Priprava na teoretični izpit pri predmetu

OSNOVE VERJETNOSTI IN STATISTIKE

Fakulteta za računalništvo in informatiko

DATUM: 3.9.2011

PRIPRAVIL: FATAL ERROR (T.R.)

1. STATISTIKA

a) Enota, populacija, vzorec, spremenljivka:

Enota = posamezni proučevani element (primer: študent na neki fakulteti)
Populacija = množica vseh proučevanih elementov; pomembna je natančna opredelitev populacije, časovno ali prostorsko (primer: vsi študenti na točno določeni fakulteti v določenem letu)

Vzorec = podmnožica iz populacije, na osnovi katere sklepamo o lastnostih cele populacije
Spremenljivka = lastnost enot; podatek ki ga proučujemo (primer: spol, uspeh pri nekem predmetu)

b) Spremenljivke glede na tip izražanja vrednosti:

Opisne (distributivne) – vrednost lahko opišemo z besedo (primer: poklic, uspeh, naziv)
Številske (numerične) – vrednost je izražena s številko (primer: starost)

c) Spremenljivke, glede na tip merjenja:

Nominalne – vrednost je lahko samo enaka ali različna (primer: spol)
Ordinalne – vrednost lahko uredimo glede na velikost (primer: uspeh, starost)
Intervalne – lahko jih primerjamo glede na interval (primer: temperatura)
Razmernostne – lahko jih primerjamo glede na razmerje med vrednostmi (primer: starost)

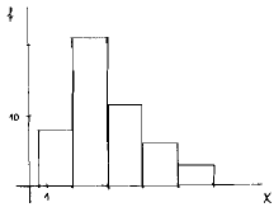
d) Frekvenčna porazdelitev:

Frekvenčna porazdelitev spremenljivke je tabela, ki jo določajo vrednosti ali skupine vrednosti in njihove frekvence. Če je spremenljivka ordinalnega značaja, vrednosti uredimo po velikosti.

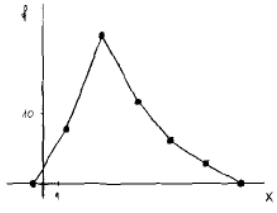
e) Prikaz frekvenčne porazdelitve:

Histogram – drug poleg drugega rišemo stolpce – pravokotnike, katerih višina je sorazmerna frekvenci v razredu. Širina vseh pravokotnikov je enaka.
Poligon – v koordinatnem sistemu zaznamujemo točke (x_i, f_i), kjer je x_i sredina tega razreda in f_i njegova frekvenca.
Ogiva – grafična predstavitev komulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke.

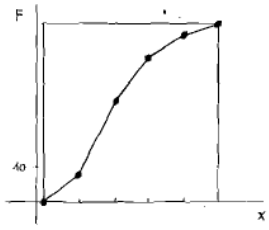
- **Histogram**



- **Poligon**



- **Ogiva**



f) Sredine:

Aritmetična sredina (ali tudi povprečje) – povprečna vrednost niza podatkov; seštevek vseh vrednost, deljen s številom enot v populaciji ali vzorcu.
Geometrijska sredina – definirana je kot n-ti koren zmnožka vseh členov množice, kjer je n število elementov v množici.

Harmonična sredina – predstavlja eno od srednjih vrednosti. Primerna je v primerih, ko je potrebno najti srednje vrednosti stopenj.
Kvadratna sredina – kvadratni koren aritmetične sredine kvadratov podatkov v vzorcu.

g) Mediana in modus:

Mediana – predstavlja srednjo vrednost, če so podatki urejeni po velikosti. Pri lihem številu elementov je to element, ki se nahaja točno na sredini. Če pa v vzorcu nastopa sodo mnogo elementov, potem vzamemo povprečje srednjih dveh.

Modus – predstavlja tisto vrednost v vzorcu, ki se največkrat ponovi. Torej, element, ki ima najvišjo frekvenco. Modus obstaja natanko takrat, kadar obstaja le en element, ki ima najvišjo frekvenco.

h) Kvartil, decil, centil:

Kvartil – katera koli od treh vrednosti kvartilov, ki delijo urejeno množico slučajnih spremenljivk na štiri enake dele. V množici nastopajo 3 kvartili, ki jih označimo z Q:
Prvi kvartil (25% manjših vrednosti in 75% večjih), Drugi kvartil (točno na sredini; mediana), Tretji kvartil (75% manjših vrednosti in 25% večjih).

Decil – katerakoli od devetih vrednosti kvantilov, ki delijo urejeno množico slučajnih spremenljivk na deset enakih delov. V množici nastopa 9 decilov, ki jih označimo z D.

Centil – katerakoli od 99 vrednosti kvantilov, ki delijo urejeno množico slučajnih spremenljivk na 100 enakih delov. V množici nastopa 99 decilov, ki jih označimo z C.

2. KOMBINATORIKA

a) Pravilo vsote:

Če imamo na voljo *m* možnosti iz prve skupine in *n* možnosti iz druge skupine, izbrati pa želimo točno eno možnost iz prve ali iz druge skupine, potem imamo na izbiro skupno *m+n* možnosti.

b) Pravilo produkta:
Pravilu produkta pravimo tudi osnovni izrek kombinatorike: Če imamo na voljo *m* možnosti iz prve skupine in *n* možnosti iz druge skupine, izbrati pa želimo eno možnost iz prve in hkrati eno iz druge skupine, potem imamo na izbiro skupno *m*n* možnosti.

c) Permutacije:

Permutacije so razporeditve danih n elementov na n prostih mest.
Če so vsi elementi med seboj različni, so to permutacije brez ponavljanja. Število permutacij brez ponavljanja izračunamo po formuli:

$$Pn = n! \sim n * (n - 1) \ldots 3 * 2 * 1$$

V formuli nastopa računska operacija fakulteta n! (ali tudi faktoriela), ki pravi, da zmnožimo vsa števila od 1 do n.
Opomba: Zaradi računskih razlogov definiramo fakulteto tudi za število 0 in sicer 0! = 1

d) Permutacije s ponavljanjem:

Permutacije s ponavljanjem so permutacije elementov, ki niso vsi med sabo različni. Pri tem lahko nastopa celo več skupin med sabo enakih elementov. Recimo, da je v prvi taki skupini k_1 enakih elementov, v drugi k_2 enakih elementov, ..., v m-ti pa k_m enakih elementov. Potem število permutacij s ponavljanjem izračunamo po formuli:

$$P_n^{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1!\ k_2! \cdots k_m!}$$

e) Variacije:

Variacije brez ponavljanja so razporeditve n različnih elementov na r prostih mest. Pri tem je $r < n$, zato ostane nekaj elementov nerazporejenih. Število variacij brez ponavljanja računamo po formuli:

$$V_n^r = \frac{n!}{(n-r)!}$$

f) Variacije s ponavljanjem:

Variacije s ponavljanjem so razporeditve, pri katerih poskušamo na r prostih mest razporediti elemente n različnih vrst. Pri tem se lahko element določene vrste v razporeditvi pojavi poljubno mnogokrat. Število variacij s ponavljanjem izračunamo po formuli:

$$(\mathrm{p})V_n^r = n^r$$

g) Kombinacije:

Če pri variacijah zanemarimo vrstni red in opazujemo samo, kateri elementi so izbrani, dobimo kombinacije. Gre za izbire r (različnih) elementov izmed n različnih elementov, ki so na voljo. Število kombinacij brez ponavljanja izračunamo s pomočjo binomskega simbola:

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} \qquad \text{oziroma} \qquad C_n^r = \binom{n}{r}$$

h) Kombinacije s ponavljanjem:

Kombinacije s ponavljanjem dobimo, če pri variacijah s ponavljanjem zanemarimo vrstni red. To so torej izbire, kjer izbiramo r elementov izmed n, vendar pa lahko isti element izberemo tudi večkrat (poljubno mnogokrat). Število kombinacij s ponavljanjem izračunamo po formuli:

$$(\mathrm{p})C_n^r = \binom{n+r-1}{r}$$

3. POSKUSI IN DOGODKI

a) Poskus:

Poskus je realizacija neke množice skupaj nastopajočih dejstev. Poskus je torej vsako dejanje, ki ga opravimo v natanko določenih pogojih.

