

Verjetnost je študij možnosti ali verjetja, da se bo nek dogodek zgodil.

Statistika preučuje podatke, jih zbira, klasificira, povzema, organizira, analizira in interpretira.

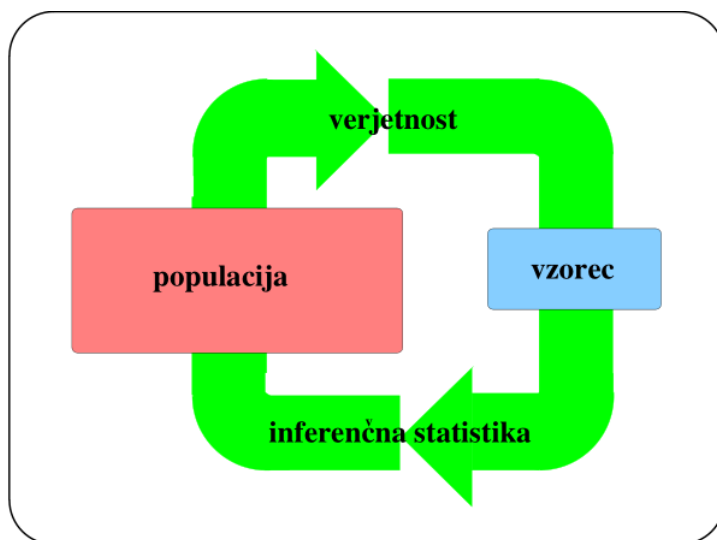
Glavni veji statistike:

Opisna statistika se ukvarja z organiziranjem, povzemanjem in opisovanjem zbirk podatkov (reduciranje podatkov na povzetke)

Analitična statistika jemlje vzorce podatkov in na osnovi njih naredi zaključke (inferenčnost) o populaciji (ekstrapolacija).

Populacija = vsi objekti ki jih opazujemo (npr. vsi registrirani glasovalci)

Vzorec je podmnožica populacije (100 registriranih glasovalcev)



Tipi podatkov:

- Kvantitativni (numerični) predstavljajo kvantiteto ali količino nečesa (npr. teža), interval (enaki intervali predstavljajo enake količine, poljubna ničla), razmerje (operacije seštevanja, odštevanja, množenja in deljenja, smiselna točka nič)
- Kvalitativni (kategorije) ni kvantitativnih interpretacij (npr. spol), nominalni (kategorije brez odgovarjajočega vrstnega reda/urejenosti), ordinalni/številski (kategorije z urejenostjo)

Frekvenca, npr. število zaposlenih (točna številka)

Relativna frekvenca, npr. delež zaposlenih (%)

Grafična predstavitev kvantitativnih podatkov: stolpčni graf, strukturni krog, histogrami, škatla z brki (box plot), steblo-list predstavitev (stem-and-leaf), zaporedje (dot plot) in runs plot(X,Y plot)

Urejeno zaporedje je zapis podatkov v vrsto po njihovi numerični velikosti (ustreznemu mestu pravimo rang)

Histogram -> izračunaj razpon podatkov -> razdeli razpon na 5 do 20 razredov enake širine -> za vsak razred preštej število vzorcev, ki spadajo v ta razred (frekvenca razreda) -> izračunaj vse relativne frekvence razredov

Modus (oznaka M_0) množice podatkov je tista vrednost, ki se pojavi z največjo frekvenco

Mediana (oznaka M_e) -> podatke uredimo v naraščajočem vrstnem redu -> če št. Podatkov liho, je mediana na sredini, če pa sodo, je mediana povprečje sredinskih dveh vrednosti

Povprečje populacije: $\mu = \frac{\sum_{i=1}^n x_i}{n}$

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

Razpon ali variacijski razmik je razlika med največjo in najmanjšo meritvijo v množici podatkov.

Centili -> 100p-ti centil (p je element [0,1]) je definiran kot število, od katerega ima 100p % meritev manjšo ali enako numerično vrednost.

Določanje 100p-tega centila: izračunaj vrednost $p(n + 1)$ in zaokroži na najbližje število. Naj bo to število enako i. Izmerjena vrednost z i-tim rangom je 100p-ti centil.

25. centil je tudi 1. kvartil, 50. centil je 2. kvartil ALI mediana, 75. centil je tudi 3. kvartil.

Varianca je kvadrat pričakovanega standardnega odklona populacije ali vsota kvadratov standardnih odklonov deljena s stopnjo prostosti vzorca.

Varianca populacije (končne populacije z N elementi):

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}.$$

Varianca vzorca (z n meritvami):

$$s^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1} = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n - 1}$$

Standardni odklon (deviacija) je pozitivni kvadratni koren variance.

Koeficient variacije je standardni odklon deljen s povprečjem.

Sredine:

- Aritmetična: $A_n = \frac{a_1 + \dots + a_n}{n}$
- Geometrična: $G_n = \sqrt[n]{a_1 \cdot \dots \cdot a_n}$
- Harmonična: $H_n = \frac{n}{\left(\frac{1}{a_1} + \dots + \frac{1}{a_n}\right)}$
- Kvadratna: $K_n = \sqrt{\frac{a_1^2 + \dots + a_n^2}{n}}$
- Potenčna: $P_{n,k} = \sqrt[k]{\frac{a_1^k + \dots + a_n^k}{n}}$

Normalna porazdelitev:

- Veliko podatkovnih množic ima porazdelitev približno zvonaste oblike (unimodalna oblika – en sam vrh)
- Če ima podatkovna množica porazdelitev približno zvonaste oblike, potem veljajo naslednja pravila:
 - o 68,3% vseh meritev leži na razdalji 1 standardnega odklona od povprečja
 - o 95,4% meritev leži do 2 standardnega odklona od njihovega povprečja
 - o Skoraj vse meritve, 99,7%, ležijo na razdalji 3 standardnih odklonov od povprečja
- Če je sprem. Približno normalno porazdeljena, potem jo povprečje in standardnih odklon zelo dobro opisujeta
- V primeru unimodalne porazdelitve sprem., ki pa je bolj asimetrična in bolj ali manj sploščena (koničasta), pa je potrebno izračunati še stopnjo asimetrije in sploščenosti.

l -ti centralni moment je: $m_l = \frac{(y_1 - \mu)^l + \dots + (y_N - \mu)^l}{N}$, kjer je $m_1 = 0$ in $m_2 = \sigma^2$

Koeficient asimetrije (s centralnimi momenti): $g_1 = m_3 / m_2^{3/2}$

Razlike med srednjimi vrednostnimi so tem večje, čim bolj je porazdelitev asimetrična:

$$KA_{M_0} = (\mu - M_0) / \sigma \text{ in pa } KA_{M_e} = 3(\mu - M_e) / \sigma$$

Koeficient sploščenosti: $K = g_2 = m_4 / m_2^2 - 3$, kjer:

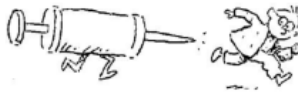
- $K = 3$ (ali 0) pomeni normalna porazdelitev zvonaste oblike (mesokurtic)
- $K < 3$ (ali negativna) bolj kopasta kot normalna porazdelitev, s krajšimi repi
- $K > 3$ (ali pozitivna) bolj špičasta kot normalna porazdelitev, z daljšimi repi

Standardizacija

- Vsaki vrednosti x_i sprem. X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ .
- $z_i = \frac{x_i - \mu}{\sigma}$
- Za novo sprem. Z bomo rekli, da je standardizirana, Z_i pa je standardizirana vrednost
- Potem je $\mu(Z) = 0$ in $\sigma(Z) = 1$

Funkcije/preslikave

Funkcija f iz množice A v množico B je predpis, ki vsakemu elementu iz množice A priredi natanko določen element iz množice B , oznaka $f : A \rightarrow B$.



Funkcija $f : A \rightarrow B$ je:

- **injektivna** (angl. one to one) če za $\forall x, y \in A$

$$x \neq y \Rightarrow f(x) \neq f(y),$$

- **surjektivna** (angl. on to), če za $\forall b \in B$

$$\exists a \in A, \text{ tako da je } f(a) = b.$$

Injektivni in surjektivni funkciji pravimo bijekcija. Množicama med katerima obstaja bijekcija pravimo bijektivni množici. Bijektivni množici imata enako število elementov (npr. končno, števno neskončno)

Permutacije

Permutacija elementov $1, \dots, n$ je bijekcija, ki slika iz množice $\{1, \dots, n\}$ v množico $\{1, \dots, n\}$.

Npr. permutacija kart je običajno premešanje kart – ko jih vrenem na kup (spremeni se vrstni red v kupu, karte pa ostanejo iste, nobene nismo ne dodali ne odvzeli).

Število permutacij n elementov, tj. razvrstitev n -tih različnih elementov, je enako

$$n! := 1 \cdot 2 \cdot \dots \cdot n$$

(oziroma definirano rekurzivno $n! = (n-1)!n$ in $0! = 1$).

Cikel je permutacija, za katero je

$$\pi(a_1) = a_2, \pi(a_2) = a_3, \dots, \pi(a_r) = a_1,$$

ostale elementi pa so fiksni (tj. $\pi(a) = a$).

Na kratko jo zapišemo z $(a_1 a_2 \dots a_r)$.

Trditev: Vsako permutacijo lahko zapišemo kot produkt disjunktnih ciklov.

Transpozicija je cikel dolžine 2. Vsak cikel pa je produkt transpozicij:

$$(a_1 a_2 a_3 \dots a_r) = (a_1 a_2) \circ (a_2 a_3) \circ \dots \circ (a_{r-1} a_r),$$

torej je tudi vsaka permutacija produkt transpozicij.

Seveda ta produkt ni nujno enolično določen, vseeno pa velja:

Trditev: Nobena permutacija se ne da zapisati kot produkt sodega števila in kot produkt lihega števila permutacij.



Permutacije s ponavljanjem

Permutacije s ponavljanjem so nekakšne permutacije, pri katerih pa ne ločimo elementov v skupinah s k_1, k_2, \dots, k_r elementi - zato delimo število vseh permutacij s številom njihovih vrstnih redov, tj. permutacij:

$$\frac{n!}{k_1! k_2! \dots k_r!}.$$

Koliko različnih besed lahko napišemo s črkami iz beseda ANANAS?

Za $k_1 = k_2 = \dots = k_r = 1$ dobimo običajne permutacije.

Za $r = 2$ dobimo v bistvu kombinacije.

Kombinacije

Binomski koeficient oz. število **kombinacij**, tj. število m -elementnih podmnožic množice moči n , je

$$\binom{n}{m} = \frac{n \cdot (n-1) \cdots (n-m+1)}{1 \cdot 2 \cdots m} = \frac{n!}{m!(n-m)!},$$

saj lahko prvi element izberemo na n načinov,
drugi na $n-1$ načinov, ...,
zadnji na $n-m+1$ načinov,
ker pa vrstni red izbranih elementov ni pomemben,
dobljeno število še delimo s številom permutacij.

Trditev: *Za binomske simbole velja*

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k \quad \text{in} \quad \binom{n}{m} + \binom{n}{m+1} = \binom{n+1}{m+1}.$$

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

Dogodek je pojav, ki v množico skupaj nastopajočih dejstev ne spada in se lahko v posameznem poskusu zgodi ali pa ne.

Dogodke označujemo z velikimi črkami iz začetka abecede (A,B,C,...), poskuse pa iz konca (X, Y,...)

Dogodek je lahko: gotov (ob vsaki ponovitvi poskusa se zgodi, G), nemogoč (nikoli se ne zgodi, N) ali slučajen (včasih se zgodi, včasih pa ne)

Dogodek A je poddogodek ali način dogodka B, kar zapišemo A podmnožica B, če se vsakič, ko se zgodi dogodek A, zagotovo zgodi dogodek B (pri metu kocke je dogodek A, da pade šest pik, način dogodka B, da pade sodo število pik.) Če je dogodek A način dogodka B in sočasno dogodek B način dogodka A, sta dogodka enaka. Vsota dogodkov A in B, označimo z $A + B$, se zgodi, če se zgodi vsaj eden od dogodkov A in B (Vsota dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je gotov dogodek.). Produkt dogodkov A in B, označimo z AB , se zgodi, če se zgodita dogodka A in B hkrati (Produkt dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je nemogoč dogodek.). Dogodku A nasproten dogodek \bar{A} imenujemo negacijo dogodka A. Dogodka A in B sta nezdružljiva, če se ne moreta zgoditi hkrati, njun produkt je torej nemogoč dogodek (N). Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A sestavljen dogodek. Dogodek, ki ni sestavljen, imenujemo osnoven ali elementaren dogodek (Pri metu kocke je šest osnovnih dogodkov: E1, da pade 1 pika, E2, da padeta 2 piki, ..., E6, da pade 6 pik. Dogodek, da pade sodo število pik je sestavljen dogodek iz treh osnovnih dogodkov (E2, E4 in E6)). Množico dogodkov $S = \{A_1, A_2, \dots, A_n\}$ imenujemo popolni sistem dogodkov, če se v vsaki ponovitvi poskusa zgodi natanko eden izmed dogodkov iz množice S (pomeni da vsak dogodek ni nemogoč, so paroma nezdružljivi (produkt dveh dogodkov je nemogoč dogodek) in njihova vsota (skupna) je gotov dogodek).

Ponovitve poskusa, v katerih se dogodek A zgodi, imenujemo ugodne za dogodek A, število $f(A) = k/n$ pa je relativna frekvenca (pogostost) dogodka A v opravljenih poskusih. Če poskus X dolgo ponavljamo, se relativna frekvenca slučajnega dogodka ustali in sicer skoraj zmeraj toliko bolj, kolikor več ponovitev poskusa napravimo. **Verjetnost dogodka A v danem poskusu je število $P(A)$, pri katerem se navadno ustali relativna frekvenca dogodka A v velikem številu ponovitev tega poskusa. (STATISTIČNA DEFINICIJA VERJETNOSTI)**

Osnovne lastnosti:

- Ker je relativna frekvenca vedno nenegativna $\rightarrow P(A) \geq 0$
- $P(\Omega) = 1$, $P(\emptyset) = 0$, če A podmnožica B $\rightarrow P(A) \leq P(B)$
- Če sta A in B nezdružljiva, velja $P(A + B) = P(A) + P(B)$

Klasična definicija verjetnosti:

Verjetnosti prostor S slučajnega pojava je množica vseh možnih izidov.

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

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

Če je nek dogodek A sestavljen iz r dogodkov iz tega popolnega sistema dogodkov, potem je njegova verjetnost $P(A) = r/s$.

Geometrijska verjetnost: verjetnost sestavljenega dogodka kot razmerje dolžin dela, ki ustreza ugodnim izidom, in dela, ki ustreza možnim izidom.

Za dogodka A in B velja tudi: $P(A + B) = P(A) + P(B) - P(AB)$, pri dogodkih A, B, C pa: $P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$ (pravilo o vključitvi in izključitvi za množice)

Negacija: $P(\bar{A}) \equiv 1 - P(A)$

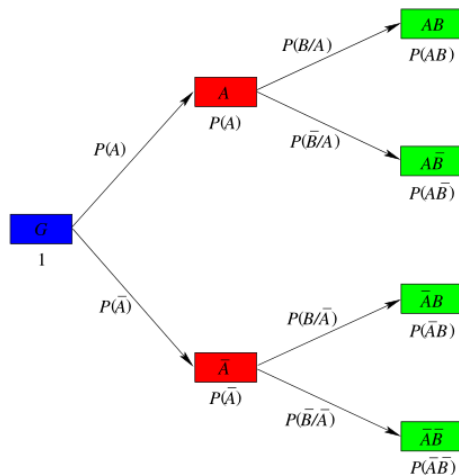
Če so dogodki A_i , i je element I paroma nezdružljivi: $P(\sum_{i \in I} A_i) = \sum_{i \in I} P(A_i)$

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

Relativna frekvenca A v opravljenih ponovitvah poskusa X': $P(A / B) = \frac{P(AB)}{P(B)}$

Grafično (oz. tekstovno 😊): $P(B) \rightarrow P(A/B) \rightarrow P(AB)$

Par formul: $P(AB) = P(B) * P(A/B)$ in $P(AB) = P(A) * P(B/A)$, velja: $P(A) * P(B/A) = P(B) * P(A/B)$



Verjetnost vsakega izmed dogodkov $AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}$ je enaka produktu verjetnosti na puščicah od začetka (koren od G) pa do samega dogodka)

Dogodka A in B sta neodvisna, če velja: $P(AB) = P(A) \cdot P(B)$, zato za neodvisna dogodka A in B velja $P(A/B) = P(A)$, za nezdružljiva dogodka A in B pa velja $P(A/B) = 0$.

Dogodka A in B sta neodvisna, če je $P(A/B) = P(A/\bar{B})$, nadalje velja: $P(ABC) = P(A) \cdot P(B/A) \cdot P(C/AB)$

Naj bo H_i razbitje gotovega dogodka G, hkrati pa paroma nezdružljivi. Zanima nas verjetnost dogodka A, če poznamo verjetnost $P(H_i)$ in pogojno verjetnost $P(A/H_i)$. Ker so dogodki AH_i paroma nezdružljivi, velja: $P(A) = \sum_{i \in I} P(AH_i) = \sum_{i \in I} P(H_i) \cdot P(A/H_i)$

V prvem koraku se zgodi natanko eden od dogodkov H_i , ki ga imenujemo domneva (hipoteza – sestavljajo popoln sistem dogodkov), dogodek A je eden izmed mogočih dogodkov na drugi stopnji.

Včasih nas zanima po uspešnem izhodu tudi druge stopnje, verjetnost tega, da se je na prvi stopnji zgodil dogodek H_i , odgovor dobimo preko Bayes-ovega obrazca: $P(H_k/A) =$

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

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 \bar{A} z verjetnostjo $P(\bar{A}) = 1 - P(A) = 1 - p = q$.

Zanima nas kolikokrat se v n zaporednih poskusih zgodi dogodek A natanko k-krat. To pomeni, da se nasprotni dogodek izvede (n-k)-krat. Zvezi pravimo Bernoullijev obrazec:

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ uporaba rekurzije: } P_n(k) = \frac{(n-k+1)p}{kq} P_n(k-1), \text{ za } k = 1, \dots$$

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

Točkovni obrazci:

- De Moivreov točkovni obrazec: $P_n(k) \approx \frac{1}{\sqrt{\pi n/2}} \cdot e^{-\frac{(k-(n/2))^2}{n/2}}$, ki je poseben primer Laplaceovega točkovnega obrazca, ki ga smemo uporabljati, ko je p blizu $1/2$:

$$P_n(k) \approx \frac{1}{\sqrt{2\pi \cdot n \cdot p \cdot q}} \cdot e^{-\left(\frac{(k-(n/2))^2}{2npq}\right)}$$

Slučajne spremenljivke in porazdelitve:

Imamo poskus, katerega izidi so števila (npr. pri metu kocke so izidi števila pik). Se pravi, da je poskusom prirejena neka količina, ki more imeti različne vrednosti. Torej je spremenljivka. Katere od mogočih vrednosti zavzame v določeni ponovitvi poskusa, je odvisno od slučaja. Zato ji rečemo slučajna spremenljivka. Potrebno je vedeti: kakšne vrednosti more imeti (zaloga vrednosti) in kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti. Predpis, ki določa te verjetnosti, imenujemo porazdelitveni zakon.

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami iz konca abecede, vrednosti spremenljivke pa z enakimi malimi črkami. Tako je npr. ($X = x$) dogodek, da slučaj. Spremlj. X zavzame vrednost x . Porazdelitveni zakon slučajne spremlj. X je poznan, če je mogoče za vsako realno število x določiti verjetnost: $F(x) = P(X \leq x)$. $F(x)$ imenujemo porazdelitvena funkcija.

