnovo 1992, verižne indekse in stopnje rasti. Rezultate grafično prikažite in obrazložite.

7. Brezposelni v Sloveniji

V tabeli prikazujemo število brezposelnih v Sloveniji po letih v obdobju 1986 - 1998.

Tabela 2-15: Število registriranih brezposelnih oseb na dan 31. 12. po letih v Sloveniji v obdobju 1986 - 1998 (Vir: SL 97, str. 207, SL - 99, str. 234)

Leto	Število brezposelnih (1000)
1986	14,0
1987	17,8
1988	25,4
1989	33,8
1990	55,4
1991	91,2
1992	118,2
1993	137,1
1994	123,5
1995	126,8
1996	124,5
1997	128,6
1998	126,6

a) Izračunajte in grafično prikažite indekse z osnovo 1986 za število brezposelnih. Obrazložite vrednost indeksov $I_{90/86}$ in $I_{93/86}$.

b) Izračunajte in grafično prikažite verižne indekse in stopnje rasti. Obrazložite njihove vrednosti. V katerem letu je bila stopnja rasti največja?

2.2 FREKVENČNA PORAZDELITEV

Frekvenčna porazdelitev je tabela, ki jo določajo vrednosti spremenljivke in pripadajoče frekvence. Spremenljivka *X*, ki določa frekvenčno porazdelitev, je opisna ali številska. Njene vrednosti v frekvenčni porazdelitvi imenujemo:

- *skupine* (*kategorije*), če je *X* opisna
- *razredi*, če je *X* številska.

a) frekvenčna porazdelitev po spolu, 2 skupini vrednosti

Tabela 2-16: Študenti po spolu

Spol	Število študentov
Moški	542
Ženske	405
Skupaj	947

b) frekvenčna porazdelitev po teži, 5 razredov vrednosti

Tabela 2-17: Otroci po teži (kg)

Teža (kg)	Število otrok
do pod 20	12
20 do pod 30	210
30 do pod 40	540
40 do pod 50	40
50 in več	2
Skupaj	804

Frekvenčne porazdelitve delamo zato, da pridobimo preglednost nad podatki, npr. podatki za težo 804 otrok bi bili popolnoma nepregledni. Frekvenčno porazdelitev dobimo takole:

1. Določimo skupine oz. razrede, tako da je vsak podatek uvrščen v natanko eno skupino oz. razred.

   Pri b) bi bili razredi za težo lahko oblikovani takole: do 20, nad 20 do 30, nad 30 do 40, itd.; narobe pa bi bilo določiti razrede takole: do 20, 20 do 30, 30 do 40 itd, saj ne bi bilo jasno, v kateri razred je uvrščen podatek 20, 30 itd.
2. Izvedemo postopek uvrščanja podatkov v skupine oz. razrede.
3. Izpis frekvenčne tabele: izpis skupin oz. razredov in pripadajočih frekvenc.

V nadaljevanju bomo pogledali, kako dopolnimo in grafično prikažemo frekvenčno porazdelitev za opisno in za številsko spremenljivko.

2.2.1 Frekvenčna porazdelitev za opisno spremenljivko

Za grafično predstavitev frekvenčne porazdelitve za opisno spremenljivko uporabljamo prikaz s stolpci ali s krogom. Frekvenčno porazdelitev za opisno spremenljivko dopolnimo z izračunom strukture. Za grafično predstavitev relativnih frekvenc uporabljamo strukturni stolpec ali strukturni krog (glej Relativna števila).

2.2.2 Frekvenčna porazdelitev za številsko spremenljivko

Ko podatke uvrstimo v razrede, zgubimo nekaj informacije, pridobimo pa pri preglednosti. Več kot je razredov, manj informacije izgubimo, preglednost pa je slabša. Pri dovolj velikem številu podatkov (nad 100) je smiselno število razredov med 6 in 12.

Frekvenčno porazdelitev za številsko spremenljivko dopolnimo s karakteristikami razredov: spodnja/zgornja meja razreda, širina razreda, sredina razreda.

Recimo, da imamo *K* razredov. Vsakemu razredu določimo *spodnjo mejo razreda* $x_{i,\min}$ in *zgornjo mejo razreda* $x_{i,\max}$, tako da velja:

$x_{i,\max} = x_{i+1,\min}$, $i = 1, 2, 3, \ldots K\text{-}1$

Zgornja meja razreda se ujema s spodnjo mejo naslednjega razreda.

Izračunamo *širino razreda* d_i in *sredino razreda* x_i:

$$d_i = x_{i,\max} - x_{i,\min}$$

$$x_i = (x_{i,\max} + x_{i,\min}) / 2$$

Razred, ki nima spodnje oz. zgornje meje, imenujemo *odprt razred*. Tak razred nima širine in sredine.

Karakteristike razredov so pomožne količine, ki jih uporabljamo pri grafični predstavitvi frekvenčne porazdelitve in pri njeni analizi. Najpomembnejša karakteristika razreda je njegova sredina, ki je predstavnik vseh vrednosti v razredu. Da dobimo sredino, potrebujemo spodnjo in zgornjo mejo. Za njuno določanje je najbolj pogosto primerno 'pravilo polovic': spodnjo mejo razreda določimo tako, da od najmanjše vrednosti v razredu odštejemo polovico; zgornjo mejo razreda pa tako, da največji vrednosti v razredu prištejemo polovico. Poglejmo primer.

Tabela 2-18: Karakteristike razredov za težo otrok

Teža (kg)	$x_{i,\min}$	$x_{i,\max}$	d_i	x_i
do pod 20	—	19,5	—	—
20 do pod 30	19,5	29,5	10	24,5
30 do pod 40	29,5	39,5	10	34,5
40 do pod 50	39,5	49,5	10	44,5
50 in več	49,5	—	—	—

Pravilo polovic ni primerno uporabljati pri starostnih razredih, če je starost izražena v t.i. dopolnjenih letih. V razredu od 0 do 9 dopolnjenih let so osebe, ki do dneva zbiranja podatkov niso dopolnile 10 let. Za ta razred je spodnja meja 0 in zgornja meja 10 in sredina razreda 5.

Tabela 2-19: Karakteristike razredov za starost oseb

Starost (dopolnjena leta)	$x_{i,\min}$	$x_{i,\max}$	d_i	x_i
0 do 9	0	10	10	5
10 do 19	10	20	10	15
20 do 29	20	30	10	25
…	…	…	…	…

Frekvenčno tabelo dopolnimo z naslednjimi izračuni: relativna frekvenca $f_i\%$, gostota frekvence g_i (pogojno), kumulativa frekvenc F_i, kumulativa relativnih frekvenc $F_i\%$.

Če so razredi različno široki, frekvence po razredih niso primerljive. Tedaj za vsak razred izračunamo še **gostoto frekvence**:

$$g_i = \frac{f_i}{d_i}$$

Gostote frekvenc so po razredih primerljive.

Za vsak razred izračunamo **kumulativo frekvenc** F_i. To je število podatkov, uvrščenih do zgornje meje *i*-tega razreda. Računamo jo po rekurzivni formuli:

$$F_0 = 0$$
$$F_{i+1} = F_i + f_{i+1}$$

Če kumulativo frekvenc F_i izrazimo relativno, dobimo **kumulativo relativnih frekvenc** $F_i\%$. Ta predstavlja odstotek podatkov uvrščenih do zgornje meje *i*-tega razreda.

Primer

V vzorcu je 30 učencev, za vsakega učenca imamo podatek za število ur odsotnosti v preteklem šolskem letu. Podatki so:

70	54	29	73	72	47	41	43	59	97
43	52	67	42	73	84	74	60	80	71
42	69	37	64	78	63	59	72	72	69

Podatke bomo uvrstili v frekvenčno porazdelitev in jih grafično prikazali. Učinkovit pripomoček za oblikovanje in prikaz frekvenčne porazdelitve je **prikaz stebla z listi**. Vsako vrednost razdelimo na t. i. steblo in list. Steblu pripišemo 'vsebinsko pomembni del' vrednosti, listu pa ostanek. Poglejmo to na naših podatkih. Vrednosti so od 29 do 97. Vsebinsko pomembna informacija je v deseticah. Zato npr. vrednost 29 razdelimo na 2 (steblo) in 9 (list). Najprej oblikujemo steblo, ki ga zapišemo navpično; v našem primeru 2, 3, ... 9. Za vsako vrednost steblu na ustreznem mestu dodamo list. Vrednosti, ki imajo enako steblo in različne liste, zapišemo v isto vrsto in jih sproti urejamo po velikosti.

Tabela 2-20: Prikaz stebla z listi

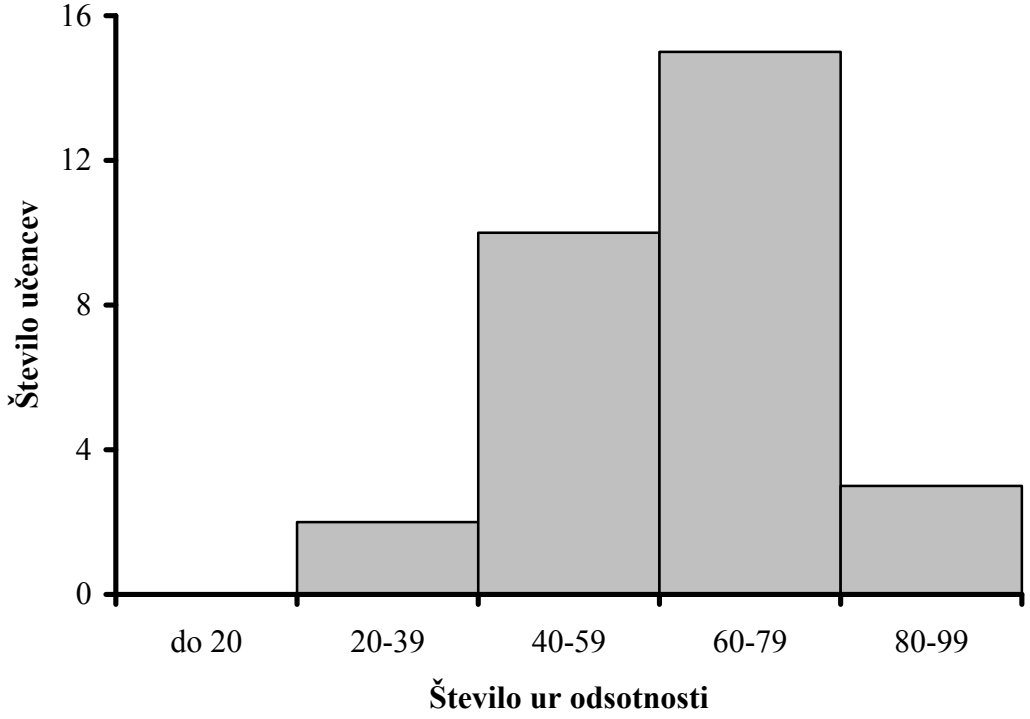
Steblo	Listi
2	9
3	7
4	122337
5	2499
6	034799
7	012223348
8	04
9	7

Ker je vzorec majhen, bomo podatke uvrstili v štiri razrede. V tabeli je frekvenčna porazdelitev dopolnjena s karakteristikami razredov.

Tabela 2-21: Učenci glede na število ur odsotnosti

Št. ur odsotnosti	$x_{i,\min}$	$x_{i,\max}$	d_i	x_i	f_i	f_i%	F_i	F_i%
20-39	19,5	39,5	20	29,5	2	6,7	2	6,7
40-59	39,5	59,5	20	49,5	10	33,3	12	40,0
60-79	59,5	79,5	20	69,5	15	50,0	27	90,0
80-99	79,5	99,5	20	89,5	3	10,0	30	100,0
Skupaj					30	100,0		

Frekvenčno porazdelitev grafično prikažemo s histogramom. **Histogram** je prikaz s stolpci, ki se držijo skupaj. Za vsak razred narišemo stolpec s širino d_i in višino f_i. Če razredi niso enako široki, je višina stolpca g_i. Na abscisni osi naj bodo opisi razredov, ne spodnje/zgornje meje.

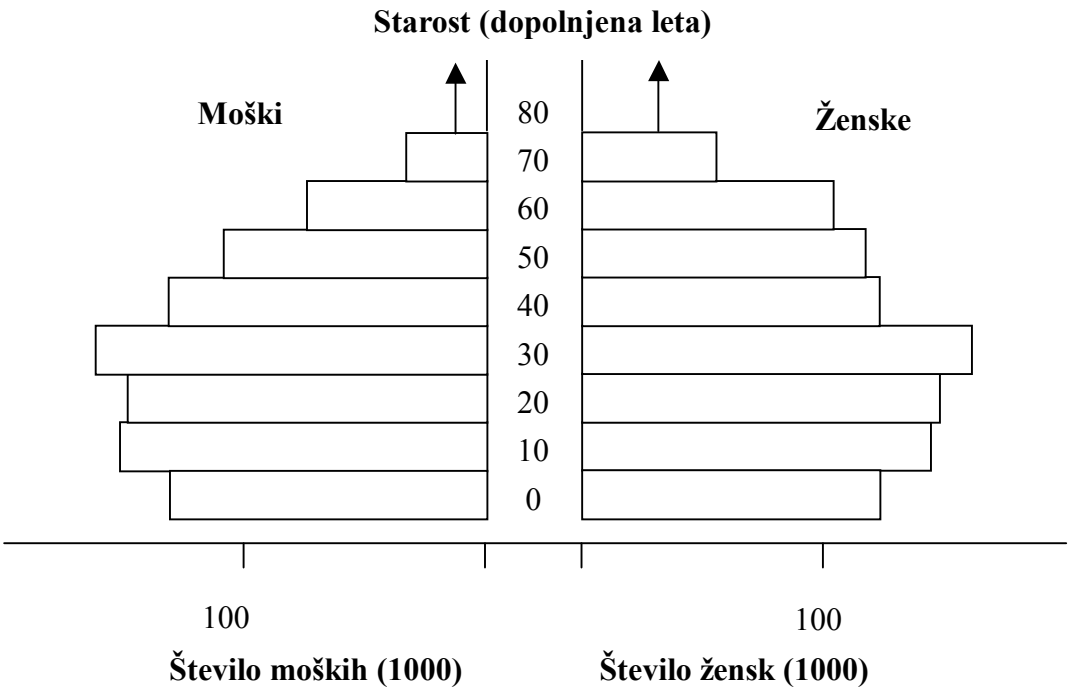


Slika 2-11: Histogram za števila ur odsotnosti za 30 učencev

Tudi prikaz stebla z listi je posebne vrste histogram. Histogram prebivalcev po starosti, narisan vertikalno posebej za ženske in posebej za moške, imenujemo **starostna piramida**. Prikazujemo starostno piramido, ki se navezuje na podatke popisa prebivalstva v Sloveniji 1991. Podatke povzemamo v naslednji tabeli.

Tabela 2-22: Prebivalci po spolu in po starosti ob popisu 1991 v Sloveniji (Vir: SL - 93, str. 49)

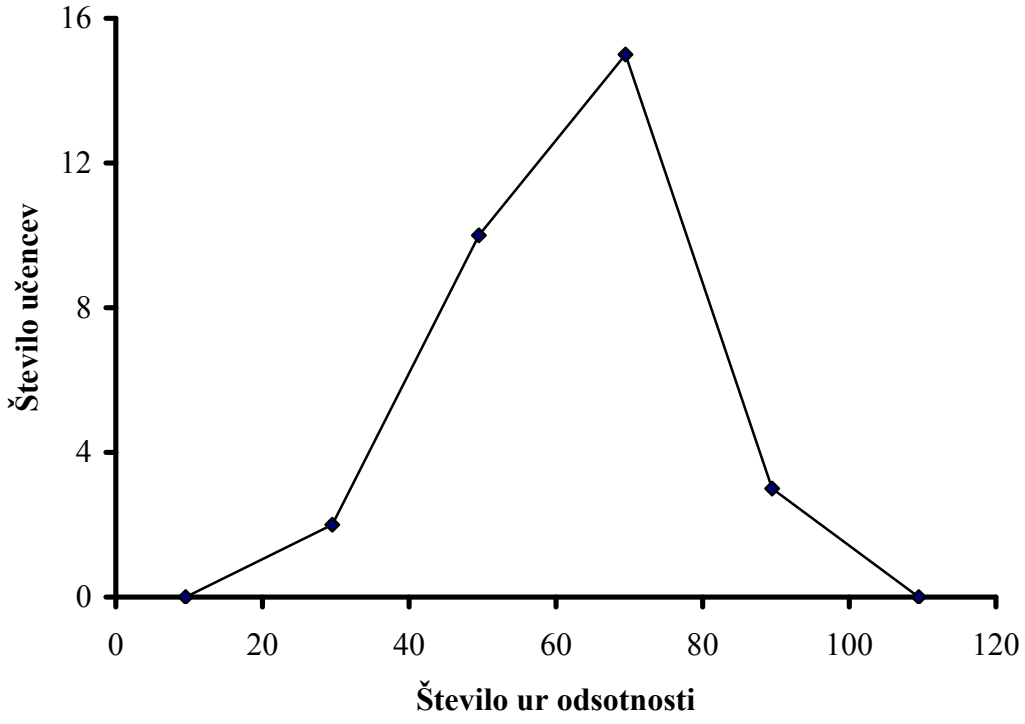
Starost (dop. leta)	Število moških	Število žensk
0 - 9	130437	123911
10 - 19	151295	143615
20 - 29	148448	148526
30 - 39	161375	155786
40 - 49	131320	126606
50 - 59	109785	116486
60 - 69	74983	109343
70 - 79	31157	57034
80 in več	13471	31718
neznano	340	350
Skupaj	952611	1013375



Slika 2-12: Starostna piramida za Slovenijo ob popisu 1991

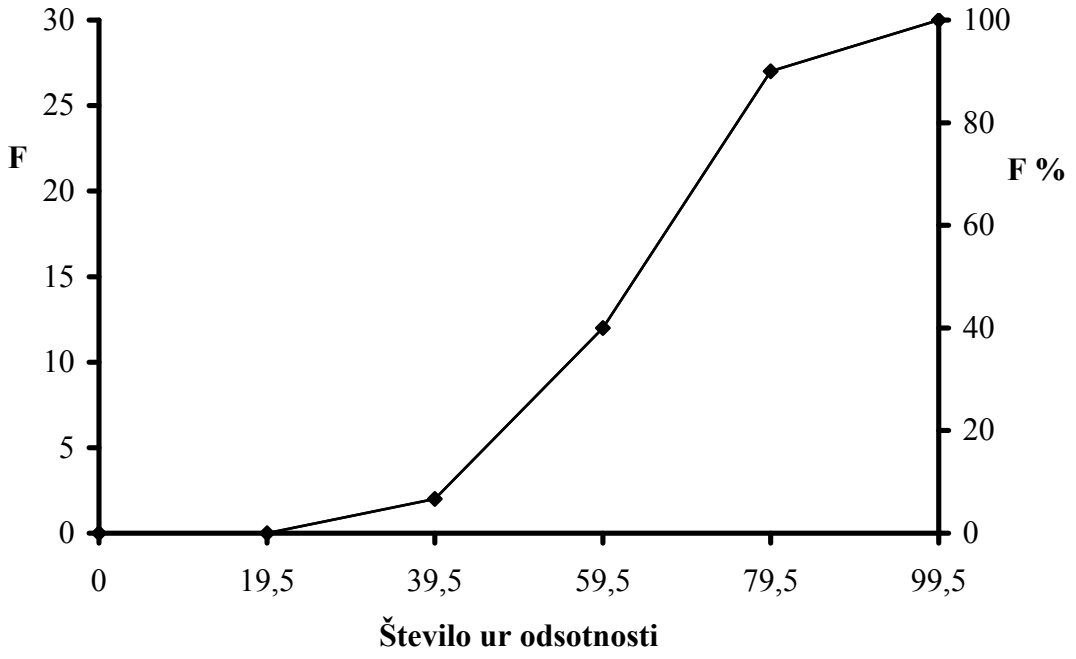
Opomba: razred na 80 let je odprt, to smo nakazali s puščico.

Alternativni prikaz histograma je poligon. To je linijski grafikon, ki povezuje točke (x_i, f_i). Dodamo dve točki: $(x_0, 0)$ in $(x_{K+1}, 0)$.



Slika 2-13: Poligon za število ur odsotnosti za 30 učencev

Za grafični prikaz kumulative frekvenc uporabljamo **ogivo**. Na abscisno os narišemo zgornje meje razredov, na ordinatno os pa pripadajoče kumulative frekvenc. Za vsak razred narišemo točko $(x_{i,\max}, F_i)$. Dodamo točko na začetku $(x_{1,\min}, 0)$. Točke povežemo z daljicami.



Slika 2-14: Kumulativa za število ur odsotnosti za 30 učencev

NALOGE

1. Razred

V razredu je učitelj za vsakega od 24 učencev ugotavljal izobrazbo očeta in matere. Vrednosti za izobrazbo so: osnovnošolska (O), srednješolska (S), višješolska ali več (V), neznano (N). Podatki so v tabeli.

Tabela 2-23: Izobrazba očeta in matere za 24 učencev

Učenec (zap. št.)	Izobrazba očeta	Izobrazba matere	Učenec (zap. št.)	Izobrazba očeta	Izobrazba matere
1	S	S	**13**	S	S
2	O	O	**14**	S	V
3	S	S	**15**	O	O
4	S	O	**16**	S	S
5	S	O	**17**	O	S
6	S	S	**18**	N	S
7	S	S	**19**	S	S
8	S	V	**20**	S	S
9	V	V	**21**	S	O
10	S	S	**22**	V	S
11	O	O	**23**	O	O
12	S	S	**24**	V	V

a) Kakšno mersko lestvico ima spremenljivka izobrazba?

b) Za vsako spremenljivko uredite vrednosti v frekvenčno porazdelitev.

c) Grafično prikažite frekvenčni porazdelitvi.

2. Starost ob diplomi

V vzorcu je bilo 35 študentov, ki so diplomirali v šolskem letu 1993/94 na Univerzi v Ljubljani. Navajamo njihovo starost ob diplomi (dopolnjena leta):

27 25 23 24 25 25 26 29 25 26
25 24 26 23 25 24 26 27 25 26
25 24 28 28 23 53 24 22 29 24
27 28 25 27 24

a) Določite, kaj je enota, spremenljivka in kakšna je njena merska lestvica.

b) Podatke uredite v frekvenčno porazdelitev in jo grafično predstavite.

3. Količina padavin

V Sloveniji je 67 meteoroloških postaj. Navajamo frekvenčno porazdelitev za količino padavin izmerjeno na teh postajah v letu 1992.

Tabela 2-24: Meteorološke postaje po količini padavin (Vir: Arhiv Hidrometeorološki zavod Slovenije)

Količina padavin (mm)	Št. postaj
800 do pod 1200	12
1200 do pod 1600	27
1600 do pod 2000	16
2000 do pod 2400	7
2400 do pod 2800	4
2800 do pod 3200	0
3200 do pod 3600	1

a) Dopolnite frekvenčno porazdelitev: za vsak razred določite spodnjo in zgornjo mejo, sredino in širino.

b) Grafično predstavite frekvenčno porazdelitev s histogramom in s poligonom.
c) Izračunajte relativno frekvenco in jo grafično prikažite.
d) Izračunajte kumulativo frekvenc in kumulativo relativnih frekvenc. Izračune grafično predstavite.
e) Grafično določite tisto količino padavin, ki je bila presežena pri polovici meteoroloških postaj.

4. Delovna doba brezposelnih

V SL - 93 so podatki o delovni dobi tistih, ki so bili v 1990 in 1992 v Sloveniji brezposelni. Število brezposelnih se nanaša na 31.12.

Tabela 2-25: Število brezposelnih v Sloveniji v letih 1990 in 1992 po dolžini delovne dobe (Vir: SL - 93, str. 148)

Dolžina delovne dobe	1990	1992
Brez delovne dobe	4175	23846
Z delovno dobo		
- **do 1 leta**	21035	16922
- **nad 1 do 2 leti**	3563	6066
- **nad 2 do 3 leta**	2654	4533
- **nad 3 do 5 let**	4188	7670
- **nad 5do 10 let**	7189	16518
- **nad 10 do 20 let**	8690	23970
- **nad 20 do 30 let**	3299	13907
- **nad 30 let**	648	4792
Skupaj	55441	118224

a) Analizirajte in grafično prikažite število brezposelnih brez delovne dobe in z delovno dobo za leti 1990 in 1992.
b) Analizirajte brezposelne z delovno dobo na dan 31. 12. 1992. Dopolnite frekvenčno porazdelitev: za vsak razred določite spodnjo mejo, zgornjo mejo, sredino razreda in širino razreda. Ali so frekvence po razredih primerljive?
c) Narišite histogram za leto 1992 za brezposelne z delovno dobo.

2.3 KVANTILI

Spremenljivka X je številska spremenljivka, njene vrednosti so: $x_1, x_2, \ldots x_n$. Vrednosti uredimo po velikosti od najmanjše do največje. Tako urejeno zaporedje imenujemo **ranžirna vrsta**. **Rang** R je zaporedno mesto vrednosti v ranžirni vrsti, $R = 1, 2, \ldots n$. Vsaki vrednosti v ranžirni vrsti priredimo njen rang.

Rang izrazimo tudi relativno, navadno ga izrazimo v deležih. **Relativni rang** označimo P. Načinov izračunavanja relativnega ranga je več. Običajno se uporablja formula:

$$P = \frac{R - 0{,}5}{n}$$

Torej je

$R = P \cdot n + 0{,}5$. Za tako izračunani relativni rang velja $0 < P < 1$. Popravek 0,5 se uporablja, ker je rang diskretna količina, relativni rang pa ne. Če je število enot n dovolj veliko, je popravek 0,5 zanemarljiv in velja

$$P \approx \frac{R}{n}$$

$$R \approx P \cdot n$$

Opomba: nekateri računajo relativni rang po formuli:

$$P = \frac{R-1}{n-1},$$

ki zagotavlja, da je $0 \le P \le 1$. Mi bomo za izračun relativnega ranga uporabljali prvi način.

Primer

Podatki: 2, 28, 1, 0, 16, 0, 7

Ranžirna vrsta	0	0	1	2	7	16	28
Rang	1	2	3	4	5	6	7
Relativni rang	$\frac{1}{14}$	$\frac{3}{14}$	$\frac{5}{14}$	$\frac{1}{2}$	$\frac{9}{14}$	$\frac{11}{14}$	$\frac{13}{14}$

Kvantil je vrednost, ki razdeli ranžirno vrsto na dva dela. Glede na to, na kolikšne dele razdeli kvantil ranžirno vrsto, ločimo mediano, kvartile, decile, centile.

Mediana *Me* razdeli ranžirno vrsto na dva enaka dela. Polovica vrednosti je manjših od mediane ali njej enakih, polovica pa večjih od mediane ali njej enakih. Relativni rang za mediano je 0,5.

$P(Me) = 0{,}5 \qquad R(Me) = 0{,}5 \cdot n + 0{,}5$

Če je izračunani rang celo število, dobimo mediano direktno iz ranžirne vrste; če izračunani rang ni celo število, je mediana povprečje dveh vrednosti na sredini ranžirne vrste.

Primer

a) 0, 0, 1, 2, 6, 12, 19 $\qquad R(Me) = 4 \qquad Me = 2$

b) 0, 0, 1, 2, 6, 12, 19, 25 $\qquad R(Me) = 4{,}5 \qquad Me = 0{,}5 \cdot (2+6) = 4$

Kvartili Q_1, Q_2, Q_3 razdelijo ranžirno vrsto na štiri enake dele.

$P(Q_1) = 0{,}25 \quad R(Q_1) = 0{,}25 \cdot n + 0{,}5$

$P(Q_2) = 0{,}50 \quad R(Q_2) = 0{,}50 \cdot n + 0{,}5$

$P(Q_3) = 0{,}75 \quad R(Q_3) = 0{,}75 \cdot n + 0{,}5$

Drugi kvartil je mediana.

Kvartilni razmik Q je razlika med tretjim in prvim kvartilom:

$Q = Q_3 - Q_1.$

Interval od Q_1 do Q_3 vsebuje 50% vrednosti.

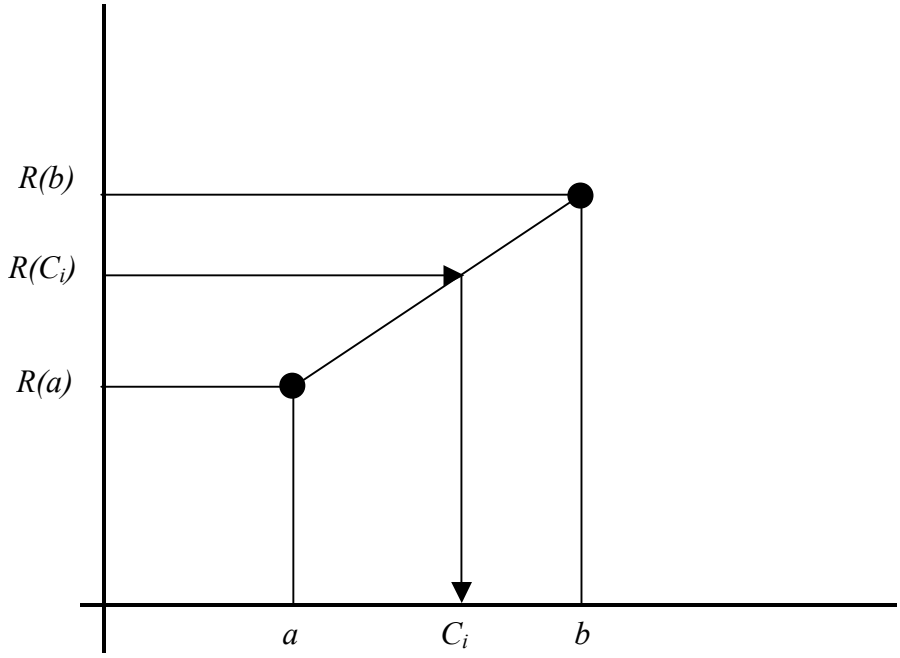
Decili $D_1, D_2, \ldots, D_9$ razdelijo ranžirno vrsto na deset enakih delov, **centili** $C_1, C_2, \ldots, C_{99}$ razdelijo ranžirno vrsto na sto enakih delov. Izračunavanje kateregakoli kvantila se da prevesti na izračunavanje pripadajočega centila, npr. $Me = C_{50}$, $D_2 = C_{20}$.

$$P(C_i) = \frac{i}{100} \qquad R(C_i) = \frac{i}{100} n + 0{,}5$$

Pri izračunavanju se uporablja **linearna interpolacija**. Poglejmo si ta princip. Radi bi izračunali C_i. Ugotovili smo, da je C_i med zaporednima vrednostma a in b v ranžirni vrsti. Pripadajoči rangi so $R(a), R(C_i), R(b)$.

Vrednosti	a	C_i	b
Rangi	$R(a)$	$R(C_i)$	$R(b)$

Grafično prikažemo situacijo takole:



Slika 2-15: Princip linearne interpolacije

Linearna interpolacija izhaja in naslednjega razmerja

$$\frac{C_i - a}{b - a} = \frac{R(C_i) - R(a)}{R(b) - R(a)}$$

Ker je $R(b) - R(a) = 1$, izračunamo C_i takole:

$$C_i = a + (R(C_i) - R(a)) \cdot (b - a)$$

Za grafični prikaz kvantilov uporabljamo linijski grafikon. Narišemo točke, ki imajo za abscise vrednosti ranžirne vrste, za ordinate pa pripadajoče range. Točke povežemo z daljicami. Za iskani kvantil izračunamo rang *R*. Narišemo vzporednico abscisni osi pri *R*. Abscisa sečišča vzporednice in poligona je grafično določena vrednost kvantila.

Primer

Podatki so: 180, 42, 10, 21, 23, 4, 62, 80. Izračunajmo drugi in šesti decil.

Ranžirna vrsta	4	10	21	23	42	62	80	180
Rang	1	2	3	4	5	6	7	8

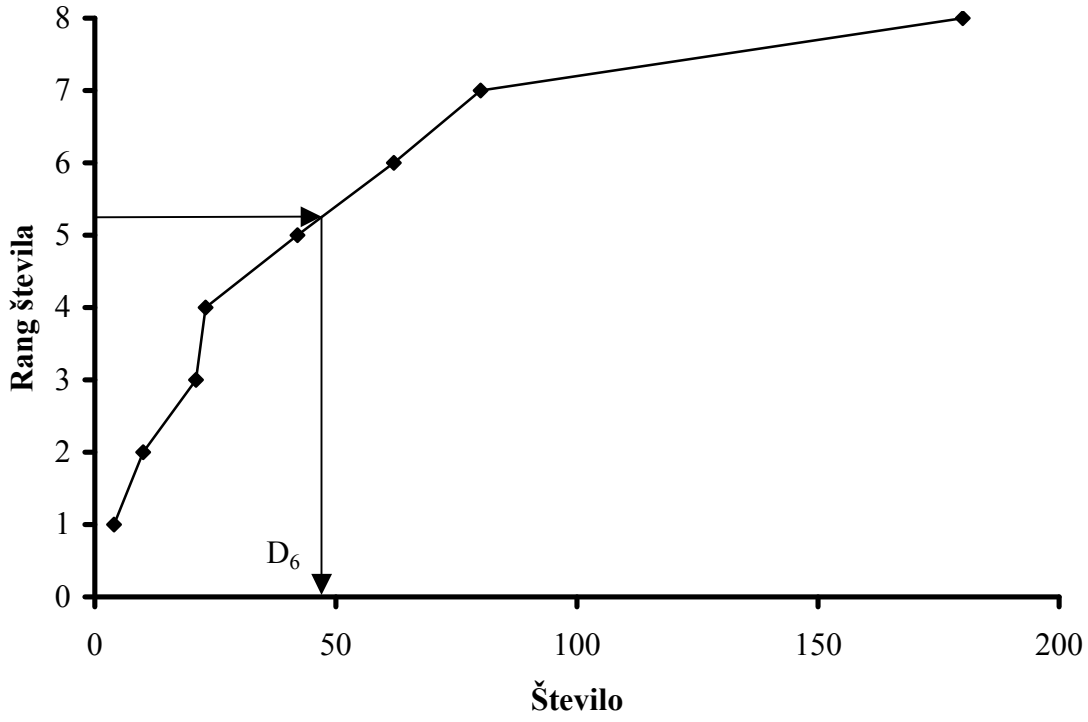
$$R(D_2) = R(C_{20}) = 0{,}20 \cdot 8 + 0{,}5 = 2{,}1$$

$$D_2 = 10 + 0{,}1 \cdot (21 - 10) = 11{,}1$$

$$R(D_6) = R(C_{60}) = 0{,}60 \cdot 8 + 0{,}5 = 5{,}3$$

$$D_6 = 42 + 0{,}3 \cdot (62 - 42) = 48$$

Za grafičen prikaz kvantilov narišemo točke: (4, 1), (10, 2), (21, 3), (23, 4), (42, 5), (62, 6), (80, 7), (180, 8) in jih povežemo z daljicami. Grafično določimo D_6. Upoštevamo, da je rang za D_6 enak 5,3. Iz slike razberemo, da je vrednost blizu 50.



Slika 2-16: Grafična določitev D_6

Če so podatki v frekvenčni porazdelitvi in osnovnih podatkov nimamo, uporabimo podoben postopek. Vlogo ranga ima kumulativa frekvenc *F*. Najprej ugotovimo, v katerem razredu je C_i. Spodnjo mejo tega razreda označimo $x_{0,min}$, zgornjo mejo pa $x_{0,max}$. Širina razreda je d_0 , frekvenca pa f_0. Približek za centil C_i dobimo z linearno interpolacijo:

$$\frac{C_i - x_{0,\min}}{x_{0,\max} - x_{0,\min}} = \frac{F(C_i) - F(x_{0,\min})}{F(x_{0,\max}) - F(x_{0,\min})}$$

Ker velja: $F(x_{0,\max}) - F(x_{0,\min}) = f_0$ in $x_{0,\max} - x_{0,\min} = d_0$, izračunamo približek C_i takole:

$$C_i = x_{0,\min} + \frac{F(C_i) - F(x_{0,\min})}{f_0} \cdot d_0$$

V tem primeru za grafično določanje kvantilov potrebujemo grafičen prikaz kumulative frekvenc, torej ogivo. Princip grafičnega določanja je enak kot v predhodnem primeru.

Primer

V tabeli imamo frekvenčno porazdelitev študentov po ocenah pri izpitu iz statistike. Izračunali bomo kvartile.

Tabela 2-26: Študenti po številu točk pri izpitu

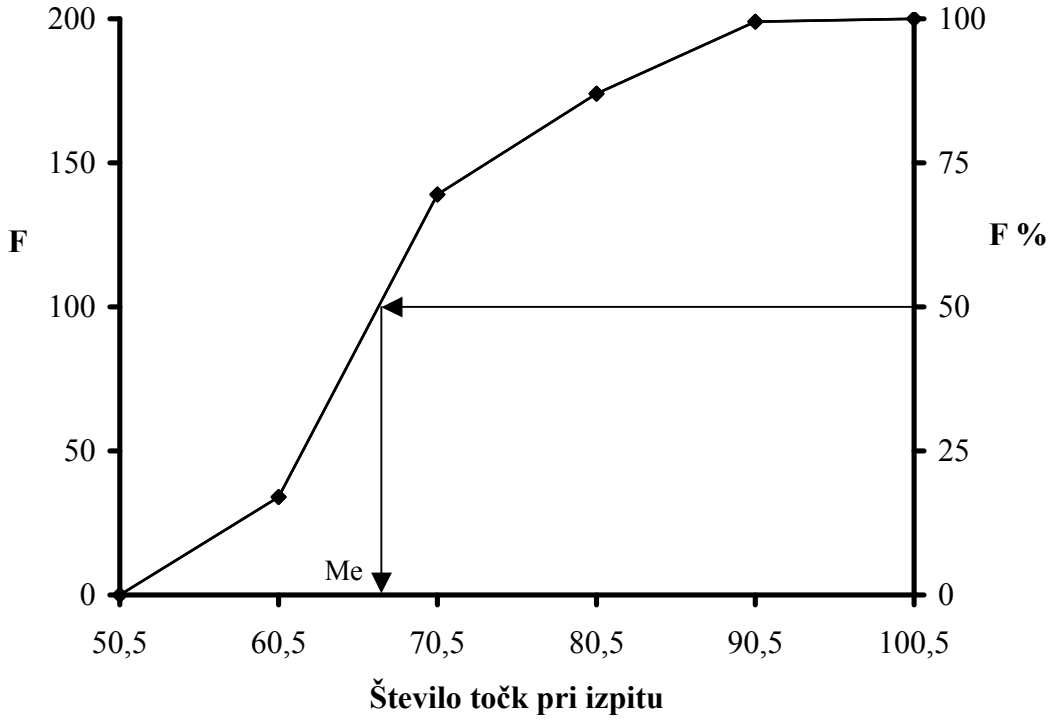
Število točk	Število študentov	Kumulativa frekvenc	Kumulativa relativnih frekvenc %	Spodnja meja	Zgornja meja	Sredina	Širina
51-60	34	34	17,0	50,5	60,5	55,5	10
61-70	105	139	69,5	60,5	70,5	65,5	10
71-80	35	174	87,0	70,5	80,5	75,5	10
81-90	25	199	99,5	80,5	90,5	85,5	10
91-100	1	200	100,0	90,5	100,5	95,5	10
Skupaj	200						

$$Q_1 = 60{,}5 + \frac{50{,}5 - 34}{105} \cdot 10 = 62{,}1$$

$$Q_2 = Me = 60{,}5 + \frac{100{,}5 - 34}{105} \cdot 10 = 66{,}8$$

$$Q_3 = 70{,}5 + \frac{150{,}5 - 139}{35} \cdot 10 = 73{,}8$$

Četrtina študentov je dosegla manj kot 62 točk, četrtina je dosegla 74 ali več točk. Mediana je 67 točk, polovica je imela slabši, polovica pa boljši uspeh. Za grafično določanje kvartilov narišemo kumulativo frekvenc oz. kumulativo relativnih frekvenc.



Slika 2-17: Kumulativa frekvenc in kumulativa relativnih frekvenc za porazdelitev študentov po uspehu pri izpitu ter grafična določitev mediane

2.3.1 Okvir z ročaji

Zelo ilustrativen grafični prikaz podatkov številske spremenljivke je **okvir z ročaji** (angl. box and whiskers plot). Ta slika prikazuje pogojni minimum in pogojni maksimum, kvartile ter osamelce. Poglejmo najprej definicije.

Osamelec je vrednost, ki bistveno odstopa od večine ostalih vrednosti. Kot osamelca opredelimo vrednost, ki je izven intervala $(Q_1 - 1{,}5 \cdot Q,\ Q_3 + 1{,}5 \cdot Q)$, pri čemer je Q kvartilni razmik. *Ekstremni osamelec* je osamelec, ki je izven intervala $(Q_1 - 3 \cdot Q,\ Q_3 + 3 \cdot Q)$.

Pogojni minimum min^* je najmanjša vrednost, ki ni spodnji osamelec**.** *Pogojni maksimum* max^* je največja vrednost, ki ni zgornji osamelec. Če v podatkih osamelcev ni, je pogojni minimum oz. pogojni maksimum najmanjša oz. največja vrednost.

Okvir z ročaji določa 5 točk: min^*, Q_1, Q_2, Q_3 ter max^*. Spodnji ročaj določata vrednosti min^* in Q_1, zgornji ročaj vrednosti Q_3 in max^*. Okvir določata kvartila Q_1 ter Q_3, njegovo prečko pa Q_2. Dolžina okvira je enaka kvartilnemu razmiku, širina okvira nima pomena. Na sliki dodamo zgornje in spodnje osamelce ter zgornje in spodnje ekstremne osamelce, ki jih predstavimo s posebnimi znaki. Običajno osamelca prikažemo s krožcem (°), ekstremnega osamelca pa z zvezdico (*).

Sliko lahko narišemo vodoravno ali navpično.

Primer

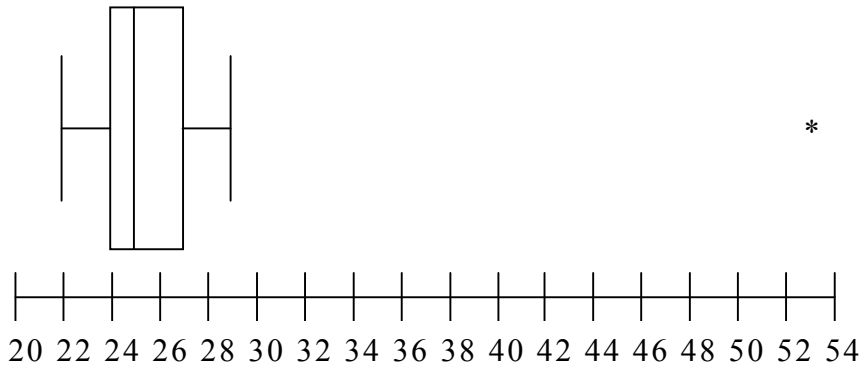
V vzorcu je bilo 35 študentov, ki so diplomirali v šolskem letu 1993/94 na Univerzi v Ljubljani. Navajamo njihovo starost ob diplomi (dopolnjena leta):

27	25	23	24	25	25	26	29	25	26
25	24	26	23	25	24	26	27	25	26
25	24	28	28	23	53	24	22	29	24
27	28	25	27	24					

Vrednosti, ki jih potrebujemo za okvir z ročaji, so:

$$\min = 22 \quad Q_1 = 24 \quad Q_2 = 25 \quad Q_3 = 27 \quad \max^* = 29 \quad \max = 53$$

Predstavljamo okvir z ročaji za starost diplomantov.



Starost ob diplomi (dopolnjena leta)

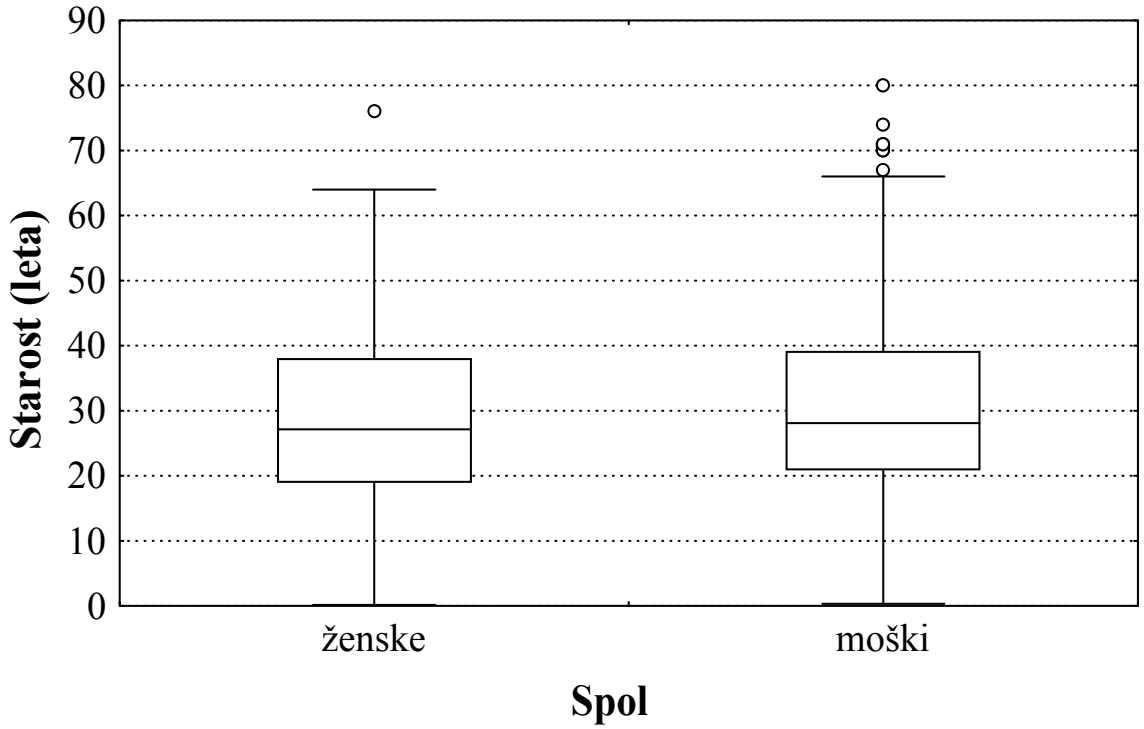
Slika 2-18: Okvir z ročaji za starost diplomantov

Iz slike lahko bralec razbere, da je bil najmlajši diplomant star 22 let, da je bila četrtina diplomantov mlajša od 24 let, četrtina pa starejša od 27 let. Polovica diplomantov je bila mlajša od 25 let, polovica pa starejša od 25 let. Starost variira od 22 do 29 let, en študent je bil ob diplomi star 53 let. Ta vrednost je zgornji osamelec.

Okvir z ročaji je zelo ilustrativen grafični prikaz, njegovo uporabo priporočamo. Še posebej koristen je v primeru, ko grafično predstavimo porazdelitev iste spremenljivke v različnih skupinah, torej ko primerjamo več okvirov na isti sliki. Tedaj vizualna primerjava omogoča, da dobimo globalno sliko o vplivu skupine na porazdelitev spremenljivke.

Primer

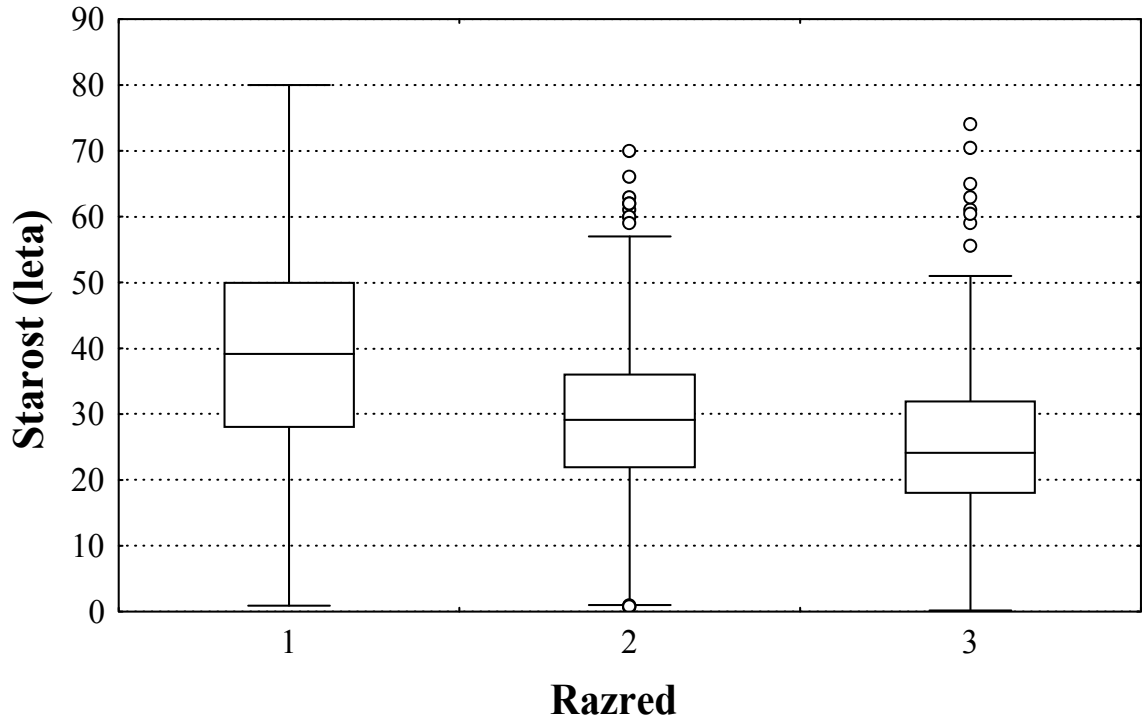
Na spletni strani Biostatistics Teaching Materials Univerze v Virginii (http://hesweb1.med.virginia.edu/biostat/s/data/titanic3.txt) so podatki za 1309 potnikov na ladji Titanic, ki je aprila 1912 potonila na svoji prvi vožnji v bližini obale Nove Fundlandije. Analizirali bomo 3 spremenljivke, ki opisujejo potnike: spol, potniški razred in starost. Spol in potniški razred sta opisni spremenljivki, spol ima imensko mersko lestvico, potniški razred pa urejenostno mersko lestvico. Na krovu ladje je bilo 466 žensk in 843 moških. V prvem, najdražjem potniškem razredu, je potovalo 323 potnikov, v drugem 277 ter v tretjem, najcenejšem, 709 potnikov. Starost je številska spremenljivka. Zanima nas porazdelitev starosti po spolu ter po potniškem razredu. Vsi podatki za starost niso bili zbrani, približno 20% je manjkajočih. Poglejmo najprej porazdelitev potnikov po starosti, posebej za ženske in za moške.



Slika 2-19: Okvir z ročaji za starost glede na spol. Podatki so za 388 žensk ter za 658 moških.

Slika kaže, da bistvenih razlik v porazdelitvi starosti glede na spol ni. Na krovu so bili tudi dojenčki. Četrtina potnikov in potnic je bila mlajša od približno 20 let, mediana za starost je 28 let, le četrtina potnikov in potnic je bila starejša od 38 let. Če izvzamemo osamelce, je največja starost za ženske 64 let in 68 let za moške. Pri ženskah je ena potnica, ki glede na starost odstopa od ostalih, njena starost je 76 let, pri moških pa je takih več, njihove starosti so od 76 do 80 let.

Poglejmo še porazdelitev starosti potnikov glede na potniški razred.



Slika 2-20: Okvir z ročaji za starost glede na potniški razred. Podatki so za 284 potnikov iz prvega razreda, 261 iz drugega razreda ter 501 iz tretjega razreda.

Slika kaže bistveno razliko v porazdelitvi starosti glede na potniški razred. Iz slike odčitamo, da je mediana v prvem razredu 40 let, v drugem 30 in v tretjem 25 let. 75% potnikov v prvem razredu je bilo mlajših od 50 let, v drugem razredu je bilo mlajših od 36 let in v tretjem razredu mlajših od 32 let. Vrednost za prvi kvartil je 30 let za potnike v prvem razredu, 22 let za potnike v drugem razredu ter 18 let za potnike v tretjem razredu. Minimalna starost je v vseh treh razredih 0 let, pogojni maksimumi pa se močno ločijo: 80 v prvem razredu (v prvem razredu ta vrednost ni osamelec), okoli 57 let v drugem razredu in 51 let v tretjem. V drugem in tretjem razredu je nekaj zgornjih osamelcev.

Sklenemo lahko, da je bila na ladji Titanic porazdelitev starosti potnikov odvisna od potniškega razreda: čim cenejši je razred, tem mlajša je bila pripadajoča populacija.

Okvir z ročaji uporabljamo za grafični prikaz porazdelitve številske spremenljivke in je alternativni prikaz histogramu. Vendar pa je med njima bistvena razlika. Histogram je odvisen od tega, kako oblikujemo razrede, zato lahko za iste podatke dobimo različne histograme. Okvir z ročaji določajo enolično opredeljene vrednosti, njegov grafični prikaz je neodvisen od analitika.

NALOGE

1. Stoli

V tržni raziskavi kvalitete stolov so v vzorec izbrali 19 stolov. Navajamo njihove prodajne cene (SIT):

11786 13200 3204 19580 3980 6800 7000 5734 10474 6281

4416 4721 7200 5781 11330 2613 3200 13750 3607

a) Izračunajte in obrazložite kvartile.

b) Ali so v podatkih osamelci?

c) Narišite okvir z ročaji.

d) Kateremu centilu pripada vrednost 10000 SIT?

2. Število točk pri maturi

Za 28 maturantov podajamo število točk, ki so jih dosegli pri maturi.

14	11	19	25	18	12	15
16	21	20	20	30	16	18
18	17	18	15	18	24	18
12	30	18	11	18	18	24

a) Izračunajte kvartile za število točk pri maturi. Obrazložite izračunane vrednosti.

b) Narišite okvir z ročaji.

3. Štipendije

Za vzorec 120 študentov so zbrali podatke o znesku štipendije, ki so jo prejeli v februarju 1996.

Tabela 2-27: Študenti po višini štipendije v februarju 1996

Višina štipendije (v tisoč SIT)	Število študentov
nad 4 do 6	9
nad 6 do 8	15
nad 8 do 10	21
nad 10 do 12	34
nad 12 do 14	24
nad 14 do 16	12
nad 16 do 18	5
Skupaj	120

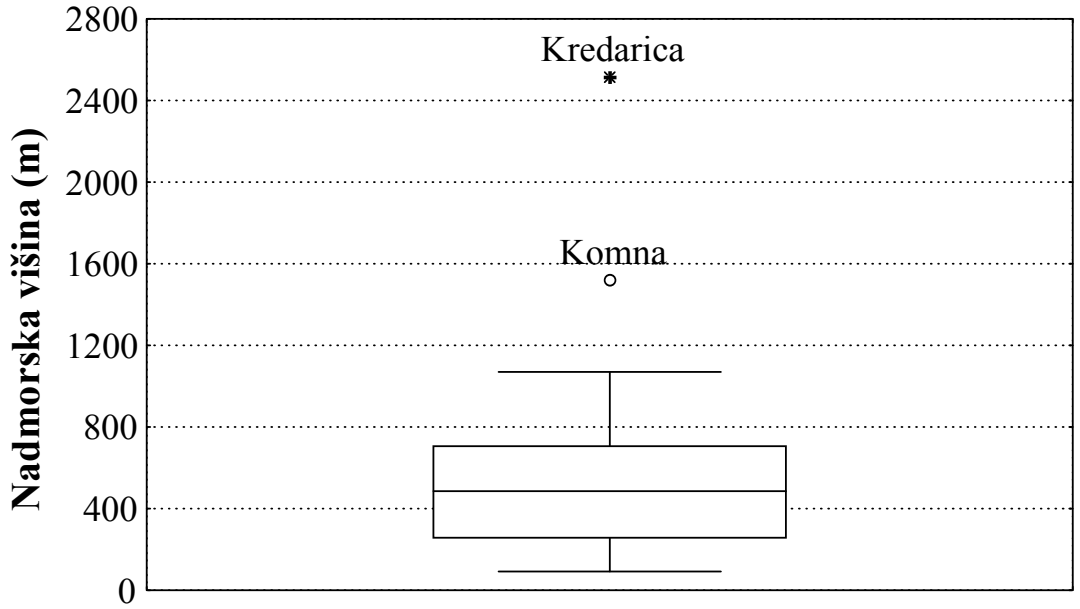
a) Izračunajte in grafično prikažite kvartile za višino štipendije.

b) Izračunajte in grafično določite odstotek študentov, ki so prejeli več kot 9000 SIT.

c) Kolikšna je višina štipendije, ki je bila presežena pri 10% študentov? To vrednost določite tudi grafično. Kako imenujemo to vrednost?

4. Nadmorska višina meteoroloških postaj

Na sliki je okvir z ročaji za nadmorsko višino 67 meteoroloških postaj v Sloveniji v letu 1992.

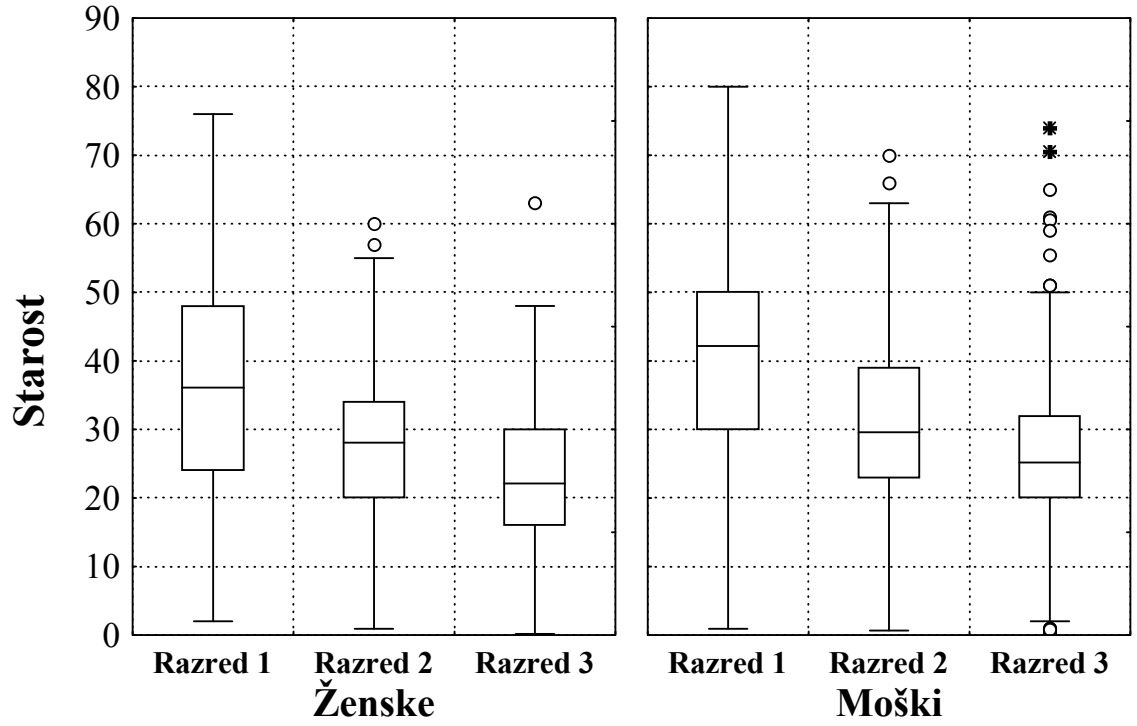


Slika 2-21: Okvir z ročaji za nadmorsko višino 67 meteoroloških postaj v Sloveniji leta 1992

Obrazložite sliko.

5. Titanic

Predstavljamo grafični prikaz porazdelitev starosti po spolu in potniškem razredu. Na sliki je 6 okvirov z ročaji.



Slika 2-22: Okviri z ročaji za starost glede na spol in potniški razred

Obrazložite sliko.

2.4 MERE SREDINE: SREDNJE VREDNOSTI

Vrednosti spremenljivke želimo predstaviti z njenimi predstavniki, da bi dobili bolj jedrnato in pregledno predstavo o spremenljivki. Najpogosteje uporabljeni predstavniki spremenljivke so njene **srednje vrednosti**, ki so **mere sredine** spremenljivke.

Ker lahko sredino spremenljivke vrednotimo na različne načine, je srednjih vrednosti več. Srednje vrednosti so: **modus**, **mediana** in **povprečje**. Katera srednja vrednost je za določeno spremenljivko primerna, je odvisno od njene merske lestvice.

2.4.1 Modus

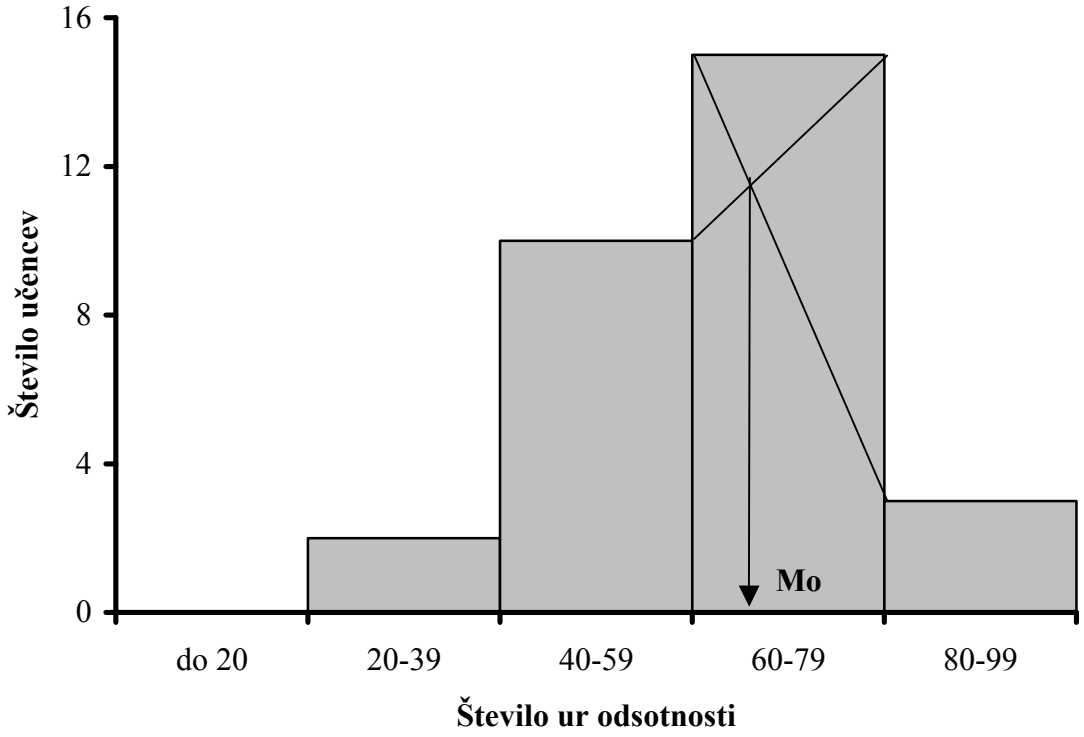
Modus *Mo* je najpogostejša vrednost spremenljivke, torej vrednost z največjo frekvenco. Modusov je lahko več. Če se vse vrednosti enako mnogokrat ponovijo, modusa ne določimo.

Primer

Za 11 družin imamo podatke o številu otrok v družini. Podatki so: 1, 2, 2, 2, 2, 3, 1, 2, 2, 5, 1 otrok. Da dobimo modus, vrednosti uvrstimo v frekvenčno porazdelitev in poiščemo vrednost, ki nastopa največkrat. Ker je število 2 v podatkih največkrat, je $Mo = 2$ otroka. Največ opazovanih družin ima po dva otroka.

Modus je primerna srednja vrednost za vsako spremenljivko, ne glede na mersko lestvico. Npr. modus določimo za kraj rojstva (imenska spremenljivka), za izobrazbo (urejenostna spremenljivka), za temperaturo (razmična spremenljivka), za dolžino delovne dobe (razmernostna spremenljivka). Poudariti pa velja, da je modus edina srednja vrednost, ki je primerna za imenske spremenljivke.

Če imamo vrednosti spremenljivke uvrščene v frekvenčno porazdelitev z enako širokimi razredi, osnovnih podatkov pa nimamo, določimo *modusni razred*. Modusni razred je razred z največjo frekvenco. Modusnih razredov je lahko več. Če je modusni razred en sam, na histogramu grafično določimo točkovni približek za modus. Za grafični približek potrebujemo modusni razred in njegovega levega in desnega soseda. Narišemo dve daljici, tako kakor je razvidno iz slike. Abscisa sečišča daljic je grafični približek za modus.



Slika 2-23: Histogram za učence glede na število ur odsotnosti ter grafična določitev modusa

Približek za modus lahko tudi izračunamo. Najprej določimo modusni razred, njegove karakteristike označimo z indeksom 0. Uporabili bomo naslednje oznake:

f_0 frekvenca modusnega razreda

f_{-1} frekvenca njegovega levega soseda

f_{+1} frekvenca njegovega desnega soseda

d širina razredov

$x_{0,\min}$ spodnja meja modusnega razreda

Točkovni približek za modus izračunamo po formuli:

$$Mo = x_{0,\min} + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_{+1}} \cdot d$$

Primer

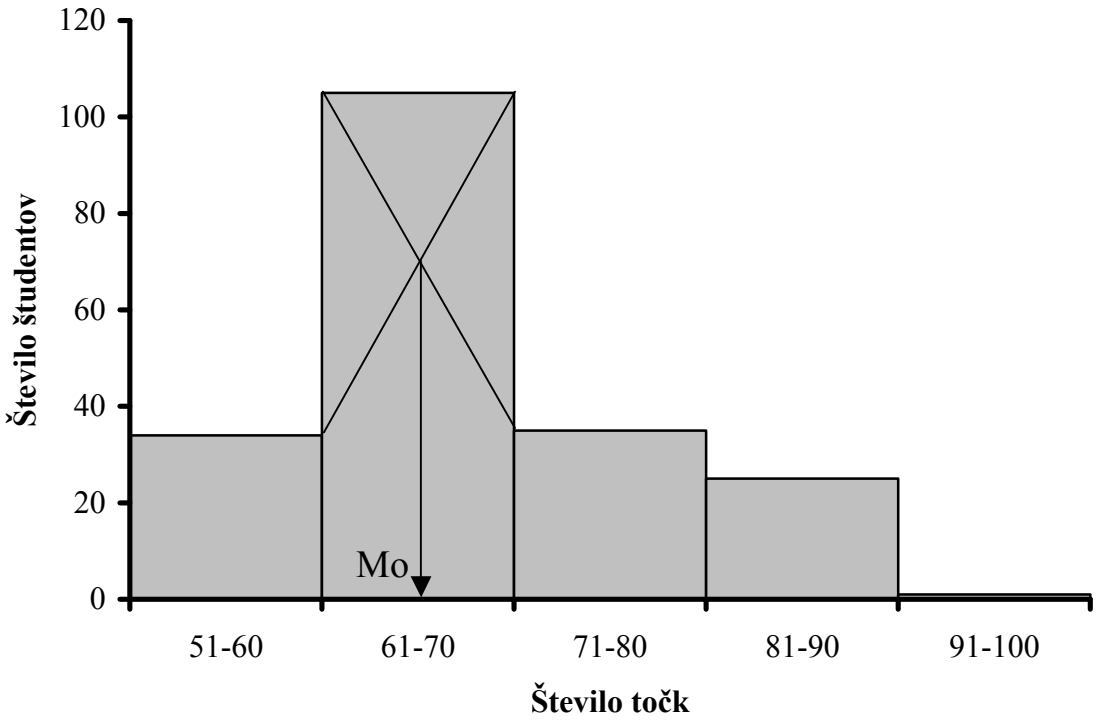
V tabeli imamo frekvenčno porazdelitev študentov po ocenah pri izpitu iz statistike. Radi bi določili število točk, ki je bilo pri izpitu najpogostejše.