Primeri: Met igralne kocke, met kovanca, iz kupčka 32 kart vzamemo 4 karte, met pikada v tarčo...

b) Dogodek:

Dogodek je izid pri nekem poskusu. Gre za pojav, ki se lahko zgodi, ali pa ne.

Primeri: na igralni kocki pade 6 pik, na kovancu pade grb, izbrane karte so asi, zadenemo v sredino tarče...

c) Poddogodek:

Dogodek A je poddogodek (način) dogodka B, če se zagotovo zgodi vsakič, ko se zgodi dogodek A. Pravimo, da je dogodek A podmnožica dogodka B.
Če je dogodek A način dogodka B in obenem dogodek B način dogodka A, potem sta oba dogodka enaka.

d) Vrste dogodkov:

Slučajni dogodek – se v izbranem poskusu zgodi z neko verjetnostjo (obvezno med 0 in 1).
Gotov dogodek – se v vsaki ponovitvi poskusa zgodi, njegova verjetnost je 1.
Nemogoč dogodek – nikoli se ne zgodi, njegova verjetnost je 0.

Gotov dogodek G ustreza univerzalni množici, nemogoč dogodek N pa prazni množici.

e) Nasprotni dogodek:

Naj bo A nek dogodek, ki se zgodi z verjetnostjo P(A) = 0.9, potem je $\overline{A}$ nasprotni (ali tudi komplementarni) dogodek in se zgodi z verjetnostjo $P(\overline{A}) = 1 - P(A) = 0.1$

Primer:
Dogodek A ... na kocki pade sodo število pik
Nasprotni dogodek A ... na kocki pade liho število pik

f) Nezdružljivi dogodki:

Dogodka A in B sta nezdružljiva, če se ne moreta zgoditi hkrati. To se zgodi natanko tedaj, ko je njun produkt nemogoč dogodek. Pišemo: $A \cap B = N$.
Dogodki $A_1, A_2, \ldots A_n$ so med seboj nezdružljivi, če velja: $P(A_1 \cap A_2 \cap \ldots \cap A_n) = N$.

Primer:
A ... izvlečena karta je kralj B … izvlečena karta je as
Dogodka A in B sta nezdružljiva, saj se lahko v vsaki ponovitvi zgodi le eden od dogodkov.

g) Osnovni in sestavljeni dogodki:

Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A sestavljen dogodek. Dogodek, ki ni sestavljen, imenujemo osnoven (ali tudi elementaren) dogodek. Elementarni dogodek vsebuje natanko en elementarni izid, sestavljen dogodek pa lahko vsebuje več elementarnih izidov.

Primer:
A … na kocki pade 6 pik (to je osnovni dogodek)
B … na kocki pade sodo število pik (sestavljen dogodek, ker lahko pade 2, 4 ali 6 pik)

h) Popoln sistem dogodkov:

Množico dogodkov S = {A_1, A_2 ... A_n} imenujemo popoln sistem dogodkov, če se v vsaki ponovitvi poskusa zgodi natanko eden od dogodkov iz množice S. To pomeni, da ni noben med njimi nemogoč. Dogodki so paroma nezdružljivi, njuna vsota pa je gotov dogodek G.

i) Računanje verjetnosti:

Verjetnost dogodka A je razmerje med številom ugodnih izidov in številom vseh možnih izidov.

$$P(A) = \frac{\text{število ugodnih izidov}}{\text{število vseh možnih izidov}} \quad \text{oziroma} \quad P(A) = \frac{m}{n}$$

Še nekaj drugih obrazcev:

Popolna verjetnost: $P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2) + \ldots + P(H_n)P(A/H_n)$

Verjetnost nasprotnega dogodka: $P(A') = 1 - P(A)$ oziroma $P(A) + P(A') = 1$

Verjetnost unije dogodkov (splošno): $P(A \cup B) = P(A) + P(B) - P(A\,B)$

Verjetnost unije nezdružljivih dogodkov ($A\,B = \mathcal{N}$): $P(A \cup B) = P(A) + P(B)$

Verjetnost produkta neodvisnih dogodkov: $P(A\,B) = P(A)\,P(B)$

Unija (A ∪ B) se zgodi takrat, kadar se zgodi vsaj en od dogodkov A in B. Presek (A ∩ B) pa se zgodi, natanko takrat, ko se zgodita oba dogodka. Pri uniji zato računamo logično funkcijo ALI, pri preseku pa logični IN.

4. POGOJNA VERJETNOST

a) Pogojna verjetnost:

Pogojna verjetnost je verjetnost, da se zgodi dogodek A, pod pogojem, da se je zgodil neki drugi dogodek B. Takšno verjetnost označimo s P(A|B).
Pogojno verjetnost lahko določimo za nezvezne (diskretne) in zvezne slučajne spremenljivke.

Verjetnost dogodka B, če vemo, da se je zgodil dogodek A: $P_B(A) = P(A \mid B)$

Za dva dogodka dobimo pogojno verjetnost po obrazcu: $P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$

b) Neodvisni dogodki:

Dogodka A in B sta neodvisna, če ne vplivata eden na drugega. Kadar sta dogodka neodvisna, velja: $P(A \mid B) = P(A)$ in tudi obratno: $P(B \mid A) = P(B)$.
Pravilo lahko posplošimo na več dogodkov: Dogodki Ai, $i \in I$ so neodvisni, če je

$$P(A_j) = P(A_j / \bigcap_{i=1}^{j-1} A_i), \quad j \in I$$

Primer:
A ... iz prve posode vzamemo 3 rdeče kroglice
B ... iz druge posode vzamemo 5 zelenih kroglic

c) Dvofazni poskus:

Dvofazni poskus se dogaja v dveh fazah. Pravimo jim tudi relejni poskusi.
(1) zgodi se natančno en od dogodkov $H_1, H_2, H_3, \ldots, H_n$ (2) opazujemo nek izbrani dogodek A.
Dogodkom H_i pravimo hipoteze. Ti dogodki morajo sestavljati popoln sistem paroma nezdružljivih dogodkov. Nemogočih hipotez ne uporabimo.

d) Bayesov obrazec:

Bayesov obrazec se uporablja za računanje dvofaznega poskusa v obratni smeri.

Primer:
Vaš kolega meče dve kocki, pove da je v drugem metu vrgel več kot v prvem. Zanima nas, kolikšna je verjetnost, da je v prvem metu vrgel 1.

Verjetnost bomo izračunali s pomočjo obrazca: $P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$

e) Zaporedje neodvisnih poskusov:

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

f) Bernoullijevo zaporedje neodvisnih poskusov:

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

V Bernoullijevem zaporedju nas zanima kolikšna je verjetnost, da se v n zaporednih poskusih zgodi dogodek A natanko k-krat. Dogodek A se lahko v n poskusih zgodi na toliko načinov, kolikor lahko izberemo k načinov iz n načinov. Teh je „n nad k“.

g) Bernoullijev obrazec:

Bernoullijev obrazec P(n,p,k) pravi, da je verjetnost, da se pri n ponovitvah poskusa dogodek A zgodi natančno k-krat, če je verjetnost A v enem poskusu enaka p. Zanima nas kolikšna je verjetnost, da se v n zaporednih poskusih zgodi dogodek A natanko k-krat.

$$P(n,p,k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad \text{kjer je binomski simbol } \binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

n ~ vsi poskusi, k ~ uspešni poskusi, p ~ verjetnost enega poskusa.

h) Stirlingov obrazec:

Stirlingov obrazec je aproksimacija za funkcijo fakultete n!. Posebno učinkovit je pri velikih številih, ker ima majhno časovno zahtevnost in število računskih operacij.

$$\text{Stirlingov obrazec:} \quad n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

i) Laplaceov točkovni obrazec:

Laplaceov obrazec se uporablja za risanje usojenih krivulj (imenovane tudi fatalke) iz Pascalovega trikotnika. Gre za to, da vsako vrstico trikotnika prenesemo v koordinatni sistem in ga narišemo od X osi, do koder sežejo števila.
Laplaceov obrazec smemo uporabiti, kadar je n velik, p pa blizu 1/2. V primeru velikih n pa uporabimo De Moivrov obrazec, ki je tudi poseben primer Laplaceovega obrazca.

$$\text{Laplaceov točkovni obrazec} \qquad P_n(k) \approx \frac{1}{\sqrt{2\pi npq}}\, e^{-\frac{(k-np)^2}{2npq}}$$

$$\text{De Moivrov točkovni obrazec} \qquad P_n(k) \approx \frac{1}{\sqrt{\pi n/2}}\, e^{-\frac{(k-n/2)^2}{n/2}}$$

j) Poissonov obrazec:

Poissonov obrazec se uporablja za računanje verjetnosti v Poissonovi porazdelitvi.

$$\textbf{Poissonov obrazec:} \quad \text{za } p \text{ blizu } 0 \qquad P_n(k) \approx \frac{(np)^k e^{-np}}{k!}$$

5. SPREMENLJIVKE

a) Slučajna spremenljivka:

S spremenljivko označimo izid nekega dogodka. Kadar je izid takega dogodka odvisen le od slučaja (drugi dejavniki ne vplivajo), potem spremenljivki pravimo slučajna spremenljivka.
Da je slučajna spremenljivka znana, je potrebno vedeti kakšna je njena zaloga vrednosti in verjetnost vsake med vrednostmi, ki jih spremenljivka lahko zavzame.

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami, njihove vrednosti pa z malimi tiskanimi (na primer: X = xi pomeni: dogodek, pri katerem spremenljivka X zavzame vrednost xi).

b) Porazdelitveni zakon in funkcija:

Porazdelitveni zakon je predpis, ki določa verjetnosti posameznih vrednostim slučajne spremenljivke. F(X) = P(X < x). F(x) pa je porazdelitvena funkcija.

c) Diskretna in zvezna slučajna spremenljivka:

- *Diskretna slučajna spremenljivka:* slučajna spremenljivka, pri kateri je zaloga vrednosti neka števna (diskretna) množica.
- *Zvezna slučajna spremenljivka:* slučajna spremenljivka, ki lahko zavzame vsako realno število znotraj določenega intervala.

Slučajna spremenljivka je zvezno porazdeljena, če obstaja taka integrabilna funkcija p, imenovana gostota verjetnosti, da za vsako število x $\in$ R, velja:

$$F(x)=P(x<X)=\int_{-\infty}^{x} p(t)\,dt \text{ , kjer je } p(x)\geq 0$$

d) Enakomerna zvezna slučajna spremenljivka:

To je slučajna spremenljivka, ki je enakomerno porazdeljena na končnem intervalu [a, b] če je njena gostota:

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{drugod} \end{cases}$$

Grafično si jo lahko predstavljamo kot pravokotnik nad intervalom [a, b] višine 1/(b-a).

e) Verjetnostna shema:

Verjetnostna shema prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti xi, pod njimi pa so pripisane pripadajoče verjetnosti. Seštevek vseh verjetnosti je enak 1.

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

f) Kontigenčna tabela:

Kontigenčna tabela je dvorazsežna frekvenčna porazdelitev. Podatki so urejeni po obeh spremenljivkah in ustrezajo prvi spremenljivki v glavi tabele in drugi spremenljivki, ki jo zapišemo na levi strani tabele.

6. VERJETNOSTNE PORAZDELITVE

a) Enakomerna porazdelitev:

Končna diskretna slučajna spremenljivka je enakomerno porazdeljena, če so vse njene vrednosti enako verjetne.

Primeri:
- Vsi izidi na kocki so enako verjetni. Ker imamo 6 izidov, je verjetnost vsakega izida enaka 1/6.
- Na kovancu je enako verjetno ali bo padel grb ali cifra. Ker sta 2 izida, je verjetnost vsakega 1/2.

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 \\ 1/2 & 1/2 \end{pmatrix}$$

b) Binomska porazdelitev:

Binomska porazdelitev je diskretna porazdelitev n uspešnih izidov zaporednih poskusov, kjer sta možna le dva izida: DA ali NE. Takšno vrsto neodvisnih poskusov imenujemo Bernoullijevi poskusi, sam postopek pa Bernoullijev postopek.
Pri binomski porazdelitvi nas zanima verjetnost, da se v zaporedju n-tih poskusov zgodi dogodek DA k-krat. Tabela binomske porazdelitve je simetrična glede na sredino, le če je p=0,5.

Verjetnost računamo po Bernoullijevem obrazcu: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

c) Poissonova porazdelitev:

Poissonova porazdelitev je nezvezna diskretna porazdelitev. Izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se dogodki pojavijo s poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek.
V praksi Poissonovo porazdelitev uporabimo, če vemo da se dogodek v povprečju zgodi 3-krat na minuto in nas zanima kolikokrat se bo zgodil v 1 uri.

Primeri: število dostopov do spletnega strežnika v 1 uri, število telefonskih klicev na bazni postaji vsako minuto, število prometnih nesreč na slovenskih cestah v 1 tednu

Verjetnost Poissonove porazdelitve računamo po obrazcu:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- e je osnova naravnih logaritmov (e = 2.71828183)
- k je število ponovitev dogodka
- λ je pozitivno realno število, ki je enako pričakovanemu številu ponovitev dogodka v danem intervalu

d) Zveza med Binomsko in Poissonovo porazdelitvijo:

V binomski porazdelitvi obravnavamo število uspehov v n ponovitvah poskusa, v binomski pa nas zanima verjetnost, da se pojavi m dogodkov v nekem časovnem obdobju.
Poissonovo porazdelitev uporabljamo tudi za izračun verjetnosti pojavljanja dogodkov na določeni razdalji, površini ali prostornini (ne samo v časovnem intervalu).

Primer:
Binomska: število uspešno opravljenih izpitov v študijskem letu (recimo, da jih opravimo 10)
Poissonova: verjetnost, da bo študent letos opravil vse (recimo, da jih ima 10) izpite

e) Pascalova porazdelitev:

Pascalova porazdelitev (imenovana tudi negativna binomska porazdelitev) opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi m-krat.
Primer: mečemo kovanec, dokler 3-krat ne pade cifra.

Verjetnostna funkcija Pascalove porazdelitve: $p_k = \binom{k-1}{m-1} p^m q^{k-m}$

f) Geometrijska porazdelitev:

Geometrijsko porazdelitev lahko uporabimo, kadar osnovne dogodke lahko predstavimo kot „enakovredne“ točke na delu premice (ravnine ali prostora). Verjetnost sestavljenega dogodka določimo kot razmerje dolžin (ploščin, prostornin) dela, ki ustreza ugodnim izidom in dela, ki ustreza vsem možnim izidom.

$$P(x=k) = (1-p)^{k-1} \cdot p$$

g) Hipergeometrijska porazdelitev:

Geometrijska porazdelitev opisuje verjetnost dogodka, da je med n izbranimi kroglicami natanko k belih, če je v posodi M belih in N-M črnih kroglic. Pri tem kroglice izbiramo n-krat in brez vračanja.

Obrazec za hipergeometrijsko porazdelitev: $$p_k = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

h) Gaussova porazdelitev:

Gaussova porazdelitev (imenovana tudi normalna porazdelitev) je verjetnostna porazdelitev vrednosti statističnih enot v statistični populaciji, ki je v grafični predstavitvi oblikovana v obliki zvona oziroma normalne krivulje. Vanjo sodi družina porazdelitev, ki imajo različne parametre (na primer aritmetično sredino in standardni odklon), a oblikujejo enake grafe porazdelitve. Standardna normalna porazdelitev je porazdelitev vrednosti s povprečjem (aritmetično sredino) 0 in standardnim odklonom 1.
Graf Gaussove porazdelitve ima obliko enogrbe kamele. Zaloga vrednosti normalno porazdeljene slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je:

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

Kadar je spremenljivka približno normalno porazdeljena, jo statistični karakteristiki povprečje in standardni odklon zelo dobro opisujeta.

i) Laplaceov intervalski obrazec:

Laplaceov intervalski obrazec uporabimo takrat, kadar nas zanima verjetnost $P_n(k_1, k_2)$, da se v Bernoullijevem zaporedju neodvisnih poskusov v n zaporednih poskusih zgodi dogodek A vsaj k_1-krat in manj kot k_2-krat.