Najpogosteje uporabljamo naslednji vrsti slučaj. Spremlj.:

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

Lastnosti porazdelitvene funkcije:

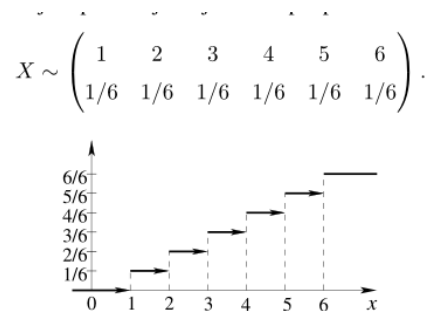
1. Funkcija F je definirana na vsem \mathbb{R} in $0 \leq F(x) \leq 1, \forall x \in \mathbb{R}$.
2. Funkcija F je ne padajoča: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$.
3. $F(-\infty) := \lim_{x \rightarrow -\infty} F(x) = 0$ in $F(\infty) := \lim_{x \rightarrow \infty} F(x) = 1$.
4. Funkcija je v vsaki točki z desne zvezna $F(x+) := \lim_{0 \leq h \rightarrow 0} F(x+h) = F(x)$.
5. Funkcija ima lahko v nekaterih točkah skok. Vseh skokov je največ števno mnogo.
6. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$.
7. $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1+)$.
8. $P(X > x) = 1 - F(x)$.
9. $P(X = x) = F(x) - F(x-)$.

Diskretne slučajne spremenljivke: Zaloga vrednosti je števna množica (x_1, x_2, \dots, x_n), torej je tudi števno neskončna, kot pri množici naravnih ali celih števil. Dogodki $X = x_k, k = 1, 2, \dots$ sestavljajo popoln sistem dogodkov. Posamezna verjetnost dogodka je enaka $P(X = x_i) = p_i$. Vsota verjetnosti vseh dogodkov je enaka 1 ($p_1 + p_2 + \dots + p_n = 1$).

Verjetnostna tabela prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti x_i , pod njimi pa pripisane pripadajoče verjetnosti p_i : $X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$.

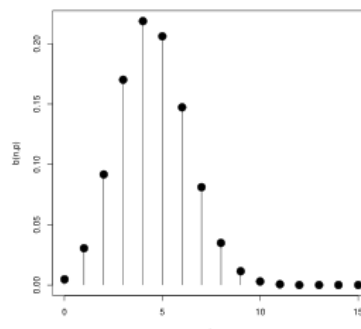
Porazdelitvena funkcija je v tem primeru: $F(x_k) = P(X \leq x_k) = \sum_{i=1}^k p_i$

Enakomerna diskretna porazdelitev: končna diskretna slučajna sprem. Se porazdeljuje enakomerno, oznaka $U(n)$, kjer je n velikost zaloge vrednosti, če so vse njene vrednosti enako verjetne.



Binomska porazdelitev: ima zalogo vrednosti $\{0, 1, \dots, n\}$ in verjetnosti, ki jih računamo po Bernoullijevemu obrazcu:

$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}$. Binomska porazdelitev je natanko določena z dvema podatkom/parametroma: n in p . Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p , zapišemo: $X \sim B(n, p)$, $E(X) = np$



Pričakovana vrednost $E(X)$ je posplošitev povprečne vrednosti diskretne spremenljivke X , tj:

$$\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^m x_i k_i = \sum_{i=1}^m x_i f_i, \text{ od koder izhaja: } E(x) = \sum_{i=1}^m x_i p_i$$

Poissonova porazdelitev: definiramo slučaj. Spremem. X kot X = število pojavitev nečesa na nek časovni interval, pri tem velja: vsak časovni interval je enak kot vsak drug časovni interval in pa da so si časovni intervali neodvisni (če je v enem časovnem intervalu veliko pojavitev nečesa, še ne pomeni da bo tudi v naslednjem). $E(X) = \lambda$. Najprej uporabimo binomsko porazdelitev $B(n, p)$, kjer je n število ponovitev poskusa (v našem primeru je enako številu manjših časovnih enot), verjetnost posameznega dogodka p pa je enaka verjetnosti, da je v dani manjši časovni enoti prišel mimo vsaj en avto. Izpeljali smo Poissonov obrazec, kjer za velike n in majhne verjetnosti

tj. p blizu 0 velja: $P_n(k) \approx \frac{(np)^k \cdot e^{-np}}{k!}$. Poissonova porazdelitev $P(\lambda)$ izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se ti dogodki pojavijo s poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek. Poissonovo porazdelitev lahko uporabimo tudi za število dogodkov v drugih intervalih, npr. razdalja, prostornina, ... Ima zalogo vrednosti $\{0, 1, 2, \dots\}$, njena verjetnostna funkcija pa je $p_k =$

$P(X = k) = \lambda^k \cdot \frac{e^{-\lambda}}{k!}$, kjer je $\lambda > 0$ dani parameter in predstavlja pričakovano pogostost nekega dogodka. $p_{k+1} = \frac{\lambda}{k+1} p_{k, p_0} = e^{-\lambda}$. Vidimo da zaloga vrednosti ni omejena, kar je bistvena razlika v primerjavi z binomsko, kjer število uspehov seveda ne more presegati števila Bernoullijevih poskusov n .

Pascalova porazdelitev: (oz. negativna binomska porazdelitev, $\text{negBin}(m, p)$). Ima zalogo vrednosti $\{m, m+1, m+2, \dots\}$, verjetnostna funkcija pa je: $p_k = \binom{k-1}{m-1} \cdot p^m \cdot q^{(k-m)}$, za $k \geq m$, kjer je $0 < p < 1$ dani parameter/verjetnost dogodka A v posameznem poskusu. Opisuje porazdelitev potrebnega števila poskusov, da se dogodek A zgodi m-krat. Če številu poskusov sledimo s slučajno sprem. X, potem verjetnost $P(X = k)$, da se bo pri k ponovitvah poskusa dogodek A zgodil v zadnjem poskusu ravno m-tič, izračunamo po zgornji formuli za p_k .

Za $m = 1$ dobimo geometrijsko porazdelitev $G(p)$, le-ta opisuje porazdelitev števila poskusov, da se dogodek A v zadnji ponovitvi poskusa zgodi prvič.

Hipergeometrijska porazdelitev $H(n, M, N)$: bolj splošno, naj bo v posodi $M = R$ rdečih in $N - M = B$ belih kroglic, kjer R, B je element naravnih števil in zato $N \geq M$. Zanima nas verjetnost, da je med n je element naravnih števil izbranimi kroglicami natanko k rdečih, če izbiramo n-krat brez vračanja. $P_k = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = \frac{\binom{R}{k} \binom{B}{n-k}}{\binom{R+B}{n}}$. Za zalogo vrednosti Hipergeometrijske porazdelitve bomo vzeli podmnožico množice naravnih števil, kjer je definirana zgornja verjetnostna funkcija. Veljajo še naslednje omejitve: $\max(0, n - B) \leq k \leq \min(M, n)$ in $n \leq N = R + B$.

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

Zvezne slučajne spremenljivke so zvezno porazdeljene, če obstaja taka integrabilna funkcija p, imenovana gostota verjetnosti, da za vsak $x \in \mathbb{R}$ velja: $F(x) = P(X \leq x) = \int_{-\infty}^x p(t) dt$, kjer $p(x) \geq 0$. To verjetnost si lahko predstavimo tudi grafično v koordinatnem sistemu, kjer na abscisno os nanašamo vrednosti slučajne spremenljivke, na ordinatno pa gostoto verjetnosti $p(x)$. Verjetnost je tedaj predstavljena kot ploščina pod krivuljo, ki jo določa $p(x)$. Velja: $\int_{-\infty}^{\infty} p(x) dx = 1$ in $P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} p(t) dt$ ter $p(x) = F'(x)$.

Enakomerna porazdelitev zvezne slučajne spremenljivke: verjetnostna gostota enakomerno porazdeljene zvezne slučajne spremenljivke $U[a, b]$ je: $p_x = \frac{1}{b-a}$, $a \leq X < b$, ter 0 drugod. Grafično si jo predstavljamo kot pravokotnik nad intervalom (a, b) višine $\frac{1}{b-a}$.

Lastnosti določenega integrala:

$$1) \int_a^b f(x) dx = - \int_b^a f(x) dx.$$

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

3) Naj bo $c \in [a, b]$

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

4) Naj bo $f(x) \geq g(x)$, $x \in [a, b]$,

$$\text{potem velja} \quad \int_a^b f(x) dx \geq \int_a^b g(x) dx.$$



Normalna porazdelitev: zaloga vrednosti normalno porazdeljene slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je: $p(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$. Porazdelitev je natanko določena z parametroma μ in σ , zapišemo: $X \sim N(\mu, \sigma)$

Funkcija napake $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{1}{2}t^2} dt$, je liha, zvezno odvedljiva, strogo naraščajoča funkcija. $\phi(\text{infinity}) = \frac{1}{2}$, $\phi(0) = 0$ in $\phi(-\text{infinity}) = -\frac{1}{2}$ ter $P_n(k_1, k_2) \sim \phi(xk_2) - \phi(xk_1)$. V sklopu normalne porazdelitve imamo: $P(x_1 \leq X < x_2) = \phi\left(\frac{x_2-\mu}{\sigma}\right) - \phi\left(\frac{x_1-\mu}{\sigma}\right)$. Porazdelitev $N(0, 1)$ je standardizirana normalna porazdelitev. Spremenljivko $X : N(\mu, \sigma)$ pretvorimo z: $z = \frac{x-\mu}{\sigma}$ v standardizirano spremenljivko $Z : N(0, 1)$.

Laplace: iz Laplace-ovega točkovnega obrazca izhaja, da za p blizu $\frac{1}{2}$ in velike n velja: $B(n, p) \approx N(np, \sqrt{npq})$.

Laplace-ov intervalski obrazec: $P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx$

Bernoulijev zakon velikih števil: naj bo k frekvenca dogodka A v n neodvisnih ponovitvah danega poskusa, v katerem ima dogodek A verjetnost p . Tedaj velja: $P\left(\left|\frac{k}{n} - p\right| \leq \varepsilon\right) \approx 2\phi\left(\varepsilon \cdot \sqrt{\frac{n}{p \cdot q}}\right)$ (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?) ($n = 3600$, $p = \frac{1}{2}$, $\varepsilon = 0,01$).

Porazdelitev Poissonovega toka, eksponentna: čas med dvema zaporednima Poissonovega dogodkoma. Gostota eksponentne porazdelitve je enaka: $p(x) = \lambda e^{-\lambda x}$, $x \geq 0$; porazdelitvena funkcija pa: $F(x) = \int_0^x \lambda \cdot e^{-\lambda t} dt = 1 - e^{-\lambda x}$

Porazdelitev Gama: naj bosta $b, c > 0$. Tedaj ima porazdelitev Gama $\Gamma(b, c)$ gostoto: $p(x) = \frac{c^b}{\Gamma(b)} \cdot x^{b-1} \cdot e^{-c \cdot x}$ in $p(x) = 0$ za $x \leq 0$. Funkcijo Gama lahko definiramo z določenim integralom za $\text{Re}(z) > 0$: $\Gamma(z) = \int_0^\infty t^{z-1} \cdot e^{-t} dt = 2 \cdot \int_0^\infty e^{-t^2} \cdot t^{2z-1} dt$. Torej je $\Gamma(1) = 1$.

Porazdelitev hi-kvadrat: je poseben primer porazdelitve gama: $\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, (n je element naravnih števil je število prostostnih stopenj) in ima gostoto: $p(x) = \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}$, kjer je $x > 0$ in 0 sicer. Hi-kvadrat test za ugotavljanje kategoričnih spremenljivk uporabimo v dveh podobnih (a različnih primerih): kako dobro se opazovana/izmerjena porazdelitev prilega pričakovani porazdelitvi (kvaliteta prilagoditve) ter ocenjevanje ali sta obe naključni spremenljivki neodvisni. Test hi-kvadrat napravi dvojce: dejanskim in teoretičnim frekvencom

priredi število, s katerim merimo odstopanje frekvenc. Čim večje je dobljeno število, tem večje je odstopanje. Za odstopanje dopuščamo dve razlagi: lahko da gre za slučajno odstopanje ali pa gre še za sistematično odstopanje (torej teoretične frekvence ne ustrezajo dejanski porazdelitvi).

$\chi^2(n-1) = \frac{(E_1-O_1)^2}{E_1} + \dots + \frac{(E_n-O_n)^2}{E_n}$, kjer so E -ji teoretične absolutne frekvence in O -ji dejanske absolutne frekvence. To pomeni, da se je pri N ponovitvah poskusa izid i dogodil O_i -krat, medtem ko smo pričakovali, da se zgodi E_i -krat. (Denimo, da ima poskus 6 izidov in je vrednost hi-kvadrat enaka 12.7. Število prostostnih stopenj je 5. Pogledamo v vrstico s petimi prostostnimi stopnjami in vidimo, da leži 12.7 med 11.1 in 15.1. Verjetnost, da so odstopanja med dejanskimi in teoretičnimi frekvencami zgolj slučajna, je manj kot 5% in več kot 1%.)

Cauchyeva porazdelitev z gostoto $p(x) = \frac{a}{\pi} \cdot \frac{1}{1+a^2(x-b)^2}$, $-\infty < x < \infty$, $a > 0$ ima porazdelitveno

$$\text{funkcijo: } F(x) = \frac{a}{\pi} \cdot \int_{-\infty}^x \frac{1}{1+a^2(x-b)^2} dx = \frac{1}{\pi} \cdot \arctg(a(x-b)) + \frac{1}{2}$$

Slučajni vektorji in neodvisnost slučajnih spremenljivk:

Verjetnostna funkcija $P(X=x, Y=y) = p(x, y)$ je definirana z 2D tabelo. Pri tem je: $P(X=x_i) = \sum_{k=1}^n p_{x_i y_k}$, enako velja za Y . Spremenljivki X in Y sta neodvisni, če za vsako celico v 2D tabeli velja: $P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$. Vsaka celica v tabeli ima verjetnost: $0 \leq p \leq 1$.

$P(A+B) = P(A) + P(B)$, za $AB = \emptyset$; $P(AB) = P(A) \cdot P(B)$ NEODVISNOST; $E(X+Y) = E(X) + E(Y)$ VEDNO; $E(XY) = E(X) \cdot E(Y)$ NEKORELIRANOST.

Slučajni vektor je n -terica sluč. Spremlj. $X = (X_1, \dots, X_n)$. Opišemo ga s porazdelitveno funkcijo:

$F(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$, pri čemer slednja oznaka pomeni: $P(\{X_1 \leq x_1\} \cap \dots \cap \{X_n \leq x_n\})$, za katero velja: $0 \leq F(x_1, \dots, x_n) \leq 1$. Funkcija F je za vsako spremenljivko napadajoča in z desne zvezna. $F(-\infty, \dots, -\infty) = 0$ in $F(\infty, \dots, \infty) = 1$.

Funkciji $F_i(x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty)$ pravimo robna porazdelitvena funkcija sprem. X_i

Diskretne večrazsežne porazdelitve – polinomska porazdelitev: $P_{(n; p_1, p_2, \dots, p_r), \sum p_i = 1, \sum k_i = n}$ je določena s predpisom: $P(X_1 = k_1, \dots, X_r = k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r}$. Koeficient šteje permutacije s ponavljanjem. Za $r = 2$ dobimo binomsko porazdelitev, tj. $B(n, p) = P(n; p, q)$. Koeficient r je število razredov. (primer: kup 52 igralnih kart, vlečemo eno in vrnemo nazaj, ponovimo 5x, verjetnost da dobimo 2x srce in 1x po pik, križ in karo? $r = 4$, $n = 5$, $p_1 = p_2 = p_3 = p_4 = 1/4$, zanima nas $P(X_1 = 1, X_2 = 2, X_3 = 1, X_4 = 1)$).

Lastnosti dvojnega integrala

1) Če je $f(x, y) \leq 0 \quad \forall (x, y) \in R$, je vrednost dvojnega integrala negativna.

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

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

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

$$\iint_R f(x, y) dx dy \leq \iint_R g(x, y) dx dy.$$

Dvojni integral predstavlja

Prostornina telesa med ploskvijo podano z $z = f(x, y)$, in ravnino $z = 0$ je enaka dvojnemu integralu

$$\iint_R f(x, y) dx dy,$$

ki ga izračunamo z uporabo dvakratnega integrala

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

prostornino pod ploskvijo.

Zvezne večrazsežne porazdelitve: slučajni vektor $X = (X_1, X_2, \dots, X_n)$ je zvezno porazdeljen, če obstaja integrabilna funkcija (gostota verjetnosti) $p(x_1, x_2, \dots, x_n) \geq 0$ z lastnostjo:

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

in

$$F(\infty, \infty, \dots, \infty) = 1.$$

Zvezne dvorazsežne porazdelitve

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

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

Kjer je p zvezna je

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

Robni verjetnostni gostoti sta

$$p_X(x) = F'_X(x) = \int_{-\infty}^{\infty} p(x, y) dy,$$

$$p_Y(y) = F'_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx.$$

Večrazsežna normalna porazdelitev

V dveh razsežnostih $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ ima gostoto

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

V splošnem pa jo zapišemo v matrični obliki

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

kjer je A simetrična pozitivno definitna matrika.

Vse robne porazdelitve so normalne.

Neodvisnost slučajnih spremenljivk: Slučajne spremenljivke X_1, X_2, \dots, X_n so med seboj neodvisne, če za poljubne vrednosti $x_1, x_2, \dots, x_n \in \mathbb{R}$ velja $F(x_1, x_2, \dots, x_n) = F_1(x_1) \cdot F_2(x_2) \cdot \dots \cdot F_n(x_n)$

Trditev: Če sta X in Y diskretni slučajni spremenljivki in p_{ij} verjetnostna funkcija slučajnega vektorja (X, Y) , potem sta X in Y neodvisni natanko takrat, ko je $p_{ij} = p_i q_j$ za vsak par i, j .

Trditev: Če sta X in Y zvezno porazdeljeni slučajni spremenljivki z gostotama $p_X(x)$ in $p_Y(y)$ ter je $p(x, y)$ gostota zvezno porazdeljenega slučajnega vektorja (X, Y) , potem:

$$X \text{ in } Y \text{ sta neodvisni} \iff p(x, y) = p_X(x) \cdot p_Y(y) \quad \forall x, y \in \mathbb{R}.$$

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

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

Torej sta komponenti X in Y neodvisni.

Če je $\rho = 0$ sta X in Y neodvisni. Zvezno porazdeljeni slučajni spremenljivki X in Y sta neodvisni natanko takrat, ko lahko gostoto verjetnosti slučajnega vektorja (X, Y) zapišemo v obliki $p(x, y) = f(x) \cdot g(y)$.

Funkcije slučajnih spremenljivk/vektorjev in pogojne porazdelitve:

Naj bo $X: G \rightarrow \mathbb{R}$ slučajna spremenljivka in $f: \mathbb{R} \rightarrow \mathbb{R}$ neka realna funkcija. Tedaj je njen kompozitum $Y = f \circ X$ določen s predpisom $Y(e) = f(X(e))$, za vsak e je element G , določa novo preslikavo $Y: G \rightarrow \mathbb{R}$. V ta namen mora biti za vsak y je element \mathbb{R} množica:

$(Y \leq y) = \{e \in G : Y(e) \leq y\} = \{e \in G : X(e) \in f^{-1}((-\infty, y])\}$ dogodek, torej v D . Če je to res, imenujemo Y funkcija slučajne spremenljivke X in jo zapišemo kar $Y = f(X)$. Njena porazdelitvena funkcija je: $F_Y(y) = P(Y \leq y)$. Kakšna mora biti množica $f^{-1}((-\infty, y])$, da je množica v D ? Mora biti Borelova množica (ali so intervali, ali števne unije intervalov, ali števnih preseki števnih unij intervalov), vsekakor pa ko je f zvezna funkcija. $F_Y = F_X \circ f^{-1} \rightarrow F_Y(y) = P(Y \leq y) = P(f(X) \leq y) = P(X \leq f^{-1}(y)) = F_X(f^{-1}(y))$

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

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

$F_{Y_j}(y) = P(Y_j \leq y) = P(f_j(X) \leq y) = P(X \in f_j^{-1}(-\infty, y])$ in če je X zvezno porazdeljen z gostoto $p(x_1, \dots, x_n)$, potem je:

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

Če sta spremenljivki X in Y neodvisni dobimo naprej zvezo:

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

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

Če so X_1, X_2, \dots, X_n neodvisne standardizirane normalne slučajne spremenljivke, je slučajna spremenljivka $Y = X_1^2 + X_2^2 + \dots + X_n^2$ porazdeljena po $\chi^2(n)$ (hi-kvadrat).