Tabela 2-28: Študenti po številu točk pri izpitu

Število točk	Število študentov	Sp. meja	Zg. meja	Sredina	Širina
51-60	34	50,5	60,5	55,5	10
61-70	105	60,5	70,5	65,5	10
71-80	35	70,5	80,5	75,5	10
81-90	25	80,5	90,5	85,5	10
91-100	1	90,5	100,5	95,5	10
Skupaj	200				

$$Mo = 60{,}5 + \frac{105 - 34}{2 \cdot 105 - 34 - 35} \cdot 10 = 65{,}5$$

Najpogostejša ocena pri izpitu je 66 točk. Grafično določimo približek za modus takole:



Slika 2-24: Histogram za študente glede na število točk pri izpitu in grafična določitev modusa

2.4.2 Mediana

Mediana *Me* deli ranžirno vrsto na polovico. Izračun in grafično določanje mediane smo spoznali v poglavju Kvantili. Na vrednost mediane vplivajo samo vrednosti v sredini ranžirne vrste.

Mediana je primerna srednja vrednost za spremenljivke, ki imajo vsaj urejenostno mersko lestvico. Npr. izobrazba, število otrok v družini, dolžina delovne dobe. Izračun mediane iz frekvenčne porazdelitve smo obravnavali v poglavju Kvantili.

2.4.3 Povprečje

Povprečje se računa samo za številske spremenljivke. Povprečje je vrednost, za katero velja: če bi bili vsi podatki enaki, bi bili enaki povprečju. Pomen povprečja bomo ilustrirali na primeru povprečne plače: povprečna plača je plača, ki bi jo prejel zaposleni v primeru, če bi denar za plače razdelili enakomerno med zaposlene, torej bi vsi imeli enako plačo.

Povprečje ima predvsem analitični pomen. Uporabljamo ga pri primerjavah, pri študiju dinamike pojava: npr. povprečna plača po državah, povprečno letno število brezposelnih ip.

Za izračun povprečja uporabljamo tri načine izračunavanja:

- aritmetično sredino,
- geometrijsko sredino,
- harmonično sredino.

Način izračunavanja povprečja je odvisen od vrste podatkov. Povprečna plača, povprečni verižni indeks, povprečna hitrost se izračunajo na različne načine.

2.4.3.1 Aritmetična sredina

Standardna oznaka za aritmetično sredino podatkov $x_1, x_2, \ldots x_n$ je $\bar{x}$. **Aritmetična sredina** leži med vrednostmi $x_1, x_2, \ldots x_n$. Vsaka posamezna vrednost x_i se od $\bar{x}$ odklanja navzgor ali navzdol: odklon $x_i - \bar{x}$ je pozitiven ali negativen. Aritmetična sredina je postavljena tako, da je vsota odklonov enaka 0:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Iz tega izraza dobimo izraz za aritmetično sredino:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Aritmetično sredino izračunamo tako, da seštejemo vse vrednosti spremenljivke in vsoto delimo s številom podatkov.

Izračunana aritmetična sredina navadno ni v zalogi vrednosti spremenljivke. Še posebej to velja za diskretne spremenljivke. Npr. povprečno število otrok za že omenjeni primer enajstih družin je 2,1 otroka.

Aritmetična sredina je najpogosteje uporabljena srednja vrednost. Na vrednost aritmetične sredine vplivajo vse vrednosti, kar za mediano in za modus ne drži. Izjemno velike oz. majhne vrednosti, torej osamelci, močno vplivajo na vrednost aritmetične sredine. Če so v podatkih osamelci, priporočamo, da se aritmetična sredina izračuna z osamelci in brez njih.

Primer

Navajamo starost 35 diplomantov

27	25	23	24	25	25	26	29	25	26	25	24
26	23	25	24	26	27	25	26	25	24	28	28
23	53	24	22	29	24	27	28	25	27	24	

Vrednost 53 je osamelec.

a) če upoštevamo vse podatke, je povprečna starost 26,2 leta;
b) če izločimo podatek za študenta, ki je imel ob diplomi 53 let, je povprečna starost 25,4 leta.

Včasih je smiselno, da imajo vrednosti $x_1, x_2, \ldots x_n$ različen vpliv pri izračunu povprečja. Vsaka vrednost ima svojo *utež* p_i, $i = 1, 2, \ldots, n$. Če upoštevamo uteži, se aritmetična sredina izračuna takole:

$$\bar{x} = \frac{1}{\sum_{i=1}^{n} p_i} \cdot \sum_{i=1}^{n} p_i x_i$$

To sredino imenujemo *tehtana aritmetična sredina*. Če imajo vse vrednosti enake uteži, je tehtana aritmetična sredina enaka navadni aritmetični sredini. Poglejmo primer uporabe tehtane aritmetične sredine.

Primer

Potnik je od kraja A do B potoval eno uro s hitrostjo 80 km/h, od kraja B do C pa dve uri s hitrostjo 90 km/h. Kolikšna je povprečna hitrost na celotni poti?

Povprečna hitrost je razmerje celotne poti s in celotnega časa t:

$$\bar{v} = \frac{s}{t} = \frac{1 \cdot 80 + 2 \cdot 90}{1 + 2} \frac{\text{km}}{\text{h}} = 86{,}7 \text{ km/h}$$

Povprečno hitrost smo izračunali kot tehtano aritmetično sredino hitrosti, uteži so časi.

Tehtano aritmetično sredino uporabljamo za izračun aritmetične sredine vrednosti, ki so uvrščene v frekvenčno porazdelitev. Za vsak razred določimo njegovega predstavnika, spoznali smo že, da je to pripadajoča sredina razreda x_i. Ker ne poznamo posamičnih vrednosti v razredu, predpostavimo, da so vse vrednosti v razredu enake sredini razreda; le-to v vsakem razredu upoštevamo f_i - krat. Torej je prispevek posamičnega razreda k vsoti vrednosti $f_i x_i$, te prispevke seštejemo po vseh razredih. Aritmetično sredino izračunamo takole:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{K} f_i x_i = \frac{1}{\sum_{i=1}^{K} f_i} \sum_{i=1}^{K} f_i x_i$$

Aritmetično sredino iz frekvenčne porazdelitve izračunamo kot tehtano aritmetično sredino sredin razredov, pripadajoče uteži so frekvence razredov.

Primer

Izračunajmo povprečno število točk pri izpitu iz statistike za primer 200 študentov:

Tabela 2-29: Študenti po številu točk pri izpitu

Število točk	Število študentov	Sp. meja	Zg. meja	Sredina	Širina
51-60	34	50,5	60,5	55,5	10
61-70	105	60,5	70,5	65,5	10
71-80	35	70,5	80,5	75,5	10
81-90	25	80,5	90,5	85,5	10
91-100	1	90,5	100,5	95,5	10
Skupaj	200				

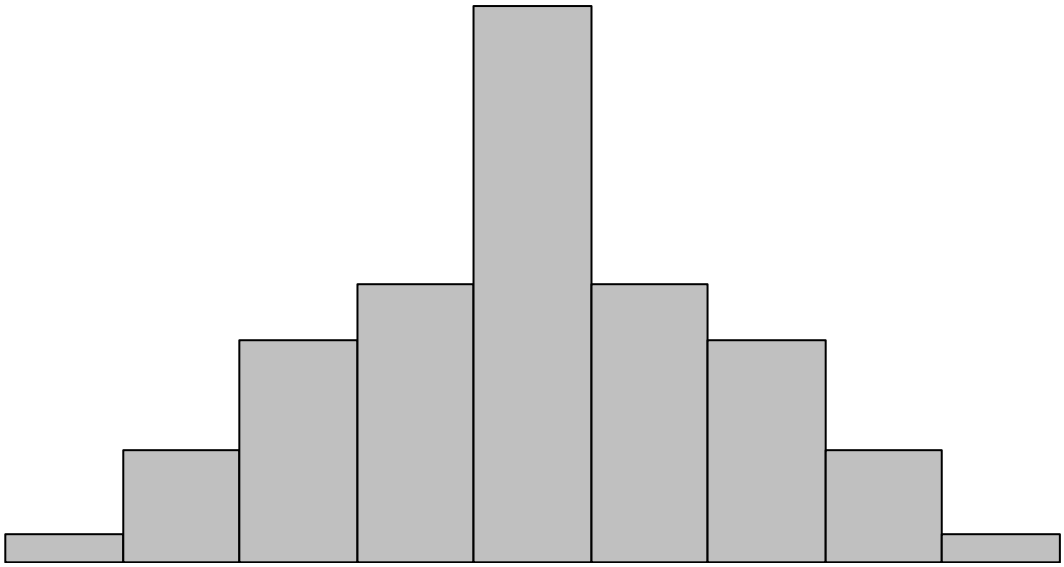
$$\bar{x} = \frac{1}{200}(34 \cdot 55{,}5 + 105 \cdot 65{,}5 + \ldots + 1 \cdot 95{,}5) = 68{,}2$$

Povprečen uspeh pri izpitu je 68 točk. Če bi imeli vsi študenti enak uspeh pri izpitu, bi le-ta bil 68 točk.

Pri izračunu aritmetične sredine iz frekvenčne porazdelitve delajo probleme odprti razredi, saj nimajo sredine. Edina možna pot za izračun približka aritmetične sredine je, da za odprte razrede sami določimo vsebinsko smiselno sredino razreda. Npr. v frekvenčni porazdelitvi oseb po starosti je pogosto zadnji razred odprt, npr. 100 in več let. Sami moramo presoditi, kaj je ustrezna sredina takega razreda.

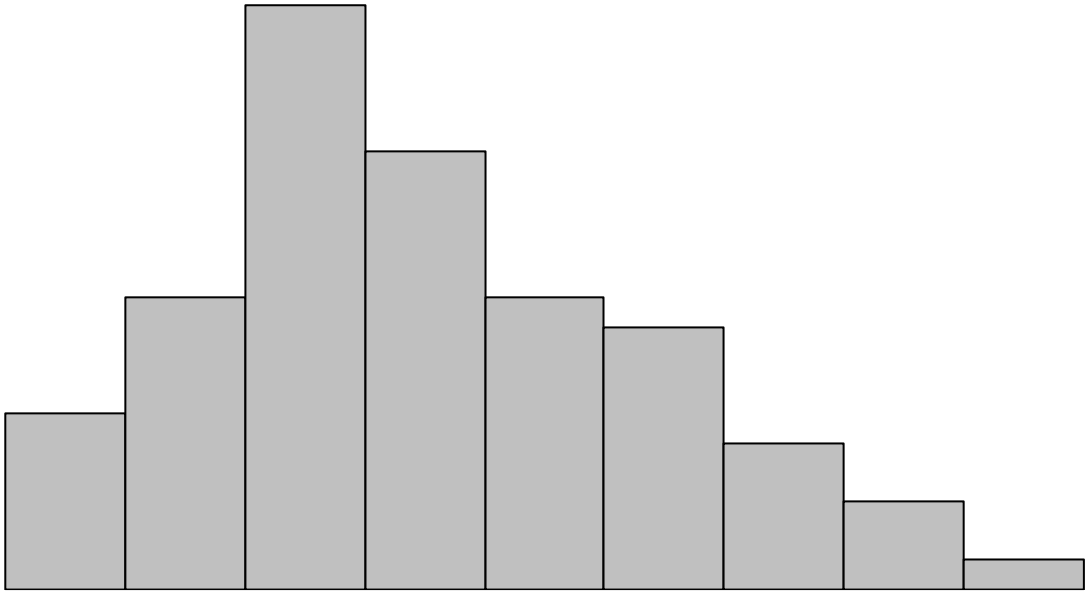
Frekvenčne porazdelitve delimo glede na obliko na simetrične in asimetrične. Oblika odraža odnos med mediano, modusom in aritmetično sredino:

- *simetrična porazdelitev* z enim modusnim razredom: $\bar{x} \approx Me \approx Mo$



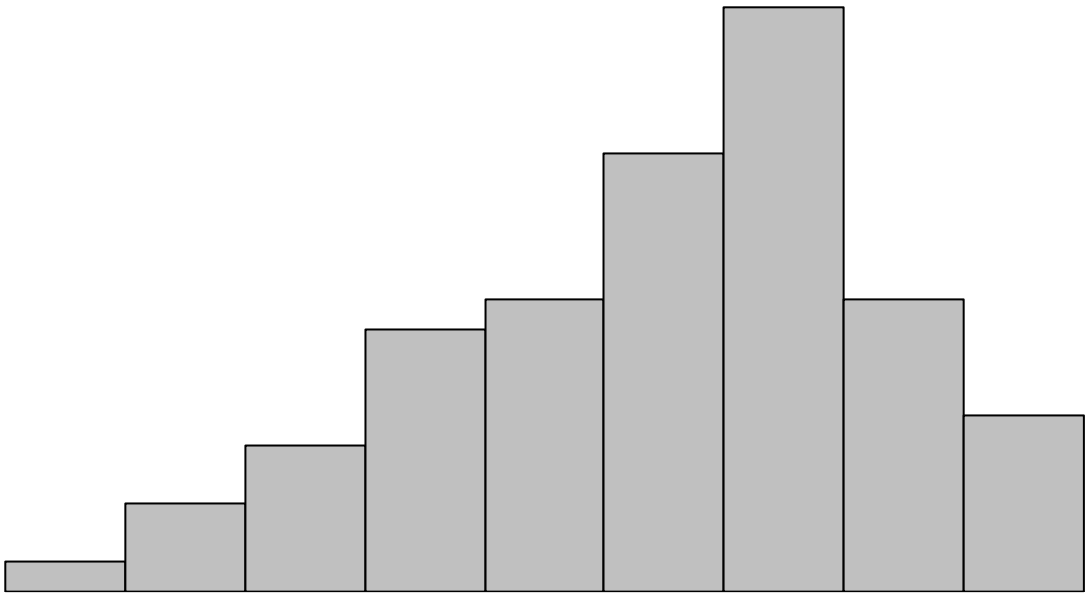
Slika 2-25: Simetrična frekvenčna porazdelitev

- porazdelitev je *asimetrična v desno*: $Mo < Me < \bar{x}$



Slika 2-26: Frekvenčna porazdelitev asimetrična v desno

- porazdelitev je *asimetrična v levo*: $\bar{x} < Me < Mo$



Slika 2-27: Frekvenčna porazdelitev asimetrična v levo

2.4.3.2 Harmonična sredina

Posebne vrste sredina je **harmonična sredina**. Opredeljena je takole: harmonična sredina vrednosti $x_1, x_2, \ldots x_n$ je recipročna vrednost aritmetične sredine vrednosti $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$.

$$H = (\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i})^{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Veljati mora $x_i \neq 0,\ i = 1, 2, ..., n$. Podobno kot pri aritmetični sredini imamo tudi pri harmonični sredini *tehtano harmonično sredino*:

$$H = (\frac{1}{\sum_{i=1}^{n} p_i} \cdot \sum_{i=1}^{n} \frac{p_i}{x_i})^{-1} = \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} \frac{p_i}{x_i}}$$

Najpogosteje jo uporabljamo za izračun povprečja koeficientov, njeno uporabo narekuje osnovna opredelitev koeficienta. Poglejmo primere uporabe harmonične sredine na primeru povprečne hitrosti.

Primer

Od kraja A do kraja B je 100 km, od B do C pa 240 km. Potnik je potoval od A do B s hitrostjo 80 km/h, od B do C pa s hitrostjo 90 km/h. Kolikšna je povprečna hitrost na celotni poti?

$$\bar{v} = \frac{s}{t} = \frac{100 + 240}{\frac{100}{80} + \frac{240}{90}} \frac{\text{km}}{\text{h}} = 86{,}8 \text{ km/h}$$

Povprečno hitrost smo izračunali kot tehtano harmonično sredino hitrosti, uteži so poti.

2.4.3.3 Geometrijska sredina

Geometrijska sredina vrednosti $x_1, x_2, \ldots x_n$ je n-ti koren iz produkta teh vrednosti:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Za izračun geometrijske sredine mora veljati: $x_i > 0,\ i = 1, 2, ..., n$. Tudi geometrijska sredina je v sorodu z aritmetično sredino, saj lahko formulo za geometrijsko sredino zapišemo takole:

$$\log G = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

Logaritem geometrijske sredine je enak aritmetični sredini logaritmiranih vrednosti. Geometrijsko sredino uporabljamo za izračun povprečnega koeficienta rasti, povprečnega verižnega indeksa ter povprečne stopnje rasti. Poglejmo, zakaj.

Časovna vrsta $y_1, y_2, \ldots, y_T$ je ekvidistantna, dolžina časovne vrste je T. Pri analizi časovnih vrst pride prav t. i. **koeficient rasti**, ki je razmerje dveh zaporednih podatkov v časovni vrsti:

$$K_{t/t-1} = \frac{y_t}{y_{t-1}}, \quad t = 2, 3, ..., T$$

Če koeficient rasti množimo s 100, dobimo pripadajoči verižni indeks.

Izračunajmo **povprečni koeficient rasti**. Izhajamo iz naslednje zveze:

$$y_T = y_1 \cdot K_{2/1} \cdot K_{3/2} \cdot ... \cdot K_{T/T-1}$$

Če bi bili vsi koeficienti rasti enaki, bi bili enaki povprečnemu koeficientu rasti $\overline{K}$, torej velja:

$$y_T = y_1 \cdot K_{2/1} \cdot K_{3/2} \cdot ... \cdot K_{T/T-1} = y_1 \cdot \overline{K}^{T-1}$$

Iz tega izraza dobimo dva načina izračunavanja povprečnega koeficienta rasti:

$$\overline{K} = \sqrt[T-1]{K_{2/1} \cdot K_{3/2} \cdot ... \cdot K_{T/T-1}} = \sqrt[T-1]{\frac{y_T}{y_1}}$$

Iz zgornjega sledi, da je **povprečni verižni indeks** $\bar{I}$ geometrijska sredina pripadajočih verižnih indeksov. Če imamo podatke o časovni vrsti, je boljša pot za izračun povprečnega verižnega indeksa formula, ki upošteva prvi in zadnji podatek časovne vrste ter njeno dolžino:

$$\bar{I} = \sqrt[T-1]{I_{2/1} \cdot I_{3/2} \cdot ... \cdot I_{T/T-1}} = \sqrt[T-1]{\frac{y_T}{y_1}} \cdot 100$$

Povprečna stopnja rasti $\overline{S}$ se izračuna iz povprečnega verižnega indeksa:

$$\overline{S} = \bar{I} - 100$$

Primer

V tabeli so verižni indeksi za pridelavo jabolk v EGS po letih v obdobju 1980-1984, osnovnih podatkov o pridelavi jabolk nimamo. Želimo izračunati povprečno stopnjo rasti za pridelavo jabolk v tem obdobju.

Tabela 2-30: Verižni indeksi za pridelavo jabolk v EGS

Leto	Verižni indeks
1980	–
1981	70
1982	170
1983	76
1984	108

$$\bar{I} = \sqrt[4]{70 \cdot 170 \cdot 76 \cdot 108} = 99{,}4$$

$$\overline{S} = -0{,}6$$

Povprečna stopnja rasti je –0,6%. V obdobju 1980-1984 se pridelava jabolk v povprečju ni spreminjala.

Primer

V tabeli je podano število brezposelnih v Sloveniji v letih od 1990 do 1994 (Vir: SL-95, str. 186) in izračunani verižni indeksi ter stopnje rasti.

Tabela 2-31: Število brezposelnih v Sloveniji v letih od 1990 do 1994 (Vir: SL-95, str. 186) in izračunani verižni indeksi ter stopnje rasti

Leto	Število brezposelnih (v tisoč)	Verižni indeks	Stopnja rasti
1990	44,2	–	–
1991	75,0	169,7	69,7
1992	102,6	136,8	36,8
1993	129,1	125,8	25,8
1994	128,1	99,2	-0,8

Izračunali bomo povprečno stopnjo rasti na dva načina:

a) iz verižnih indeksov $\overline{S} = \sqrt[4]{169,7 \cdot 136,8 \cdot 125,8 \cdot 99,2} - 100 = 30,5$

b) iz podatkov časovne vrste $\overline{S} = \sqrt[4]{\frac{128,1}{44,2}} \cdot 100 - 100 = 30,5$

V opisanem obdobju je stopnja rasti v povprečju naraščala za 30,5% letno. Če bi bilo v obravnavanem obdobju naraščanje enakomerno, bi število brezposelnih naraščalo za 30,5% letno.

Pri izračunu povprečja spremenljivke moramo upoštevati ustrezno sredino. Izrek opisuje odnos med aritmetično, harmonično in geometrijsko sredino. Navajamo ga brez dokaza.

Izrek:

$H \le G \le \overline{x}$

Premislite, kdaj so vse tri sredine enake.

NALOGE

1. Katere srednje vrednosti?

Za spodaj navedene spremenljivke napišite v tabelo mersko lestvico in srednje vrednosti (modus, mediana, povprečje), ki so zanje primerne.

Spremenljivka	Merska lestvica	Srednje vrednosti
Spol		
Datum rojstva		
Temperatura zraka (^{0}C)		
Plača (SIT)		
Dolžina delovne dobe (število dni)		
Gostota (kg/m^3)		

2. Tri sredine

Za podatke 4, 5, 2, 6, 3, 10, 8 izračunajte aritmetično, harmonično in geometrijsko sredino.

3. Število točk pri maturi

Izračunajte in obrazložite srednje vrednosti za število točk pri maturi. Podatki so:

14	11	19	25	18	12	15
16	21	20	20	30	16	18
18	17	18	15	18	24	18
12	30	18	11	18	18	24

4. Starost delavcev

V neki organizaciji je bilo januarja 1995 anketiranih 30 delavcev. Njihove starosti (v dopolnjenih letih) so:

39	33	20	45	56	63	32	25	18	66
21	45	54	59	43	50	43	21	60	45
20	45	21	62	56	29	58	33	27	21

a) Izračunajte povprečje in mediano za starost anketiranih. Obrazložite izračunani vrednosti.
b) Podatke uredite v frekvenčno porazdelitev.
c) Izračunajte povprečje in mediano iz frekvenčne porazdelitve.
d) Primerjajte vrednosti, dobljene pod a) in c).
e) Grafično določite modus iz frekvenčne porazdelitve.
f) Grafično določite mediano iz frekvenčne porazdelitve.

5. Vpis študentov v Sloveniji

Tabela 2-32: Število študentov, vpisanih na visokih šolah, fakultetah in umetniških akademijah v Sloveniji v šolskih letih 1986/87 do 1995/96 (Vir: SL-97, str. 118)

Šolsko leto	Število vpisanih (v tisoč)
1986/87	26,5
1987/88	27,8
1988/89	27,6
1989/90	30,6
1990/91	30,3
1991/92	33,7
1992/93	35,2
1993/94	38,4
1994/95	40,6
1995/96	46,0

Izračunajte povprečno stopnjo rasti za število vpisanih študentov na dva načina.

6. Poroke

V tabeli je število porok v Sloveniji po letih v obdobju 1991 – 1998.

Tabela 2-33: Število porok v Sloveniji po letih 1991 do 1998 (Vir SL-99, stran 95)

Leto	Število porok
1991	8173
1992	9119
1993	9022
1994	8314
1995	8245
1996	7555
1997	7500
1998	7528

a) Izračunajte verižne indekse in stopnje rasti. Rezultate grafično prikažite.
b) Izračunajte povprečno stopnjo rasti za število porok in jo obrazložite

7. Rast bakterij

Na začetku poskusa je bilo 1000 bakterij, ob koncu tretjega dne poskusa pa 4000 bakterij. Kolikšna je bila povprečna dnevna stopnja rasti za število bakterij v tem poskusu?

2.5 MERE VARIABILNOSTI

Srednja vrednost je predstavnik vseh vrednosti spremenljivke, vendar je njena informativnost sorazmerno majhna. Npr. podatek, da je bila v avgustu 1998 povprečna neto plača v Sloveniji 99 488 SIT, pove nekaj o srednji vrednosti plač, nič pa o tem, kako so plače variirale. Zanimivo bi bilo vedeti, za koliko največja plača presega najmanjšo plačo, je variiranje plač v avgustu enako kot mesec prej ipd. Take in podobne količine opisujejo **mere variabilnosti**, ki dopolnjujejo mere sredine.

Tako kot sredino spremenljivke vrednotimo na različne načine, tudi variabilnost spremenljivke vrednotimo z različnih izhodišč. Predstavljamo mere variabilnosti samo za številske spremenljivke. Ločimo *absolutne* in *relativne mere variabilnosti*.

2.5.1 Absolutne mere variabilnosti

Absolutne mere variabilnosti razdelimo v dve skupini glede na to, kako vrednotijo variabilnost spremenljivke.

- razmiki: variacijski razmik, kvartilni razmik;
- odkloni: povprečni absolutni odklon, povprečni kvadrirani odklon.

2.5.1.1 Razmiki

Najpreprostejša mera variabilnosti je **variacijski razmik** VR, ki je razlika med maksimalno in minimalno vrednostjo:

$$VR = x_{\max} - x_{\min}$$

Variacijski razmik pove, za koliko je največja vrednost večja od najmanjše vrednosti. Na njegovo vrednost vplivajo osamelci.

Kvartilni razmik Q je razlika med tretjim in prvim kvartilom:

$$Q = Q_3 - Q_1$$

V intervalu od Q_1 do Q_3 je 50% vseh vrednosti, četrtina vrednosti je manjših od Q_1, četrtina večjih od Q_3. Na Q ne vplivajo osamelci.

Razmika sta grafično prikazana v okviru z ročaji.

Primer

Izračunajmo oba razmika za starost diplomantov

VR = 53 - 22 = 31

Q = 27 - 24 = 3

Starost diplomantov je od 22 do 53 let. Najstarejši diplomant je 31 let starejši od najmlajšega. Polovica diplomantov je stara od 24 do 27 let, četrtina je mlajših od 24 let, četrtina starejših od 27 let. Če iz podatkov izločimo vrednost 53, ki je osamelec, se variacijski razmik bistveno spremeni, kvartilni pa ne:

VR = 29 - 22 = 7

Q = 26,75 - 24 = 2,75

2.5.1.2 Odkloni

Pri izračunu razmikov določata mero variabilnosti le dve vrednosti: najmanjša in največja oz. prvi in tretji kvartil. Za izhodišče vrednotenja variabilnosti, kjer upoštevamo vse vrednosti spremenljivke, se uporablja odklone vrednosti od aritmetične sredine $(x_i - \bar{x})$, $i = 1, 2, ..., n$.

Nekateri odkloni so pozitivni, drugi negativni. Kot že vemo, je vsota odklonov enaka nič:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Zato vsota odklonov od aritmetične sredine ni dobro izhodišče za mero variabilnosti. Nadomestimo jo z vsoto absolutnih odklonov:

$$\sum_{i=1}^{n}|x_i - \bar{x}|$$

ali z vsoto kvadriranih odklonov:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Izvedene mere variabilnosti iz vsote absolutnih odklonov se v praksi manj uporabljajo in jih tu ne navajamo.

Najpomembnejši meri variabilnosti sta **varianca** in **standardni odklon**. Opredeljeni na osnovi **vsote kvadriranih odklonov od aritmetične sredine**, ki jo krajše označimo *VKO*:

$$VKO = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Varianca σ^2 je povprečni kvadrirani odklon od aritmetične sredine:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n} \cdot VKO$$

Če imamo podatke iz vzorca in ne iz celotne populacije, izračunamo t. i. **vzorčno varianco** s^2 takole:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Opomba: dokaz, da je pri izračunu vzorčne variance bolj primerno *VKO* deliti z *n*-1 kot z *n*, posreduje matematična statistika. Kot bomo videli kasneje, je tako izračunana vzorčna varianca nepristranska ocena populacijske variance.

Varianca je izražena v kvadriranih merskih enotah obravnavane spremenljivke (npr. kg^2, SIT^2, ...) in je vsebinsko težje obrazložljiva mera. Če varianco korenimo, dobimo standardni odklon σ:

$$\sigma = \sqrt{\sigma^2}$$

oz. **vzorčni standardni odklon** *s*:

$$s = \sqrt{s^2}\ .$$

Standardni odklon ima iste merske enote kot spremenljivka. Tudi vsebinski pomen ima. Izrek pove, da za spremenljivko, ki ima približno simetrično frekvenčno porazdelitev z enim modusnim razredom, pričakujemo v intervalu

- $\bar{x} \mp s$ približno 2/3 vrednosti spremenljivke;
- $\bar{x} \mp 2s$ približno 95% vrednosti spremenljivke;

- $\bar{x} \mp 3s$ skoraj vse vrednosti spremenljivke.

Za poljubno frekvenčno porazdelitev obstaja bolj ohlapna ocena. V poglavju Verjetnostni račun in statistika bomo spoznali, od kje izhajajo te ocene.

Na osnovi izreka (glej Cedilnik: Uvod v verjetnostni račun, str. 99) ugotovimo, da velja

$$s \le \frac{VR}{2}$$

Standardni odklon je manjši ali enak polovici variacijskega razmika. Ta ugotovitev je koristna pri ugotavljanju pričakovane vrednosti za standardni odklon.

Opomba: večina žepnih kalkulatorjev ima statistične funkcije in zna izračunati aritmetično sredino, standardni odklon in vzorčni standardni odklon. Opozarjamo, da so oznake za standardni odklon in vzorčni standardni odklon na različnih kalkulatorjih različne.

Primer

V vzorcu je 7 učencev. Njihova oddaljenost od šole (km) je: 10, 8, 9, 7, 9, 13, 14. Izračunajmo vzorčno varianco in vzorčni standardni odklon za oddaljenost od šole.

Pomožni računi: $\sum_{i=1}^{7} x_i = 70$ $\bar{x} = 10$ $VKO = 40$

$$s^2 = \frac{1}{6} \cdot 40 = 6{,}667$$

$$s = \sqrt{6{,}667} \approx 2{,}6$$

Izračunajte to vrednost s svojim kalkulatorjem.

Pri interpretaciji izračunanega standardnega odklona smo v zadregi, saj nič ne vemo o obliki porazdelitve oddaljenosti od šole. Na osnovi sedmih podatkov je nemogoče karkoli reči o tem. Če bi smeli privzeti, da je ta porazdelitev simetrična z enim modusnim razredom, bi veljala ocena:

- približno dve tretjini učencev imata do šole med 7,4 km in 12,6 km;
- približno 95% učencev med 4,8 km in 15,2 km,
- skoraj vsi med 2,2 km in 17,8 km.

Če so podatki uvrščeni v frekvenčno porazdelitev s K razredi; x_i, i =1, 2, …, K so sredine razredov, VKO dobimo tako, da za vsak razred izračunamo kvadrat razlike med sredino razreda in aritmetično sredino ter le-to upoštevamo f_i-krat. Prispevke po razredih seštejemo:

$$VKO = \sum_{i=1}^{K} f_i (x_i - \bar{x})^2$$

Če je razredov več, je za »peš« računstvo bolj primerna izpeljana formula, ki je ne navajamo.

Primer

Izračunajmo povprečje in standardni odklon za starost ob popisu, posebej za moške in za ženske.

Tabela 2-34: Prebivalci po spolu in po starosti (v dopolnjenih letih) ob popisu 1991 v Sloveniji (Vir: SL - 93, str. 49)

Starost (dop. leta)	Število moških (v tisoč)	Število žensk (v tisoč)	Sp. meja	Zg. meja	Sredina
0 - 9	130,4	123,9	0	10	5
10 – 19	151,3	143,6	10	20	15
20 – 29	148,4	148,5	20	30	25
30 – 39	161,4	155,8	30	40	35
40 – 49	131,3	126,6	40	50	45
50 – 59	109,8	116,5	50	60	55
60 – 69	75,0	109,3	60	70	65
70 – 79	31,2	57,0	70	80	75
80 in več	13,5	31,7	80	—	—
Skupaj	952,3	1013,0			

Za sredino zadnjega razreda bomo upoštevali starost 85 let za oba spola.

Moški:

Pomožni računi:

$$\sum_{i=1}^{9} f_i x_i = (130{,}4 \cdot 5 + 151{,}3 \cdot 15 + ... + 13{,}5 \cdot 85) = 32590{,}5$$

$$\bar{x} = \frac{32590{,}5}{952{,}3} \approx 34{,}2$$

$$VKO = \sum_{i=1}^{K} f_i (x_i - \bar{x})^2 = 400365$$

$$s = \sqrt{\frac{400365}{952{,}3}} \approx 20{,}5$$

Ženske:

Pomožni računi:

$$\sum_{i=1}^{9} f_i x_i = (123{,}9 \cdot 5 + 143{,}6 \cdot 15 + ... + 31{,}7 \cdot 85) = 38117{,}5$$

$$\bar{x} = \frac{38117{,}5}{1013{,}0} \approx 37{,}6$$

$$VKO = \sum_{i=1}^{K} f_i (x_i - \bar{x})^2 = 504863$$

$$s = \sqrt{\frac{504863}{1013{,}0}} \approx 22{,}3$$

Povprečna starost moških ob popisu leta 1991 je bila 34,2 leta, žensk pa 37,6 leta. Pripadajoča standardna odklona sta 20,5 leta za moške in 22,3 leta za ženske.

2.5.2 Relativne mere variabilnosti

Relativne mere variabilnosti omogočajo primerjavo variabilnosti različnih spremenljivk. Najpomembnejša relativna mera variabilnosti je **koeficient variacije**

$$KV\% = \frac{s}{\bar{x}} \cdot 100$$

Koeficient variacije meri, kolikšen odstotek aritmetične sredine predstavlja standardni odklon. Koeficient variacije je relativna mera variabilnosti in omogoča primerjavo variabilnosti različnih spremenljivk.

Primer

Tabela 2-35: Statistike za starost ob popisu za moške in za ženske

Statistike	Moški	Ženske
Aritmetična sredina (dop. leta)	34,2	37,6
Standardni odklon (dop. leta)	20,5	22,3
Koeficient variacije (%)	59,9	59,3

Relativna variabilnost starosti moških in žensk je približno enaka.

NALOGE

1. Število točk pri maturi

Izračunajte in obrazložite mere variabilnosti za število točk pri maturi. Podatki so:

14	11	19	25	18	12	15
16	21	20	20	30	16	18
18	17	18	15	18	24	18
12	30	18	11	18	18	24

2. Meteorološke postaje

V vzorcu je 67 meteoroloških postaj. Navajamo frekvenčno porazdelitev za količino padavin, izmerjeno na teh postajah v letu 1992.

Tabela 2-36: Meteorološke postaje po količini padavin (Vir: Arhiv HMZ)

Količina padavin (mm)	Št. postaj
800 do pod 1200	12
1200 do pod 1600	27
1600 do pod 2000	16
2000 do pod 2400	7
2400 do pod 2800	4
2800 do pod 3200	0
3200 do pod 3600	1

Izračunajte:

a) Povprečno vrednost, modus in mediano ter vrednosti obrazložite.

b) Izračunajte koeficient variacije za količino padavin.

3. Poraba bencina

V 10 slučajno izbranih 4-članskih gospodinjstvih v kraju Gornji Dol so opazovali porabo bencina marca 1995 in septembra 1995.

Tabela 2-37: Poraba bencina (l) za 10 izbranih gospodinjstev v kraju Gornji Dol

Gospodinjstvo	1	2	3	4	5	6	7	8	9	10
Marec	22	15	5	34	33	25	20	34	12	15
September	21	19	6	34	35	20	22	30	11	24

Primerjajte variabilnost porabe bencina v marcu 1995 in v septembru 1995

4. Smreke

V vzorcu je 190 smrek, ki so razvrščene v razrede po debelini debla. Podatki so v frekvenčni tabeli.

Tabela 2-38: Smreke po debelini debla

Debelina debla (cm)	Število smrek
9 do pod 13	6
13 do pod 17	20
17 do pod 21	41
21 do pod 25	34
25 do pod 29	45
29 do pod 33	22
33 do pod 37	18
37 do pod 41	4
Skupaj	190

a) Grafično določite modus. Vrednost obrazložite.
b) Grafično določite mediano. Vrednost obrazložite.
c) Izračunajte povprečno debelino debla.
d) Izračunajte koeficient variacije za debelino debla.

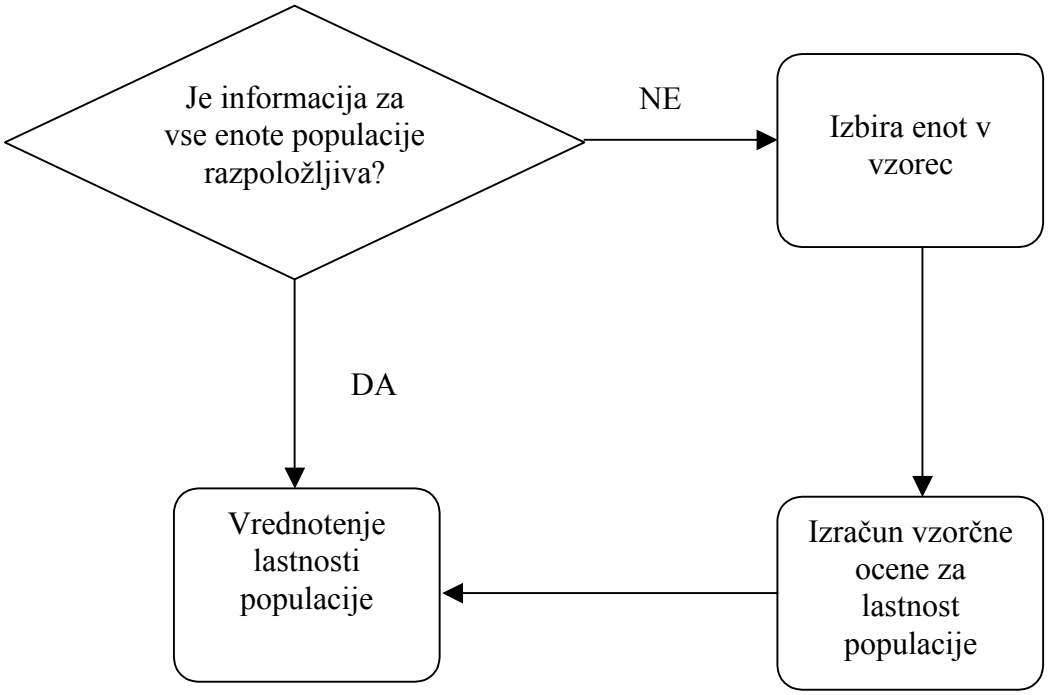
3 OSNOVE VERJETNOSTNEGA RAČUNA

3.1 VERJETNOSTNI RAČUN IN STATISTIKA

Statistika proučuje lastnosti populacije tako, da analizira spremenljivke, ki opisujejo to populacijo. Statistiko zanimajo porazdelitev spremenljivke in določene karakteristike te porazdelitve, kot npr. povprečje, standardni odklon.

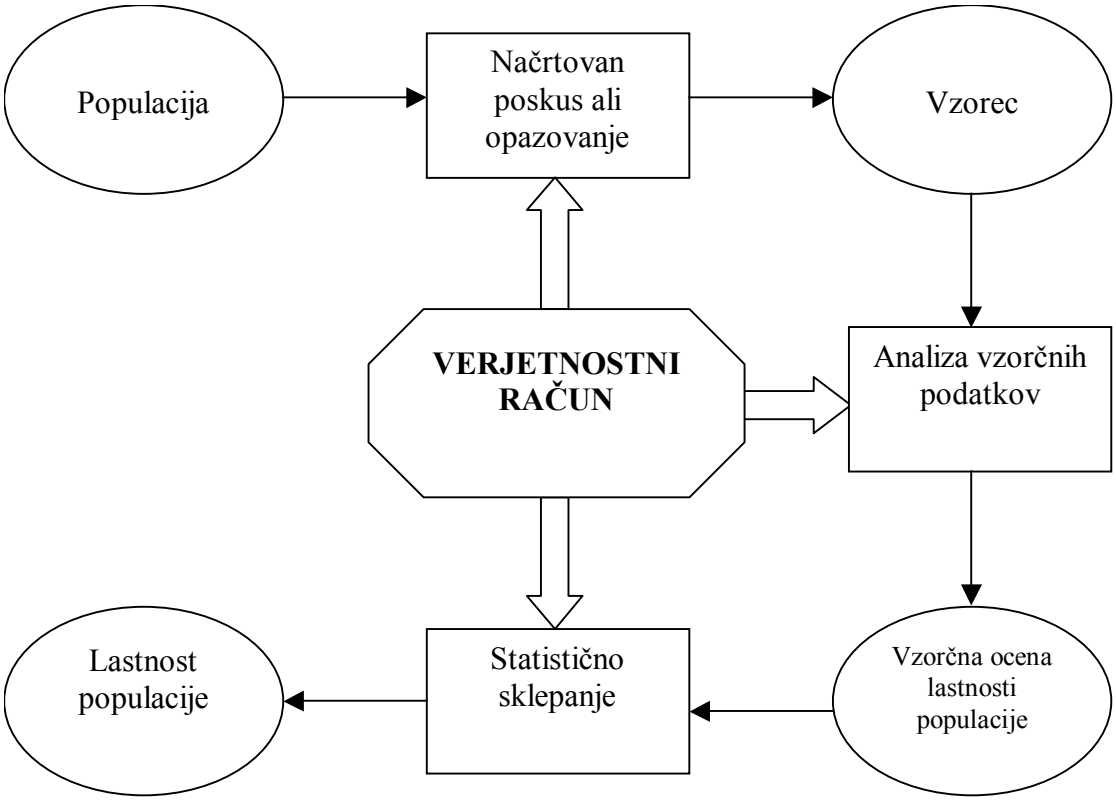
Poglejmo si primer. Proučujemo aktivnost študentov Biotehniške fakultete v šolskem letu 1998/99. Zanima nas, koliko časa na teden študent v povprečju nameni študiju, koliko športni dejavnosti, in, ali sta ti dve količini povezani. V tem primeru proučujemo tri karakteristike: povprečje dveh spremenljivk ter mero povezanosti teh dveh spremenljivk. Če bi imeli možnost pridobiti informacijo od vseh študentov fakultete, bi te količine izračunali. Praktični problemi (dosegljivost študentov, zavračanje anketiranja, preveliki stroški, itd.) narekujejo, da izberemo določeno število študentov v vzorec in jih anketiramo. Na osnovi dobljenih vrednosti izračunamo vzorčne ocene. Iz vzorčnih ocen sklepamo o tem, kaj velja za celotno populacijo.

Naslednja shema ponazarja situacijo in nakazuje našo dejavnost pri vrednotenju lastnosti populacije. V primeru, da je informacija o vseh enotah populacije razpoložljiva, iz le-te vrednotimo proučevano lastnost populacije. Če pa celovite informacije o populaciji nimamo, iz populacije izberemo vzorec in iz njega izračunamo vzorčno oceno. Le-ta služi kot izhodišče za vrednotenje lastnosti populacije.



Slika 3-1: Vrednotenje lastnosti populacije

Verjetnostni račun je matematična disciplina, ki predstavlja osnovno orodje statistike pri delu z nepopolno informacijo. Njegova vloga je razvidna iz naslednje sheme: metode za načrtovanje poskusov oz. opazovanj, metode za izračun vzorčnih ocen in metode za sklepanje iz vzorčnih vrednosti na populacijske vrednosti temeljijo na verjetnostnih predpostavkah.



Slika 3-2: Vloga verjetnostnega računa v statistiki

Ker spada verjetnostni račun med težja poglavja matematike, podajamo nekaj osnov iz verjetnostnega računa, ki jih bomo potrebovali pri statističnem sklepanju. Bralcu, ki se z verjetnostnim računom še ni srečal, svetujemo, da bolj podrobne razlage poišče v literaturi (Čibej, Vadnal, Jamnik).

3.1.1 Slučajna spremenljivka

Verjetnostni račun proučuje t. i. verjetnostne poskuse in njihove izide. Take poskuse opisujejo slučajne spremenljivke. **Slučajna spremenljivka** je količina, ki dobi v vsakem poskusu neko vrednost, ta pa je odvisna od slučaja. Slučajna spremenljivka je določena z dvema podatkoma:

- z zalogo vrednosti,
- s porazdelitvenim zakonom.

Zaloga vrednosti slučajne spremenljivke X so vrednosti, ki jih X more zavzeti. Glede na zalogo vrednosti delimo slučajne spremenljivke v:

- **diskretne**: te imajo končno ali števno neskončno zalogo vrednosti,
- nediskretne; te imajo neštevno zalogo vrednosti. Med njimi so najpomembnejše **zvezne** slučajne spremenljivke.

Porazdelitveni zakon pove, kako je verjetnost porazdeljena po zalogi vrednosti. Najbolj splošna oblika porazdelitvenega zakona je **porazdelitvena funkcija** (**kumulativa verjetnosti**) F, ki je definirana takole:

$$F(x) = P(X \le x)$$

Pri danem x je vrednost funkcije $F(x)$ enaka verjetnosti P, da slučajna spremenljivka X zavzame vrednosti, ki so manjše ali enake x. Porazdelitvena funkcija ima naslednje lastnosti:

- $F(-\infty) = 0 \qquad F(+\infty) = 1$
- F je naraščajoča funkcija

- $P(X > x) = 1 - F(x)$

Za zapis diskretne slučajne spremenljivke uporabljamo **porazdelitveno shemo**:

$$X: \begin{bmatrix} x_1, & x_2, & x_3, & \dots \\ p_1, & p_2, & p_3, & \dots \end{bmatrix}$$

pri čemer velja:

$P(X = x_i) = p_i \qquad 0 \le p_i \le 1$

$$\sum_i p_i = 1$$

Primer

Mečemo kocko, izid X je število pik na kocki. Izid meta je lahko 1, 2, 3, 4, 5, 6, pripadajoče verjetnosti so enake 1/6. Porazdelitvena shema slučajne spremenljivke X je:

$$X: \begin{bmatrix} 1, & 2, & 3, & 4, & 5, & 6 \\ \frac{1}{6}, & \frac{1}{6}, & \frac{1}{6}, & \frac{1}{6}, & \frac{1}{6}, & \frac{1}{6} \end{bmatrix}$$

Porazdelitev s porazdelitveno funkcijo $F(x)$ je zvezna, če obstaja taka funkcija $p(x)$, da je

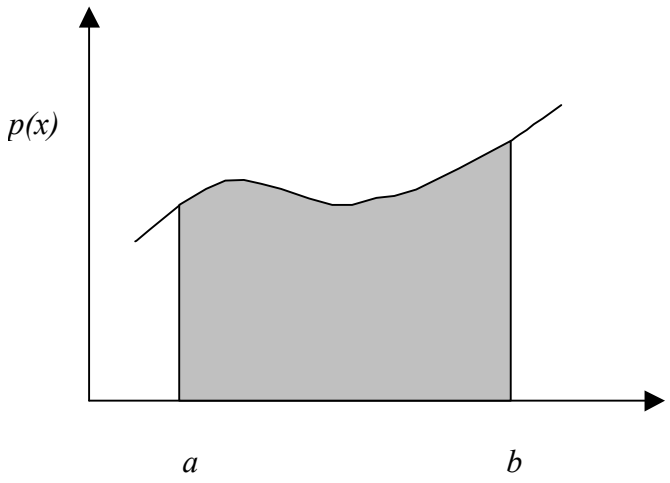
$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)\, dt$$

Funkcijo $p(x)$ imenujemo **gostota verjetnosti**. Ima naslednje lastnosti:

- $p(x) \ge 0$
- $\int_{-\infty}^{\infty} p(x)\, dx = 1$

  Graf funkcije $p(x)$ je nad abscisno osjo, ploščina pod njim pa je enaka 1. Velja še:
- $P(a < X < b) = \int_{a}^{b} p(t)\, dt$

  Ploščina lika, ki ga omejujeta abscisi a in b ter gostota verjetnosti $p(x)$, geometrijsko upodablja iskano verjetnost.



Slika 3-3: Izračun verjetnosti iz gostote verjetnosti

Najpomembnejši karakteristiki slučajne spremenljivke sta njena povprečna vrednost in varianca. Povprečno vrednost slučajne spremenljivke označimo $E(X)$, včasih jo imenujemo pričakovana vrednost slučajne spremenljivke .Varianca slučajne spremenljivke označimo *Var (X)*, nekateri jo imenujejo tudi disperzija slučajne spremenljivke. Poglejmo njuni definiciji.

Povprečna vrednost slučajne spremenljivke $E(X)$ je aritmetična sredina vrednosti spremenljivke X v velikem številu poskusov. Izračun povprečne vrednosti $E(X)$:

$E(X) = \sum_i x_i \cdot p_i$ za diskretne slučajne spremenljivke,

$E(X) = \int_{-\infty}^{\infty} x \cdot p(x)\, dx$ za zvezne slučajne spremenljivke

Velja:

$E(cX) = cE(X), \quad c \in R$

$E(X+Y) = E(X) + E(Y)$

Varianca slučajne spremenljivke $Var(X)$ je aritmetična sredina kvadratov odklonov od $E(X)$ v velikem številu poskusov. Izračun variance $Var(X)$:

$Var(X) = \sum_i (x_i - E(X))^2 \cdot p_i$ za diskretne slučajne spremenljivke

$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot p(x)\, dx$ za zvezne slučajne spremenljivke

Velja:

$Var(cX) = c^2\, Var(X), \quad c \in R$

$Var(X \pm Y) = Var(X) + Var(Y) \pm 2\, Cov(X, Y)$

pri čemer je $Cov(X,Y)$ kovarianca med spremenljivkama X in Y.

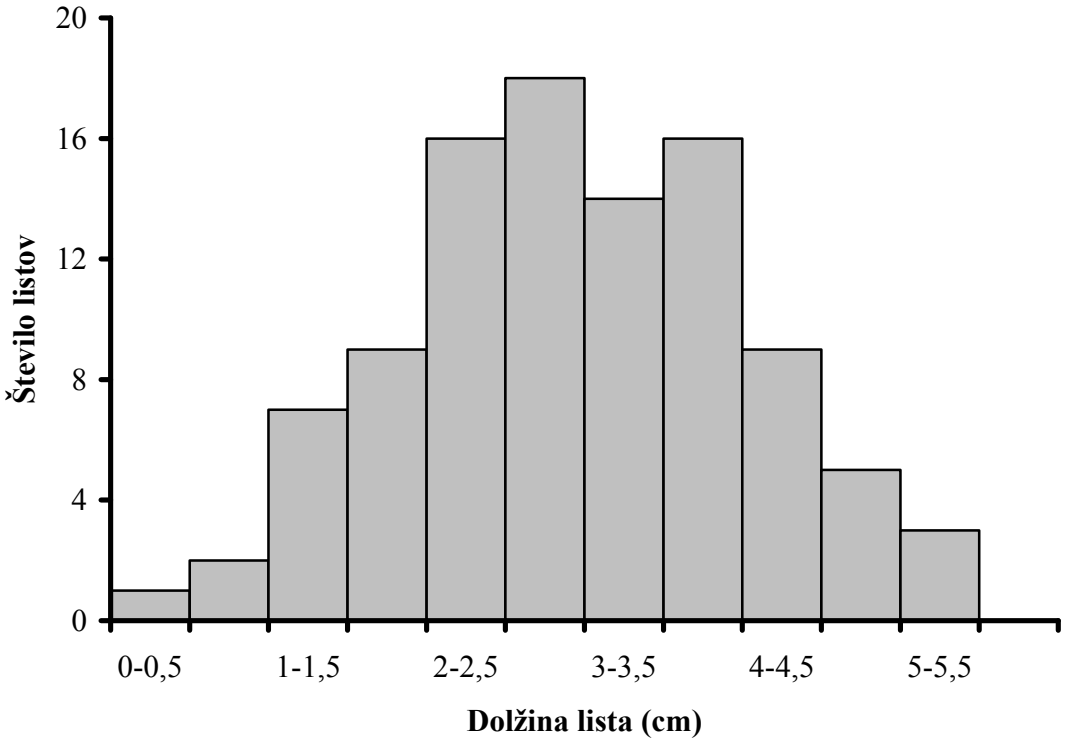
Standardni odklon slučajne spremenljivke X je $\sqrt{Var(X)}$.

3.1.2 Statistična spremenljivka in slučajna spremenljivka

Za razumevanje statistike je pomemben *odnos med statistično spremenljivko in slučajno spremenljivko*. Če je bila izbira enot v vzorec slučajna, je vrednost, ki jo statistična spremenljivka dobi pri neki enoti, odvisna od slučaja. Navadno nekaj vemo o zalogi vrednosti

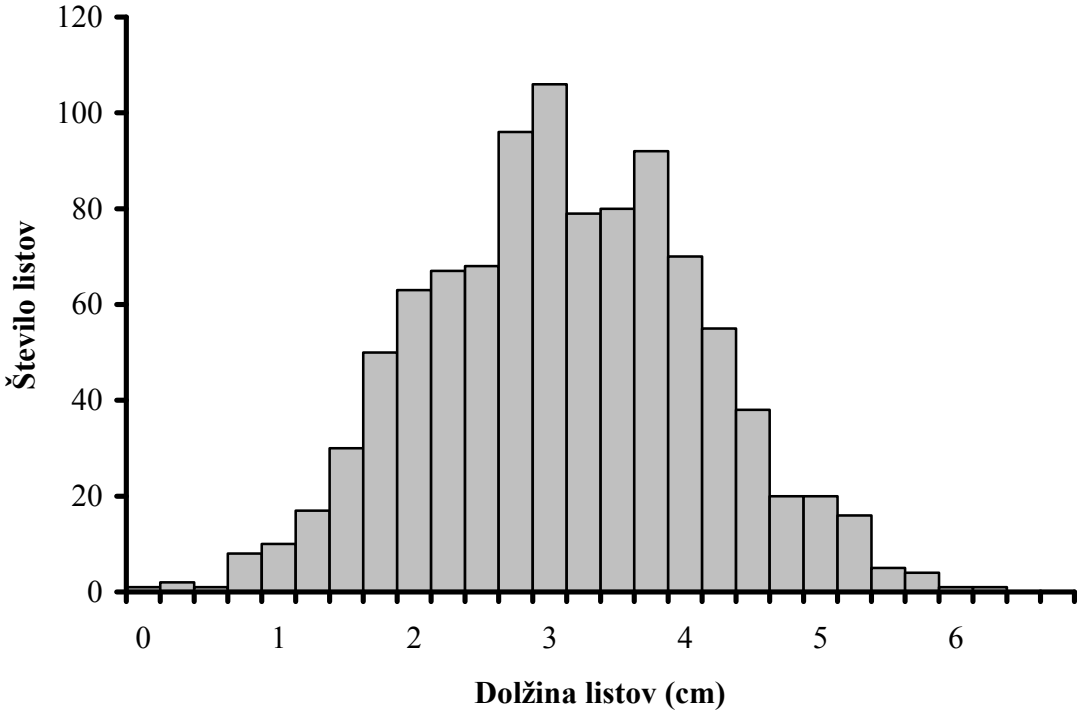
spremenljivke. Če lahko privzamemo še ustrezni porazdelitveni zakon, torej porazdelitev verjetnosti po zalogi vrednosti, imamo slučajno spremenljivko. Povedano drugače: slučajna spremenljivka je *verjetnostni model* za statistično spremenljivko.

Ilustrirajmo to na primeru. Proučujemo dolžino listov nekega drevesa. Dolžino izmerimo na 100 listih. Podatke za statistično spremenljivko uredimo v frekvenčno porazdelitev in jih grafično prikažemo s histogramom.

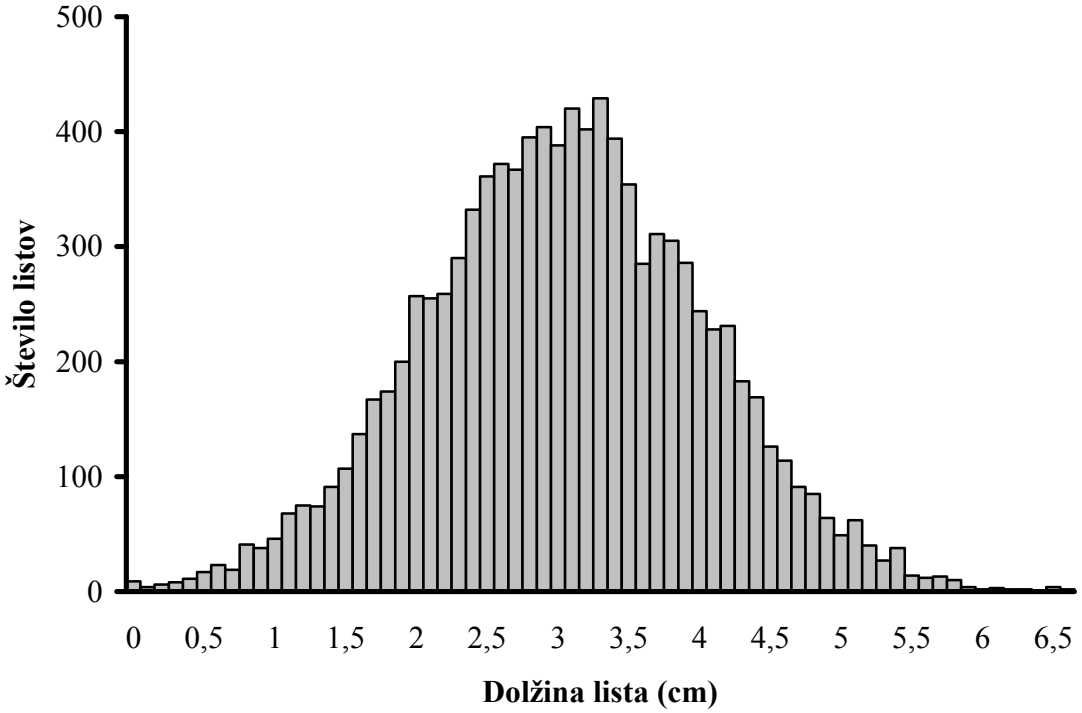


Slika 3-4: Histogram za 100 listov

Slika kaže, da je porazdelitev simetrična in ima en modusni razred. Če število listov v vzorcu povečujemo, se stopničke v histogramu zmanjšujejo.

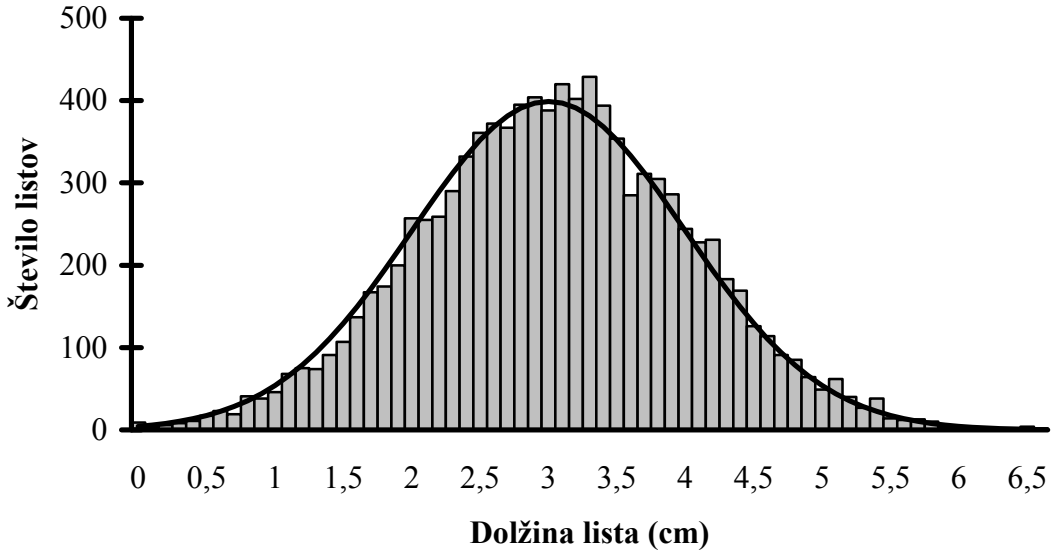


Slika 3-5: Histogram za 1000 listov



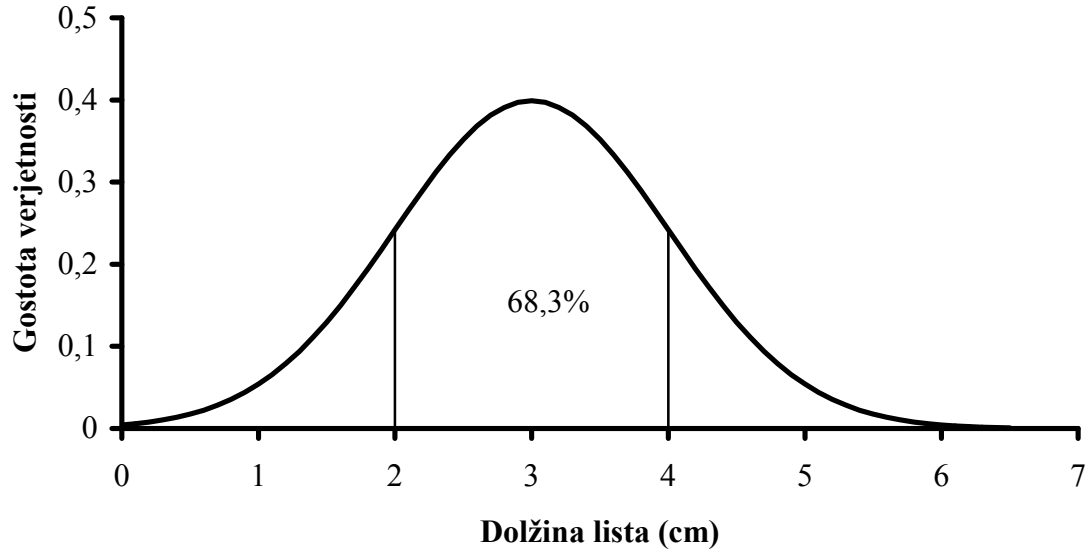
Slika 3-6: Histogram za 10000 listov

Grafični prikaz frekvenčne porazdelitve na velikem vzorcu nakazuje, da lahko histogramu priredimo gladko krivuljo, ki se mu dovolj dobro prilega. Rečemo, da frekvenčno porazdelitev modeliramo z ustrezno matematično funkcijo.



Slika 3-7: Histogram za 10000 listov in matematičen model zanj

To funkcijo narišemo v verjetnostni skali: s tem smo statistični spremenljivki priredili slučajno spremenljivko. V tem primeru smo dolžino lista modelirali s slučajno spremenljivko, ki ima normalno porazdelitev N(3,1).



Slika 3-8: Verjetnost dolžine listov med 2 cm in 4 cm

Dejstvo, da imamo ustrezno verjetnostno porazdelitev za dolžino listov, omogoča, da lahko o dolžini listov povemo še kaj. Npr. izračunamo lahko, kolikšen odstotek listov v celotni populaciji tega drevesa pričakujemo, da ima dolžino med 2 cm in 4 cm. Iskano verjetnost izračunamo takole:

$$\int_{2}^{4} p(x)dx$$

pri čimer je *p(x)* gostota verjetnosti normalne porazdelitve N(3, 1). Kot bomo videli v naslednjem poglavju, je vrednost tega integrala 0,683. Povedano drugače: v proučevani populaciji listov pričakujemo 68,3% listov z dolžino med 2 cm in 4 cm.

3.2 VERJETNOSTNE PORAZDELITVE

Največkrat uporabljene zvezne verjetnostne porazdelitve so:

- normalna (Gaussova) porazdelitev,
- Studentova oz. *t*-porazdelitev,
- χ^2-porazdelitev (hi-kvadrat porazdelitev),
- *F*-porazdelitev;

največkrat uporabljene diskretne verjetnostne porazdelitve so:

- enakomerna diskretna porazdelitev,
- binomska porazdelitev,
- Poissonova porazdelitev.

Pri našem delu bomo obravnavali najprej normalno porazdelitev, nato pa še binomsko porazdelitev. V poglavju o porazdelitvi vzorčnih ocen bomo spoznali Studentovo porazdelitev in χ^2-porazdelitev.

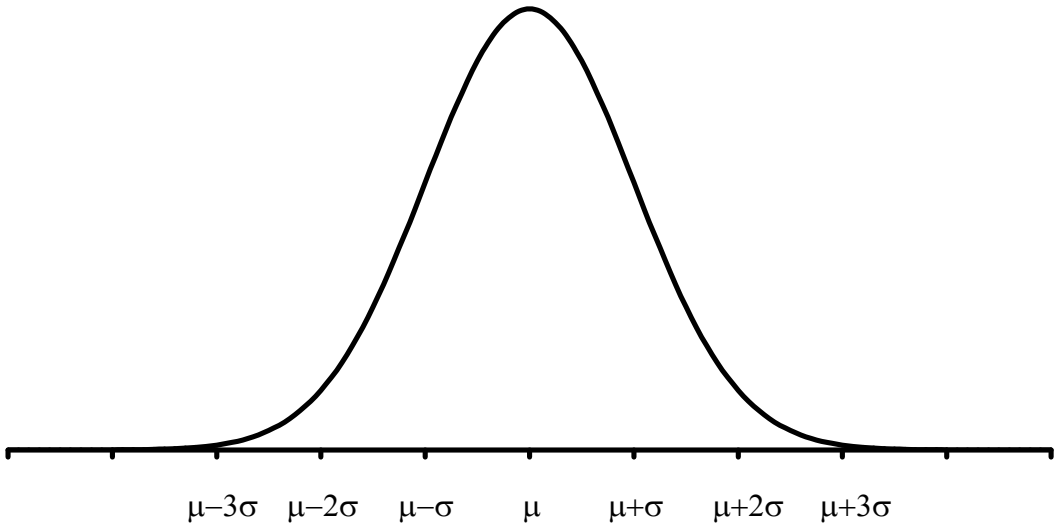
3.2.1 Normalna porazdelitev

Najpomembnejša izmed vseh verjetnostnih porazdelitev je **normalna porazdelitev**. Gostoto verjetnosti za normalno porazdelitev je prvi zapisal francoski matematik Albert de Moivre (1733); izpeljal jo je kot aproksimacijo binomske porazdelitve. Leta 1809 je Carl Friedrich Gauss proučeval porazdelitev izmerkov neke količine, če na izmerke vplivajo le slučajni vplivi. Meritve naj bi bile izvedene pod t. i. normalnimi pogoji, torej brez sistematičnih napak. Verjetnostna porazdelitev izmerkov je znana pod imenom normalna ali Gaussova porazdelitev. Leta 1810 je Laplace izpeljal Laplaceovo lokalno formulo, ki je prva verzija t. i. centralnega limitnega izreka.