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

Za (zelo) velike n lahko vsoto zamenjamo z integralom $$P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx$$

j) Eksponentna porazdelitev:

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

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

porazdelitvena funkcija pa je $$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}$$

k) Funkcija napake:

Funkcija napake je liha, zvezno odvedljiva, strogo naraščajoča funkcija, za katero velja $\Phi(0) = 0$, $P_n(k_1 , k_2) \sim \Phi(xk_2) - \Phi(xk_1)$.

Funkcija napake imenujemo funkcijo $$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2}\, \mathrm{d}t$$

Spremenljivko X ~ N(μ , σ) lahko pretvorimo v standardizirano spremenljivko X ~ N(0, 1) tako, da od njene vrednosti odštejemo povprečje in delimo s standardnim odklonom:

$$z = \frac{x_i - y}{\sigma}$$

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

l) Porazdelitev gama:

Eksponentno porazdelitev lahko še precej posplošimo. Naj bosta b, c > 0. Tedaj ima porazdelitev Gama Γ(b, c) gostoto:

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

p(x) = 0 za x ≤ 0. Za b = 1 seveda dobimo eksponentno porazdelitev. Funkcijo gama lahko definiramo z določenim integralom za R[z] > 0 (Eulerjeva integralna forma) Za naravno število x = n ∈ {1, 2, ...} torej dobimo Γ(n) = (n − 1)! Za b = 1 seveda dobimo eksponentno porazdelitev.

z integralom za R[z]>0 $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt = 2\int_0^\infty e^{-t^2} t^{2z-1} dt$

m) Porazdelitev Hi-kvadrat:

Hi-kvadrat je poseben primer Gama porazdelitve: $\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$

n) Cauchyjeva porazdelitev:

Cauchyjeva porazdelitev je porazdelitev z gostoto $p(x) = \frac{a}{\pi} \frac{1}{1 + a^2(x-b)^2}$

o) Bernoullijev zakon velikih števil (1713):

Naj bo k frekvenca dogodka A v n neodvisnih ponovitvah danega poskusa, v katerem ima dogodek A verjetnost p. Tedaj za vsak ε > 0 velja:

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

Primer: Kolikšna je verjetnost, da se pri metu kovanca relativna frekvenca grba v 3600 metih ne razlikuje od 0,5 za več kot 0,01, se pravi, da grb pade med 1764 in 1836-krat?

7. SLUČAJNI VEKTORJI

a) Slučajni vektor:

Slučajni vektor je n-terica slučajnih spremenljivk X = (X_1, X_2 … X_n). Opišemo ga s porazdelitveno

funkcijo (Xi ∈ R)

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

Funkcija F je za vsako spremenljivko naraščajoča in od leve zvezna, veljati pa mora tudi:

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

Funkciji Fi(Xi) = (∞, ..., ∞, Xi, ∞, ... ∞) pravimo robna porazdelitvena funkcija spremenljivke Xi. Funkcija je lahko porazdelitvena funkcija nekega vektorja, samo če zavzame vrednosti na [0, 1].

b) Robni gostoti slučajnega vektorja:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dy \qquad p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dx$$

Zvezno porazdeljeni slučajni spremenljivki sta neodvisni natanko tedaj, ko je tudi slučajni vektor (x,y) porazdeljen zvezno z gostoto:

$$p_{XY}(x,y) = p_X(x)\,p_Y(y)$$

c) Odvisnost spremenljivk:

Kako pokazati odvisnost (povezanost) dveh spremenljivk? Podatke lahko zapišemo v kontigenčno tabelo, kjer po X osi navedemo vrednosti prve spremenljivke, po Y pa vrednosti druge. Vrednosti v vsaki vrstici in stolpcu med seboj seštejemo in na konec vrstice/stolpca zapišemo vsoto. Spremenljivki X in Y sta odvisni le v primeru, kadar velja:

$$P(x = x, y = y) = x,y \;=\; x \cdot y = P(x = x) \cdot P(y = y)$$

Na kratko rečeno, preveriti moramo ali je Cov(X, Y) = 0. Ko ta pogoj drži, sta X in Y neodvisni.
Primer: Spremenljivka X šteje število naprav, Y pa število zaporednih operacij:

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

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0,08 & 0,37 & 0,40 & 0,15 \end{pmatrix}$$

$$Y : \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0,40 & 0,25 & 0,20 & 0,15 \end{pmatrix}$$

Ali sta slučajni spremenljivki X in Y neodvisni?

Ne nista, saj velja npr.: $P(x = 4, y = 3) = 0 \neq 0,15 \cdot 0,15 = P(x = 4) \cdot P(y = 3)$.

d) Pogojna porazdelitev:

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

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

V diskretnem primeru je: $p_{ik} = P(X = x_i, Y = y_k)$, $B = (Y = y_k)$ in $P(B) = P(Y = y_k) = q_k$

Tedaj je pogojna porazdelitvena funkcija

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

e) Večrazsežne porazdelitve:

Slučajni vektor X = (X_1, X_2 ... X_n) je zvezno porazdeljen, če obstaja integrabilna funkcija (gostota verjetnosti) p(x_1, x_2 … x_n) ≥ 0 z lastnostjo:

$$F(x_1, x_2, x_3, \ldots, x_n) = \int_{-\infty}^{x_1} \left(\int_{-\infty}^{x_2} \left(\ldots \left(\int_{-\infty}^{x_n} p(t_1, t_2, \ldots, t_n)\, dt_n \right) \ldots \right) dt_2 \right) dt_1 \quad \text{in} \quad F(\infty, \infty, \infty, \ldots, \infty) = 1$$

Robni verjetnostni gostoti sta: $p_X(x) = F_X'(x) = \int_{-\infty}^{\infty} p(x, y)\, dy$ in $p_Y(y) = F_Y'(y) = \int_{-\infty}^{\infty} p(x, y)\, dx$

Zaloga vrednosti je kvečjemu števna množica.
Primer večrazsežne porazdelitve je polinomska porazdelitev, ki je določena s predpisom:

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

Kvocient šteje permutacije s ponavljanjem. Za vrednost r=2 dobimo Binomsko porazdelitev tj. B(n, p) = P(n, p, q).

f) Neenakost Čebiševa:

Neenačba Čebiševa ocenjuje kakšna je verjetnost, da se slučajna spremenljivka veliko razlikuje od matematičnega upanja. Če ima slučajna spremenljivka X končno disperzijo, tj. DX < 1, velja za vsak ε > 0 neenakost:

$$P(|X - \mathsf{E}X| \geq \varepsilon) \quad \leq \quad \frac{\mathsf{D}X}{\varepsilon^2}$$

Izrek: Če so slučajne spremenljivke Xi paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C, tj. Dxi < C (za vsak I), velja za zaporedje šibki zakon velikih števil.

Uporaba: 1000-krat vržemo kovanec. Oceni da bo število grbov med 400 in 600.

g) Neenakost Markova:

Nastane kot posledica neenakosti Čebiševa.

Izrek: Če gre za zaporedje slučajnih spremenljivk Xi izraz $\frac{\mathsf{D}S_n}{n^2} \to 0$, ko gre $n \to \infty$, velja za zaporedje šibki zakon velikih števil.

Naj velja X ≥ 0. Potem za vsak a > 0 velja neenakost P(X ≥ a) ≤ E(X)/a

8. MOMENTI IN KOVARIANCA

a) Matematično upanje:

Matematično upanje EX (ali tudi pričakovana vrednost) je posplošitev povprečne vrednosti diskretne spremenljivke X.