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

Transformacije: Jacobijeva determinanta: za gostoto $q(u, v)$ vektorja (U, V) dobimo od tu: $q(u, v) = p(x(u, v), y(u, v)) * |J(u, v)|$.

Pogojne porazdelitve: naj bo B nek mogoč dogodek, tj. $P(B) > 0$. Potem lahko vpeljemo pogojno porazdelitveno funkcijo: $F(x | B) = P(X \leq x | B) = P(X \leq x, B) / P(B)$. V diskretnem primeru je: $p_{ik} = P(X = x_i, Y = y_k) / P(B) = P(Y = y_k) = q_k$. Tedaj je pogojna porazdelitvena funkcija:

$$\begin{aligned} F_X(x | y_k) &= F_X(x | Y = y_k) = P(X \leq x | Y = y_k) = \\ &= \frac{P(X \leq x, Y = y_k)}{P(Y = y_k)} = \frac{1}{q_k} \sum_{x_i \leq x} p_{ik} \end{aligned}$$

Vpeljimo **pogojno verjetnostno funkcijo** s $p_{i|k} = \frac{p_{ik}}{q_k}$.

Tedaj je $F_X(x | y_k) = \sum_{x_i \leq x} p_{i|k}$.

Primer: zapiši pogojno verjetnostno porazdelitev slučajne sprem. X glede na pogoj $Y = 2$: verjetnosti v vrstici pri $Y = 2$ moramo deliti s $P(Y = 2)$.

Zvezne pogojne porazdelitve: postavimo $B = (y < Y \leq y + h)$ za $h > 0$ in zahtevajmo $P(B) > 0$. Če obstaja limita (za $h \rightarrow 0$) jo imenujemo pogojna porazdelitvena funkcija slučajne spremenljivke X

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

glede na **Gostota zvezne pogojne porazdelitve**

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

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

oziroma, če vpeljemo **pogojno gostoto**

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

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

V primeru dvorazsežne normalne porazdelitve dobimo

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

Momenti in kovarianca:

Pričakovana vrednost $E(X)$ (matematično upanje) je poslošitev povprečne vrednosti diskretne spremenljivke X , tj. $\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^m x_i k_i = \sum_{i=1}^m x_i f_i$, od koder izhaja $E(x) = \sum_{i=1}^m x_i p_i$.

Diskretna slučajna spremenljivka X z verjetnostno funkcijo p_k ima pričakovano vrednost $E(x) = \sum_{i=1}^m x_i p_i$, če je $\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty$. Zvezna slučajna spremenljivka X z gostoto $p(x)$ ima pričakovano vrednost: $E(x) = \int_{-\infty}^{\infty} x \cdot p(x) dx$, če je: $\int_{-\infty}^{\infty} |x| \cdot p(x) dx < \infty$.

Primeri slučajnih spremenljivk, za katere pričakovana vrednost ne obstaja:

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

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

Lastnosti pričakovane vrednosti: naj bo a realna konstanta. Če je $P(X = a) = 1$, je $E(X) = a$. Velja, da je: $|E(X)| \leq E(|X|)$. Splošno: $E(f(X))$ obstaja in je v diskretnem primeru enaka: $\sum_{i=1}^m f(x_i) p_i$, v zveznem pa $\int_{-\infty}^{\infty} f(x) \cdot p(x) dx$. $E(aX) = a \cdot E(X)$. Če imata slučajni spremenljivki X in Y pričakovano vrednost, ga ima tudi njuna vsota $X + Y$ in velja $E(X + Y) = E(X) + E(Y)$. Torej je pričakovana vrednost E linearen funkcional, tj. $E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$. Če sta slučajni spremenljivki neodvisni: $E(XY) = E(X) \cdot E(Y)$, vendar opomba: obstajajo tudi odvisne spremenljivke, za katere velja gornja zveza. Spremenljivki, za kateri velja $E(XY) \neq E(X) \cdot E(Y)$ imenujemo korelirani.

Disperzija ali varianca $D(X)$ je določena z izrazom: $D(X) = E(X^2) - (E(X))^2$. Naj bo a realna konstanta. Če je $P(X = a) = 1$, je $D(X) = 0$. $D(aX) = a^2 \cdot D(X)$. Količino $\sigma_x = \sqrt{D(x)}$ imenujemo standardna deviacija ali standardni odklon.

Slučajno spremenljivko X standardiziramo s transformacijo: $X_s = \frac{x - \mu}{\sigma}$, kjer sta $\mu = E(X)$ in $\sigma = \sqrt{D(X)}$. Za X_s velja $E(X_s) = 0$ in $D(X_s) = 1$

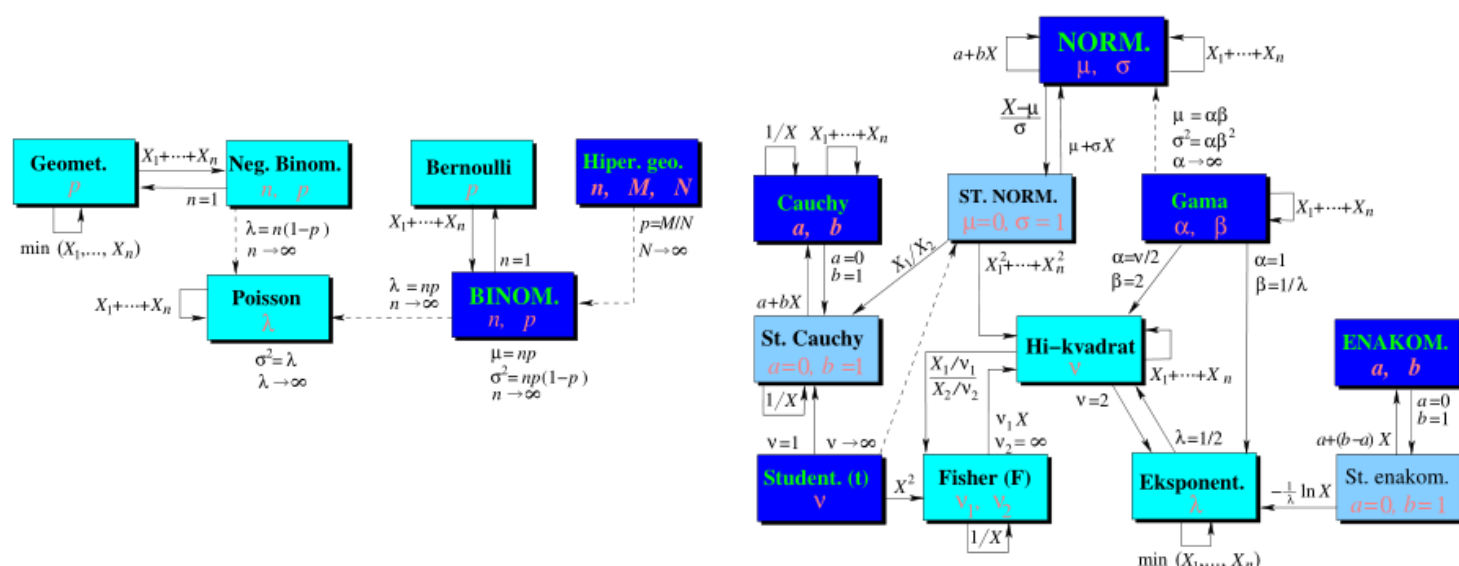
$$E(X_s) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X - \mu)}{\sigma} = \frac{\mu - \mu}{\sigma} = 0.$$

$$D(X_s) = D\left(\frac{X - \mu}{\sigma}\right) = \frac{D(X - \mu)}{\sigma^2} = \frac{\sigma^2 - 0}{\sigma^2} = 1.$$

Pričakovana vrednost in disperzije porazdelitev

porazdelitev	$E(X)$	$D(X)$
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	$2n$

Urejene porazdelitve (kako eno porazdelitev zapisat v drugo (transformacija):



Kovarianca $\text{Cov}(X, Y)$ slučajnih spremenljivk X in Y je določena z izrazom:

$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$ in velja:
 $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (simetričnost) in $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$ (bilinearnost). Spremenljivki X in Y sta nekorelirani natanko takrat, ko je $\text{Cov}(X, Y) = 0$

Če obstajata $D(X)$ in $D(Y)$, obstaja tudi $\text{Cov}(X, Y)$ in velja

$$|\text{Cov}(X, Y)| \leq \sqrt{D(X)D(Y)} = \sigma_X \sigma_Y.$$

Enakost velja natanko takrat, ko je

$$Y - E(Y) = \pm \frac{\sigma_Y}{\sigma_X} (X - E(X))$$

Če imata spremenljivki X in Y končni disperziji,

jo ima tudi njuna vsota $X + Y$ in velja: $D(X + Y) = D(X) + D(Y) + 2\text{Cov}(X, Y)$. Če pa sta spremenljivki nekorelirani, je enostavno $D(X + Y) = D(X) + D(Y)$.

Korelacijski koeficient slučajnih spremenljivk X in Y je določen z izrazom: $r(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$. Za

$(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ je $r(X, Y) = \rho$. Torej sta normalno porazdeljeni slučajni spremenljivki X in Y neodvisni natanko takrat, ko sta nekorelirani. Velja še: $-1 \leq r(X, Y) \leq 1$. $r(X, Y) = 0 \Leftrightarrow X$ in Y sta nekorelirani. $r(X, Y) = 1 \Leftrightarrow Y = \frac{\sigma_Y}{\sigma_X} \cdot (X - E(X)) + E(Y)$ z verjetnostjo 1. $r(X, Y) = -1 \Leftrightarrow Y = -\frac{\sigma_Y}{\sigma_X} \cdot (X - E(X)) + E(Y)$ z verjetnostjo 1. Torej, če $|r(X, Y)| = 1$, obstaja med X in Y linearna zveza z verjetnostjo 1.

Pogojna pričakovana vrednost je pričakovana vrednost pogojne porazdelitve: diskretna slučajna spremenljivka X ima pri pogoju $Y = y_k$ pogojno verjetnostno funkcijo $p_{i|k} = p_{ik}/q_k$, $i = 1, 2, \dots$ in potemtakem pogojno pričakovano vrednost:

Slučajna spremenljivka

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

ima enako pričakovano vrednost kot spremenljivka X :

$$\begin{aligned} E(X|y_k) &= \sum_{i=1}^{\infty} x_i p_{i|k} = \frac{1}{q_k} \sum_{i=1}^{\infty} x_i p_{ik} \\ E(X) &= \sum_{k=1}^{\infty} q_k E(X|y_k) = \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} x_i p_{ik} \\ &= \sum_{i=1}^{\infty} x_i \sum_{k=1}^{\infty} p_{ik} = \sum_{i=1}^{\infty} x_i p_i = E(X). \end{aligned}$$

Pogojna pričakovana vrednost zvezne spremenljivke*

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

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

Slučajna spremenljivka $E(X|Y)$ z gostoto $p_Y(y)$ ima enako pričakovano vrednost kot spremenljivka X

$$\begin{aligned} E(E(X|Y)) &= \int_{-\infty}^{\infty} E(X|y)p_Y(y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xp(x, y) dx dy \\ &= \int_{-\infty}^{\infty} xp_X(x) dx = E(X). \end{aligned}$$

Kovariančna matrika:

Pričakovana vrednost slučajnega vektorja $X = (X_1, X_2, \dots, X_n)$ je vektor $E(X) = (E(X_1), E(X_2), \dots, E(X_n))$. Primer: Za $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ je $E(X, Y) = (\mu_X, \mu_Y)$.

$K = [\text{Cov}(X_i, X_j)]$ je kovariančna matrika vektorja X .

Kovariančna matrika $K = [K_{ij}]$ je simetrična: $K_{ij} = K_{ji}$

Diagonalne vrednosti so disperzije spremenljivk: $K_{ii} = D(X_i)$

Če je kaka lastna vrednost enaka 0, je vsa verjetnost skoncentrirana na neki hiperravnini – porazdelitev je izrojena. To se zgodi natanko takrat, ko je kovariančna matrika K ni obrnljiva, oz. ko je $\det K = 0$.

Primer: Za $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ je $K = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$. Ker je $|\rho| < 1$, je $\det K = \sigma_X^2\sigma_Y^2(1-\rho^2) > 0$.

Višji momenti so posplošitev pojmov pričakovane vrednosti in disperzije. Moment reda $k \in \mathbb{N}$ glede na točko

$a \in \mathbb{R}$ imenujemo količino: $m_k(a) = E((X - a)^k)$. Moment obstaja, če obstaja pričakovana vrednost $E(|X - a|^k) < \infty$.

Za $a = 0$ dobimo začetni moment $z_k = m_k(0)$, za $a = E(X)$ pa centralni moment $m_k = m_k(E(X))$. Primera: $E(X) = z_1$ in $D(X) = m_2$. Če obstaja moment $m_n(a)$, obstajajo vsi $m_k(a)$, za $k < n$

Če obstaja moment z_n , obstaja tudi moment $m_n(a)$ za vse a je element \mathbb{R} .

$$m_n(a) = E((X - a)^n) = \sum_{k=0}^n \binom{n}{k} (-a)^{n-k} z_k.$$

Regresijska funkcija*

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

Primer: Naj bo $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$.

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

Torej je pogojna pričakovana vrednost

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

in prirejena spremenljivka

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

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

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

Obe regresijski funkciji sta **linearni**.

Pričakovana vrednost slučajne spremenljivke Y , ki je linearna kombinacija spremenljivk X_1, X_2, \dots, X_n , je potem za $\mathbf{a} = (a_1, a_2, \dots, a_n)^T$

$$E(Y) = E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i E(X_i) = E(\mathbf{X}) \mathbf{a}^T.$$

Za disperzijo spremenljivke Y pa dobimo $D(Y) = E(Y - E(Y))^2 =$

$$E\left(\sum_{i=1}^n \sum_{j=1}^n a_i a_j (X_i - E(X_i))(X_j - E(X_j))\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j),$$

kjer je $\text{Cov}(X_i, X_j) = E((X_i - E(X_i))(X_j - E(X_j)))$ kovarianca spremenljivk X_i in X_j oziroma $D(Y) = \mathbf{a}^T \mathbf{K} \mathbf{a}$,

... Lastnosti kovariančne matrike

Poglejmo še, kako se spremeni kovariančna matrika pri linearni transformaciji vektorja $X' = \mathbf{A}X$, kjer je \mathbf{A} poljubna matrika reda $n \times n$.

Vemo, da je $D(\mathbf{a}^T X) = \mathbf{a}^T \mathbf{K} \mathbf{a}$.

Tedaj je, če označimo kovariančno matriko vektorja X' s \mathbf{K}' ,

$$\begin{aligned} \mathbf{a}^T \mathbf{K}' \mathbf{a} &= D(\mathbf{a}^T X') = D(\mathbf{a}^T \mathbf{A} X) = D((\mathbf{A}^T \mathbf{a})^T X) \\ &= (\mathbf{A}^T \mathbf{a})^T \mathbf{K} (\mathbf{A}^T \mathbf{a}) = \mathbf{a}^T \mathbf{A} \mathbf{K} \mathbf{A}^T \mathbf{a} \end{aligned}$$

in potemtakem

$$\mathbf{K}' = \mathbf{A} \mathbf{K} \mathbf{A}^T.$$

Posebej za centralni moment velja

$$m_n = m_n(z_1) = \sum_{k=0}^n \binom{n}{k} (-z_1)^k z_{n-k}$$

$$m_0 = 1, m_1 = 0, m_2 = z_2 - z_1^2, m_3 = z_3 - 3z_2 z_1 + 2z_1^3, \dots$$

Asimetrija spremenljivka X imenujemo količino $A(X) = \frac{m_3}{\sigma^3}$.

Sploščenost spremenljivke X imenujemo količino $K(X) = \frac{m_4}{\sigma^4} - 3$, kjer je $\sigma = \sqrt{m_2}$

Za simetrično glede na $z_1 = E(X)$ porazdeljene spremenljivke so vsi lihi centralni momenti enaki 0.

... Višji momenti

Za $X \sim N(\mu, \sigma)$ so $m_{2k+1} = 0$ in $m_{2k} = (2k-1)!!\sigma^{2k}$.
Zato sta tudi $A(X) = 0$ in $K(X) = 0$.

Če sta spremenljivki X in Y neodvisni, je $m_3(X+Y) = m_3(X) + m_3(Y)$.

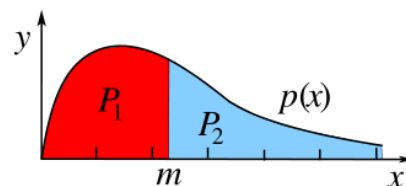
Za binomsko porazdeljeno spremenljivko $X \sim B(n, p)$ je

$$m_3(X) = npq(q-p) \quad \text{in dalje} \quad A(X) = \frac{q-p}{\sqrt{npq}}.$$

Kadar spremenljivka nima momentov, uporabljamo kvantile.

Kvantil reda $p \in (0, 1)$ je vsaka vrednost $x \in \mathbb{R}$, za katero velja $P(X \leq x) \geq p$ in $P(X \geq x) \geq 1-p$ oziroma $F(x) \leq p \leq F(x+)$.
Kvantil reda p označimo z x_p . Za zvezno spremenljivko je $F(x_p) = p$.

Kvantil $x_{\frac{1}{2}}$ imenujemo **mediana**;



$x_{\frac{i}{4}}$, $i = 0, 1, 2, 3, 4$ so **kvantili**.

Kot nadomestek za standardni odklon uporabljamo **kvartilni razmik** $\frac{1}{2}(x_{\frac{3}{4}} - x_{\frac{1}{4}})$.

Karakteristične funkcije in limitni izreki:

Karakteristična funkcija realne slučajne spremenljivke X je kompleksna funkcija realne spremenljivke t določena z zvezo: $\varphi_X(t) = E(e^{it \cdot X})$. Posebej pomembni lastnosti sta: Če obstaja začetni moment z_n , je karakteristična funkcija n -krat odvedljiva v vsaki točki in velja: $\varphi_X^{(k)}(0) = i^k \cdot z^k$. Za neodvisni X in Y je $\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t)$

Reprodukcijska lastnost normalne porazdelitve: vsaka linearna kombinacija neodvisnih in normalno porazdeljenih slučajnih spremenljivk je tudi sama normalno porazdeljena. Če so slučajne spremenljivke X_1, \dots, X_n neodvisne in normalno porazdeljene $N(\mu_i, \sigma_i)$, potem je njihova vsota tudi normalno porazdeljena: $N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right)$. Da ne bi vsota povprečij rastla z n ,

nadomestimo vsoto spremenljivk X_i z njihovim povprečjem \bar{X} in dobimo: $N\left(\bar{\mu}, \sqrt{\sum \left(\frac{\sigma_i}{n}\right)^2}\right)$. Če privzamemo $\mu_i = \mu$ in $\sigma_i = \sigma$, dobimo $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Limitni izreki: zaporedje slučajnih spremenljivk X_n verjetnostno konvergira k slučajni spremenljivki X , če za vsak $\xi > 0$ velja $\lim_{n \rightarrow \infty} P(|X_n - X| < \xi) = 1$. Zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X , če velja: $P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$. Če zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X , potem za vsak $\xi > 0$ velja: $\lim_{m \rightarrow \infty} P(|x_n - x| < \xi \text{ za vsak } n \geq m) = 1$. Od tu izhaja: če konvergira skoraj gotovo $X_n \rightarrow X$, potem konvergira tudi verjetnostno $X_n \rightarrow X$.