Gostota verjetnosti za normalno porazdeljeno slučajno spremenljivko X je:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty$$

Gostota je definirana na celotni realni osi, je zvonaste oblike, je simetrična okoli $x = \mu$, ima maksimum v točki $x = \mu$ in se asimptotično bliža abscisni osi, ko $x \to \infty$ in $x \to -\infty$.



Slika 3-9: Gostota verjetnosti za normalno porazdelitev

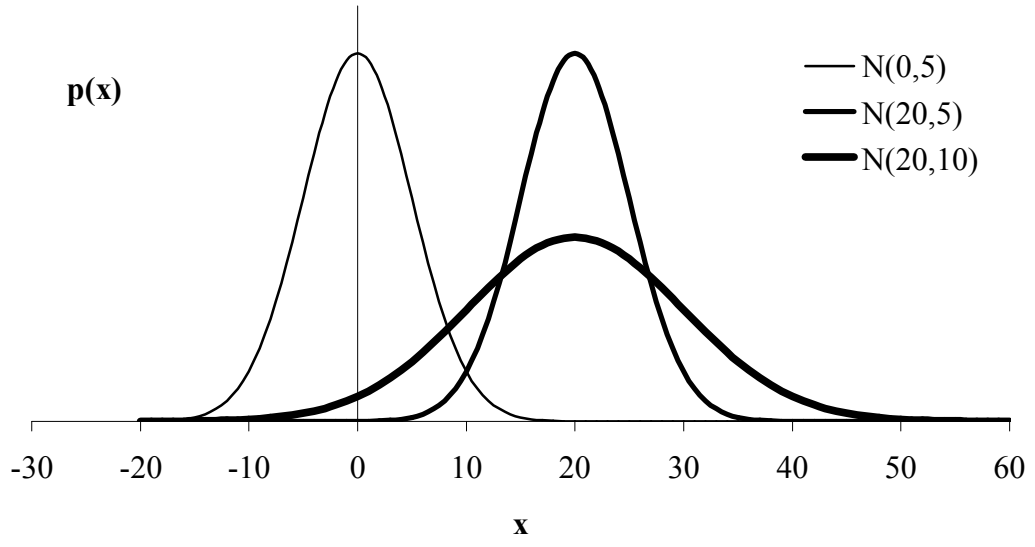
Normalna porazdelitev ima dva parametra μ in σ. Dejstvo, da je slučajna spremenljivka X porazdeljena po normalni porazdelitvi s parametroma μ in σ, krajše zapišemo $X \sim N(\mu, \sigma)$. Izračun pokaže, da velja:

$E(X) = \mu$

$Var(X) = \sigma^2$

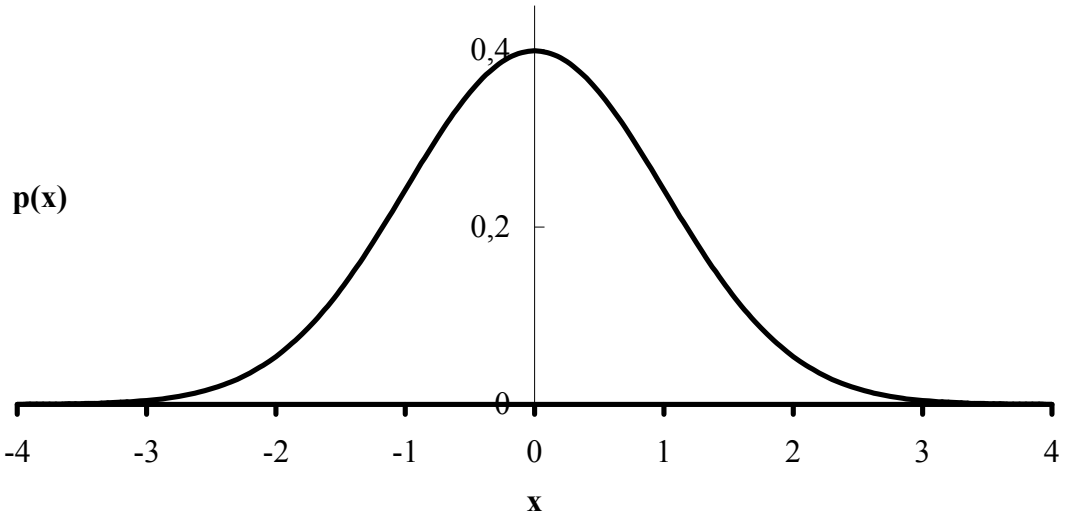
Parametra normalne porazdelitve sta povprečna vrednost $\mu \in R$ in standardni odklon $\sigma > 0$. Parameter μ vpliva na lego krivulje; parameter σ vpliva na obliko krivulje. Večji σ pomeni večjo raztegnjenost v smeri abscisne osi.

Predstavljamo gostoto porazdelitve za tri normalne porazdelitve: $N(0, 5)$, $N(20, 10)$, $N(20, 5)$. Primerjajte njihove grafe.



Slika 3-10: Gostota verjetnosti za tri normalne porazdelitve

Izmed vseh normalnih porazdelitev je najbolj enostavna **standardizirana normalna porazdelitev**, to je porazdelitev $N(0, 1)$.



Slika 3-11: Gostota verjetnosti za standardizirano normalno porazdelitev

Zelo uporabno zvezo med normalnimi porazdelitvami posreduje naslednji izrek.

Izrek

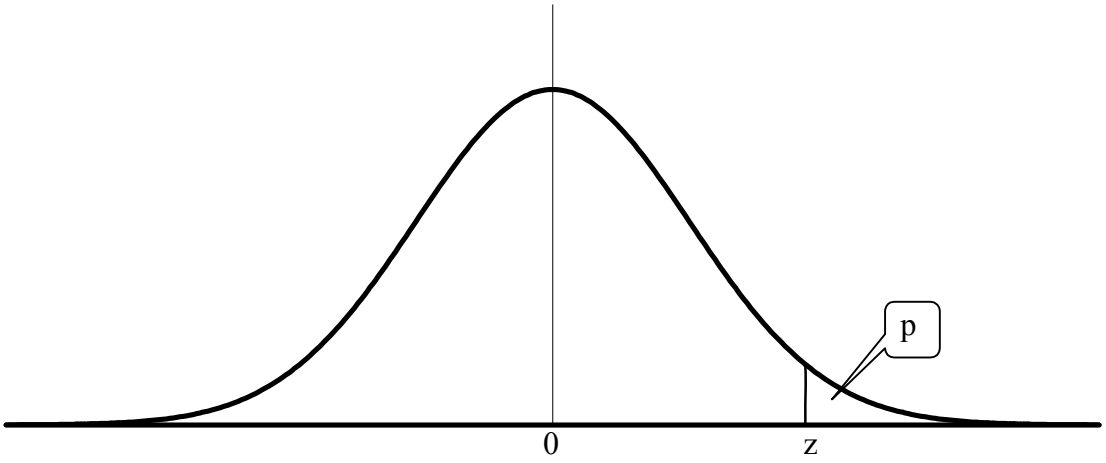
Vsako normalno porazdelitev lahko prevedemo v standardizirano normalno porazdelitev. Če je $X \sim N(\mu, \sigma)$*, je slučajna spremenljivka Z,*

$$Z = \frac{X - \mu}{\sigma}$$

porazdeljena po standardizirani normalni porazdelitvi, torej je $Z \sim N(0, 1)$.

Gornji izrek pove, da za poznavanje vseh normalnih porazdelitev zadošča poznavanje $N(0, 1)$. Izračunavanje verjetnosti za $N(0, 1)$ z določenim integralom nadomeščajo statistične tabele, ki so podane na različne načine. V prilogi v Tabeli 2 je za $N(0, 1)$ za izbrano vrednost z, $z > 0$, podana verjetnost p v 'desnem repu' porazdelitve:

$$p = P(Z > z) = \int_z^\infty p(t)\, dt = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{t^2}{2}} dt$$



Slika 3-12: Uporaba tabele standardizirane normalne porazdelitve

Vrednosti z so tabelirane na intervalu [0, 4] v koraku 0,01. Če potrebujemo večjo natančnost, uporabimo linearno interpolacijo.

Zapišimo še pomembne tabelarične vrednosti standardizirane normalne porazdelitve, ki jih bomo potrebovali kasneje. Uporabili bomo naslednji zapis: za vrednost z_α velja: $P(Z > z_\alpha) = \alpha$. Za verjetnost α izberemo vrednosti 0,05, 0,01 in 0,001 ter njihove polovice, razlog za tako izbiro bomo spoznali kasneje.

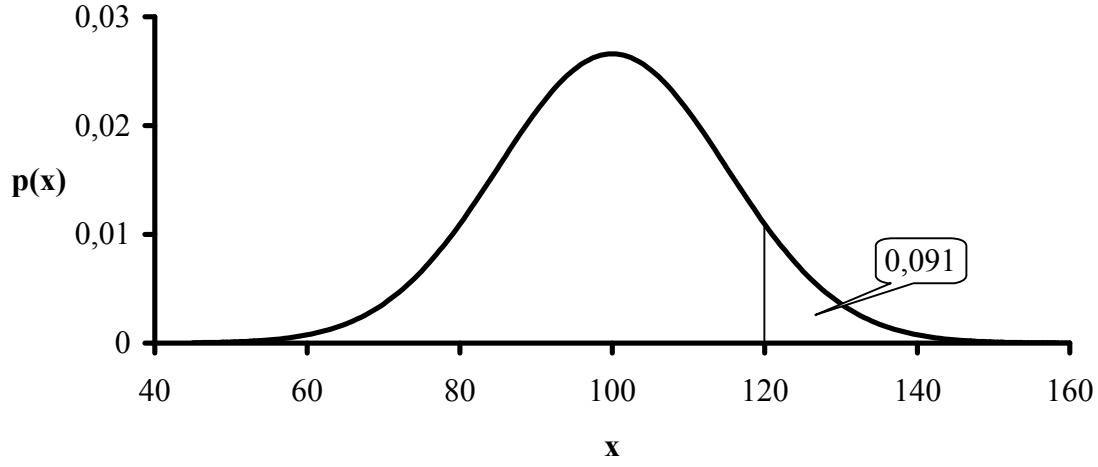
$z_{0,05} = 1{,}645$ $z_{0,025} = 1{,}960$

$z_{0,01} = 2{,}326$ $z_{0,005} = 2{,}576$

$z_{0,001} = 3{,}090$ $z_{0,0005} = 3{,}291$

Primer

Psihologi trdijo, da je v populaciji oseb inteligenčni kvocient IQ normalno porazdeljena slučajna spremenljivka s povprečno vrednostjo 100 in standardnim odklonom 15. Denimo, da je njihova trditev pravilna. Kolikšna je verjetnost, da je IQ nad 120?



Slika 3-13: Porazdelitev inteligenčnega kvocienta

$$P(IQ > 120) = P(Z > \frac{120-100}{15} = 1{,}33\overline{3}) = 0{,}0912$$

Verjetnost, da je IQ nad 120, je 0,0912. Povedano drugače: v populaciji ima 9,1% oseb IQ nad 120.

Za poljubno normalno porazdelitev velja:

$$P(\mu-\sigma<X<\mu+\sigma)=P(-1<Z<+1)=0{,}6827\approx\frac{2}{3}$$

$$P(\mu-2\sigma<X<\mu+2\sigma)=P(-2<Z<+2)=0{,}9545\approx 0{,}95$$

$$P(\mu-3\sigma<X<\mu+3\sigma)=P(-3<Z<+3)=0{,}9973$$

Preverite te zveze s pomočjo tabele.

Zapomnimo si, da sta pri normalni porazdelitvi približno dve tretjini vrednosti v intervalu $(\mu-\sigma,\ \mu+\sigma)$, približno 95% vrednosti v intervalu $(\mu-2\sigma,\ \mu+2\sigma)$ in skoraj vse vrednosti v intervalu $(\mu-3\sigma,\ \mu+3\sigma)$.

NALOGE

Rezultati naj bodo na vsaj 3 mesta pravilni. Uporabite linearno interpolacijo, če je potrebno.

1. Ocene verjetnosti

Za $X\sim N(35,5)$ narišite skico in ocenite verjetnosti:

$P(X>35)$, $P(X\geq 35)$, $P(30<X<40)$, $P(X>40)$, $P(X<25)$, $P(20<X<40)$

2. Standardizirana normalna porazdelitev

Izračunajte verjetnosti:

$P(Z>2{,}5)$ $P(Z\geq 2{,}5)$

$P(Z\leq 2{,}5)$ $P(Z>2{,}25)$

$P(Z>2{,}256)$ $P(Z<-3{,}0)$

$P(Z>-1{,}5)$ $P(-1{,}5<Z<1)$

3. Inteligenčni kvocient

Inteligenčni kvocient IQ je normalno porazdeljena slučajna spremenljivka s povprečno vrednostjo 100 in standardnim odklonom 15.

a) Kolikšna je verjetnost, da je IQ nad 130?
b) Kolikšna je verjetnost, da je IQ nad 200?
c) Kolikšna je verjetnost, da je IQ med 70 in 130?
d) Kolikšna je verjetnost, da je IQ vsaj 110?
e) Kolikšna je verjetnost, da je IQ med 95 in 115?
f) Za koliko izmed 2 milijona ljudi (npr. Slovenija) pričakujemo, da bodo imeli IQ nad 150?

4. Stroj

Da bi dosegli pravilno delovanje stroja, vsak teden prekinejo njegovo delovanje in ga v času prekinitve ustrezno umerijo. Privzeti smemo, da je čas prekinitve porazdeljen N(20 min, 2 min). Kolikšna je verjetnost, da bo delovanje stroja prekinjeno za največ 21 minut?

5. Starost ob prvi zaposlitvi

Predpostavimo, da za starost oseb ob prvi zaposlitvi lahko privzamemo normalno porazdelitev s povprečno vrednostjo 21 let in standardnim odklonom 2 leti. Kolikšna je verjetnost, da bo starost osebe ob prvi zaposlitvi

a) manjša od 18 let
b) manjša od 25 let
c) večja od 19 let
d) med 20 in 24 leti?

3.2.2 *Binomska porazdelitev

Najprej poglejmo, kaj je Bernoullijevo zaporedje poskusov. Zaporedje verjetnostnih poskusov imenujemo **Bernoullijevo zaporedje**, če velja naslednje:

- v vsakem poskusu se lahko zgodi dogodek A ali njegova negacija $\overline{A}$;
- $P(A) = p$, $P(\overline{A}) = 1 - p = q$. Verjetnost p imenujemo **Bernoullijeva verjetnost**;
- izid v posameznem poskusu je neodvisen od izidov v ostalih poskusih.

Zamislimo si Bernoullijevo zaporedje n-tih poskusov. Naj slučajna spremenljivka X šteje, kolikokrat se v Bernoullijevem zaporedju poskusov zgodi dogodek A, torej je *frekvenca dogodka* A. Ta se lahko zgodi 0-krat, 1-krat, 2-krat,…, n-krat; pripadajoče verjetnosti označimo $p_0, p_1, ..., p_n$. Porazdelitvena shema slučajne spremenljivke X je:

$$X = \begin{bmatrix} 0, & 1, & ..., & n \\ p_0 & p_1 & ..., & p_n \end{bmatrix}$$

pri čemer verjetnosti p_k izračunamo po *Bernoullijevi formuli:*

$$p_k = \binom{n}{k} p^k q^{n-k}$$

Frekvenca dogodka A v Bernoullijevem zaporedju n-tih poskusov, torej X, je porazdeljena po **binomski porazdelitvi** s parametroma n in p, kar krajše zapišemo: $X \sim b(n, p)$. Pomen parametrov:

n označuje število poskusov,

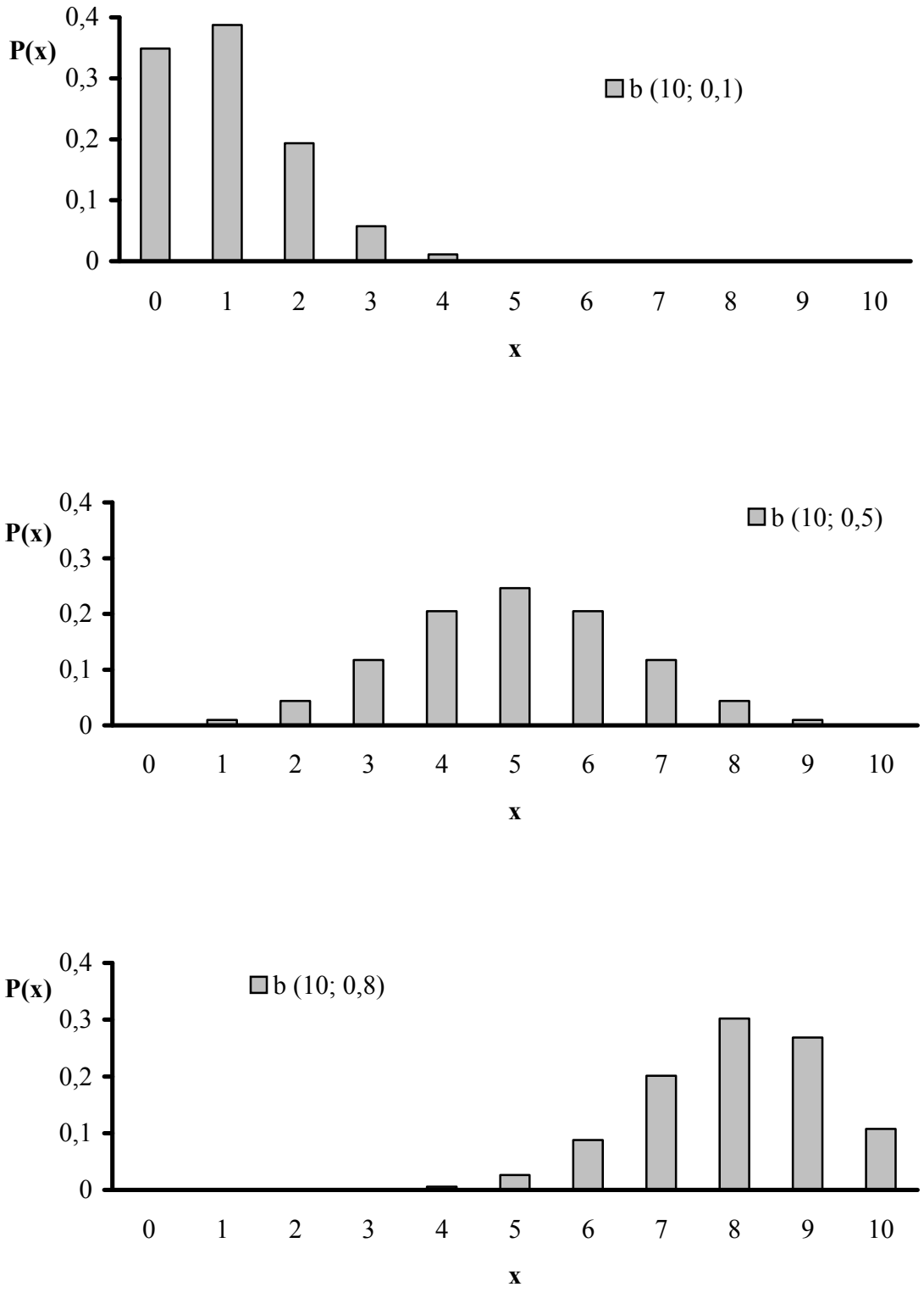
p je Bernoullijeva verjetnost, torej verjetnost za izid A v posameznem poskusu.

Binomska porazdelitev je najpomembnejša diskretna porazdelitev. Ima končno zalogo vrednosti. Za binomsko porazdelitev velja:

$$\mathrm{E}(X) = np$$

$$\mathrm{Var}(X) = npq$$

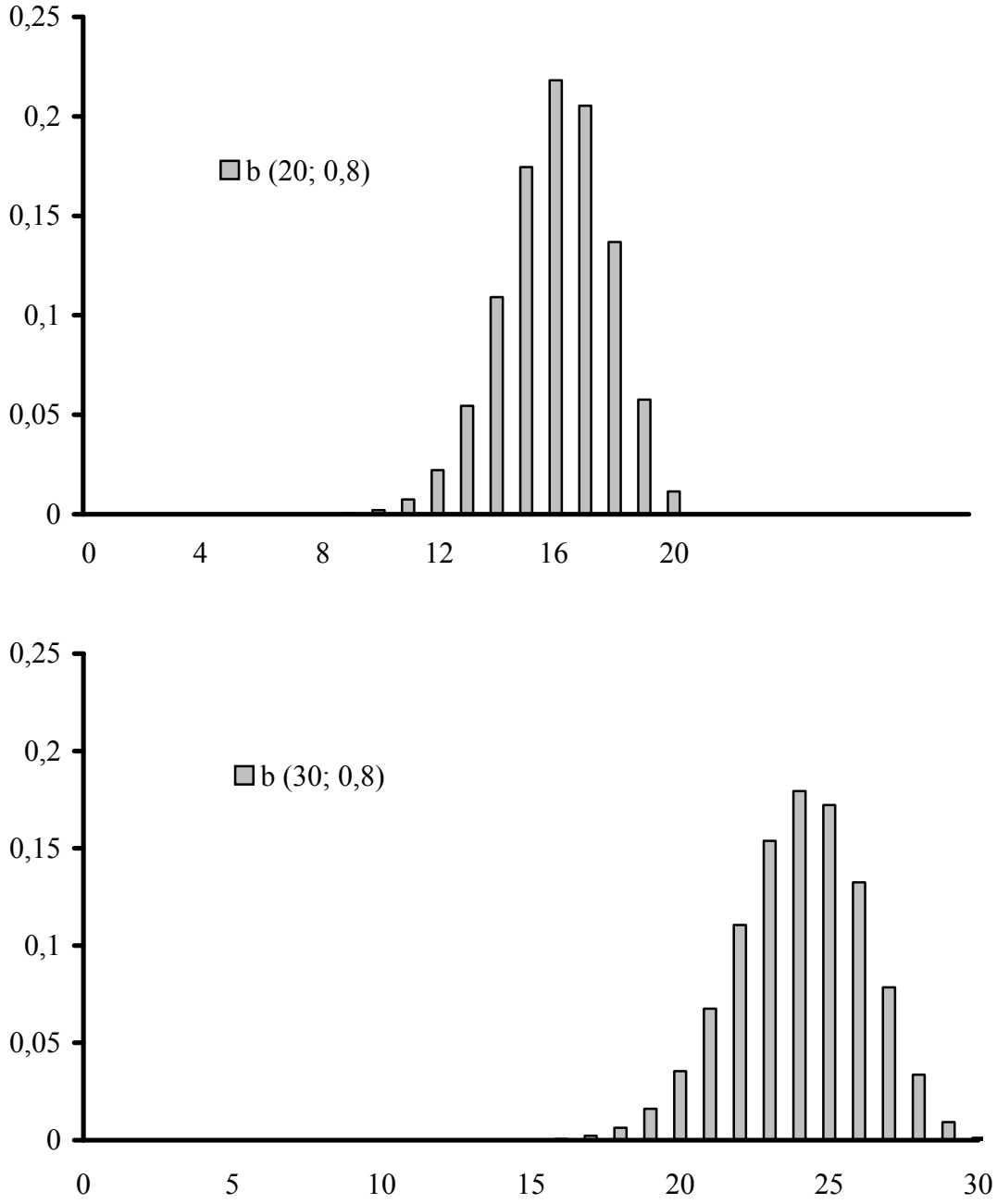
Predstavljamo binomsko porazdelitev za $n = 10$ in za različne vrednosti parametra p, $X \sim b(10, p)$.



Slika 3-14: Binomske porazdelitve b(10, p)

Pri $p = \frac{1}{2}$ slika nakazuje podobnost oblike binomske porazdelitve z obliko normalne porazdelitve. Če je $X \sim b(n, p)$ in je $p = \frac{1}{2}$, postaja binomska porazdelitev z naraščajočim *n*

čedalje bolj podobna normalni porazdelitvi. Pri $p \neq \frac{1}{2}$ se podobnost oblike binomske porazdelitve z obliko normalne porazdelitve nakazuje pri večjih *n*, kar ilustrira spodnja slika.



Slika 3-15: Nekatere binomske porazdelitve za p=0,8

Več o tem pove naslednji izrek.

Izrek

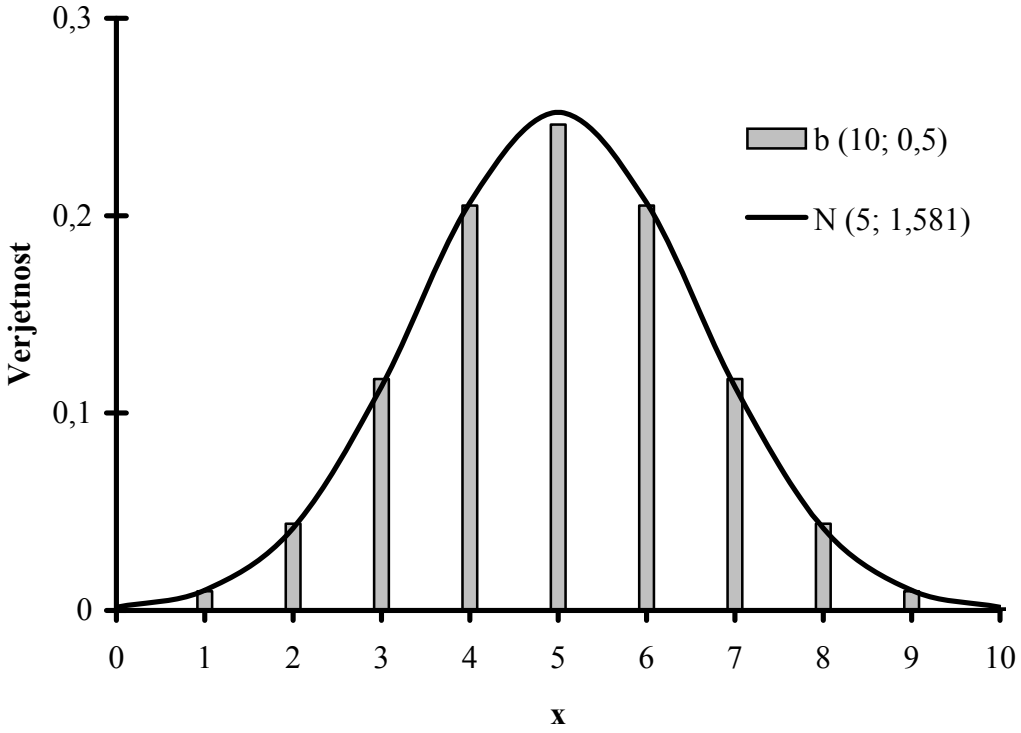
Z naraščanjem števila poskusov preko vsake meje postaja binomska porazdelitev $b(n, p)$ *čedalje bolj podobna normalni porazdelitvi z enako povprečno vrednostjo in enakim standardnim odklonom; torej*

$$n \to \infty \Rightarrow b(n, p) \to N(np, \sqrt{npq})$$

Dokaz temelji na t. i. **Laplaceovi lokalni formuli**:

$$P(X=k) \approx \frac{1}{\sqrt{2\pi npq}} \cdot e^{-\frac{(k-np)^2}{2npq}}$$

ki pove, da pri dovolj velikem *n* binomsko porazdelitev lokalno (v točki) aproksimira normalna porazdelitev. Ilustracijo Laplaceove lokalne formule kaže slika, na kateri je grafično predstavljena binomska porazdelitev b(10; 0,5) in njej prilegajoča se normalna porazdelitev $N(10 \cdot 0{,}5; \sqrt{10 \cdot 0{,}5 \cdot 0{,}5})$.



Slika 3-16: Ilustracija Laplaceove lokalne formule

Binomska porazdelitev je diskretna, normalna pa je zvezna. V praksi smemo binomsko porazdelitev aproksimirati z normalno, če sta hkrati izpolnjena dva pogoja:

- $np \geq 5$
- $nq \geq 5$

Opomba: v literaturi najdemo tudi druge pogoje za aproksimacijo, ki so strožji.

Primer

Verjetnost, da dobimo pri križanju pšenice sort A in B sorto A, je 0,25. Kolikšna je verjetnost, da dobimo pri 4800 križanjih manj kot 1000-krat sorto A?

Naj X šteje križanja, pri katerih dobimo sorto A. Velja: $X \sim b(4800; 0{,}25)$, iskana verjetnost pa je:

$$P(X<1000) = \sum_{k=0}^{999} \binom{4800}{k} \cdot \left(\frac{1}{4}\right)^k \cdot \left(\frac{3}{4}\right)^{4800-k}$$

Temu napornemu računstvu se bomo izognili, saj smemo binomsko porazdelitev aproksimirati z normalno: $np \geq 5$ in $nq \geq 5$. Torej:

$b(4800; 0{,}25) \approx N(1200, 30)$

$P(X < 1000) = P(Z < -6{,}\overline{6}) = 0{,}0000$

Verjetnost, da dobimo pri 4800 križanjih manj kot 1000-krat sorto A, je 0,0000.

Poglejmo uporabnost binomske porazdelitve v statistiki. Bernoullijevo zaporedje *n*-tih neodvisnih poskusov je ustrezen verjetnostni model, kadar proučujemo dvojiško spremenljivko na slučajnem vzorcu velikosti *n*. Vsaka enota ima eno od dveh možnosti (dogodek A ali dogodek $\overline{A}$), vse enote imajo isto verjetnost dogodka A (to je verjetnost *p*), slučajna izbira enot v vzorec pa omogoča, da privzamemo, da sta vrednosti pri dveh različnih enotah neodvisni. Slučajna spremenljivka *X*, ki je porazdeljena po binomski porazdelitvi b(*n*, *p*), šteje enote v slučajnem vzorcu velikosti *n*, ki imajo izid A. Npr. za slučajni vzorec oseb velikosti 100 proučujemo, ali oseba kadi ali ne kadi. Slučajna spremenljivka *X*, ki je porazdeljena po b(100, *p*), šteje število kadilcev v tem vzorcu, pri čemer je *p* verjetnost, da oseba kadi.

NALOGE

1. Družina

Verjetnost, da je posamezni otrok deček, je 0,5. V družini je 6 otrok. Izračunajte verjetnost, da so:

a) 3 dečki in 3 deklice
b) da je manj deklic kot dečkov.

2. Škart

Verjetnost, da je v množični proizvodnji izdelek slab, je 0,02. V pošiljkah je po 100 izdelkov.

a) Zapišite porazdelitev za število slabih izdelkov v pošiljki.
b) Kolikšno je pričakovano število slabih izdelkov v pošiljki?
c) Ocenite verjetnost, da je v pošiljki 20 slabih izdelkov.
d) Ali lahko to porazdelitev aproksimiramo z normalno porazdelitvijo?

3. Francoska ruleta

Kolo pri francoski ruleti ima 37 krožnih izsekov, 18 črnih, 18 rdečih in 1 zelenega. Izberimo stavo, pri kateri stavimo 1 tolar na rdeče. Če kroglica pade na rdeč izsek, dobimo nazaj našo stavo in še tolar zraven, sicer (če kroglica pade na črno ali na zeleno) pa izgubimo našo stavo.

a) Zapišite slučajno spremenljivko, ki opisuje izid posamezne stave. Njene vrednosti sta dve: v posamezni igri lahko priigramo 1 tolar ali pa izgubimo en tolar.
b) Izračunajte pričakovano vrednost te slučajne spremenljivke.
c) Kolikšno finančno stanje lahko pričakujemo po 250 zaporednih stavah na rdeče?

4. Drevesnica

V drevesnici cepijo jablane. Sadjarji navajajo, da je verjetnost, da se cepič posuši, enaka 0,10.

a) Za koliko % cepičev pričakujemo, da se bodo posušili?
b) Kolikšna je verjetnost, da se bo pri 1000 cepljenih posušilo natanko 10 cepičev?

3.3 PORAZDELITVE VZORČNIH STATISTIK

Razložimo najprej naslov. Imamo vzorec velikosti *n*. Na tem vzorcu ima spremenljivka *X* vrednosti: $x_1, x_2, ..., x_n$. **Vzorčna statistika** *u* je poljubna funkcija vzorčnih vrednosti:

$u = f(x_1, x_2, ..., x_n)$. Poznamo že precej vzorčnih statistik, npr. kvantili, mere sredine, mere variabilnosti.

Nekatere vzorčne statistike so še posebej pomembne, saj so ocene parametrov pripadajoče verjetnostne porazdelitve. Poglejmo nekaj primerov.

X je številska statistična spremenljivka, za katero sta najpomembnejši vzorčni statistiki vzorčna aritmetična sredina $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

in vzorčni standardni odklon s

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Če za *X* privzamemo normalno porazdelitev $N(\mu, \sigma)$, je $\bar{x}$ ocena za μ, s pa ocena za σ. To spoznanje posreduje matematična statistika.

Posebno vlogo pri statističnem sklepanju imata ***z*-statistika**

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

in ***t*-statistika**, ki je znana tudi pod imenom **Studentova statistika**

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Natančneje ju bomo spoznali kasneje.

Spomnimo se populacije vseh vzorcev velikosti *n,* ki smo jo spoznali v uvodu. Na vsakem od teh vzorcev ima vzorčna statistika *u* svojo vrednost. Njenih vrednosti je toliko, kolikor je vseh možnih vzorcev velikosti *n*. Da je to število ogromno že pri skromni velikosti osnovne populacije, smo spoznali v uvodu.

Vzorčni statistiki *u* priredimo slučajno spremenljivko *U*. Zanima nas verjetnostna porazdelitev slučajne spremenljivke *U*, njeno poznavanje nam bo potrebno za statistično sklepanje iz vzorca na populacijo.

Pogledali bomo verjetnostno porazdelitev za nekaj najpomembnejših vzorčnih statistik. Za prve korake v uporabno statistiko zadošča poznavanje porazdelitve vzorčnih aritmetičnih sredin in porazdelitve *t*-statistik.

3.3.1 Porazdelitev vzorčnih aritmetičnih sredin

Na populaciji proučujemo številsko spremenljivko *X*. Izbrana velikost vzorca je *n*. Zamislimo si, da na vsakem vzorcu velikosti *n* iz podatkov izračunamo vzorčno aritmetično sredino $\bar{x}$. Vsak vzorec generira svojo vrednost $\bar{x}$. Pripadajočo slučajno spremenljivko označimo $\bar{X}$, njene vrednosti so vse vzorčne aritmetične sredine $\bar{x}$, ki jih dobimo na populaciji vzorcev velikosti *n*. Verjetnostno porazdelitev vzorčnih aritmetičnih sredin podajata naslednja dva izreka.

Izrek

Če je slučajna spremenljivka X na osnovni populaciji porazdeljena $N(\mu, \sigma)$*, potem je slučajna spremenljivka* $\bar{X}$ *na populaciji vseh vzorcev velikosti n porazdeljena* $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Izrek pove, da vzorčne aritmetične sredine podedujejo normalno porazdelitev in prvi parameter izhodiščne spremenljivke X; povprečna vrednost vzorčnih aritmetičnih sredin je enaka povprečni vrednosti izhodiščne spremenljivke. Varianca vzorčnih aritmetičnih sredin je n-krat manjša od variance izhodiščne spremenljivke X. Če izbiramo večje vzorce, se varianca vzorčnih aritmetičnih sredin zmanjšuje, pripadajoča porazdelitev je bolj zgoščena okoli povprečne vrednosti.

Opomba: ta izrek velja, kadar gre za vzorčenje z vračanjem enot. Če enot po vzorčenju ne vračamo, teorija pokaže, da se varianca vzorčnih aritmetičnih sredin izrazi takole:

$$Var(\overline{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

Ker je člen

$$\frac{N-n}{N-1} < 1$$

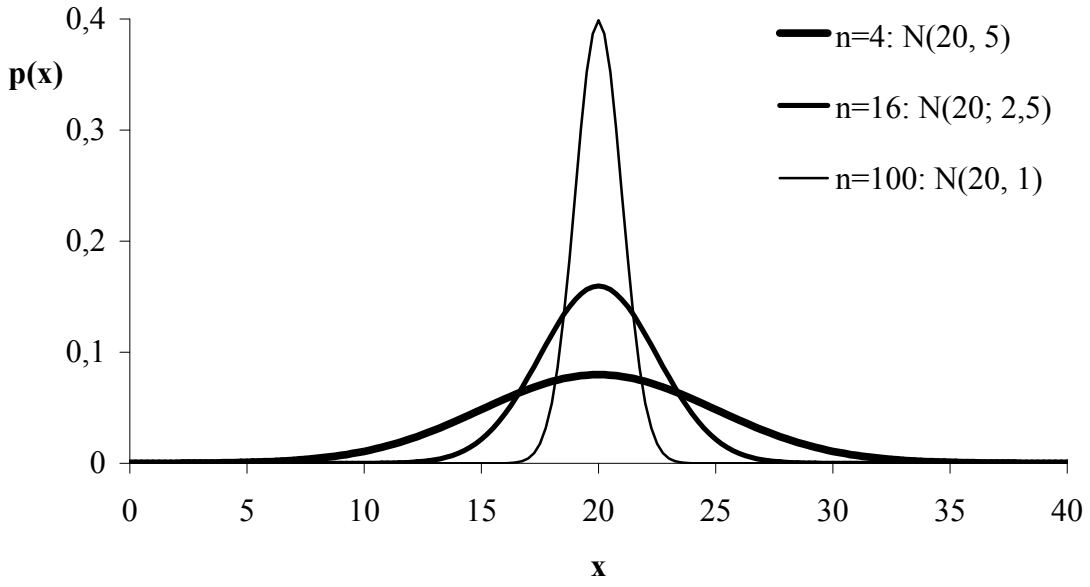
je varianca v primeru, ko enote ne vračamo, manjša kot v primeru, da enote vračamo. Pokažemo lahko, da je razlika med formulama zanemarljivo majhna, če je osnovna populacija velika in je vzorec majhen del populacije:

$$\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \approx \frac{N-n}{N} \cdot \frac{\sigma^2}{n} = \left(1 - \frac{n}{N}\right) \cdot \frac{\sigma^2}{n} \approx \frac{\sigma^2}{n}$$

Ker je v praksi običajno tako, v nadaljevanju ne bomo ločili med vzorci z vračanjem in vzorci brez vračanja enot.

Primer

Naj velja: $X \sim N(20, 10)$. Porazdelitev vzorčnih aritmetičnih sredin v vzorcih velikosti 4 je: $\overline{X} \sim N(20, 5)$, v vzorcih velikosti 16 je $\overline{X} \sim N(20; 2{,}5)$, v vzorcih velikosti 100 pa je: $\overline{X} \sim N(20, 1)$.



Slika 3-17: Porazdelitev vzorčnih aritmetičnih sredin pri različnih velikostih vzorcev

Centralni limitni izrek

$\overline{X}$ se pri velikih vzorcih porazdeljuje približno normalno tudi tedaj, ko verjetnostna porazdelitev slučajne spremenljivke X na osnovni populaciji ni normalna. Velja:

$$\overline{X} \approx N\left(E(X), \sqrt{Var(X)/n}\right)$$

Izrek pove, da pri velikih vzorcih ni treba poznati porazdelitve izhodiščne spremenljivke *X*, zadošča poznavanje njene povprečne vrednosti $E(X)$ in variance $Var(X)$. Velja isto kot prej: porazdelitev vzorčnih aritmetičnih sredin je (približno) normalna, povprečje se podeduje od izhodiščne spremenljivke, varianca pa je *n*-krat manjša.

Centralni limitni izrek lahko uporabimo v praksi, če je velikost vzorca večja od 30. Kot bomo videli kasneje, je centralni limitni izrek izjemno pomemben za statistiko.

Primer

Psihologi trdijo, da je v populaciji inteligenčni kvocient $IQ \sim N(100, 15)$. Predpostavimo, da je njihova trditev pravilna. Izračunajmo naslednje verjetnosti in obrazložimo njihov pomen.

a) $P(IQ > 115) = P\left(Z > \frac{115-100}{15}\right) = P(Z > 1) = 0{,}1587$

V populaciji ima 15,9% oseb IQ nad 115.

b) Tvorimo vzorce velikosti 4. Potem velja: $\overline{IQ} \sim N(100,\ 7{,}5)$.

$$P(\overline{IQ} > 115) = P\left(Z > \frac{115-100}{7{,}5}\right) = P(Z > 2) = 0{,}0228$$

2,3% vzorcev velikosti 4 ima vzorčno aritmetično sredino nad 115.

c) Tvorimo vzorce velikosti 25. Potem velja: $\overline{IQ} \sim N(100,\ 3)$.

$$P(\overline{IQ} > 115) = P(Z > 5) = 0{,}0000$$

Obstaja izjemno majhna verjetnost tega dogodka. 0,00% vzorcev velikosti 25 ima vzorčno aritmetično sredino nad 115. Bolj natančno: iz Tabele 2 v prilogi (zadnja vrstica spodaj) je razvidno, da na deset milijonov vzorcev velikosti 25 pričakujemo 3 vzorce, katerih vzorčna aritmetična sredina presega 115.

3.3.2 *Porazdelitev vzorčnih deležev

X je dvojiška statistična spremenljivka, ki jo proučujemo v slučajnem vzorcu velikosti *n*. Njena najpomembnejša vzorčna statistika je **vzorčni delež**:

$$\hat{p} = \frac{x}{n}$$

pri čemer označuje *x* število enot, ki imajo lastnost A. Vsak vzorec velikosti *n* generira svojo vrednost $\hat{p}$. Zanima nas verjetnostna porazdelitev pripadajoče slučajne spremenljivke $\frac{X}{n}$.

V razdelku o binomski porazdelitvi smo spoznali, da je Bernoullijevo zaporedje *n*-tih neodvisnih poskusov ustrezni verjetnostni model, kadar proučujemo dvojiško spremenljivko na slučajnem vzorcu velikosti *n*. Naj bo *X* frekvenca dogodka A v Bernoullijevem zaporedju *n*-tih poskusov, torej velja: $X \sim b(n, p)$. Če lahko binomsko porazdelitev $b(n, p)$

aproksimiramo z normalno $N(np, \sqrt{npq})$, je porazdelitev slučajne spremenljivke $\frac{X}{n}$ približno normalna:

$$\frac{X}{n} \approx N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Njena povprečna vrednost je enaka Bernoullijevi verjetnosti *p*, na njeno varianco pa vpliva tudi velikost vzorca.

Primer

Verjetnost, da je izdelek slab, je 0,05. V kontrolnih vzorcih je po 250 izdelkov. Koliko odstotkov kontrolnih vzorcev velikosti 250 vsebuje več kot 10% slabih izdelkov?

Naj bo *X* frekvenca slabih izdelkov v kontrolnih vzorcih velikosti 250. Velja: $X \sim b\,(250;\ 0{,}05)$. Ker je: $n \cdot p = 12{,}5 > 5$ in $n \cdot q = 237{,}5 > 5$, je aproksimacija binomske porazdelitve z normalno sprejemljiva. Velja:

$$\frac{X}{n} \approx N\left(0{,}05; \sqrt{\frac{0{,}05 \cdot 0{,}95}{250}}\right) = N(0{,}05;\ 0{,}0138)$$

iskana verjetnost pa:

$$P\left(\frac{X}{n} > 0{,}10\right) \approx 0{,}0001$$

Približno 0,01% vzorcev velikosti 250 vsebuje več kot 10% slabih izdelkov. Povedano drugače: na 10000 vzorcev velikosti 250 pričakujemo en vzorec, ki vsebuje več kot 10% slabih izdelkov.

3.3.3 Porazdelitev *t*-statistik

Spoznali smo že naslednje dejstvo: če je $X \sim N(\mu, \sigma)$, je $\overline{X} \sim N\,(\mu, \frac{\sigma}{\sqrt{n}})$ in posledično je spremenljivka

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)\,.$$

Če poznamo oba parametra normalne porazdelitve μ in σ, iz vzorca velikosti *n* izračunamo $\bar{x}$ in nato vrednost ***z*-statistike**:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Angleški statistik William Gosset, znan pod psevdonimom Student, je v izrazu za *z* nadomestil parameter σ z njegovo vzorčno oceno *s* in tako opredelil ***t*-statistiko**:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Za njen izračun potrebujemo parameter μ in vzorčni oceni $\bar{x}$ in *s*. Gosset je ugotovil, da se pri majhnih vzorcih verjetnostna porazdelitev *t*-statistike bistveno loči od standardizirane normalne porazdelitve, pri velikih vzorcih pa je ta porazdelitev zelo blizu standardizirane

normalne porazdelitve. Izpeljal je verjetnostno porazdelitev *t*-statistik in jo imenoval **Studentova** oz. ***t*-porazdelitev**. Navajamo naslednji izrek.

Izrek

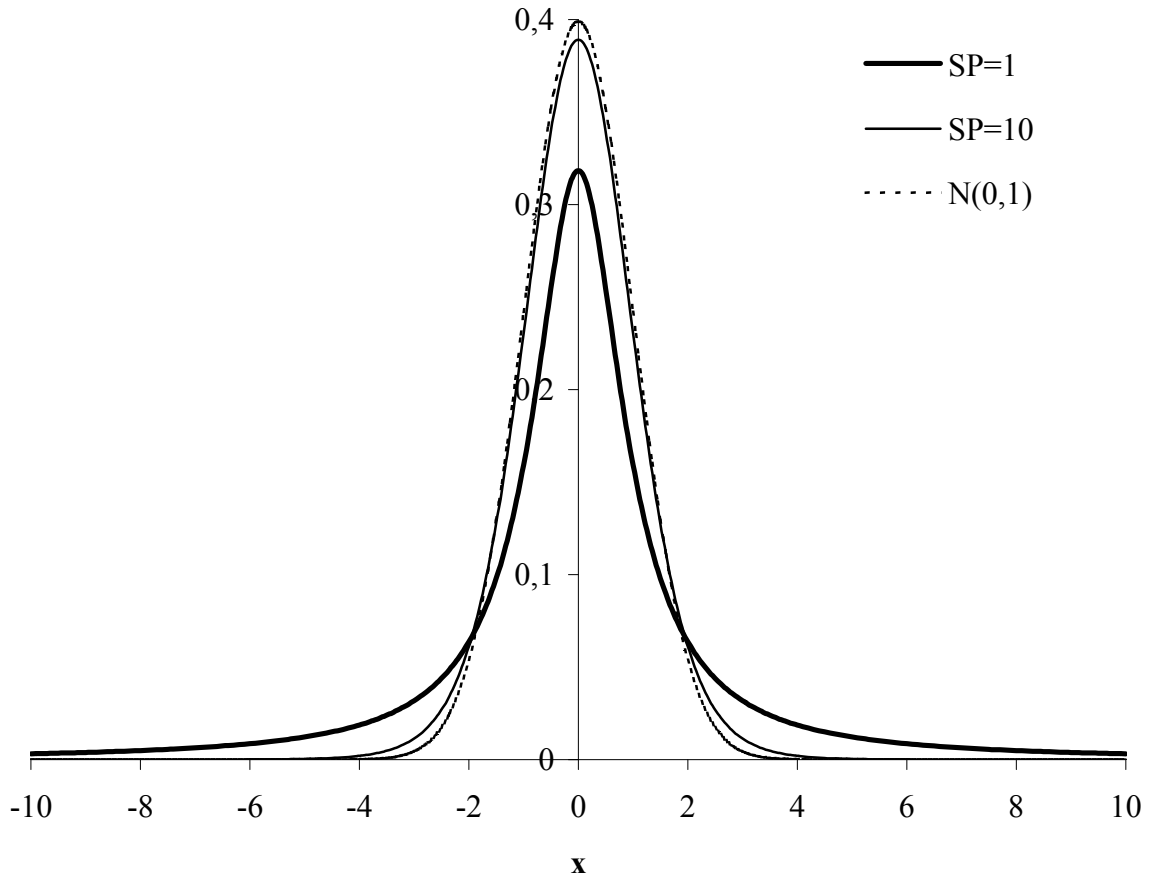
Na populaciji vzorcev velikosti n je slučajna spremenljivka

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

porazdeljena po Studentovi porazdelitvi z n-1 stopinjami prostosti. To krajše zapišemo takole: $T \sim t \;\; (SP = n-1)$.

Lastnosti Studentove porazdelitve so:

- gostota verjetnosti za *t*-porazdelitev je po obliki podobna gostoti verjetnosti za *N*(0, 1). Funkcija je zvezna, definirana na celotni realni osi, je simetrična okoli 0. Formule za gostoto verjetnosti ne navajamo;
- parameter Studentove porazdelitve imenujemo *stopinje prostosti*, $SP = 1, 2, ..., \infty$; stopinje prostosti določajo obliko porazdelitve;
- če je *SP* majhno število, je *t*-porazdelitev bolj raztegnjena v levo in desno kot *N* (0, 1), ko se *SP* povečuje, postaja *t*-porazdelitev čedalje bolj podobna *N* (0, 1) ;
- v limiti je Studentova porazdelitev enaka standardizirani normalni porazdelitvi: $t(SP = \infty) = N(0, 1)$.



Slika 3-18: Studentova porazdelitev z različnimi stopinjami prostosti

Vrednosti *t*-porazdelitve so tabelirane. Tabele so drugačne kot pri normalni porazdelitvi. V Tabeli 3 v prilogi je Studentova porazdelitev. Oznaka α predstavlja verjetnost v desnem repu porazdelitve. Za izbrane verjetnosti α (0,10, 0,05, 0,025, 0,01 itd.) in za izbrane stopinje prostosti SP je v tabeli navedena vrednost *t*, za katero velja $P(T \geq t) = \alpha$. Zadnja vrstica tabele pri $SP = \infty$ se ujema s tabelo za $N(0, 1)$. Iz nje najlažje odčitamo pomembne vrednosti standardizirane normalne porazdelitve, ki smo jih navedli v poglavju o normalni porazdelitvi.

3.3.4 *Porazdelitev vzorčnih varianc

Naj bo $X \sim N(\mu, \sigma)$. Zamislimo si, da na vsakem vzorcu velikosti *n* izračunamo vzorčno varianco s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Vsak vzorec velikosti *n* generira svojo vrednost s^2. Tem vrednostim priredimo slučajno spremenljivko S^2. Zanima nas njena verjetnostna porazdelitev. Informacijo o tem podaja naslednji izrek.

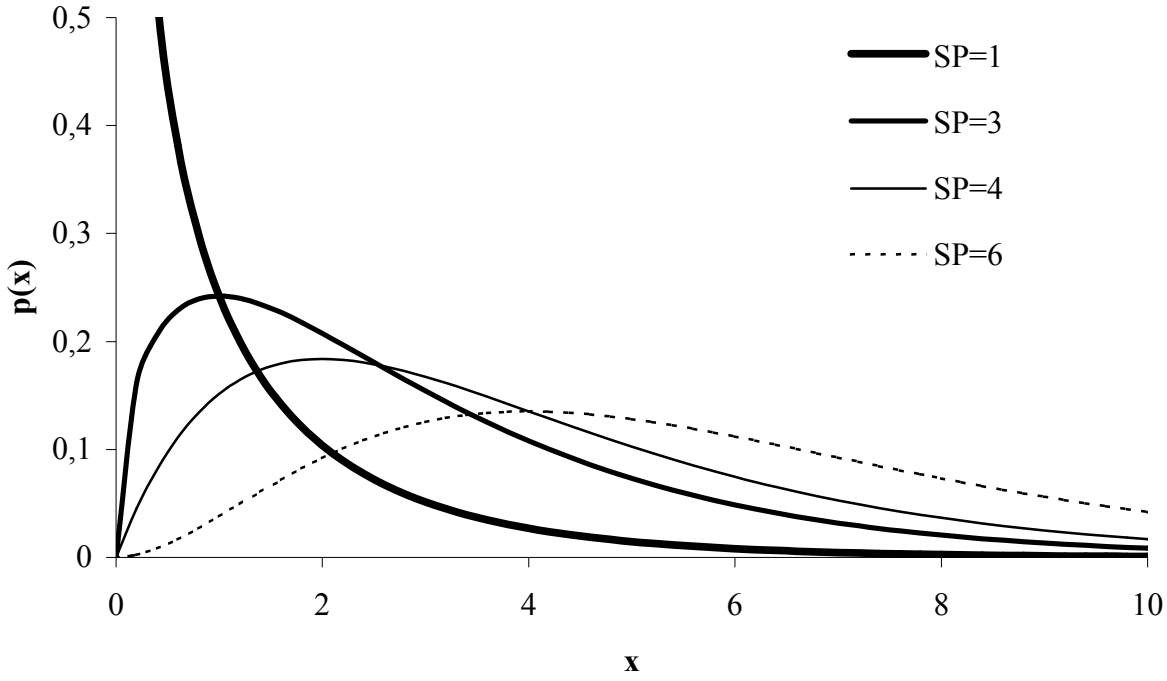
Izrek

Slučajna spremenljivka X je porazdeljena normalno s povprečno vrednostjo μ *in standardnim odklonom* σ. *Na populaciji vzorcev velikosti n je porazdelitev za* S^2 *podana s* χ^2*-porazdelitvijo z n-1 stopinjami prostosti takole:*

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2 \ (SP = n-1)$$

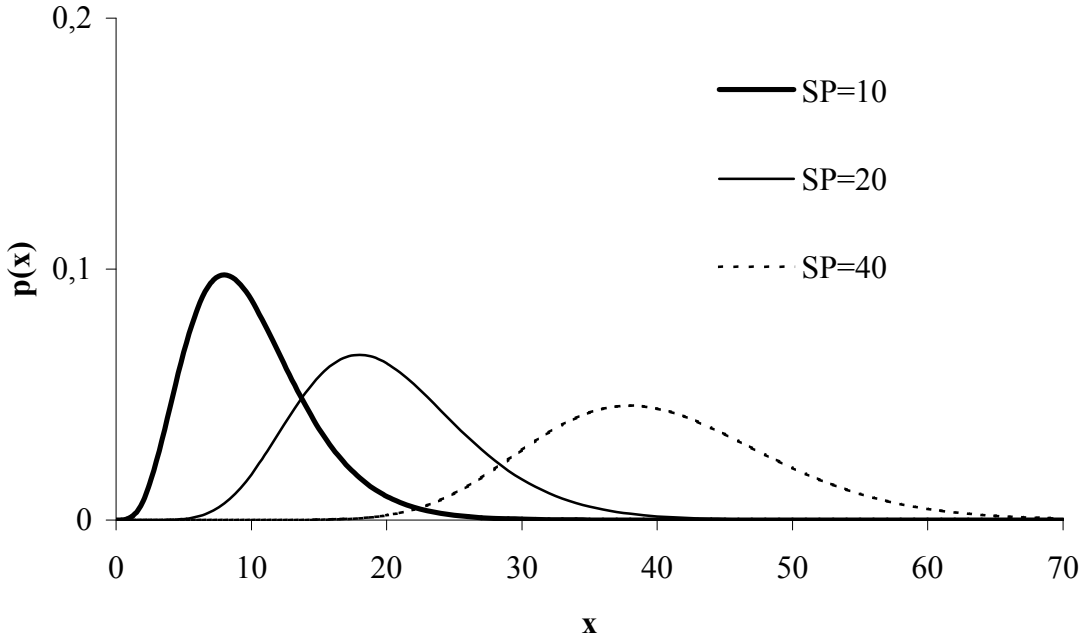
Lastnosti χ^2 **-porazdelitve** so:

- je zvezna porazdelitev, definirana na pozitivnem delu realne osi;
- parameter χ^2-porazdelitve imenujemo *stopinje prostosti*, $SP = 1, 2, ..., \infty$. Le-te določajo obliko porazdelitve. Pri $SP = 1$ in $SP = 2$ ima porazdelitev posebno obliko;
- ko je *SP* majhno število, je porazdelitev asimetrična v desno; ko se število *SP* povečuje, se asimetrija zmanjšuje;



Slika 3-19: χ^2-porazdelitve s stopinjami prostosti 1, 3, 4 in 6

- s povečevanjem $SP \to \infty$, postaja χ^2-porazdelitev čedalje bolj podobna normalni porazdelitvi $N(SP, \sqrt{2SP})$. Slika kaže ta proces.



Slika 3-20: χ^2-porazdelitve s stopinjami prostosti 10, 20 in 40

Vrednosti χ^2-porazdelitve so tabelirane, v prilogi je Tabela 4. Organizirana je tako kot tabela za Studentovo porazdelitev.

NALOGE

1. Porazdelitev vzorčnih aritmetičnih sredin

Slučajna spremenljivka X je porazdeljena po $N(0, 1)$.

a) Kako imenujemo to porazdelitev? Napišite in narišite gostoto verjetnosti za to porazdelitev.

b) Kakšna je pripadajoča porazdelitev vzorčnih aritmetičnih sredin v vzorcih velikosti 16?

c) Kakšna je pripadajoča porazdelitev vzorčnih aritmetičnih sredin v vzorcih velikosti 36?

2. Maslo

Mlekarna pakira zavitke masla, za katere velja predpis, da je masa normalno porazdeljena s parametroma $\mu = 250$ g in $\sigma = 10$ g. Privzemimo, da se proizvodnja sklada s predpisom.

a) Kolikšen odstotek zavitkov bo imel maso vsaj 242 g?

b) Mlekarna dnevno kontrolira zavitke tako, da s slučajno izbiro vzame kontrolni vzorec s 25 zavitki, vsak zavitek stehta, iz podatkov pa izračuna vzorčno aritmetično sredino. Kolikšen odstotek kontrolnih vzorcev bo imel vzorčno aritmetično sredino med 248 g in 252,5 g?

3. Avtomat

V tovarni imajo avtomat, ki razdeljuje neko snov v stekleničke, in sicer je norma 100 mg na stekleničko. V resnici odmerki niso vsi enaki, temveč se zaradi slučajnih vplivov razlikujejo. Predpostaviti smemo, da je odmerek normalno porazdeljena slučajna spremenljivka. V tovarni imajo delovanje stroja za zadovoljivo, če ima ta slučajna spremenljivka povprečno vrednost 100 mg in standardni odklon ne večji od 2 mg.

a) Če velja $\mu = 100$ mg in $\sigma = 2$ mg, za koliko odstotkov stekleničk lahko pričakujemo, da bodo vsebovale več kot 103 mg snovi?

b) V tovarni preverjajo delovanje stroja tako, da s slučajno izbiro vsak dan vzamejo kontrolni vzorec, ki vsebuje 16 stekleničk. Stehtajo vsebino v vsaki steklenički, nato pa iz teh vrednosti izračunajo vzorčno aritmetično sredino. Če velja $\mu = 100$ mg in $\sigma = 2$ mg, za koliko odstotkov vzorcev velikosti 16 lahko pričakujemo, da bo imelo vzorčno aritmetično sredino več kot 103 mg?

c) Nekega dne so dobili v vzorcu naslednje vrednosti (v mg):

99,2	95,8	99,0	101,3	102,1	100,3	98,4	97,6
103,2	101,9	100,5	99,5	101,2	104,2	100,4	99,4

Za ta vzorec izračunajte aritmetično sredino in standardni odklon in ju primerjajte s parametroma normalne porazdelitve.

4. Mlekarna

Predpostavimo, da je vsebnost beljakovin v mleku, ki ga mlekarna odkupuje od svojih proizvajalcev, normalno porazdeljena s povprečno vrednostjo 3,15% in standardnim odklonom 0,30%.

V kontroli kakovosti mleka mlekarna od vsakega proizvajalca vzame po 4 stekleničke mleka in za vsako stekleničko ugotovi vsebnost beljakovin. Iz teh štirih števil izračuna vzorčno aritmetično sredino.

Mlekarna od proizvajalcev mleka, pri katerih je vzorčna aritmetična sredina:

- manjša od 3,00%, odkupuje mleko po znižani ceni;
- med 3,00% in 3,20%, odkupuje mleko po normalni ceni;
- večja od 3,20%, odkupuje mleko po zvišani ceni.

Izračunajte, od koliko % proizvajalcev bo mlekarna odkupovala mleko po znižani, normalni oz. zvišani ceni.

5. Pesticid

Koncentracija določene snovi v tekočem pesticidu ne sme preseči 12 ppm, nad tem pragom postane pesticid toksičen. Učinek pesticida pa je izgubljen, če je koncentracija te snovi manjša od 3 ppm. Predpostavimo, da proizvajalec pošilja pesticid v posodah, v katerih je koncentracija te snovi normalno porazdeljena s povprečno vrednostjo 8 ppm in standardnim odklonom 1,5 ppm.

a) Kolikšen % pošiljk presega dovoljeni maksimum ?

b) Kolikšen % pošiljk ima koncentracijo pod sprejemljivim pragom?

6. *Volitve

V populaciji volilnih upravičencev je 30% neopredeljenih. Kolikšna je verjetnost, da bo v vzorcih velikosti 200 manj kot 40 neopredeljenih?

7. *Križanja

Verjetnost, da pri križanju sort A in B dobimo sorto A, je 0,25. Kolikšna je verjetnost, da dobimo sorto A pri več kot 26% križanj, če:

a) izvedemo 100 križanj?

b) izvedemo 1000 križanj?

4 OSNOVE STATISTIČNEGA SKLEPANJA

V okviru tega poglavja si bomo ogledali dve temi: ocenjevanje parametrov in preizkušanje statističnih domnev.

Kot **parameter** bomo pojmovali lastnost verjetnostne porazdelitve. Npr. normalna porazdelitev $N(\mu,\sigma)$ ima dva parametra, μ in σ, ki natančno določata njeno obliko. Ta dva parametra sta vsebinsko zanimiva, parameter μ je povprečje pripadajoče slučajne spremenljivke, parameter σ pa njen standardni odklon.

4.1 *OCENJEVANJE PARAMETROV

V splošnem obstajata dva načina ocenjevanja neznanih količin:

- **točkovno ocenjevanje**: ena vrednost ocenjuje neznano količino;
- **intervalno ocenjevanje**: dve vrednosti določata interval, ki ocenjuje neznano količino.

Primer: merjenje hitrosti

- izmerjena hitrost je 5,4 m/s. To je točkovna ocena hitrosti;
- izmerjena hitrost je $5{,}4 \mp 0{,}05$ m/s. To je intervalna ocena hitrosti. Intervalna ocena podaja tudi neke vrste mero za napako meritve.

Tudi v statistiki poznamo točkovno in intervalno ocenjevanje parametrov, le da se intervalno ocenjevanje po vsebini bistveno razlikuje od npr. fizikalnega ocenjevanja.

4.1.1 *Točkovna ocena parametra

Za točkovno oceno parametra so zaželene določene matematične lastnosti. Ena najpomembnejših je nepristranskost.

*Ocena je **nepristranska**, če je povprečje vseh vzorčnih ocen v vzorcih določene velikosti n enako ocenjevanemu parametru.* Spoznali smo že nekaj nepristranskih ocen.

Če $X \sim N(\mu,\sigma)$, potem je vzorčna aritmetična sredina $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

nepristranska ocena za μ, saj velja $E(\bar{X}) = \mu$. Teorija pokaže, da je vzorčna varianca s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

nepristranska ocena za σ^2, saj velja $E(S^2) = \sigma^2$. Prav zahteva po nepristranskosti je narekovala, da se vzorčna varianca izračuna tako, da je v imenovalcu $(n-1)$ in ne n. Vzorčni standardni odklon s pa je pristranska ocena parametra σ, saj teorija pokaže, da $E(S) \neq \sigma$.

Primer

Stroj polni stekleničke z zdravilom. Predpostavljati smemo, da je masa zdravila v stekleničkah normalno porazdeljena, $X \sim N(\mu,\sigma)$. V vzorcu je 9 stekleničk. V njih so ugotovili naslednje mase zdravila (mg):

10,8 9,0 10,1 10,9 10,1 11,0 9,8 11,6 11,2

Iz podatkov izračunamo:

$\bar{x} = 10{,}50$ mg je točkovna ocena za povprečno maso zdravila μ. Ta ocena je nepristranska.

$s^2 = 0{,}6575$ mg^2 je točkovna ocena za varianco mase σ^2. Ta ocena je nepristranska.

$s = 0{,}811\,\text{mg}$ je točkovna ocena za standardni odklon mase σ. Ta ocena je pristranska.

Če $X \sim b(n, p)$, potem je vzorčni delež $\hat{p}$

$$\hat{p} = \frac{x}{n}$$

nepristranska ocena za Bernoullijev *p*, saj velja:

$$E\left(\frac{X}{n}\right) = p$$

Primer

V slučajnem vzorcu je 200 semen sorte A, od teh je kalilo 120 semen. Vzorčni delež

$\hat{p} = \frac{120}{200} = 0{,}60$ je točkovna ocena za verjetnost *p*, da seme kali. Ta ocena je nepristranska.

Poleg nepristranskosti obstajajo še druge pomembne lastnosti točkovnih ocen, ki jih ne bomo obravnavali.

Slabost točkovne ocene je njena nezanesljivost, saj ne vemo, na katerega od možnih vzorcev smo naleteli oz. kje v porazdelitvi vzorčnih ocen leži dobljena točkovna ocena. Zato pogosto točkovno oceno parametra dopolnimo z intervalno oceno. Ta se izraža z intervalom, ki z vnaprej predpisano verjetnostjo pokriva vrednost parametra. Poglejmo bolj natančno njegovo opredelitev.