Diskretna slučajna spremenljivka X z verjetnostno funkcijo p_k ima matematično upanje:

$$\mathsf{E}X = \sum_{i=1}^{\infty} x_i p_i \qquad \text{če je} \qquad \sum_{i=1}^{\infty} |x_i| p_i < \infty$$

Zvezna slučajna spremenljivka X z gostoto p(x) ima matematično upanje:

$$\mathrm{E}X = \int_{-\infty}^{\infty} xp(x)\,dx \quad \text{če je} \quad \int_{-\infty}^{\infty} |x|p(x)\,dx < \infty$$

Matematično upanje ne obstaja za diskretno in zvezno (Cauchyjevo) porazdelitev.

b) Lastnosti matematičnega upanja:

Če sta slučajni spremenljivki, ki imata matematično upanje neodvisni, obstaja tudi matematično upanje njunega produkta in velja EXY = EX * EY.
Spremenljivki, za kateri velja EXY ≠ EX * EY imenujemo korelirani.

c) Disperzija:

Disperzija (razpršenost ali tudi varianca) DX slučajne spremenljivke, ki ima matematično upanje, je določena z izrazom: $DX = E(X - EX)^2$

Disperzija je vedno nenegativna, DX ≥ 0, je pa lahko tudi neskončna. Velja: $DX = EX^2 - (EX)^2$.

d) Standardni odklon:

Standardni odklon (ali tudi standardna deviacija) σ slučajne spremenljivke je statistični kazalec, največkrat uporabljen za merjenje statistične razpršenosti enot. Z njim je moč izmeriti, kako razpršene so vrednosti, vsebovane v populaciji.

Standardni odklon spremenljivke X: $\sigma X = \sqrt{\mathrm{D}X}$

Približno 68,3% vseh meritev leži na razdalji 1x standardnega odklona od njihovega povprečja.

e) Standardizacija:

Slučajno spremenljivko standardiziramo (normaliziramo) s transformacijo $X_S = \dfrac{X-\mu}{\sigma}$

kjer sta $\mu = \mathrm{E}X$ in $\sigma = \sqrt{\mathrm{D}X}$.

Za X_S velja $\mathrm{E}X_S = 0$ in $\mathrm{D}X_S = 1$, saj je $\mathrm{E}X_S = \mathrm{E}\dfrac{X-\mu}{\sigma} = \dfrac{\mathrm{E}(X-\mu)}{\sigma} = \dfrac{\mu-\mu}{\sigma} = 0$

kjer smo upoštevali linearnost matematičnega upanja, ter $\mathrm{D}X_S = \mathrm{D}\dfrac{X-\mu}{\sigma} = \dfrac{\mathrm{D}(X-\mu)}{\sigma^2} = \dfrac{\sigma^2-0}{\sigma^2} = 1$

f) Matematična upanja in disperzija porazdelitev:

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

g) Kovarianca:

Kovarianca C(X, Y) slučajnih spremenljivk X in Y je definirana z izrazom *Cov(X, Y) = EXY – EXEY,* kar je izpeljava iz Cov(X, Y) = E((X - EX) (Y - EY))

Če obstaja DX in DY, potem obstaja tudi Cov(X, Y) in velja: $|K(x,y)| \leq \sqrt{DX DY} = \sigma X \sigma Y$

Izpeljave: EXY = E(EY) = E(X) * E(Y) EXEY = E(X) * E(Y)
Pri kovarianci velja simetričnost, kar pomeni, da vrstni red spremenljivk X in Y lahko tudi obrnemo.
Spremenljivki X in Y sta nekorelirani (neodvisni) natanko takrat, ko je Cov(X, Y) = 0.
Če imata spremenljivki X in Y končni disperziji, jo ima tudi njuna vsota X+Yin velja:
D(X+Y) = DX + DY + 2Cov(X, Y)

Zveza med disperzijo in kovarianco: D(X + Y) = DX + DY + 2Cov(X, Y)
Če pa sta spremenljivki nekorelirani, je enostavno: D(X + Y) = DX + DY
Torej sta normalno porazdeljeni slučajni spremenljivki X in Y neodvisni natanko takrat, ko sta nekorelirani.

h) Kovariančna matrika:

Kovariančna matrika (ali tudi variančno-kovariančna matrika) je matrika, katere elementi so kovariance i-tega in j-tega elementa vektorja slučajne spremenljivke.

- Kovariančna matrika K = [Kij] je simetrična: Kij = Kji.
- Diagonalne vrednosti so disperzije spremenljivk Kii = DXi.
- Če je determinanta matrike detK=0, potem kovariančna matrika K ni obrnljiva.

i) Korelacijski koeficient:

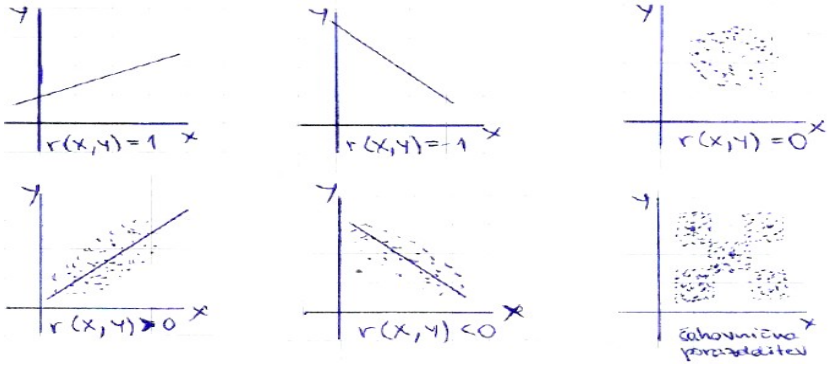
Korelacija ali korelacijski koeficient predstavlja moč linearne povezanosti dveh spremenljivk. Definiramo ga kot razmerje med kovarianco in standardnim odklonom spremenljivk X in Y. Najbolj znan je Pearsonov koeficient, ki je definiran:

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

Korelacijski koeficient lahko zavzame vrednosti na intervalu [-1 ,1]. Če se z večanjem vrednosti prve spremenljivke večajo tudi druge, gre za pozitivno povezanost. Takrat je koeficient povezanosti blizu 1.

j) Pomen korelacijskega koeficienta:

Pomen korelacijskega koeficienta v grafu:



Pomen korelacijskega koeficienta za slučajni spremenljivki:

relacijski koeficient	pomen relacijskega koeficienta
r(X, Y) = 1	Y = a*X + b, a > 0
r(X, Y) = -1	Y = -a*X + b, a > 0
r(X, Y) = 0	spremenljivki sta nekorelirani, sta neodvisni
r(X, Y) > 0	spremenljivki sta pozitivno korelirani
r(X, Y) < 0	spremenljivki sta negativno korelirani

k) Pomen standardizacije:

Standardizacija spremenljivk je postopek, s katerim vrednosti spremenljivke transformiramo, in sicer tako, da od vsake vrednosti spremenljivke odštejemo aritmetično sredino μx in delimo z njenim standardnim odklonom σx. Dobimo standardizirano vrednost. Tako iz velikega števila različnih spremenljivk naredimo eno novo spremenljivko – če imamo na primer različne merske enote, spremenljivke najprej standardiziramo in jih tako spravimo na skupni imenovalec oziroma na isto mersko raven (vrednost). Standardizirana spremenljivka Z ima vedno aritmetično sredino μz = 0 in standardni odklon σz = 1. Standardizacijo uporabljamo, kadar želimo, da ima vsaka od spremenljivk enak vpliv oziroma enako težo na novo, skupno oceno.

l) Centralni limitni izrek:

Za zaporedje slučajnih spremenljivk Xi velja centralni limitni zakon, če porazdelitvene funkcije za Zn gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve,
to je, če za vsak x ∈ R velja:

$$\lim_{n\to\infty} P\left(\frac{S_n - \mathrm{E}S_n}{\sigma(S_n)} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Osnovni CLI: Če so slučajne spremenljivke Xi neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon.

$$\frac{Sn - E(Sn)}{\sigma(Sn)}$$ je razporejena približno kot N(0, 1)

- CLI velja tudi, če niso enako porazdeljene, če Xi nimajo vse iste disperzije in/ali matematična upanja, če za Xi velja E(Xi) = n in D(Xi) = σ^2
- CLI ne velja, če spremenljivke niso neodvisne.