Šibki in krepki zakon velikih števil

Naj bo X_1, \dots, X_n zaporedje spremenljivk, ki imajo pričakovano vrednost. Označimo $S_n = \sum_{k=1}^n X_k$ in

$$Y_n = \frac{S_n - E(S_n)}{n} = \frac{1}{n} \sum_{k=1}^n (X_k - E(X_k)) = \frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n E(X_k).$$

Pravimo, da za zaporedje slučajnih spremenljivk X_k velja:

- **šibki zakon velikih števil**, če gre verjetnostno $Y_n \rightarrow 0$, tj., če $\forall \varepsilon > 0$ velja

$$\lim_{n \rightarrow \infty} P\left(\left|S_n - E(S_n)\right|/n < \varepsilon\right) = 1;$$

- **krepki zakon velikih števil**, če gre skoraj gotovo $Y_n \rightarrow 0$, tj., če velja

$$P\left(\lim_{n \rightarrow \infty} (S_n - E(S_n))/n = 0\right) = 1.$$

Če za zaporedje X_1, \dots, X_n velja krepki zakon, velja tudi šibki.

Dokaz Bernoullijevega izreka

Za Bernoullijevo zaporedje X_i so spremenljivke paroma neodvisne, $D(X_i) = pq$, $S_n = k$. Pogoji izreka Čebiševa so izpolnjeni in dobimo:

(Bernoulli 1713) Za vsak $\varepsilon > 0$ velja

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

...Še nekaj izrekov

(Kolmogorov) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene in imajo pričakovano vrednost $E(X_i) = \mu$, potem velja krepki zakon velikih števil

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu\right) = 1.$$

(Borel 1909) Za Bernoullijevo zaporedje velja

$$P\left(\lim_{n \rightarrow \infty} \frac{k}{n} = p\right) = 1.$$

Centralni limitni izrek: »vsaka vsota ali povprečje, če je število členov dovolj veliko, je približno normalno porazdeljena«.

Osnovni centralni limitni izrek

(CLI): Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon (v praksi uporabimo $n > 30$).

Neenakost Čebiševa

Če ima slučajna spremenljivka X končno disperzijo, tj. $D(X) < \infty$, velja za vsak $\varepsilon > 0$ **neenakost Čebiševa**

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



Neenakost Čebiševa – posledice

(Markov) Če gre za zaporedje slučajnih spremenljivk X_i izraz

$$\frac{D(S_n)}{n^2} \rightarrow 0,$$

ko gre $n \rightarrow \infty$, velja za zaporedje šibki zakon velikih števil.

(Čebišev) Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C , tj.

$$D(X_i) < C \quad \text{za vsak } i,$$

velja za zaporedje šibki zakon velikih števil.

Še nekaj izrekov

(Hinčin) Če so neodvisne slučajne spremenljivke X_i enako porazdeljene in imajo pričakovano vrednost $E(X_i) = a$ za vsak i , potem velja zanje šibki zakon velikih števil, tj. za vsak $\varepsilon > 0$ je

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - a\right| < \varepsilon\right) = 1.$$

(Kolmogorov) Če so slučajne spremenljivke X_i neodvisne, imajo končno disperzijo in velja $\sum_{n=1}^{\infty} \frac{D(S_n)}{n^2} < \infty$, potem velja krepki zakon velikih števil:

$$P\left(\lim_{n \rightarrow \infty} \frac{S_n - E(S_n)}{n} = 0\right) = 1.$$

$$Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)}, \text{ kjer je } S_n = X_1 + \dots + X_n.$$

Za zaporedje slučajnih spremenljivk X_i velja **centralni limitni zakon**, če porazdelitvene funkcije za Z_n gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve, to je, če za vsak $x \in \mathbb{R}$ velja

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

Iz konvergence karakterističnih funkcij ϕ_{Y_n} proti karakteristični funkciji standardizirano normalne porazdelitve lahko sklepamo po obratnem konvergenčnem izreku, da tudi porazdelitvene funkcije za Y_n konvergirajo proti porazdelitveni funkciji standardizirano normalne porazdelitve. Torej velja centralni limitni zakon.

Uporaba verjetnosti: Ko vstopi v sobo k-ta oseba, je verjetnost, da je vseh k rojstnih dnevov različnih enaka: $\frac{365}{365} \times \frac{364}{365} \times \dots \times \frac{365-k+1}{365}$.

Enostavnejše kode za odpravljanje napak: V splošnem lahko odpravimo n napak z $(2n + 1)$ -kratnim ponavljanjem in uporabo večinskega pravila. Toda ta metoda je preveč potratna.

Ramsey-eva teorija: Vsaka dovolj velika struktura vsebuje urejeno podstrukturo (popoln nered je nemogoč).

Ramsey-evo število: $r(k, l)$ je najmanjše število za katerega vsak graf na $r(k, l)$ vozliščih vsebuje bodisi k-kliko bodisi l-antikliko.

Ramsey-ev izrek: $\forall k, l \in \mathbb{N}: r(k, l) \leq r(k, l-1) + r(k-1, l)$. Če sta obe števili na desni strani neenakosti sodi, potem velja stroga neenakost. Zgled uporabe: $r(3, 3) \leq r(3, 2) + r(2, 3) = 3 + 3 = 6$.

Erdosev izrek: $\forall k \in \mathbb{N}: r(k, k) \geq 2^{k/2}$. Zgled uporabe: $r(3, 3) \geq 3$ in $r(4, 4) \geq 4$.

Statistika:

Statistika je veda, ki proučuje množične pojave.

Dve veji: opisna statistika (zbiranje in urejanje podatkov o nekem pojavu) in inferenčna statistika (poskuša spoznanja iz zbranih podatkov posplošiti (razširiti, napovedovati,...) in oceniti kakovost teh posplošitev). Lahko jo razdelimo tudi na uporabno in teoretično (računalniško in matematično) statistiko.

(Statistična) enota: posamezna proučevana stvar ali pojav. (npr. redni študent v UL leta 2009)

Populacija je množica vseh proučevanih enot, pomembna je natančna opredelitev populacije (npr. časovno in prostorsko), primer: vsi redni študentje v UL leta 2009.

Vzorec: podmnožica populacije, na osnovi katere po navadi sklepamo o lastnostih celotne populacije.

Spremenljivka: lastnost enot, označujemo jih npr. z X, Y, X_1 . Vrednost spremenljivke X na i -ti enoti označimo z x_i .

Posamezne spremenljivke in odnose med njimi opisujejo ustrezne porazdelitve. Parameter je značilnost populacije in običajno jih označujemo z malimi grškimi črkami. Statistika je značilnost vzorca in običajno jih označujemo z malimi latinskimi črkami. Vrednost statistike je lahko za različne vzorce različna. **Eno izmed osnovnih vprašanj statistike je, kako z uporabo ustreznih statistik oceniti vrednosti izbranih parametrov.**

Vrste spremenljivk glede na vrsto vrednosti:

- Opisne (ali atributivne) spremenljivke: vrednosti opišemo z imeni razredov (npr. poklic)
- Številске (ali numerične) spremenljivke: vrednosti lahko izrazimo s števili (npr. starost)

Vrste spremenljivk glede na vrsto merske lestvice:

- Imenske (ali nominalne) spremenljivke: vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol)
- Urejenostne (ali ordinalne) spremenljivke: vrednosti lahko uredimo od najmanjše do največje (npr. uspeh)
- Razmične (ali intervalne) spremenljivke: lahko primerjamo razlike med vrednostnima dvojic enot (npr. temperatura)
- Razmernostne spremenljivke: lahko primerjamo razmerja med vrednostnima dvojic enot (npr. starost)
- Absolutne spremenljivke: štetja (npr. število prebivalcev)

... Vrste spremenljivk

dovoljene transformacije	vrsta lestvice	primeri
$f(x) = x$ (identiteta)	absolutna	štetje
$f(x) = a \cdot x, a > 0$ podobnost	razmernostna	masa temperatura (K)
$f(x) = a \cdot x + b, a > 0$	razmična	temperatura (C,F) čas (koledar)
$x \geq y \Leftrightarrow f(x) \geq f(y)$ strogo naraščajoča	urejenostna	šolske ocene, kakovost zraka, trdost kamnin
f je povratno enolična	imenska	barva las, narodnost

Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi. Urejenostne spremenljivke zadoščajo lastnostim, ki jih imajo imenske spremenljivke; in podobno razmernostne spremenljivke zadoščajo lastnostim, ki jih imajo razmične, urejenostne in imenske spremenljivke: absolutna \subset razmernostna \subset razmična \subset urejenostna \subset imenska.

Posamezne statistične metode predpostavljajo določeno vrsto spremenljivk. Največ učinkovitih statističnih metod je razvitih za številske spremenljivke.

V teoriji merjenja pravimo, da je nek stavek smiseln, če ohranja resničnost/lažnost pri zamenjavi meritev z enakovrednimi (glede na dovoljene transformacije) meritvami.

Frekvenčna porazdelitev:

Število vseh možnih vrednosti proučevane spremenljivke je lahko preveliko za pregledno prikazovanje podatkov. Zato sorodne vrednosti razvrstimo v skupine. Posamezni skupini priredimo ustrezno reprezentativno vrednost, ki je nova vrednost spremenljivke. Skupine vrednosti morajo biti določene enolično: vsaka enota s svojo vrednostjo je lahko uvrščena v natanko eno skupino vrednosti. Frekvenčna porazdelitev spremenljivke je tabela, ki jo določajo vrednosti ali skupine vrednosti in njihove frekvence. Če je spremenljivka vsaj urejenostna, vrednosti (ali skupine vrednosti) uredimo od najmanjše do največje. Skupine vrednosti številske spremenljivke imenujemo razredi.

x_{\min} in x_{\max} – najmanjša in največja vrednost spremenljivke X . $x_{i,\min}$ in $x_{i,\max}$ sta spodnji in zgornji meji i -tega razreda. Meje razredov so določene tako, da velja $x_{i,\max} = x_{i+1,\min}$. Širina i -tega razreda je $d_i = x_{i,\max} - x_{i,\min}$. Če je le mogoče, vrednosti razvrstimo v razrede enake širine. Sredina i -tega razreda je $x_i = (x_{i,\min} + x_{i,\max}) / 2$ in je značilna vrednost/predstavnik tega razreda. Kumulativna (ali nakopičena frekvenca) je frekvenca do spodnje meje določenega razreda. Velja $F_{i+1} = F_i + f_i$, kjer je F_i kumulativa in f_i frekvenca v i -tem razredu.

Slikovni prikazi:

- Stolpčni prikaz: na eni osi prikažemo (urejene) razrede. Nad vsakim naredimo stolpec/črto višine sorazmerne frekvenci razreda

- Krožni prikaz: vsakemu razredu priredimo krožni izsek s kotom $\alpha_i = \frac{f_i}{n} \cdot 360$ stopinj.
- Histogram: drug poleg drugega rišemo stolpce – pravokotnike, katerih ploščina je sorazmerna frekvenci v razredu. Če so razredi enako široki, je višina sorazmerna tudi frekvenci.
- Poligon: v koordinatnem sistemu zaznamujemo točke (x_i, f_i) , kjer je x_i sredina i-tega razreda in f_i njegova frekvenca. K tem točkam dodamo še točki $(x_0, 0)$ in $(x_{k+1}, 0)$, če je v frekvenčni porazdelitvi k razredov. Točke zvežemo z daljicami.
- Ogiva: grafična predstavitev kumulative frekvenčne porazdelitve s poligonom, kjer v koordinatnem sistemu nanašamo točke $(x_{i,\min}, F_i)$.

Škatle (box-and-whiskers plot, grafikon kvantilov) boxplot: škatla prikazuje notranja kvartila razdeljena z mediansko črto. Daljici – brka vodita do robnih podatkov, ki sta največ za 1.5 dolžine škatle oddaljena od nje. Ostali podatki so prikazani posamično.

Q-Q prikaz (qqnorm) je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatke izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici. Premica qqline nariše premico skozi prvi in tretji kvartil. Obstaja tudi splošnejši ukaz qqplot, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametroma datax=T zamenjamo vlogo koordinatnih osi.

Vzorčenje:

Analitična statistika je veja statistike, ki se ukvarja z uporabo vzorčnih podatkov, da bi z njimi naredili zaključek (inferenco) o populaciji. Zakaj vzorčenje? Cena, čas in destruktivno testiranje.

Glavno vprašanje statistike je: kakšen mora biti vzorec, da lahko iz podatkov zbranih na njem veljavno sklepamo o lastnostih celotne populacije.

Vzorec dobro predstavlja populacijo, če je izbran nepristransko in je dovolj velik. Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor (X_1, \dots, X_n) , ki mu rečemo vzorec. Število n je velikost vzorca.

Ker v vzorcu merimo isto spremenljivko in posamezna meritev ne sme vplivati na ostale, lahko predpostavimo: 1. vsi členi X_i vektorja imajo isto porazdelitev, kot spremenljivka X in 2. členi X_i so med seboj neodvisni. Takemu vzorcu rečemo enostavni slučajni vzorec. Večina statistične teorije temelji na predpostavki, da imamo opravka enostavnim slučajnim vzorcem. Če je populacija končna, lahko dobimo enostavni slučajni vzorec, tako da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo.

Z vprašanjem, kako sestaviti dobre vzorce v praksi, se ukvarja posebno področje statistike – teorija vzorčenja. Načini vzorčenja:

- Ocena: priročnost
- Naključnost:
 - o Enostavno: pri enostavnem naključnem vzorčenju je vsak član populacije izbran/vključen z enako verjetnostjo

- Deljeno : razdeljen naključni vzorec dobimo tako, da razdelimo populacijo na disjunktne množice oziroma dele (razrede) in nato izberemo enostavne naključne vzorce za vsak del posebej
- Grozdno: takšno vzorčenje je enostavno naključno vzorčenje skupin ali klastrov/grozdov elementov

Osnovni izrek statistike: spremenljivka X ima na populaciji G porazdelitev $F(x) = P(X \leq x)$. Toda tudi vsakemu vzorcu ustreza neka porazdelitev. Za realizacijo vzorca (x_1, x_2, \dots, x_n) in $x \in \mathbb{R}$ postavimo $K(x) = |\{x_i : x_i \leq x, i = 1, \dots, n\}|$ in $V_n(x) = K(x)/n$. Slučajni spremenljivki $V_n(x)$ pravimo vzorčna porazdelitvena funkcija. Ker ima, tako kot tudi $K(x)$, $n+1$ možnih vrednosti k/n , $k = 0, \dots, n$, je njena verjetnostna funkcija $B(n, F(x))$: $P(V_n(x) = k/n) = \binom{n}{k} \cdot F(x)^k \cdot (1 - F(x))^{n-k}$.

... Osnovni izrek statistike

Če vzamemo n neodvisnih Bernoullijevih spremenljivk

$$Y_i(x) \sim \begin{pmatrix} 1 & 0 \\ F(x) & 1 - F(x) \end{pmatrix},$$

velja

$$V_n(x) = \frac{1}{n} \sum_{i=1}^n Y_i(x).$$

Krepki zakon velikih števil tedaj zagotavlja, da za vsak x velja

$$P\left(\lim_{n \rightarrow \infty} V_n(x) = F(x)\right) = 1.$$

To je v bistvu Borelov zakon, da relativna frekvenca dogodka $(X \leq x)$ skoraj gotovo konvergira proti verjetnosti tega dogodka.

... Osnovni izrek statistike

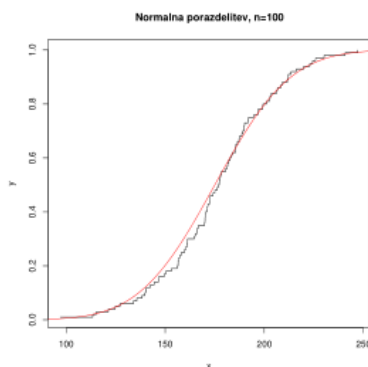
Velja pa še več. $V_n(x)$ je stopničasta funkcija, ki se praviloma dobro prilega funkciji $F(x)$.

Odstopanje med $V_n(x)$ in $F(x)$ lahko izmerimo s slučajno spremenljivko

$$D_n = \sup_{x \in \mathbb{R}} |V_n(x) - F(x)|$$

za $n = 1, 2, 3, \dots$. Zanje lahko pokažemo *osnovni izrek statistike*

$$P\left(\lim_{n \rightarrow \infty} D_n = 0\right) = 1.$$



Torej se z rastjo velikosti vzorca $V_n(x)$ enakomerno vse bolj prilega funkciji $F(x)$ – vse bolj povzema razmere na celotni populaciji.

Vzorčne ocene: najpogostejša parametra, ki bi ju radi ocenili, sta:

- Sredina populacije μ glede na izbrano lastnost – pričakovano vrednost spremenljivke X na populaciji
- Povprečni odklon od sredine σ – standardni odklon spremenljivke X na populaciji

Statistike/ocene za te parametre so izračunane iz podatkov vzorca. Zato jim tudi rečemo vzorčne ocene.

Kot sredinske mere se pogosto uporabljajo:

- Vzorčni modus: najpogostejša vrednost (smiselna tudi za imenske)
- Vzorčna mediana: srednja vrednost, glede na urejenost (smiselna tudi za urejenostne)
- Vzorčno povprečje: povprečna vrednost (smiselna za vsaj razmične): $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Vzorčna geometrijska sredina (smiselna za vsak razmerostne): $G_n = \sqrt[n]{\prod_{i=1}^n x_i}$

Mere razpršenosti uporabimo za oceno populacijskega odklona:

- Vzorčni razmah: $\max_i x_i - \min_i x_i$
- Vzorčna disperzija: $s_0^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$
- Popravljen vzorčna disperzija: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$ ter ustrezna vzorčna odklona s_0 in s

Porazdelitve vzorčnih statistik: Denimo, da je v populaciji N enot in da iz te populacije slučajno izbiramo n enot v enostavni slučajni vzorec ali na kratko slučajni vzorec (vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. $1/N$). Če hočemo dobiti slučajni vzorec, moramo izbrane enote pred ponovnim izbiranjem vrniti v populacijo (vzorec s ponavljanjem). Če je velikost vzorca v primerjavi s populacijo majhna, se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec, ki nastane s slučajnim izbiranjem brez vračanja. Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce. Dobili smo populacijo vseh možnih vzorcev. Teh je v primeru enostavnih slučajnih vzorcev s ponavljanjem N^n ; kjer je N število enot v populaciji in n število enot v vzorcu. Število slučajnih vzorcev brez ponavljanja pa je:

$\binom{N}{n}$, če ne upoštevamo vrstnega reda izbranih enot v vzorcu in $\binom{N+n-1}{n}$, če upoštevamo vrstni red.

Vzorčna porazdelitev povprečja: naj bo x_1, x_2, \dots, x_n naključni vzorec, ki je sestavljen iz n meritev populacije s končno pričakovano vrednostjo μ in končnim standardnim odklonom σ . Potem sta povprečje in standardni odklon vzorčne porazdelitve \bar{X} enaka: $\mu_{\bar{X}} = \mu$ in $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Centralni limitni izrek (pri vzorčenju): če je naključni vzorec velikosti n izbran iz populacije s končno pričakovano vrednostjo μ in končno varianco σ^2 ter če je n dovolj velik ($n > 30$), potem je porazdelitev standardiziranega vzorčnega povprečja \bar{X} , tj. $\frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu) \cdot \sqrt{n}}{\sigma}$, aproksimirana z $N(0,1)$

Dokaz prvega izreka

Oglejmo si **vzorčno povprečje**, določeno z zvezo $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$, ki je tudi slučajna spremenljivka. Tedaj zaradi linearnosti $E(X)$, homogenosti σ_X , 'Pitagorovega izreka za σ_X (za paroma nekorrelirane spremenljivke X_i) in dejstva, da je $E(X_i) = E(X)$ in $(\sigma_{X_i})^2 = \sigma^2$ za $i = 1, \dots, n$, velja

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{nE(X)}{n} = \mu,$$

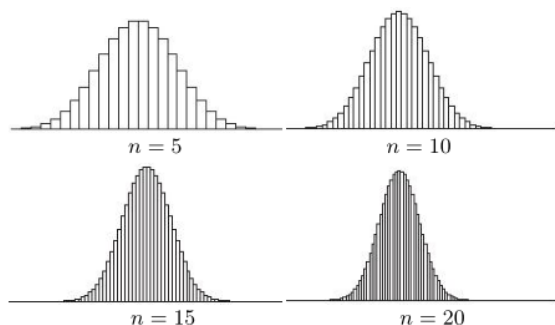
$$(\sigma_{\bar{X}})^2 = \frac{1}{n^2} \sum_{i=1}^n (\sigma_{X_i})^2 = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

Iz druge zveze vidimo, da standardna napaka $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ statistike \bar{X} pada proti 0 z naraščanjem velikosti vzorca, tj. $\bar{X} \rightarrow \mu$; (enako nam zagotavlja tudi krepki zakon velikih števil).