4.1.2 *Intervalna ocena parametra

Intervalna ocena parametra je t. i. **interval zaupanja**. To je slučajni interval, vezan na pripadajoči slučajni vzorec. Poglejmo definicijo intervala zaupanja:

Naj θ označuje parameter, ki ga ocenjujemo, vrednost α je vnaprej predpisana verjetnost, $0 < \alpha < 1$. Interval (L_1, L_2) imenujemo interval zaupanja za parameter θ, če velja:

$$P(L_1 < \theta < L_2) = 1 - \alpha$$

Komentar

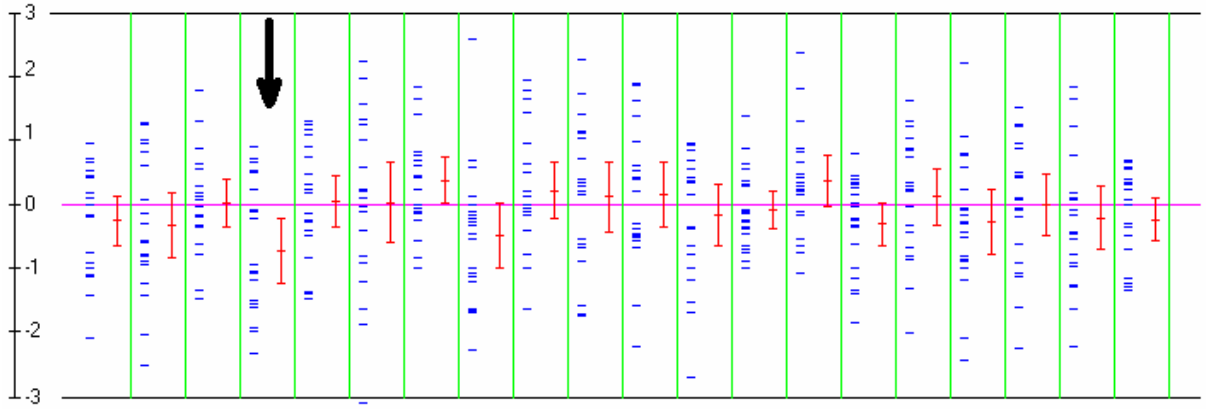
- Standardne vrednosti, ki jih uporabljamo za verjetnost α, so: 0,05; 0,01 ali 0,001. Verjetnost $1 - \alpha$ imenujemo **zaupanje**. Standardne vrednosti za zaupanje torej so: 0,95; 0,99 oz. 0,999. Običajno izražamo zaupanje v odstotkih, govorimo npr. o 95-odstotnem (95%) intervalu zaupanja.
- L_1 oz. L_2 je spodnja oz. zgornja meja intervala zaupanja. L_1 in L_2 sta slučajni spremenljivki: pri vsakem vzorcu imata drugo vrednost. Drugače povedano: vsak slučajni vzorec velikosti *n* generira svoj interval zaupanja (l_1, l_2). V populaciji vseh vzorcev velikosti *n* je odstotek intervalov zaupanja, ki vsebujejo parameter θ, enak $100 \cdot (1 - \alpha)$. Za posamezni interval zaupanja ne vemo, ali je θ vsebovan v tem intervalu ali ne. Trdimo lahko le, da je ta interval z verjetnostjo $(1 - \alpha)$ eden tistih, ki vsebujejo parameter θ.

Primer

Ocenjujemo parameter μ iz $N(\mu, \sigma)$. Naredimo 100 vzorcev določene velikosti *n*. Iz podatkov vsakega vzorca izračunamo interval zaupanja za μ pri verjetnosti $\alpha = 0{,}05$.

Pričakujemo, da bo izmed 100 intervalov zaupanja 95 takih, ki pokrijejo neznani parameter μ, in 5 takih, ki ga ne pokrijejo. Za posamezni interval zaupanja ne vemo, ali vsebuje μ ali ne.

Slika ilustrira to dogajanje na konkretnih podatkih. Število vzorcev je 20, podatki (vodoravne črtice) so generirani iz N(0,1). Ocenjujemo povprečje, ki je 0. Izmed 20 intervalov zaupanja (navpične črte) pri stopnji zaupanja 0,95 je 19 takih, ki pokrivajo vrednost 0. En interval (označen s puščico) povprečja ne pokriva.



Slika 4-1: Intervali zaupanja za povprečje normalne porazdelitve

---

Intervalno ocenjevanje v statistiki je verjetnostno ocenjevanje, zato moramo biti pri interpretaciji intervalov zaupanja dovolj pazljivi. Izračun intervala zaupanja za posamezni parameter temelji na verjetnostni porazdelitvi njegovih vzorčnih ocen. Poglejmo izpeljave intervalov zaupanja za nekaj parametrov.

4.1.2.1 *Interval zaupanja za povprečno vrednost

Izračun intervala zaupanja za povprečno vrednost E(X) temelji na porazdelitvi vzorčnih aritmetičnih sredin $\bar{x}$ v populaciji vzorcev velikosti n. Ponovimo teorijo.

a) Če $X \sim N(\mu,\sigma)$ in je σ znana, je $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

b) Če $X \sim N(\mu,\sigma)$ in σ ni znana, je $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(SP = n-1)$

c) Če so vzorci dovolj veliki, je $\overline{X} \sim N(E(X), \sqrt{\dfrac{Var(X)}{n}})$ ne glede na porazdelitev izhodiščne spremenljivke X.

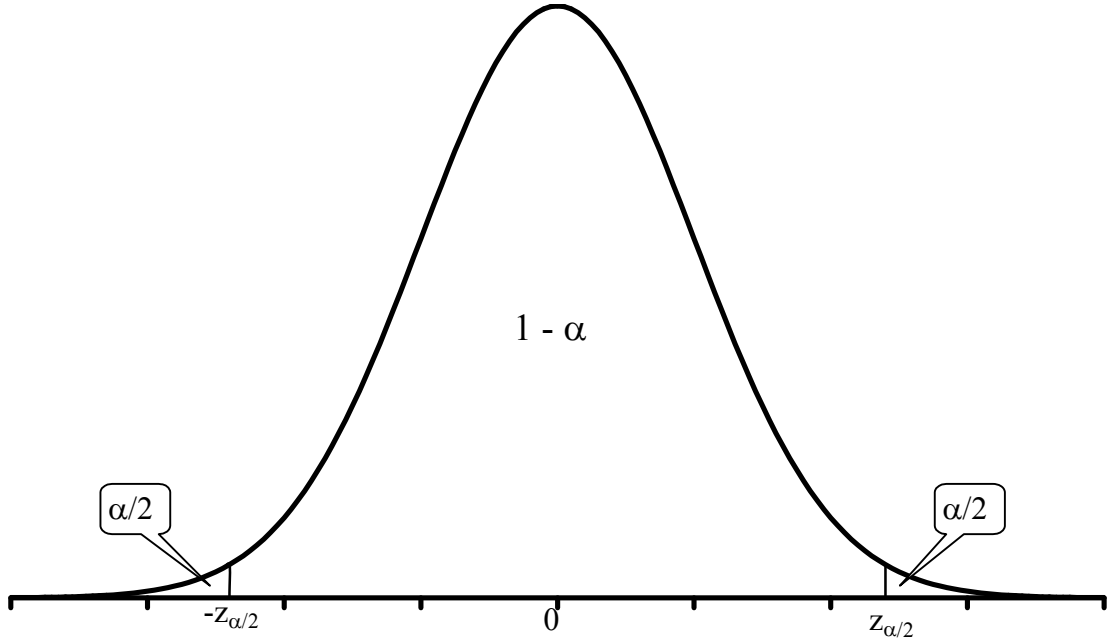
Nakazali bomo izpeljavo intervala zaupanja za a). Ta situacija v praksi zelo redko nastopa, vendar je za razumevanje konstrukcije intervala zaupanja najlažja.

a) $X \sim N(\mu,\sigma)$ in je σ znana

S pomočjo tabel za standardizirano normalno porazdelitev N(0, 1) določimo vrednosti $\mp z_{\frac{\alpha}{2}}$, da velja:

$P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$

Slika ilustrira to konstrukcijo.



Slika 4-2: Izpeljava intervala zaupanja za povprečno vrednost za primer, ko je σ znana

Za Z ustavimo gornji izraz

$$P(-z_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

rešimo dve neenačbi in dobimo:

$$P(\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Iz tega izraza odčitamo meji intervala zaupanja za μ:

$$l_1 = \overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$l_2 = \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Zapišimo to krajše v obliki:

$$l_{1,2} = \overline{x} \mp z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Primer

Izračunajmo 90% in 95% interval zaupanja za povprečno maso zdravila v stekleničkah, pri čemer upoštevajmo, da je: $\sigma = 1\,\text{mg}$.

90% interval zaupanja za povprečno maso zdravila v stekleničkah izračunamo iz naslednjih vrednosti:

$\overline{x} = 10{,}50$; $z_{0{,}05} = 1{,}645$; $\sigma = 1$; $n = 9$

$$l_{1,2} = 10{,}5 \mp 1{,}645 \cdot \frac{1}{\sqrt{9}} = 10{,}5 \mp 0{,}55$$

$l_1 = 9{,}95$ mg $\quad\quad$ $l_2 = 11{,}05$ mg

Obrazložitev: Interval (9,95 mg; 11,05 mg) je z verjetnostjo 0,90 eden tistih, ki vsebujejo vrednost za povprečno maso zdravila v stekleničkah.

95% interval zaupanja za povprečno maso zdravila v stekleničkah izračunamo takole:

$\bar{x} = 10{,}50$; $z_{0,025} = 1{,}960$; $\sigma = 1$; $n = 9$

$$l_{1,2} = 10{,}5 \mp 1{,}960 \cdot \frac{1}{\sqrt{9}} = 10{,}5 \mp 0{,}653$$

$l_1 = 9{,}85$ mg $\quad\quad$ $l_2 = 11{,}15$ mg

---

b) $X \sim N(\mu, \sigma)$ in σ ni znana

Standardni odklon σ ni podan, ampak ga ocenimo iz podatkov. Interval zaupanja izpeljemo na enak način kot pod a), le da standardizirano normalno porazdelitev nadomesti Studentova porazdelitev z *n*-1 stopinjami prostosti:

$$l_{1,2} = \bar{x} \mp t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$$

Formula za izračun intervala zaupanja je po obliki enaka kot pod a).

Primer

Izračunajmo 95% in 99% interval zaupanja za povprečno maso zdravila v stekleničkah.

95% interval zaupanja za povprečno maso zdravila v stekleničkah izračunamo iz naslednjih vrednosti:

$\bar{x} = 10{,}50$; $t_{0,025}(8) = 2{,}306$; $s = 0{,}811$; $n = 9$

$$l_{1,2} = 10{,}5 \mp 2{,}306 \cdot \frac{0{,}811}{\sqrt{9}} = 10{,}5 \mp 0{,}62$$

$l_1 = 9{,}88$ mg $\quad\quad$ $l_2 = 11{,}13$ mg

Obrazložitev: Interval (9,9 mg; 11,1 mg) je z verjetnostjo 0,95 eden tistih, ki vsebujejo vrednost za povprečno maso zdravila v stekleničkah.

99% interval zaupanja za povprečno maso zdravila v stekleničkah:

$\bar{x} = 10{,}50$; $t_{0,005}(8) = 3{,}355$; $s = 0{,}811$; $n = 9$

$$l_{1,2} = 10{,}5 \mp 3{,}355 \cdot \frac{0{,}811}{\sqrt{9}} = 10{,}5 \mp 0{,}91$$

$l_1 = 9{,}59$ mg $\quad$ $l_2 = 11{,}41$ mg

Mimogrede ugotovimo: večje ko je zaupanje, širši je interval zaupanja.

---

c) Veliki vzorci.

Če so vzorci tako veliki, da velja centralni limitni izrek, izračunamo interval zaupanja za povprečno vrednost takole:

$$l_{1,2} = \bar{x} \mp z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Primer

Raziskovalci so izbrali 100 gomoljev semenskega krompirja in jih stehtali. Iz teh podatkov so izračunali vzorčno aritmetično sredino, ki je bila 58,9 g, in vzorčni standardni odklon, ki je bil 10,0 g. Izračunajmo 95% interval zaupanja za povprečno maso gomolja semenskega krompirja.

$$l_{1,2} = 58{,}9 \mp 1{,}960 \cdot \frac{10}{\sqrt{100}} = 58{,}9 \mp 1{,}960$$

$$l_1 = 56{,}9 \text{ g} \qquad l_2 = 60{,}9 \text{ g}$$

Pri 95% zaupanju interval (56,9 g; 60,9 g) vsebuje vrednost za povprečno maso gomolja semenskega krompirja.

Opomba: v tem primeru predpostavka o normalni porazdelitvi mase gomoljev ni bila potrebna, saj je vzorec dovolj velik.

*4.1.2.1.1 *Velikost vzorca*

Interval zaupanja za povprečno vrednost lahko zapišemo v obliki:

$$l_{1,2} = \bar{x} \mp \Delta$$

Meji intervala zaupanja za povprečno vrednost sta simetrični okoli ocene $\bar{x}$. Δ imenujemo **odklon zaupanja**. Ta količina je bistvena informacija, ki jo dobimo iz intervala zaupanja. Poglejmo si njen pomen. Pri zaupanju $(1-\alpha)$ je odklon zaupanja največja možna razlika med $\bar{x}$ in med μ, torej med oceno za povprečno vrednost $\bar{x}$ in njeno pravo vrednostjo μ.

Izračun odklona zaupanja je odvisen od tega, ali imamo velike ali majhne vzorce.

Veliki vzorci	Majhni vzorci
$\Delta = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$	$\Delta = t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$

Na širino odklona zaupanja vpliva

- zaupanje $(1-\alpha)$: večje zaupanje pomeni širši odklon zaupanja. Na to vplivamo sami, saj α določimo sami vnaprej;
- variabilnost proučevane spremenljivke, ki jo izraža *s*. Večja variabilnost pomeni širši odklon zaupanja;
- število enot v vzorcu *n*. Število enot je v izrazu $\frac{1}{\sqrt{n}}$ za velike vzorce oz. v izrazu $t_{\frac{\alpha}{2}}(n-1) \cdot \frac{1}{\sqrt{n}}$ za majhne vzorce. Več enot v vzorcu pomeni ožji odklon zaupanja. Zveza ni linearna. Če želimo, da se širina odklona zaupanja prepolovi, moramo zvečati število enot v vzorcu vsaj 4-krat.

V izrazu za odklon zaupanja je izraz

$$s(\bar{x}) = \frac{s}{\sqrt{n}}$$

ki ga imenujemo **standardna napaka ocene** $\bar{x}$, označimo jo $s(\bar{x})$. Kot je iz povedanega razvidno, le-ta vpliva na širino odklona zaupanja. Standardna napaka je mera za natančnost vzorčne ocene. Večja standardna napaka pomeni slabšo natančnost vzorčne ocene in obratno.

Najpogostejše vprašanje, ki ga raziskovalci postavijo statistikom, se glasi: 'Kako velik vzorec potrebujemo?' Odgovor na vprašanje o **velikosti vzorca** je odvisen od tega, kolikšno natančnost vzorčne ocene želimo. Za izračun potrebne velikosti vzorca moramo:

- sami vnaprej izbrati vrednost za zaupanje $(1-\alpha)$;
- sami vnaprej določiti največjo še sprejemljivo vrednost za odklon zaupanja Δ;
- moramo dobiti oceno variabilnosti proučevane spremenljivke, ki jo izrazimo s standardnim odklonom *s*. To oceno običajno poznamo iz predhodnih raziskav ali iz t. i. 'tipalnih' proučevanj.

Iz vrednosti $(1-\alpha)$, Δ ter *s* izračunamo potrebno velikost vzorca *n*. Izračunavanje poteka v več korakih. Najprej izračunamo začetni približek n_0 iz zveze:

$$n_0 = (z_{\frac{\alpha}{2}} \cdot \frac{s}{\Delta})^2$$

V naslednjem koraku upoštevamo to oceno in izračunamo nov približek n_1 iz zveze:

$$n_1 = (t_{\frac{\alpha}{2}}(n_0 - 1) \cdot \frac{s}{\Delta})^2$$

Postopek nadaljujemo toliko korakov, da se velikost vzorca ne spreminja več. Običajno se postopek po nekaj korakih ustavi.

Primer

Koliko žarnic moramo vključiti v poskus, če želimo pri 95% zaupanju oceniti povprečno življenjsko dobo žarnic tako, da se bo ocena za povprečno življenjsko dobo ločila od prave vrednosti za povprečno življenjsko dobo za največ 5 ur? Za življenjsko dobo žarnic lahko privzamemo normalno porazdelitev, ocena za njen standardni odklon je 10 ur.

Iz besedila razberemo naslednje količine:

$1-\alpha = 0{,}95$

$\Delta = 5$ h

$s = 10$ h

Začetni približek za število žarnic je:

$$n_0 = (z_{\frac{\alpha}{2}} \cdot \frac{s}{\Delta})^2 = (1{,}96 \cdot \frac{10}{5})^2 = 15{,}37 = 16$$

Nadaljnji koraki:

$$n_1 = (t_{\frac{\alpha}{2}}(15) \cdot \frac{s}{\Delta})^2 = (2{,}131 \cdot \frac{10}{5})^2 = 18{,}16 = 19$$

$$n_2 = (t_{\frac{\alpha}{2}}(18) \cdot \frac{s}{\Delta})^2 = (2{,}101 \cdot \frac{10}{5})^2 = 17{,}66 = 18$$

$$n_3 = (t_{\frac{\alpha}{2}}(17) \cdot \frac{s}{\Delta})^2 = (2{,}110 \cdot \frac{10}{5})^2 = 17{,}81 = 18$$

Za poskus potrebujemo 18 žarnic.

4.1.2.2 *Interval zaupanja za varianco in za standardni odklon

Naj bo $X \sim N(\mu, \sigma)$. Nepristranska ocena za σ^2 je vzorčna varianca:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Iz poglavja o porazdelitvi vzorčnih varianc vemo, da velja:

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2\ (n-1)$$

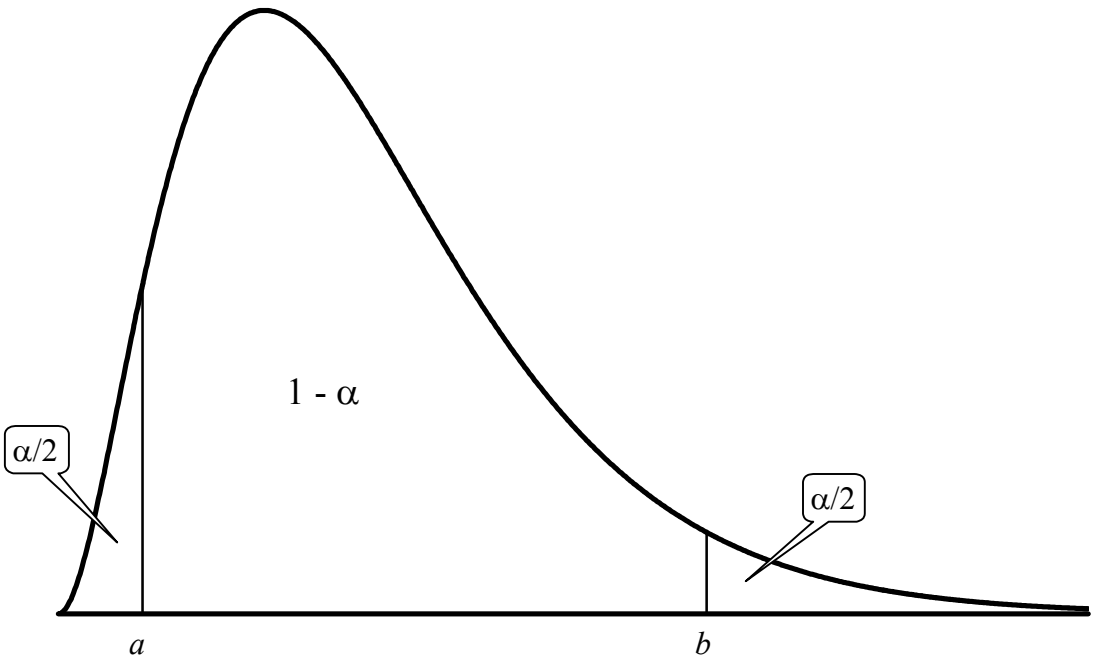
Pri predpisanem zaupanju $1-\alpha$ konstruiramo interval zaupanja za σ^2 tako, da velja

$$P(a \le \frac{n-1}{\sigma^2}S^2 \le b) = 1-\alpha$$

Vrednosti *a* in *b* odčitamo iz tabele za χ^2-porazdelitev (Tabela 4 iz priloge):

$$a = \chi^2_{1-\frac{\alpha}{2}}(n-1)$$

$$b = \chi^2_{\frac{\alpha}{2}}(n-1)$$



Slika 4-3: Izpeljava intervala zaupanja za varianco

Rešimo dve neenačbi in dobimo

$$P(\frac{n-1}{b}S^2 \le \sigma^2 \le \frac{n-1}{a}S^2) = 1-\alpha$$

Meji intervala zaupanja za σ^2 sta:

$$l_1 = \frac{n-1}{b}\cdot s^2 \qquad\qquad l_2 = \frac{n-1}{a}\cdot s^2$$

Meji intervala zaupanja za σ^2 nista simetrični okoli ocene s^2. Iz tega izraza dobimo meji intervala zaupanja za σ:

$$l_1 = \sqrt{\frac{n-1}{b}} \cdot s \qquad l_2 = \sqrt{\frac{n-1}{a}} \cdot s$$

Tudi meji intervala zaupanja za σ nista simetrični okoli ocene *s*.

Primer

Izračunajmo 95% interval zaupanja za standardni odklon mase zdravila v stekleničkah. Za izračun potrebujemo naslednje vrednosti:

$s = 0{,}811$; $n = 9$; $a = \chi^2_{0{,}975}(8) = 2{,}180$; $b = \chi^2_{0{,}025}(8) = 17{,}535$

$$l_1 = \sqrt{\frac{8}{17{,}535}} \cdot 0{,}811 = 0{,}548$$

$$l_2 = \sqrt{\frac{8}{2{,}180}} \cdot 0{,}811 = 1{,}553$$

Pri 95% zaupanju interval (0,55 mg; 1,55 mg) pokrije vrednost za standardni odklon σ.

4.1.2.3 *Interval zaupanja za Bernoullijevo verjetnost

Naj bo $X \sim b(n, p)$ in naj velja, da lahko binomsko porazdelitev aproksimiramo z $N(np, \sqrt{npq})$. Potem je

$$\frac{X}{n} \approx N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Interval zaupanja za *p* izpeljemo na osnovi normalne porazdelitve tako, kot smo izpeljali interval zaupanja za μ za primer a) in dobimo:

$$l_{1,2} = \hat{p} \mp z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{pq}{n}}$$

Žal pa je neznani *p* tudi pod korenom, zato intervala zaupanja ne moremo izračunati. Preostane nam le to, da vrednost *p* pod korenom nadomestimo z oceno $\hat{p}$, ki jo izračunamo iz vzorca. Tako dobimo približni interval zaupanja za *p*:

$$l_{1,2} \approx \hat{p} \mp z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Teorija pove, da je ta približek utemeljen, če je število enot v vzorcu vsaj 40.

Primer

V slučajnem vzorcu je 200 semen sorte A, od teh je kalilo 120 semen. Točkovna ocena za kalivost je 0,60 oz. 60%. Izračunajmo 95% interval zaupanja za verjetnost, da seme kali.

$$l_{1,2} = 0{,}60 \mp 1{,}96 \cdot \sqrt{\frac{0{,}60 \cdot 0{,}40}{200}} = 0{,}60 \mp 0{,}068$$

$l_1 = 0{,}532$ oz. 53,2%

$l_2 = 0{,}668$ oz. 66,8%

*4.1.2.3.1 *Velikost vzorca*

Potrebno velikost vzorca n izračunamo iz odklona zaupanja Δ za Bernoullijev p. Vrednost za zaupanje in za odklon zaupanja Δ določi uporabnik.

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{\Delta} \right)^2 \cdot p \cdot q = \left(\frac{z_{\frac{\alpha}{2}}}{\Delta} \right)^2 \cdot p \cdot (1 - p)$$

Ker vrednosti za Bernoullijev p ne poznamo, poiščemo maksimum zgornjega izraza. Ta je pri $p = q = 0{,}5$. Zato velja:

$$n_{\max} = \left(\frac{z_{\frac{\alpha}{2}}}{\Delta} \right)^2 \cdot \frac{1}{4}$$

Če imamo vsaj grobo oceno za $\hat{p}$, uporabimo njeno vrednost v izrazu za n.

Primer

Pred volitvami nas zanima odstotek neopredeljenih v populaciji. Koliko volilnih upravičencev moramo anketirati, če želimo pri 95% zaupanju oceniti odstotek neopredeljenih v populaciji, tako da bo razlika med oceno in pravo vrednostjo največ 5 odstotnih točk?

$$n_{\max} = \left(\frac{1{,}96}{0{,}05} \right)^2 \cdot \frac{1}{4} = 385$$

Za tolikšno natančnost potrebujemo v vzorcu 385 oseb.

Recimo, da na osnovi predhodnih anket ugotovimo, da je v populaciji približno 20% neopredeljenih. Pri zahtevi, da je razlika med oceno in pravo vrednostjo za verjetnost največ 0,05, izračunajmo potrebno velikost vzorca.

$$n = \left(\frac{1{,}96}{0{,}05} \right)^2 \cdot 0{,}80 \cdot 0{,}20 = 246$$

V tem primeru moramo anketirati vsaj 246 oseb. Dodatna informacija je omogočila, da se je potrebno število anketirancev zmanjšalo iz 385 na 246.

NALOGE

1. *Uspeh pri izpitih

Izračunajte 95 % interval zaupanja za povprečno vrednost in za standardni odklon za število točk pri maturi. Podatki za vzorec 28 maturantov so:

14	11	19	25	18	12	15
16	21	20	20	30	16	18
18	17	18	15	18	24	18
12	30	18	11	18	18	24

2. *Količina padavin

V vzorcu je 67 meteoroloških postaj. Navajamo frekvenčno porazdelitev za količino padavin, izmerjeno na teh postajah v letu 1992.

Tabela 4-1: Meteorološke postaje po količini padavin (Vir: Arhiv Hidrometeorološki zavod Slovenije)

Količina padavin (mm)	Št. postaj
800 do pod 1200	12
1200 do pod 1600	27
1600 do pod 2000	16
2000 do pod 2400	7
2400 do pod 2800	4
2800 do pod 3200	0
3200 do pod 3600	1

Izračunajte 95% interval zaupanja za povprečno vrednost in za standardni odklon količine padavin.

3. *Prirast telet

V poskus je bilo vključenih 25 telet. Ocena za povprečno težo telet v 7. tednu starosti je bila 72,0 kg. Ocena variance prirasta telet med 7. tednom in 11. tednom je znašala 121 kg^2, relativna variabilnost prirasta je bila ocenjena z vrednostjo 36%.

a) Izračunajte 95% interval zaupanja za povprečni prirast med sedmim in enajstim tednom in ga obrazložite.

b) Koliko so v povprečju tehtala teleta ob dopolnjenem enajstem tednu starosti?

c) Koliko telet bi morali vključiti v poskus, da bi se širina intervala zaupanja za povprečni prirast zmanjšala za polovico?

4. *Računalnik

Računalnik generira vrednosti slučajne spremenljivke *X*, ki je normalno porazdeljena, vrednosti parametrov μ in σ pa sta znani le programerju. Programer opredeli velikost vzorca *n* in vsakemu od 55 študentov generira *n* vrednosti spremenljivke *X*. Iz teh vrednosti vsak študent izračuna 95% interval zaupanja za μ.

a) Za koliko od teh 55 intervalov zaupanja pričakujemo, da vsebujejo vrednost μ?

b) Ali je odstotek intervalov zaupanja, ki vsebujejo vrednosti μ, odvisen od števila enot v vzorcu *n*? Utemeljite odgovor.

c) Podatki za 55 vzorcev velikosti 12 so:

Vzorec št.	Vzorčne vrednosti:											
1	98,5	96,3	124,0	108,0	100,5	94,8	102,1	113,9	81,7	91,7	100,5	97,1
2	92,4	77,8	98,9	98,4	105,3	92,1	89,9	101,5	100,3	102,1	90,8	93,4
3	109,3	109,6	89,1	96,7	96,4	97,1	93,0	93,2	104,1	117,9	88,1	106,3
4	90,7	89,9	114,4	104,9	96,6	107,3	91,9	102,5	97,7	105,7	105,9	95,3
5	104,6	102,2	107,5	92,5	113,9	95,7	83,7	100,9	76,7	91,7	94,6	92,8
6	111,3	84,0	108,7	94,5	106,1	89,5	105,5	112,0	90,0	116,8	104,5	96,4
7	110,3	94,6	100,8	85,5	97,0	98,2	105,3	129,3	99,3	103,1	99,9	114,7
8	108,2	91,5	116,2	92,5	100,2	108,3	104,9	87,3	106,1	82,0	98,8	83,2
9	96,6	106,2	101,4	94,8	100,2	98,8	88,2	99,5	107,4	96,8	99,5	96,9
10	113,9	97,0	129,6	93,8	97,0	111,4	104,2	110,9	96,2	115,9	120,1	111,9
11	91,5	115,8	101,7	91,6	92,7	97,1	99,9	78,8	90,5	99,3	101,0	84,8
12	90,6	98,7	105,3	100,0	119,6	89,8	98,0	93,7	99,6	99,6	93,3	91,4
13	95,0	113,9	122,8	96,7	98,7	99,5	100,6	110,4	89,3	117,5	96,6	89,1
14	103,3	84,8	93,1	106,0	91,0	98,5	117,1	94,4	101,8	103,5	98,4	103,5

15	99,4	101,4	91,2	95,9	79,6	109,6	106,3	91,6	97,8	108,7	91,8	99,6
16	96,8	95,9	117,6	95,7	108,1	111,3	94,8	108,1	99,3	97,2	109,0	98,6
17	109,7	102,1	111,0	103,5	92,4	91,4	105,1	118,2	105,4	95,2	104,8	118,1
18	114,7	113,3	103,7	102,3	88,2	114,7	96,5	88,0	99,3	99,5	84,0	92,6
19	90,8	115,4	121,1	117,1	116,9	95,5	97,1	114,8	103,2	99,8	85,9	104,7
20	92,1	86,2	92,9	89,3	83,4	95,9	104,8	97,9	104,0	105,9	101,6	113,9
21	83,9	103,3	101,3	91,0	94,9	97,1	85,6	105,5	97,7	98,9	105,4	85,4
22	93,4	80,0	103,3	88,9	109,1	109,4	95,8	92,9	88,5	97,5	97,5	106,8
23	81,7	89,4	99,3	97,4	109,8	96,7	98,4	119,6	103,9	103,5	93,3	115,1
24	103,5	101,8	98,7	105,6	114,1	94,4	85,3	93,1	98,1	89,9	106,4	98,2
25	98,6	112,1	92,9	97,5	118,5	104,6	103,7	102,3	101,8	113,4	104,7	109,7
26	105,1	98,9	111,2	98,4	101,4	99,3	95,6	92,0	99,9	96,8	116,9	94,1
27	89,8	102,2	78,0	102,7	90,3	94,2	107,9	78,8	94,5	101,0	109,3	110,7
28	90,9	106,0	90,0	105,5	92,2	121,9	100,2	96,8	96,7	81,6	91,7	104,5
29	88,1	97,3	113,9	101,5	93,7	111,8	119,5	91,1	109,5	102,5	94,6	81,8
30	100,0	88,8	111,4	98,7	84,4	101,4	111,7	91,8	89,5	100,3	100,0	107,1
31	109,9	107,5	111,5	95,2	112,4	105,7	89,0	73,7	100,2	97,6	92,6	101,1
32	115,3	93,5	96,9	105,8	78,6	108,8	96,4	87,8	102,9	95,0	80,3	98,3
33	107,7	85,9	109,6	93,4	106,4	98,6	116,8	98,1	107,1	104,8	106,5	111,1
34	102,9	110,9	93,0	109,3	118,5	98,1	94,9	81,8	81,6	95,5	108,4	110,8
35	85,0	104,7	100,4	97,5	88,5	99,1	92,6	99,2	77,7	102,6	96,3	97,5
36	96,6	103,1	93,3	96,7	95,5	98,7	102,1	115,4	92,7	95,4	94,8	112,4
37	99,7	99,5	104,8	93,8	102,8	106,8	86,8	110,9	100,1	97,4	87,0	93,7
38	108,7	101,0	93,1	90,9	97,8	88,2	92,6	108,7	117,2	94,1	102,3	116,2
39	101,0	93,9	105,4	96,8	85,0	106,9	93,3	88,5	96,8	104,9	105,8	108,4
40	91,3	92,5	84,0	108,0	81,3	107,3	98,8	94,2	119,8	106,8	93,1	111,6
41	95,9	117,8	97,2	103,8	92,7	91,5	102,3	101,0	99,7	83,6	92,4	108,1
42	111,9	97,2	103,5	95,5	123,3	93,0	122,3	85,6	89,3	98,6	98,3	116,3
43	105,4	107,9	104,8	97,0	93,5	86,7	99,4	104,6	95,2	97,1	106,8	91,4
44	83,2	98,2	101,3	115,2	99,0	105,7	87,3	93,6	94,1	90,4	99,9	97,4
45	105,6	81,7	110,3	113,5	98,5	100,2	104,9	110,0	82,1	96,5	88,4	96,9
46	77,4	84,2	91,2	96,4	98,3	108,6	100,8	85,5	98,9	96,9	106,6	100,1
47	104,5	95,3	86,0	95,2	91,7	92,9	117,2	86,9	128,0	98,1	77,1	81,3
48	99,3	111,6	92,9	94,1	96,4	104,0	93,4	84,9	89,9	108,7	98,0	89,7
49	86,7	86,1	98,8	97,8	106,2	92,6	106,8	94,7	100,7	101,4	95,1	93,4
50	107,9	96,8	87,2	103,1	100,7	96,0	89,1	102,5	88,2	88,5	91,3	96,7
51	106,4	106,0	109,2	93,6	81,6	113,5	110,3	100,4	97,3	92,1	77,7	96,0
52	118,6	85,5	96,0	83,5	101,3	95,9	99,5	112,7	100,9	110,0	89,0	103,8
53	107,7	100,5	81,3	76,3	107,9	86,0	97,2	92,3	91,2	98,2	111,8	103,8
54	114,6	79,6	117,2	102,1	97,5	110,1	111,9	101,0	90,8	66,4	125,4	101,5
55	100,5	99,3	103,9	91,0	83,4	95,4	99,2	83,2	102,1	92,7	100,1	115,8

Za vsak vzorec izračunajte 95% interval zaupanja za povprečno vrednost in za standardni odklon.

d) Podatki so generirani iz normalne porazdelitve N (100, 10). Ugotovite, kateri IZ ne vsebujejo povprečne vrednosti 100 in kateri standardnega odklona 10.

5. *Kroglice za ležaje

Premeri kroglic za ležaje, ki so izdelani z določenim postopkom, so normalno porazdeljeni z znano varianco $\sigma^2 = 0{,}25\ \text{mm}^2$. Kako velik vzorec moramo vzeti, da lahko ocenimo povprečni premer kroglic, tako da se bo ocena razlikovala od prave vrednosti premera za največ 0,05 mm s 95% zaupanjem?

6. *Gripa

Izračunajte 90% in 95% interval zaupanja za verjetnost okužbe z gripo v proučevani populaciji, če je v slučajnem vzorcu velikosti 120 zbolelo 15% oseb.

7. *Volitve

Pred volitvami nas zanima odstotek neopredeljenih v populaciji. Koliko volilnih upravičencev moramo anketirati, če želimo pri 95% zaupanju oceniti odstotek neopredeljenih v populaciji, tako da bo razlika med oceno in pravo vrednostjo največ

a) 1 odstotna točka?
b) 5 odstotnih točk?
c) 10 odstotnih točk?

4.2 PREIZKUŠANJE STATISTIČNIH DOMNEV

4.2.1 Princip preizkušanja statističnih domnev

Raziskovalna domneva (hipoteza) je še nedokazana trditev, ki jo želimo potrditi ali zavrniti z raziskovalnim delom.

Statistična domneva je še nedokazana trditev o lastnosti slučajne spremenljivke. Opredelimo dve statistični domnevi: **ničelno domnevo** H_0 in **alternativno domnevo** H_1.

Na enostavnem primeru bomo ilustrirali princip preizkušanja statističnih domnev.

Stroj polni neko snov v stekleničke in sicer je norma 50 mg na stekleničko. Zaradi slučajnih vplivov odmerki nihajo. Privzeti smemo, da so odmerki porazdeljeni normalno. Če stroj dela v skladu s predpisom, za maso odmerka velja: $X \sim N(50\,\text{mg}, 5\,\text{mg})$.

Kontrola kakovosti zahteva, da preverimo, ali stroj dela v skladu s predpisom. Za zdaj nas zanima, ali je povprečna masa zdravila v stekleničkah 50 mg, torej ali je $\mu = 50\,\text{mg}$. Če bi se izkazalo, da to ni res, je potrebno stroj ustaviti in ga ponovno nastaviti.

Izvedemo naslednji postopek. S slučajno izbiro izberemo določeno število stekleničk v t. i. kontrolni vzorec. Naj bo v kontrolnem vzorcu 25 stekleničk. V vsaki steklenički odmerek stehtamo, nato iz dobljenih vrednosti izračunamo vzorčno aritmetično sredino $\bar{x}$. Statistični sklep temelji na naslednji tezi: če $\bar{x}$ 'dovolj' odstopa od vrednosti 50 mg, se nakazuje, da povprečna vrednost ni 50 mg, torej da $\mu \neq 50\,\text{mg}$.

Na osnovi povedanega formuliramo dve statistični domnevi takole:

Ničelna domneva: povprečje je 50 mg.

$H_0 : \mu = 50\,\text{mg}$

Ničelni domnevi nasprotna je alternativna domneva.

Alternativna domneva: Povprečje ni 50 mg.

$H_1 : \mu \neq 50\,\text{mg}$

Opomba: tako določeno alternativno domnevo imenujemo dvostranska sestavljena alternativna domneva, saj obsega dve območji: $\mu < 50\,\text{mg}$ in $\mu > 50\,\text{mg}$. Vsako od teh dveh območij je sestavljeno iz množice vrednosti, ki so v intervalu.

Za zdaj privzemimo, da poznamo vrednost za standardni odklon: $\sigma = 5$ mg.

Preizkušanje statističnih domnev izhaja iz predpostavke, da je ničelna domneva pravilna. Če je to res, je porazdelitev vzorčnih aritmetičnih sredin $\bar{x}$ v kontrolnih vzorcih velikosti 25 normalna, njeno povprečje je 50 mg, standardni odklon pa 1 mg:

$$\bar{X} \sim N(50\,\text{mg}, \frac{5}{\sqrt{25}}\,\text{mg} = 1\,\text{mg})$$

To porazdelitev imenujemo **ničelna porazdelitev**, saj velja v primeru, da je ničelna domneva pravilna. Za to porazdelitev velja: približno dve tretjini kontrolnih vzorcev velikosti 25 ima $\bar{x}$ med 49 mg in 51 mg; približno 95% kontrolnih vzorcev ima $\bar{x}$ med 48 mg in 52 mg; skoraj vsi (99,7%) kontrolni vzorci imajo $\bar{x}$ med 47 in 53 mg.

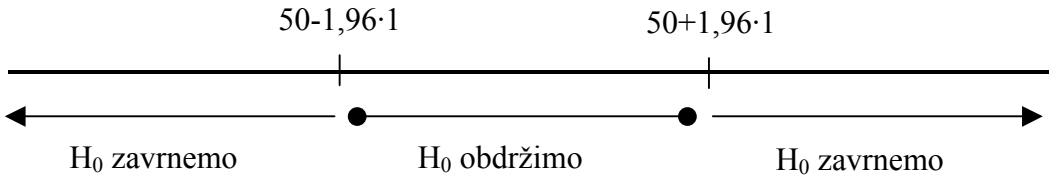
- Če bi za določen vzorec dobili $\bar{x} = 55$ mg, bi zagotovo zavrnili ničelno domnevo v korist alternativne domneve. Res seveda je, da ob pogoju, da je ničelna domneva pravilna, ni nemogoče dobiti kontrolnega vzorca, katerega $\bar{x} = 55$ mg ali celo več kot to, vendar je verjetnost takega dogodka izjemno majhna, manjša od milijoninke: $P(\bar{X} \geq 55) = P(Z \geq 5) < 3 \cdot 10^{-7}$. Torej, če je ničelna domneva pravilna, pričakujemo na deset milijonov kontrolnih vzorcev tri, ki bi imeli $\bar{x}$ 55 mg ali več. Ker je ta verjetnost izjemno majhna, izjemno malo tvegamo, ko zavrnemo ničelno domnevo v korist alternativne domneve.
- Pa recimo, da je $\bar{x} = 53$ mg. Če v tem primeru ničelno domnevo zavrnemo v korist alternativne domneve, tvegamo malce več, čeprav se zdi tveganje še vedno majhno: $P(\bar{X} \geq 53) = P(Z \geq 3) < 2 \cdot 10^{-4}$

Kaj narediti v posamičnem primeru, je odvisno od tega, kolikšno tveganje smo pripravljeni sprejeti. Pri statističnem sklepanju ta problem rešimo tako, da vnaprej opredelimo verjetnost, ki predstavlja zgornjo mejo za omenjeno tveganje. To verjetnost označimo α, njeno 'tehnično' ime je **stopnja značilnosti**. Na osnovi verjetnosti α, domnev H_0 in H_1 razdelimo vrednosti za $\bar{x}$ na dve območji:

- *območje, kjer H_0 obdržimo***.** Območje je določeno tako, da vsebuje $100 \cdot (1 - \alpha)$ odstotkov vzorčnih aritmetičnih sredin $\bar{x}$;
- *območje, kjer H_0 zavrnemo v korist H_1*. To območje vsebuje $100 \cdot \alpha$ odstotkov vzorčnih aritmetičnih sredin $\bar{x}$.

Vrednost na abscisi, ki razločuje območje, kjer H_0 obdržimo in kjer H_0 zavrnemo, se imenuje **kritična vrednost**.

Za zgornji primer naj velja $\alpha = 0{,}05$. Na sliki je predstavljeno območje, kjer H_0 obdržimo, in območje, kjer H_0 zavrnemo v korist H_1 za obravnavani primer. Ničelno domnevo obdržimo, če je $\bar{x}$ v intervalu $50\,\text{mg} \mp 1{,}96 \cdot 1\,\text{mg}$, torej približno v intervalu 48 mg do 52 mg. Če je $\bar{x}$ izven tega intervala, torej manjši od 48 mg ali večji od 52 mg, ničelno domnevo zavrnemo v korist alternativne domneve. Tedaj $\bar{x}$ dovolj odstopa od vrednosti 50 mg, da pri predpisani stopnji značilnosti 0,05 zavrnemo ničelno domnevo v korist alternativne domneve.

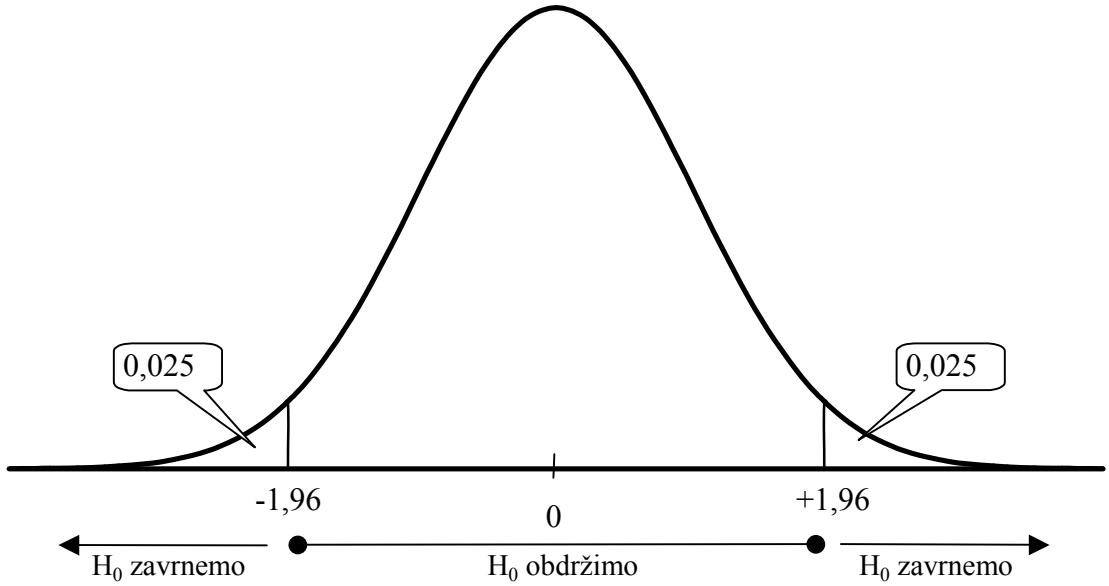


Slika 4-4: Območje za $\bar{x}$, kjer H_0 obdržimo in kjer H_0 zavrnemo

Spoznali smo že, da je bolj enostavno normalno porazdelitev vzorčnih aritmetičnih sredin $\bar{x}$ standardizirati:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Na sliki je prikazana ničelna porazdelitev za *z*-statistiko, to je N(0, 1), kritični vrednosti $\mp z_{0,025} = \mp 1{,}96$ in območje, kjer H_0 obdržimo in kjer H_0 zavrnemo v korist H_1.



Slika 4-5: Ničelna porazdelitev za *z*-statistiko, kritični vrednosti in območje sprejema in zavrnitve ničelne domneve

Primer

Poglejmo podatke za maso (mg) iz enega kontrolnega vzorca:

61,0 51,2 47,8 49,9 50,3 49,0 50,1 49,9 47,5 51,2
52,1 60,1 46,6 52,1 62,2 54,2 53,1 51,1 49,9 47,9
53,3 53,0 49,0 49,8 50,2

Upoštevajmo, da je $\sigma = 5$ mg .

Ničelna porazdelitev je N(0, 1), kritični vrednosti $\mp z_{0,025} = \mp 1{,}96$. Izračun *z*-statistike:

$$z = \frac{51{,}7 - 50}{5/\sqrt{25}} = 1{,}7$$

Vrednost 1,7 je v območju, kjer H_0 obdržimo.

Statistični sklep: ničelno domnevo obdržimo. V statističnem žargonu se odgovor glasi: *rezultati niso statistično značilni*.

Sedaj pa izhajamo iz dejstva, da σ ne poznamo, kar je bistveno bolj realističen primer. Njeno vrednost ocenimo iz vzorca. Iz prejšnjega poglavja vemo, da v tem primeru *z*-statistiko nadomesti Studentova *t*-statistika

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

ki je porazdeljena po Studentovi porazdelitvi z *n*-1 stopinjami prostosti.

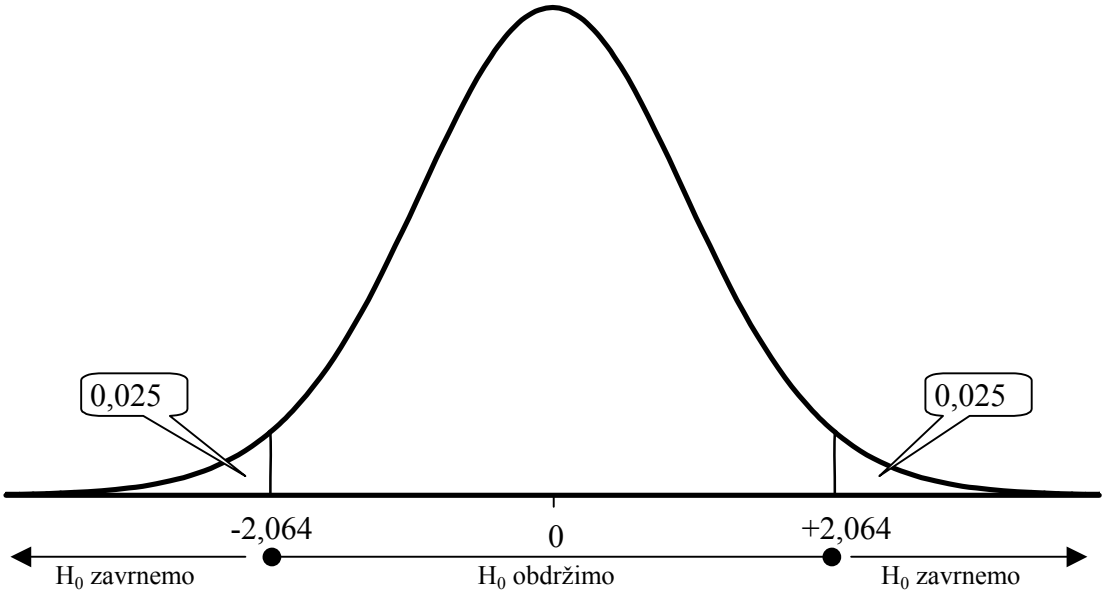
Primer

Za podatke iz prejšnjega primera ocenimo σ iz podatkov: $s = 4{,}026\,\text{mg}$.

Ničelna porazdelitev za Studentovo *t*-statistiko je *t*(24), kritični vrednosti odčitamo iz tabele za Studentovo porazdelitev: $\mp t_{0{,}025}(24) = \mp 2{,}064$. Izračun *t*-statistike:

$$t = \frac{51{,}7 - 50}{4{,}026/\sqrt{25}} = 2{,}112$$

Vrednost 2,112 je v območju, kjer H_0 zavrnemo v korist H_1.



Slika 4-6: Ničelna porazdelitev za *t*-statistiko, kritični vrednosti in območje sprejema in zavrnitve ničelne domneve

Statistični sklep: ničelno domnevo zavrnemo v korist alternativne domneve. V statističnem žargonu se odgovor glasi: *rezultati so statistično značilni* pri $\alpha = 0{,}05$.

4.2.2 Postopek pri preizkušanju statističnih domnev

Postopek pri preizkušanju statističnih domnev ima 6 korakov.

1. Izberemo vrednost za stopnjo značilnosti α. Standardne vrednosti, ki jih uporabljamo, so: 0,05; 0,01; 0,001.

   Opomba: te vrednosti so izbrali vodilni statistiki v časih, ko računalnikov še ni bilo. Izračun kritičnih vrednosti je računsko zahteven, zato so izdelali tabele le za nekaj izbranih vrednosti za stopnjo značilnosti.
2. Postavimo H_0 in H_1. Zapišemo ju z besedami.
3. Predpostavimo, da H_0 velja. Izberemo ustrezno testno statistiko. Narišemo skico: na ničelni porazdelitvi določimo območje, kjer H_0 obdržimo, in območje, kjer H_0 zavrnemo v korist H_1. Kritične vrednosti odčitamo iz tabel.
4. Iz vzorca izračunamo vrednost testne statistike. Pogledamo, v katero območje ta vrednost pade.
5. Statistični sklep:
   - ničelno domnevo obdržimo, rezultati niso statistično značilni;
   - ničelno domnevo zavrnemo v korist alternativne domneve, rezultati so statistično značilni.
6. Vsebinski sklep:
   - vzorčni podatki ne nasprotujejo ničelni domnevi.
   - pri stopnji značilnosti α trdimo, da je alternativna domneva pravilna. Verjetnost, da smo se zmotili, je največ α.

Da bomo znali korektno obrazložiti dobljene rezultate, poglejmo naslednje poglavje.

4.2.3 Napake pri preizkušanju statističnih domnev

Kot smo videli, se pri preizkušanju statističnih domnev postavimo na izhodišče, da je ničelna domneva pravilna. Pogledamo ničelno porazdelitev ustrezne vzorčne statistike v populaciji vseh vzorcev določene velikosti, velikost določa naš vzorec. Iz vzorca izračunamo vrednost vzorčne statistike in jo umestimo v njeno ničelno porazdelitev.

Sklep o tem, ali ničelno domnevo obdržimo ali zavrnemo, temelji na enem vzorcu, zato so pri preizkušanju statističnih domnev možne napake. Poglejmo možne napake. Dejansko je lahko ničelna domneva pravilna ali pa je pravilna alternativna domneva; kaj je res, seveda ne vemo. Na osnovi podatkov iz enega vzorca naredimo statistični sklep: ničelno domnevo obdržimo ali ničelno domnevo zavrnemo v korist alternativne domneve.

Če ničelno domnevo obdržimo tedaj, ko je le-ta pravilna, smo storili prav. Isto velja v primeru, ko ničelno domnevo zavrnemo v korist alternativne in je le-ta dejansko pravilna.

Napako storimo v primeru, ko ničelno domnevo zavrnemo v korist alternativne domneve, pa je ničelna domneva pravilna. V tem primeru naredimo t. i. **napako I. vrste**. Napako storimo tudi v primeru, ko ničelno domnevo obdržimo, pa ničelna domneva ni pravilna, pravilna je alternativna domneva. Tedaj storimo t. i. **napako II. vrste**.

Primer

Razmislimo, kaj pomeni napaka I. in II. vrste za primer stroja, ki polni stekleničke.

H_0 : Stroj dela v skladu s predpisom.

H_1 : Stroj ne dela v skladu s predpisom.

Napaka I. vrste pomeni, da trdimo, da stroj ne dela v skladu s predpisom, v resnici pa dela v skladu s predpisom. Posledica te napake je, da stroj po nepotrebnem ustavimo in mu ponovno nastavimo doziranje.

Napaka II. vrste pomeni, da trdimo, da stroj dela v skladu s predpisom, v resnici pa ne dela tako. Posledica je, da stroj napačno dozira, ne da bi mi vedeli za to.

Verjetnost za napako I. vrste označimo α, verjetnost za napako II. vrste označimo β.
Spodnja tabela povzema besedilo.

Tabela 4-2: Možne napake pri statističnem sklepanju in oznake za verjetnost napake

STATISTIČNI SKLEP	**DEJANSKO STANJE**	
	Velja H_0	**Velja** H_1
H_0 **obdržimo**	Napake ni	Napaka II. vrste (β)
H_0 **zavrnemo v korist** H_1	Napaka I. vrste (α)	Napake ni

Torej je α verjetnost za napako, ki jo storimo v primeru, ko s statističnim sklepom ničelno domnevo zavrnemo v korist alternativne domneve, pa je ničelna domneva pravilna. Pri preizkušanju statističnih domnev smo to verjetnost že spoznali, imenovali smo jo stopnja značilnosti. Njeno vrednost izberemo sami glede na to, kolikšno je za nas največje še sprejemljivo tveganje za napako I. vrste. Pomembno je, da je vrednost za α izbrana vnaprej, s tem je območje sprejema in zavrnitve ničelne domneve določeno pred izvedbo računskega postopka.

Problem pri preizkušanju statističnih domnev je dejstvo, da ne poznamo verjetnosti za napako II. vrste β. Če ničelno domnevo obdržimo, ne vemo, kolikšna je verjetnost za to, da dejansko velja alternativna domneva. Poglejmo, zakaj pride do tega.

Grafično predstavimo verjetnosti α in β za primer stroja, ki polni stekleničke. Zaradi enostavnosti bomo zopet privzeli, da je standardni odklon $\sigma = 5\,\text{mg}$. Za tveganje izberemo $\alpha = 0{,}05$.

$H_0 : \mu = 50\,\text{mg}$

$H_1 : \mu \neq 50\,\text{mg}$

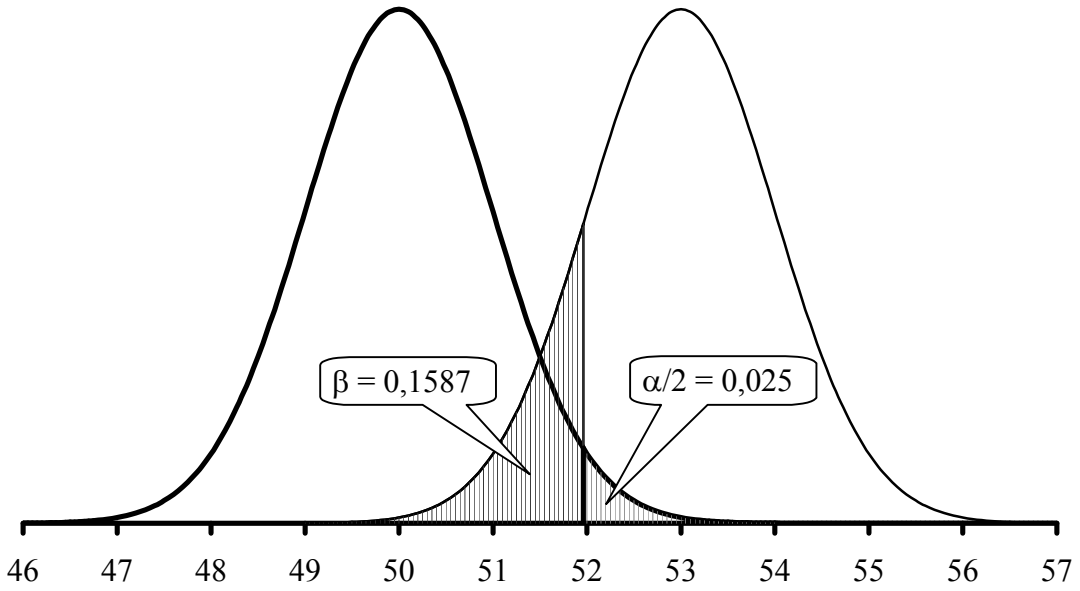
Ta alternativna domneva je sestavljena iz množice enostavnih alternativnih domnev oblike:

$H_1^* : \mu = \mu_1,\ \mu_1 \in R, \mu_1 \neq 50\,\text{mg}$.

Pa vzemimo eno izmed teh domnev, npr. $\mu_1 = 53\,\text{mg}$:

$H_1^* : \mu = 53\,\text{mg}$

Če H_0 velja, je ničelna porazdelitev $\overline{X}_0 \sim N(50,1)$. Če H_1^* velja, je alternativna porazdelitev $\overline{X}_1^* \sim N(53,1)$. Na sliki sta grafično predstavljeni ti dve porazdelitvi in verjetnosti α in β.



Slika 4-7: Grafični prikaz verjetnosti za napako I. vrste in II. vrste

Izračunajmo vrednost za β:

$$\beta = P(\overline{X}_1^* < 50 + 1{,}96 \cdot 1) = P(Z < -1) = 0{,}1587$$

Če je H_1^* drugačna, se spremeni β. Za določene izbrane vrednosti μ_1 je v tabeli izračunana verjetnost za napako II. vrste β.

Tabela 4-3: Verjetnost za napako II. vrste za določene izbrane vrednosti μ_1

μ_1	β
46	0,0228
47	0,1587
48	0,5000
49	0,8413
51	0,8413
52	0,5000
53	0,1587
54	0,0228

Ker je naša izhodiščna alternativna domneva sestavljena, ne vemo, kolikšna je verjetnost za napako II. vrste β, le-ta je lahko izjemno velika. Zato moramo biti v primeru, ko ničelno domnevo obdržimo, pri vsebinskem sklepu zelo previdni. Trdimo lahko le, da *vzorčni rezultati ne nasprotujejo ničelni domnevi*, kar še zdaleč ni dokaz, da je ničelna domneva pravilna. To pomeni, da na osnovi vzorčnih podatkov ničelne domneve ne moremo zavrniti v korist alternativne domneve pri predpisani stopnji značilnosti.

S statističnega stališča je najbolje, če uspemo ničelno domnevo zavrniti v korist alternativne domneve pri čim manjši stopnji značilnosti α. Kako izbrati vrednost za α? Splošnega odgovora ni. Odgovor je odvisen od posledic, ki jih ima statistični sklep v življenju. Če

informacije o tem nimamo, navadno privzamemo pri statističnem sklepanju za α vrednost 0,05.

Ker je β verjetnost, da H_0 obdržimo, ko velja H_1, je $1-\beta$ verjetnost, da H_0 zavrnemo, ko velja H_1. Količino $1-\beta$ imenujemo **moč preizkusa**. Razumljivo je, da želimo čim večjo moč preizkusa. Na moč preizkusa bistveno vpliva velikost vzorca. Večji vzorec pomeni več informacije in večjo možnost, da zavrnemo ničelno domnevo v korist alternativne domneve, če je alternativna domneva pravilna. To dejstvo ilustrira naslednji primer.

Primer

Pri stopnji značilnosti 0,05 preverimo domnevo, da je povprečna plača v občini 90 000 SIT.

$\alpha = 0{,}05$

$H_0 : \mu = 90\,000\,\text{SIT}$ Povprečna plača je 90 000 SIT.

$H_1 : \mu \neq 90\,000\,\text{SIT}$ Povprečna plača ni 90 000 SIT.

a) Na slučajnem vzorcu 10 zaposlenih so izračunali vzorčno aritmetično sredino 87 530 SIT in vzorčni standardni odklon 5 530 SIT.

Testna statistika: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Ničelna porazdelitev: $t(SP = 9)$

Kritični vrednosti: $\mp t_{0,025}(SP = 9) = \mp 2{,}262$

Vrednost *t*-statistike:

$$t = \frac{87530 - 90000}{5530/\sqrt{10}} = -1{,}412$$

H_0 obdržimo.

Vsebinski sklep: podatki iz našega vzorca ne nasprotujejo domnevi, da je povprečna plača v občini 90 000 SIT.

Opomba: ta sklep ni dokaz, da je povprečna plača v občini 90 000 SIT.

b) Na slučajnem vzorcu 40 zaposlenih so izračunali isto vzorčno aritmetično sredino 87 530 SIT in isti vzorčni standardni odklon 5 530 SIT.

Testna statistika $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Ničelna porazdelitev: $t(SP = 39)$

Kritični vrednosti: $\mp t_{0,025}(SP = 39) \approx \mp 2{,}021$

Vrednost *t*-statistike: $t = \dfrac{87530 - 90000}{5530/\sqrt{40}} = -2{,}825$

H_0 zavrnemo v korist H_1.

Vsebinski sklep: pri stopnji značilnosti 0,05 trdimo, da povprečna plača v občini ni 90 000 SIT. Podatki *nakazujejo*, da je povprečna plača pod 90 000 SIT.

4.2.4 *Dvostranske in enostranske alternativne domneve

Dvostranska alternativna domneva je npr.

$H_1 : \mu \neq 50\,\text{mg}$

Enostranska alternativna domneva je oblike:

$H_1 : \mu > 50\,\text{mg}$

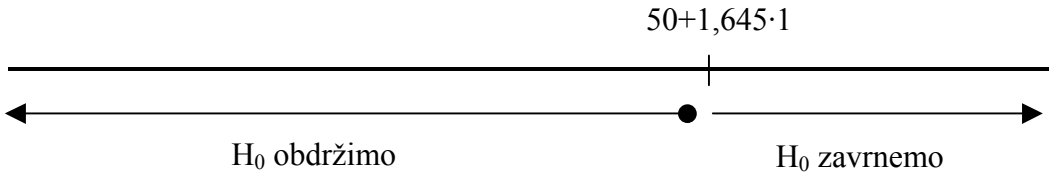
ali

$H_1 : \mu < 50\,\text{mg}$

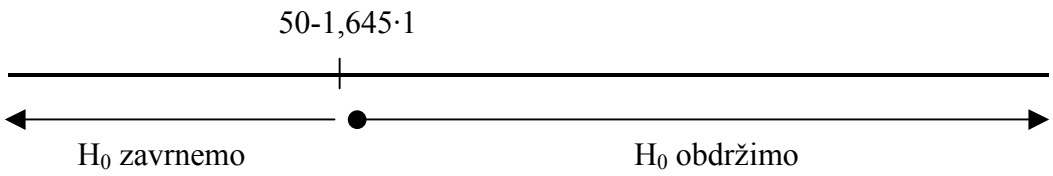
Pri enostranskih alternativnih domnevah je območje sprejema in zavrnitve ničelne domneve drugačno. Območje sprejema in zavrnitve določa ena kritična vrednost, ki je določena tako, da je verjetnost napake I. vrste α v enem repu porazdelitve.

Slika prikazuje območja sprejema in zavrnitve ničelne domneve pri enostranski alternativni domnevi pri pogojih, da je $\alpha = 0{,}05$ in $\sigma = 1$. Iz tabel odčitana vrednost je $z_{0,05} = 1{,}645$.

a) $H_1 : \mu > 50\,\text{mg}$



b) $H_1 : \mu < 50\,\text{mg}$



Slika 4-8: Območje za $\bar{x}$, kjer H_0 obdržimo in kjer H_0 zavrnemo pri enostranski alternativni domnevi

Za enostranske alternativne domneve se odločamo samo v primeru, kadar je iz vsebine problema jasno, da so možni odkloni od ničelne domneve le v eno smer. Zavedati se moramo, da pri enostranski alternativni domnevi prej zavrnemo ničelno domnevo v korist alternativne domneve kot pri dvostranski.

Primer

Izdelovalec trdi, da je zdravilo učinkovito sredstvo proti alergiji vsaj v 90 odstotkov primerov. V slučajnem vzorcu je bilo 200 alergikov, zdravilo je učinkovalo pri 160 alergikih. Preverite trditev proizvajalca pri stopnji značilnosti 0,01.

$\alpha = 0{,}01$

$H_0 : p = 0{,}9$ Verjetnost, da je zdravilo učinkovito, je 0,9.

$H_1 : p < 0{,}9$ Verjetnost, da je zdravilo učinkovito, je manjše od 0,9.

Naj bo X število alergikov v vzorcu velikosti n, pri katerih je zdravilo učinkovalo. Ničelna porazdelitev te slučajne spremenljivke je binomska $b(n, p_0)$, pri čemer je p_0 verjetnost, ki jo predpostavlja ničelna domneva. Binomsko porazdelitev smemo aproksimirati z normalno $b(n, p_0) \approx N(np_0, np_0q_0)$. Iz tega sledi, da je testna statistika

$$z = \frac{x - np_0}{\sqrt{np_0q_0}}$$

porazdeljena po $N(0,1)$.

Ker je alternativna domneva enostranska, je kritična vrednost samo ena: $-z_{0,01} = -2{,}326$. To vrednost najhitreje najdemo v prilogi v Tabeli 3: $t_{0,01}(SP = \infty)$.

Izračun *z*-statistike: $z = \frac{160 - 180}{\sqrt{200 \cdot 0{,}9 \cdot 0{,}1}} = -4{,}714$

Izračunana vrednost je v območju, kjer se ničelna domneva zavrne v korist alternativne domneve.

Sklep: Pri stopnji značilnosti 0,01 trdimo, da je učinkovitost zdravila manj kot 90-odstotna.

4.2.5 *p*-vrednost

Predstavimo še drug način predstavitve rezultatov statističnih domnev, ki je v uporabi pri računalniški analizi. Za statistični preizkus se izračuna dve vrednosti: testno statistiko in pripadajočo ***p*-vrednost**.

p-vrednost je verjetnost, da ob predpostavki, da je ničelna domneva pravilna, dobimo za testno statistiko vrednosti, ki so bolj 'ekstremne' (bolj v korist alternativne domneve) od izračunane vrednosti testne statistike.