9. INTERVALI ZAUPANJA

a) Interval zaupanja:

Za aritmetično sredino (ali tudi povprečje): Denimo, da s slučajnim vzorcem ocenjujemo parameter γ. Potem je interval zaupanja Z (ki je določen s spodnjo Zmin in zgornjo mejo Zmax) tak interval, v katerem se z dano verjetnostjo (ponavadi 95-odstotno) nahaja ocenjevani parameter.
Interpretacija: Z verjetnostjo tveganja α se parameter nahaja v tem intervalu.
Pišemo: [Zmin < α < Zmax] >= stopnja zaupanja

b) Povprečje populacije – velik vzorec:

Določanje intervala zaupanja za povprečje populacije, če imamo vzorec z več kot 20 enotami:

- določimo stopnjo zaupanja ɣ (običajno je že znan)
- Izračunamo c (tabela Standardna normalna oblika): $c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$
- Izračunamo a: $a = \frac{c \cdot \sigma}{\sqrt{n}} = \frac{c \cdot \hat{s}}{\sqrt{n}} = \frac{c \cdot \hat{s} \cdot \sqrt{\frac{N-n}{N-1}}}{\sqrt{n}}$
- interval zaupanja: I = [$\overline{X}$ - a , $\overline{X}$ + a]

c) Povprečje populacije – majhen vzorec:

Določanje intervala zaupanja za povprečje populacije, če imamo vzorec z največ 20 enotami:

- določimo stopnjo zaupanja ɣ (običajno je že znan)
- Izračunamo c (tabela Standardna normalna oblika): $c = t_{\frac{1+\gamma}{2}}(n-1)$
- Izračunamo a: $a = \frac{c \cdot \sigma}{\sqrt{n}} = \frac{c \cdot \hat{s}}{\sqrt{n}}$
- interval zaupanja: I = [$\overline{X}$ - a , $\overline{X}$ + a]

d) Standardni odklon populacije:

Najprej določimo stopnjo zaupanja ɣ, ki je običajno že znana. Potem pa s pomočjo formul izračunamo meje intervala:

$\chi^2_1 = \chi_{\frac{1-\gamma}{2}}(n-1)$ spodnja meja: $a = \frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_1}}$

$\chi^2_2 = \chi_{\frac{1+\gamma}{2}}(n-1)$ zgornja meja: $b = \frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_2}}$

Zdaj lahko določimo interval zaupanja za standardni odklon: I = [a , b]

e) Razlika povprečij – velik vzorec:

- določimo stopnjo zaupanja ɣ (običajno je že znan)
- določimo povprečje obeh vzorcev: $\overline{Y}_1$ (prvi vzorec) in $\overline{Y}_2$ (drugi vzorec)

$z_{\frac{\gamma}{2}} = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$ $H = z_{\frac{\gamma}{2}} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$

- interval zaupanja: I = [$\overline{Y}_1 - \overline{Y}_2 - H$, $\overline{Y}_1 - \overline{Y}_2 + H$]

f) Razlika povprečij – majhen vzorec:

- določimo stopnjo zaupanja ɣ (običajno je že znan)
- določimo povprečje obeh vzorcev: $\overline{Y}_1$ (prvi vzorec) in $\overline{Y}_2$ (drugi vzorec)

$$\sqrt{} = \left[\frac{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}{\left(\frac{\hat{S}_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{\hat{S}_2^2}{n_2}\right)^2/(n_2-1)} \right]$$

$$z_{\frac{\alpha}{2}} = t_{\frac{1+\gamma}{2}}(\sqrt{})$$

$$H = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

- interval zaupanja: $I = [\overline{Y}_1 - \overline{Y}_2 - H \ , \ \overline{Y}_1 - \overline{Y}_2 + H]$

g) Sklepanje iz vzorca na populaciji:

Vzorčna aritmetična sredina $\overline{X}$ je ocena populacijske aritmetične sredine μ. Vzorčno aritmetično sredino imenujemo tudi cenilka populacijske aritmetične sredine. Vrednost cenilke se od ocenjevanja parametra bolj ali manj odklanja. Rečemo, da je cenilka parametra dobra, če ima nekaj dobrih lastnosti:

Nepristranska cenilka – povprečje vseh vzorčnih ocen (matematično upanje cenilk) je enako ocenjevanemu parametru. Velja: $E(X) = \mu$.
Dosledna cenilka – z večanjem vzorca se vzorčna ocena bliža parametru.
Točkovna cenilka – pove, kako izračunati numerično oceno parametra populacije na osnovi merjenj vzorca.

h) Vzorčno povprečje:

Naj bo X spremenljivka na populaciji. Izberemo vzorec velikosti n, pridobimo vrednosti $X_1, X_2 \dots X_n$. Vzorčno povprečje je $\overline{X} = 1/n * (X_1 + X_2 + \dots + X_n)$.
Izrek: Matematično upanje vzorčnega povprečja je enako povprečni vrednosti (matematičnemu upanju) na celotni populaciji: $E(\overline{Y}) = E(Y)$.

Lastnosti:
Naj bo Y spremenljivka na populaciji velikosti N, $\mu = E(Y)$, $\sigma^2 = D(Y)$. Vzorčimo brez ponavljanja, vzorci velikosti n. Velja:
- $E(\overline{Y}) = E(Y) = \mu$
- Ne glede na to, ali je Y na populaciji normalno porazdeljena z $N(\mu, \sigma/\mathrm{Sqrt}[n])$, je $\overline{Y}$ vedno porazdeljena normalno z $N(\mu, \sigma/\mathrm{Sqrt}[n])$.

Trditev: Naj bo $X_1, X_2 \dots X_n$ naključni vzorec, ki je sestavljen iz n meritev populacije s končnim povprečjem in končnim standardnim odklonom . Potem sta povprečje in standardni odklon vzorčnega povprečja X enaka:

10. PREVERJANJE DOMNEV

a) Statistične hipoteze:

Statistična hipoteza je hipoteza o porazdelitvi slučajne spremenljivke. Statističen test je lahko:
- Parametričen: znan je tip porazdelitve, hipoteza govori o parametru
- Neparametričen: porazdelitev je neznana, hipoteza govori o vrsti porazdelitve.

Parametrični testi so enostranski in dvostranski.

b) Ničelna in alternativna hipoteza:

Ničelna hipoteza H_0 je trditev o lastnosti populacije, za katero predpostavljamo, da drži (verjamemo, da je ta trditev resnična). S pomočjo testov, pa poskušamo to hipotezo zavreči. Alternativna hipoteza H_A je trditev, ki nasprotuje ničelni hipotezi. S pomočjo testiranja poskušamo dokazati, da je ta trditev drži.
Primer:
Ničelna hipoteza - H_0: obtoženec je nedolžen , voznik je vozil po predpisih
Alternativna hipoteza - H_A: obtoženec je kriv, voznik je vozil prehitro

c) Postopek testiranja domneve:

Domnevo testiramo po točno določenem postopku, vrstni red je pomemben:
(1) Postavimo ničelno in alternativno domnevo, (2) Izberemo testno statistiko, (3) Določimo zavrnitveni kriterij, (4) Izberemo naključni vzorec, (5) Izračunamo vrednost na osnovi testne statistike, (6) Sprejmemo odločitev, (7) Naredimo ustrezen zaključek

d) Odločitev in zaključek:

Zaključek naredimo na osnovi rezultatov, ki jih pokaže test:
- obtoženca spoznamo za krivega, saj večina dokazov kaže, da je storil kaznivo dejanje
- obtoženca spoznamo za nedolžnega, saj nimamo dovolj dokazov, ki bi kazali, da je storil kaznivo dejanje

e) Napaka prve vrste:

Zavrnemo ničelno hipotezo, če je le ta pravilna (praktičen primer: nedolžnega spoznamo za krivega). Napako prve vrste označimo z α. Verjetnost, da naredimo napako 1. vrste je merljiva in jo lahko poljubno zmanjšamo.

f) Napaka druge vrste:

Zavrnemo ničelno hipotezo, če je le ta napačna (praktičen primer: krivega spoznamo za nedolžnega). Napako druge vrste označimo z β. Verjetnosti, da naredimo napako 2. vrste ni mogoče oceniti, zato teh napak ne delamo. To pomeni, da ničelnih hipotez nikoli ne sprejmemo.

g) Stopnja značilnosti:

Stopnja značilnosti je verjetnost napake 1. vrste (α), ki jo uporabljamo pri postopku preverjanja domnev (jo vnaprej določimo sami), in na njeni osnovi določimo kritično območje (območje zavračanja ničelne domneve). Če eksperimentalna vrednost testne statistike, ki smo jo izračunali iz vzorčnih podatkov, pade v kritično območje, ničelno domnevo zavrnemo in sprejmemo osnovno ob stopnji.
Interpretacija: Največja stopnja zavrnitve H_0, ki jo je vodja eksperimenta pripravljen sprejeti.

h) Stopnja tveganja:

Stopnja tveganja nam pove, s kolikšno verjetnostjo tveganja α se ocenjevani parameter γ nahaja v intervalu s spodnjo mejo Zmin in zgornjo mejo Zmax. Stopnjo tveganja določimo sami (ponavadi je α = 0.05).
Stopnja tveganja nam pove, kolikšna je verjetnost, da bomo naredili napako prve vrste.

i) Stopnja zaupanja:

Stopnja zaupanja (1-α) nam daje obratne informacije od stopnje tveganja. Torej nam pove, kolikšna je verjetnost, da napake prve vrste ne bomo naredili.

j) Pomen stopnje tveganja/zaupanja:

Za vsak slučajni vzorec lahko ob izbrani stopnji tveganja α izračunamo interval zaupanja za parameter γ. Ker se podatki vzorcev razlikujejo, se razlikujejo vzorčne ocene parametrov in zato tudi izračunani intervali zaupanja za parameter γ. To pomeni, da se intervali zaupanja od vzorca do vzorca razlikujejo. Pri 5% stopnji tveganja približno 95% intervalov pokrije parameter γ (oz. 5 intervalov zaupanja od 100 ne pokrije iskanega populacijskega parametra).

k) Zavrnitveni kriterij:

Zavrnitveni kriterij je interval (lahko si predstavljamo kot območje na grafu) za katerega velja, da če vrednost testne statistike pade znotraj njega, ničelno hipotezo zavrnemo. Hipotezo zavržemo ali sprejmemo na podlagi izračuna.

l) Kritično območje testa:

Kritično območje (ali tudi območje zavrnitve) je del zaloge vrednosti vzorcev, ki ga izberemo zato, da ničelno hipotezo zavrnemo, če se eksperimentalna vrednost vzorca nahaja v njem.
Interpretacija: Kritično območje testa je tisti del prostora parametrov, v katerem se mora nahajati eksperimentalna vrednost q_1, q_2 … q_n , da H_0 zavrnemo.

m) Enostranski in dvostranski test:

Primer:
Ničelna hipoteza H_0: verjetnost, da se bo rodil deček, je enaka 0,5
Dvostranski test – H_A: verjetnost, da se bo rodil deček, je različna od 0,5
Enostranski test – H_A: verjetnost, da se bo rodil deček, je manjša (ali večja) od 0,5

n) Moč zavrnitvenega testa:

Moč zavrnitvenega testa (1-beta) nam pove kolikšna je verjetnost, da zavrnemo ničelno hipotezo v primeru, ko je le-ta v resnici napačna.

o) P-vrednost:

P-vrednost (ali ugotovljena bistvena stopnja za določen statistični test) je verjetnost (ob predpostavki, da drži H_0), da ugotovimo vrednost testne statistike, ki je vsaj toliko v protislovju s H_0 in podpira H_A kot tisto, ki je izračunana iz vzorčnih podatkov.

p) Formalen postopek za preverjanje domnev:

- Postavi domnevi o parametrih (ničelno H0 in alternativno H1).
- Za parameter poiščemo kar se da dobro cenilko (npr. nepristransko) in njeno porazdelitev ali porazdelitev ustrezne statistike (izraz, v katerem nastopa cenilka).
- Določi odločitveno pravilo. Izberemo stopnjo značilnosti α. Na osnovi stopnje značilnosti in porazdelitve statistike določimo kritično območje.

- Zberi/manipuliraj podatke ter na vzorčnih podatkih izračunaj (eksperimentalno) vrednost testne statistike.
- Primerjaj in naredi zaključek.

Zaključek preverjanja:
- če eksperimentalna vrednost pade v kritično območje, ničelno domnevo zavrni in sprejmi osnovno domnevo ob stopnji značilnosti.
- če eksperimentalna vrednost ne pade v kritično obmocje, pa pravimo da vzorčni podatki kažejo na statistično neznačilne razlike med parametrom in vzorčno oceno.

r) Testiranje hipotez – formule:

Dvostranski test uporabimo, če: $H_A : \mu \neq \mu_0$
Enostranski test uporabimo, če: $H_A : \mu > \mu_0$ ali $\mu < \mu_0$

Dvostranski test uporabimo, če: $H_A : \mu \neq \mu_0$
Enostranski test uporabimo, če: $H_A : \mu > \mu_0$ ali $\mu < \mu_0$

$$U = \frac{\bar{x} - \mu_0}{\hat{s}} \sqrt{n}$$

Dvostranski test:

* Velik vzorec: $z_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ * Majhen vzorec: $z_\alpha = t_{1-\frac{\alpha}{2}}(n-1)$

$W_0 = (\infty , - Z\alpha] \cup [Z\alpha, \infty)$

Enostranski test:

* Velik vzorec: $z_\alpha = \Phi^{-1}(1 - \alpha)$ * Majhen vzorec: $z_\alpha = t_{1-\alpha}(n-1)$

če $\mu > \mu_0 : W_0 = [Z\alpha, \infty)$ ali če $\mu < \mu_0 : W_0 = (\infty , - Z\alpha]$

Zaključek:
* $U \in W_0$ ---> ovržemo ničelno hipotezo H_0, sprejmemo alternativno hipotezo H_A
* $U \notin W_0$ ---> ne ovržemo ničelne hipoteze H_0, ne sprejmemo alternativne hipoteze H_A

11. BIVARIATNA ANALIZA IN REGRESIJA

a) Regresijska analiza:

Regresijska funkcija Y 0 = f(X) kaže, kakšen bi bil vpliv spremenljivke X na Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y. Ker pa so ponavadi še drugi vplivi na proučevano spremenljivko Y, se točke, ki predstavljajo enote v razsevnem grafikonu, odklanjajo od idealne regresijske krivulje.

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

X = neodvisna spremenljivka, Y = odvisna spremenljivka, E = člen napake oz. motnja

b) Regresija:

Preslikavo X ---> E(Y|X) imenujemo regresija slučajne spremenljivke Y, glede na slučajno spremenljivko X.
Regresija je linearna in regresijska krivulja premic, ki gre skozi točko (σx, σy). Med X in Y ni linearne zveze, sta le „v povprečju“ linearno odvisni.

c) Regresijska premica:

Regresijska premica je enolično določena. Premici izračunamo s pomočjo regresijske funkcije, po pravilu:

Prva regresijska premica: $$Y = \mu_Y + \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2}(X - \mu_X)$$

Druga regresijska premica: $$X = \mu_X + \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}(Y - \mu_Y)$$

Regresijski premici se srečata v točki, določeni z aritmetičnima sredinama spremenljivke X in Y.

d) Metoda najmanjših kvadratov:

Želimo poiskati tako premico, ki se n točkam najbolj prilega. V tem primeru uporabimo metodo najmanjših kvadratov, s katero poiščemo oceni za regresijska parametra a in b. Parametra sta izbrana tako, da je vsota kvadratov napak modela najmanjša.
Regresijske koeficiente določimo po tej metodi tako, daje vsota kvadratov odklonov stvarnih vrednosti odvisne spremenljivke, y_i ,od ocenjenih vrednosti, y'_i, minimalna.

$$S = \sum_{i=1}^{N} (y_i - y_i')^2 = \sum_{i=1}^{N} e_i^2 \Rightarrow \min$$

e) Časovne vrste:

Časovna vrsta je niz istovrstnih podatkov, ki se nanašajo na zaporedne razmike ali trenutke. Podatke izmerimo na različnih časovnih točkah, ki so navadno določeni z intervali.

f) Trend:

Trend (dolgoročno gibanje) je sestavina dinamike v časovni vrsti. Trend Xt podaja dolgoročno smer razvoja. Običajno ga je mogoče izraziti s preprostimi rahlo ukrivljenimi krivuljami.

g) Koeficient variacije:

Koeficient variacije je statistični kazalec, ki prikazuje razpršitev statističnih enot okoli aritmetične sredine njihove statistične populacije. Definiran je kot razmerje med standardnim odklonom in aritmetično sredino; od standardnega odklona, ki prav tako prikazuje razpršenost statističnih enot, pa se razlikuje po tem, da je merjen v odstotkih in ga je zato moč uporabiti za primerjavo razpršenosti enot različnih statističnih populacij.