Hitrost centralne tendence pri CLI: dokaz CLI je precej tehničen, kljub temu pa nam ne da občutka kako velik mora biti n , da se porazdelitev slučajne spremenljivke $X_1 + \dots + X_n$ približa normalni porazdelitvi. Hitrost približevanja k normalni porazdelitvi je odvisna od tega kako simetrična je porazdelitev. To lahko potrdimo z eksperimentom: mečemo (ne)pošteno kocko, X_k naj bo vrednost, ki jo kocka pokaže pri k -tem metu.

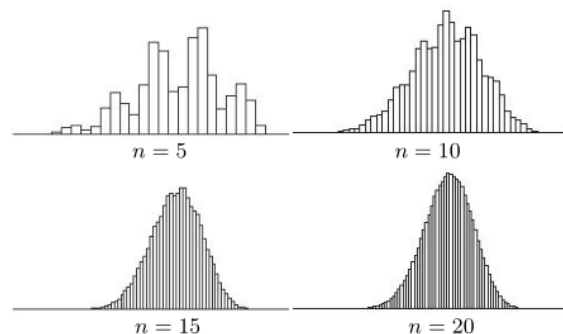
$$p_1 = 1/6, \quad p_2 = 1/6, \quad p_3 = 1/6, \quad p_4 = 1/6, \quad p_5 = 1/6, \quad p_6 = 1/6.$$

in slučajno spremenljivko $X_1 + X_2 + \dots + X_n$:



$$p_1 = 0.2, \quad p_2 = 0.1, \quad p_3 = 0, \quad p_4 = 0, \quad p_5 = 0.3, \quad p_6 = 0.4.$$

in slučajno spremenljivko $X_1 + X_2 + \dots + X_n$:



Cenilke:

Vzorčna statistika je poljubna simetrična funkcije (njena vrednost je neodvisna od permutacije argumentov) vzorca: $Y = g(X_1, X_2, \dots, X_n)$. Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitev vzorca. Najzanimivejši sta značilni vrednosti njene pričakovani vrednosti $E(Y)$ ter standardni odklon σ_Y , ki mu pravimo tudi standardna napaka statistike Y (standard error – zato oznaka $SE(Y)$).

Vzorčno povprečje: porazdelitev vzorčnih povprečij je normalna, kjer je pričakovana vrednost vzorčnih povprečij enako pričakovani vrednosti slučajne spremenljivke na populaciji: $E(\bar{X}) = \mu$ ter standardni odklon vzorčnih povprečij je:

$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon

vzorčnih povprečij: $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$. Za

dovolj velike vzorce ($n > 30$) je porazdelitev vzorčnih povprečij približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se statistika X porazdeljuje vsaj približno normalno s standardno napako $SE(X)$, potem se

z $Z = \frac{\bar{X} - E(\bar{X})}{SE(\bar{X})}$ porazdeljuje standardizirano normalno.

Naj bo $X \sim N(\mu, \sigma)$. Tedaj velja $\sum_{i=1}^n X_i \sim N(n\mu, \sigma\sqrt{n})$ in dalje $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Za standardizirano vzorčno statistiko Z velja

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

Kaj pa če porazdelitev X ni normalna? Izračun porazdelitve se lahko zelo zaplete. Toda pri večjih vzorcih ($n > 30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka Z porazdeljena skoraj standardizirano normalno. Vzorčno povprečje

$$\bar{X} = \frac{\sigma}{\sqrt{n}} Z + \mu$$

je tedaj porazdeljeno približno $N(\mu, \sigma/\sqrt{n})$.

Vzorčna disperzija:

Imejmo normalno populacijo $N(\mu, \sigma)$.

Kako bi določili porazdelitev za vzorčno disperzijo $S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

ali popravljeno vzorčno disperzijo $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$?

Raje izračunamo porazdelitev za statistiko

$$\chi^2 = \frac{nS_0^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

Ker vemo, da je $E(\chi^2(n)) = n$ in $D(\chi^2(n)) = 2n$, lahko takoj izračunamo

$$E(S_0^2) = E\left(\frac{\sigma^2 \chi^2}{n}\right) = \frac{(n-1)\sigma^2}{n} \quad E(S^2) = E\left(\frac{\sigma^2 \chi^2}{n-1}\right) = \sigma^2$$

in

$$D(S_0^2) = D\left(\frac{\sigma^2 \chi^2}{n}\right) = \frac{2(n-1)\sigma^4}{n^2} \quad D(S^2) = D\left(\frac{\sigma^2 \chi^2}{n-1}\right) = \frac{2\sigma^4}{n-1}$$

Če je n zelo velik, je po centralnem limitnem izreku statistika χ^2 porazdeljena približno normalno in sicer po zakonu: $N\left(n-1, \sqrt{2(n-1)}\right)$, cenilka za vzorčno disperzijo S_0^2 približno po:

$N\left(\frac{(n-1) \cdot \sigma^2}{n}, \sqrt{\frac{2(n-1) \cdot \sigma^2}{n}}\right)$ in cenilka za popravljeno vzorčno disperzijo S^2 približno po:

$N\left(\sigma^2, \sqrt{\frac{2}{(n-1)}} \cdot \sigma^2\right)$.

Studentova porazdelitev: pri normalno porazdeljeni slučajni spremenljivki X je tudi porazdelitev \bar{X} normalna, in sicer $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Statistika $z = \frac{\bar{X}-\mu}{\sigma} \cdot \sqrt{n}$ je potem porazdeljena standardizirano normalno. Pri ocenjevanju parametra μ z vzorčnim povprečjem \bar{X} to lahko uporabimo le, če poznamo σ , sicer ne moremo oceniti standardne napake – ne vemo kako dobra je ocena za μ . Parameter σ lahko ocenimo s vzorčnima S_0 in S . Toda S je slučajna spremenljivka in porazdelitev statistike $\frac{\bar{X}-\mu}{s} \cdot \sqrt{n}$ ni več normalna $N(0, 1)$ (razen, če je n zelo velik in s skoraj enak σ).

Porazdelitev nove vzorčne statistike $T = \frac{\bar{X}-\mu}{s} \cdot \sqrt{n}$ je $t(n-1)$ z gostoto :

$$p(x) = \frac{\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}}{\sqrt{n-1} B\left(\frac{n-1}{2}, \frac{1}{2}\right)}$$

in $n-1$ prostostnimi stopnjami. Tej porazdelitvi pravimo Studentova. Za $t(1)$ dobimo Cauchyovo porazdelitev, za $n \rightarrow \infty$ pa ima porazdelitev gostoto standardizirane normalne porazdelitve.

Beta funkcija: vpeljemo jo lahko z Gama funkcijo na naslednji način: $B(x, y) = \frac{\Gamma(x) \cdot \Gamma(y)}{\Gamma(x+y)} = B(y, x)$.

Posebne vrednosti: $B\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$ in za $m, n \in \mathbb{N}$ še: $B(m, n) = \frac{(m-1)! \times (n-1)!}{(m+n-1)!}$.

Fisherjeva ali Snedecorjeva porazdelitev: poskusimo najti še porazdelitev kvocienta $Z = \frac{U}{V}$, kjer sta $U \sim \chi^2(m)$ in $V \sim \chi^2(n)$ in U in V neodvisni. Za $x > 0$ je gostota ustrezne porazdelitve $F(m, n)$ enaka: $p(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{x^{\frac{m}{2}-1}}{(n + mx)^{\frac{m+n}{2}}}$ in 0 drugje. Porazdelitvi $F(m, n)$ pravimo Fisherjeva ali tudi Snedecorjeva porazdelitev F z (m, n) prostostnimi stopnjami.

Točkovna cenilka je pravilo ali formula, ki nam pove, kako izračunati numerično oceno parametra populacije na osnovi merjenj vzorca. Število, ki je rezultat izračuna, se imenuje točkovna ocena (in mu ne moremo zaupati – v smislu verjetnosti).

Cenilka parametra ζ je vzorčna statistika $C = C(X_1, \dots, X_n)$, katere porazdelitveni zakon je odvisen od parametra ζ , njene vrednosti pa ležijo v prostoru parametrov. Cenilka je simetrična funkcija: njena vrednost je enaka za vse permutacije argumentov. Seveda je odvisna tudi od velikosti vzorca n . Primer: vzorčna mediana \bar{X} in vzorčno povprečje \bar{X} sta cenilki za populacijsko povprečje μ .

Cenilka C parametra ζ je dosledna, če z rastočim n zaporedje G_n verjetnostno konvergira k ζ , to je, za vsak $\varepsilon > 0$ velja: $\lim_{n \rightarrow \infty} P(|c_n - \zeta| < \varepsilon) = 1$. Primer: vzorčno povprečje \bar{X} je dosledna cenilka

za populacijsko povprečje μ . Tudi vsi vzorčni začetni momenti $Z_k = \frac{1}{n} \cdot \sum_{i=1}^n X_i^k$ so dosledne cenilke ustreznih začetnih populacijskih momentov $z_k = E(X^k)$, če le-ti obstajajo. Vzorčna mediana \bar{X} je dosledna cenilka za populacijsko mediano. Izrek: Če za $n \rightarrow \infty$ velja $E(C_n) \rightarrow \zeta$ in $D(C_n) = 0$, je C_n dosledna cenilka parametra ζ .

Nepistranska cenilka z najmanjšo varianco: Cenilka C_n parametra ζ je nepistranska, če je $E(C_n) = \zeta$ (za vsak n), in je asimptotično nepistranska, če je $\lim_{n \rightarrow \infty} E(C_n) = \zeta$. Količino $B(C_n) = E(C_n) - \zeta$ imenujemo pristranost (angl. bias) cenilke C_n . Primer: vzorčno povprečje \bar{X} je nepistranska cenilka za populacijsko povprečje μ , vzorčna disperzija S_0^2 je samo asimptotično nepistranska cenilka za σ^2 , popravljena vzorčna disperzija S^2 pa je nepistranska cenilka za σ^2 .

Izmed nepistranskih cenilk istega parametra ζ je boljše tista, ki ima manjšo disperzijo – v povprečju daje boljše ocene. Če je razred cenilk parametra ζ konveksen (vsebuje tudi njihove konveksne kombinacije), obstaja v bistvu ena sama cenilka z najmanjšo disperzijo.

Srednja kvadratična napaka: Včasih je celo bolje vzeti pristransko cenilko z manjšo disperzijo, kot jo ima druga, sicer nepistranska, cenilka z veliko disperzijo. Mera učinkovitosti cenilk parametra ζ je srednja kvadratična napaka $q(C) = E(C - \zeta)^2$, oz. zapišemo v obliki: $q(C) = D(C) + B(C)^2$. Za nepistranske cenilke je $B(C) = 0$ in zato $q(C) = D(C)$. Če pa je disperzija cenilke skoraj 0, je $q(C) \sim B(C)^2$.

Rao-Cramerjeva ocena: Naj bo p gostotna ali verjetnostna funkcija slučajne spremenljivke X in naj bo odvisna še od parametra a , tako da je $p(x; a)$ njena vrednost v točki x . Združeno gostotno ali verjetnostno funkcijo slučajnega vzorca (X_1, \dots, X_n) označimo z L in ji pravimo funkcija verjetja (tudi zanesljivosti, angl. likelihood): $L(x_1, \dots, x_n; a) = p(x_1; a) \cdot \dots \cdot p(x_n; a)$. $L(X_1, \dots, X_n)$ je funkcija vzorca – torej slučajna spremenljivka.

Učinkovitost cenilk: Rao-Cramerjeva ocena da absolutno spodnjo mejo disperzije za vse nepistranske cenilke parametra a (v dovolj gladkih porazdelitvah). Ta meja ni nujno dosežena. Cenilka, ki jo doseže, se imenuje najučinkovitejša cenilka parametra a in je ena sama (z verjetnostjo 1).

Naj bo C_0 najučinkovitejša cenilka parametra a in C kaka druga nepistranska cenilka. Tedaj je učinkovitost cenilke C določena s predpisom: $e(C) = \frac{D(C_0)}{D(C)}$. Učinkovitost najučinkovitejše cenilke je $e(C_0) = 1$. Če najučinkovitejša cenilka ne obstaja, vzamemo za vrednost $D(C_0)$ desno stran v Rao-Cramerjevi oceni.

Metoda momentov: Parametre populacije (ki jih ne poznamo) določimo tako, da so momenti slučajne spremenljivke X (npr. $\bar{X} = \mu_X, S_X^2 = \sigma_X^2$) enaki ocenam teh momentov, ki jih izračunamo

iz vzorca. Če iščemo en parameter (tj. a) in velja $\mu_x = f(a)$ izračunamo: $a = f^{-1}(\mu_x)$ oz. $\hat{a} = f^{-1}(\bar{X})$. Cenilke, ki jih dobimo po metodi momentov so dosledne.

Metoda največjega verjetja: $V = (X_1, \dots, X_n)$ je slučajni vzorec in je odvisen od porazdelitve ter od njenih parametrov a_1, \dots, a_m . Želimo določiti ocene $\hat{a}_1, \dots, \hat{a}_m$ tako, da je verjetnost, da se je zgodil vzorec V , največja. VELJA NEODVISNOST. Funkcija verjetja $L(a_i)$ je verjetnost, da se je zgodil nek vzorec, tj. $L(a_i) = p(\text{vzorec}) = \prod_{i=1}^n p_X(x_i)$. Ocena parametra \hat{a}_i (za parameter a) določimo tako, da bo imela funkcija verjetja $L(a_i)$ maksimum. Če najučinkovitejša cenilka obstaja, jo dobimo s to metodo.

Porazdelitev vzorčnih povprečij:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Porazdelitev vzorčnih deležev: denimo, da želimo na populaciji oceniti delež enot π z določeno lastnostjo. Zato na vsakem vzorcu poiščemo vzorčni delež p . Pokazati se da, da je za dovolj velike slučajne vzorce s ponavljanjem vzorčni deleži porazdeljujejo približno normalno s pričakovano vrednostjo vzorčnih deležev enako deležu na populaciji tj. $E(\hat{P}) = \pi$ in standardnim odklonom vzorčnih deležev

$$SE(\hat{P}) = \sqrt{\frac{\pi(1-\pi)}{n}}. \text{ Za manjše vzorce se}$$

vzorčni deleži porazdeljujejo binomsko. Cenilka populacijskega deleža je nepristranska cenilka.

Izračunajmo, kolikšna vzorčna povprečja ima 90% vzorcev (simetrično na pričakovano vrednost). 90% vzorčnih povprečij se nahaja na intervalu:

$$P(\bar{X}_1 < \bar{X} < \bar{X}_2) = 0.90$$

$$P(-z_1 < z < z_1) = 0.90 \implies 2\Phi(z_1) = 0.90$$

$$\Phi(z_1) = 0.45 \implies z_1 = 1.65$$

Potem se vzorčna povprečja nahajajo v intervalu

$$P\left(\mu - z_1 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_1 \frac{\sigma}{\sqrt{n}}\right) = 0.90$$

oziroma konkretno

$$P\left(100 - 1.65 < \bar{X} < 100 + 1.65\right) = 0.90$$

90% vseh slučajnih vzorcev velikosti 225 enot bo imelo povprečja za inteligenčni kvocient na intervalu

$$(98.35, 101.65).$$

Porazdelitev razlik vzorčnih povprečij: Denimo, da imamo dve populaciji velikosti N_1 in N_2 in se spremenljivka X na prvi populaciji porazdeljuje normalno $N(\mu_1, \sigma)$, na drugi populaciji pa $N(\mu_2, \sigma)$ (**standardna odklona sta na obeh populacijah enaka!**). V vsaki od obeh populacij tvorimo neodvisno slučajne vzorce velikosti n_1 in n_2 . Na vsakem vzorcu (s ponavljanjem) prve populacije izračunamo vzorčno povprečje \bar{X}_1 in podobno na vsakem vzorcu druge populacije \bar{X}_2 . Dokazati se da, da je porazdelitev razlik vzorčnih povprečij normalna, kjer je pričakovana vrednost razlik vzorčnih povprečij enaka: $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$ in disperzija razlike vzorčnih povprečij pa: $D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}$.

Porazdelitev razlik vzorčnih deležev: Podobno kot pri porazdelitvi razlik vzorčnih povprečij naj bosta dani dve populaciji velikosti N_1 in N_2 z deležema enot z neko lastnostjo π_1 in π_2 . Iz prve populacije tvorimo slučajne vzorce velikosti n_1 in na vsakem izračunamo delež enot s to lastnostjo p_1 . Podobno naredimo tudi na drugi populaciji: tvorimo slučajne vzorce velikosti n_2 in na njih določimo deleže p_2 . Pokazati se da, da se za dovolj velike vzorce razlike vzorčnih deležev

porazdeljujejo približno normalno z matematičnim upanjem razlik vzorčnih deležev

$$E(\hat{P}_1 - \hat{P}_2) = E(\hat{P}_1) - E(\hat{P}_2) = \pi_1 - \pi_2 \text{ in disperzijo razlik vzorčnih deležev: } D(\hat{P}_1 - \hat{P}_2) = D(\hat{P}_1) + D(\hat{P}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}.$$

Intervali zaupanja:

Denimo, da s slučajnim vzorcem ocenjujemo parameter γ . Poskušamo najti cenilko G , ki je nepristranska, tj. $E(G) = \gamma$ in se na vseh možnih vzorcih vsaj približno normalno porazdeljuje s standardno napako $SE(G)$. Nato poskušamo najti interval, v katerem se bo z dano gotovostjo $(1 - \alpha)$ nahajal ocenjevalni parameter: $P(a < \gamma < b) = 1 - \alpha$. a je spodnja meja zaupanja, b je zgornja meja zaupanja, α verjetnost tveganja oziroma $1 - \alpha$ verjetnost gotovosti. Ta interval imenujemo interval zaupanja in ga interpretiramo takole: z verjetnostjo tveganja α se parameter γ nahaja v tem intervalu.

Konstrukcija intervala zaupanja: na osnovi predpostavk o porazdelitvi cenilke G zapišemo standardizirano slučajno spremenljivko $Z = \frac{G - E(G)}{SE(G)} = \frac{G - \gamma}{SE(G)}$, ki se porazdeljuje normalno, $N(0, 1)$. Tveganje α porazdelimo simetrično polovico na levo in polovico na desno na konce normalne porazdelitve. Ti konci so označeni z $\pm z_{\alpha/2}$ oz. poltraka: $(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$.

Vrednosti $z_{\alpha/2}$ lahko razberemo iz tabele za verjetnosti za standardizirano normalno porazdelitev, ker velja: $\Phi(z_{\alpha/2}) = 0.5 - \frac{\alpha}{2}$.

Pomen stopnje tveganja pri intervalih zaupanja: Za vsak slučajni vzorec lahko ob omenjenih predpostavkah izračunamo ob izbrani stopnji tveganja α interval zaupanja za parameter γ . Ker se podatki vzorcev razlikujejo, se razlikujejo vzorčne ocene parametrov in zato tudi izračunani intervali zaupanja za parameter γ . To pomeni, da se intervali zaupanja od vzorca do vzorca razlikujejo. **Meji intervala sta slučajni spremenljivki.** Vzemimo stopnjo tveganja $\alpha = 0.05$. Denimo, da smo izbrali 100 slučajnih vzorcev in za vsakega izračunali interval zaupanja za parameter γ . Tedaj lahko pričakujemo, da 5 intervalov zaupanja od 100 ne bo pokrilo iskanega parametra γ .