Ilustrirajmo izračun *p*-vrednosti in njen grafični prikaz na primeru.

Primer

Izračunajmo *p*-vrednost za statistični preizkus podatkov o stroju, ki polni stekleničke. Obravnavamo primer: $\sigma = 5$ mg.

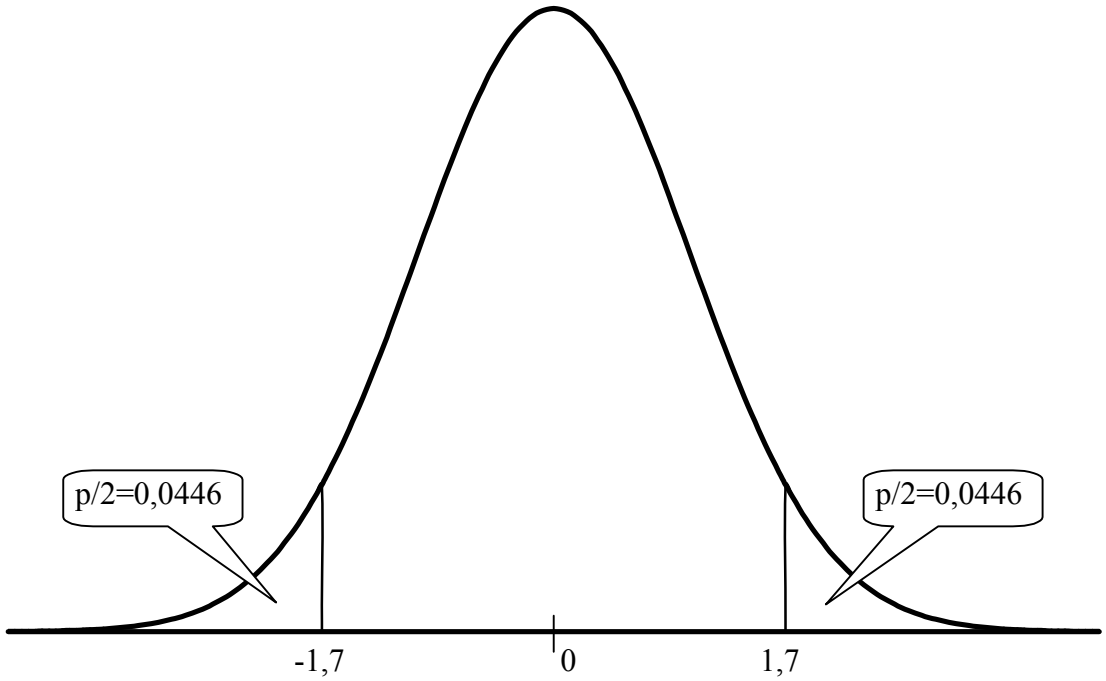
$H_0 : \mu = 50$ mg

$H_1 : \mu \neq 50$ mg

Izračunana vrednost *z*-statistike je: $z = 1{,}7$. *p*-vrednost izračunamo takole:

$p = 2 \cdot P(Z > 1{,}7) = 2 \cdot 0{,}0446 = 0{,}0891$

Opomba: faktor 2 nastopa zaradi dvostranske alternativne domneve. Grafična predstavitev *p*-vrednosti je na sliki



Slika 4-9: Grafična predstavitev *p*-vrednosti

Računalniški izpis rezultatov: $z = 1{,}7$; p=0,0891.

---

p-vrednost je vezana na vzorec in izraža, v kolikšni meri so vzorčni podatki v skladu z ničelno domnevo. Večja vrednost za *p* pomeni večjo podporo ničelni domnevi, majhen *p* pa govori v prid alternativne domneve. Stopnja značilnosti α je zgornja meja za napako I. vrste. Če je *p*-vrednost manjša od predpisane vrednosti za α, ničelno domnevo zavrnemo. Če je *p*-vrednost večja od α, ničelno domnevo obdržimo.

Za zgornji primer bi se pri $\alpha = 0{,}05$ odločili, da se ničelna domneva obdrži, saj je $p > \alpha$.

Za standardizirano normalno porazdelitev lahko izračunamo *p*-vrednost z našimi tabelami. Za ostale porazdelitve pa lahko s tabelami *p*-vrednost bolj ali manj grobo ocenimo. Računalniški programi dajo točne *p*-vrednosti.

p-vrednost navajamo pri vsebinskem sklepu, kakor kaže naslednji primer.

Primer

Izračunajmo *p*-vrednost za statistični preizkus podatkov o stroju, ki polni stekleničke, ob pogoju, da je $\alpha = 0{,}05$, vrednost za standardni odklon ni podana.

Izračunana vrednost *t*-statistike je: $t = 2{,}112$. *p*-vrednost izračunamo takole:

$p = 2 \cdot P(T(SP = 24) > 2{,}112)$

S tabelami to vrednost ocenimo: $0{,}02 < p < 0{,}05$.

Računalniški izpis rezultatov: $t = 2{,}112$; $p = 0{,}0453$.

Statistični sklep: ničelno domnevo zavrnemo v korist alternativne domneve.

Vsebinski sklep: pri stopnji značilnosti 0,05 trdimo, da povprečna vrednost ni 50 mg ($p = 0{,}0453$).

Pri raziskavah pogosto navajajo rezultate preizkušanja statističnih domnev v obliki: vrednost testne statistike, *p*-vrednost. Posebno je to v uporabi tedaj, ko sklepi v obliki: obdržimo/zavrnemo za prakso niso potrebni. To še posebej velja za mnoga področja znanstveno raziskovalnega dela, kjer je statistično sklepanje le eno izmed orodij, ki določa nadaljnji potek raziskave.

4.2.6 Pregled preizkusov o povprečju in o Bernoullijevi verjetnosti

Povzeli bomo statistične preizkuse in izračune intervalov zaupanja za povprečje in za Bernoullijevo verjetnost, ki smo jih uporabili v tem poglavju.

4.2.6.1 Povprečje

a) $X \sim N(\mu,\sigma)$, σ je znana

$H_o : \mu = \mu_0$

Testna statistika: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Ničelna porazdelitev: $N(0,1)$

b) $X \sim N(\mu,\sigma)$, σ ni znana

$H_o : \mu = \mu_0$

Testna statistika: $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

Ničelna porazdelitev: $t(SP = n-1)$

c) veliki vzorci, izhodiščna porazdelitev je poljubna

$H_o : E(X) = \mu_0$

Testna statistika: $z = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

Ničelna porazdelitev: $N(0,1)$

4.2.6.2 *Bernoullijeva verjetnost

$X \sim b(n,p)$

Analiziramo primere, ko lahko binomsko porazdelitev aproksimiramo z normalno:

$b(n,p) \approx N(np, npq)$

$H_o : p = p_0$

Testna statistika: $z = \dfrac{x - np_0}{\sqrt{np_0 q_0}}$

Ničelna porazdelitev: $N(0,1)$

NALOGE

1. Sodstvo

Razmislite, kaj pomeni napaka I. vrste in napaka II. vrste za ugotavljanje krivde obtoženega. Ničelno in alternativno domnevo postavimo takole:

H_0 : obtoženi ni kriv.

H_1 : obtoženi je kriv.

Kakšne so posledice napake I. vrste in napake II. vrste?

2. Avtomat

Avtomat polni stekleničke. Privzeti smemo, da je masa odmerka porazdeljena po normalni porazdelitvi. Stroj dela v skladu s predpisi, če je povprečna masa v stekleničkah 100 g. Nekega dne so dobili v kontrolnem vzorcu velikosti 16 naslednje vrednosti (v mg):

99,2 95,8 99,0 101,3 102,1 100,3 98,4 97,6

103,2 101,9 100,5 99,5 101,2 104,2 100,4 99,4

Pri stopnji značilnosti 0,05 preverite domnevo, da stroj dela v skladu s predpisi.

3. Starost gledalcev

Pri stopnji značilnosti 0,05 preverite domnevo, da je povprečna starost gledalcev določene televizijske oddaje 35 let.

V vzorec je bilo vključenih 150 slučajno izbranih gledalcev. Podatki o njihovi starosti (dopolnjena leta) so v frekvenčni porazdelitvi:

Tabela 4-4: Frekvenčna porazdelitev gledalcev po starosti

Starost (dop. leta)	Število
20-29	20
30-39	75
40-49	45
50 in več	10

4. *Mežiška dolina

Da bi ugotovili vsebnost žvepla v smrekovih iglicah, so raziskovalci na območju Mežiške doline s slučajno izbiro izbrali 20 dreves. Na vsakem drevesu so izbrali sedemletno vreteno. Za analizo so uporabili potrgane enoletne in dvoletne iglice. Podatki o količini žvepla v iglicah (mg/m^3) so v tabeli. Za količino žvepla bomo privzeli normalno porazdelitev.

Tabela 4-5: Podatki o količini žvepla v iglicah (mg/m^3)

Drevo številka	Žveplo (mg/m^3) enoletne iglice	Žveplo (mg/m^3) dvoletne iglice
1	16,3	16,3
2	25,4	25,5
3	10,6	10,5
4	30,1	33,9
5	18,1	26,1
6	20,5	33,9
7	11,8	14,7
8	13,5	20,6
9	6,6	11,1
10	12,5	13,6
11	36,6	40,6
12	6,8	7,7
13	9,4	12,5
14	35,9	40,5
15	8,9	10,5
16	12,9	18,3
17	20,6	21,5
18	10,8	43,4
19	24,4	36,7
20	29,1	40,8

a) Izračunajte 95% interval zaupanja za povprečno količino žvepla v enoletnih iglicah.

b) Izračunajte 95% interval zaupanja za povprečno količino žvepla v dvoletnih iglicah.

c) Maksimalna dovoljena vrednost za povprečno količino žvepla v iglicah je 12,5 mg/m^3. Komentirajte rezultate pod a) in b).

5. *Križanje graha

V skladu z Mendelovimi genetskimi zakoni bi moralo biti pri križanju dveh vrst graha v F1 generaciji razmerje med rumenimi in zelenimi zrni 3:1. V poskusu so ugotovili, da je izmed 1064 grahovih zrn v F1 generaciji 787 rumenih. Presodite pri stopnji značilnosti 0,05, ali so eksperimentalni rezultati v skladu z Mendelovo teorijo. Izračunajte *p*-vrednost. Komentirajte rezultate.

6. *Reklama cigaret

Raziskovalci so želeli ugotoviti, ali se prebivalci določenega območja strinjajo s prepovedjo reklame cigaret. Anketirali so 430 oseb. S prepovedjo se je strinjalo 280 oseb. Izračunajte 95% interval zaupanja za odstotek oseb, ki se strinjajo s prepovedjo reklam za cigarete.

5 PRIMERJAVA DVEH POPULACIJ

Primerjali bomo parametre verjetnostnih porazdelitev za dve populaciji in sicer povprečji dveh normalnih porazdelitev ter Bernoullijevi verjetnosti dveh binomskih porazdelitev. Poglejmo najprej nekaj zgledov za te probleme:

- ali je povprečna življenjska doba procesorjev A in B enaka?
- kolikšna je razlika povprečne količine žvepla v enoletnih iglicah in v dvoletnih smrekovih iglicah?
- ali je odstotek kadilcev pri študentih in pri študentkah enak?
- kolikšna je razlika kalivosti med standardnim in novim kultivarjem?

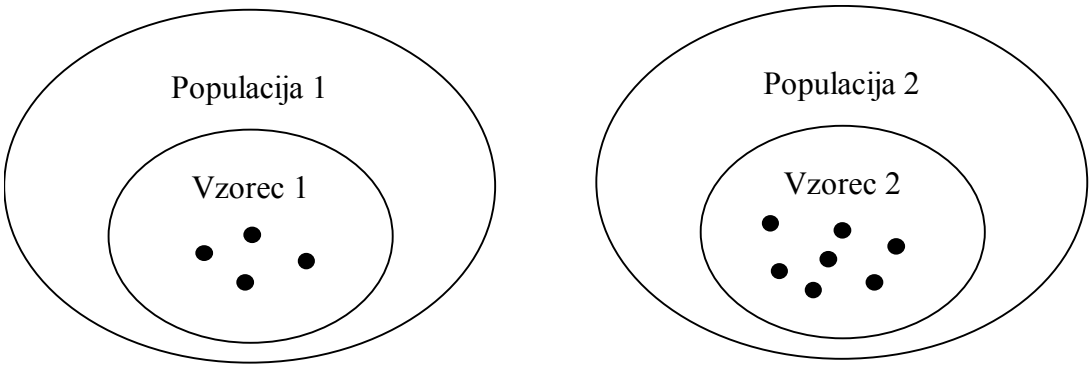
Če bi lahko pogledali vse enote v obeh populacijah, bi dobili točen odgovor na zastavljeno vprašanje. Običajno tega ne moremo, zato izvedemo vzorčenje oz. načrtovan poskus. Vzorčimo prvo populacijo, vzorčimo drugo populacijo. Na osnovi informacije obeh vzorcev skušamo sklepati o tem, kaj velja v pripadajočih populacijah.

Najprej poglejmo oznake. Naj bo X slučajna spremenljivka, ki jo proučujemo v dveh populacijah (v gornjih primerih: življenjska doba, količina žvepla, kadilec/nekadilec, kali/ne kali), X_1 označuje X v prvi populaciji, X_2 v drugi populaciji (v gornjih primerih: procesorji A in B, enoletne in dvoletne iglice, študenti in študentke, standardni in novi kultivar).

Vzorca, ki zastopata prvo in drugo populacijo, sta lahko **neodvisna** ali **odvisna**. Obravnavali bomo oba primera, najprej neodvisna vzorca.

5.1 DVA NEODVISNA VZORCA

Vzorca imenujemo **neodvisna**, če informacija, ki jo dobimo iz prvega vzorca, ni povezana z informacijo, ki jo dobimo iz drugega vzorca. Slika nakazuje to situacijo.



Slika 5-1: Shema dveh neodvisnih vzorcev

5.1.1 Razlika povprečij

Najprej bomo obravnavali primer, ko je možno za spremenljivko X v obeh populacijah privzeti normalno porazdelitev, pri čemer je v obeh populacijah varianca ista:

$X_1 \sim N(\mu_1, \sigma)$

$X_2 \sim N(\mu_2, \sigma)$

Iz prve populacije s slučajno izbiro izberemo vzorec velikosti n_1, iz druge vzorec velikosti n_2. Kot že vemo, za pripadajoče vzorčne aritmetične sredine velja:

$$\overline{X}_1 \sim N(\mu_1, \frac{\sigma}{\sqrt{n_1}})$$

$$\overline{X}_2 \sim N(\mu_2, \frac{\sigma}{\sqrt{n_2}})$$

Zanima nas verjetnostna porazdelitev razlike $\overline{X}_1 - \overline{X}_2$. S pomočjo izreka o povprečju razlike in izreka o varianci razlike dveh neodvisnih slučajnih spremenljivk ugotovimo:

$$E\left(\overline{X}_1 - \overline{X}_2\right) = \mu_1 - \mu_2$$

$$Var\left(\overline{X}_1 - \overline{X}_2\right) = \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Teorija pokaže, da je porazdelitev razlike vzorčnih aritmetičnih sredin naslednja:

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Pripadajoča vzorčna statistika:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

je porazdeljena po standardizirani normalni porazdelitvi $N(0,1)$.

Ko smo proučevali domnevo o povprečju enega vzorca, smo ugotovili, da σ^2 ne poznamo in da jo moramo oceniti. V tem primeru velja isto, to oceno imenujemo *skupna varianca* s_{sk}^2. Matematična statistika pokaže, kako se ta ocena izračuna. Ocena je opredeljena na osnovi obeh vzorčnih varianc s_1^2 in s_2^2, upošteva pa tudi velikost pripadajočih vzorcev:

$$s_{sk}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{SP_1 \cdot s_1^2 + SP_2 \cdot s_2^2}{SP_1 + SP_2}$$

Kot vidimo, se skupna varianca s_{sk}^2 izračuna kot tehtana aritmetična sredina vzorčnih varianc, uteži so pripadajoče stopinje prostosti. Če sta vzorca enako velika, torej $n_1 = n_2$, je skupna varianca izračunana kot navadna aritmetična sredina obeh vzorčnih varianc:

$$s_{sk}^2 = \frac{s_1^2 + s_2^2}{2}$$

Če σ^2 nadomestimo z njeno oceno s_{sk}^2, *z*-statistiko nadomesti *t*-statistika:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_{sk}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Pripadajoča porazdelitev je Studentova porazdelitev, stopinje prostosti so vsota stopinj prostosti: $SP = SP_1 + SP_2 = n_1 + n_2 - 2$.

To dejstvo bomo uporabili pri preizkušanju domneve o razliki povprečnih vrednosti in pri izračunavanju pripadajočega intervala zaupanja. Nekaj komentarjev pred tem:

- Predpostavili smo, da je varianca v obeh populacijah enaka. Za preverjanje te predpostavke uporabimo ohlapno 'pravilo palca'. Če velja:

$$\frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} < 2,$$

torej je razmerje večje in manjše variance pod 2, je uporaba zgornjega testa praviloma dopustna. Statistično neoporečen odgovor na to vprašanje dobimo s pomočjo F -preizkusa, ki ga navajamo na koncu tega razdelka.

- Če se izkaže, da predpostavka o enakih variancah ni sprejemljiva, uporabimo posplošeno t-statistiko, ki upošteva obe vzorčni varianci:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Pripadajoča porazdelitev je Studentova porazdelitev, stopinje prostosti se izračunajo po precej zapleteni formuli, ki je tu ne navajamo. Kot sprejemljiv približek za stopinje prostosti lahko vzamemo $SP = \min(SP_1, SP_2) = \min(n_1 - 1, n_2 - 1)$.

5.1.1.1 Preizkus domneve o razliki povprečij

Preizkušali bomo domnevo o razliki dveh povprečnih vrednosti. Ničelna domneva predvideva, da je njuna razlika δ, torej $\delta = \mu_1 - \mu_2$, enaka neki vnaprej določeni vrednosti δ_0, $\delta_0 \in R$. Najpogosteje nas zanima, ali je ta razlika enaka nič, torej ali je $\delta_0 = 0$.

Zapišimo ničelno in alternativno domnevo.

$H_0: \delta = \mu_1 - \mu_2 = \delta_0$ Razlika povprečij je δ_0.

Alternativna domneva je negacija ničelne domneve.

$H_1: \delta = \mu_1 - \mu_2 \neq \delta_0$ Razlika povprečij ni δ_0.

Opomba: alternativna domneva je lahko enostranska.

Če lahko privzamemo, da sta varianci v obeh populacijah enaki, je testna statistika:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_{sk}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Ničelna porazdelitev je $t(SP = n_1 + n_2 - 2)$.

V primeru različnih varianc je testna statistika

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Približna ničelna porazdelitev je $t(SP = \min(n_1 - 1, n_2 - 1))$.

Primer

Pedagog je proučeval, ali je povprečno število ur študija za določen kolokvij pri študentih in pri študentkah enako. V raziskavo je bilo vključenih 12 slučajno izbranih študentk in 15 slučajno izbranih študentov. Povprašali so jih, koliko ur so študirali za kolokvij. Podatki so (ure):

Študentke: 15, 18, 1, 10, 20, 13, 12, 5, 18, 12, 0, 12

Študenti: 14, 15, 10, 2, 0, 4, 17, 6, 19, 11, 10, 12, 3, 4, 16

Pri stopnji značilnosti 0,05 preverimo domnevo, da je povprečno število ur za študij pri študentkah in študentih enako. Privzeli bomo, da je porazdelitev spremenljivke število ur študija normalna.

Rešitev

$\alpha = 0{,}05$

Ničelna domneva: povprečno število ur študija je pri študentih in študentkah enako.

Alternativna domneva: povprečno število ur študija pri študentih in študentkah ni enako.

Izračuni:

	Študentke	**Študenti**
n	12	15
$\bar{x}$	11,3333	9,5333
s^2	41,6970	36,4095

Razmerje vzorčnih varianc je pod 2, zato bomo izračunali skupno varianco:

$$s_{sk}^2 = \frac{11 \cdot 41{,}6970 + 14 \cdot 36{,}4095}{25} = 38{,}7360$$

$$t = \frac{11{,}333 - 9{,}533}{\sqrt{38{,}7360 \cdot \left(\frac{1}{12} + \frac{1}{15}\right)}} = \frac{1{,}800}{2{,}410} = 0{,}747$$

Kritični vrednosti sta: $\mp t_{0{,}025}(25) = \mp 2{,}060$

Statistični sklep: ničelno domnevo obdržimo.

Odgovor: podatki ne nasprotujejo domnevi, da je povprečno število ur študija pri študentih in študentkah enako ($p > 0{,}2$).

5.1.1.2 *Interval zaupanja za razliko povprečij

Iz teorije sledi, da izračunamo interval zaupanja za $\delta = \mu_1 - \mu_2$ takole:

$$l_{1,2} = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}(SP = n_1 + n_2 - 2) \cdot \sqrt{s_{sk}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Če predpostavka o enakih variancah ni sprejemljiva, se formula ustrezno spremeni.

Primer

V Mežiški dolini in na Pohorju so ugotavljali količino žvepla v smrekovih iglicah (mg/m^3).

Na vsaki lokaciji je bilo s slučajno izbiro izbranih po 90 smrek. Podatkov ne navajamo. Izračuni so v tabeli:

Tabela 5-1: Statistike vsebnosti žvepla v Mežiški dolini in na Pohorju

	Mežiška dolina	**Pohorje**
n	90	90
$\bar{x}$	16,4	12,3
s^2	162,20	105,59

Razmerje vzorčnih varianc je pod 2, zato bomo izračunali skupno varianco:

$$s_{sk}^2 = \frac{162{,}20 + 105{,}59}{2} = 133{,}895$$

Izračunajmo 95% interval zaupanja za razliko med povprečnima vrednostma žvepla v Mežiški dolini in na Pohorju:

$$l_{1,2} = 4{,}1 \pm 1{,}96 \cdot \sqrt{133{,}895 \cdot \frac{2}{90}} = 4{,}1 \pm 3{,}38$$

$l_1 = 0{,}7$ mg/m^3

$l_2 = 7{,}5$ mg/m^3

Pri 95% zaupanju interval (0,7 mg/m^3;7,5 mg/m^3) pokriva razliko povprečnih vrednosti. Povedano drugače: pri 95% zaupanju je povprečna količina žvepla v smrekovih iglicah v Mežiški dolini od 0,7 mg/m^3 do 7,5 mg/m^3 večja kot na Pohorju.

Opomba: predpostavka o normalni porazdelitvi količine žvepla ni potrebna, saj sta vzorca tako velika, da se po centralnem limitnem izreku lahko sklicujemo na normalno porazdelitev vzorčnih aritmetičnih sredin.

5.1.1.3 *Razmerje varianc

Pri primerjanju povprečij smo izhajali iz domneve, da sta varianci v obeh populacijah enaki. Ali je to za proučevani primer sprejemljiva predpostavka, je treba na danih podatkih preveriti. Pri tem uporabljamo naslednji izrek:

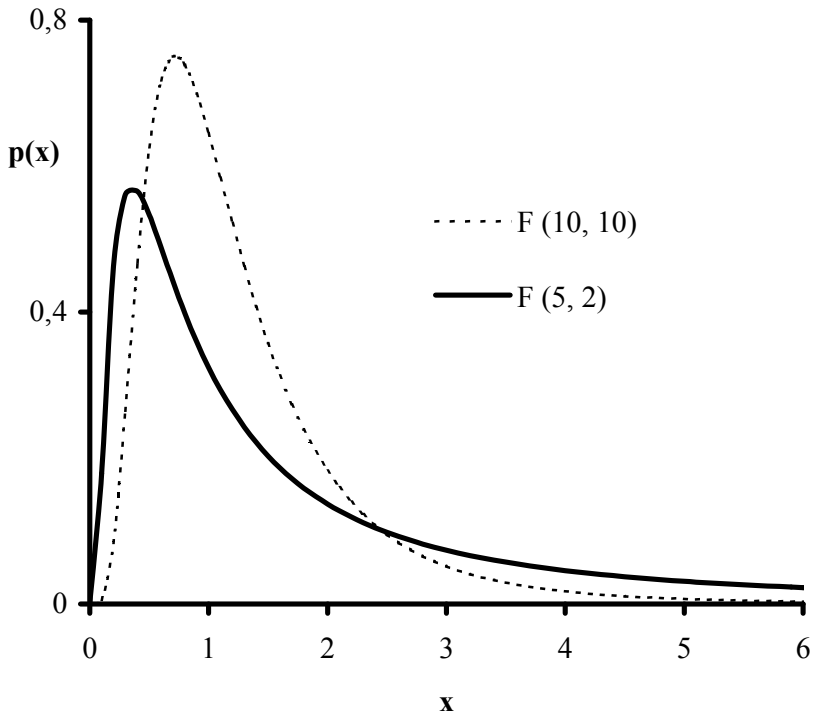
Če velja: $X_1 \sim N(\mu_1, \sigma)$ *in* $X_2 \sim N(\mu_2, \sigma)$*, torej je varianca v obeh populacijah enaka, je pri vzorcih velikosti* n_1 *in* n_2 *razmerje vzorčnih varianc*

$$\frac{S_1^2}{S_2^2}$$

porazdeljeno po $F -$*porazdelitvi s* $(SP_1 = n_1 - 1, SP_2 = n_2 - 1)$ *stopinjami prostosti.*

Na kratko navajamo nekaj lastnosti $F -$**porazdelitve**:

- je zvezna porazdelitev, definirana na pozitivnem delu realne osi
- ima dvoje stopinj prostosti SP_1 in SP_2
- je asimetrična v desno, njena oblika je odvisna od števila stopinj prostosti



Slika 5-2: Dve F porazdelitvi

- velja: $F_{\alpha}(SP_1, SP_2) = \dfrac{1}{F_{1-\alpha}(SP_2, SP_1)}$
- velja: $F_{\alpha}(1, n) = t^2_{\frac{\alpha}{2}}(n)$

Vrednosti za *F*-porazdelitev so tabelirane. Glej tabele 5, 6, 7 in 8 v prilogi: za $\alpha = 0{,}05$; 0,025; 0,001; 0,0005.

Preizkus o enakosti varianc v proučevanih dveh populacijah temelji na navedenemu izreku. Ničelno in alternativno domnevo zapišemo takole:

$H_0: \sigma_1^2 = \sigma_2^2$ Varianci sta enaki.

$H_1: \sigma_1^2 \neq \sigma_2^2$ Varianci nista enaki.

Teorija pokaže, da je testna statistika naslednja:

$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

Ničelna porazdelitev je $F(SP_{\max}, SP_{\min})$, $SP_{\max}$ se navezuje na maksimalno vzorčno varianco, $SP_{\min}$ pa na minimalno vzorčno varianco.

Primer:

Ali je domneva o enakosti variance za število ur študija pri študentkah in študentih sprejemljiva? Poglejmo, kaj kažejo podatki ($\alpha = 0{,}05$).

$H_0: \sigma_1^2 = \sigma_2^2$ Varianci sta enaki.

$H_1: \sigma_1^2 \neq \sigma_2^2$ Varianci nista enaki.

Testna statistika:

$$F = \frac{41{,}6970}{36{,}4095} = 1{,}145$$

Kritična vrednost : $F_{0,025}(SP_1 = 11, SP_2 = 14) \approx 3{,}098$

Statistični sklep: ničelno domnevo obdržimo.

Odgovor: podatki ne nasprotujejo domnevi o enakosti variance pri študentkah in študentih (p=0,7976).

Opomba: p-vrednost smo izračunali z računalnikom.

NALOGE

1. Pridelek paradižnika

Ali je povprečni pridelek sort paradižnika A in B enak? Da bi preverili to domnevo, so izvedli poskus: na razpolago je bilo 12 parcel na homogenem zemljišču. S slučajno izbiro je bilo izbranih 6 parcel za A in šest za B.

Shema poskusa in pridelek na parcelo (kg):

A	A	B	B	A	B	B	B	A	A	B	A
18,0	19,0	19,0	23,6	22,9	25,3	24,0	20,0	21,1	21,2	24,1	18,6

Pri stopnji značilnosti 0,05 preverite domnevo o enakosti povprečnega pridelka. Privzeti smemo, da je pridelek normalno porazdeljena slučajna spremenljivka.

2. Krma

Raziskovalci so imeli dve skupini svinj, v vsaki je bilo po 8 živali. Kontrolna skupina je bila krmljena klasično, eksperimentalna pa s siliranimi krompirjevimi olupki in dopolnilno krmo. Podana je teža živali pred zakolom (kg).

Kontrolna skupina

112 118 122 110 107 134 97 121

Eksperimentalna skupina

113 104 109 97 87 103 120 124

Privzeti smemo, da je porazdelitev teže normalna.

a) Za vsako skupino izračunajte kvartile.
b) Narišite okvir z ročaji za obe skupini na eni sliki. Komentirajte sliko.
c) Preverite domnevo, da je povprečna teža pred zakolom v obeh skupinah enaka (stopnja značilnosti 0,05). Obrazložite rezultate.
d) Ocenite p-vrednost.

3. Ocene na izpitu

Ocene na izpitu so bile v točkah od 0 do 100. V slučajnem vzorcu je bilo 32 fantov in 40 deklet. Pri izpitu je bila povprečna ocena fantov 72 točk in standardni odklon 8 točk, povprečna ocena deklet pa 75 točk in standardni odklon 6 točk.

Preizkusite domnevo, da sta povprečna uspeha deklet in fantov enaka (stopnja značilnosti 0,05).

4. *Starost študentov

V vzorcu je 17 študentov izrednega študija v šolskem letu 1996/97. Od njih je 8 nezaposlenih, drugi so zaposleni. Za vse navajamo rojstne letnice:

Nezaposleni:

1972 1975 1964 1975 1965 1974 1975 1972

Zaposleni:

1970 1975 1964 1966 1965 1965 1968 1963 1970

a) Izračunajte starost študentov v letu 1997.

b) Izračunajte 95% interval zaupanja za razliko povprečne starosti nezaposlenih in zaposlenih. Obrazložite rezultate. Privzeti smemo, da je porazdelitev starosti normalna.

5.1.2 *Razlika Bernoullijevih verjetnosti

V tem razdelku bomo pogledali, kako primerjamo verjetnost določenega dogodka v dveh populacijah. Včasih ima ta verjetnost kakšno drugo (strokovno) ime, pogosto jo izražamo v odstotkih.

Če govorimo o verjetnosti določenega dogodka, ki se zgodi ali pa se ne zgodi, je verjetnostna porazdelitev za pripadajočo slučajno spremenljivko binomska porazdelitev. Obravnavali bomo samo primere, pri katerih lahko binomsko porazdelitev aproksimiramo z normalno porazdelitvijo. Kdaj lahko to storimo, smo spoznali v poglavju o binomski porazdelitvi.

Najprej poglejmo tista spoznanja iz verjetnostnega računa, ki nam bodo pomagala pri reševanju tega problema.

Spremenljivka X_1 je porazdeljena po binomski porazdelitvi s parametroma n_1 in p_1, analogno X_2. Privzemimo situacijo, da lahko vsako od njiju aproksimiramo z ustrezno normalno porazdelitvijo:

$$X_1 \sim b(n_1, p_1) \approx N(n_1 p_1, \sqrt{n_1 p_1 q_1})$$

$$X_2 \sim b(n_2, p_2) \approx N(n_2 p_2, \sqrt{n_2 p_2 q_2})$$

Porazdelitev deležev je normalna:

$$\frac{X_1}{n_1} \approx N\left(p_1, \sqrt{\frac{p_1 q_1}{n_1}}\right)$$

$$\frac{X_2}{n_2} \approx N\left(p_2, \sqrt{\frac{p_2 q_2}{n_2}}\right)$$

Za situacijo, kjer sta X_1 in X_2 neodvisna, je porazdelitev razlike vzorčnih deležev normalna s parametroma:

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Situacija se poenostavi, če sta parametra p_1 in p_2 enaka. Recimo, da velja: $p_1 = p_2 = p_{sk}$, posledično je $q_{sk} = 1 - p_{sk}$. Potem zgornjo formulo zapišemo takole:

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\left(0, \sqrt{p_{sk}\ q_{sk}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Poglejmo, kako izračunamo vzorčne ocene za omenjene verjetnosti p_1, p_2, p_{sk}. Ta spoznanja posreduje matematična statistika:

- iz prvega vzorca izračunamo oceno za p_1, označimo jo $\hat{p}_1$:

$$\hat{p}_1 = \frac{x_1}{n_1} \qquad \hat{q}_1 = 1 - \hat{p}_1$$

pri čemer je n_1 velikost vzorca, x_1 pa število enot v tem vzorcu, pri katerih se je zgodil dogodek A;

- iz drugega vzorca izračunamo oceno za p_2, označimo jo $\hat{p}_2$:

$$\hat{p}_2 = \frac{x_2}{n_2} \qquad \hat{q}_2 = 1 - \hat{p}_2$$

- iz obeh vzorcev ocena za p_{sk}, označimo jo $\hat{p}_{sk}$:

$$\hat{p}_{sk} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \hat{q}_{sk} = 1 - \hat{p}_{sk}$$

5.1.2.1 *Preizkus domneve o razliki Bernoullijevih verjetnosti

Poglejmo naprej enostavni primer: ali je verjetnost dogodka A v prvi in v drugi populaciji enaka, torej, ali sta verjetnosti p_1 in p_2 enaki. Prvo populacijo reprezentira vzorec velikosti n_1, drugo vzorec velikosti n_2. Spremenljivka, ki jo proučujemo na enotah, je dvojiška, torej se na vsaki enoti lahko zgodi dogodek A ali $\overline{A}$. Ničelna domneva trdi, da je verjetnost dogodka A v obeh populacijah enaka, alternativna pa, da ni enaka.

H_0: $p_1 = p_2$

H_1: $p_1 \neq p_2$

Opomba: alternativna domneva je lahko enostranska.

Na osnovi spoznanj iz verjetnostnega računa je testna statistika *z* zapisana v obliki:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{sk} \cdot \hat{q}_{sk} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

njena ničelna porazdelitev je $N(0,1)$.

Primer

Izveden je bi poskus kalivosti sorte A in sorte B. Pri sorti A je od 150 semen kalilo 57 semen, pri sorti B je od 100 semen kalilo 33 semen. Pri $\alpha = 0{,}05$ preverite domnevo, da je kalivost sorte A in sorte B enaka.

Premislimo situacijo. Posamezno seme kali ali ne kali. Število semen, ki kalijo v slučajnem vzorcu velikosti n, opišemo s slučajno spremenljivko X, za katero privzamemo binomsko porazdelitev: $X \sim b(n, p)$.

Za sorto A velja: $X_A \sim b(n_A, p_A)$, za sorto B pa velja: $X_B \sim b(n_B, p_B)$. Zapišimo ničelno in alternativno domnevo.

H_0: $p_A = p_B$ Verjetnost, da seme kali, je pri sortah A in B enaka.

H_1: $p_A \neq p_B$ Verjetnost, da seme kali, pri sortah A in B ni enaka.

Ocene iz vzorcev:

$$\hat{p}_A = \frac{57}{150} = 0{,}380$$

V vzorcu semen sorte A je kalilo 38 % semen.

$$\hat{p}_B = \frac{33}{100} = 0{,}330$$

V vzorcu semen sorte B je kalilo 33 % semen.

$$\hat{p}_{sk} = \frac{90}{250} = 0{,}36 \qquad \hat{q}_{sk} = 0{,}64$$

Če bi združili vzorca semen sorte A in B, bi bila v tem vzorcu kalivost 36%.

$$z = \frac{0{,}380 - 0{,}330}{\sqrt{0{,}36 \cdot 0{,}64 \cdot \left(\frac{1}{150} + \frac{1}{100}\right)}} = \frac{0{,}05}{0{,}0620} = 0{,}807$$

Kritični vrednosti: $\mp z_{0{,}025} = \mp 1{,}96$

Ničelno domnevo obdržimo. Eksperimentalni rezultati ne nasprotujejo domnevi o enaki kalivosti sorte A in B ($p = 0{,}4197$).

Če preverjamo ničelno domnevo, da je razlika verjetnosti p_0 različna od nič, $p_0 \neq 0$:

H_0: $p_1 - p_2 = p_0$, $p_0 \neq 0$

ne smemo uporabiti skupne verjetnosti p_{sk}. Tedaj je testna statistika:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

Ničelna porazdelitev je $N(0,1)$.

5.1.2.2 *Interval zaupanja za razliko Bernoullijevih verjetnosti

Zgornje dejstvo uporabljamo tudi pri izračunu intervala zaupanja za razliko verjetnosti. Približni interval zaupanja za razliko Bernoullijevih verjetnosti je:

$$l_{1,2} \approx (\hat{p}_1 - \hat{p}_2) \mp z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Primer

V vzorec je bilo po principu slučajnosti izbranih 150 moških, od tega jih kadi 50, in 120 žensk, od tega jih kadi 20. Izračunajmo 95% interval zaupanja za razliko odstotkov kadilcev v populaciji moških in žensk.

Ocene iz vzorcev:

$$\hat{p}_M = \frac{50}{150} = \frac{1}{3}$$

V vzorcu moških je ena tretjina kadilcev.

$$\hat{p}_Z = \frac{20}{120} = \frac{1}{6}$$

V vzorcu žensk je ena šestina kadilcev.

Točkovna ocena za razliko verjetnosti pri moških in ženskah je ena šestina (16,7%). Pripadajoča intervalna ocena za to razliko je:

$$l_{1,2} \approx (\frac{1}{3} - \frac{1}{6}) \mp 1{,}96 \cdot \sqrt{\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{150} + \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{1}{120}} = \frac{1}{6} \mp 0{,}1007$$

$l_1 \approx 0{,}066$

$l_2 \approx 0{,}267$

Pri 95% zaupanju je v populaciji moških od 7 do 27 odstotkov več kadilcev kot v populaciji žensk.

Pa poglejmo še, kakšni bi bili rezultati, če bi imeli vzorec 1500 moških in 1200 žensk, pri katerih bi veljalo, da je v pripadajočih vzorcih 1/3 kadilcev pri moških in 1/6 pri ženskah (enaka situacija kot v našem primeru, le da sta vzorca 10-krat večja). Kakšni so tu rezultati?

$l_1 \approx 0{,}1504$

$l_2 \approx 0{,}1829$

Pri 95% zaupanju je v populaciji moških od 15 do 18 odstotkov več kadilcev kot v populaciji žensk.

Komentar: vidimo, da ima velikost vzorcev bistven vpliv na širino intervala zaupanja, dejstvo, ki smo ga spoznali že prej.

NALOGE

1. Obstojnost paštete

Raziskovalci so proučevali obstojnost jetrne paštete, če jo vzamemo iz pločevinke in postavimo za 4 dni v hladilnik. V poskusu so imeli 200 pločevink od proizvajalca A in 180 pločevink od proizvajalca B. Po 4 dneh je bilo od proizvajalca A pokvarjenih 16 pločevink, od proizvajalca B pa 21 pločevink. Pri stopnji značilnosti 0,05 preverite domnevo, da je obstojnost paštete obeh proizvajalcev enaka.

2. Učinkovitost zdravil

Primerjali so učinkovitost dveh zdravil proti visokemu krvnemu tlaku. Imeli so 240 pacientov z visokim krvnim tlakom. S slučajno izbiro so bolnike razdelili na polovico, 120 pacientom so dali zdravilo A, 120 pacientom pa zdravilo B. Stanje se je izboljšalo 85 bolnikom, ki so dobili zdravilo A, in 95 bolnikom, ki so dobili zdravilo B. Pri stopnji značilnosti 0,01 primerjajte učinkovitost zdravil A in B.

3. Cepljenje proti gripi

V neki organizaciji so se delavci prostovoljno odločali, ali se želijo cepiti proti gripi. Za cepljenje se je odločilo 80 zaposlenih, od teh je kasneje za gripo zbolelo 12 zaposlenih. Od 160 necepljenih zaposlenih je zbolelo za gripo 55 oseb.

a) Podatke grafično predstavite.
b) Izračunajte 95% interval zaupanja za razliko verjetnosti obolenja z gripo pri necepljenih in pri cepljenih osebah.

4. Bakterije v mleku

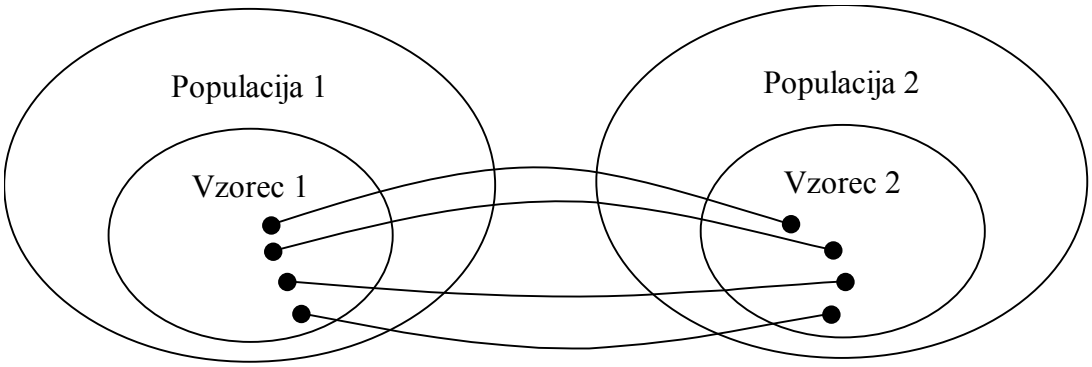
Mlekarna je opravila test na koliformne bakterije v mleku na različnih progah zbiranja. Rezultat pregleda 133 vzorcev mleka na progi A je bil 42 pozitivnih, ostali negativni; rezultat pregleda 117 vzorcev mleka na progi B pa 29 pozitivnih, ostali negativni.

Izračunajte 90% interval zaupanja za razliko okuženosti z omenjenimi bakterijami na progi A in B.

5.2 DVA ODVISNA VZORCA

Pri neodvisnih vzorcih smo vzorčenje izvedli tako, da informacija iz prvega vzorca ni povezana z informacijo iz drugega vzorca. Sedaj bomo obravnavali primer dveh odvisnih vzorcev, pri katerih velja: vsaka enota iz prve populacije je *v paru* z enoto iz druge populacije. Slika shematsko prikazuje to situacijo.

Najpogosteje gre za iste enote, ki so obravnavane v dveh različnih situacijah, npr. ob dveh različnih časih. Npr. proučujemo študente ob začetku šolskega leta in ob koncu šolskega leta; bolnike pred operacijo in po operaciji; gospodinjstva ob začetku leta in ob koncu leta ipd. Poskus, s katerim bi primerjali povprečen pridelek dveh sort pšenice A in B, je izvedljiv na različnih lokacijah. Na vsaki lokaciji damo sorto A in sorto B, torej sta sorti na vsaki lokaciji v parih. Včasih pa so pari npr. oče-sin, bolnik-kontrola, leva roka-desna roka in podobno.



Slika 5-3: Shema dveh odvisnih vzorcev: enote so v parih

5.2.1 Razlika povprečij

Izhodišče je isto kot pri dveh neodvisnih vzorcih. Predpostavimo, da je porazdelitev proučevane spremenljivke normalna:

$X_1 \sim N(\mu_1, \sigma_1)$

$X_2 \sim N(\mu_2, \sigma_2)$

Na vsakem paru izračunamo razliko vrednosti $D = X_1 - X_2$. Njena porazdelitev je

$D \sim N(\mu_1 - \mu_2 = \mu_D, \sigma_D)$

Ker sta vzorca odvisna, σ_D ne moremo izraziti s σ_1 in σ_2 (kovariančni člen ni enak 0).

Izberemo slučajni vzorec parov velikosti n. Kot smo spoznali v poglavju o enem vzorcu, je porazdelitev povprečne razlike iz n – tih parov naslednja:

$$\overline{D} \sim N(\mu_1 - \mu_2 = \mu_D, \frac{\sigma_D}{\sqrt{n}})$$

Iz gornjega sledi, da je vzorčna statistika

$$z = \frac{\overline{d} - \mu_D}{\sigma_d / \sqrt{n}}$$

porazdeljena po N(0, 1). Če pa v tej formuli σ_d nadomestimo z njeno oceno s_d, standardizirano normalno porazdelitev nadomesti Studentova porazdelitev. Torej je vzorčna statistika zapisana v obliki:

$$t = \frac{\overline{d} - \mu_D}{s_d / \sqrt{n}}$$

njena porazdelitev pa $t(SP = n-1)$, stopinje prostosti so enake številu parov, zmanjšano za 1, torej n-1.

Ostane še vprašanje, kako izračunati oceno $\overline{d}$ in s_d. Odgovor na to vprašanje že poznamo. Na vsakem paru $i, i = 1, ..., n$, izračunamo razliko vrednosti $d_i = x_{i1} - x_{i2}$. Iz teh razlik izračunamo aritmetično sredino razlik

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

in standardni odklon razlik:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} d_i^2 - \frac{1}{n}\cdot\left(\sum_{i=1}^{n} d_i\right)^2\right)}$$

5.2.1.1 Preizkus domneve o razliki povprečij

Preizkus domneve o razliki povprečij v tej situaciji prevedemo na preizkus domneve o povprečni vrednosti, le da gre tu za povprečje razlik in ne za povprečje podatkov. Torej še enkrat: v tem primeru je analizirana spremenljivka razlika vrednosti na parih. Zato velikokrat imenujemo ta preizkus tudi **preizkus parov**.

$H_0: \mu_D = \mu_{D_0}$ Povprečje razlik je $\mu_{D_0} \in R$.

$H_1: \mu_D \neq \mu_{D_0}$ Povprečje razlik ni μ_{D_0}.

Opomba: alternativna domneva je lahko enostranska.

Testna statistika je:

$$t = \frac{\overline{d} - \mu_{D_0}}{s_d / \sqrt{n}}$$

Ničelna porazdelitev je $t(SP = n-1)$. Stopinje prostosti so vezane na število parov.

Primer

V laboratoriju sta dve tehtnici A in B, za kateri laboranti sumijo, da v povprečju ne kažeta enako. Da bi ugotovili, ali je to res, so izvedli poskus: 12 predmetov so stehtali na obeh tehtnicah in dobili naslednje rezultate (g):

Tabela 5-2: Masa predmetov na tehtnici A in na tehtnici B ter njuna razlika

Predmet	A	B	razlika A-B
1	12,13	12,17	-0,04
2	17,56	17,61	-0,05
3	9,33	9,35	-0,02
4	11,40	11,42	-0,02
5	28,62	28,61	0,01
6	10,25	10,27	-0,02
7	23,37	23,42	-0,05
8	16,27	16,26	0,01
9	12,40	12,45	-0,05
10	24,78	24,75	0,03
11	12,55	12,60	-0,05
12	26,33	26,38	-0,05

Pri stopnji značilnosti 0,05 preverite domnevo laborantov.

Rešitev:

$H_0: \mu_D = 0$ Povprečje razlik je 0.

$H_1: \mu_D \neq 0$ Povprečje razlik ni 0.

Izračuni:

$\overline{d} = -0{,}0250$

$s_d = 0{,}02844$

$$t = \frac{-0{,}0250}{0{,}02844/\sqrt{12}} = -3{,}04$$

Kritični vrednosti: $\mp t_{0{,}025}(11) = \mp 2{,}201$

Statistični sklep: ničelno domnevo zavrnemo v korist alternativne.

Odgovor: pri stopnji značilnosti 0,05 trdimo, da tehtnici v povprečju ne kažeta enako ($p = 0{,}0111$).

5.2.1.2 *Interval zaupanja za razliko povprečij

Na osnovi predhodnega je razvidno, da interval zaupanja za $\mu_D = \mu_1 - \mu_2$ izračunamo takole:

$$l_{1,2} = \overline{d} \mp t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s_d}{\sqrt{n}}$$

Primer

Da bi ugotovili vsebnost žvepla v smrekovih iglicah, so raziskovalci na območju Mežiške doline s slučajno izbiro izbrali 20 dreves. Na vsakem drevesu so izbrali sedemletno vreteno. Za analizo so uporabili potrgane enoletne in dvoletne iglice. Podatki o količini žvepla v iglicah (mg/m^3) so v tabeli.

Tabela 5-3: Podatki o količini žvepla v iglicah (mg/m^3)

Drevo	enoletne iglice	dvoletne iglice
1	16,3	16,3
2	25,4	25,5
3	10,6	10,5
4	30,1	33,9
5	18,1	26,1
6	20,5	33,9
7	11,8	14,7
8	13,5	20,6
9	6,6	11,1
10	12,5	13,6
11	36,6	40,6
12	6,8	7,7
13	9,4	12,5
14	35,9	40,5
15	8,9	10,5
16	12,9	18,3
17	20,6	21,5
18	10,8	43,4
19	24,4	36,7
20	29,1	40,8

Izračunajmo 95% interval zaupanja za razliko povprečne količine žvepla v dvoletnih in v enoletnih iglicah.

Izračuni:

n=20

$\bar{d} = 5{,}895 \;\; s_d = 7{,}51955 \;\; t_{0{,}025}(19) = 2{,}093$

$l_{1,2} = 5{,}895 \mp 3{,}5193$

$l_1 = 2{,}3757$

$l_2 = 9{,}4143$

Pri 95% zaupanju trdimo, da je povprečna količina žvepla v dvoletnih iglicah od 2,4 do 9,4 mg/m^3 večja kot v enoletnih iglicah.

NALOGE

1. Koncentracija škroba v krompirju

Z dvema različnima metodama so izmerili koncentracijo škroba v krompirjih. Privzamemo lahko, da so izmerki pri obeh metodah normalno porazdeljeni. V vzorcu je bilo 15 krompirjev. Vsak krompir so razpolovili, polovico analizirali po prvi metodi, polovico po drugi metodi. Dobili so naslednje razlike med rezultati meritev po prvi in po drugi metodi: -2, 1, 0, 1, 2, 2, 3, 3, 1, 2, 5, 0, 1, -2, 4.

a) Preizkusite s temi podatki domnevo, da dobimo z obema metodama isto povprečno vrednost (stopnja značilnosti 0,05).

b) Ocenite *p*-vrednost.

c) Obrazložite rezultate naloge.

2. Nabrek ivernih plošč

Raziskovalci so proučevali, kako vlaga vpliva na nabrek ivernih plošč, ki jih vgrajujejo v okolje. V obratu, kjer izdelujejo iverne plošče, so naključno izbrali 10 plošč. Poskus so začeli s tem, da so plošče potopili v kad z destilirano vodo, tako da se niso dotikale med seboj. Po 30 minutah so plošče odcedili in jim izmerili debelino. Nato so jih potopili še enkrat in izmerili njihovo debelino po 60 minutah od začetka poskusa.

Tabela 5-4: Debelinski nabrek (mm) v odvisnosti od časa namakanja

Plošča	Čas namakanja 30 min	Čas namakanja 60 min
1	4,88	7,54
2	3,65	6,55
3	3,84	6,88
4	4,24	6,39
5	3,52	6,98
6	3,03	5,19
7	2,41	4,20
8	3,46	6,55
9	3,46	5,93
10	3,89	6,87

Preizkusite domnevo, da se v času namakanja od 30 min na 60 min debelinski nabrek v povprečju poveča za 3 mm (stopnja značilnosti 0,05). Obrazložite rezultate.

3. *Plavanje

Prostovoljci so se udeležili trimesečnega tečaja plavanja. Njihove plavalne sposobnosti so preverili na začetku in na koncu tečaja, tako da so neprekinjeno plavali 12 minut v poljubni plavalni tehniki. Merili so število metrov, ki so jih plavalci preplavali v 12 minutah. Njihovi rezultati (v metrih) so:

Tabela 5-5: Preplavana dolžina ob začetku in koncu tečaja

Plavalec	Na začetku tečaja	Ob koncu tečaja
1	375	400
2	500	600
3	475	525
4	550	575
5	375	400
6	300	435
7	500	575
8	425	450
9	500	550
10	480	525
11	475	560
12	400	500
13	500	575
14	575	600
15	400	425

a) Izračunajte 95% interval zaupanja za povprečno razliko preplavane dolžine na koncu in na začetku tečaja in ga obrazložite.

b) Trenerka domneva, da udeleženci takega tečaja v povprečju izboljšajo rezultat neprekinjenega plavanja na 12 minut za 50 metrov. Ali so rezultati v skladu s trditvijo trenerke?

4. *Prirast debla

Raziskovalci so proučevali prirast debla pri breskvah. Podatki za obseg debla leta 1995 in 1996 za vzorec 12 dreves breskev cv. 'elegant lady' v nasadu Bilje (Vir: M. Brus, Diplomsko delo) so v tabeli.

Tabela 5-6: Obseg debla leta 1995 in leta 1996

Drevo	Obseg 1995 (mm)	Obseg 1996 (mm)
1	95	159
2	94	167
3	85	155
4	82	148
5	55	123
6	63	110
7	35	91
8	65	122
9	50	130
10	52	127
11	67	112
12	35	78

a) Za vsako drevo izračunajte prirast debla od leta 1995 do 1996.

b) Izračunajte 90% interval zaupanja za prirast debla in ga obrazložite.

5.2.2 *Razlika Bernoullijevih verjetnosti

Izhajamo iz istih izhodišč kot pri neodvisnih vzorcih, vendar so enote prve in druge populacije odvisne, so v parih, n označuje velikost vzorca teh parov. Predpostavke so naslednje:

$X_1 \sim b(n, p_1) \approx N(np_1, \sqrt{np_1q_1})$

$X_2 \sim b(n, p_2) \approx N(np_2, \sqrt{np_2q_2})$

Za vsak par imamo informacijo o stanju pri X_1 (A ali $\overline{A}$) in informacijo o stanju pri X_2 (A ali $\overline{A}$). Torej ima posamezni par 4 možna stanja. Verjetnost za stanje $X_1 = A$ in $X_2 = A$ označimo p_{11}, verjetnost za stanje $X_1 = A$ in $X_2 = \overline{A}$ označimo p_{12}, itd. Bolj pregledno so oznake predstavljene v tabeli:

Tabela 5-7: Verjetnosti za štiri možna stanja

	$X_2 = A$	$X_2 = \overline{A}$	**Skupaj**
$X_1 = A$	p_{11}	p_{12}	p_1
$X_1 = \overline{A}$	p_{21}	p_{22}	q_1
Skupaj	p_2	q_2	1

Verjetnostni račun pokaže, da je razlika deležev porazdeljena normalno:

$$\frac{X_1}{n} - \frac{X_2}{n} \approx N\left(p_1 - p_2, \sqrt{\frac{p_1q_1}{n} + \frac{p_2q_2}{n} - 2\frac{p_{11}p_{22} - p_{12}p_{21}}{n}} \right)$$

Opomba: če primerjamo ta izraz z izrazom, ki smo ga dobili pri neodvisnih vzorcih, vidimo, da se prvi parameter ujema, drugi pa v prvih dveh členih pod korenom. Odvisnost vzorcev se izraža s tretjim členom pod korenom.

Poglejmo, kako izračunamo ocene za te verjetnosti. Podatke uredimo v frekvenčno tabelo. Število parov, pri katerih je stanje $X_1 = A$ in $X_2 = A$, označimo f_{11}, itd. Oznake so v tabeli.

Tabela 5-8: Shema frekvenčne tabele

	$X_2 = A$	$X_2 = \overline{A}$	**Skupaj**
$X_1 = A$	f_{11}	f_{12}	$f_{11} + f_{12}$
$X_1 = \overline{A}$	f_{21}	f_{22}	$f_{21} + f_{22}$
Skupaj	$f_{11} + f_{21}$	$f_{12} + f_{22}$	n

Izračun ocen:

- ocena za p_1: $\hat{p}_1 = \frac{f_{11} + f_{12}}{n}$
- ocena za p_2: $\hat{p}_2 = \frac{f_{11} + f_{21}}{n}$
- ocena za $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 = \frac{f_{12} - f_{21}}{n}$

- ocena za $\sqrt{\frac{p_1q_1}{n}+\frac{p_2q_2}{n}-\frac{p_{11}p_{22}-p_{12}p_{21}}{n}}$: $s(\hat{p}_1-\hat{p}_2)=\sqrt{\frac{f_{12}+f_{21}}{n^2}}$

5.2.2.1 *Preizkus domneve o razliki Bernoullijevih verjetnosti

Z dobljenimi spoznanji poglejmo testno statistiko.

H_0: $p_1 - p_2 = p_0$

H_1: $p_1 - p_2 \neq p_0$

Testna statistika je:

$$z=\frac{\frac{f_{12}-f_{21}}{n}-p_0}{\sqrt{\frac{f_{12}+f_{21}}{n^2}}}$$

Ničelna porazdelitev je $N(0,1)$.

Testna statistika se poenostavi v primeru, kadar nas zanima, ali sta verjetnosti enaki.

H_0: $p_1 = p_2$

H_1: $p_1 \neq p_2$

$$z=\frac{f_{12}-f_{21}}{\sqrt{f_{21}+f_{12}}}$$

Informacija, ki jo uporabimo pri tem testu, je v enotah, pri katerih je stanje $(A, \overline{A})$ oz. $(\overline{A}, A)$, torej v izvendiagonalnih elementih frekvenčne tabele. Enote, pri katerih je stanje (A, A) oz. $(\overline{A}, \overline{A})$, ne vsebujejo informacije za ta preizkus. Ta preizkus se imenuje **McNemarjev preizkus simetrije**.

Primer

Raziskovalci so preučevali univerzitetno izobrazbo očetov in njihovih sinov. Zanimalo jih je, ali je verjetnost, da imata oče in sin univerzitetno izobrazbo, enaka. V vzorcu je bilo 230 parov oče – sin. Podatki so v tabeli:

Tabela 5-9: Univerzitetna izobrazba očetov in sinov

	Oče DA	**Oče NE**	**Skupaj**
Sin DA	50	50	100
Sin NE	20	110	130
Skupaj	70	160	230

Pri stopnji značilnosti 0,05 preizkusite zgornjo domnevo.

H_0: $p_1 = p_2$ Verjetnost univerzitetne izobrazbe je enaka pri očetu in sinu.

H_1: $p_1 \neq p_2$ Verjetnost univerzitetne izobrazbe ni enaka pri očetu in sinu.

Izračuni:

$$\hat{p}_1=\frac{70}{230}=0{,}3043$$

$$\hat{p}_2=\frac{100}{230}=0{,}4348$$

V vzorcu 230 parov oče – sin je imelo univerzitetno izobrazbo 30,4% očetov in 43,5% sinov.

$$z = \frac{50 - 20}{\sqrt{50 + 20}} = 3{,}59$$

Kritični vrednosti: $\mp z_{0{,}025} = \mp 1{,}96$

Ničelno domnevo zavrnemo v korist alternativne domneve. Pri stopnji značilnosti 0,05 trdimo, da verjetnost univerzitetne izobrazbe očeta in sina ni enaka (p=0,0004).

5.2.2.2 *Interval zaupanja za razliko Bernoullijevih verjetnosti

Interval zaupanja za $p_1 - p_2$ je:

$$l_{1,2} = (\frac{f_{12} - f_{21}}{n}) \mp z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f_{12} + f_{21}}{n^2}}$$

Primer

V vzorec je bilo izbranih 1600 volilnih upravičencev. Postavili so jim vprašanje: 'Ste zadovoljni s predsednikom?' Dopuščena odgovora sta bila Da oz. Ne. Prvo anketo so ponovili po enem letu na istih volilnih upravičencih in dobili naslednje odgovore:

Tabela 5-10: Odgovori 1600 anketirancev na vprašanje: 'Ste zadovoljni s predsednikom?'

	Druga anketa DA	Druga anketa NE	Skupaj
Prva anketa DA	794	150	944
Prva anketa NE	86	570	656
Skupaj	880	720	1600

Na osnovi odgovorov so želeli oceniti, za koliko se je spremenilo zadovoljstvo volilnega telesa s predsednikom (stopnja značilnosti 0,05).

Izračunali bomo ustrezni 95% interval zaupanja.

$$\hat{p}_1 = \frac{944}{1600} = 0{,}59$$

Ob prvi anketi je bilo s predsednikom zadovoljnih 59% anketiranih.

$$\hat{p}_2 = \frac{880}{1600} = 0{,}55$$

Ob drugi anketi je bilo s predsednikom zadovoljnih 55% anketiranih.

$$\hat{p}_1 - \hat{p}_2 = \frac{150 - 86}{1600} = 0{,}04$$

$$s(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{150 + 86}{1600^2}} = 0{,}0096$$

Pri anketiranih je zadovoljstvo padlo za 4% od prve do druge ankete.

95% interval zaupanja za razliko odstotkov zadovoljnih s predsednikom v volilnem telesu:

$$l_{1,2} = 0{,}04 \mp 1{,}96 \cdot 0{,}0096$$

$$l_1 = 0{,}0212$$

$$l_2 = 0{,}0588$$

Pri 95% zaupanju je odstotek zadovoljnih s predsednikom v času od prve do druge ankete padel, in sicer od 2% do 6%.

NALOGE

1. Vremenska napoved

Vremenoslovci so preučevali, kako se verjetnost napovedi padavin ujema z dejanskim stanjem. Za vzorec so vzeli eno koledarsko leto in za vsak dan pogledali, ali je napoved predvidela padavine ali ne ter ali so dejansko padavine bile ali ne. Podatki so:

	Napoved DA	**Napoved NE**	**Skupaj**
Dejansko DA	140	25	165
Dejansko NE	40	160	200
Skupaj	180	185	365

Pri stopnji značilnosti 0,05 preverite domnevo, da se verjetnost napovedi padavin ujema z dejansko verjetnostjo padavin. Izračunajte *p*-vrednost.

2. Vpliv kajenja

Zdravniki so izvedli t. i. študijo primerov s kontrolami. V vzorcu je bilo 490 parov bolnik – kontrola. Bolnik je bil izbran iz populacije moških, obolelih za rakom na pljučih. Njegova kontrola je bila izbrana iz populacije zdravih, vendar je bila z bolnikom usklajena po starosti. Namen študije je bil ugotavljati vpliv kajenja na pojavnost raka na pljučih. Podatki so:

	Bolnik kadi	**Bolnik ne kadi**	**Skupaj**
Kontrola kadi	20	29	49
Kontrola ne kadi	431	10	441
Skupaj	451	39	490

Izračunajte 95% interval zaupanja za razliko odstotkov kadilcev pri bolnikih in kontrolah.

5.3 PREGLED PREIZKUSOV ZA NEODVISNA IN ZA ODVISNA VZORCA

Izbira statističnega preizkusa mora upoštevati:

- smiselne predpostavke za slučajno spremenljivko
- namen analize
- način zbiranja podatkov.

Podajamo kratek pregled statističnih preizkusov, ki smo jih obravnavali v tem poglavju.

Tabela 5-11: Pregled statističnih preizkusov za dve populaciji

Predpostavke	**Neodvisna vzorca velikosti n_1 in n_2**	**Odvisna vzorca: n parov**
$X_i \sim N(\mu_i, \sigma_i)$, $i = 1, 2$	$H_0: \delta = \mu_1 - \mu_2 = \delta_0$ Predpostavka : $\sigma_1 = \sigma_2$ $t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_{sk}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathrm{t}(n_1 + n_2 - 2)$ $H_0: \delta = \mu_1 - \mu_2 = \delta_0$ Predpostavka : $\sigma_1 \neq \sigma_2$ $t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t(\min(n_1 - 1, n_2 - 1))$	$H_0: \mu_D = \mu_{D_0}$ $t = \frac{\bar{d} - \mu_{D_0}}{s_d / \sqrt{n}} \sim \mathrm{t}(n-1)$
$X_i \sim b(n_i, p_i) \approx$ $\approx N(n_i p_i, \sqrt{n_i p_i q_i})$, $i = 1, 2$	$H_0: p_1 = p_2$ $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{sk} \cdot \hat{q}_{sk} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathrm{N}(0,1)$ $H_0: p_1 - p_2 = p_0, \ p_0 \neq 0$ $z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \sim \mathrm{N}(0,1)$	$H_0: p_1 = p_2$ $z = \frac{f_{12} - f_{21}}{\sqrt{f_{21} + f_{12}}} \sim \mathrm{N}(0,1)$ $H_0: p_1 - p_2 = p_0, p_0 \neq 0$ $z = \frac{\frac{f_{12} - f_{21}}{n} - p_0}{\sqrt{\frac{f_{12} + f_{21}}{n^2}}} \sim \mathrm{N}(0,1)$

6 REGRESIJA IN KORELACIJA

6.1 UVOD

Sedaj bomo proučevali primere, ko na vsaki enoti gledamo po dve številski spremenljivki hkrati. Našo prvo pozornost bomo usmerili v relacijo med spremenljivkama. Pomembni relaciji sta: **odvisnost** in **povezanost** (**soodvisnost**).

Pod pojmom odvisnost razumemo relacijo, kjer vrednosti ene spremenljivke vplivajo na vrednosti druge spremenljivke, v drugo smer pa vpliva ni. Rečemo, da je ena spremenljivka odvisna od druge. Standardne oznake in poimenovanja za relacijo odvisnosti so: z oznako Y označujemo odvisno spremenljivko, z oznako X pa neodvisno spremenljivko.

Primer: nadmorska višina kraja in količina padavin v kraju. Nadmorska višina vpliva na količino padavin, količina padavin pa ne vpliva na nadmorsko višino kraja. Količina padavin je odvisna (tudi) od nadmorske višine. Pri izbrani vrednosti nadmorske višine x je količina padavin različna, modeliramo jo lahko s slučajno spremenljivko $Y \mid x$, ki ima določeno verjetnostno porazdelitev. Npr. pri krajih z nadmorsko višino 500 m količina padavin variira, morda je normalna porazdelitev sprejemljiva verjetnostna porazdelitev za to spremenljivko.

Namen študija odvisnosti je:

- pridobiti nova spoznanja o odvisnosti
- napovedovanje vrednosti odvisne spremenljivke Y pri izbrani vrednosti neodvisne spremenljivke x_0.

Pod pojmom povezanost oz. soodvisnost razumemo relacijo, ko se vrednosti obeh spremenljivk spreminjajo hkrati. Rečemo, da sta spremenljivki povezani ali soodvisni.

Primer: višina in teža otroka. Ko otrok raste, se spreminjata višina in teža otroka hkrati. Višina in teža otroka sta povezani (soodvisni) spremenljivki.

V primeru povezanosti sta obe spremenljivki enakovredni. Privzeli bomo oznake X in Y, vendar je v tem primeru vseeno, kateri spremenljivki damo oznako X in kateri Y. Namen študija povezanosti je izračunati ustrezno mero, ki vrednoti jakost povezanost dveh spremenljivk. Teh mer je veliko, ogledali si bomo dve.