Koeficient variacije je izračunan po formuli: $$KV = \frac{100\sigma}{\overline{x}}$$

kjer je σ standardni odklon, $\overline{x}$ pa aritmetična sredina.

$$\mu_{\overline{X}} = \mu \qquad in \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

TABELA PORAZDELITEV:

Porazdelitev	Oznaka	Opis	$E(X)$	$D(X)$	Izvor
Bernoullijeva	$\mathrm{Be}(p)$	$P(X=0)=1-p$ $P(X=1)=p$	p	pq	Indikator dogodka
Binomska	$\mathrm{B}(n,p)$	$P(X=k)=\binom{n}{k}p^k q^{n-k}$	np	npq	Število uspešnih izidov v n neodvisnih poskusih; vsota n neodv. Bernoullijevih sl. spr.
Geometrijska	$\mathrm{G}(p)$	$P(X=k)=pq^{k-1}$ $k=1,2,\ldots$	$\frac{1}{p}$	$\frac{q}{p^2}$	Število poskusov do prvega uspešnega izida
Pascalova	$\mathrm{P}(n,p)$	$P(X=k)=\binom{k-1}{n-1}p^n q^{k-n}$ $k=n,\ n+1,\ \ldots$	$\frac{n}{p}$	$\frac{nq}{p^2}$	Število poskusov do n-tega uspešnega izida; vsota n neodv. geom. sl. spr.
Poissonova	$\mathrm{P}(\lambda)$	$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$ $k=0,1,\ldots$	λ	λ	Število telefonskih klicev, nesreč ipd. v določenem času
Hipergeometrijska	$\mathrm{H}(s;r,n)$ $\mathrm{H}(r;s,n)$	$P(X=k)=\frac{\binom{s}{k}\binom{n-s}{r-k}}{\binom{n}{r}}$	$\frac{rs}{n}$	$\frac{rs(n-r)(n-s)}{n^2(n-1)}$	Število rdečih kroglic v vzorcu velikosti s, če je v škatli skupaj n kroglic, od tega r rdečih
Enakomerna na točkah $x_1,\ldots x_n$	$\mathrm{E}(x_1\ldots x_n)$	$P(X=x_k)=\frac{1}{n}$	$\bar{x}:=\frac{\sum_{k=1}^{n}x_k}{n}$	$\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2=\frac{\sum_{k=1}^{n}x_k^2-n\bar{x}^2}{n}$	Vrednost srečke, izbrane povsem na slepo
Enakomerna na intervalu	$\mathrm{E}[a,b]$	$p_X(x)=\frac{1}{b-a},\ a\le x\le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	Faza periodičnega pojava
Normalna	$\mathrm{N}(\mu,\sigma)$	$p_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2	Če je X vsota veliko (vsaj 30) neodvisnih sl. spr., je približno $X\sim\mathrm{N}(\mu,\sigma)$, kjer je $\mu=E(X)$ in $\sigma=\sqrt{D(X)}$.
Standardizirana normalna	$\mathrm{N}(0,1)$	$p_X(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ $P(a<X<b)=\Phi(b)-\Phi(a)$	0	1	$X\sim\mathrm{N}(\mu,\sigma)\Rightarrow\frac{X-\mu}{\sigma}\sim\mathrm{N}(0,1)$
Eksponentna	$\mathrm{Exp}(\lambda)$	$p_X(x)=\lambda e^{-\lambda x},\ x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Čas prvega klica, življenjska doba radioaktivnega delca
Gama	$\mathrm{Gama}(n,\lambda)$	$p_X(x)=\frac{\lambda^n x^{n-1}e^{-\lambda x}}{\Gamma(n)}$ $x>0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	Za $n\in\mathbb{N}$: čas n-tega klica
Hi kvadrat	$\chi^2(n)=$ $\mathrm{Gama}(\frac{n}{2},\frac{1}{2})$	$p_X(x)=\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}$ $x>0$	n	$2n$	Vsota kvadratov n neodvisnih stand. normalnih slučajnih spremenljivk

Opomba: $q=1-p$.

KLJUČNE BESEDE:

1. Kombinatorika (permutacije, variacije, kombinacije s ponavljanjem in brez)
2. Popoln sistem dogodkov
3. Verjetnost in računanje z dogodki
4. Pogojna verjetnost
5. Dvofazni poskusi, formula za popolno verjetnost in Bayesov obrazec
6. Bernullijevo zaporedje neodvisnih poskusov in Laplaceov obrazec
7. Slučajne spremenljivke
8. Diskretna slučajna spremenljivka (enakomerna in binomska porazdelitev)
9. Zvezna slučajna spremenljivka (porazdelitvena funkcija, gostota verjetnosti)
10. Normalna porazdelitev (funkcija napake in standardizirana normalna porazdelitev), tudi večrazsežna gostota porazdelitve in opis kovariančne matrike
11. Matematično upanje (pričakovana vrednost) slučajne spremenljivke (kdaj obstaja)
12. Disperzija (razpršenost) slučajne spremenljivke in odklon
13. Standardizacija slučajne spremenljivke
14. Povezanost dveh številskih slučajnih spremenljivk
15. Kovarianca, Pearsonov koeficient korelacije
16. Funkcije slučajnih spremenljivk in slučajni vektorji (robna porazdelitvena funkcija, verjetnostna funkcija, neodvisnost, pogojne porazdelitve)
17. Sredine (aritmetična, geometrijska, harmonična, kvadratna)
18. Momenti, centralni limitni izrek (ni dovolj napisati, da gre za CLI, pač pa je potrebno razumeti kaj je centralni limitni zakon) in neenakost Čebiševa
19. Statistika (osnovni pojmi kot npr. mediana, kvantil, kvartil in kvartilni razmik, vrste spremenljivk, tipi analiz)
20. Opisna statistika (koraki statistične analize ter urejanje in prikazovanje podatkov, standardizacija)
21. Mere razpršenosti (povprečni absolutni odklon, varianca, standardni odklon)
22. Mere asimetrije in sploščenosti, momenti (centralni, začetni)
23. Relativne mere razpršenosti (relativni variacijski razmik, relativni kvartilni odklon, relativni povprečni absolutni odklon, koeficient variacije)
24. Porazdelitve vzorčnih statistik (aritmetičnih sredin, deležev, razlike aritm. sredin in razlike deležev)
25. Sklepanje iz vzorca na populacijo (vzorčne statistike, cenilke: nepristranske in dosledne)
26. Intervali zaupanja (za povprečje, odklon, delež, tudi njihove razlike oz. kvocijenti, majhen in velik vzorec)
27. Preverjanje domnev (tj. testiranje hipotez: alternativa, stopnja zaupanja/tveganja, napake, stopnja značilnosti testa, P-vrednost, kritično območje, postopek)
28. Ocenjevanje parametrov z majhnimi vzorci (tudi na nove vpeljane porazdelitve kot so Studentova, Fisherjeva in hi-kvadrat)
29. Regresija (regresijska premica z metodo najmanjših kvadratov)
30. Časovne vrste in določanje trenda