Intervalsko ocenjevanje parametrov: Naj bo porazdelitev sl. spremenljivke X odvisna od nekega parametra a . Slučajna množica $M \subset \mathbb{R}$, ki je odvisna le od sl. vzorca, ne pa od a , se imenuje množica zaupanja za parameter a , če velja $\exists \alpha \in (0, 1)$, da velja: $P(a \in M) = 1 - \alpha$. Število $1 - \alpha$ imenujemo stopnja zaupanja, α pa stopnja tveganja ($1 - \alpha$ nam pove, kakšna je verjetnost, da M vsebuje vrednost parametra a , ne glede na to, kakšna je njegova dejanska vrednost). Običajno je α enak 0.1, 0.05 ali 0.01. Če je $M = [A, B]$ je M interval zaupanja (za a). A in B sta funkciji slučajnega vzorca – torej statistiki. Pokazati je mogoče, da mora biti $a = -b$ in $\phi(b) = (1 - \alpha) / 2$, kar pomeni $b = z_{\alpha/2}$. Iskani interval je določen nato s točkama:

Primer. $X \sim N(\mu, \sigma)$, ocenjujemo parameter μ pri znanem σ . Izberimo taki konstanti a in b ($b > a$), da bo

$$P(a \leq Z \leq b) = 1 - \alpha,$$

kjer je $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Tedaj je

$$P\left(\bar{X} - \frac{b\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{a\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Označimo

$$A = \bar{X} - \frac{b\sigma}{\sqrt{n}} \text{ in } B = \bar{X} - \frac{a\sigma}{\sqrt{n}}.$$

$$A = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ in } B = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Od tu dobimo, da mora za to, da bo napaka manjša od ξ z verjetnostjo $1 - \alpha$ veljati: $n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\xi}\right)^2$.

Če pri porazdelitvi $X \sim N(\mu, \sigma)$ tudi parameter σ ni znan, ga nadomestimo s cenilko s in moramo zato uporabiti Studentovo statistiko $T = \frac{\bar{X} - \mu}{s} \cdot \sqrt{n}$. Interval je tedaj: $A = \bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}}$, $B = \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}}$, kjer je $P(T > t_{\alpha/2}) = \alpha / 2$.

Teoretična interpretacija koeficienta zaupanja ($1 - \alpha$): Če zaporedoma izbiramo vzorce velikosti n iz dane populacije in konstruiramo $[(1 - \alpha)100]\%$ interval zaupanja za vsak vzorec, potem lahko pričakujemo, da bo $[(1 - \alpha)100]\%$ intervalov vsebovalo pravo vrednost parametra. Stopnja tveganja = $1 - \text{stopnja zaupanja}$.

- I. $(1 - \alpha)\%$ -ni interval zaupanja za povprečje μ populacije, kadar poznamo standardni odklon σ : točki $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ predstavljata krajišči intervala zaupanja, pri čemer je $z_{\alpha/2}$ vrednost spremenljivke, ki zavzame površino $\alpha/2$ na svoji desni; σ je standardni odklon za populacijo; n je velikost vzorca; \bar{x} je vrednost vzorčnega povprečja.
- II. Velik vzorec za $(1 - \alpha)\%$ -ni interval zaupanja za povprečje μ populacije: $\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, kjer je s standardni odklon vzorca.
- III. Majhen vzorec za $(1 - \alpha)\%$ -ni interval zaupanja za povprečje μ populacije: $\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, kjer je porazdelitev spremenljivke X vzeta na osnovi $(n - 1)$ prostostnih stopenj. Privzeli smo: populacija, iz katere smo izbrali vzorec, ima približno normalno porazdelitev.
- IV. $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_1 - \mu_2$, če poznamo odklona σ_1 in σ_2 in sta vzorca izbrana neodvisno: $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- V. Velika vzorca za $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2 , vzorce pa izbiramo neodvisno: $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- VI. Majhni vzorci za $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2 , ki pa sta si enaka, vzorci so izbrani neodvisno: $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, kjer je $s_p^2 = \frac{(n_1-1) \cdot s_1^2 + (n_2-1) \cdot s_2^2}{n_1+n_2-2}$. Privzeli smo: obe populaciji sta približno normalni, varianci sta enaki in naključni vzorci so izbrani neodvisno.
- VII. Majhni vzorci za $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2 , vzorci so izbrani neodvisno: $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, \nu} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, kjer je $\nu = \left\lfloor \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \right\rfloor$
- VIII. Velika vzorca za $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_d = \mu_1 - \mu_2$ ujemajočih se parov: $\bar{d} \pm z_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$, kjer je n število parov.
- IX. Majhni vzorci za $(1 - \alpha)\%$ -ni interval zaupanja za razliko $\mu_d = \mu_1 - \mu_2$ ujemajočih se parov: $\bar{d} \pm t_{\alpha/2, n-1} \cdot \frac{s_d}{\sqrt{n}}$, kjer je n število parov. Privzeli smo: populacija razlik parov je normalno porazdeljena.
- X. ZA DELEŽE: π = delež populacije, p = delež vzorca, kjer je $p = x/n$ in je x število uspehov v n poskusih: $(1 - \alpha)\%$ -ni interval zaupanja za delež populacije π , kadar poznamo σ : $p \pm z_{\alpha/2} \cdot \sigma$

- XI. Velik vzorec za $(1 - \alpha)\%$ -ni interval zaupanja za delež populacije π : $p \pm z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}$, privzeli smo: velikost vzorca n je dovolj velika, da je aproksimacija veljavna. Dobro pravilo (angl. rule of thumb) za izpolnitev pogoja "dovolj velik vzorec" je $np \geq 4$ in $nq \geq 4$.
- XII. $(1 - \alpha)\%$ -ni interval zaupanja za razliko deležev $\pi_1 - \pi_2$, kadar poznamo $\sigma_{\pi_1 - \pi_2}$: $(p_1 - p_2) \pm z_{\alpha/2} \cdot \sigma_{\pi_1 - \pi_2}$
- XIII. Velika vzorca za $(1 - \alpha)\%$ -ni interval zaupanja za razliko deležev $\pi_1 - \pi_2$: $(p_1 - p_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$, privzeli smo: velikost vzorca n je dovolj velika, da je aproksimacija veljavna. Kot splošno pravilo za dovolj velika vzorca privzamemo naslednje: $n_x p_x \geq 4$ in $n_x q_x \geq 4$ za $x = 1, 2$.
- XIV. Majhen vzorec za $(1 - \alpha)\%$ -ni interval zaupanja za varianco populacije σ^2 : $\frac{(n-1) \cdot s^2}{\chi^2_{(1-\alpha/2, n-1)}} \leq \sigma^2 \leq \frac{(n-1) \cdot s^2}{\chi^2_{(\alpha/2, n-1)}}$, privzeli smo: populacija iz katere izbiramo vzorce, ima približno normalno porazdelitev.
- XV. $(1 - \alpha)\%$ -ni interval zaupanja za kvocient varianc dveh populacij σ_1^2 / σ_2^2 : Privzeli smo: obe populaciji iz katerih izbiramo vzorce, imata približno normalni porazdelitvi relativnih frekvenc in naključni vzorci so izbrani neodvisno iz obeh populacij.
$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2, n_1-1, n_2-1}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2, n_2-1, n_1-1}}$$
- XVI. Izbira velikosti vzorca za oceno populacijskega povprečja μ znotraj ε enot z verjetnostjo $(1 - \alpha)$: $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{\varepsilon} \right)^2$. Populacijski odklon mora biti običajno aproksimiran.
- XVII. Izbira velikosti vzorca za oceno razlike $\mu_1 - \mu_2$ med parom populacijskih povprečij, ki je pravilna znotraj ε enot z verjetnostjo $(1 - \alpha)$: $n_1 = n_2 = \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \cdot (\sigma_1^2 + \sigma_2^2)$.
- XVIII. Izbira velikosti vzorca za oceno deleža populacije π , ki je pravilna znotraj ε enot z verjetnostjo $(1 - \alpha)$: $n = \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \cdot pq$, v tem primeru potrebujemo oceni za p in q . Če nimamo nobene na voljo, potem uporabimo $p = q = 0.5$ za konzervativno izbiro števila n .
- XIX. Izbira velikosti vzorca za cenilko razlike $\pi_1 - \pi_2$ med dvema deležema populacije, ki je pravilna znotraj ε enot z verjetnostjo $(1 - \alpha)$: $n_1 = n_2 = \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \cdot (p_1 q_1 + p_2 q_2)$

Preverjanje statističnih domnev:

Načrt: postopek -> elementi (napake 1. in 2. vrste, značilno razlikovanje, moč statističnega testa)
-> testi (centralna tendenca, delež, varianca).

Uvod: postavimo domnevo (hipotezo) o populaciji -> izberemo vzorec, s katerim bomo preverili domnevo -> zavrnemo ali sprejmemo domnevo. Domneva je testirana z določanjem verjetja, da dobimo določen, rezultat kadar jemljemo vzorce iz populacije s predpostavljenimi vrednostnimi parametrom.

Statistična domneva (ali hipoteza) je vsaka domneva o porazdelitvi slučajne spremenljivke X na populaciji.

Če poznamo vrsto (obliko) porazdelitve $p(x; a)$ in postavljamo/raziskujemo domnevo o parametru a , govorimo o parametrični domnevi. Če pa je vprašljiva tudi sama vrsta porazdelitve, je domneva neparametrična. Domneva je enostavna, če natančno določa porazdelitev (njeno vrsto

in točno vrednost parametra); sicer je sestavljena. (primer: če poznamo parameter σ , je domneva $H: \mu = 0$ enostavna; če pa parametra σ ne poznamo, je sestavljena. primer sestavljene domneve je tudi $H: \mu > 0$.).

Statistična domneva je lahko pravilna ali napačna. Želimo seveda sprejeti pravilno domnevo in zavrniti napačno. Težava je v tem, da o pravilnosti/napačnosti domneve ne moremo biti gotovi, če jo ne preverimo na celotni populaciji. Po navadi se odločamo le na podlagi vzorca. Če vzorčni podatki preveč odstopajo od domneve, rečemo, da niso skladni z domnevo, oziroma, da so razlike značilne, in domnevo zavrnemo. Če pa podatki domnevo podpirajo, jo ne zavrnemo – včasih jo celo sprejmemo. To ne pomeni, da je domneva pravilna, temveč da ni zadostnega razloga za zavrnitev.

Postopek preverjanja domneve: postavimo ničelno in alternativno domnevo -> izberemo testno statistiko -> določimo zavrnitveni kriterij -> izberemo naključni vzorec -> izračunamo vrednost na osnovi testne statistike -> sprejmemo odločitev -> naredimo ustrezen zaključek.

Domneva:

- Ničelna domneva (H_0): je trditev o lastnosti populacije za katero predpostavimo, da drži (oz. za katero verjamemo, da je resnična). Je trditev, ki jo test skuša zavreči. (primer: obtoženec je nedolžen)
- Alternativna (nasprotna) domneva (H_a): je trditev nezdružljiva z ničelno domnevo, je trditev, ki jo s testiranjem skušamo dokazati. (primer: obtoženec je kriv)

Odločitev in zaključek:

		odločitev	
		nedolžen	kriv
<ul style="list-style-type: none"> - Porota je spoznala obtoženca za krivega: Zaključimo, da je bilo dovolj dokazov, ki nas prepričajo, da je obtoženec storil kaznivo dejanje. - Porota je spoznala obtoženca za nedolžnega: Zaključimo, da je ni bilo dovolj dokazov, ki bi nas prepričali, da je obtoženec storil kaznivo dejanje. 	nedolžen	pravilna odločitev	napaka 1. vrste (α)
	kriv	napaka 2. vrste (β)	moč testa ($1 - \beta$)

dejansko stanje

Elementi preverjanja domneve:

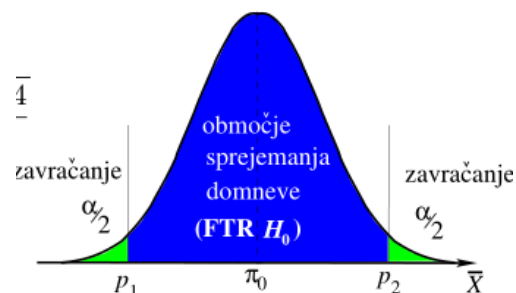
Verjetnost napake 1. vrste (α) je verjetnost za obtožbo nedolžnega obtoženca. Značilno razlikovanje (signifikantno) oziroma stopnja značilnosti. Količina dvoma (α), ki ga bo porota še sprejela.

Verjetnost napake 2. vrste (β): verjetnost, da spoznamo krivega obtoženca za nedolžnega.

Moč testa ($1 - \beta$): verjetnost, da obtožimo krivega obtoženca.

Sodba: breme dokazov, potrebno prepričati poroto, da je obtoženi kriv preko določene stopnje značilnosti.

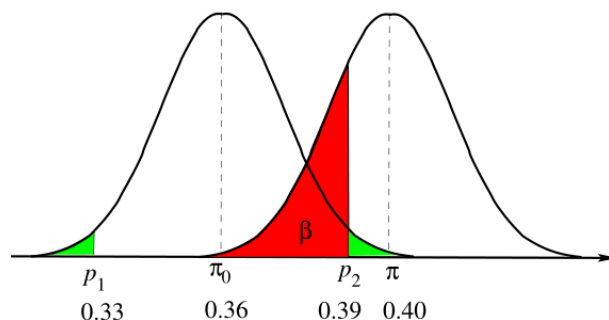
Obramba: Ni bremena dokazovanja, povzročiti morajo dovolj dvoma pri poroti, če je obtoženi resnično kriv.



Kot smo že omenili, je sprejemanje ali zavračanje domnev po opisanem postopku lahko napačno v dveh smislih:

Napaka 1. vrste (α): Če vzorčna vrednost deleža pade v območje zavračanja, domnevo $\pi = \pi_0$ zavrnilo. Pri tem pa vemo, da ob resnični domnevi $\pi = \pi_0$ obstajajo vzorci, ki imajo vrednosti v območju zavračanja. Število α je verjetnost, da vzorčna vrednost pade v območje zavračanja, ob predpostavki, da je domneva resnična. Zato je α verjetnost, da zavrnilo pravilno domnevo – napaka 1. vrste. Ta napaka je merljiva in jo lahko poljubno manjšamo.

Napaka 2. vrste (β): Vzorčna vrednost lahko pade v območje sprejemanja, čeprav je domnevna vrednost parametra napačna. (primer: Ker je območje sprejemanja, domneve v intervalu $0.33 < \pi < 0.39$, lahko izračunamo verjetnost, da bomo sprejeli napačno domnevo takole: $\beta = P(0.33 < \pi < 0.39) = 0.27$. Napako 2. vrste lahko izračunamo le, če imamo znano resnično vrednost parametra π . Ker ga po navadi ne poznamo, tudi ne poznamo napake 2. vrste. Zato ne moremo sprejemati domnev.



Statistična ničelna domneva: $H_0: \mu = \text{vrednost}$

Neusmerjena alternativna domneva: $H_a: \mu \neq \text{vrednost}$

»Manj kot« alternativna domneva: $H_a: \mu < \text{vrednost}$

»Več kot« alternativna domneva: $H_a: \mu > \text{vrednost}$

- ničelna domneva

- $H_0: q = q_0$

- alternativna domneva

- $H_a: q \neq q_0$

- $H_a: q > q_0$

- $H_a: q < q_0$

Definicije:

- Zavrnitev ničelne domneve, če je le-ta pravilna, je napaka 1. vrste. Verjetnost, da naredimo napako 1. vrste, označimo s simbolom α in ji pravimo stopnja tveganja, $(1 - \alpha)$ pa je stopnja zaupanja.
- Če ne zavrnilo ničelno domnevo, v primeru, da je napačna, pravimo, da gre za napako 2. vrste. Verjetnost, da naredimo napako 2. vrste, označimo s simbolom β .
- Moč statističnega testa, $(1 - \beta)$ je verjetnost zavrnitve ničelne domneve v primeru, ko je le-ta v resnici napačna.
- Zaključimo, da se z večanjem števila α manjša verjetnost α , medtem ko za β , z nižanjem vrednosti le-tega, se manjša verjetnost β .

Formalen postopek preverjanja domnev

- Postavi domnevi:
 - ničelna,
 - alternativna.
- 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 (TS).

5. Primerjaj in naredi zaključek.

- če TS pade v kritično območje, ničelno domnevo zavrnemo in sprejmi osnovno domnevo ob stopnji značilnosti α .
- če TS ne pade v kritično območje, pa pravimo da vzorčni podatki kažejo na statistično neznačilne razlike med parametrom in vzorčno oceno.

Elementi preverjanja domneve:

- verjetnost napake 1. vrste (α): Če domneva H_0 drži, kakšna je možnost, da jo zavržemo.
- stopnja značilnosti testa (signifikantnosti): Največji α , ki ga je vodja eksperimenta pripravljen sprejeti (zgornja meja za napako 1. vrste).
- verjetnost napake 2. vrste (β): Če domneva H_0 ne drži, kakšna je možnost, da je ne zavržemo.
- moč statističnega testa: $(1 - \beta)$: Če domneva H_0 ne drži, kakšna je možnost, da jo zavržemo

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. Poenostavljeno: P -vrednost je najmanjša stopnja značilnosti, pri kateri še zavrnemo ničelno domnevo pri danih podatkih:

- Sprejemljivost domneve H_0 na osnovi vzorca: Verjetnost, da je opazovani vzorec (ali podatki) bolj ekstremni, če je domneva H_0 pravilna.
- Najmanjši α pri katerem zavrnemo domnevo H_0 :
 - o če je P-vrednost $> \alpha$, potem FTR H_0 ,
 - o če je P-vrednost $< \alpha$, potem zavrnemo H_0

... Elementi preverjanja domneve

GLEJ DE
NA n !

velikost vzorca	napaka 1. vrste	napaka 2. vrste	moč testa
n	α	β	$1 - \beta$
konst.	↑	↓	↑
konst.	↓	↑	↓
povečanje	↓	↓	↑
zmanjšanje	↑	↑	↓

Predznačni test:

Predpostavke: naključni vzorec (neodvisen, enako porazdeljen (kot celotna populacija), vzorčenje iz zvezne porazdelitve, verjetnostna porazdelitev ima mediano.

Postavitev statistične domneve in izbira testne statistike:

- Ničelna domneva: $H_0 = \tau = \text{vrednost}$
- Alternativna domneva: $H_a = \tau < \text{vrednost}$
- Testna statistika (TS):
 - o S_+ = število vzorcev, ki so večji od mediane τ_0 iz domneve
 - o S_- = število vzorcev, ki so manjši od mediane τ_0 iz domneve

Porazdelitev testne statistike: Vsak poskus je bodisi uspeh bodisi neuspeh. Fiksen vzorec, velikosti n . Naključni vzorci (neodvisni poskusi in konstantna verjetnost uspeha). Torej gre za binomsko porazdelitev: $S_+ \sim B(n, p)$.

Določimo zavrnitveni kriterij: stopnja značilnosti testa α , kritična vrednost $S_+ = \text{vrednost}$, območje zavrnitve $S_+ \leq \text{vrednost}$.

Izberemo naključni vzorec in izračunamo vrednost iz testne statistike in naredimo odločitev (če je x manjši od \bar{X} , daš -, če ne pa +. S_+ je enak število plusov, in pogledaš če pade/ne pade v zavrnitveno območje, da veš ali rabiš sprejeti H_0 ali pa ga zavrniti.

Odločitev in zaključek: Pogledamo, če P-vrednost večje ali manjše od α . S tem vemo ali zavrnemo (zaključimo, da empirični podatki sugerirajo, da velja alternativna trditev) ali sprejmemo (FTR – fail to reject, zaključimo, da nimamo dovolj osnov, da bi dokazali, da velja alternativna trditev) ničelne domneve H_0 .

pH	predznak
5.93	-
6.08	+
5.86	-
5.91	-
6.12	+
5.90	-
5.95	-
5.89	-
5.98	-
5.96	-

$$\bar{X} = 6$$
$$S_+ = 2$$



- I. $H_0: \mu = \mu_0$, če poznamo odklon σ , potem $TS = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ sledi z porazdelitev. P-vrednost: sprejemljivost domneve H_0 na osnovi vzorca (možnost za opazovanje vzorca, če je domneva H_0 pravilna) \rightarrow P-vrednost = $2 * P(Z > [\text{vrednost testne statistike}]) \rightarrow$ dobimo najmanjši α , pri katerem zavrnemo domnevo H_0 . Za $H_a: \mu > \mu_0$ zavrnemo H_0 , če je $TS \geq z_\alpha$.