6.2 ENOSTAVNA LINEARNA REGRESIJA

6.2.1 Izračun ocen in napovedi

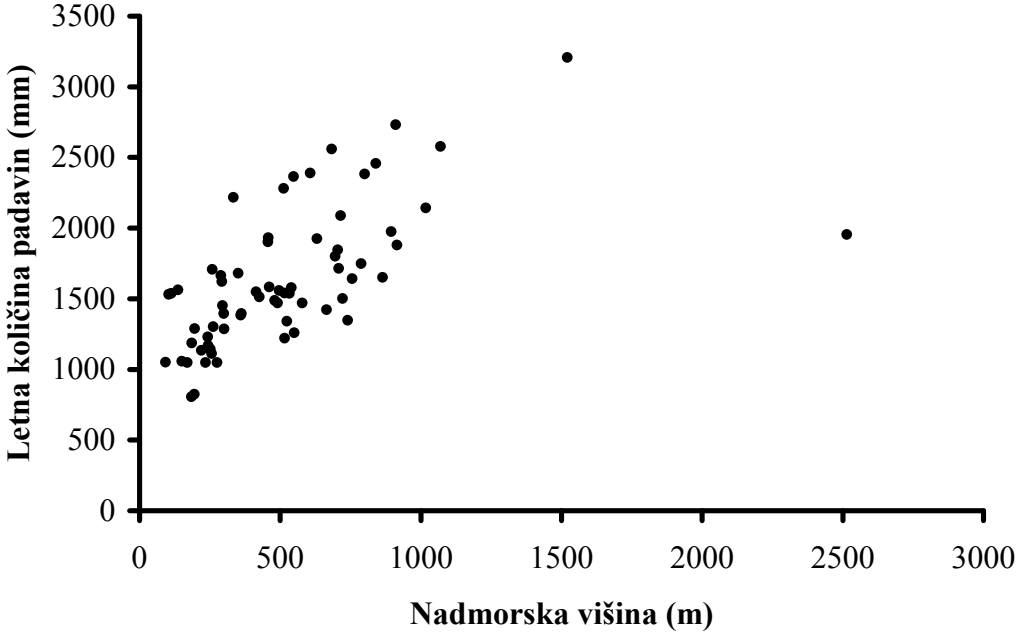
Regresija je prilagajanje ustrezne matematične funkcije empiričnim podatkom. To funkcijo imenujemo regresijska funkcija. Lahko je zelo enostavna (npr. linearna), lahko pa bolj kompleksna.

V tem delu bomo obravnavali primer, ko sta odvisna in neodvisna spremenljivka številski. Regresijska funkcija je linearna (linearna regresija), proučujemo vpliv samo ene neodvisne spremenljivke X na Y (enostavna regresija).

V vzorcu je n enot, na i-ti enoti imamo par vrednosti (x_i, y_i): x_i za neodvisno spremenljivko in za y_i za odvisno spremenljivko. Podatke najprej grafično prikažemo z **razsevnim grafikonom**: na ordinati je odvisna spremenljivka, na abscisi pa neodvisna. Vsaka enota je predstavljena z eno točko.

Primer

Za 67 meteoroloških postaj v Sloveniji imamo podatke za nadmorsko višino (m) in letno količino padavin v letu 1992 (mm). Nadmorska višina je neodvisna spremenljivka, njene vrednosti so na abscisni osi, količina padavin je odvisna spremenljivka, njene vrednosti so na ordinati. Vsaka meteorološka postaja je grafično predstavljena z eno točko.



Slika 6-1: Letna količina padavin v odvisnosti od nadmorske višine za 67 meteoroloških postaj v Sloveniji (Vir: Arhiv Hidrometeorološki zavod Slovenije)

Slika kaže, da se s povečevanjem nadmorske višine povečuje tudi količina padavin. Slika nakazuje, da je odvisnost linearna. Dve točki štrlita ven iz oblaka točk: ena pri nadmorski višini okoli 1500 m (Komna), ena pri nadmorski višini okoli 2500 m (Kredarica). Količina padavin na Kredarici je bistveno nižja, kot bi pričakovali na osnovi ostalih podatkov in na osnovi splošnega znanja o padavinah v gorah. Dodatno poizvedovanje je pokazalo, da gre za meritveno napako, ki je posledica posebnih vremenskih razmer na Kredarici (veter odpiha sneg). Zato je smiselno to postajo izločiti iz analize.

Na osnovi slike poskusimo ugotoviti, katera matematična funkcija najbolje opiše odvisnost spremenljivke Y od spremenljivke X. Če slika kaže, da se premica dovolj dobro prilega točkam, uporabimo *model enostavne linearne regresije*. Ta pravi, da je v opazovani populaciji vrednost odvisne spremenljivke vsota treh členov: konstante α, večkratnika neodvisne spremenljivke βX in t. i. slučajnih (neznanih, nepojasnjenih) vplivov ε:

$$Y = \alpha + \beta \cdot X + \varepsilon$$

Količini α in β sta *parametra enostavnega linearnega modela*. Opomba: doslej smo oznaki α in β spoznali v smislu verjetnosti za napako I. vrste in II. vrste, v tem primeru imata drugačen pomen.

Populacijo predstavlja vzorec velikosti *n*, na osnovi katerega izračunamo oceno za α in oceno za β. Ti dve oceni označimo a in b. **Ocena regresijske premice** $\hat{Y}$ je:

$$\hat{Y} = a + b \cdot X$$

Kako dobiti oceni parametrov? Izhajali bomo iz kriterija, naj se premica 'čim bolje' prilega točkam. Bolj natančno povedano: odkloni točk od premice $(y_i - \hat{y}_i)$ naj bodo čim manjši. Nekateri odkloni so pozitivni, nekateri pa negativni, zato je smiselno (tako kot pri izračunu variance) odklone kvadrirati. Carl Friedrich Gauss (1777-1855) je postavil kriterij, ki pravi, da naj bo premica postavljena tako, da je vsota kvadratov odklonov minimalna. Ta kriterij imenujemo **metoda najmanjših kvadratov**. Matematično ga zapišemo takole:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2 = \min S(a,b)$$

Točko, v kateri ima funkcija $S(a,b)$ minimalno vrednost, dobimo s parcialnim odvajanjem:

$$\frac{\partial S(a,b)}{\partial a} = 0$$

$$\frac{\partial S(a,b)}{\partial b} = 0$$

Rešitev za a in b dobimo iz sistema dveh linearnih enačb z dvema neznankama. Sistema ne navajamo, njegova rešitev je:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Količino v imenovalcu za *b* že poznamo, to je vsota kvadriranih odklonov za neodvisno spremenljivko, ki smo jo spoznali pri merah variabilnosti, označimo jo VKO_{xx}**.** Vemo, da je vedno pozitivna količina. V števcu pa je podobna količina, označimo jo VKO_{xy}, le-ta je lahko pozitivna ali negativna. S temi oznakami krajše zapišemo izračun ocen linearnega regresijskega modela:

$$b = \frac{VKO_{xy}}{VKO_{xx}}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Vrednost a je odsek na ordinati: če je vrednost neodvisne spremenljivke 0, je vrednost odvisne spremenljivke a. V določenih primerih nima vsebinskega pomena.Vrednost b predstavlja tangens naklonskega kota premice in pomeni: če se X poveča za eno enoto, se povprečje za Y spremeni za b enot.

Kako najlažje narišemo regresijsko premico? Za grafični prikaz regresijske premice potrebujemo dve točki. Iz izraza za izračun a sledi: $\bar{y} = a + b\bar{x}$, kar pomeni, da točka $(\bar{x}, \bar{y})$ leži na regresijski premici. Za drugo točko izberemo poljubno točko x_0 na razsevnem grafikonu in izračunamo $\hat{y}(x_0)$.

Opomba: mnogi kalkulatorji znajo izračunati ocene linearnega regresijskega modela. Če imate tak kalkulator, se naučite uporabljati njegove funkcije.

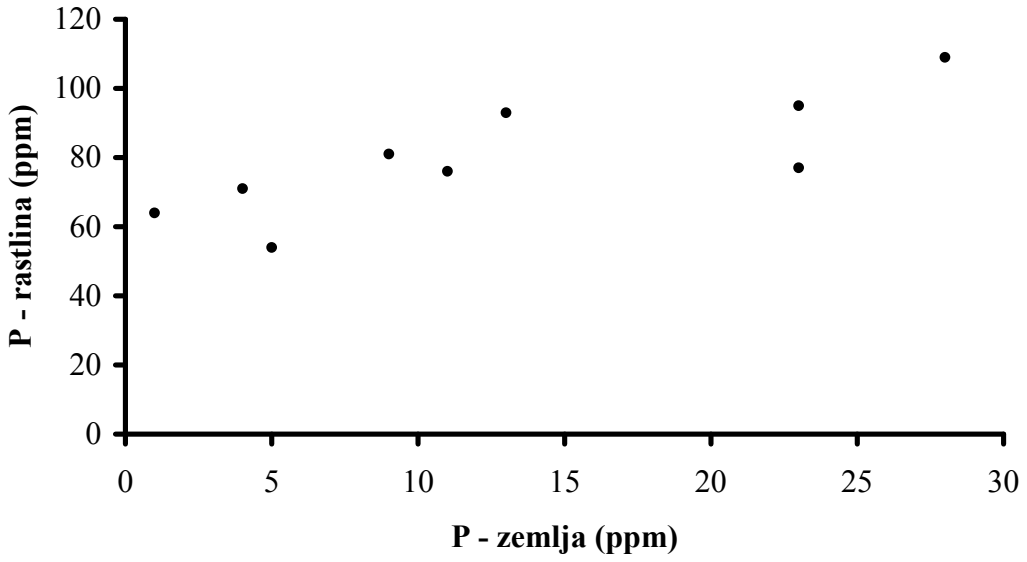
Primer (1.del)

V poskus je bilo vključenih 9 lončkov z zemljo, vanje so posadili rastline. Vsebnost fosforja v zemlji so izmerili na začetku poskusa, vsebnost fosforja v rastlini pa so merili po 38 dneh rasti. Analiziramo odvisnost vsebnosti fosforja v rastlini od vsebnosti fosforja v zemlji.

Tabela 6-1: Rezultati meritev vsebnosti fosforja v zemlji in v rastlini

Lonček	P-zemlja (ppm)	P-rastlina (ppm)
1	1	64
2	4	71
3	5	54
4	9	81
5	13	93
6	11	76
7	23	77
8	23	95
9	28	109

Podatke najprej grafično predstavimo z razsevnim grafikonom:



Slika 6-2: Vsebnost fosforja v rastlini v odvisnosti od vsebnosti fosforja v zemlji za 9 poskusnih lončkov

Grafični prikaz kaže, da se z večanjem vsebnosti fosforja v zemlji veča tudi vsebnost fosforja v rastlini. Točk je malo, premalo za sklep o vrsti odvisnosti. Slika nakazuje, da je zveza linearna. Izračunajmo oceni parametrov linearnega regresijskega modela a in b.

Pomožni računi:

$n = 9$

$$\sum x = 117 \quad \sum x^2 = 2255 \quad \bar{x} = 13$$

$$VKO_{xx} = 2255 - \frac{117^2}{9} = 734$$

$$\sum y = 720 \quad \sum y^2 = 59874 \quad \bar{y} = 80$$

$$VKO_{yy} = 59874 - \frac{720^2}{9} = 2274$$

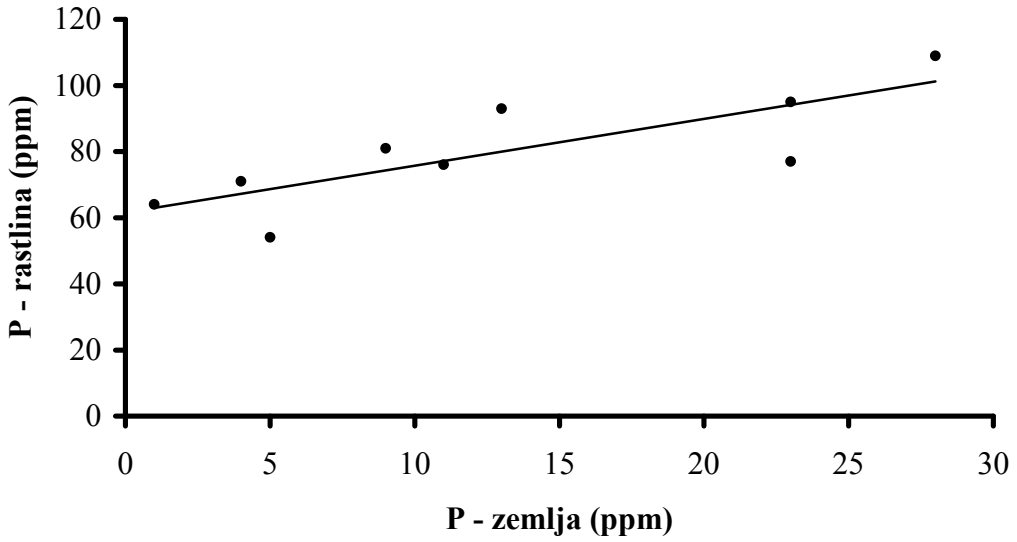
$$\sum xy = 10400$$

$$VKO_{xy} = 10400 - \frac{117 \cdot 720}{9} = 1040$$

$$b = \frac{1040}{734} = 1{,}417$$

$$a = 80 - 1{,}417 \cdot 13 = 61{,}580$$

Ocena linearnega regresijskega modela: $\hat{Y} = 61{,}580 + 1{,}417 \cdot X$



Slika 6-3: Vsebnost fosforja v rastlini v odvisnosti od vsebnosti fosforja v zemlji za 9 poskusnih lončkov in linearni regresijski model, dobljen na osnovi teh podatkov

Interpretacija ocen linearnega regresijskega modela:

- vrednost a=61,580 nima vsebinskega pomena, saj strokovnjaki pravijo, da je v zemlji vedno nekaj fosforja;
- če se vsebnost fosforja v zemlji poveča za 1 ppm, se vsebnost fosforja v rastlini poveča za 1,42 ppm.

Eden od bistvenih namenov regresijske analize so *napovedi* odvisne spremenljivke pri izbranih vrednostih neodvisne spremenljivke. Preden se lotimo napovedi, razmislimo, kje je dobljeni regresijski model veljaven. Ocene za parametre modela izračunamo iz vrednosti spremenljivk na določenem območju. Zato je regresijski model veljaven le na območju, ki ga določajo vrednosti neodvisne spremenljivke v podatkih. To

pomeni, da smemo napovedi računati le na tem območju, izven tega območja pa je lahko zveza drugačna. To spoznanje velja splošno: model je veljaven le na območju, kjer imamo podatke, ki smo jih uporabili za ocenjevanje parametrov modela.

Primer (2. del)

Dopolnimo zgornje spoznanje: če se vsebnost fosforja v zemlji poveča za 1 ppm, se vsebnost fosforja v rastlini poveča za 1,42 ppm. Ta relacija velja na območju, kjer imamo podatke za vsebnost fosforja v zemlji, torej na območju med 1 ppm in 28 ppm. Na tem območju lahko izračunamo napovedi po dobljenem modelu. Izven tega območja pa napovedi niso upravičene.

Npr. kolikšna je napovedana vsebnost fosforja v rastlini, če je vsebnost fosforja v zemlji 10 ppm?

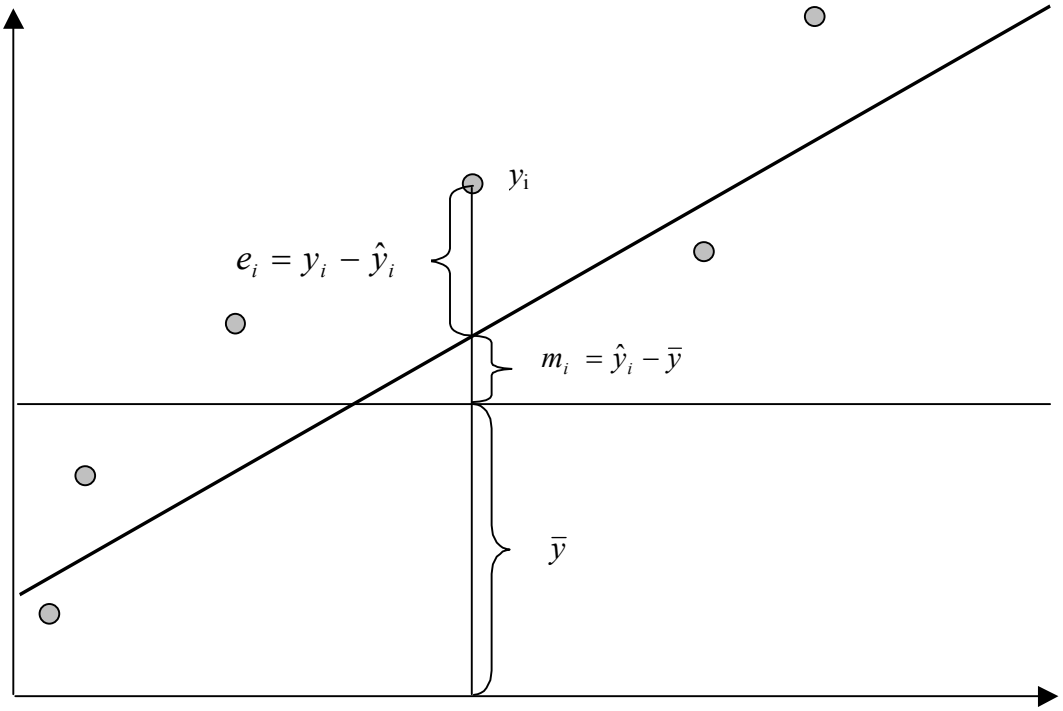
$\hat{y}(x = 10) = 61{,}58 + 1{,}417 \cdot 10 = 75{,}75$

Napovedana vsebnost fosforja v rastlini je 75,8 ppm.

Kolikšna je napovedana vsebnost fosforja v rastlini, če je vsebnost fosforja v zemlji 50 ppm? Na osnovi dobljenega modela bi bilo nekorektno izračunati to napoved, saj informacije o tem, kaj se dogaja s fosforjem v rastlini, ko je vsebnost fosforja v zemlji nad 28 ppm, nimamo. Poglejmo možno situacijo. Ko se vsebnost fosforja v zemlji povečuje preko določene vrednosti (npr. preko 40 ppm), se vsebnost fosforja v rastlini ne povečuje več. Seveda bi bil sklep: 'če se vsebnost fosforja v zemlji poveča za 1 ppm, se vsebnost fosforja v rastlini poveča za 1,42 ppm', za te vrednosti neodvisne spremenljivke napačen.

6.2.2 Koeficient determinacije

Regresijski model je lahko boljši ali slabši. Kakovost modela lahko vrednotimo na različne načine. Ena izmed najenostavnejših mer, ki vrednoti kakovost regresijskega modela, je **koeficient determinacije**. Prikazali bomo, kako pridemo do izraza za izračun koeficienta determinacije.



Slika 6-4: Grafični prikaz enačbe $y_i = \bar{y} + m_i + e_i$

Vrednost y_i zapišemo kot vsoto treh členov: povprečja odvisne spremenljivke $\bar{y}$, odseka m_i, ki ga določa premica, ter odklona od premice e_i:

$$y_i = \bar{y} + m_i + e_i$$

Ta izraz lahko zapišemo takole:

$$y_i - \bar{y} = m_i + e_i = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Kvadriramo levo in desno stran, seštejemo po vseh točkah in po nekaj korakih dobimo:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Člen na levi prestavlja vsoto kvadriranih odklonov za odvisno spremenljivko VKO_{yy}. Prvi člen na desni je vsota kvadriranih odklonov, ki ga določa regresijski model, označili ga bomo VKO_{mod}. Drugi člen na desni pa je vsota kvadriranih odklonov točk od premice. To količino imenujemo vsota kvadriranih odklonov za ostanek, označimo jo VKO_{ost}, in predstavlja nekakšno napako regresijskega modela. Če bi vse točke ležale na regresijski premici, bi bil $VKO_{ost} = 0$.

Tako smo dobili osnovno relacijo:

$$VKO_{yy} = VKO_{\text{mod}} + VKO_{\text{ost}}$$

Kratek povzetek izračunov:

$$VKO_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$VKO_{\text{mod}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b^2 \cdot VKO_{xx} = \frac{VKO_{xy}^2}{VKO_{xx}}$$

$$VKO_{\text{ost}} = VKO_{yy} - VKO_{\text{mod}}$$

Koeficient determinacije je razmerje

$$r^2 = \frac{VKO_{\text{mod}}}{VKO_{yy}},$$

navadno se izraža v odstotkih:

$$r^2\% = \frac{VKO_{\text{mod}}}{VKO_{yy}} \cdot 100$$

Koeficient determinacije izraža odstotek variabilnosti odvisne spremenljivke, ki je pojasnjen z regresijskim modelom. Preostali del variabilnosti odvisne spremenljivke z regresijskim modelom ni pojasnjen.

Koeficient determinacije lahko izračunamo tudi direktno iz vrednosti VKO:

$$r^2\% = \frac{VKO_{xy}^2}{VKO_{xx} \cdot VKO_{yy}} \cdot 100$$

Primer (3. del)

Izračunajmo koeficient determinacije za primer z lončki.

$$VKO_{yy} = 59874 - \frac{720^2}{9} = 2274$$

$$VKO_{\text{mod}} = \frac{1040^2}{734} = 1473{,}57$$

$$VKO_{ost} = 2274 - 1473{,}57 = 800{,}43$$

Koeficient determinacije

$$r^2\% = \frac{1473{,}57}{2274} \cdot 100 = 64{,}8\%$$

Približno 65% variabilnosti vsebnosti fosforja v rastlinah pojasni vsebnost fosforja v zemlji, 35% variabilnosti vsebnosti fosforja v rastlinah pa ostane s tem modelom nepojasnjene.

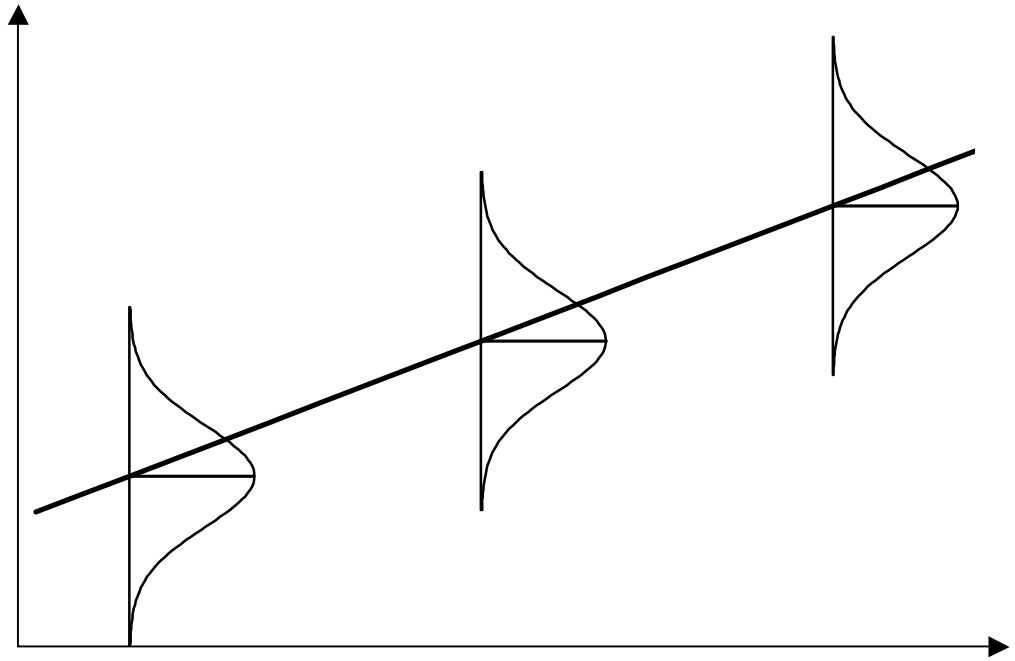
6.2.3 *Statistično sklepanje pri enostavni linearni regresiji

Pri enostavni linearni regresiji lahko izvedemo statistično sklepanje iz vzorca na populacijo, če privzamemo naslednje matematične predpostavke:

- X ni slučajna spremenljivka: njene vrednosti so izbrane (določene).
- Y je slučajna spremenljivka. Pri vsaki vrednosti x iz definicijskega območja je slučajna spremenljivka $Y \mid x$ porazdeljena normalno:
  $Y \mid x \sim N(\alpha + \beta \cdot x, \sigma)$

  Njena povprečna vrednost je na regresijski premici $\alpha + \beta \cdot x$, standardni odklon je σ, ki je za vse vrednosti x enak.
- Vrednosti za Y so pri različnih x med seboj neodvisne.

Grafično prikažemo to situacijo takole:



Slika 6-5: Prikaz porazdelitve odvisne spremenljivke pri posameznih vrednostih neodvisne spremenljivke

Zgornje predpostavke lahko povemo drugače: porazdelitev slučajnih vplivov ε je normalna, s povprečjem 0 in standardnim odklonom σ, torej $\varepsilon \sim N(0,\sigma)$, in so vrednosti za ε med seboj neodvisne.

6.2.3.1 *Standardna napaka regresije

Standardni odklon normalne porazdelitve spremenljivke $Y \mid x$, torej količino σ, imenujemo **standardna napaka regresije**. Oceno za standardno napako regresije dobimo na osnovi VKO_{ost}, ki meri razpršenost točk okoli premice. Teorija pokaže, da nepristransko oceno variance regresije izračunamo takole:

$$s^2 = \frac{1}{n-2} \cdot VKO_{ost} = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Opomba: iz opisne statistike poznamo izraz za izračun nepristranske ocene variance, kjer VKO delimo z $n-1$. V primeru linearnega regresijskega modela je v imenovalcu količina $n-2$, saj sta za regresijsko premico potrebni dve točki.

Standardna napaka regresije pa je kvadratni koren iz variance regresije:

$$s = \sqrt{\frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Glede na lastnost normalne porazdelitve lahko trdimo: v pasu $\hat{Y} \pm 2 \cdot s$ pričakujemo približno 95% točk populacije.

6.2.3.2 *Domneve in intervali zaupanja za parametre modela

Če veljajo opisane predpostavke, lahko izpeljemo porazdelitev vzorčnih ocen za parameter α in za parameter β ter za napovedi. Teorija pokaže, da so vse te

porazdelitve normalne. Posledica tega je preverjanje statističnih domnev, ki temelji na *t*-statistiki:

$$t = \frac{\text{ocena parametra} - \text{predpostavljena vrednost parametra}}{\text{standardna napaka ocene}}$$

Ničelna porazdelitev je Studentova porazdelitev, stopinje prostosti so $SP = n - 2$.

Izračun intervala zaupanja sledi naslednjemu pravilu:

$l_{1,2} = \text{ocena parametra} \mp t_{tab} \cdot \text{standardna napaka ocene}$

Brez dokazov navajamo porazdelitve vzorčnih ocen regresijskega modela. V vzorcih velikosti *n* je porazdelitev vzorčnih ocen za parameter α, torej porazdelitev ocen *a*, normalna:

$$N\left(\alpha, \sqrt{\sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{VKO_{xx}}\right)}\right)$$

Standardna napaka za *a* je:

$$s(a) = \sqrt{s^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{VKO_{xx}}\right)}$$

V vzorcih velikosti *n* je porazdelitev vzorčnih ocen za parameter β, torej porazdelitev vrednosti *b*, normalna:

$$N\left(\beta, \sqrt{\frac{\sigma^2}{VKO_{xx}}}\right)$$

Standardna napaka za *b* je:

$$s(b) = \sqrt{\frac{s^2}{VKO_{xx}}}$$

Za obe standardni napaki velja, da sta tem manjši, čim večja je variabilnost neodvisne spremenljivke.

Primer (4. del)

Predpostavimo, da je porazdelitev vsebnosti fosforja v rastlinah pri vsaki vrednosti vsebnosti fosforja v zemlji normalna, da je varianca pri vseh normalnih porazdelitvah enaka in da so izmerjene količine za fosfor v rastlini med seboj neodvisne.

Izračunajmo standardno napako regresije in obrazložimo njen pomen.

$$s = \sqrt{\frac{800{,}43}{7}} = \sqrt{114{,}347} = 10{,}69$$

V pasu 'premica $\mp 21{,}4$' pričakujemo približno 95% točk populacije.

Ali se vsebnost fosforja v rastlini spreminja z vsebnostjo fosforja v zemlji (stopnja značilnosti 0,05)? Za odgovor na to vprašanje moramo na naših podatkih preveriti ničelno domnevo, da odvisnosti ni. Povedano drugače, regresijska premica je vzporedna abscisni osi, torej $\beta = 0$.

$H_o : \beta = 0$ Vsebnost fosforja v rastlini ni odvisna od vsebnosti fosforja v zemlji.

$H_1 : \beta \neq 0$ Vsebnost fosforja v rastlini je odvisna od vsebnosti fosforja v zemlji.

Izračuni:

$$s(b) = \sqrt{\frac{114{,}35}{734}} = 0{,}395$$

$$t = \frac{1{,}417 - 0}{0{,}395} = 3{,}590 < t_{0{,}025}(7) = 2{,}365$$

Ničelno domnevo zavrnemo v korist alternativne domneve. Pri stopnji značilnosti 0,05 trdimo, da je vsebnost fosforja v rastlini odvisna od vsebnosti fosforja v zemlji (p=0,0089).

Izračunajmo še 95 % interval zaupanja za naklon premice β:

$$l_{1,2} = 1{,}417 \mp 2{,}365 \cdot 0{,}395$$

$$l_1 = 0{,}484 \text{ ppm} \quad l_2 = 2{,}350 \text{ ppm}$$

Dopolnimo znanje od prej:

Pri 95% zaupanju pričakujemo, da bo interval od 0,48 ppm do 2,35 ppm pokril vrednost β, torej povečanje vsebnosti fosforja v rastlini, če se vsebnost fosforja v zemlji poveča za 1 ppm.

Komentar: širina intervala zaupanja je prevelika, da bi lahko imeli ta rezultat za posebno uporaben. Seveda je to posledica dejstva, da je podatkov za analizo malo.

6.2.3.3 *Intervali zaupanja za napovedi

Ločimo dve vrsti napovedi odvisne spremenljivke pri izbrani vrednosti neodvisne spremenljivke x_0: povprečno napoved in posamično napoved.

Povprečna napoved pri x_0 je $\alpha + \beta \cdot x_0$.

Posamična napoved pri x_0 je $\alpha + \beta \cdot x_0 + \varepsilon_0$.

Teorija pokaže, da sta njuni porazdelitvi normalni, izraza za varianco sta dokaj kompleksna:

$$\alpha + \beta \cdot x_0 \sim N\left(\alpha + \beta \cdot x_0, \sqrt{\sigma^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{VKO_{xx}}\right)}\right)$$

$$\alpha + \beta \cdot x_0 + \varepsilon \sim N\left(\alpha + \beta \cdot x_0, \sqrt{\sigma^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{VKO_{xx}} + 1\right)}\right)$$

Pri izbrani točki x_0 je ocena za povprečno napoved na regresijski premici: $a + bx_0$, tudi ocena za posamično napoved je na regresijski premici: $a + bx_0$. Njuni standardni napaki pa sta različni.

Standardna napaka povprečne napovedi pri x_0 je:

$$s(a + bx_0) = \sqrt{s^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{VKO_{xx}}\right)}$$

Standardna napaka posamične napovedi pri x_0 pa je večja za en člen:

$$s(a+bx_0+e_0)=\sqrt{s^2\cdot\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{VKO_{xx}}+1\right)}$$

Interval zaupanja za napoved je:

$l_{1,2}=\hat{y}_0 \mp t_{tab}\cdot$ standardna napaka napovedi

Za obe vrsti napovedi velja:

- interval zaupanja za napoved je najožji, ko je $x_0=\bar{x}$;
- ko se oddaljujemo od $\bar{x}$ v levo in desno, se širina intervala zaupanja veča;
- meji intervala zaupanja za povprečno napoved in za posamično napoved ležita na hiperbolah: notranje hiperbole so za povprečno napoved, zunanje hiperbole pa za posamično napoved (glej sliko).

Primer (5. del)

Naj bo vsebnost fosforja v zemlji 10 ppm. Povprečna napoved se nanaša na povprečno vrednost fosforja v rastlini za lončke, ki so imeli 10 ppm fosforja v zemlji. Posamična napoved se nanaša na en določen lonček in pove, kolikšna je vrednost fosforja v rastlini za en lonček, ki je imel v zemlji 10 ppm fosforja.

Ocena za povprečno napoved in za posamično napoved je ista:

$61{,}58+1{,}417\cdot 10=75{,}75$

Standardna napaka za povprečno napoved pri tej vsebnosti fosforja v zemlji:

$$\sqrt{114{,}347\cdot\left(\frac{1}{9}+\frac{(10-13)^2}{734}\right)}=3{,}756$$

Interval zaupanja za povprečno napoved:

$l_{1,2}=75{,}75\mp 2{,}365\cdot 3{,}756$

$l_1=66{,}87$ ppm $\quad l_2=84{,}63$ ppm

Pri 95% zaupanju pričakujemo, da bo interval od 66,9 ppm do 84,6 ppm pokril vrednost za povprečno vsebnost fosforja v rastlinah za lončke, ki imajo 10 ppm fosforja v zemlji.

Standardna napaka za posamično napoved pri tej vsebnosti fosforja v zemlji:

$$\sqrt{114{,}347\cdot\left(\frac{1}{9}+\frac{3^2}{734}+1\right)}=11{,}334$$

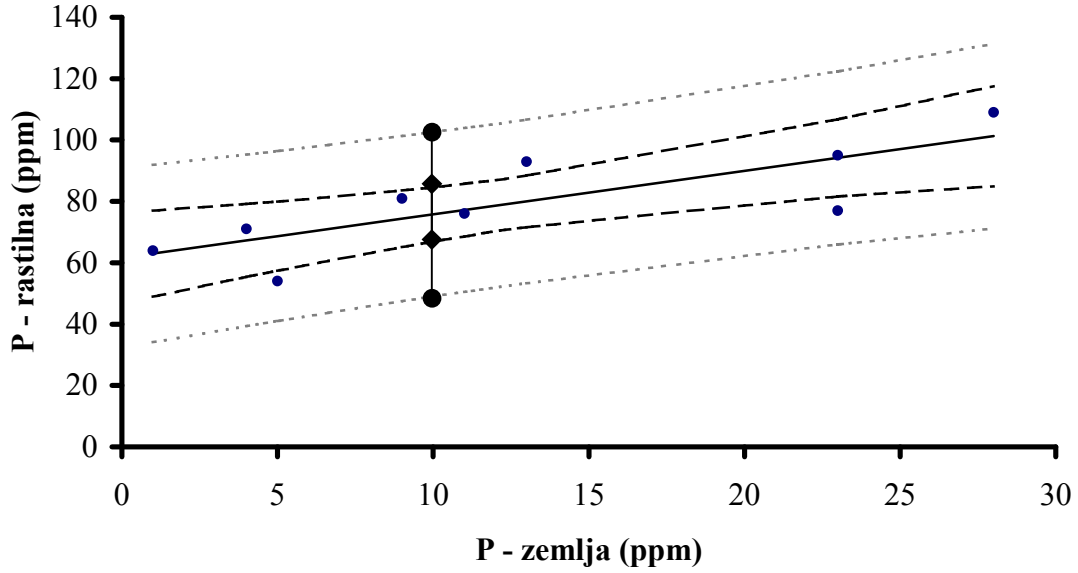
Interval zaupanja za posamično napoved:

$l_{1,2}=75{,}75\mp 2{,}365\cdot 11{,}334$

$l_1=48{,}95$ ppm $\quad l_2=102{,}55$ ppm

Pri 95% zaupanju pričakujemo, da bo interval od 49,0 ppm do 102,6 ppm pokril vrednost za vsebnost fosforja v rastlini za posamezni lonček, ki ima 10 ppm fosforja v zemlji.

Na sliki so predstavljene izračunane meje intervalov zaupanja za povprečno in za posamično napoved.



Slika 6-6: Meje intervalov zaupanja za povprečno napoved in za posamično napoved

Analizo primera z lončki lahko sklenemo z opažanjem, da je sklepanje o odvisnosti na osnovi 9 lončkov sicer možno, rezultati pa kažejo, da kakovost sklepov ni zadovoljiva. To se še najbolj vidi po širokih intervalih zaupanja za parametre modela in za napovedi.

NALOGE

1. Vzdrževanje osebnih vozil

Proučevali so letne stroške vzdrževanja za osebno vozilo v odvisnosti od starosti vozila. Podatki za 9 osebnih vozil so v tabeli:

Tabela 6-2: Stroški vzdrževanja osebnega vozila v odvisnosti od starosti vozila

Vozilo	Starost vozila(leta)	Stroški(DE)
1	8	859
2	5	682
3	3	471
4	9	708
5	11	1094
6	2	224
7	1	320
8	8	651
9	12	1049

a) Grafično prikažite podatke.

b) Kako starost vpliva na stroške vzdrževanja? Ali je linearni regresijski model primeren za te podatke?

c) Izračunajte ocene linearnega regresijskega modela in obrazložite dobljene vrednosti.

d) Izračunajte koeficient determinacije in ga obrazložite.

e) Izračunajte napovedane stroške za osebno vozilo, ki je staro 5 let.

f) Ali je smiselno izračunati napovedane stroške za 25 let staro vozilo?

2. Smrtno ponesrečeni

V tabeli je navedeno število smrtno ponesrečenih v prometnih nesrečah v obdobju 1988 - 1998 v Sloveniji.

Tabela 6-3: Število smrtno ponesrečenih v prometnih nesrečah v Sloveniji po letih v obdobju 1988 - 1998 (Vir: SL - 93, str. 274, SL - 99, str. 364)

Leto	Število smrtno ponesrečenih
1988	551
1989	554
1990	517
1991	462
1992	492
1993	493
1994	505
1995	415
1996	389
1997	358
1998	309

a) Grafično prikažite podatke.

b) Kako se število smrtno ponesrečenih spreminja s časom? Izračunajte ocene linearnega regresijskega modela. Obrazložite dobljene vrednosti.
Opomba: izračune si poenostavite tako, da časovno skalo transformirate: namesto odvisnosti od spremenljivke Leto izračunajte odvisnost od spremenljivke T, ki je opredeljena takole: T=Leto-1993.

c) Izračunajte napoved za leto 1999.

d) V letu 1999 je bilo 334 smrtno ponesrečenih oseb (SL-2000). Izračunajte absolutno in relativno napako napovedi.

3. Črvivost jabolk

Sadjarji so ugotavljali odvisnost % črvivih plodov od števila plodov. Opazovali so 12 jablan in dobili naslednje podatke:

Tabela 6-4: Število plodov na drevo (100) in % črvivih plodov na drevo (Vir: Snedecor, Cochran: Statistical Methods, str. 150)

Drevo	Število plodov (v 100)	% črvivih plodov
1	8	59
2	6	58
3	11	56
4	22	53
5	14	50
6	17	45
7	18	43
8	24	42
9	19	39
10	23	38
11	26	30
12	40	27

a) Grafično prikažite podatke.

b) Izračunajte ocene parametrov ustreznega modela in jih obrazložite.

c) *Izračunajte 95% intervale zaupanja za parametre modela in jih obrazložite.

d) Izračunajte koeficient determinacije in ga obrazložite.

e) Kolikšen % črvivih plodov lahko pričakujemo za drevo, ki bi imelo 200 plodov, kolikšen za drevo, ki bi imelo 2000 plodov?

4. Starost in krvni tlak

Za vzorec desetih pacientov imamo podatke o njihovi starosti in o njihovem sistoličnem krvnem tlaku. Podatki so:

Tabela 6-5: Podatki o starosti in krvnem tlaku

Pacient	Starost (leta)	Sistolični tlak (mm)
1	24	134
2	28	117
3	35	132
4	40	129
5	48	162
6	50	150
7	52	171
8	54	147
9	63	146
10	72	154

a) Analizirajte odvisnost krvnega tlaka od starosti. Obrazložite dobljene vrednosti.

b) *Preverite domnevo, da je naklon premice 45^0 (stopnja značilnosti 0,05). Kaj to pomeni?

c) Izračunajte napoved za osebo, staro 50 let.

d) *Izračunajte še 95% interval zaupanja za to napoved.

5. Hitrost vetra in čas teka čez ovire

Britanska atletska zveza je proučevala rezultate atleta Colina Jacksona v teku na 110 m čez ovire. Njen namen je bil ugotoviti, kako hitrost vetra vpliva na čas teka. Negativne vrednosti hitrosti pomenijo veter v prsi, pozitivne pa veter v hrbet. Podatki za 21 tekov so v tabeli.

Tabela 6-6: Hitrost vetra in čas teka na 110 m čez ovire za atleta Colina Jacksona (Vir: Daly et al.:Elements of Statistics, str. 525)

Tek	Hitrost vetra (m/s)	Čas teka (s)
1	-2,9	13,53
2	-2,0	13,63
3	-1,6	13,39
4	-1,4	13,53
5	-0,8	13,63
6	-0,4	13,17
7	-0,4	13,25
8	-0,2	13,23
9	-0,1	13,63
10	0,2	13,09
11	0,2	13,18
12	0,5	13,08
13	0,5	13,38
14	0,8	13,52
15	1,0	13,11
16	1,1	13,10
17	1,1	13,20
18	1,2	13,23
19	2,2	13,22
20	2,8	13,14
21	2,9	13,12

a) Grafično prikažite podatke.

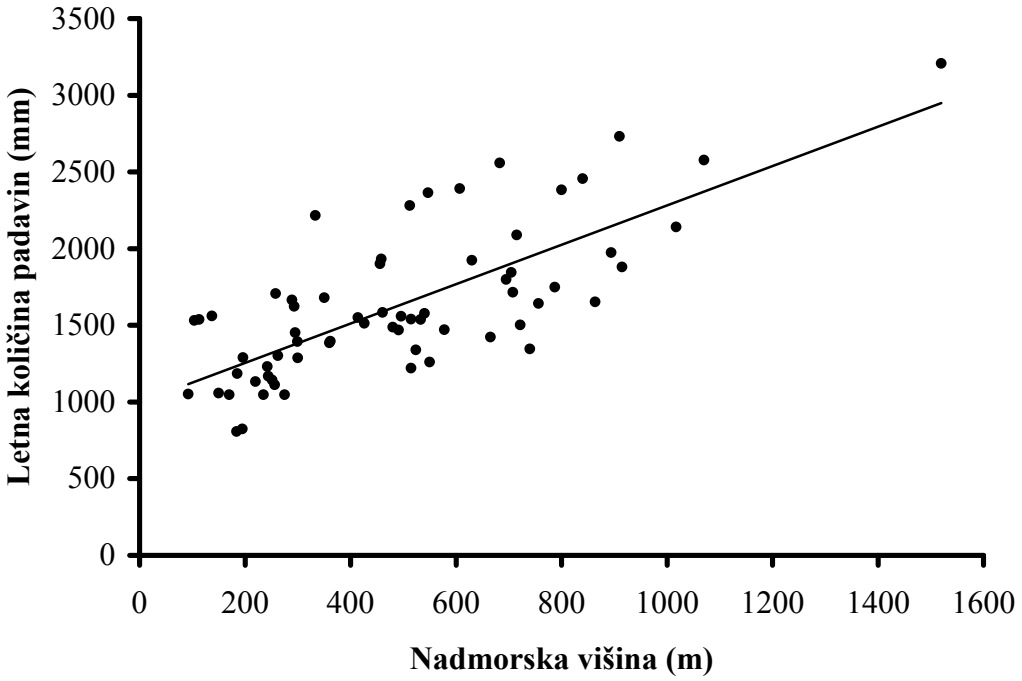
b) Izračunajte ocene parametrov linearnega regresijskega modela. Pri računanju si lahko pomagate z naslednjimi delnimi izračuni:
$\sum x = 4{,}7 \ \sum y = 279{,}36 \ \sum x^2 = 45{,}11 \ \sum y^2 = 3717{,}068 \ \sum xy = 58{,}796$

c) *Na osnovi vzorca preverite domnevo, da veter ne vpliva na Jacksonov čas teka (stopnja značilnosti 0,05).

d) *Izračunajte povprečno napoved časa teka v brezvetrju in njen 95% interval zaupanja.

6. Nadmorska višina in količina padavin

S programom MS Excel smo analizirali odvisnost letne količine padavin od nadmorske višine za vzorec 66 meteoroloških postaj v Sloveniji. Postajo Kredarica smo iz analize izločili. Grafični prikaz podatkov in linearnega regresijskega modela je na sliki.



Slika 6-7: Odvisnost količine padavin od nadmorske višine

Rezultati so:

r^2	0,553
s	325,142
n	66

	Ocena parametra	**Standardna napaka**	***t*-stat**	***p*-vrednost**	l_1	l_2
a	999,246	81,233	12,301	1,624E-18	836,963	1161,528
b	1,282	0,144	8,902	8,403E-13	0,994	1,570

a) Kako nadmorska višina vpliva na letno količino padavin?
b) Obrazložite ocene linearnega regresijskega modela.
c) *Obrazložite pripadajoče intervale zaupanja.
d) Kolikšen del variabilnosti letne količine padavin pojasni nadmorska višina? Kaj pa preostanek?
e) Kakšna je povprečna napoved za letno količino padavin na morju?
f) *Kakšen je pripadajoči 95% interval zaupanja?
g) *Izračunajte povprečno napoved pri nadmorski višini 1000 m in pripadajoči 95% interval zaupanja.

6.3 KORELACIJA

X in Y sta povezani spremenljivki, obe spremenljivki sta številski. Pogledali bomo dve meri povezanosti:

- **Pearsonov koeficient korelacije,**
- **Spearmanov koeficient korelacije** znan tudi pod imenom **koeficient korelacije rangov**.

6.3.1 Pearsonov koeficient korelacije

6.3.1.1 Dvorazsežna normalna porazdelitev

Oglejmo si najprej **dvorazsežno normalno porazdelitev**. Naj za slučajni vektor (X, Y) velja, da je porazdeljen po dvorazsežni normalni porazdelitvi. To zapišemo takole:

$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$$

Gostota verjetnosti za dvorazsežno normalno porazdelitev je zapisana v obliki:

$$p(x) = \frac{1}{2\,\pi\,\sigma_x\,\sigma_y\,\sqrt{1-\rho^2}} \cdot$$

$$\exp\left[-\frac{1}{2\,(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\,\rho\,(x-\mu_x)(y-\mu_y)}{\sigma_x\,\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right],$$

$$-\infty < x < \infty, -\infty < y < \infty$$

Dvorazsežna normalna porazdelitev ima pet parametrov: $\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$. Prvi štirje so nam znani. ρ je peti parameter dvorazsežne normalne porazdelitve, imenujemo ga **koeficient korelacije**.

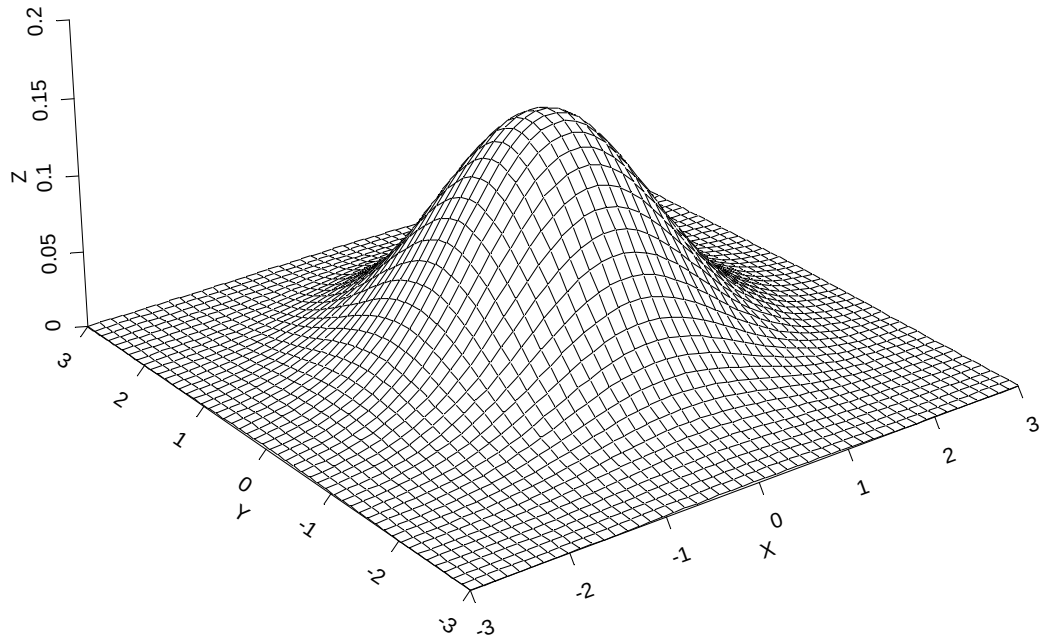
Teorija pokaže naslednje lastnosti koeficienta korelacije:

a) $-1 \le \rho \le 1$ Vrednosti koeficienta korelacije so na intervalu od –1 do +1.

b) $\rho_{xy} = \rho_{yx}$ Koeficient korelacije je simetričen; spremenljivki sta enakovredni.

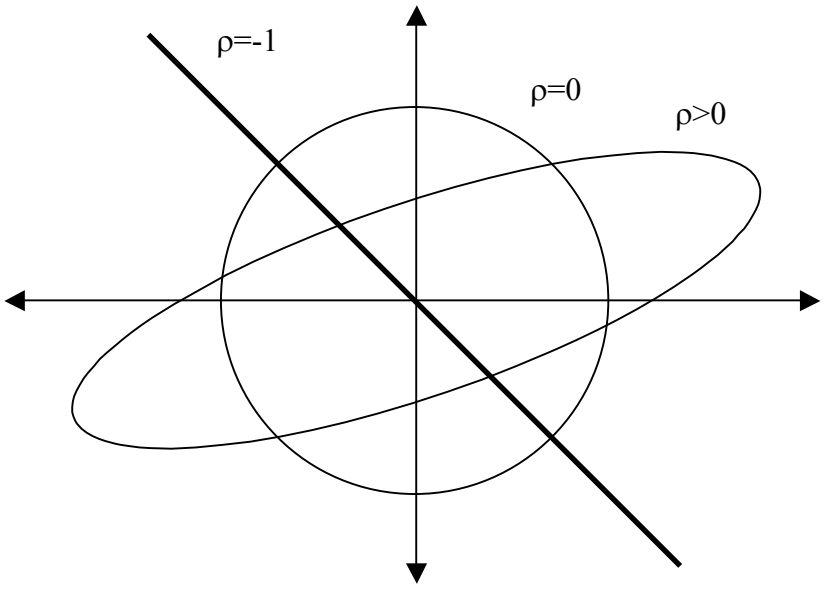
c) $\rho_{xx} = 1$ Korelacija spremenljivke same s seboj je 1.

Gostota verjetnosti ima obliko ‘hriba’. Na sliki je prikazana standardizirana dvorazsežna normalna porazdelitev N(0, 0, 1, 1, 0), pri kateri je vrednost koeficienta korelacije 0.



Slika 6-8: Gostota verjetnosti za dvorazsežno normalno porazdelitev N(0,0,1,1,0)

Koeficient korelacije vpliva na obliko porazdelitve. Če je vrednost koeficienta korelacije 0, so prerezi, ki so vzporedni z ravnino (x,y), krogi, sicer so elipse, orientirane v določeno stran. Če je koeficient korelacije +1 ali –1, porazdelitev degenerira v enorazsežno normalno porazdelitev.



Slika 6-9: Prerezi pri dvorazsežni normalni porazdelitvi glede na vrednost korelacijskega koeficienta

6.3.1.2 Izračun Pearsonovega koeficienta korelacije

Poglejmo, kje je ta verjetnostna porazdelitev uporabna v statistiki. V vzorcu je n enot. Na vsaki enoti imamo podatek za X in podatek za Y. Na i-ti enoti ju označimo

(x_i, y_i). Najprej podatke grafično predstavimo z razsevnim grafikonom. Za razliko od grafičnega prikaza pri regresiji je v tem primeru vseeno, katera spremenljivka je na abscisi, katera na ordinati, saj sta spremenljivki enakovredni.

Če sta spremenljivki povezani in je povezava linearna, zanju lahko privzamemo dvorazsežno normalno porazdelitev. Potem je smiselno izračunati oceno koeficienta korelacije. To oceno označimo r. Izpeljal jo je Pearson, zato jo imenujemo **Pearsonov koeficient korelacije**. Ocena koeficienta korelacije po Pearsonu je:

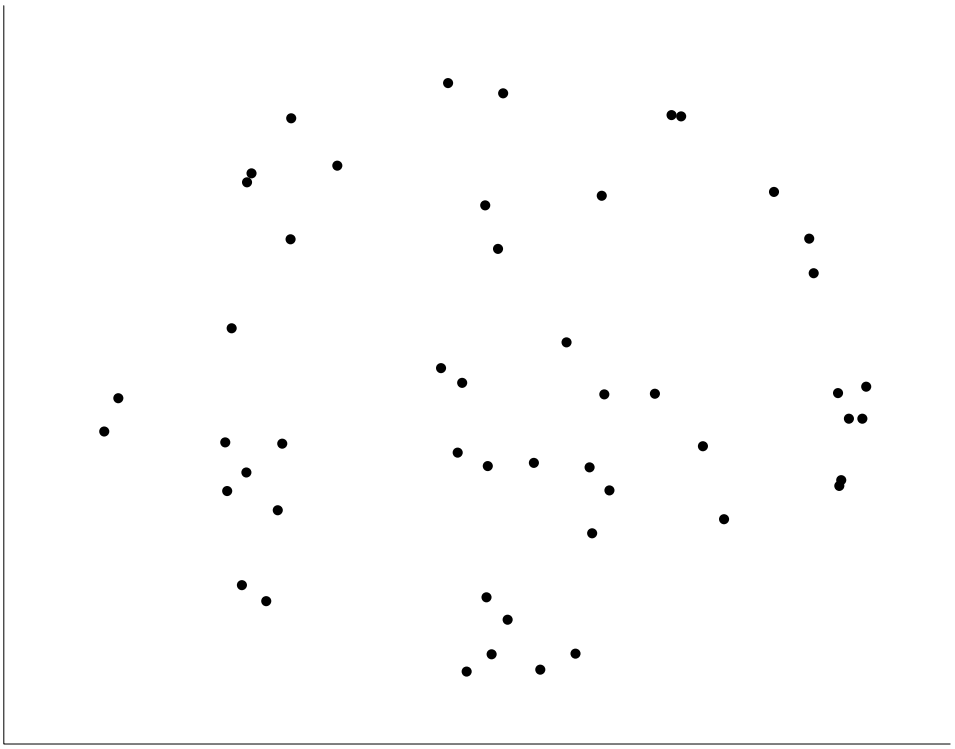
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{VKO_{xy}}{\sqrt{VKO_{xx} \cdot VKO_{yy}}}$$

V izračunu so količine, ki jih že poznamo. VKO_{xy} v števcu določa predznak korelacijskega koeficienta. Povejmo še enkrat: vrednost korelacijskega koeficienta je na intervalu [-1, +1].

Za ilustracijo predstavljamo nekaj razsevnih grafikonov za podatke z različnimi vrednostmi Pearsonovega koeficienta korelacije.

Slika 6-10: Pozitivna korelacija r=0,76

Slika 6-11: Močna negativna korelacija, r=-0,97



Slika 6-12: Korelacija nič, r=0,02

Ponavadi nas zanima, ali sta spremenljivki v populaciji sploh povezani. Sklepamo na osnovi vrednosti koeficienta korelacije v vzorcu. Ničelna domneva pravi, da povezave ni, torej da je $\rho = 0$, alternativna domneva pa je negacija tega.

$H_o : \rho = 0$ Spremenljivki nista linearno povezani.

$H_1 : \rho \neq 0$ Spremenljivki sta linearno povezani.

Teorija pokaže, da je testna statistika za tako ničelno domnevo naslednja:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Njena ničelna porazdelitev je $t(SP = n-2)$.

Izračunavanju vrednosti testne statistike se lahko izognemo, če imamo na razpolago ustrezne statistične tabele (glej Tabelo 9 v prilogi). V teh tabelah je podana kritična vrednost Pearsonovega koeficienta korelacije pri določeni stopnji značilnosti (0,05 ali 0,01) in določeni velikosti vzorca. Če je izračunana ocena *r* po absolutni vrednosti večja od kritične vrednosti, ničelno domnevo zavrnemo v korist alternativne domneve.

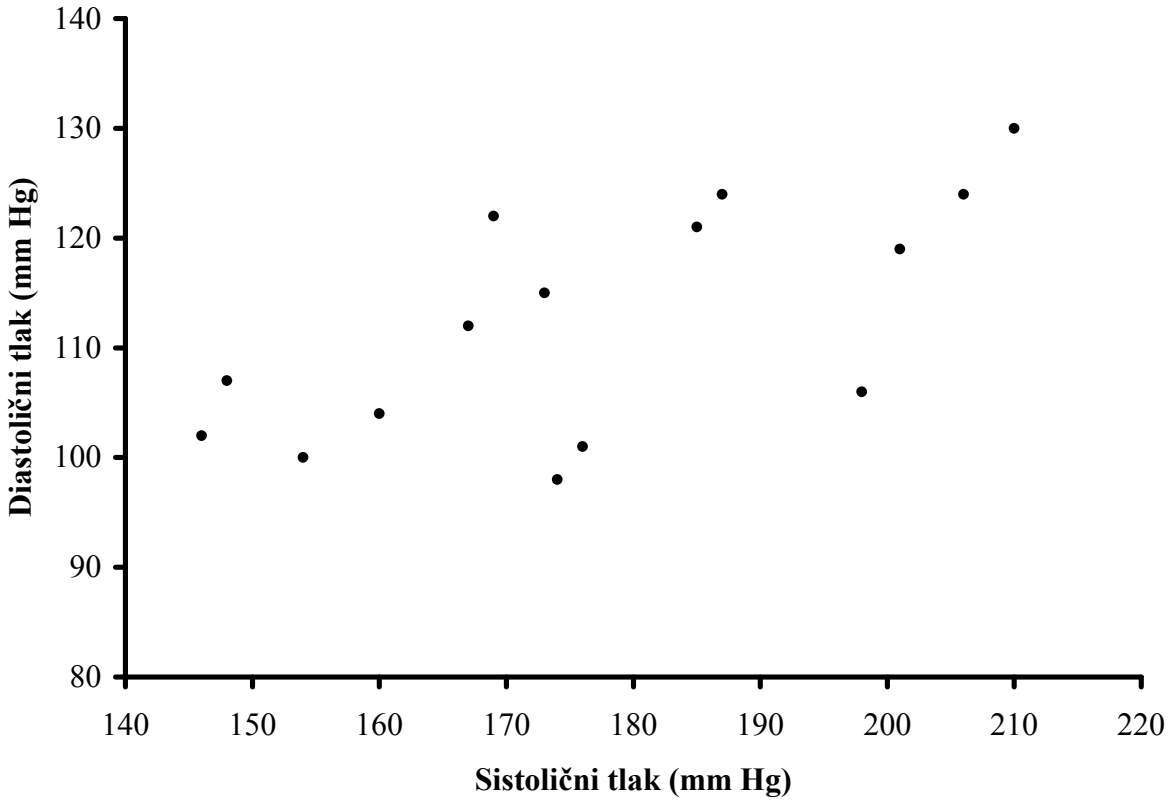
Primer

Zdravniki merijo dve vrednosti krvnega tlaka, sistolični krvni tlak in diastolični krvni tlak. Zanima nas, ali sta ti dve spremenljivki povezani. V vzorcu je 15 oseb, podatki za njih so v tabeli.

Tabela 6-7: Sistolični in diastolični krvni tlak za 15 oseb (Vir: Daly et al., str 426)

Oseba	Sistolični tlak (mm Hg)	Diastolični tlak (mm Hg)
1	210	130
2	169	122
3	187	124
4	160	104
5	167	112
6	176	101
7	185	121
8	206	124
9	173	115
10	146	102
11	174	98
12	201	119
13	198	106
14	148	107
15	154	100

Podatke najprej grafično prikažemo.



Slika 6-13: Sistolični in diastolični krvni tlak za 15 oseb

Slika nakazuje linearno soodvisnost diastoličnega in sistoličnega tlaka. Izračunajmo ustrezno mero povezanosti.

Pomožni računi:

$n = 15$

$\sum x = 2654 \quad \sum x^2 = 475502 \quad VKO_{xx} = 88814$

$\sum y = 1685 \quad \sum y^2 = 190817 \quad VKO_{yy} = 23030$

$\sum xy = 300137 \quad VKO_{xy} = 30065$

$$r = \frac{30065}{\sqrt{88814 \cdot 23030}} = 0{,}6647$$

Pri stopnji značilnosti 0,05 preverimo domnevo o povezanosti sistoličnega in diastoličnega krvnega tlaka.

$H_0: \ \rho = 0$ Povezave ni.

$H_1: \ \rho \neq 0$ Povezava je.

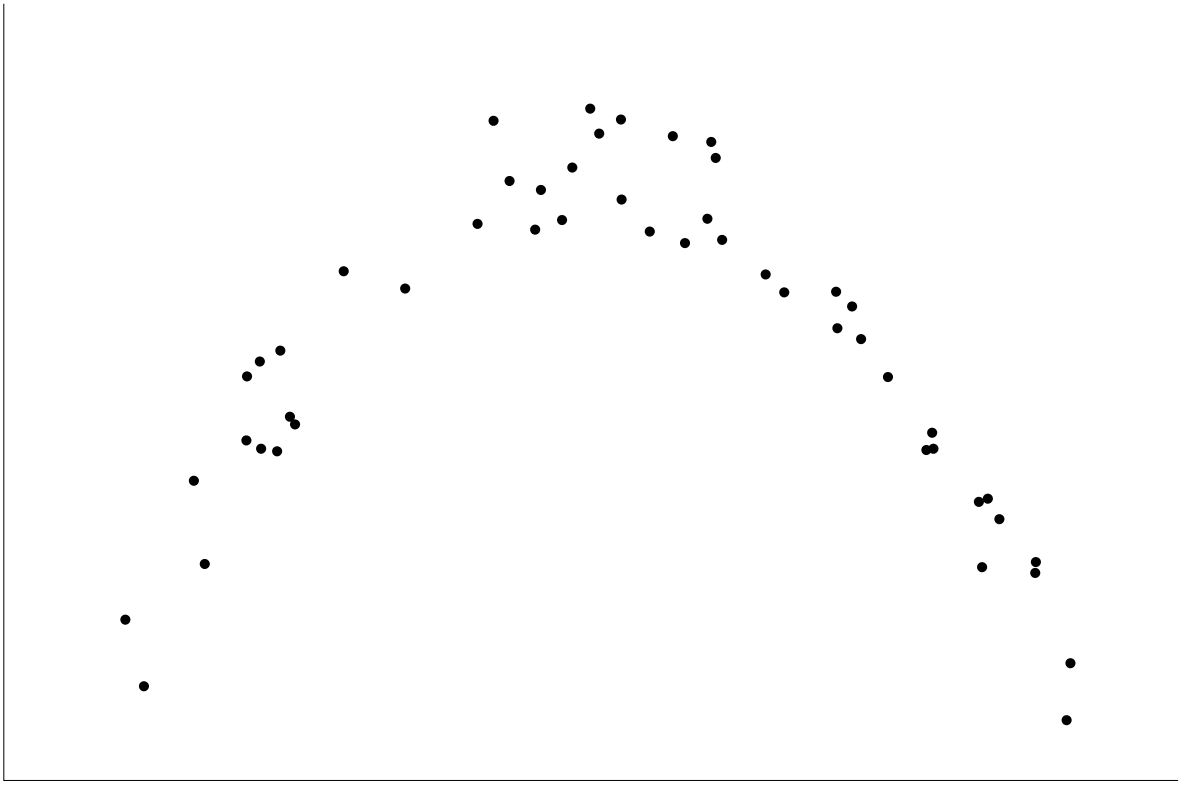
$r = 0{,}6647 > r_{tab} = 0{,}5140$

Vrednost r_{tab} smo odčitali iz tabel pri velikosti vzorca $n = 15$ in stopnji značilnosti 0,05.

Statistični sklep: ničelno domnevo zavrnemo v korist alternativne domneve.

Pri stopnji značilnosti 0,05 trdimo, da obstaja pozitivna linearna povezava med sistoličnim in diastoličnim krvnim tlakom ($p<0,01$).

Pri uporabi korelacijskega koeficienta moramo biti zelo previdni. Če povezava med spremenljivkama ni linearna, uporaba korelacijskega koeficienta kot mere povezanosti dveh številskih spremenljivk ni ustrezna. Naslednja slika prikazuje kvadratno povezavo. Dvorazsežna normalna porazdelitev ni ustrezen verjetnostni model za ta slučajni vektor. Za ta primer je izračunana vrednost Pearsonovega korelacijskega koeficienta okoli nič. Obrazložitev, da povezave med spremenljivkami ni, bi bila napačna. Povezava obstaja, vendar ni linearna.



Slika 6-14: Grafičen prikaz nelinearne povezave. Koeficient korelacije ni ustrezna mera povezanosti.

Na koncu namenimo še nekaj besed primerjavi enostavnega linearnega regresijskega modela in korelacijskega modela. Predpostavke pri enostavnem linearnem regresijskem modelu in pri korelacijskem modelu so različne. Pri regresiji je Y slučajna spremenljivka, ki ima za vsako vrednost x določeno verjetnostno porazdelitev. Pri korelaciji pa je par (X,Y) slučajni vektor z ustrezno dvorazsežno verjetnostno porazdelitvijo. Obstajajo pa določene matematične zveze med regresijo in korelacijo.

- Koeficient korelacije je ustrezno predznačen koren iz koeficienta determinacije: $r = \pm\sqrt{r^2}$
- Obstaja naslednja zveza med koeficientom korelacije in koeficientom regresije b :

$$r = \frac{s_y}{s_x} b,$$

pri čemer sta s_y in s_x standardna odklona spremenljivk *Y* in *X*. Predznak koeficienta korelacije in predznak koeficenta b se ujemata.