Za $H_a: \mu < \mu_0$ zavrնemo H_0 , če je $TS \leq -z_{\alpha}$. Za $H_a: \mu \neq \mu_0$ zavrնemo H_0 , če je $TS \leq -z_{\alpha/2}$ ALI $TS \geq z_{\alpha/2}$.

- II. $H_0: \mu = \mu_0$, če ne poznamo odklona σ in je $n \geq 30$, potem $TS = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ sledi t porazdelitev z $n-1$ prostostnimi stopnjami. (Velja omeniti še, da se pri tako velikem n z- in t porazdelitev tako ne razlikujeta kaj dosti).
- III. $H_0: \mu = \mu_0$, če ne poznamo odklona σ , populacija je normalna in je $n < 30$, potem $TS = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ sledi t porazdelitev z $n-1$ prostostnimi stopnjami.

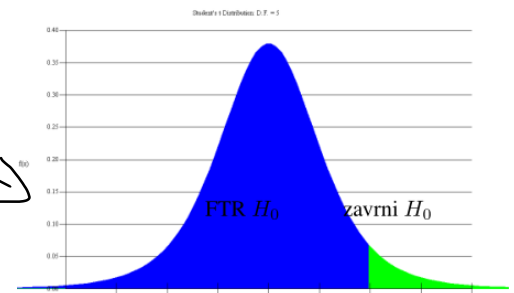
T-test: predpostavke: naključno vzorčenje, ne poznamo varianco populacije in izbiramo vzorce iz normalne porazdelitve in/ali imamo vzorec, pri katerem je n velik. P-vrednost

izračunamo kot: P-vrednost = $P(T > [\text{vrednost TS}])$,

ostalo ostane enako kot pri Z-testu.

Razlaga P-vrednosti: izberi največjo vrednost za α , ki smo jo pripravljene tolerirati. Če je P-vrednost testa manjša kot maksimalna vrednost parametra α , potem zavrni ničelno domnevo.

Določimo zavrnitveni kriterij



Razlika povprečij dveh populacij:

- IV. $H_0: \mu_1 - \mu_2 = D_0$, če poznamo odklona σ_1 in σ_2 ter jemljemo vzorce neodvisno, potem: $TS = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ sledi z porazdelitev.

- V. $H_0: \mu_1 - \mu_2 = D_0$, če ne poznamo odklona σ_1 in/ali σ_2 , vzorce jemljemo neodvisno, $n_1 \geq 30$ in/ali $n_2 \geq 30$, potem: $TS = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ sledi z porazdelitev.

- VI. $H_0: \mu_1 - \mu_2 = D_0$, če ne poznamo odklona σ_1 in/ali σ_2 , vzorce jemljemo neodvisno, populaciji sta normalno porazdeljeni, varianci obeh populacij sta enaki, $n_1 < 30$ ali $n_2 < 30$, potem $TS = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ sledi t porazdelitev z $n_1 + n_2 - 2$ prostostnimi

stopnjami. Privzeli smo: populaciji iz katerih jemljemo vzorce imata obe približno normalno relativno porazdelitev frekvenc, varianci obeh populacij sta enaki, naključni vzorci so izbrani neodvisno iz obeh populacij.

- VII. $H_0: \mu_1 - \mu_2 = D_0$, če ne poznamo odklona σ_1 in/ali σ_2 , vzorce jemljemo neodvisno, spremenljivki sta vsaka na svoji populaciji normalno porazdeljeni, njuni varianci nista enaki, $n_1 < 30$ ali $n_2 < 30$, potem: $TS = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ sledi t porazdelitev z ν prostostnimi

stopnjami. Če v ni naravno število, zaokroži v navzdol do najbližjega naravnega števila za uporabo t-tabele. ν je enak:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- VIII. $H_0 : \mu_d = D_0$, če vzorce jemljemo neodvisno in če je $n \geq 30$ potem: $TS = \frac{\bar{D} - D_0}{\frac{s_d}{\sqrt{n}}}$ sledi z porazdelitev.
- IX. $H_0 : \mu_d = D_0$, če vzorce ne jemljemo neodvisno, če je populacija razlik normalno porazdeljena in če je $n \geq 30$ potem: $TS = \frac{\bar{D} - D_0}{\frac{s_d}{\sqrt{n}}}$ sledi t porazdelitev z $n - 1$ prostostnimi stopnjami.
- X. $H_0 : \pi = \pi_0$, če je n dovolj velik, potem: $TS = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ sledi z porazdelitev. Kot splošno pravilo bomo zahtevali, da velja $np \geq 4$ in $nq \geq 4$.

Razlika deležev dveh populacij: velika vzorca za testiranje domneve o $\pi_1 - \pi_2$. Kot splošno pravilo bomo zahtevali, da velja: $n_x p_x \geq 4$ in $n_x q_x \geq 4$ za $x = 1, 2$.

- XI. Velika vzorca za preverjanje domneve o $\pi_1 - \pi_2$, kadar je $D_0 = 0$. $TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ sledi z porazdelitev, kjer je $p = \frac{y_1 + y_2}{n_1 + n_2}$ in $q = 1 - p$.
- XII. Velika vzorca za preverjanje domneve o $\pi_1 - \pi_2$, kadar je $D_0 \neq 0$. $TS = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}}$ sledi z porazdelitev.
- XIII. Preverjanje domneve o varianci populacije $H_0 : \sigma^2 = \sigma_0^2$. $TS = \frac{(n-1) \cdot s^2}{\sigma_0^2}$ sledi porazdelitev χ^2 . Če je:
- $H_a : \sigma^2 > \sigma_0^2$, potem je odločitveno pravilo: zavrni ničelno domnevo, če je test statistike večji ali enak $\chi_{\alpha}^2(n-1)$.
 - $H_a : \sigma^2 < \sigma_0^2$, potem je odločitveno pravilo: zavrni ničelno domnevo, če je test statistike manjši ali enak $\chi_{1-\alpha}^2(n-1)$.
 - $H_a : \sigma^2 \neq \sigma_0^2$, potem je odločitveno pravilo: zavrni ničelno domnevo, če je test statistike večji ali enak $\chi_{\alpha}^2(n-1)$ ali manjši ali enak $\chi_{1-\alpha}^2(n-1)$.

P-vrednost se tukaj izračuna enako kot pri T porazdelitvi: P-vrednost = $P(\chi^2 > [\text{vrednost testne statistike}])$, pravila in predpostavke enako kot pri Z in T porazdelitvi.

- XIV. Preverjanje domneve o kvocientu varianc neodvisnih vzorcev $H_0 : \sigma_1^2 / \sigma_2^2 = 1$. Če velja $H_0 : \sigma_1^2 / \sigma_2^2 = 1$, potem je testna statistika enaka s_1^2 / s_2^2 , odločitveno pravilo pa je: zavrni ničelno domnevo, če velja: $TS \geq F_{\alpha}(n_1 - 1, n_2 - 1)$. Če pa velja $H_0 : \sigma_1^2 / \sigma_2^2 < 1$ je potem testna statistika enaka: (varianca večjega vzorca)/(varianca manjšega vzorca), odločitveno pravilo pa je: zavrni ničelno domnevo, če velja $s_1^2 > s_2^2$ in $TS \geq F_{\alpha}(n_1 - 1, n_2 - 1)$ oz. zavrni ničelno domnevo, če velja $s_1^2 < s_2^2$ in $TS \geq F_{\alpha}(n_2 - 1, n_1 - 1)$.

Preverjanje domnev o porazdelitvi spremenljivke: Do sedaj smo ocenjevali in preverjali domnevo o parametrih populacije kot μ , σ in π . Sedaj pa bomo preverjali, če se spremenljivka porazdeljuje po določeni porazdelitvi. Test je zasnovan na dejstvu, kako dobro se prilegajo empirične (eksperimentalne) frekvence vrednosti spremenljivke hipotetičnim (teoretičnim) frekvencam, ki so določene s predpostavljeno porazdelitvijo.

Primer za enakomerno porazdelitev:

Preverjanje domneve o enakomerni porazdelitvi

Za primer vzemimo met kocke in za spremenljivko število pik pri metu kocke. Preverimo domnevo, da je kocka pošena, kar je enakovredno domnevi, da je porazdelitev spremenljivke enakomerna. Tedaj sta ničelna in osnovna domneva

H_0 : spremenljivka se porazdeljuje enakomerno,

H_1 : spremenljivka se ne porazdeljuje enakomerno.

Denimo, da smo 120-krat vrgli kocko ($n = 120$) in štejemo kolikokrat smo vrgli posamezno število pik.

To so empirične ali opazovane frekvence, ki jih označimo s f_i .

Teoretično, če je kocka pošena, pričakujemo, da bomo dobili vsako vrednost z verjetnostjo $1/6$ oziroma 20 krat.

To so teoretične ali pričakovane frekvence, ki jih označimo s f'_i .

Podatke zapišimo v naslednji tabeli

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6
f'_i	20	20	20	20	20	20
f_i	20	22	17	18	19	24

S primerjavo empiričnih frekvenc z ustreznimi teoretičnim frekvencami se moramo odločiti, če so razlike posledica le vzorčnih učinkov in je kocka pošena ali pa sp razlike prevelike, kar kaže, da je kocka nepošena. Testna statistika, ki meri prilagojenost empiričnih frekvenc teoretičnim je

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - f'_i)^2}{f'_i},$$

V našem primeru smo uporabili le eno količino in sicer skupno število metov kocke ($n = 120$). Torej število prostostnih stopenj je $m = k - 1 = 6 - 1 = 5$.

Ničelna in osnovna domneva sta tedaj

$$H_0 : \chi^2 = 0 \quad \text{in} \quad H_1 : \chi^2 > 0.$$

Doomnevo preverimo pri stopnji značilnosti $\alpha = 5\%$.

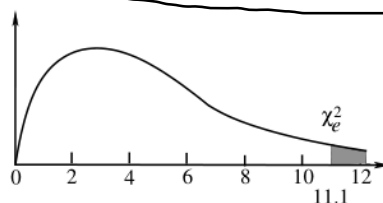
Ker gre za enostranski test, je kritična vrednost enaka

$$\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(5) = 11.1.$$

Izračunamo

$$\begin{aligned} TS &= \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} \\ &+ \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} \\ &= \frac{4+9+4+1+16}{20} = \frac{34}{20} = 1.7. \end{aligned}$$

ki se porazdeljuje po χ^2 porazdelitvi z $m = k - 1$ prostostnimi stopnjami, ki so enake številu vrednosti spremenljivke ali celic (k) minus število količin dobljenih iz podatkov, ki so uporabljene za izračun teoretičnih frekvenc.



Ker TS ne pade v kritično območje, ničelne domneve ne moremo zavrniti. Empirične in teoretične frekvence niso statistično značilno različne med seboj.

Bi variantna analiza in regresija:

$X \leftrightarrow Y$ je povezanost, medtem ko je $X \rightarrow Y$ odvisnost. Mere povezanosti ločimo glede na tip spremenljivk:

- NOMINALNI tip para spremenljivk (ena od spremenljivk je nominalna): χ^2 , kontingenčni koeficienti, koeficienti asociacije;
- ORDINALNI tip para spremenljivk (ena spremenljivka je ordinalna druga ordinalna ali boljša) koeficient korelacije rangov;
- ŠTEVILSKI tip para spremenljivk (obe spremenljivki sta številske): koeficient korelacije

Preverjanje domneve o povezanosti dveh nominalnih spremenljivk: (PRIMER:)

- Enota: dodiplomski študent neke fakultete v letu 1993
- Vzorec: 200 študentov
- 1. spremenljivka: spol
- 2. spremenljivka: stanovanje v času študija

Zanima nas ali študentke drugače stanujejo kot študentje oziroma: ali sta spol in stanovanje v času študija povezana. V ta namen podatke študentov po obeh spremenljivkah uredimo v 2-razsežno frekvenčno porazdelitev. To tabelo imenujemo kontingenčna tabela.

	starši	štud. dom	zasebno	skupaj
moški	16	40	24	80
ženske	48	36	36	120
skupaj	64	76	60	200

Ker nas zanima ali študentke drugače stanujejo v času študija kot študentje, moramo porazdelitev stanovanja študentk primerjati s porazdelitvijo študentov.

Kontingenčna tabela kaže podatke za slučajni vzorec. Zato nas zanima, ali so razlike v porazdelitvi tipa stanovanja v času študija po spolu statistično značilne in ne le učinek vzorca. H_0 : spremenljivki nista povezani in H_1 : spremenljivki sta povezani.

Ker je število študentk različno od števila študentov, moramo zaradi primerjave izračunati relativne frekvence:

	starši	št. dom	zasebno	skupaj
moški	20	50	30	100
ženske	40	30	30	100
skupaj	32	38	30	100

Za preverjanje domneve o povezanosti med dvema nominalnima spremenljivkama na osnovi vzorčnih podatkov, podanih v dvorazsežni frekvenčni porazdelitvi, lahko uporabimo χ^2 test. Ta test sloni na primerjavi empiričnih (dejanskih) frekvenc s teoretičnimi frekvencami, ki so v tem primeru frekvence, ki bi bile v kontingenčni tabeli, če spremenljivki ne bi bili povezani med seboj. To pomeni, da bi bili porazdelitvi stanovanja v času študija deklet in fantov enaki.

Če spremenljivki nista povezani med seboj, so verjetnosti hkratne zgoditve posameznih vrednosti prve in druge spremenljivke enake produktu verjetnosti posameznih vrednosti. Npr., če označimo moške z M in stanovanje pri starših s S, je:

$$P(M) = \frac{80}{200} = 0.40;$$

$$P(S) = \frac{64}{200} = 0.32;$$

$$P(M \cap S) = P(M) \cdot P(S) = \frac{80}{200} \cdot \frac{64}{200} = 0.128.$$

Teoretična frekvenca je verjetnost $P(M \cap S)$ pomnožena s številom enot v vzorcu:

$$f'(M \cap S) = n \cdot P(M \cap S) = 200 \cdot \frac{80}{200} \cdot \frac{64}{200} = 25.6.$$

Spomnimo se tabel empiričnih (dejanskih) frekvenc f_i : χ^2 statistika, ki primerja dejanske in teoretične

Podobno izračunamo teoretične frekvence tudi za druge celice kontingenčne tabele.

frekvence je: $\chi^2 = \sum_{i=1}^k \frac{(f_i - f'_i)^2}{f'_i}$, kjer je k število celic v kontingenčni tabeli. Statistika χ^2 se porazdeljuje po

	starši	št. dom	zasebno	skupaj
moški	26	30	24	80
ženske	38	46	36	120
skupaj	64	76	60	200

χ^2 porazdelitvi s $(s - 1)(v - 1)$ prostostnimi stopnjami, kjer je v število vrstic v kontingenčni tabeli in s število stolpcev.

Ničelna in osnovna domneva sta v primeru tega testa sta:

$H_0: \chi^2 = 0$ (spremenljivki nista povezani) in $H_1: \chi^2 > 0$

(spremenljivki sta povezani). Iz tabele za porazdelitev χ^2

lahko razberemo kritične vrednost te statistike pri 5% stopnji značilnosti: $\chi_{1-\alpha}^2[(s - 1)(v - 1)] = \chi_{0.95}^2(2) = 5.99$. Izračunamo še TS.

$$TS = \frac{(16 - 26)^2}{26} + \frac{(40 - 30)^2}{30} + \frac{(24 - 24)^2}{24} + \frac{(48 - 38)^2}{38} + \frac{(36 - 46)^2}{46} + \frac{(36 - 36)^2}{36} = 12.$$

Ker je eksperimentalna vrednost večja od kritične vrednosti, pomeni, da pade v kritično območje. To pomeni, da ničelno domnevo zavrnamo. Pri 5% stopnji značilnosti lahko sprejmemo osnovno domnevo, da sta spremenljivki statistično značilno povezani med seboj.

Statistika χ^2 je lahko le pozitivna. Zavzame lahko vrednosti v intervalu $[0, \chi_{\max}^2]$, kjer je $\chi_{\max}^2 = n(k - 1)$, če je $k = \min(v, s)$. χ^2 statistika v splošnem ni primerljiva. Zato je definiranih več kontingenčnih koeficientov, ki so bolj ali manj primerni. Omenimo naslednje:

1. Pearsonov koeficient: $\phi = \frac{\chi^2}{n}$, ki ima zgornjo mejo $\phi_{\max}^2 = k - 1$.
2. Cramerjev koeficient: $\alpha = \sqrt{\frac{\phi^2}{k-1}} = \sqrt{\frac{\chi^2}{n(k-1)}}$, ki je definiran na intervalu $[0, 1]$.

3. Kontingenčni koeficient: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$, ki je definiran na intervalu $[0, C_{\max}]$, kjer je $C_{\max} = \sqrt{k / (k - 1)}$.

4. Koeficienti asociacije:

Koeficienti asociacije

5. Yulov koeficient asociacije:

$$Q = \frac{ad-bc}{ad+bc} \in [-1,1]$$

6. Sokal Michenerjev

$$\text{koeficient: } S = \frac{a+d}{a+b+c+d} =$$

$$\frac{a+d}{N} \in [0,1]$$

7. Jaccardov koeficient: $J =$

$$\frac{a}{a+b+c} \in [0,1]$$

8. Pearsonov koeficient: $\phi =$

$$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1,1], \text{ velja } \chi^2 = N \cdot \phi^2.$$

Denimo, da imamo dve nominalni spremenljivki, ki imata le po dve vrednosti (sta dihotomni). Povezanost med njima lahko računamo poleg kontingenčnih koeficientov s **koeficienti asociacije** na osnovi frekvenc iz kontingenčne tabele 2×2 :

$Y \backslash X$	x_1	x_2	
y_1	a	b	$a + b$
y_2	c	d	$c + d$
	$a + c$	$b + d$	N

Preverjanje domneve o povezanosti dveh ordinalnih spremenljivk: (PRIMER):

Vzemimo slučajni vzorec šestih poklicev in ocenimo, koliko so odgovorni (O) in koliko fizično naporni (N). V tem primeru smo poklice uredili od najmanj odgovornega do najbolj odgovornega in podobno od najmanj fizično napornega do najbolj napornega. Poklicem smo torej priredili range po odgovornosti (R_O) in po napornosti (R_N) od 1 do 6.

poklic	R_O	R_N
A	1	6
D	2	4
C	3	5
D	4	2
E	5	3
F	6	1

Povezanost med spremenljivkama lahko merimo s koeficientom korelacije

rangov r_s (Spearman), ki je definiran takole: $r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}$, kjer je d_i

razlika med rangoma v i -ti enoti. Koeficient korelacije rangov lahko zavzame vrednosti v intervalu $[-1, 1]$. Če se z večanjem rangov po prvi spremenljivki večajo rangi tudi po drugi spremenljivki, gre za pozitivno povezanost. Tedaj je koeficient pozitiven in blizu 1. Če pa se z večanjem rangov po prvi spremenljivki rangi po drugi spremenljivki manjšajo, gre za negativno povezanost. Koeficient je tedaj negativen in blizu -1. V našem preprostem primeru gre negativno povezanost. Če ne gre za pozitivno in ne za negativno povezanost, rečemo, da spremenljivki nista povezani.

poklic	R_O	R_N	d_i	d_i^2
A	1	6	-5	25
B	2	4	-2	4
C	3	5	-2	4
D	4	2	2	4
E	5	3	2	4
F	6	1	5	25
vsota			0	66

Res je koeficient blizu, kar kaže na močno negativno povezanost teh 6-ih poklicev.

$$r_s = 1 - \frac{6 \cdot 66}{6 \cdot 35} = 1 - 1.88 = -0.88.$$

Omenili smo, da obravnavamo 6 slučajno izbranih poklicev. Zanima nas, ali lahko na osnovi tega vzorca splošimo na vse poklice, da sta odgovornost in fizična napornost poklicev (negativno) povezana med seboj. Upoštevajmo 5% stopnjo značilnosti.