6.3.1.3 *Interval zaupanja za koeficient korelacije

Če želimo preveriti bolj splošno domnevo, da je $H_o : \rho = \rho_0$ in $\rho_0 \neq 0$, prej omenjeni preizkus ne velja. Da dobimo uporaben preizkus, moramo koeficient korelacije transformirati. Teorija pokaže, da je ustrezna transformacija naslednja:

$$w = w(r) = \frac{1}{2} \ln \frac{1+r}{1-r}$$

Ta transformacija prevede vrednosti koeficienta korelacije iz intervala $[-1, +1]$ na celo realno os, pripadajoča porazdelitev ocen w je približno normalna. Inverzna transformacija, ki w prevede v r, pa je:

$$r = r(w) = \frac{e^{2w} - 1}{e^{2w} + 1}$$

Teorija tudi pokaže, da je standardna napaka za w naslednja:

$$s(w) = \sqrt{\frac{1}{n-3}}$$

Za preverjanje domneve $H_o : \rho = \rho_0 \neq 0$ transformiramo tudi vrednost ρ_0 v $w(\rho_0)$, nato uporabimo statistiko z:

$$z = \frac{w - w(\rho_0)}{s(w)}$$

Ničelna porazdelitev je N(0,1).

Da bi izračunali interval zaupanja za koeficient korelacije ρ, izračunamo meji intervala zaupanja za $w(\rho)$:

$$l_{1,2}(w(\rho)) = w \mp z_{\frac{\alpha}{2}} \cdot s(w)$$

Zatem na tako izračunanih mejah uporabimo inverzno transformacijo in dobimo meji intervala zaupanja za koeficient korelacije ρ. Interval zaupanja za ρ ni simetričen okoli ocene r.

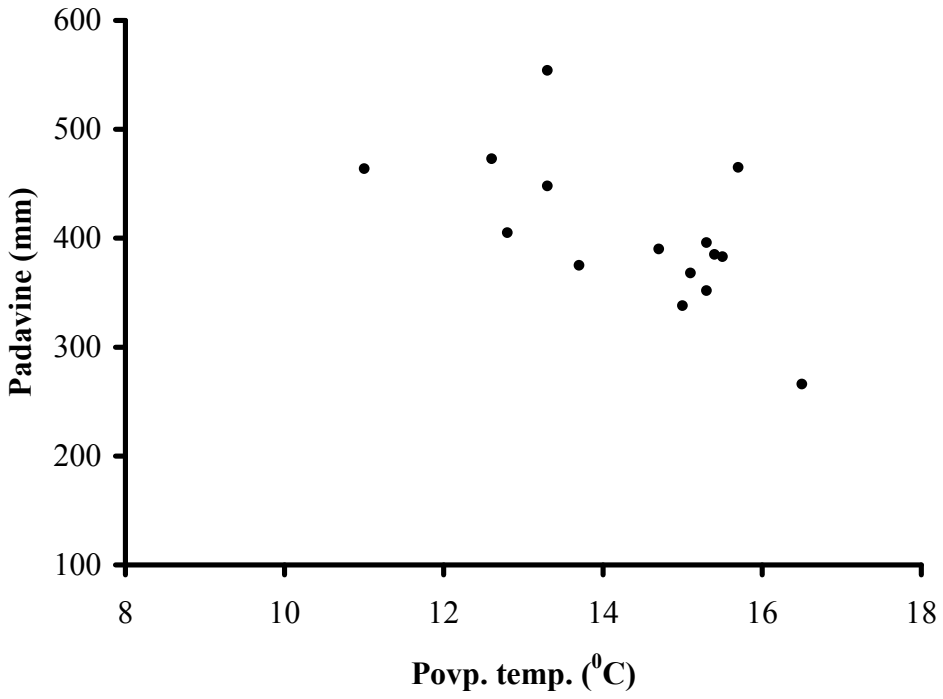
Primer

V vzorcu imamo 15 meteoroloških postaj. Za njih imamo podatke za dve spremenljivki: povprečna temperatura zraka in količina padavin v obdobju april-junij 1996. Podatki so v tabeli. Zanima nas 95% interval zaupanja za koeficient korelacije.

Tabela 6-8: Povprečna temperatura zraka in količina padavin v obdobju april-junij v letu 1996 za 15 meteoroloških postaj v Sloveniji (Vir: SL-97)

Postaja	Povp. temp. (^{0}C)	Padavine (mm)
Celje	15,0	338
Ljubljana	15,4	385
Maribor	15,5	383
Murska Sobota	15,3	352
Novo mesto	15,1	368
Portorož	16,5	266
Slap pri Vipavi	15,7	465
Brnik	13,7	375
Črnomelj	15,3	396
Kočevje	12,8	405
Lesce	13,3	554
Postojna	12,6	473
Rateče-Planica	11,0	464
Šmartno pri Sl. Gradcu	13,3	448
Velenje	14,7	390

Najprej grafično predstavimo podatke in si oglejmo sliko.



Slika 6-15: Povprečna temperatura zraka in količina padavin za 15 meteoroloških postaj v Sloveniji

Slika nakazuje, da je povezava linearna. Izračunajmo koeficient korelacije:

Izračuni:

$n = 15$

$$\sum x = 215{,}2 \quad \sum x^2 = 3118{,}5 \quad VKO_{xx} = 3118{,}5 - \frac{215{,}2^2}{15} = 31{,}097$$

$$\sum y = 6062 \quad \sum y^2 = 2515678 \quad VKO_{yy} = 2515678 - \frac{6062^2}{15} = 65821{,}733$$

$$\sum xy = 86071{,}1 \quad VKO_{xy} = 86071{,}1 - \frac{215{,}2 \cdot 6062}{15} = -898{,}393$$

$$r = \frac{-898{,}393}{\sqrt{31{,}097 \cdot 65821{,}733}} = -0{,}62794$$

$$w = 0{,}5 \ln \frac{1 + 0{,}62794}{1 - 0{,}62794} = -0{,}73801$$

$$s(w) = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{12}} = 0{,}288675$$

$$l_{1,2} = -0{,}73801 \mp 1{,}96 \cdot 0{,}288675 = -0{,}737941 \mp 0{,}565803$$

$l_1 = -1{,}3038$ $\qquad$ $l_2 = -0{,}1722$

Iskani 95% interval zaupanja za koeficient korelacije ρ po inverzni transformaciji je:

$l_1 = -0{,}8627$ $\qquad$ $l_2 = -0{,}1705$

Pri 95% zaupanju trdimo, da interval od –0,86 do –0,17 pokrije koeficient korelacije med povprečno količino padavin in temperaturo zraka.

6.3.2 *Spearmanov koeficient korelacije

Spoznali smo, da je Pearsonov koeficient korelacije mera linearne povezanosti za dve številski spremenljivki. Velikokrat pa predpostavka o linearnosti ni utemeljena. Pogledali bomo še eno mero povezanosti, ki meri intenzivnost povezanosti dveh spremenljivk in ne temelji na predpostavki linearnosti. Za to mero zadošča, da je povezava med spremenljivkama *monotona*. Ta koeficient je **koeficient korelacije rangov**. Njegovo oceno je izpeljal Spearman, zato jo imenujemo **Spearmanov koeficient korelacije**, označimo jo r_s.

Spearmanova ideja je nadomestiti podatke z njihovimi rangi in izračunati Pearsonov koeficient korelacije na rangih. Izračun koeficienta korelacije rangov gre po zelo preprostem postopku.

1. Za spremenljivko X enotam priredimo range. Če imata dve enoti enako vrednost, jima priredimo njun povprečni rang.
2. Isto za spremenljivko Y.
3. Na tako opredeljenih vrednostih, torej na rangih, izračunamo Pearsonov koeficient korelacije.

Pri statističnem sklepanju nas zanima, ali sta v populaciji spremenljivki monotono povezani. Ničelna porazdelitev pravi, da povezave ni, torej da je $\rho_s = 0$, alternativna domneva pa je negacija tega.

$H_o : \rho_s = 0$ Spremenljivki nista monotono povezani.

$H_1 : \rho_s \neq 0$ Spremenljivki sta monotono povezani.

Kritične vrednosti koeficienta korelacije rangov so v tabelah (glej Tabelo 10 v prilogi). Ta tabela je narejena po istem principu kot tabela za Pearsonov koeficient korelacije.

Spearmanov koeficient korelacije uporabljamo v naslednjih primerih:

- zveza med spremenljivkama ni linearna, ampak monotona,
- podatki ene ali obeh spremenljivk so rangi,
- podatki so nezanesljivi, za njihove predstavnike uporabimo pripadajoče range.

Alternativna mera povezanosti je *Kendallov* τ (tau), ki ga tu ne navajamo.

Opomba: če je povezava med spremenljivkama linearna, je razlika med Pearsonovim in Spearmanovim koeficientom korelacije minimalna. V tem primeru uporabljamo Pearsonov koeficient korelacije, ker ima pripadajoči statistični test večjo moč.

Primer

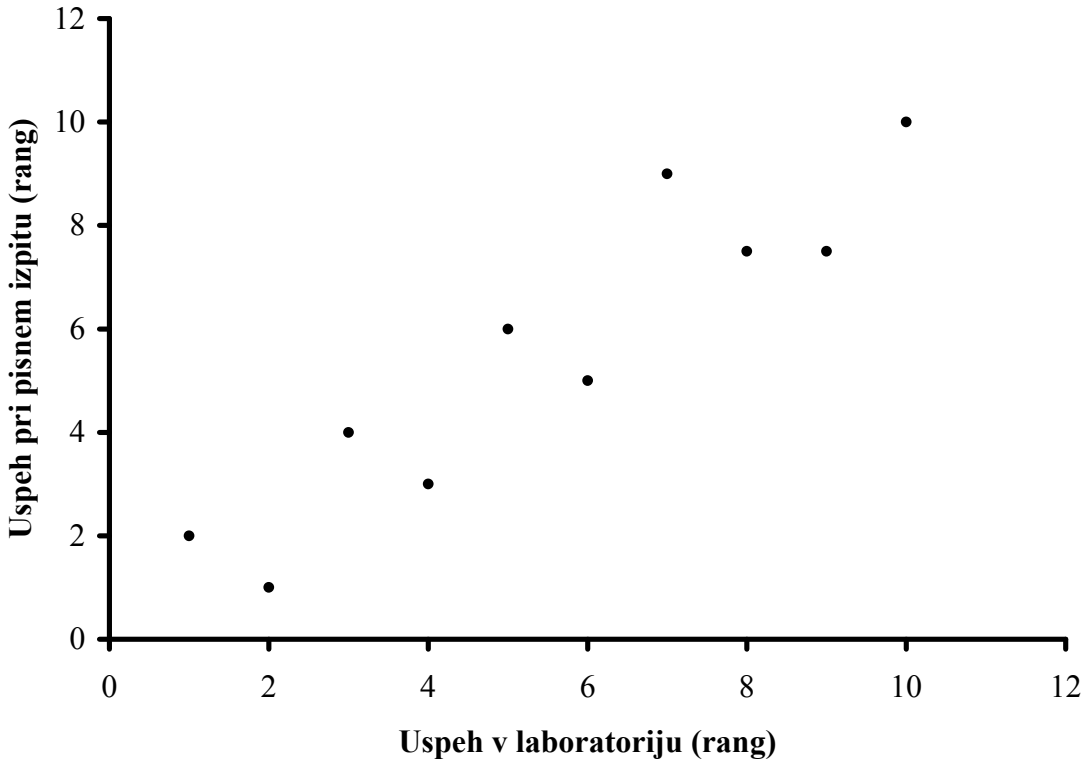
Zanima nas, ali obstaja povezava med uspehom pri laboratorijskem delu in uspehom pri pisnem izpitu. V vzorcu je 10 študentov. Pri laboratorijskem delu so bili študenti rangirani, rang 1 ima najboljši študent, rang 10 najslabši. Pri pisnem izpitu so bile ocene v točkah od 0 do 100. Podatki so v tabeli. Za pisni izpit imamo številske vrednosti, za uspeh v laboratoriju so rangi. Da bi lahko ovrednotili povezanost, moramo podatkom za število točk prirediti range.

Tabela 6-9: Uspeh študentov pri laboratorijskem delu in pri ustnem izpitu

Študent	Uspeh v laboratoriju (rang)	Uspeh pri pisnem izpitu (število točk)	Uspeh pri pisnem izpitu (rang)
1	3	72	4
2	8	35	7,5
3	2	100	1
4	9	35	7,5
5	4	83	3
6	1	89	2
7	7	32	9
8	5	44	6
9	10	25	10
10	6	45	5

Dva študenta sta zbrala 35 točk, priredili smo jima povprečni rang.

Tako prirejene podatke, torej range, grafično prikažemo z razsevnim grafikonom. Slika kaže močno pozitivno povezanost.



Slika 6-16: Rang uspeha v laboratoriju in rang uspeha pri pisnem izpitu za 10 študentov

Izračuni:

$n = 10$

$$\sum x = 55 \quad \sum x^2 = 385 \quad VKO_{xx} = 385 - \frac{55^2}{10} = 82{,}5$$

$$\sum y = 55 \quad \sum y^2 = 384{,}5 \quad VKO_{yy} = 384{,}5 - \frac{55^2}{10} = 82{,}0$$

$$\sum xy = 378{,}5 \quad VKO_{xy} = 378{,}5 - \frac{55 \cdot 55}{10} = 76{,}0$$

$$r_s = \frac{76{,}0}{\sqrt{82{,}5 \cdot 82{,}0}} = 0{,}9240$$

Kritična vrednost za koeficent korelacije rangov pri $\alpha = 0{,}05$ in velikosti vzorca 10 je 0,6485 (Tabela 10 priloge). Na osnovi tega ničelno domnevo zavrnemo v korist alternativne domneve.

Pri stopnji značilnosti 0,05 trdimo, da obstaja močna pozitivna povezava med uspehom v laboratoriju in uspehom pri pisnem izpitu ($p < 0{,}01$).

Naloge

1. Primerjava merskih metod

Koncentracijo nitrata so merili s klasično metodo A. To metodo so želeli nadomestiti z novo metodo B. Da bi ugotovili, ali dajeta metodi skladne rezultate, so nekaj časa uporabljali obe metodi. Rezultati so v tabeli.

Tabela 6-10: Količina nitrata v vodi (mikrogram/l vode), izmerjena na istih vzorcih vode z metodo A in z metodo B

Vzorec št.	A	B
1	25	30
2	40	80
3	120	150
4	75	80
5	150	200
6	300	350
7	270	240
8	400	320
9	450	470
10	575	583

a) Podatke grafično predstavite z razsevnim grafikonom.

b) Ali so rezultati, dobljeni po metodi A, povezani z rezultati, dobljenimi po metodi B (stopnja značilnosti 0,05)? Obrazložite rezultate.

2. Višina in teža otrok

Ali sta višina in teža otrok povezani? V vzorcu je 12 otrok starih 10 let. Podatki so:

Tabela 6-11: Otroci po višini in teži

Otrok	Višina (cm)	Teža (kg)
1	135	26
2	146	33
3	153	55
4	154	50
5	139	32
6	131	25
7	149	44
8	137	31
9	143	36
10	146	35
11	141	28
12	136	28

a) Podatke grafično predstavite z razsevnim grafikonom.

b) Preverite, ali sta višina in teža otrok povezani spremenljivki (stopnja značilnosti 0,05). Obrazložite rezultate.

3. Stroški oglaševanja

Imamo podatke o stroških oglaševanja in o prometu za vzorec 10 podjetij.

Tabela 6-12: Promet in stroški oglaševanja za deset mesecev

Podjetje	Promet (DE)	Stroški oglaševanja (DE)
1	101	1,20
2	92	0,80
3	110	1,00
4	120	1,30
5	90	0,70
6	82	0,80
7	93	1,00
8	75	0,60
9	91	0,90
10	105	1,10

a) Podatke grafično predstavite z razsevnim grafikonom.

b) Preverite, ali sta spremenljivki povezani (stopnja značilnosti 0,05). Obrazložite rezultate.

4. Ptiči

Iz literature povzemamo podatke za širino krila in dolžino repa pri določeni vrsti ptic. Podatki za vzorec 12 ptic so v tabeli.

Tabela 6-13: Širina krila in dolžina repa za 12 ptic

Št.	Širina krila (cm)	Dolžina repa (cm)
1	10,4	7,4
2	10,8	7,6
3	11,1	7,9
4	10,2	7,2
5	10,3	7,4
6	10,2	7,1
7	10,7	7,4
8	10,5	7,2
9	10,8	7,8
10	11,2	7,7
11	10,6	7,8
12	11,4	8,3

a) Grafično prikažite podatke.

b) Preverite, ali obstaja linearna povezava med širino krila in dolžino repa (stopnja značilnosti 0,05).

c) *Izračunajte 95% interval zaupanja za koeficient korelacije.

5. *Primerjava dveh ocenjevalcev

V vzorcu je bilo deset vrst jogurtov. Ocenjevala sta jih ocenjevalca A in B, tako da sta jih rangirala od 1 do 10 (1 najboljši, 10 najslabši). Delo sta opravila neodvisno drug od drugega. Podatki so v tabeli.

Tabela 6-14: Rangi jogurtov, ki sta jih določila ocenjevalca A in B

Jogurt	A	B
1	6	5
2	4	6
3	9	10
4	1	2
5	2	3
6	7	8
7	3	1
8	8	7
9	5	4
10	10	9

Zanima nas usklajenost ocenjevalca A in B. Pri stopnji značilnosti 0,01 preverite ustrezno domnevo.

6. *Primerjava Pearsonovega in Spearmanovega koeficienta korelacije

Iz tabel odčitajte kritično vrednost Pearsonovega in Spearmanovega koeficienta korelacije pri stopnji značilnosti 0,05 in velikosti vzorca 5, 10, 20, 40, 60 in 100. Primerjajte vrednosti in jih komentirajte.

7 χ^2-PREIZKUSI

7.1 PREIZKUŠANJE DOMNEVE O PORAZDELITVI SPREMENLJIVKE

Statistično spremenljivko modeliramo s slučajno spremenljivko in zanjo privzamemo določeno verjetnostno porazdelitev. Ali je predpostavka o privzeti verjetnostni porazdelitvi sprejemljiva? S tem problemom se do sedaj sploh nismo ukvarjali, navadno smo kar privzeli določeno verjetnostno porazdelitev. Rekli smo npr.: porazdelitev pridelka je normalna, porazdelitev inteligenčnega kvocienta je normalna. Sedaj pa bomo pogledali, kako na osnovi zbranih podatkov preverjamo domnevo o verjetnostni porazdelitvi. Najprej poglejmo dva primera za ilustracijo.

Primer 1

Ali je porazdelitev nesreč pri delu po dnevih v tednu *enakomerna*?

Naj bo X pojavnost nesreče po dnevih v tednu, njene vrednosti so ponedeljek, torek, sreda, četrtek in petek. Če je porazdelitev enakomerna, je njena porazdelitvena shema taka:

$$X: \begin{bmatrix} \text{P}, \text{T}, ..., \text{P} \\ \frac{1}{5}, \frac{1}{5}, ..., \frac{1}{5} \end{bmatrix}$$

Primer 2

Ali je v naši populaciji inteligenčni kvocient porazdeljen po normalni porazdelitvi s povprečno vrednostjo 100 in standardnim odklonom 15, torej: $IQ \sim N(100;15)$?

Če ta verjetnostna porazdelitev velja, je porazdelitvena shema slučajne spremenljivke IQ taka:

$$IQ: \begin{bmatrix} \leq 70 & \text{nad 70 do 85} & \text{nad 85 do 100} & \text{nad 100 do 115} & \text{nad 115 do 130} & > 130 \\ 0{,}0228 & 0{,}1359 & 0{,}3413 & 0{,}3413 & 0{,}1359 & 0{,}0228 \end{bmatrix}$$

Opomba: razrede lahko naredimo tudi drugače.

Naj $x_i, i = 1, ..., k$ označuje vrednost oz. razred spremenljivke X, x_i bomo imenovali celica.

Označimo *dejansko verjetnost* za to celico: $p(x_i) = p_i$. Teh verjetnosti seveda ne poznamo.

Teorija predvideva verjetnost za posamezno celico. To verjetnost bomo označili $p_{i0}, i = 1, ..., k$ in jo imenovali *teoretična verjetnost*. Indeksu smo dodali ničlo, ker bodo te verjetnosti zapisane z ničelno domnevo.

H_0 : verjetnostna porazdelitev slučajne spremenljivke X je:

$$X: \begin{bmatrix} x_1 & x_2 & ... & x_k \\ p_{10} & p_{20} & ... & p_{k0} \end{bmatrix}$$

Povejmo še enkrat: $p_{10}, p_{20}, ... p_{k0}$ so teoretične oz. pričakovane verjetnosti, ki izhajajo iz teorije. Zanje seveda velja:

$$\sum_{i=1}^{k} p_{i0} = 1$$

Alternativna domneva je negacija ničelne domneve.

H_1: H_0 ne velja.

Preverjali bomo domnevo o predpostavljeni verjetnostni porazdelitvi spremenljivke na osnovi podatkov iz vzorca. Iz vzorca bomo ocenili dejansko verjetnost in jo primerjali s teoretično verjetnostjo. Zato potrebujemo ustrezno vzorčno statistiko in njeno ničelno porazdelitev. Nadaljnja pot je taka:

V vzorcu je n enot, velikost vzorca je vnaprej določena. Predpostavimo, da je ničelna domneva pravilna. Potem v vzorcu velikosti n pričakujemo np_{10} enot v celici x_1, np_{20} enot v celici x_2 itd. Pripadajoče frekvence imenujemo **teoretične** oz. **pričakovane frekvence**, označimo jih $f_1^{'}, f_2^{'}, ..., f_k^{'}$, izračunamo jih z naslednjim izrazom:

$$f_i^{'} = np_{i0}, i = 1, ..., k$$

Zanje velja:

$$\sum_{i=1}^{k} f_i^{'} = n \cdot \sum_{i=1}^{k} p_{i0} = n$$

Poglejmo stanje v vzorcu. Ugotovimo, koliko enot v vzorcu uvrstimo v celico x_1, koliko v celico x_2 itd. Pripadajoče frekvence označimo $f_1, f_2, ..., f_k$, imenujemo jih **dejanske frekvence**. Zanje velja:

$$\sum_{i=1}^{k} f_i = n$$

Dejanske in pričakovane frekvence zapišimo v tabelo. Vsota dejanskih in pričakovanih frekvenc je enaka številu enot v vzorcu.

Tabela 7-1: Dejanske in pričakovane frekvence

	x_1	x_2	x_3	…	x_k	**Skupaj**
Dejanske frekvence	f_1	f_2	f_3	…	f_k	n
Pričakovane frekvence	$f_1^{'}$	$f_2^{'}$	$f_3^{'}$	…	$f_k^{'}$	n

Ideja statističnega preizkusa je naslednja: če se dejanske in pričakovane frekvence dovolj dobro ujemajo, H_0 obdržimo, sicer jo zavrnemo v korist H_1. Mera ujemanja temelji na razlikah $f_i - f_i^{'}$. Ker so določene razlike pozitivne, določene pa negativne, mera upošteva kvadrate razlik $\left(f_i - f_i^{'}\right)^2$. Karl Pearson (1857-1936) je razvil mero ujemanja dejanskih in pričakovanih frekvenc. Imenujemo jo **Pearsonova χ^2-statistika**:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i^{'})^2}{f_i^{'}}$$

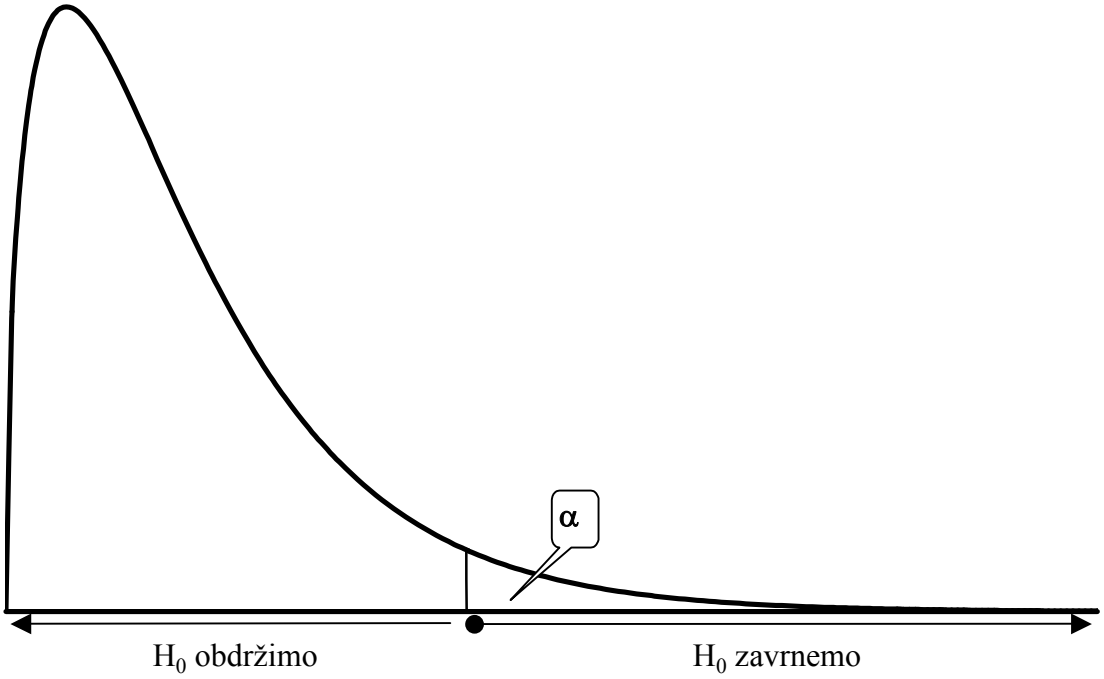
Očitno velja: bolj ko dejanske frekvence odstopajo od pričakovanih frekvenc, večja je vrednost Pearsonove χ^2-statistike. Pearson je izpeljal ničelno porazdelitev te statistike. Izrek pravi:

Če je ničelna domneva pravilna, je Pearsonova χ^2-statistika porazdeljena približno po χ^2-porazdelitvi s $k-1$ stopinjami prostosti.

Torej je sklep naslednji:

$\chi^2 \le \chi^2_\alpha(SP = k-1) \Rightarrow H_0$ obdržimo.

$\chi^2 > \chi^2_\alpha(SP = k-1) \Rightarrow H_0$ zavrnemo v korist H_1.



Slika 7-1: Hi-kvadrat porazdelitev in sklepanje

Opozorimo na naslednja dejstva:

- Stopinje prostosti se vežejo na število celic in ne na velikost vzorca. Velikost vzorca uporabimo pri izračunu pričakovanih frekvenc.
- Alternativna domneva je enostranska. Celotna α je v zgornjem repu porazdelitve.
- χ^2-porazdelitev je le aproksimacija prave ničelne porazdelitve. Teorija pove, da je aproksimacija sprejemljiva, če je pričakovana frekvenca v vsaki celici 5 ali več: $f_i^{'} \ge 5, i = 1, ..., k$
  Če ta pogoj ni izpolnjen, združimo sosednje celice (če je to vsebinsko smiselno).
- Vrednost χ^2-statistike je izračunana kot vsota prispevkov posameznih celic. Večji prispevek pomeni večje odstopanje dejanske frekvence od teoretične frekvence. Pomembno je analizirati te prispevke. Prav analiza teh odstopanj je bistvena za razumevanje obravnavanega pojava. Primeri bodo ilustrirali to dejstvo.

Primer 1 (nadaljevanje)

V določenem časovnem obdobju je bilo 500 nesreč pri delu, od tega 130 ob ponedeljkih, 90 ob torkih, 100 ob sredah, 90 ob četrtkih in 90 ob petkih. Na podlagi teh podatkov bomo preverili domnevo, da je porazdelitev nesreč pri delu po dnevih v tednu enakomerna ($\alpha = 0{,}05$).

H_0: Porazdelitev nesreč je enakomerna.

$$p_{po} = p_{to} = ... = p_{pe} = \frac{1}{5}$$

V vzorcu velikosti 500 so pričakovane frekvence:

$$f'_{po} = f'_{to} = \ldots = f'_{pe} = 500 \cdot \frac{1}{5} = 100$$

H_1 : Porazdelitev nesreč ni enakomerna.

Tabela 7-2: Dejanska in pričakovana frekvenca, njuna razlika ter prispevek k χ^2-statistiki za dan v tednu

Dan v tednu	Dejanska frekvenca	Pričakovana frekvenca	Razlika	χ^2-prispevek
Po	130	100	30	9
To	90	100	-10	1
Sr	100	100	0	0
Če	90	100	-10	1
Pe	90	100	-10	1
Skupaj	500	500	0	12

$$\chi^2 = \frac{30^2}{100} + \frac{10^2}{100} + \frac{0^2}{100} + \frac{10^2}{100} + \frac{10^2}{100} = 12$$

$\chi^2_{0,05}(4) = 9{,}488$

$p = 0{,}0174$

Ničelno domnevo zavrnemo.

Pri stopnji značilnosti 0,05 trdimo, da število nesreč ni enakomerno porazdeljeno po dnevih v tednu ($p = 0{,}0174$). Največja odstopanja so ob ponedeljkih. Podatki nakazujejo, da je nesreč ob ponedeljkih več, kot bi pričakovali pri enakomerni porazdelitvi.

Primer

Na Biotehniški fakulteti so izvedli poskus, v katerem so križali dvoredni ječmen s črnimi plevami s šestrednim ječmenom z rumenimi plevami. V F2 generaciji je bilo od skupno 1264 rastlin 735 črnih dvorednih (Č2), 232 črnih šestrednih (Č6), 223 rumenih dvorednih (R2) in 74 rumenih šestrednih (R6). Ali so dobljeni eksperimentalni rezultati v skladu s teoretičnim razmerjem, ki ga podaja Mendelova teorija: Č2:Č6:R2:R6= 9:3:3:1? ($\alpha = 0{,}05$)?

H_0 : razmerje je 9:3:3:1, oz.

$p_0(\text{Č2}) = \frac{9}{16}$, $p_0(\text{Č6}) = \frac{3}{16}$, $p_0(R2) = \frac{3}{16}$, $p_0(R6) = \frac{1}{16}$.

H_1 : razmerje ni 9:3:3:1.

Tabela 7-3: Dejanska in pričakovana frekvenca, njuna razlika ter prispevek k χ^2-statistiki za izid križanja

Izid križanja	Dejanska frekvenca	Pričakovana frekvenca	Razlika	χ^2-prispevek
Č2	735	711	24	0,8101
Č6	232	237	-5	0,1055
R2	223	237	-14	0,8270
R6	74	79	-5	0,3165
Skupaj	1264	1264	0	2,0591

$\chi^2 = 2{,}0591 \quad \chi^2_{0,05}(3) = 7{,}815 \qquad p = 0{,}5602$

Ničelno domnevo obdržimo. Eksperimentalni podatki ne nasprotujejo razmerju 9:3:3:1.

Poudarimo naj, da s tem poskusom seveda nismo dokazali pravilnosti Mendelove teorije, dobili smo rezultate, ki so v skladu z njegovo teorijo. Mero te skladnje ilustrira p-vrednost. V tem primeru ga interpretiramo takole: če je razmerje 9:3:3:1, dobimo v 56% poskusov z 1264 rastlinami zgolj zaradi slučajnih dejavnikov za χ^2-statistiko vrednost, ki je več kot 2,0591.

Primer

V tovarni vozil trdijo, da je verjetnost, da je izdelano vozilo brezhibno, enaka 0,9. V oddelku za kontrolo kakovosti izberejo s slučajno izbiro v kontrolni vzorec 4 vozila. Za vsako vozilo ugotovijo, ali je brezhibno ali ne. V kontrolnem vzorcu so lahko 0, 1, 2, 3 ali 4 brezhibna vozila.

V 200 kontrolnih vzorcih je bilo ugotovljeno naslednje stanje:

- 1 vzorec z 1 brezhibnim vozilom,
- 19 vzorcev z 2 brezhibnima voziloma,
- 78 vzorcev s 3 brezhibnimi vozili,
- 102 vzorca s 4 brezhibnimi vozili.

Pri $\alpha = 0{,}01$ preverimo trditev, da je verjetnost brezhibnega vozila 0,9.

Naj X označuje število brezhibnih vozil v kontrolnem vzorcu velikosti 4. Njegove možne vrednosti so: 0, 1, 2, 3, 4. Vozilo je brezhibno ali ne, torej lahko za X privzamemo binomsko porazdelitev $b(4, p)$. Ničelno domnevo zapišemo takole:

$H_0: \ p_0 = 0{,}9$

$H_1: \ p_0 \neq 0{,}9$

Izračunajmo verjetnostno shemo za X.

$$P(X=0) = \binom{4}{0} \cdot 0{,}9^0 \cdot 0{,}1^4 = 0{,}0001$$

$$P(X=1) = \binom{4}{1} \cdot 0{,}9^1 \cdot 0{,}1^3 = 0{,}0036$$

$$P(X=2) = \binom{4}{2} \cdot 0{,}9^2 \cdot 0{,}1^2 = 0{,}0486$$

$$P(X=3)=\binom{4}{3}\cdot 0{,}9^3\cdot 0{,}1^1=0{,}2916$$

$$P(X=4)=\binom{4}{4}\cdot 0{,}9^4\cdot 0{,}1^0=0{,}6561$$

Tabela 7-4: Dejansko in pričakovano število kontrolnih vzorcev glede na število brezhibnih vozil v kontrolnem vzorcu

Število brezhibnih vozil v kontrolnem vzorcu	Dejansko število kontrolnih vzorcev	Pričakovano število kontrolnih vzorcev
0	0	0,02
1	1	0,72
2	19	9,72
3	78	58,32
4	102	131,22
Skupaj	200	200,00

V prvih dveh razredih so pričakovane frekvence premajhne, zato prve tri razrede združimo in izračunamo vrednost χ^2-statistike.

Tabela 7-5: Dejansko in pričakovano število kontrolnih vzorcev ter prispevek k χ^2-statistiki za število brezhibnih vozil v kontrolnem vzorcu, razredi so združeni

Število brezhibnih vozil v kontrolnem vzorcu	Dejansko število kontrolnih vzorcev	Pričakovano število kontrolnih vzorcev	Prispevek k χ^2-statistiki
0, 1, 2	20	10,46	8,70
3	78	58,32	6,64
4	102	131,22	6,51
Skupaj	200	200,00	21,85

$\chi^2 = 21{,}85 \qquad \chi^2_{0,01}(2) = 9{,}210 \qquad p = 0{,}00002$

Ničelno domnevo zavrnemo. Pri stopnji značilnosti 0,01 trdimo, da verjetnost brezhibnega vozila ni 0,9 ($p = 0{,}00002$).

Oceno za verjetnost brezhibnega vozila izračunamo na osnovi dobljenih podatkov takole:

$$\hat{p}=\frac{1+2\cdot 19+3\cdot 78+4\cdot 102}{800}=0{,}851$$

Podatki nakazujejo, da je verjetnost brezhibnega vozila, ki jo podaja tovarna, precenjena.

Za zvezne spremenljivke je potrebno podatke uvrstiti v razrede, da lahko uporabimo χ^2-statistiko. Kako naredimo razrede, je prepuščeno nam. Nekateri statistiki predlagajo uporabo kvantilov (npr. kvartile), včasih pa so meje vsebinsko določene. Poglejmo primer.

Primer

Zanima nas, ali lahko za maso določenega izdelka privzamemo normalno porazdelitev. V vzorcu je bilo 300 izdelkov, njihove mase (g) so:

498, 494, 501, 506, 506, 509, 489, 499, 505, 495, 497, 492, 491, 495, 496, 489, 497, 498, 501, 498, 498, 498, 507, 500, 499, 497, 510, 504, 512, 497, 508, 492, 503, 505, 510, 500, 497, 503,

498, 504, 493, 496, 492, 498, 500, 500, 498, 511, 491, 496, 487, 507, 494, 497, 504, 502, 504, 503, 493, 494, 503, 502, 495, 499, 501, 503, 501, 495, 509, 502, 500, 504, 504, 497, 495, 506, 494, 492, 504, 503, 511, 507, 507, 501, 500, 502, 500, 495, 491, 504, 502, 503, 501, 495, 506, 498, 496, 496, 498, 498, 497, 504, 503, 497, 507, 491, 503, 499, 500, 497, 497, 504, 504, 502, 503, 508, 502, 503, 509, 498, 505, 501, 506, 499, 496, 505, 497, 503, 503, 498, 498, 493, 510, 497, 500, 499, 514, 506, 504, 507, 501, 503, 499, 494, 506, 499, 493, 504, 504, 502, 502, 497, 504, 503, 505, 486, 502, 507, 491, 500, 507, 508, 501, 501, 509, 500, 497, 505, 500, 505, 502, 507, 499, 495, 492, 487, 501, 501, 500, 501, 501, 507, 500, 494, 493, 501, 505, 492, 497, 503, 499, 498, 499, 498, 496, 503, 500, 506, 501, 500, 501, 501, 500, 494, 501, 501, 504, 496, 503, 501, 501, 504, 509, 493, 494, 503, 506, 496, 489, 496, 501, 502, 495, 496, 502, 501, 497, 501, 495, 496, 505, 508, 497, 499, 501, 509, 503, 508, 501, 508, 497, 495, 506, 505, 491, 506, 499, 494, 491, 496, 502, 505, 500, 500, 501, 498, 505, 498, 512, 504, 503, 495, 504, 507, 496, 498, 493, 507, 505, 501, 489, 503, 503, 506, 495, 506, 497, 496, 491, 507, 495, 508, 494, 498, 493, 508, 500, 502, 505, 510, 500, 495, 497, 504, 504, 503, 493, 493, 499, 497.

Podatke imamo v datoteki, vse izračune izvedemo z računalnikom.

Na osnovi podatkov bomo poskušali ugotoviti, ali je normalna porazdelitev sprejemljiva za to spremenljivko. Parametrov normalne porazdelitve ne poznamo, zato jih bomo ocenili iz vzorca. Osnovne statistike so:

min = 486 g

max = 514 g

povprečje = 500,28 g

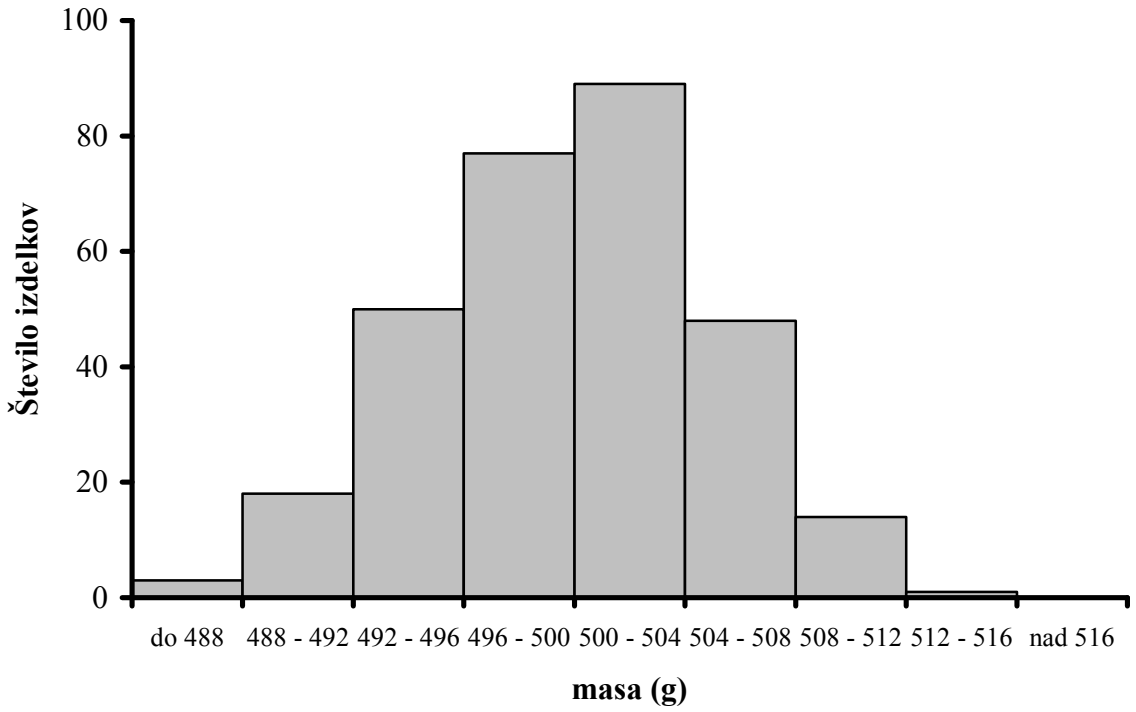
standardni odklon = 5,17 g

Z računalnikom smo podatke uvrstili v razrede širine 4 g in dobili naslednjo frekvenčno porazdelitev:

Tabela 7-6: Frekvenčna porazdelitev za maso izdelka

Masa (g)	Dejanska frekvenca
do 488	3
nad 488 do 492	18
nad 492 do 496	50
nad 496 do 500	77
nad 500 do 504	89
nad 504 do 508	48
nad 508 do 512	14
nad 512 do 516	1
nad 516	0
Skupaj	300

Grafični prikaz frekvenčne porazdelitve nakazuje, da je oblika empirične porazdelitve podobna normalni porazdelitvi.



Slika 7-2: Histogram za maso izdelka

Izvedimo formalni preizkus. Za parametra normalne porazdelitve bomo vzeli njuni oceni, ki sta izračunani iz vzorčnih podatkov.

H_0: porazdelitev je normalna $N(500{,}28;\ 5{,}17)$.

H_1: H_0 ne velja.

Z računalnikom smo izračunali teoretične verjetnosti za posamezne razrede in na osnovi le-teh pripadajoče teoretične frekvence. Razrede s premajhnimi teoretičnimi frekvencami smo združili. Prispevki k χ^2-statistiki so v tabeli:

Tabela 7-7: Dejanska in pričakovana frekvenca ter prispevek k χ^2-statistiki za maso izdelka

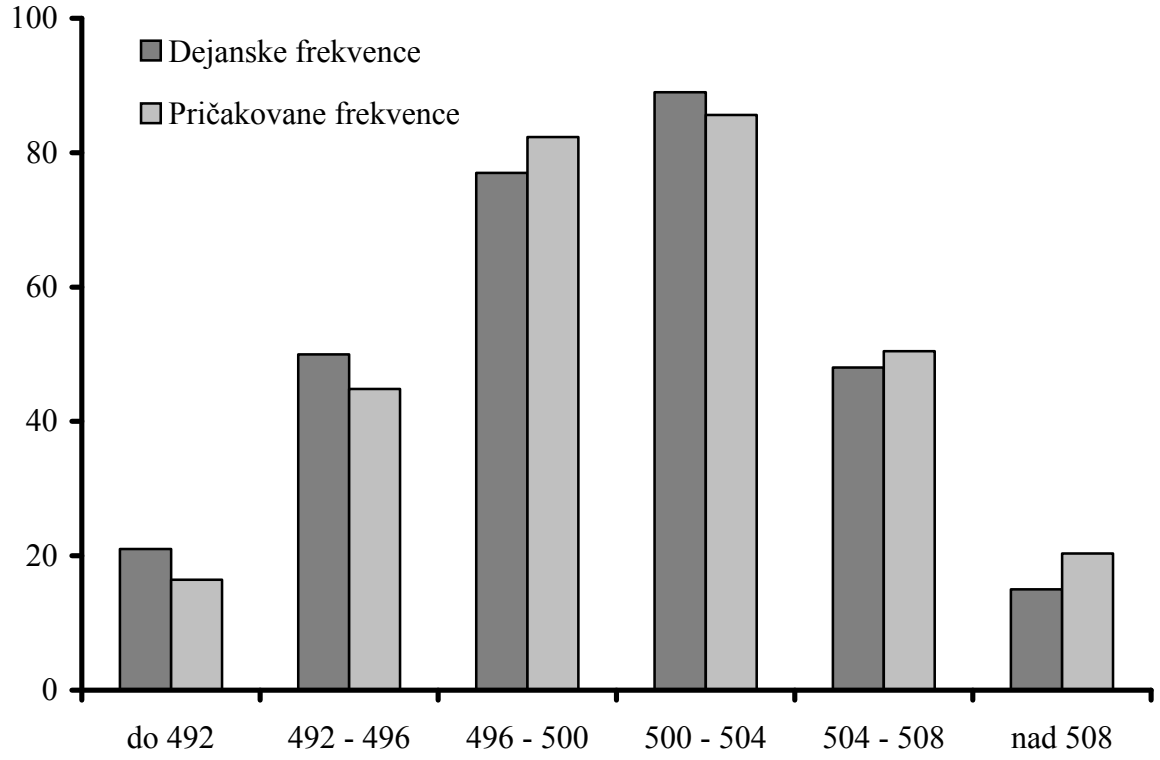
Masa (g)	**Dejanske frekvence**	**Pričakovane Frekvence**	**Prispevek k χ^2-statistiki**
do 492	21	16,45	1,26
nad 492 do 496	50	44,82	0,60
nad 496 do 500	77	82,33	0,35
nad 500 do 504	89	85,63	0,13
nad 504 do 508	48	50,44	0,12
nad 508	15	20,33	1,40
Skupaj	300	300,00	3,85

V tem primeru smo iz podatkov ocenili tudi parametra predpostavljene normalne porazdelitve in na tej osnovi izračunali pričakovane frekvence. V takem primeru se stopinje prostosti za χ^2-porazdelitev dodatno zmanjšajo za število ocenjenih parametrov. Ker smo iz podatkov

ocenili oba parametra normalne porazdelitve, ima ničelna porazdelitev 2 stopinji prostosti manj, torej: SP=6-1-2=3.

$\chi^2 = 3{,}85 \qquad \chi^2_{0,05}(3) = 7{,}615 \qquad p = 0{,}2784$

Ničelno domnevo obdržimo. Podatki nakazujejo, da je normalna porazdelitev sprejemljiv verjetnostni model za maso artikla ($p = 0{,}2784$).



Slika 7-3: Primerjava dejanskih in pričakovanih frekvenc

In končno: kaj lahko naredimo, če podatki niso skladni z domnevno verjetnostno porazdelitvijo? Npr. če podatki nakazujejo, da porazdelitev slučajne spremenljivke ni normalna, kar je pogoj za določene preizkuse, ki so bili predstavljeni. Poti je več, omenimo dve:

- transformacija podatkov z ustrezno funkcijo (npr. logaritemsko, korensko), ki asimetrično porazdelitev prevede v 'dovolj normalno';
- uporaba t. i. neparametričnih statističnih preizkusov. Ti preizkusi ne temeljijo na verjetnostni porazdelitvi, ne preizkušajo parametrov verjetnostne porazdelitve. Predpostavke za uporabo neparametričnih preizkusov so blažje kot predpostavke pripadajočih parametričnih preizkusov, po drugi strani pa je moč teh preizkusov manjša.

Obravnavanje te problematike presega obseg tega učbenika.

NALOGE

1. Pošteni kocki

Igralec je metal hkrati dve kocki in zapisal, kolikokrat je pri metu padla šestica: to zgodilo nobenkrat, enkrat ali dvakrat.

a) Če privzamemo, da je verjetnost, da vrže s posamezno kocko šestico, enaka 1/6, kolikšna je verjetnost, da bo pri metu dveh kock dobil 0 šestic, 1 šestico, 2 šestici?

b) Metal je 180-krat in dobil naslednje rezultate:

Tabela 7-8: Frekvenca metov šestic z dvema kockama

Izid pri metu dveh kock	**0 šestic**	**1 šestica**	**2 šestici**
Število metov	105	70	5

Pri stopnji značilnosti 0,05 preverite domnevo, da sta obe kocki 'pošteni', torej, da je verjetnost za met šestice 1/6 pri vsaki od kock.

2. Slučajne števke

Da bi ugotovili, ali Tabela slučajnih števk zadošča ustrezni verjetnostni porazdelitvi, so pregledali 1000 slučajnih števk in ugotovili, kolikokrat nastopajo števke 0, 1, 2,....9.

Tabela 7-9: Frekvenca števk

Števka	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frekvenca	98	104	105	92	99	100	102	104	104	92

a) Grafično prikažite podatke.

b) Kateri verjetnostni porazdelitvi mora biti zadoščeno, da je Tabela slučajnih števk ustrezna?

c) Kaj lahko sklepamo na osnovi ugotovljenih frekvenc? (stopnja značilnosti 0,01)

3. Kakovost nogavic

Proizvajalec nogavic trdi, da je 80% njegovih izdelkov prve kakovosti, 15% izdelkov druge kakovosti, preostalih 5% pa neprimernih za prodajo. Pregledali so 200 parov nogavic in ugotovili, da je bilo 165 parov prve kakovosti, 30 parov druge kakovosti, 5 parov neprimernih za prodajo.

Ali pregled izdelkov potrjuje izjavo proizvajalca? Postavite ustrezno domnevo in jo preverite (stopnja značilnosti 0,05). Obrazložite rezultate.

4. Beljakovine v mleku

Kmetijska zadruga mesečno analizira vsebnost beljakovin v mleku pri vseh kmetih, od katerih odkupuje mleko. Analizo opravijo tako, da pri vsakem kmetu 4-krat mesečno vzamejo stekleničko s po 100 ml mleka in določijo vsebnost beljakovin. Iz teh štirih vrednosti izračunajo mesečno povprečje za količino beljakovin. Glede na vrednost tega povprečja so kmetje razdeljeni v 5 cenovnih razredov, kot je razvidno iz tabele spodaj.

Navajamo podatke za vzorec 246 kmetov, ki jih je v oktobru 1994 dobila Kmetijska zadruga Medvode.

Tabela 7-10: Število kmetov v Kmetijski zadrugi Medvode v oktobru 1994 glede na cenovni razred

Cenovni razred	**Vsebnost beljakovin v mleku (%)**	**Število kmetov**
1	nad 3,35	46
2	nad 3,25 do 3,35	32
3	nad 3,15 do 3,25	39
4	nad 2,80 do 3,15	121
5	2,80 in manj	8
Skupaj		246

a) Vsebnost beljakovin naj bi bila porazdeljena N(3,15%; 0,30%). Če ta predpostavka velja, kolikšen odstotek kmetov pričakujemo po cenovnih razredih?
b) Grafično prikažite podatke. Kakšna je oblika porazdelitve?
c) Ali podatki potrjujejo domnevo o porazdelitvi N(3,15%; 0,30%)? (stopnja značilnosti 0,05)

7.2 ANALIZA KONTINGENČNIH TABEL

7.2.1 Uvod

Analiziramo dve spremenljivki, X in Y. Za razliko od situacije, ki smo jo obravnavali pri regresiji in korelaciji, ko smo analizirali številske spremenljivke, bomo sedaj pozornost posvetili opisnim spremenljivkam. Analizirali bomo dve opisni spremenljivki z imensko ali urejenostno mersko lestvico. Naj ima X k vrednosti, Y pa r vrednosti. Podatki so urejeni v t. i. **kontingenčni tabeli**. V celicah tabele so frekvence. Tabela ima robno vrstico in robni stolpec. V nadaljnji teoriji bomo uporabljali oznake, ki so razvidne spodaj.

Tabela 7-11: Shema kontingenčne tabele z oznakami

	Y						
X	y_1	y_2	…	y_j	…	y_r	**Skupaj**
x_1	f_{11}	f_{12}		f_{1j}		f_{1r}	$f_{1\cdot}$
x_2	f_{21}	f_{22}		f_{2j}		f_{2r}	$f_{2\cdot}$
…							…
x_i	f_{i1}	f_{i2}		f_{ij}		f_{ir}	$f_{i\cdot}$
…							…
x_k	f_{k1}	f_{k2}		f_{kj}		f_{kr}	$f_{r\cdot}$
Skupaj	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot j}$	…	$f_{\cdot r}$	n

f_{ij} je število enot v i-ti vrstici in j-tem stolpcu. To je število enot, ki ima pri X vrednost x_i in pri Y vrednost y_j. Te frekvence imenujemo **dejanske frekvence**. Tabela ima robne vsote, ki jih izračunamo za vrstice takole:

$$f_{i\cdot} = \sum_{j=1}^{r} f_{ij},\ i = 1, ..., k$$

Ta števila tvorijo robni stolpec. Robne vsote za stolpce tvorijo robno vrstico:

$$f_{\cdot j} = \sum_{i=1}^{k} f_{ij},\ j = 1, ..., r$$

Vsota vseh frekvenc je enaka številu obravnavanih enot:

$$n = \sum_{i=1}^{k}\sum_{j=1}^{r} f_{ij}$$

Poglejmo nekaj primerov kontingenčnih tabel in z njimi povezanih vprašanj.

Primer 1: povezanost dveh opisnih spremenljivk

V vzorcu je bilo 6800 slučajno izbranih oseb. Za vsako so ugotavljali barvo las in barvo oči. Podatki so:

Tabela 7-12: Barva oči in barva las za 6800 slučajno izbranih oseb

	Barva las				
Barva oči	**Svetla**	**Rjava**	**Črna**	**Rdeča**	**Skupaj**
Modra	1768	807	189	47	2811
Siva/zelena	946	1387	746	53	3132
Rjava	115	438	288	16	857
Skupaj	2829	2632	1223	116	**6800**

Tabela kaže, da je od 6800 vzorčenih oseb imelo 1768 modre oči in svetlo barvo las, 807 modro barvo oči in rjavo barvo las itd. Z zasnovo opazovanja je bila vnaprej določena samo velikost vzorca. Obe spremenljivki sta slučajni spremenljivki, iz velikosti vzorca ne vemo, kakšne so robne vsote. Zanju lahko privzamemo dvorazsežno verjetnostno porazdelitev. Zanima nas, ali sta barva las in barva oči povezani spremenljivki. Na osnovi tega vzorca bomo sklepali na to, kaj velja v populaciji, katere reprezentant je analizirani vzorec.

Problem povezanosti dveh spremenljivk smo že obravnavali, vendar za dve številski spremenljivki. Mera njune povezanosti je koeficient korelacije (če smo za spremenljivke lahko privzeli dvorazsežno normalno porazdelitev) oz. koeficient korelacije rangov (če smo gledali povezanost dveh spremenljivk z urejenostno mersko lestvico).

Primer 2: homogenost struktur

S slučajno izbiro so izbrali 100 žensk in 160 moških in ugotavljali njihovo izobrazbo. Podatki so v tabeli.

Tabela 7-13: Število oseb po spolu in izobrazbi

Izobrazba	**Ženske**	**Moški**	**Skupaj**
Osnovna	10	15	25
Srednja	70	100	170
Višja ali več	20	45	65
Skupaj	**100**	**160**	260

Zanima nas, ali je v proučevani populaciji izobrazbena struktura moških enaka izobrazbeni strukturi žensk. Povedano drugače: ali sta strukturi homogeni. Tu imamo dva slučajna vzorca, slučajni vzorec moških, ki reprezentira populacijo moških, in slučajni vzorec žensk, ki reprezentira populacijo žensk. Vzorca sta neodvisna. V tem primeru sta stolpčni robni vsoti določeni z zasnovo študije.

Primer 3: podatki v parih

Zanima nas, kako se vremenska napoved ujema z dejanskim stanjem vremena pri različnih vrstah vremena (sončno, oblačno, padavine). V vzorcu je bilo 141 opazovanih dni. Podatki so:

Tabela 7-14: Dejansko in napovedano vreme za vzorec 141 dni

	Napovedano vreme			
Dejansko vreme	**Sončno**	**Oblačno**	**Padavine**	**Skupaj**
Sončno	34	24	17	75
Oblačno	4	21	3	28
Padavine	6	9	23	38
Skupaj	44	54	43	**141**

Za vsak izbrani dan imamo dva podatka: napovedano vreme in dejansko vreme. Podatki so v parih, torej sta vzorec, za katerega imamo napoved, in vzorec, za katerega imamo dejansko stanje, odvisna. Zanima nas, ali se verjetnost za napovedano sončno vreme ujema z verjetnostjo za sončno vreme; ali se verjetnost za napovedano oblačno vreme ujema z verjetnostjo oblačnega vremena in ali se verjetnost napovedanih padavin ujema z verjetnostjo dejanskih padavin.

V nadaljevanju bomo pogledali, kako rešujemo te tri vsebinsko različne probleme. V prvem primeru bomo analizirali **povezanost dveh opisnih spremenljivk**, v drugem **homogenost struktur**, v tretjem pa **analizo podatkov v parih**. Izkazalo se bo, da je testna statistika v vseh treh primerih ista.

7.2.2 Povezanost dveh opisnih spremenljivk

Analiziramo dve opisni spremenljivki X in Y. Zanima nas, ali sta povezani ali ne. V vzorcu je n enot, velikost vzorca je vnaprej določena. Na vsaki enoti imamo vzorčni podatek za X in za Y. Podatke uredimo v dvorazsežno kontingenčno tabelo. V celicah tabele so dejanske frekvence. Zapišimo ničelno in alternativno domnevo:

H_0: X in Y nista povezani.

H_1: X in Y sta povezani.

Če ničelna domneva velja, pričakujemo v (*i,j*)-ti celici določeno število enot izmed *n* enot. To število imenujemo **pričakovana frekvenca** in jo označimo $f_{ij}^{'}$. Verjetnost za izid v (*i,j*)-ti celici označimo p_{ij}.

Tabela 7-15: Verjetnostna shema pri analizi povezanosti

	Y						
X	y_1	y_2	…	y_j	…	y_r	**Skupaj**
x_1	p_{11}	p_{12}		p_{1j}		p_{1r}	$p_{1\cdot}$
x_2	p_{21}	p_{22}		p_{2j}		p_{2r}	$p_{2\cdot}$
…							
x_i	p_{i1}	p_{i2}		p_{ij}		p_{ir}	$p_{i\cdot}$
…							
x_k	p_{k1}	p_{k2}		p_{kj}		p_{kr}	$p_{k\cdot}$
Skupaj	$p_{\cdot 1}$	$p_{\cdot 2}$		$p_{\cdot j}$		$p_{\cdot r}$	1

Iz verjetnostnega računa poznamo izrek o neodvisnosti dogodkov, ki pravi: dogodka A in B sta neodvisna natanko takrat, kadar je verjetnost produkta njunih dogodkov enaka produktu pripadajočih verjetnosti: $P(AB) = P(A) \cdot P(B)$. Iz tega izreka sledi: če ničelna domneva velja, je verjetnost za izid v (*i,j*)-ti celici p_{ij} produkt pripadajočih robnih verjetnosti $p_{i\cdot}$ in $p_{\cdot j}$:

$$P(X = x_i, Y = y_j) = p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$$

Ničelno domnevo matematično zapišemo takole:

$$H_0: \; p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, i = 1, ..., k, j = 1, ..., r$$

Izrazimo te verjetnosti s pričakovanimi frekvencami. Verjetnost p_{ij} je pričakovana frekvenca, deljena z velikostjo vzorca:

$$p_{ij} = \frac{f'_{ij}}{n}$$

Isto velja tudi za robne verjetnosti. Pokažemo lahko, da se robne vsote pričakovanih frekvenc ujemajo z robnimi vsotami dejanskih frekvenc:

$$f_{i\cdot} = f'_{i\cdot} \qquad f_{\cdot j} = f'_{\cdot j}$$

Iz zgornjega izhaja:

$$p_{i\cdot} = \frac{f'_{i\cdot}}{n} = \frac{f_{i\cdot}}{n} \qquad p_{\cdot j} = \frac{f'_{\cdot j}}{n} = \frac{f_{\cdot j}}{n}$$

$$\frac{f'_{ij}}{n} = \frac{f_{i\cdot}}{n} \cdot \frac{f_{\cdot j}}{n}$$

Končno: pričakovano frekvenco v (*i,j*)-ti celici izrazimo z robnima vsotama dejanskih frekvenc:

$$f'_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{n}$$

Povejmo to pravilo z besedami: *za izračun pričakovane frekvence v (i,j)-ti celici zmnožimo robno vsoto v i-ti vrstici z robno vsoto v j-tem stolpcu in produkt delimo z velikostjo vzorca.*

Za vsako celico tabele znamo izračunati pričakovano frekvenco, torej frekvenco, ki bi jo pričakovali v vzorcu velikosti *n*, če bi bila ničelna domneva pravilna. Iz podatkov pa imamo

dejanske frekvence. Poznamo mero ujemanja pričakovanih frekvenc z dejanskimi frekvencami, to je **Pearsonova χ^2-statistika**:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(f_{ij} - f_{ij}^{'})^2}{f_{ij}^{'}}$$

Vemo, da je ničelna porazdelitev Pearsonove χ^2-statistike približno χ^2-porazdelitev. Poglejmo še, koliko stopinj prostosti ima ničelna porazdelitev. Število stopinj prostosti je enako številu neodvisnih podatkov minus število neodvisnih parametrov, ki jih ocenjujemo iz podatkov. Imamo $kr-1$ neodvisnih podatkov (en podatek je določen z vsoto *n*), ocenjujemo $k-1$ vrstičnih robnih verjetnosti (ena verjetnost je določena z vsoto) ter $r-1$ stolpčnih robnih verjetnosti (ena verjetnost je določena z vsoto). Število stopinj prostosti torej je:

$$SP = rk - 1 - (k-1) - (r-1) = (k-1)(r-1)$$

Stopinje prostosti določa dimenzija tabele in ne število enot v vzorcu. Torej je ničelna porazdelitev za Pearsonovo χ^2-statistiko:

$$\chi^2(SP = (k-1)\cdot(r-1))$$

Pogoj za uporabo Pearsonove χ^2-statistike je:

$$f_{ij}^{'} \geq 5,\ i = 1,...,k,\ j = 1,...,r$$

Če ta pogoj ni izpolnjen, združimo sosednje vrstice oz. stolpce (če je to vsebinsko smiselno).

Opomba: χ^2-statistika se vedno računa na frekvencah, nikoli na odstotkih ali na deležih.

Primer 1 (nadaljevanje)

Pri $\alpha = 0{,}05$ preverimo domnevo o povezanosti barve las in barve oči.

H_0: Barva las in barva oči nista povezani.

H_1: Barva las in barva oči sta povezani.

Za vsako celico tabele moramo izračunati pričakovano frekvenco. Začnimo v prvi vrstici in prvem stolpcu:

$$f_{11}^{'} = \frac{2811\cdot 2829}{6800} = 1169{,}5$$

Obrazložimo to vrednost. Če povezave med barvo las in barvo oči ne bi bilo, bi v vzorcu velikosti 6800 pričakovali 1170 oseb z modro barvo oči ter svetlo barvo las. V našem vzorcu pa je bilo 1768 takih oseb, torej skoraj 600 več.

$$f_{12}^{'} = \frac{2811\cdot 2632}{6800} = 1088{,}0$$

Če povezave med barvo las in barvo oči ne bi bilo, bi v vzorcu velikosti 6800 pričakovali 1088 oseb z modro barvo oči in rjavo barvo las. V našem vzorcu pa je bilo 807 takih oseb, torej približno 200 manj.

Tako z izračuni nadaljujemo. V prvo tabelo smo zapisali dejanske in pričakovane frekvence ter razlike med njimi, v drugo pa prispevke k χ^2-statistiki.

Tabela 7-16: Dejanske in pričakovane frekvence ter njihove razlike

	Barva las				
Barva oči	**Svetla**	**Rjava**	**Črna**	**Rdeča**	**Skupaj**
Modra	1768	807	189	47	2811
	1169,5	1088,0	505,6	48,0	2811,1
	598,5	-281,0	-316,6	-1	
Siva/zelena	946	1387	746	53	3132
	1303,0	1212,3	563,3	53,4	3132,0
	-357,0	174,7	182,7	-0,4	
Rjava	115	438	288	16	857
	356,5	331,7	154,1	14,6	856,9
	-242,0	106,0	134,0	1,4	
Skupaj	2829	2632	1223	116	**6800**

Vidimo, da je največja razlika med dejanskimi in pričakovanimi frekvencami v prvi celici tabele, pri modrookih svetlolascih. Izračunajmo prispevke k χ^2-statistiki:

Tabela 7-17: Prispevki k χ^2-statistiki

	Barva las			
Barva oči	**Svetla**	**Rjava**	**Črna**	**Rdeča**
Modra	306,3	72,6	198,2	0,0
Siva/zelena	97,8	25,2	59,3	0,0
Rjava	163,6	34,1	116,3	0,1

$\chi^2 = 1074 \quad \chi^2_{0,05}(SP = 6) = 12{,}592 \quad p = 0{,}0000$

Ničelno domnevo zavrnemo. Vrednost za χ^2-statistiko je izjemno velika. Brez tveganja trdimo, da sta barva oči in barva las povezani spremenljivki ($p = 0{,}0000$). Očitna povezava je modra barva oči in svetla barva las.

Ugotovili smo, da so največji prispevki k χ^2-statistiki v prvi vrstici (modrooki) in prvem stolpci (svetlolasi). Za ilustracijo nadaljujemo z analizo tako, da iz podatkov izločimo modrooke in svetlolase. S tem se velikost vzorca zmanjša na 2928. Zopet nas bo zanimalo, ali sta barva las in barva oči povezani spremenljivki ($\alpha = 0{,}05$). Pri tem pa bomo upoštevali samo rjavo, črno in rdečo barvo las ter sivo/zeleno ter rjavo barvo oči.

H_0: Barva las in barva oči nista povezani.

H_1: Barva las in barva oči sta povezani.

Tabela 7-18: Dejanske in pričakovane frekvence ter njihove razlike

	Barva las			
Barva oči	**Rjava**	**Črna**	**Rdeča**	**Skupaj**
Siva/zelena	1387	746	53	2186
	1362,5	772,0	51,5	2186
	24,5	-26,0	1,5	
Rjava	438	288	16	742
	462,5	262,0	17,5	742
	-24,5	26,0	-1,5	
Skupaj	1825	1034	69	**2928**

Prispevki k χ^2-statistiki niso prav veliki. Navajamo jih spodaj.

Tabela 7-19: Prispevki k χ^2-statistiki

	Barva las		
Barva oči	**Rjava**	**Črna**	**Rdeča**
Siva/zelena	0,439931	0,873568	0,042846
Rjava	1,296077	2,573612	0,126228

$\chi^2 = 5{,}35 \quad \chi^2_{0,05}(SP = 2) = 5{,}991 \quad p = 0{,}0688$

Ničelno domnevo obdržimo.

Na osnovi tega vzorca ni razvidna povezanost med navedenimi barvami las in barvami oči ($p = 0{,}07$).

Opomba: poudariti je treba, da je drugi del analize, ko smo izločili modrooke in svetlolase, le ilustrativen. Pri statističnem sklepanju ni pravilno, da postavljamo domneve, ko smo si že ogledali podatke, saj pri takem testiranju verjetnost za napako I. vrste ni enaka predpisani vrednosti. Pravilno bi bilo, da bi za preverjanje druge domneve zbrali nove podatke.

7.2.2.1 2 × 2 kontingenčna tabela

Izkaže se, da lahko za 2×2 kontingenčno tabelo vrednost za χ^2-statistiko izračunamo po krajši poti. Uporabili bomo naslednje oznake:

Tabela 7-20: 2 × 2 kontingenčna tabela

	Y		
X	y_1	y_2	**Skupaj**
x_1	a	b	$a+b$
x_2	c	d	$c+d$
Skupaj	$a+c$	$b+d$	n

Vrednost χ^2-statistike lahko izračunamo direktno iz podatkov:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(f_{ij} - f_{ij}^{'})^2}{f_{ij}^{'}} = \frac{n \cdot (ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Izračun pričakovanih frekvenc vseeno priporočamo, ker pripomore k boljšemu razumevanju problema ter h kontroli velikosti pričakovanih frekvenc. Če se zgodi, da je v 2×2 kontingenčni tabeli pričakovana frekvenca pod 5, priporočamo uporabo Fisherjevega eksaktnega preizkusa, ki ga tu ne navajamo.

Primer

Zanima nas, ali je pri otrocih pojavnost bolezni A povezana s pojavnostjo bolezni B. V slučajnem vzorcu je 172 otrok, zanimalo nas je, ali je otrok prebolel bolezen A in ali je prebolel bolezen B. Podatki kažejo naslednje:

Tabela 7-21: Število otrok glede na prebolelo bolezen A in B in pričakovane frekvence

	Prebolel B		
Prebolel A	**Da**	**Ne**	**Skupaj**
Da	112	33	145
	98,6	46,4	145
Ne	5	22	27
	18,4	8,4	27
Skupaj	117	55	**172**

Pri stopnji značilnosti $\alpha = 0{,}05$ preverimo domnevo o povezanosti bolezni A in B.

H_0: Bolezen A in bolezen B nista povezani.

H_1: Bolezen A in bolezen B sta povezani.

Pričakovane frekvence smo zapisali zgoraj. Kažejo nam, da je izračun χ^2-statistike upravičen. Krajša pot za izračun χ^2-statistike je:

$$\chi^2 = \frac{172 \cdot (112 \cdot 22 - 33 \cdot 5)^2}{145 \cdot 27 \cdot 117 \cdot 55} = 36{,}08495$$

$\chi^2_{0,05}(SP = 1) = 3{,}841 \quad p = 0{,}0000$

Ničelno domnevo zavrnemo.

Pri stopnji značilnosti 0,05 trdimo, da sta bolezni A in B povezani ($p = 0{,}0000$).

7.2.3 Homogenost struktur

Druga uporaba χ^2-kvadrat preizkusa je prikazana v Primeru 2. Z zasnovo študije je dana robna vrstica, vsaka njena celica predstavlja velikost slučajnega vzorca določene velikosti. Ti slučajni vzorci so med seboj neodvisni. Zanima nas, ali je struktura za vsak stolpec enaka, ali je **struktura homogena**. Zapišimo ničelno in alternativno domnevo:

H_0: Struktura po stolpcih je enaka.

H_1: Struktura po stolpcih ni enaka.

Uredimo pripadajoče verjetnosti v tabelo.

Tabela 7-22: Verjetnostna shema pri analizi homogenosti struktur

	Y						
X	y_1	y_2	…	y_j	…	y_r	**Skupaj**
x_1	p_{11}	p_{12}		p_{1j}		p_{1r}	p_1
x_2	p_{21}	p_{22}		p_{2j}		p_{2r}	p_2
…							…
x_i	p_{i1}	p_{i2}		p_{ij}		p_{ir}	p_i
…							…
x_k	p_{k1}	p_{k2}		p_{kj}		p_{kr}	p_k
Skupaj	1	1	…	1	…	1	1

Predpostavimo, da ničelna domneva velja. Matematično ničelno domnevo zapišemo takole:

$H_0: \ p_{i1} = p_{i2} = ... = p_{ir} = p_i, \ i = 1, ..., k$

Za vsako vrstico tabele izrazimo vrstično verjetnost p_i s frekvenco:

$$p_i = \frac{f_{i\cdot}^{'}}{n} = \frac{f_{i\cdot}}{n}$$

Pričakovana frekvenca v (*i*,*j*)-ti celici je na osnovi te ugotovitve in na osnovi velikosti pripadajočega vzorca $f_{\cdot j}$ izražena takole:

$$f_{ij}^{'} = p_i \cdot f_{\cdot j} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{n}$$

Prišli smo torej do enakega izračuna pričakovanih frekvenc kot v primeru testiranja povezanosti dveh opisnih spremenljivk. Tudi Pearsonova χ^2-statistika ima enako obliko:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(f_{ij} - f_{ij}^{'})^2}{f_{ij}^{'}}$$

Poglejmo še stopinje prostosti. Preizkus homogenosti lahko štejemo kot preverjanje domneve o porazdelitvi, vendar za vsak stolpec posebej. Za vsak stolpec imamo $(k-1)$ neodvisnih podatkov, teh stolpcev je *r*. Ocenili smo $(k-1)$ neodvisnih verjetnosti. Število stopinj prostosti je:

$SP = r(k-1) - (k-1) = (r-1)(k-1)$

Stopinje prostosti so iste kot pri testu povezanosti. Torej gre izračun v tem primeru natanko po istem postopku kot v primeru preverjanja povezanosti dveh opisnih spremenljivk, le ničelna domneva se zapiše drugače.

Primer 2 (nadaljevanje)

Pri stopnji značilnosti 0,05 preverimo domnevo o enaki izobrazbeni strukturi po spolu.

H_0: Izobrazbena struktura po spolu je enaka.

H_1: Izobrazbena struktura po spolu ni enaka.

Tabela 7-23: Dejanske in pričakovane frekvence ter prispevki k χ^2-statistiki

Izobrazba	Ženske	Moški	Skupaj
Osnovna	10	15	25
	9,6	15,4	25
	0,015	0,010	
Srednja	70	100	170
	65,4	104,6	170
	0,326	0,204	
Višja ali več	20	45	65
	25	40	65
	1,000	0,625	
Skupaj	**100**	**160**	260
	100	160	260
	1,341	0,838	2,179

$\chi^2 = 2{,}179 \quad \chi^2_{0,05}(SP = 2) = 5{,}992 \quad p = 0{,}3363$

Ničelno domnevo obdržimo.

Podatki ne nakazujejo različne izobrazbene strukture žensk in moških (p=0,3363).

7.2.3.1 *2 × 2 kontingenčna tabela

Pri preizkusu homogenosti v kontingenčni tabeli dimenzije $r \times k$ lahko za vsak stolpec privzamemo t. i. multinomsko porazdelitev s *k*-kategorijami. Zanima nas, ali so pripadajoče verjetnosti $p_i, i = 1, ..., k$ pri teh porazdelitvah enake. Podatki so dobljeni tako, da so pripadajoči vzorci neodvisni.

Poseben primer multinomske porazdelitve je binomska porazdelitev, ki jo že poznamo. Če je tabela 2 × 2, imamo dve binomski porazdelitvi. Preverjanje domneve o enakih verjetnostih p_1 in p_2 že poznamo iz poglavja, ko smo primerjali verjetnosti za dve binomski porazdelitvi, pripadajoča vzorca sta bila neodvisna, pri čemer smo binomsko porazdelitev aproksimirali z normalno porazdelitvijo. Spomnimo se testne statistike *z:*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{sk} \cdot \hat{q}_{sk} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Ničelna porazdelitev je $N(0,1)$.

Matematična teorija pokaže, da je kvadrat slučajne spremenljivke, ki je porazdeljena po standardizirani normalni porazdelitvi, porazdeljen po χ^2-porazdelitvi z eno stopinjo prostosti:

$Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2(SP = 1)$

2 x 2 kontingenčno tabelo, kjer nas zanima enakost pripadajočih verjetnosti, lahko analiziramo na dva načina: ali z *z*-statistiko ali s χ^2-statistiko. Vrednost χ^2-statistike je kvadrat vrednosti z-statistike. Tudi ničelni porazdelitvi sta v tej relaciji. Poglejmo si primer.

Primer

Izveden je bil poskus kalivosti sorte A in B. Pri sorti A je od 150 semen kalilo 57 semen, pri sorti B je od 100 semen kalilo 33 semen. Pri $\alpha = 0{,}05$ preverite domnevo, da je kalivost sorte A in B enaka.

Tabela 7-24: Število semen, ki kalijo oz. ne kalijo za sorto A in B

Kalivost	Sorta A	Sorta B	Skupaj
Kali	57	33	90
Ne kali	93	67	160
Skupaj	**150**	**100**	250

Pristop 1: binomska porazdelitev

Posamezno seme kali ali ne kali. Število semen, ki kalijo v slučajnem vzorcu velikosti n, modeliramo s slučajno spremenljivko X, za katero privzamemo binomsko porazdelitev: $X \sim b(n, p)$. Za sorto A velja: $X_A \sim b(n_A, p_A)$, za sorto B pa velja: $X_B \sim b(n_B, p_B)$

H_0: $p_A = p_B$ Verjetnost, da seme kali, je pri sortah A in B enaka.

H_1: $p_A \neq p_B$ Verjetnost, da seme kali, pri sortah A in B ni enaka.

$\hat{p}_A = \frac{57}{150} = 0{,}380$ V vzorcu semen sorte A je kalilo 38 % semen.

$\hat{p}_B = \frac{33}{100} = 0{,}330$ V vzorcu semen sorte B je kalilo 33 % semen.

$\hat{p}_{sk} = \frac{90}{250} = 0{,}36 \qquad \hat{q}_{sk} = 0{,}64$

Če bi združili vzorca semen sorte A in B, bi bila kalivost 36%.

Binomski porazdelitvi lahko aproksimiramo z normalno porazdelitvijo in uporabimo z-statistiko:

$$z = \frac{0{,}380 - 0{,}330}{\sqrt{0{,}36 \cdot 0{,}64 \cdot \left(\frac{1}{150} + \frac{1}{100}\right)}} = \frac{0{,}05}{0{,}0620} = 0{,}807$$

Kritični vrednosti: $z_{0{,}025} = \pm 1{,}96$

Ničelno domnevo obdržimo. Pri stopnji značilnosti 0,05 trdimo, da eksperimentalni rezultati ne nasprotujejo domnevi o enaki kalivosti sorte A in B ($p = 0{,}4197$).

Pristop 2: χ^2-statistika

Tabela 7-25: Dejanske in pričakovane frekvence ter prispevki k χ^2-statistiki

Sorta	A	B	Skupaj
Kali	57	33	90
	54	36	90
	0,167	0,250	
Ne kali	93	67	160
	96	64	160
	0,094	0,141	
Skupaj	**150**	**100**	250
	150	100	250
	0,260	0,391	0,651

$\chi^2 = 0{,}651 \quad \chi^2_{0,05}(SP=1) = 3{,}841 \quad p = 0{,}4197$

Ničelno domnevo obdržimo. Na osnovi teh podatkov ne moremo trditi, da se kalivost sort A in B razlikuje ($p = 0{,}4197$).

Kot smo pričakovali, smo v obeh pristopih dobili isti odgovor. Primerjava rezultatov kaže naslednjo relacijo pri izračunanih vrednostih testnih statistik:

$z^2 = 0{,}807^2 = 0{,}651 = \chi^2$

in naslednjo relacijo pri pripadajočih kritičnih vrednostih:

$z^2_{0,025} = 1{,}96^2 = 3{,}841 = \chi^2_{0,05}(SP=1)$

7.2.4 *Podatki v parih

Pri preizkusu homogenosti v kontingenčni tabeli dimenzije $r \times k$ lahko za vsak stolpec privzamemo t. i. multinomsko porazdelitev s k-kategorijami. Zanima nas, ali so pripadajoče verjetnosti $p_i, i = 1,...,k$ pri teh porazdelitvah enake.

Pri testu homogenosti smo upoštevali, da so podatki dobljeni tako, da so pripadajoči vzorci neodvisni. Imamo pa tudi situacije, ko so vzorci odvisni. Poseben primer je primer dveh multinomskih porazdelitev s k kategorijami, *podatki so v parih*. Tabela podatkov je kvadratna, dimenzije $k \times k$. Postavimo najprej oznake.

Tabela 7-26: Verjetnostna shema pri analizi dveh multinomskih porazdelitev, podatki so v parih

	Y						
X	y_1	y_2	…	y_i	…	y_k	**Skupaj**
x_1	p_{11}			p_{1i}		p_{1k}	$p_{1\cdot}$
x_2	p_{21}			p_{2i}		p_{2k}	$p_{2\cdot}$
…							…
x_i	p_{i1}			p_{ii}		p_{ik}	$p_{i\cdot}$
…							…
x_k	p_{k1}			p_{ki}		p_{kk}	$p_{k\cdot}$
Skupaj	$p_{\cdot 1}$	$p_{\cdot 2}$		$p_{\cdot i}$		$p_{\cdot k}$	1

Zanima nas, ali je $p_{i\cdot} = p_{\cdot i}, i = 1, ..., k$. Zapišimo ničelno domnevo:

H_0: $p_{1\cdot} = p_{\cdot 1}, p_{2\cdot} = p_{\cdot 2}, ..., p_{k\cdot} = p_{\cdot k}$

Če ničelna domneva velja, veljajo naslednje enakosti:

$p_{11} + p_{12} + ... + p_{1k} = p_{11} + p_{21} + ... + p_{k1}$

$p_{21} + p_{22} + ... + p_{2k} = p_{12} + p_{22} + ... + p_{k2}$

…

$p_{k1} + p_{k2} + ... + p_{kk} = p_{1k} + p_{2k} + ... + p_{kk}$

Rešitev tega sistema pokaže, da lahko ničelno domnevo zapišemo krajše:

$H_0 : p_{ij} = p_{ji}, i < j$

To pomeni, da so verjetnosti v celicah, ki so simetrične na diagonalo, enake. Torej v simetrično ležečih izvendiagonalnih celicah pričakujemo enako število enot. Pričakovane frekvence izračunamo takole:

$$f_{ij}^{'} = f_{ji}^{'} = \frac{f_{ij} + f_{ji}}{2}, i < j$$

Diagonalni elementi ne posredujejo nobene informacije. Velja:

$f_{ii}^{'} = f_{ii}$

Izkaže se, da dobi ob teh pogojih χ^2-statistika naslednjo obliko:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{(f_{ij} - f_{ij}^{'})^2}{f_{ij}^{'}} = \sum_{i<j}\frac{(f_{ij} - f_{ji})^2}{f_{ij} + f_{ji}}$$

Število stopinj prostosti je enako številu celic na zgornjem trikotniku, to je $k \cdot (k-1)/2$.

Ničelna porazdelitev je torej $\chi^2(SP = \frac{k(k-1)}{2})$.

Ta preizkus se imenuje **McNemarjev preizkus simetrije**.

Primer 3 (nadaljevanje)

Zanima nas, kako se vremenska napoved ujema z dejanskim stanjem vremena pri različnih vrstah vremena (sončno, oblačno, padavine). Populacijo dni predstavlja vzorec 141 opazovanih dni. Pri stopnji značilnosti 0,05 preverimo ujemanje napovedi z dejanskim stanjem.

H_0 : Napovedi se ujemajo z dejanskim stanjem.

H_1 : Napovedi se ne ujemajo z dejanskim stanjem.

Če bi bilo ujemanje dejanskega stanja in napovedi idealno, bi bile neničelne frekvence samo na diagonali. Če pa sta ustrezni verjetnosti enaki, sta v simetričnih izvendiagonalnih celicah isti pričakovani frekvenci. Izračunajmo pričakovane frekvence:

$$f'_{12} = f'_{21} = \frac{24+4}{2} = 14$$

$$f'_{13} = f'_{31} = \frac{17+6}{2} = 11{,}5$$

$$f'_{23} = f'_{32} = \frac{3+9}{2} = 6$$

Pričakovane frekvence so ustrezno velike.

Tabela 7-27: Dejanske in pričakovane frekvence

	Napovedano vreme		
Dejansko vreme	**Sončno**	**Oblačno**	**Padavine**
Sončno	34	24	17
	34	14	11,5
Oblačno	4	21	3
	14	21	6
Padavine	6	9	23
	11,5	6	23

$$\chi^2 = \frac{(24-4)^2}{24+4} + \frac{(17-6)^2}{17+6} + \frac{(3-9)^2}{3+9} = 14{,}29 + 5{,}26 + 3{,}00 = 22{,}55$$

$\chi^2(SP=3) = 7{,}81 \qquad p = 0{,}0001$

Pri stopnji značilnosti 0,05 trdimo, da se napovedi vremena ne ujemajo z dejanskim stanjem. Največje odstopanje je pri kombinaciji 'sončno, oblačno'.

7.2.4.1 *2 x 2 kontingenčna tabela

2 × 2 kontingenčno tabelo določata dve binomski porazdelitvi. Podatki so zbrani v parih. Zanima nas, ali sta verjetnosti enaki. McNemarjev preizkus simetrije temelji na χ^2-statistiki:

$\chi^2 = \frac{(f_{12} - f_{21})^2}{f_{12} + f_{21}}$, pripadajoča ničelna porazdelitev je $\chi^2(SP=1)$.

V razdelku 'Primerjava dveh populacij' smo ta primer že obravnavali. Dobili smo *z*-statistiko:

$z = \frac{f_{12} - f_{21}}{\sqrt{f_{21} + f_{12}}}$, pripadajoča ničelna porazdelitev je $N(0,1)$.

Primerjajte pristopa.

Primer

V vzorec je bilo izbranih 1600 volilnih upravičencev. Postavili so jim vprašanje: ‘Ste zadovoljni s predsednikom?’ Dopuščena odgovora sta bila Da oz. Ne. Prvo anketo so ponovili

po enem letu na istih volilnih upravičencih. Na osnovi odgovorov so želeli oceniti, ali se je spremenilo zadovoljstvo volilnega telesa s predsednikom (stopnja značilnosti 0,05).

Tabela 7-28: Odgovori 1600 anketirancev na vprašanje: 'Ste zadovoljni s predsednikom?'

	Druga anketa		
Prva anketa	**Da**	**Ne**	**Skupaj**
Da	794	150	944
Ne	86	570	656
Skupaj	880	720	**1600**

Vzorec volilnih upravičencev, ki je bil anketiran ob prvi anketi, je isti kot ob drugi anketi. Podatki kažejo, da je 794 anketirancev ohranilo zadovoljstvo s predsednikom (Da, Da), 570 je ohranilo nezadovoljstvo s predsednikom (Ne, Ne). Svoje mnenje je spremenilo 150 anketirancev (Da, Ne) in 86 anketirancev (Ne, Da). Prav ti, ki so spremenili svoje mnenje, vplivajo na to, ali se je zadovoljstvo s predsednikom spremenilo.

H_0 : Zadovoljstvo s predsednikom je ob prvi in drugi anketi enako.

H_1 : Zadovoljstvo s predsednikom ob prvi in drugi anketi ni enako.

$$\chi^2 = \frac{(150-86)^2}{150+86} = 17{,}356$$

$$\chi^2_{0,05}(SP=1) = 3{,}841 \quad p = 0{,}0000$$

Ničelno domnevo zavrnemo v korist alternativne domneve. Pri stopnji značilnosti 0,05 trdimo, da se je zadovoljstvo s predsednikom spremenilo ($p = 0{,}0000$). Podatki nakazujejo, da je padlo.

Opomba: če bi bila vzorca anketirancev ob prvi in ob drugi anketi neodvisna, bi bili podatki predstavljeni takole:

	Ste zadovoljni s predsednikom?		
	Da	**Ne**	**Skupaj**
Prva anketa			
Druga anketa			

NALOGE

1. Ocene pri matematiki in pri statistiki

V tabeli so podatki za vzorec 528 študentov glede na njihov uspeh pri matematiki in pri statistiki. Uspeh je bil vrednoten A, B in C (A pomeni odličen uspeh, C pa slab uspeh).

Tabela 7-29: Število študentov glede na uspeh pri matematiki in pri statistiki

	Matematika		
Statistika	**A**	**B**	**C**
A	56	71	12
B	47	163	38
C	14	42	85

Na osnovi vzorca preverite domnevo o povezanosti ocen pri matematiki in pri statistiki (stopnja značilnosti 0,05).

2. Gripa

V slučajni vzorec so izbrali 120 oseb, ki so se cepili proti gripi. Od le-teh je zbolelo 35, ostali pa niso zboleli. V drugi slučajni vzorec so izbrali 220 oseb, ki se niso cepili proti gripi. Od le-teh je zbolelo 85 oseb, ostali pa niso zboleli. Na osnovi teh podatkov preverite domnevo, da je verjetnost gripe pri cepljenih in pri necepljenih osebah enaka (stopnja značilnosti 0,05).

3. C-vitamin

Na smučarjih so izvedli raziskavo o zaščitnem vplivu vitamina C. Pred začetkom smučarske sezone so 140 smučarjem dajali teden dni placebo, 139 pa vitamin C. Od tistih, ki so dobivali placebo, jih je zbolelo 31; od tistih, ki so dobivali vitamin C, jih je zbolelo 17. Pri stopnji značilnosti 0,05 preverite domnevo o zaščitnem vplivu vitamina C. Kakšni bi bili rezultati pri stopnji značilnosti 0,01?

4. Zorenje sira

Mlekarna je opravila test na koliformne bakterije v mleku na progah zbiranja A, B in C. Rezultat pregleda 133 vzorcev mleka na progi A je bil 42 pozitivnih,ostali negativni; rezultat pregleda 117 vzorcev mleka na progi B pa 29 pozitivnih, ostali negativni; rezultat pregleda 105 vzorcev na progi C je 11 pozitivnih, ostali negativni. Z ustreznim preizkusom ugotovite, ali se proge razlikujejo po intenziteti okužbe s koliformnimi bakterijami (stopnja značilnosti 0,05).

5. Predvolilna anketa

V 4 volilnih enotah so anketirali volilne upravičence pred volitvami. Iz prve volilne enote so izbrali slučajni vzorec velikosti 120, iz druge slučajni vzorec velikosti 115, iz tretje velikosti 145 in iz četrte velikosti 120. Postavili so jim vprašanje: Kako bi se odločili, če bi bile volitve v nedeljo? Možni odgovori so bili: Kandidat A, Kandidat B, Ne vem. Rezultati ankete so v tabeli:

Tabela 7-30: Število anketirancev glede na volilno enoto (I, II, III, IV) in glede na odgovor

Odgovor	I	II	III	IV
Kandidat A	30	35	45	20
Kandidat B	30	40	50	60
Ne vem	60	40	50	40
Skupaj	**120**	**115**	**145**	**120**

Na osnovi teh podatkov preverite domnevo, da je struktura odgovorov po volilnih enotah enaka (stopnja značilnosti 0,05). Komentirajte rezultate.

6. *Izobrazba ženina in neveste

V letu 1992 je bilo v Republiki Sloveniji sklenjenih 9119 zakonskih zvez. V tabeli so predstavljeni pari po šolski izobrazbi ženina in neveste.

Tabela 7-31: Sklenjene zakonske zveze v letu 1992 po šolski izobrazbi ženina in neveste. (Vir: SL-93, str. 66)

	Ženin			
Nevesta	**Osnovna ali manj**	**Srednja**	**Višja ali več**	**Neznano**
Osnovna ali manj	753	1042	32	25
Srednja	692	4343	462	35
Višja ali več	38	701	622	3
Neznano	14	14	5	338

Privzemimo, da podatki za leto 1992 predstavljajo ustrezen vzorec, na osnovi katerega bomo sklepali na celotno populacijo zakonskih zvez v Sloveniji. Iz podatkov izločite pare, pri katerih je izobrazba vsaj enega partnerja neznana. Ali je izobrazbena struktura ženinov in nevest enaka (stopnja značilnosti 0,05)? Obrazložite rezultate.

8 PRILOGA: STATISTIČNE TABELE

8.1 SLUČAJNE ŠTEVKE

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62

8.2 STANDARDIZIRANA NORMALNA PORAZDELITEV

V tabeli je za vrednost standardiziranega odklona z navedena verjetnost $p = P(Z \geq z)$.

Primer: za $z = 0{,}93$ odčitamo $p = 0{,}1762$.

z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
4,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Dodatek: z = 3,5 p = 0,000233 z = 4,0 p = 0,0000317 z = 5,0 p = 0,000000287

8.3 STUDENTOVA PORAZDELITEV

V tabeli je za verjetnost α in za stopinje prostosti SP navedena vrednost t_α, za katero velja $P(T \geq t_\alpha) = \alpha$.

Primer: za $\alpha = 0{,}025$ in SP = 10 odčitamo: $t_\alpha = 2{,}228$

	α						
SP	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**	**0,0005**
1	3,078	6,314	12,706	31,821	63,656	318,289	636,578
2	1,886	2,920	4,303	6,965	9,925	22,328	31,600
3	1,638	2,353	3,182	4,541	5,841	10,214	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,689
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,660
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

8.4 χ^2-PORAZDELITEV

V tabeli je za verjetnost α in za stopinje prostosti SP navedena vrednost χ^2_α, za katero velja $P(\chi^2 \geq \chi^2_\alpha) = \alpha$. Primer: za $\alpha = 0{,}05$ in SP = 1 odčitamo $\chi^2_\alpha = 3{,}841$

	α							
SP	**0,995**	**0,99**	**0,975**	**0,95**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,000	0,000	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
60	35,534	37,485	40,482	43,188	79,082	83,298	88,379	91,952
80	51,172	53,540	57,153	60,391	101,879	106,629	112,329	116,321
100	67,328	70,065	74,222	77,929	124,342	129,561	135,807	140,170

8.5 F-PORAZDELITEV, $\alpha = 0{,}05$

V tabeli je za stopinje prostosti SP_1 in SP_2 navedena vrednost F_α, za katero velja $P(F \geq F_\alpha) = \alpha = 0{,}05$.

Primer: za $\alpha = 0{,}05$ in $SP_1 = 2$ ter $SP_2 = 10$ odčitamo $F_\alpha = 4{,}103$

	SP_1												
SP_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**	**20**	**25**	**∞**
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,40	19,41	19,45	19,46	19,50
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,785	8,745	8,660	8,634	8,527
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,964	5,912	5,803	5,769	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,735	4,678	4,558	4,521	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,060	4,000	3,874	3,835	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,637	3,575	3,445	3,404	3,230
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,347	3,284	3,150	3,108	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,137	3,073	2,936	2,893	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	2,978	2,913	2,774	2,730	2,538
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,854	2,788	2,646	2,601	2,405
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,753	2,687	2,544	2,498	2,297
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,671	2,604	2,459	2,412	2,207
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,602	2,534	2,388	2,341	2,131
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,544	2,475	2,328	2,280	2,066
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,494	2,425	2,276	2,227	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,450	2,381	2,230	2,181	1,961
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,412	2,342	2,191	2,141	1,917
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,378	2,308	2,155	2,106	1,879
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,348	2,278	2,124	2,074	1,844
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,321	2,250	2,096	2,045	1,812
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,297	2,226	2,071	2,020	1,784
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,275	2,204	2,048	1,996	1,758
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,255	2,183	2,027	1,975	1,734
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,236	2,165	2,007	1,955	1,712
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,220	2,148	1,990	1,938	1,691
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,204	2,132	1,974	1,921	1,672
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,190	2,118	1,959	1,906	1,655
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,177	2,104	1,945	1,891	1,638
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,165	2,092	1,932	1,878	1,623
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,077	2,003	1,839	1,783	1,510
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	1,993	1,917	1,748	1,690	1,390
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,910	1,834	1,659	1,598	1,255
∞	3,842	2,996	2,605	2,372	2,214	2,099	2,010	1,939	1,831	1,752	1,571	1,506	1,025

8.6 F-PORAZDELITEV, $\alpha = 0{,}025$

V tabeli je za stopinje prostosti SP_1 in SP_2 navedena vrednost F_α, za katero velja $P(F \geq F_\alpha) = \alpha = 0{,}025$.

Primer: za $\alpha = 0{,}025$ in $SP_1 = 2$ ter $SP_2 = 10$ odčitamo $F_\alpha = 5{,}456$.

	SP_1												
SP_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**	**20**	**25**	**∞**
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,40	39,41	39,43	39,46	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,42	14,34	14,25	14,11	13,90
4	12,22	10,65	9,979	9,604	9,364	9,197	9,074	8,980	8,844	8,751	8,657	8,501	8,258
5	10,01	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,619	6,525	6,428	6,268	6,016
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,461	5,366	5,269	5,107	4,850
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,761	4,666	4,568	4,405	4,143
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,295	4,200	4,101	3,937	3,671
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	3,964	3,868	3,769	3,604	3,334
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,717	3,621	3,522	3,355	3,081
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,526	3,430	3,330	3,162	2,884
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,374	3,277	3,177	3,008	2,726
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,250	3,153	3,053	2,882	2,596
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,147	3,050	2,949	2,778	2,488
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,060	2,963	2,862	2,689	2,396
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	2,986	2,889	2,788	2,614	2,317
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,922	2,825	2,723	2,548	2,248
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,866	2,769	2,667	2,491	2,188
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,817	2,720	2,617	2,441	2,134
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,774	2,676	2,573	2,396	2,086
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,735	2,637	2,534	2,356	2,043
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,700	2,602	2,498	2,320	2,004
23	5,750	4,349	3,750	3,408	3,183	3,023	2,902	2,808	2,668	2,570	2,466	2,287	1,969
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,640	2,541	2,437	2,257	1,936
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,613	2,515	2,411	2,230	1,906
26	5,659	4,265	3,670	3,329	3,105	2,945	2,824	2,729	2,590	2,491	2,387	2,205	1,879
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,568	2,469	2,364	2,183	1,854
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,547	2,448	2,344	2,161	1,830
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,529	2,430	2,325	2,142	1,808
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,511	2,412	2,307	2,124	1,788
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,388	2,288	2,182	1,994	1,638
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,270	2,169	2,061	1,869	1,483
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,157	2,055	1,945	1,746	1,312
∞	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,048	1,945	1,833	1,626	1,029

8.7 F-PORAZDELITEV, $\alpha = 0{,}01$

V tabeli je za stopinje prostosti SP_1 in SP_2 navedena vrednost F_α, za katero velja $P(F \geq F_\alpha) = \alpha = 0{,}01$.

Primer: za $\alpha = 0{,}01$ in $SP_1 = 2$ ter $SP_2 = 10$ odčitamo $F_\alpha = 7{,}559$.

	SP_1												
SP_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**	**20**	**25**	**∞**
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,40	99,42	99,45	99,46	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,23	27,05	26,69	26,58	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,55	14,37	14,02	13,91	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,05	9,888	9,553	9,449	9,022
6	13,75	10,93	9,780	9,148	8,746	8,466	8,260	8,102	7,874	7,718	7,396	7,296	6,881
7	12,25	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,620	6,469	6,155	6,058	5,651
8	11,26	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,814	5,667	5,359	5,263	4,860
9	10,56	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,257	5,111	4,808	4,713	4,312
10	10,04	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,849	4,706	4,405	4,311	3,910
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,539	4,397	4,099	4,005	3,604
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,296	4,155	3,858	3,765	3,362
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,100	3,960	3,665	3,571	3,166
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	3,939	3,800	3,505	3,412	3,005
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,805	3,666	3,372	3,278	2,870
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,691	3,553	3,259	3,165	2,754
17	8,400	6,112	5,185	4,669	4,336	4,101	3,927	3,791	3,593	3,455	3,162	3,068	2,654
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,508	3,371	3,077	2,983	2,567
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,434	3,297	3,003	2,909	2,490
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,368	3,231	2,938	2,843	2,422
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,310	3,173	2,880	2,785	2,361
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453	3,258	3,121	2,827	2,733	2,307
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,211	3,074	2,780	2,686	2,257
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,168	3,032	2,738	2,643	2,212
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,129	2,993	2,699	2,604	2,171
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,094	2,958	2,664	2,569	2,133
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,062	2,926	2,632	2,536	2,098
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,032	2,896	2,602	2,506	2,065
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198	3,005	2,868	2,574	2,478	2,035
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	2,979	2,843	2,549	2,453	2,008
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,801	2,665	2,369	2,271	1,806
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,632	2,496	2,198	2,098	1,602
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,472	2,336	2,035	1,932	1,383
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,640	2,511	2,321	2,185	1,878	1,773	1,035

8.8 F-PORAZDELITEV, $\alpha = 0{,}001$

V tabeli je za stopinje prostosti SP_1 in SP_2 navedena vrednost F_α, za katero velja $P(F \geq F_\alpha) = \alpha = 0{,}01$.

Primer: za $\alpha = 0{,}01$ in $SP_1 = 2$ ter $SP_2 = 10$ odčitamo $F_\alpha = 14{,}90$.

	SP_1												
SP_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**	**20**	**25**	**∞**
2	998,4	998,8	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,5	999,5
3	167,1	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,2	128,3	126,4	125,8	123,5
4	74,13	61,25	56,17	53,43	51,72	50,52	49,65	49,00	48,05	47,41	46,10	45,70	44,05
5	47,18	37,12	33,20	31,08	29,75	28,83	28,17	27,65	26,91	26,42	25,39	25,08	23,79
6	35,51	27,00	23,71	21,92	20,80	20,03	19,46	19,03	18,41	17,99	17,12	16,85	15,75
7	29,25	21,69	18,77	17,20	16,21	15,52	15,02	14,63	14,08	13,71	12,93	12,69	11,70
8	25,41	18,49	15,83	14,39	13,48	12,86	12,40	12,05	11,54	11,19	10,48	10,26	9,336
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	9,894	9,570	8,898	8,689	7,815
10	21,04	14,90	12,55	11,28	10,48	9,93	9,517	9,204	8,754	8,446	7,803	7,604	6,765
11	19,69	13,81	11,56	10,35	9,579	9,047	8,655	8,355	7,923	7,625	7,008	6,815	6,000
12	18,64	12,97	10,80	9,633	8,892	8,378	8,001	7,711	7,292	7,005	6,405	6,217	5,422
13	17,82	12,31	10,21	9,073	8,355	7,856	7,489	7,206	6,799	6,519	5,934	5,751	4,969
14	17,14	11,78	9,730	8,622	7,922	7,436	7,078	6,802	6,404	6,130	5,557	5,377	4,606
15	16,59	11,34	9,335	8,253	7,567	7,091	6,741	6,471	6,081	5,812	5,249	5,071	4,309
16	16,12	10,97	9,006	7,944	7,272	6,805	6,460	6,195	5,812	5,547	4,992	4,817	4,061
17	15,72	10,66	8,727	7,683	7,022	6,562	6,224	5,962	5,584	5,324	4,775	4,602	3,852
18	15,38	10,39	8,487	7,460	6,808	6,355	6,021	5,763	5,390	5,132	4,590	4,418	3,672
19	15,08	10,16	8,280	7,265	6,622	6,175	5,845	5,591	5,222	4,967	4,430	4,259	3,516
20	14,82	9,953	8,098	7,096	6,461	6,019	5,692	5,440	5,075	4,823	4,290	4,121	3,380
21	14,59	9,773	7,938	6,947	6,318	5,881	5,557	5,308	4,946	4,696	4,167	3,999	3,260
22	14,38	9,612	7,796	6,814	6,191	5,758	5,437	5,190	4,832	4,583	4,058	3,891	3,153
23	14,20	9,469	7,669	6,696	6,078	5,649	5,331	5,085	4,730	4,483	3,961	3,794	3,057
24	14,03	9,340	7,554	6,589	5,977	5,551	5,235	4,991	4,638	4,393	3,873	3,707	2,971
25	13,88	9,222	7,451	6,493	5,885	5,462	5,148	4,906	4,555	4,311	3,794	3,629	2,892
26	13,74	9,117	7,357	6,406	5,802	5,381	5,070	4,829	4,480	4,238	3,723	3,558	2,821
27	13,61	9,019	7,271	6,326	5,726	5,308	4,998	4,759	4,412	4,170	3,658	3,493	2,756
28	13,50	8,930	7,193	6,253	5,657	5,241	4,933	4,695	4,349	4,109	3,598	3,434	2,697
29	13,39	8,848	7,121	6,186	5,592	5,179	4,873	4,636	4,292	4,053	3,543	3,380	2,642
30	13,29	8,773	7,054	6,125	5,534	5,122	4,817	4,582	4,239	4,001	3,493	3,330	2,591
40	12,61	8,251	6,595	5,698	5,128	4,731	4,436	4,207	3,874	3,643	3,145	2,984	2,235
60	11,97	7,768	6,171	5,307	4,757	4,372	4,086	3,865	3,542	3,315	2,826	2,667	1,893
120	11,38	7,321	5,781	4,947	4,416	4,044	3,767	3,552	3,237	3,016	2,534	2,375	1,546
∞	10,83	6,909	5,422	4,617	4,103	3,743	3,475	3,266	2,959	2,743	2,266	2,105	1,047

8.9 PEARSONOV KOEFICIENT KORELACIJE, α = 0,05 IN 0,01

V tabeli je za število enot v vzorcu *n* in za verjetnost α navedena kritična absolutna vrednost ocene Pearsonovega koeficienta korelacije *r*, pri kateri zavrnemo ničelno domnevo H_0: $\rho = 0$ pri dvostranskem preizkusu.

Primer: za α = 0,05 in *n* = 8 odčitamo *r* = 0,7067 oziroma *r* = -0,7067.

	α			α			α	
n	**0,05**	**0,01**	**n**	**0,05**	**0,01**	**n**	**0,05**	**0,01**
1	-	-	31	0,3550	0,4556	61	0,2521	0,3274
2	-	-	32	0,3494	0,4487	62	0,2500	0,3248
3	0,9969	0,9999	33	0,3440	0,4421	63	0,2480	0,3223
4	0,9500	0,9900	34	0,3388	0,4357	64	0,2461	0,3198
5	0,8783	0,9587	35	0,3338	0,4296	65	0,2441	0,3173
6	0,8114	0,9172	36	0,3291	0,4238	66	0,2423	0,3150
7	0,7545	0,8745	37	0,3246	0,4182	67	0,2404	0,3126
8	0,7067	0,8343	38	0,3202	0,4128	68	0,2387	0,3104
9	0,6664	0,7977	39	0,3160	0,4076	69	0,2369	0,3081
10	0,6319	0,7646	40	0,3120	0,4026	70	0,2352	0,3060
11	0,6021	0,7348	41	0,3081	0,3978	71	0,2335	0,3038
12	0,5760	0,7079	42	0,3044	0,3932	72	0,2319	0,3017
13	0,5529	0,6835	43	0,3008	0,3887	73	0,2303	0,2997
14	0,5324	0,6614	44	0,2973	0,3843	74	0,2287	0,2977
15	0,5140	0,6411	45	0,2940	0,3801	75	0,2272	0,2957
16	0,4973	0,6226	46	0,2907	0,3761	76	0,2257	0,2938
17	0,4821	0,6055	47	0,2876	0,3721	77	0,2242	0,2919
18	0,4683	0,5897	48	0,2845	0,3683	78	0,2227	0,2900
19	0,4555	0,5751	49	0,2816	0,3646	79	0,2213	0,2882
20	0,4438	0,5614	50	0,2787	0,3610	80	0,2199	0,2864
21	0,4329	0,5487	51	0,2759	0,3575	82	0,2172	0,2830
22	0,4227	0,5368	52	0,2732	0,3542	84	0,2146	0,2796
23	0,4132	0,5256	53	0,2706	0,3509	86	0,2120	0,2764
24	0,4044	0,5151	54	0,2681	0,3477	88	0,2096	0,2732
25	0,3961	0,5052	55	0,2656	0,3445	90	0,2072	0,2702
26	0,3882	0,4958	56	0,2632	0,3415	92	0,2050	0,2673
27	0,3809	0,4869	57	0,2609	0,3385	94	0,2028	0,2645
28	0,3739	0,4785	58	0,2586	0,3357	96	0,2006	0,2617
29	0,3673	0,4705	59	0,2564	0,3328	98	0,1986	0,2591
30	0,3610	0,4629	60	0,2542	0,3301	100	0,1966	0,2565

8.10 SPEARMANOV KOEFICIENT KORELACIJE, α = 0,05 IN 0,01

V tabeli je za število enot v vzorcu n in za verjetnost α dana kritična absolutna vrednost ocene Spearmanovega koeficienta korelacije r_s, pri kateri zavrnemo ničelno domnevo H_0: $\rho_s = 0$ pri dvostranskem preizkusu.

Primer: za α = 0,05 in n = 8 odčitamo r_s = 0,7381 oziroma r_s = -0,7381.

	α			α			α	
n	**0,05**	**0,01**	**n**	**0,05**	**0,01**	**n**	**0,05**	**0,01**
1	-	-	31	0,3560	0,4593	61	0,2524	0,3287
2	-	-	32	0,3504	0,4523	62	0,2503	0,3260
3	-	-	33	0,3449	0,4455	63	0,2483	0,3234
4	-	-	34	0,3396	0,4390	64	0,2463	0,3209
5	1,0000	-	35	0,3347	0,4328	65	0,2444	0,3185
6	0,8857	1,0000	36	0,3300	0,4268	66	0,2425	0,3161
7	0,7857	0,9286	37	0,3253	0,4211	67	0,2407	0,3137
8	0,7381	0,8810	38	0,3209	0,4155	68	0,2389	0,3114
9	0,7000	0,8333	39	0,3168	0,4103	69	0,2372	0,3092
10	0,6485	0,7939	40	0,3128	0,4051	70	0,2354	0,3070
11	0,6182	0,7545	41	0,3087	0,4002	71	0,2337	0,3048
12	0,5874	0,7273	42	0,3051	0,3955	72	0,2321	0,3027
13	0,5604	0,7033	43	0,3014	0,3908	73	0,2305	0,3006
14	0,5385	0,6791	44	0,2978	0,3865	74	0,2289	0,2986
15	0,5214	0,6536	45	0,2945	0,3822	75	0,2274	0,2966
16	0,5029	0,6353	46	0,2913	0,3781	76	0,2259	0,2947
17	0,4877	0,6176	47	0,2880	0,3741	77	0,2244	0,2928
18	0,4716	0,5996	48	0,2850	0,3702	78	0,2229	0,2909
19	0,4596	0,5842	49	0,2820	0,3664	79	0,2215	0,2891
20	0,4466	0,5699	50	0,2791	0,3628	80	0,2201	0,2872
21	0,4364	0,5558	51	0,2764	0,3592	82	0,2174	0,2837
22	0,4252	0,5438	52	0,2736	0,3558	84	0,2147	0,2804
23	0,4160	0,5316	53	0,2710	0,3524	86	0,2122	0,2771
24	0,4070	0,5209	54	0,2685	0,3492	88	0,2097	0,2740
25	0,3977	0,5108	55	0,2659	0,3460	90	0,2074	0,2709
26	0,3901	0,5009	56	0,2636	0,3429	92	0,2051	0,2680
27	0,3828	0,4915	57	0,2612	0,3400	94	0,2029	0,2651
28	0,3755	0,4828	58	0,2589	0,3370	96	0,2008	0,2623
29	0,3685	0,4749	59	0,2567	0,3342	98	0,1987	0,2597
30	0,3624	0,4670	60	0,2545	0,3314	100	0,1967	0,2571

9 REŠITVE NALOG

9.1 Uvod

OSNOVNI POJMI

1. Bruci

a) stvarna opredelitev: vpisani v prvi letnik visokošolskega študija;
časovna opredelitev: šolsko leto 1997/98;
krajevna opredelitev: Univerza v Ljubljani.

g) Enota populacije je študent, ki je vpisan v prvi letnik visokošolskega študija Univerze v Ljubljani v šolskem letu 1997/98.

c) do e)

Spremenljivka	Vrednost	Vrsta		Merska lestvica
Priimek	Novak	opisna		imenska
Spol	moški	opisna		imenska
Poštna št. kraja	4260	opisna		imenska
Izobrazba očeta	srednješolska	opisna		urejenostna
Izobrazba matere	srednješolska	opisna		urejenostna
Število otrok v družini	2	številska	diskretna	razmernostna
Oddaljenost od fakultete	55 km	številska	zvezna	razmernostna
Štipendist	ne	opisna		imenska
Višina štipendije	-	številska	zvezna	razmernostna

f) 'Spol' in 'štipendist' sta dvojiški spremenljivki.

3. Populacija vseh vzorcev

a) 9 vzorcev			b) 6 vzorcev			c) 6 vzorcev			d) 3 vzorci		
AA	BA	CA	- -	BA	CA	AA	- -	- -	- -	- -	- -
AB	BB	CB	AB	- -	CB	AB	BB	- -	AB	- -	- -
AC	BC	CC	AC	BC	- -	AC	BC	CC	AC	BC	- -

4. Gornji Dol

5 461 512

NAČINI PROUČEVANJA MNOŽIČNIH POJAVOV

1. Vaje

a) Beremo po dve slučajni števki skupaj. Upoštevamo števila od 1 do 49, ostale izpustimo.

b) Korak je 10. Prvi je izbran slučajno med študenti z oznako od 1 do 10.

3. Cvetličarne

a) Osebe dosegljive po telefonu

b) Okvir vzorčenja

c) Korak je 6

4. Gledanje TV

a) Stratume.

b) Stratificirano vzorčenje.

c) Kvotno vzorčenje.

d) Ne.

5. Travnik

Koordinatno izhodišče postavimo na rob parcel, ki so primerne za vzorčenje: x=1 do 58, y=1 do 38. Za vsako parcelo izberemo po dve dvomestni slučajni števki. Za x upoštevamo števila od 1 do 58, za y pa od 1 do 38.

9.2 Opisna statistika

RELATIVNA ŠTEVILA

1. Zemljišča v Sloveniji

Skupna površina zemljišč se deli na nerodovitna zemljišča, na gozdna zemljišča ter na kmetijska zemljišča. Kmetijska zemljišča se delijo na travinje, to so travniki in pašniki, na njive in vrtove, in na sadovnjake in vinograde. V tabeli so podatki za Slovenijo za leto 1990.

Tabela 9-1: Površina zemljišč in površina kmetijskih zemljišč, struktura zemljišč ter struktura kmetijskih zemljišč v Sloveniji leta 1990

Vrsta zemljišč	Površina zemljišč (ha)	Površina kmet. zemljišč (ha)	Struktura zemljišč (%)		Struktura kmetijskih zemljišč (%)
Nerodovitna	136754		6,8		
Gozdna	1024535		50,6		
Kmetijska	864184		42,7		
-travinje		559264		27,6	64,7
-njive, vrtovi		247083		12,2	28,6
-sadov., vinog.		57837		2,9	6,7
Skupaj	2025473	864184	100,0	42,7	100,0

3. Zasedenost ležišč

Tabela 9-2: Zasedenost ležišč po vrstah krajev v Sloveniji leta 1993

Vrste krajev	Zasedenost (%)
Ljubljana	26,9
Zdraviliški kraji	41,2
Obmorski kraji	18,2
Gorski kraji	12,6
Drugi turistični kraji	15,0
Drugi kraji	8,7
Slovenija	19,7

6. Prometne nesreče

a)

Tabela 9-3: Število prometnih nesreč s smrtnim izidom ter izračunani indeksi z osnovo 1988, verižni indeksi ter stopnje rasti

Leto	Prometne nesreče	1988=100	Verižni indeks	St. rasti
1988	6085	100,0	-	-
1989	5825	95,7	95,7	-4,3
1990	5177	85,1	88,9	-11,1
1991	5479	90,0	105,8	5,8
1992	5882	96,7	107,4	7,4
1993	6290	103,4	106,9	6,9
1994	6586	108,2	104,7	4,7
1995	6540	107,5	99,3	-0,7
1996	6348	104,3	97,1	-2,9
1997	6951	114,2	109,5	9,5
1998	5864	96,4	84,4	-15,7

FREKVENČNA PORAZDELITEV

1. Razred

a) Urejenostna.

b)

Tabela 9-4: Število učencev glede na izobrazbo očeta in glede na izobrazbo matere

Izobrazba	Oče	Mati
Osnovnošolska	5	7
Srednješolska	15	13
Višješolska in več	3	4
Neznano	1	0
Skupaj	24	24

3. Količina padavin

Tabela 9-5: Meteorološke postaje po količini padavin

Količina padavin (mm)	f_i	f_i %	F_i	F_i %	$x_{i,min}$	$x_{i,max}$	x_i	d_i
800 do pod 1200	12	17,9	12	17,9	800	1200	1000	400
1200 do pod 1600	27	40,3	39	58,2	1200	1600	1400	400
1600 do pod 2000	16	23,9	55	82,1	1600	2000	1800	400
2000 do pod 2400	7	10,4	62	92,5	2000	2400	2200	400
2400 do pod 2800	4	6,0	66	98,5	2400	2800	2600	400
2800 do pod 3200	0	0,0	66	98,5	2800	3200	3000	400
3200 do pod 3600	1	1,5	67	100,0	3200	3600	3400	400
Skupaj	67	100,0						

4. Delovna doba brezposelnih

b)

Tabela 9-6: Brezposelni z delovno dobo

Delovna doba (leta)	$x_{i,min}$	$x_{i,max}$	x_i	d_i	g_i 1990	g_i 1992
do 1 leta	0	1	0,5	1	21035	16922
nad 1 do 2	1	2	1,5	1	3563	6066
nad 2 do 3	2	3	2,5	1	2654	4533
nad 3 do 5	3	5	4,0	2	2094	3835
nad 5 do 10	5	10	7,5	5	1438	3304
nad 10 do 20	10	20	15,0	10	869	2397
nad 20 do 30	20	30	25,0	10	330	1391
nad 30	30	-	-	-	-	-

KVANTILI

1. Stoli

a) $Q_1 = 4089$ SIT, $Q_2 = 6281$ SIT, $Q_3 = 11116$ SIT

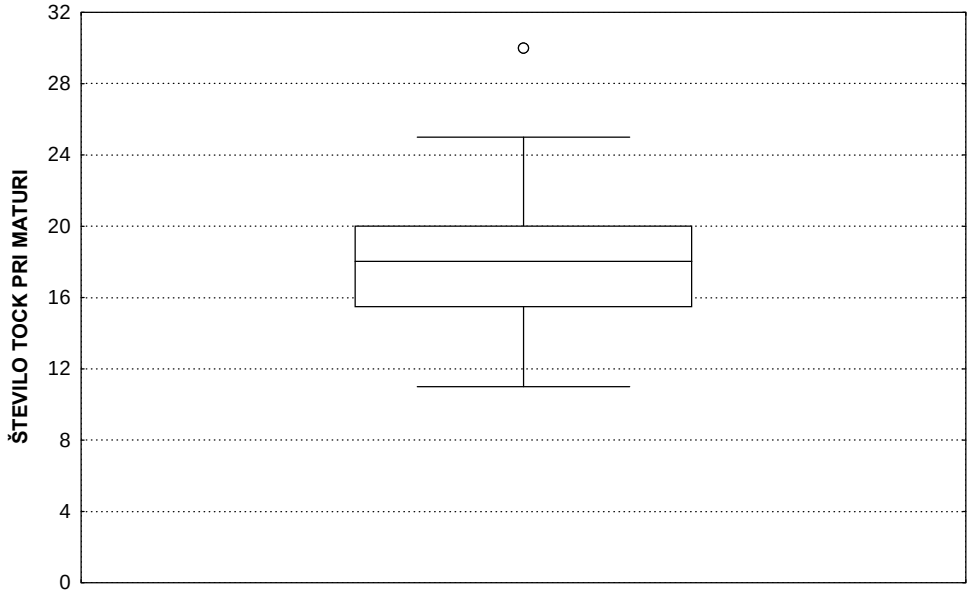
b) Ne.

c) Okvir z ročaji določajo vrednosti: 2613 SIT, 4089 SIT, 6281 SIT, 11116 SIT, 19580 SIT.

d) 70,3. Vrednost 10000 SIT je 70-ti centil.

2. Število točk pri maturi

Okvir z ročaji določajo vrednosti: 11, 15,5, 18, 20, 25. V podatkih sta dva osamelca, njuna vrednost je 30.



Slika 9-1: Okvir z ročaji za število točk pri maturi

MERE SREDINE: SREDNJE VREDNOSTI

1. Katere srednje vrednosti?

Spremenljivka	Merska lestvica	Srednje vrednosti
Spol	Imenska	Mo
Datum rojstva	Razmična	Mo, Me
Temperatura zraka (^{0}C)	Razmična	Mo, Me, povprečje
Plača (SIT)	Razmernostna	Mo, Me, povprečje
Dolžina delovne dobe (število dni)	Razmernostna	Mo, Me, povprečje
Gostota (kg/m^3)	razmernostna	Mo, Me, povprečje

2. Tri sredine

Vse tri sredine so enake, kadar so vsi podatki enaki.

3. Število točk pri maturi

Aritmetična sredina = 18,4, mediana = 18, modus = 18

5. Vpis študentov v Sloveniji

6,3%

6. Poroke

b) –1,2%

7. Rast bakterij

Povprečna dnevna stopnja rasti bakterij je bila 58,7%.

MERE VARIABILNOSTI

1. Število točk pri maturi

VR = 19, Q = 4,5 s = 4,86 KV% = 26,4

2. Meteorološke postaje

c) Povprečje je 1609, modus je 1430,8, mediana 1525,9

d) Standardni odklon je 489,195 KV % = 30,4

3. Poraba bencina

KV (marec) = 46,7% KV (september) = 41,7%

4. Smreke

e) 24,3

f) 27,2%

9.3 Osnove verjetnostnega računa

NORMALNA PORAZDELITEV

1. Ocene verjetnosti

$P(X > 35) = 0{,}5$, $P(X \geq 35) = 0{,}5$, $P(30 < X < 40) \approx 2/3$, $P(X > 40) \approx 1/6$, $P(X < 25) \approx 0{,}025$, $P(20 < X < 40) \approx 5/6$

2. Standardizirana normalna porazdelitev

$P(Z > 2{,}5) = 0{,}0062$

$P(Z \geq 2{,}5) = 0{,}0062$

$P(Z \leq 2{,}5) = 1 - 0{,}0062 = 0{,}9938$

$P(Z > 2{,}25) = 0{,}0122$

$P(Z > 2{,}256) = 0{,}0120$

$P(Z < -3{,}0) = 0{,}0013$

$P(Z > -1{,}5) = 1 - 0{,}0668 = 0{,}9332$

$P(-1{,}5 < Z < 1) = 1 - 0{,}1587 - 0{,}0668 = 0{,}7745$

3. Inteligenčni kvocient

a) 0,0228
b) 0,0000
c) 0,9544
d) 0,2525
e) 0,4711
f) približno 800

4. Stroj

0,6915

5. Starost ob prvi zaposlitvi

a) 0,0668
b) 0,9772
c) 0,8413
d) 0,6247

BINOMSKA PORAZDELITEV

1. Družina

X je število deklic v družini s 6 otroki; $X \sim b\,(6; 0{,}5)$

c) $P(X = 3) = 5/16$

d) $P(X = 0) + P(X = 1) + P(X = 2) = 0{,}5 \cdot (1 - P(X = 3)) = 11/32$

2. Škart

X je število slabih izdelkov v pošiljki s 100 izdelki.

a) $X \sim b\,(100; 0{,}2)$

b) 2

c) Laplace: 0,0000

d) Ne.

3. Francoska ruleta

a) p = 18/37; q = 19/37:

b) -1/37

c) -6,8 SIT

4. Drevesnica

a) 10%

b) Laplace: 0,0418

PORAZDELITEV VZORČNIH STATISTIK

1. Porazdelitev vzorčnih aritmetičnih sredin

d) Standardizirana normalna porazdelitev N(0,1). $p(x) = \frac{1}{2\pi} e^{-\frac{1}{2}x^2}$

e) N(0, 1/4)
f) N(0, 1/6)

2. Maslo

a) 78,81%
b) 73,57%

3. Avtomat

a) 6,68%
b) 0,00%
c) aritmetična sredina: 100,25 g, standardni odklon: 2,11 g

4. Mlekarna

Znižana cena: 15,87%
Normalna cena: 47,19%
Zvišana cena: 36,94%

5. Pesticid

a) 0,38%
b) 0,04%

6. Volitve

0,0010

7. Križanja

a) 0,4086
b) 0,2327

9.4 Osnove statističnega sklepanja

OCENJEVANJE PARAMETROV

1. Uspeh pri izpitih

95% interval zaupanja za povprečje: (16,5; 20,3), 95% interval zaupanja za standardni odklon: (3,8; 6,6).

2. Količina padavin

95% interval zaupanja za povprečje: (1489,6; 1728,3), 95% interval zaupanja za standardni odklon: (418,1; 589,6).

3. Prirast telet

a) (26,1; 35,1)
b) 102,6 kg
c) 100

4. Računalnik

a) 52
b) Ne, odvisen je od zaupanja.
d) Vzorca številka 10 in 25 ne vsebujeta povprečne vrednosti 100. Vzorci številka 9, 47, 54 ne vsebujejo standardnega odklona 10.

5. Kroglice za ležaje

385

6. Gripa

90% interval zaupanja: (0,0964; 0,204), 95% interval zaupanja: (0,086; 0,214)

7. Volitve

a) 9604

b) 385

c) 96

PREIZKUŠANJE STATISTIČNIH DOMNEV

2. Avtomat

$t = 0{,}474; \mp t_{0{,}025}(15) = \mp 2{,}131$, $p = 0{,}6422$.

3. Starost gledalcev

Za sredino zadnjega razreda smo privzeli starost 55 let.

$\bar{x} = 38{,}0 \quad s^2 = 61{,}4094, z = 4{,}689, \mp z_{0{,}025} = \mp 1{,}96$

Pri stopnji značilnosti 0,05 trdimo, da povprečna starost v populaciji gledalcev določene oddaje ni 35 let (p < 0,0000).

4. Mežiška dolina

a) (13,64; 22,44)

b) (18,25; 29,62)

5. Križanje graha

z=-0,779 $\mp z_{0{,}025} = \mp 1{,}96$ p=0,4360

6. Reklama cigaret

(60,6%; 69,6%)

9.5 Primerjava dveh populacij

DVA NEODVISNA VZORCA - RAZLIKA POVPREČIJ

1. Pridelek paradižnika

$t = 1{,}961 \quad \mp t_{0{,}025}(10) = \mp 2{,}228 \quad 0{,}05 < p < 0{,}10$

2. Krma

t=1,37 $\mp t_{0{,}025}(14) = \mp 2{,}145$ 0,10<p<0,20

3. Ocene na izpitu

t=1,82 $\mp t_{0{,}025}(70) \approx \mp 1{,}99$ 0,05<p<0,10

4. Starost študentov

(0,7; 9,9)

DVA NEODVISNA VZORCA - RAZLIKA BERNOULLIJEVIH VERJETNOSTI

1. Obstojnost paštete

z=1,204 $\mp z_{0{,}025} = \mp 1{,}96$

2. Učinkovitost zdravil

z=-1,491 $\mp z_{0{,}005} = \mp 2{,}576$

3. Cepljenje proti gripi

b) (8,6%; 30,1%)

4. Bakterije v mleku

(2,5%; 16,1%)

DVA ODVISNA VZORCA - RAZLIKA POVPREČIJ

1. Koncentracija škroba v krompirju

t=2,77 $\mp t_{0,025}(14) = \mp 2{,}145$ $0{,}01 < p < 0{,}02$ p=0,015

2. Nabrek ivernih plošč

t=2,01 $\mp t_{0,025}(9) = \mp 2{,}26$ $0{,}05 < p < 0{,}10$, $p = 0{,}0756$

3. Plavanje

(38,0 m, 77,3 m)

4. Prirast debla

(55,6 mm, 68,4 mm)

DVA ODVISNA VZORCA - RAZLIKA BERNOULLIJEVIH VERJETNOSTI

1. Vremenska napoved

z=-2,828 p=0,0046

2. Vpliv kajenja

(73,5%, 90,6%)

9.6 Regresija in korelacija

ENOSTAVNA LINEARNA REGRESIJA

1. Vzdrževanje osebnih vozil

b= 70,918 a= 208,20 $r^2 = 0{,}879$

Napoved za 5 let staro vozilo: 562,79 DE

Napoved za 25 let staro vozilo ni smiselna.

2. Smrtno ponesrečeni

a=458,6 b=-22,35

Napoved za leto 1999: 325

3. Črvivost jabolk

b) a=64,247 b=-1,013

c) Interval zaupanja za α: (56,22; 72,72), interval zaupanja za β: (-1,397; -0,6294)

d) 77,6%

e) Napoved za drevo, ki bi imelo 2000 plodov: y(20)=44%

4. Starost in krvni tlak

a) a=110,8 b=0,717

b) t=-2,768, $t_{tab} = \mp 2{,}306$

c) 146,6

d) (115,2; 178,1)

5. Hitrost vetra in čas teka čez ovire

b) a=13,322 b=-0,0846

c) s(b)=0,0236 t=-3,5839 p=0,002

d) a=13,32179201 s(a)=0,034598153 (13,2; 13,4)

KORELACIJA

1. Primerjava merskih metod

$$r = \frac{300503,5}{\sqrt{322372,5 \cdot 292988,1}} = 0,978 \quad r_{tab} = 0,6319$$

2. Višina in teža otrok

r=0,9159 r_{tab}=0,5760

3. Stroški oglaševanja

r=0,8754 r_{tab}=0,6319

4. Ptiči

r=0,866, r_{tab}=0,576

w=1,3169; meji za w: l_1=0,664; l_2=1,970

meji za r: l_1=0,581 l_2=0,962

5. Primerjava dveh ocenjevalcev

r_s=0,903 r_{tab}=0,7939

9.7 χ^2-preizkusi

PREIZKUŠANJE DOMNEV O PORAZDELITVI SPREMENLJIVKE

1. Pošteni kocki

$\chi^2 = 11,2 \qquad \chi^2_{0,05}(2) = 5,991 \qquad p = 0,0037$

2. Slučajne števke

$\chi^2 = 2,1 \qquad \chi^2_{0,05}(9) = 21,666 \qquad p = 0,9898$

3. Kakovost nogavic

$\chi^2 = 2,656 \qquad \chi^2_{0,05}(2) = 5,9991 \qquad p = 0,2650$

4. Beljakovine v mleku

25,25% 11,69% 13,06% 37,83% 12,17%

$\chi^2 = 35,0 \qquad \chi^2_{0,05}(3) = 7,815 \qquad p = 0,0000$

ANALIZA KONTINGENČNIH TABEL

1. Ocene pri matematiki in pri statistiki

$\chi^2 = 145,781 \quad \chi^2_{0,05}(SP = 4) = 9,488 \quad p = 0,0000$

2. Gripa

$\chi^2 = 1,406 \quad \chi^2_{0,05}(SP = 1) = 3,841 \quad p = 0,2357$

3. C-vitamin

$\chi^2 = 4,811 \quad \chi^2_{0,05}(SP = 1) = 3,841 \quad \chi^2_{0,01}(SP = 1) = 6,634 \quad p = 0,0282$

4. Zorenje sira

$\chi^2 = 14,99 \quad \chi^2_{0,05}(SP = 2) = 5,991 \quad p = 0,0006$

5. Predvolilna anketa

$\chi^2 = 23,08 \quad \chi^2_{0,05}(SP = 6) = 12,591 \quad p = 0,0008$

6. Izobrazba ženina in neveste

$\chi^2 = 120{,}28 \quad \chi^2_{0,05}(SP = 3) = 7{,}814 \quad p = 0{,}0000$

10 LITERATURA

- BLEJEC, M. 1972. Statistične metode za ekonomiste. Univerza v Ljubljani, Ekonomska fakulteta.
- CARLSON W. L., THORNE B. 1997. Applied Statistical Methods for Business, Economics, and the Social Sciences, Prentice Hall, Inc.
- CEDILNIK, A. 1995. Matematični priročnik, Didakta, Radovljica
- CEDILNIK, A. 2003. Uvod v verjetnostni račun, Metodološki zvezki 20, Fakulteta za družbene vede, Ljubljana
- ČIBEJ, J. A. 1994. Matematika. Kombinatorika, verjetnostni račun, statistika. Ljubljana: DZS.
- DALY, F., HAND, D.J., JONES M.C., LUNN A.D., MCCONWAY K.J. Elements of Statistics, The Open University, Addision-Wesley Publishing Company, 1995
- EVANS M., HASTINGS, N., PEACOCK B. 1993. Statistical Distributions. John Wiley and Sons Inc.
- FERLIGOJ, A. 1994. Osnove statistike na prosojnicah. Samozaložba, Ljubljana.
- FREUND, J. E. 1988. Modern Elementary Statistics. Prentice Hall International.
- HAYS W. L. 1994. Statistics. Holt, Rinehart and Winston Inc.
- HOGG R. V., LEDOLTER J. 1989. Engineering Statistics. Macmillan Publishing Company.
- JAMNIK, R. 1980. Matematična statistika. DZS, Ljubljana.
- JAMNIK, R. 1992. Matematika. Društvo matematikov, fizikov in astronomov Slovenije.
- KASTELEC, D., KOŠMELJ, K. 1998. Pogled skozi OKNA: oblikovanje besedila, grafični prikazi in analiza podatkov: Windows 95, Microsoft Office 95/97. Samozaložba, Ljubljana.
- KLEINBAUM, D. G., KUPPER, L. L, MULLER, K. E. 1988. Applied Regression Analysis and Other Multivariable Methods. PWS-KENT Publishing Company.
- KOŠMELJ, K. 1995. Opisna statistika na zgledih: naloge in rešitve. Didakta, Radovljica.
- KOŠMELJ, K. 1996. Izpitne naloge iz statistike. Biotehniška fakulteta, Univerza v Ljubljani, Ljubljana.
- KOŠMELJ, K., KASTELEC, D. 2000. Statistične tabele. Biotehniška fakulteta, Univerza v Ljubljani, Ljubljana, 2. ponatis
- MENDELHALL, W., BEAVER R. J. A. 1992. Course in Business Statistics. PWS-KENT Publishing Company.
- RICE J. A. 1988. Mathematical Statistics and Data Analysis. Wadsworth, Inc.
- SNEDECOR G.W., COCHRAN W.G. 1996. Statistical Methods. Iowa State University Press, 8. izd.
- SPIEGEL, M. R. 1980. Probability and Statistics, Schaum's Outline Series, McGraw-Hill Book Co,

- Statistični terminološki slovar. 1993. Statistično društvo Slovenije in Društvo matematikov, fizikov in astronomov, Ljubljana
- STEEL, R. G. D., TORRIE, J. H., DICKEY, D. A. 1994. Principles and Procedures of Statistics. A Biometrical Approach. The McGraw-Hill Companies, Inc.
- VADNAL, A. 1972. Elementarni uvod v verjetnostni račun. Državna založba Slovenije, Ljubljana
- WEISBERG S. 1985. Applied Linear Regression. John Wiley and Sons.

Viri podatkov so navedeni v besedilu.

11 STVARNO KAZALO

12 KAZALO TABEL

13 KAZALO SLIK