Za populacijski koeficient ρ_s postavimo ničelno in osnovno domnevo: $H_0: \rho_s = 0$ (spremenljivki nista povezani) in $H_1: \rho_s \neq 0$ (spremenljivki sta povezani). Naj bo R_s slučajna spremenljivka, ki

sledi vrednostim r_s . Potem se $TS = \frac{R_s \cdot \sqrt{n-2}}{\sqrt{1-R_s^2}}$ porazdeljuje približno po t porazdelitvi z $m = (n-2)$

prostostnimi stopnjami. Ker gre za dvostranski test, sta kritični vrednosti enaki: $\pm t_{\alpha/2} = \pm t_{0.025}(4) = \pm 2.776$,

Konstanta n je število poklicev. Nadalje je $TS = -3.71$, kar pomeni da pade v kritično območje. Pri 5% stopnji značilnosti lahko rečemo, da sta odgovornost in fizična napornost (negativno) povezani med seboj. Če je ena od obeh spremenljivk številska, moramo vrednosti pred izračunom d_i rangirati. Če so kakšne vrednosti enake, zanje izračunamo povprečne pripadajoče range.

Preverjanje domneve o povezanosti dveh številskih spremenljivk: (PRIMER:)

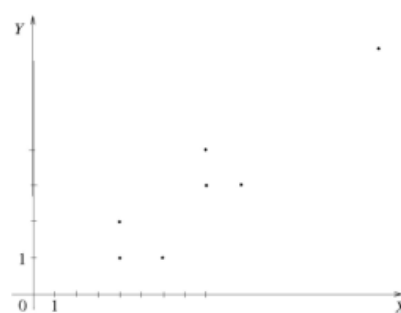
Vzemimo primer dveh številskih spremenljivk: X - izobrazba (število priznanih let šole) in Y - število ur branja dnevnih časopisov na teden.

X	10	8	16	8	6	4	8	4
Y	3	4	7	3	1	2	3	1

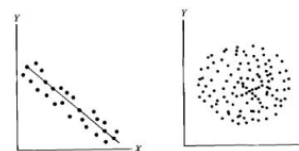
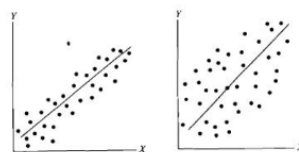
Grafično lahko ponazorimo povezanost med dvema številskima spremenljivkama z razsevnim grafikonom. To je, da v koordinatni sistem, kjer sta koordinati obe spremenljivki, vrišemo enote s pari vrednosti.

Tipi povezanosti:

- funkcijska povezanost: vse točke ležijo na krivulji
- korelacijska (stohastična) povezanost: točke so od krivulje bolj ali manj odklanjajo (manjša ali večja povezanost).



Tipični primeri linearne povezanosti spremenljivk:

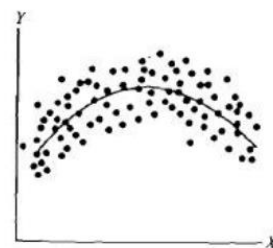


Vzorčna ko varianca (male črke za vrednosti, velike za slučajne spremenljivke): $k(X, Y) = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$ meri linearno povezanost med spremenljivkama.

- $k(X, Y) > 0$ pomeni pozitivno linearno povezanost,
- $k(X, Y) = 0$ pomeni da ni linearne povezanosti,
- $k(X, Y) < 0$ pomeni negativno linearno povezanost.

(Pearsonov koeficient korelacije): $r_{X,Y} = \frac{k(X,Y)}{s_X s_Y}$. Koeficient korelacije lahko zavzame vrednosti v intervalu $[-1, 1]$. Če se z večanjem vrednosti prve spremenljivke večajo tudi vrednosti druge spremenljivke, gre za pozitivno povezanost. Tedaj je $\rho_{X,Y}$ pozitiven in blizu 1. Če pa se z večanjem vrednosti prve spremenljivke vrednosti druge spremenljivke manjšajo, gre za negativno povezanost. Tedaj je $\rho_{X,Y}$ negativen in blizu -1 . Če ne gre za pozitivno in ne za negativno povezanost, rečemo da spremenljivki nista povezani in $\rho_{X,Y}$ je blizu 0.

Primer nelinearne povezanosti spremenljivk:



Statistično sklepanje o korelacijski povezanosti: postavimo torej ničelno in osnovno domnevo za korelacijski koeficient $\rho = \rho_{X,Y}$: $H_0: \rho = 0$ (spremenljivki nista linearno povezani) in $H_1: \rho \neq 0$ (spremenljivki sta linearno povezani). Če slučajna spremenljivka $R_{X,Y}$ spremlja vrednosti indeksa

$r_{X,Y}$, se $TS = \frac{R_{X,Y} \cdot \sqrt{n-2}}{\sqrt{1-R_{X,Y}^2}}$ porazdeljuje po t porazdelitvi z $m = (n - 2)$ prostostnimi stopnjami.

Primer: Preverimo domnevo, da sta izobrazba (število priznanih let šole) in število ur branja dnevnih časopisov na teden povezana med seboj pri 5% stopnji značilnosti. Najprej izračunajmo vzorčni koeficient korelacije:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
10	3	2	0	4	0	0
8	4	0	1	0	1	0
16	7	8	4	64	16	32
8	3	0	0	0	0	0
6	1	-2	-2	4	4	4
4	2	-4	-1	16	1	4
8	3	0	0	0	0	0
4	1	-4	-2	16	4	8
64	24	0	0	104	26	48

$$r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{48}{\sqrt{104 \cdot 26}} = 0.92.$$

Ker gre za dvostranski test, je kritično območje določeno s kritičnima vrednostima $\pm t_{\alpha/2}(n-2) = \pm t_{0.025}(6) = \pm 2.447$.

Izračunajmo še

$$TS = \frac{r_{X,Y} \cdot \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} = \frac{0.92\sqrt{8-2}}{\sqrt{1-0.92^2}} = 2.66,$$

ki pade v kritično območje.

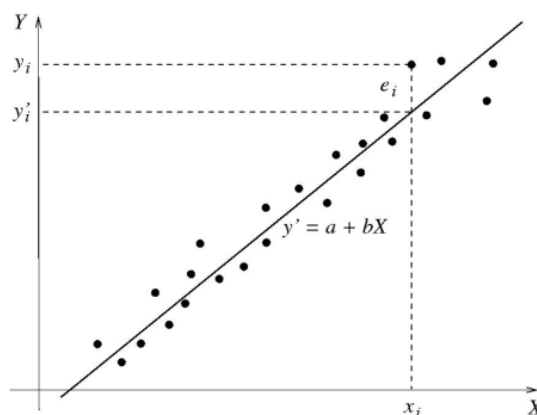
Zaključek: ob 5% stopnji značilnosti lahko rečemo, da je izobrazba linearno povezana z branjem dnevnih časopisov.

Parcialna korelacija: Včasih je potrebno meriti zvezo med dvema spremenljivkama in odstraniti vpliv vseh ostalih spremenljivk. To zvezo dobimo z uporabo koeficienta parcialne korelacije. Pri tem seveda predpostavljamo, da so vse spremenljivke med seboj linearno povezane. Če hočemo iz zveze med spremenljivkama X in Y odstraniti vpliv tretje spremenljivke Z, je koeficient parcialne korelacije. Tudi ta koeficient, ki zavzema vrednosti v intervalu $[-1, 1]$, interpretiramo podobno kot običajni koeficient korelacije. S pomočjo tega obrazca lahko razmišljamo naprej, kako bi izločili vpliv naslednjih spremenljivk.

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1-r_{XZ}^2}\sqrt{1-r_{YZ}^2}}.$$

Regresijska analiza: regresijska funkcija $Y' = f(X)$ kaže, kakšen bi bil vpliv spremenljivke X na Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y. Ker pa so ponavadi še drugi vplivi na proučevano spremenljivko Y, se točke, ki predstavljajo enote v razsevnem grafikonu, odklanjajo od idealne regresijske krivulje: $Y = Y' + E = f(X) + E$, kjer X imenujemo neodvisna spremenljivka, Y odvisna spremenljivka in E člen napake (ali motnja, disturbanca). Če je regresijska funkcija linearna: $Y' = f(X) = a + bX$ je regresijska odvisnost $Y = Y' + E = a + bX + E$ oziroma za i-to enoto $y_i = y'_i + e_i = a + b x_i + e_i$.

Regresijsko odvisnost si lahko zelo nazorno predstavimo v razsevnem grafikonu:



Regresijsko funkcijo lahko v splošnem zapišemo: $Y' = f(X, a, b, \dots)$, kjer so a, b, \dots parametri funkcije. Po navadi se moramo na osnovi pregleda razsevnega grafikona odločiti za tip regresijske funkcije in nato oceniti parametre funkcije, tako da se regresijska krivulja kar se da dobro prilega točkam v razsevnem grafikonu.

Regresijska analiza: regresija je linearna in regresijska premica, ki gre skozi točko (μ_x, μ_y) . Med Y in X ni linearne zveze, sta le 'v povprečju' linearno odvisni. Če označimo z $\beta = \rho \frac{\sigma_y}{\sigma_x}$ regresijski koeficient, $\alpha = \mu_y - \beta \mu_x$ in $\sigma^2 = \sigma_y \cdot \sqrt{1 - \rho^2}$, lahko zapišemo zvezo v obliki: $y = \alpha + \beta x$.

Linearni model: pri proučevanju pojavov pogosto teorija postavi določeno funkcijsko zvezo med obravnavanimi spremenljivkami. Oglejmo si primer linearnega modela, ko je med spremenljivkama x in y linearna zveza $y = \alpha + \beta x$. Za dejanske meritve se pogosto izkaže, da zaradi različnih vplivov, ki jih ne poznamo, razlika $u = y - \alpha - \beta x$ v splošnem ni enaka 0, čeprav je model točen. Zato je ustreznejši verjetnostni linearni model $Y = \alpha + \beta X + U$, kjer so X , Y in U slučajne spremenljivke in $E(U) = 0$ – model je vsaj v povprečju linearen.

Slučajni vzorec (meritve) $(x_1, y_1), \dots, (x_n, y_n)$ je realizacija slučajnega vektorja. Vpeljimo spremenljivke $U_i = Y_i - \alpha - \beta X_i$ in predpostavimo, da so spremenljivke U_i med seboj neodvisne in enako porazdeljene z matematičnim upanjem 0 in disperzijo σ^2 . Torej je: $E(U_i) = 0$, $D(U_i) = \sigma^2$ in $E(U_i U_j) = 0$, za $i \neq j$. Običajno privzamemo še, da lahko vrednosti X_i točno določamo – X_i ima vedno isto vrednost. Poleg tega naj bosta vsaj dve vrednosti X različni. Težava je, da (koeficientov) premice $y = \alpha + \beta x$ ne poznamo. Recimo, da je približek zanjo premica $y = a + bx$.

Določimo jo po načelu najmanjših kvadratov z minimizacijo funkcije: $f(a, b) = \sum_{i=1}^n (y_i - (bx_i + a))^2$. Vpeljemo oznako $\bar{z} = \frac{1}{n} \sum z$ in dobimo: $b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$, $a = \bar{y} - b\bar{x}$.

Seveda sta parametra a in b odvisna od slučajnega vzorca – torej slučajni spremenljivki. Iz dobljenih zvez za a in b dobimo že znani cenilki za koeficient α in β :

$$B = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \bar{Y} - B\bar{X}.$$

Iz prej omenjenih predpostavk lahko (brez poznavanja porazdelitve Y in U) pokažemo

$$E(A) = \alpha \quad \text{in} \quad D(A) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{C_x^2} \right), \quad E(B) = \beta \quad \text{in} \quad D(B) = \frac{\sigma^2}{C_x^2},$$

$$K(A, B) = -\sigma^2 \frac{\bar{X}}{C_x^2}.$$

Cenilki za A in B sta najboljši linearni nepristranski cenilki za α in β .

Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

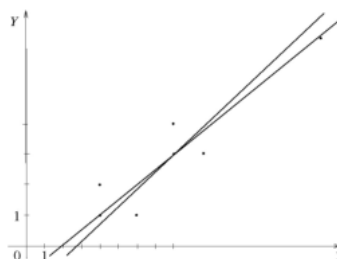
$$Y = \bar{Y} + \frac{K(X, Y)}{S_X^2} (X - \bar{X}).$$

To funkcijo imenujemo tudi prva regresijska funkcija. Podobno bi lahko ocenili linearno regresijsko funkcijo: $X = a^* + b^* Y$. Če z metodo najmanjših kvadratov podobno ocenimo parametra a^* in b^* , dobimo, kar to funkcijo imenujemo druga regresijska funkcija.

$$X = \bar{X} + \frac{K(X, Y)}{S_Y^2} (Y - \bar{Y}).$$

Regresijski premici se sečeta v točki, določeni s pričakovanimi vrednostmi spremenljivk X in Y , (\bar{X}, \bar{Y}) .

Obe regresijski premici lahko vrisemo v razsevni grafikon in preverimo, če se res najbolje prilegata točkam v grafikonu:



Statistično sklepanje o regresijskem koeficientu: vpeljimo naslednje oznake:

$Y = \alpha + \beta X$ regresijska premica na populaciji
in $y = a + bx$ regresijska premica na vzorcu.

Denimo, da želimo preveriti domnevo o regresijskem koeficientu β . Postavimo ničelno in osnovno domnevo takole: $H_0: \beta = \beta_0$ in $H_1: \beta \neq \beta_0$

Npristranska cenilka za regresijski koeficient β je $B = K(X, Y)/S_X^2$. Njena pričakovana vrednost in standardna napaka sta:

$$E(B) = \beta; \quad SE(B) = \frac{S_Y \sqrt{1 - R_{X,Y}^2}}{S_X \sqrt{n - 2}}.$$

Potem se:

$$TS = \frac{S_Y \sqrt{n-2}}{S_X \sqrt{1-R_{X,Y}^2}} (B - \beta_0),$$

porazdeljuje po t porazdelitvi z $m = (n - 2)$ prostostnimi stopnjami.

Pojasnjena varianca (ang. ANOVA)

Vrednost odvisne spremenljivke Y_i lahko razstavimo na tri komponente:

$$y_i = \mu_Y + (y'_i - \mu_Y) + (y_i - y'_i),$$

kjer so pomeni posameznih komponent

μ_Y : rezultat splošnih vplivov,

$(y'_i - \mu_Y)$: rezultat vpliva spremenljivke X (regresija),

$(y_i - y'_i)$: rezultat vpliva drugih dejavnikov (napake/motnje).

Če zgornjo enakost najprej na obeh straneh kvadriramo, nato seštejemo in delež pojasnjene variance spremenljivke Y s spremenljivko X je vseh enotah in končno delimo s številom enot (N), dobimo:

$$\frac{1}{N} \sum_{i=1}^N (y_i - \mu_Y)^2 = \frac{1}{N} \sum_{i=1}^N (y'_i - \mu_Y)^2 + \frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2.$$

To lahko zapišemo takole:

$$\sigma_Y^2 = \sigma_{Y'}^2 + \sigma_e^2,$$

kjer posamezni členi pomenijo:

σ_Y^2 : celotna varianca spremenljivke Y ,

$\sigma_{Y'}^2$: pojasnjena varianca spremenljivke Y ,

σ_e^2 : nepojasnjena varianca spremenljivke Y .

$$R = \frac{\sigma_{Y'}^2}{\sigma_Y^2}.$$

Imenujemo ga **determinacijski koeficient** in je definiran na intervalu $[0, 1]$.

Pokazati se da, da je v primeru linearne regresijske odvisnosti determinacijski koeficient enak

$$R = \rho^2,$$

kjer je ρ koeficient korelacije.

Kvadratni koren iz nepojasnjene variance σ_e imenujemo **standardna napaka regresijske ocene**. Meri razpršenost točk okoli regresijske krivulje oziroma kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo.

V primeru linearne regresijske odvisnosti je standardna napaka enaka:

$$\sigma_e = \sigma_Y \sqrt{1 - \rho^2}$$

saj velja:

$$\sigma_e^2 = \sigma_Y^2 - \sigma_Y^2 R^2 = \sigma_Y^2 - \sigma_Y^2 \frac{\sigma_{Y'}^2}{\sigma_Y^2} = \sigma_Y^2 - \sigma_{Y'}^2.$$

Časovne vrste:

Časovne vrste so niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke. Osnovni namen analize časovnih vrst je opazovati časovni razvoj pojavov, iskati njihove zakonitosti in predvidevati nadaljnji razvoj. Časovne vrste prikazujejo individualne vrednosti neke spremenljivke v času. Čas lahko interpretiramo kot trenutek ali razdobje; skladno s tem so časovne vrste: trenutne (npr. število zaposlenih v določenem trenutku) ali intervalne (npr. družbeni proizvod v letu 1993). Časovne vrste analiziramo tako, da opazujemo spreminjanje vrednosti členov v časovnih vrstah in iščemo zakonitosti tega spreminjanja. Naloga enostavne analize časovnih vrst je primerjava med členi v isti časovni vrsti. Z metodami, ki so specifične za analizo časovnih vrst, analiziramo zakonitosti dinamike ene same vrste, s korelacijsko analizo pa zakonitosti odvisnosti v dinamiki več pojavov, ki so med seboj v zvezi.

Primerljivost členov v časovni vrsti: Kljub temu, da so členi v isti časovni vrsti istovrstne količine, dostikrat niso med seboj neposredno primerljivi. Osnovni pogoj za primerljivost členov v isti časovni vrsti je pravilna in nedvoumna opredelitev pojava, ki ga časovna vrsta prikazuje. Ta opredelitev mora biti vso dobo opazovanja enaka in se ne sme spreminjati. Ker so spremembe pojava, ki ga časovna vrsta prikazuje bistveno odvisne od časa, je zelo koristno, če so časovni razmiki med posameznimi členi enaki. Na velikost pojavov dostikrat vplivajo tudi administrativni ukrepi, ki z vsebino proučevanja nimajo neposredne zveze. En izmed običajnih vzrokov so upravno teritorialne spremembe, s katerimi se spremeni geografska opredelitev pojava, ki onemogoča primerljivost podatkov v časovni vrsti. V tem primeru je potrebno podatke časovne vrste za nazaj preračunati za novo območje.

Grafični prikaz časovne vrste: Kompleksen vpogled v dinamiko pojavov dobimo z grafičnim prikazom časovnih vrst v koordinatnem sistemu, kjer nanašamo na abscisno os čas in na ordinatno vrednosti dane spremenljivke. V isti koordinatni sistem smemo vnašati in primerjati le istovrstne časovne vrste.

Indeksi: Denimo, da je časovna vrsta dana z vrednostmi neke spremenljivke v časovnih točkah takole: X_1, X_2, \dots, X_n , o indeksih govorimo, kadar z relativnimi števili primerjamo istovrstne podatke. Glede na to, kako določimo osnovo, s katero primerjamo člene v časovni vrsti, ločimo dve vrsti indeksov:

- Indeksi s stalno osnovo: Člene časovnih vrst primerjamo z nekim stalnim členom v časovni vrsti, ki ga imenujemo osnova X_0 : $I_{k/0} = \frac{X_k}{X_0} \cdot 100$
- Verižni indeksi: Za dano časovno vrsto računamo vrsto verižnih indeksov tako, da za vsak člen vzamemo za osnovo predhodni člen: $I_{k/0} = \frac{X_k}{X_{k-1}} \cdot 100$, člene časovne vrste lahko primerjamo tudi z absolutno in relativno razliko med členi.
- Absolutna razlika: $D_k = X_k - X_{k-1}$
- Stopnja rasti (relativna razlika med členi): $T_k = \frac{X_k - X_{k-1}}{X_{k-1}} \cdot 100 = I_k - 100$

Interpretacija indeksov

indeks	pojav		
	raste	stagnira	pada
s stalno osnovo	$I_{k+1/0} > I_{k/0}$	$I_{k+1/0} = I_{k/0}$	$I_{k+1/0} < I_{k/0}$
verižni	$I_k > 100$	$I_k = 100$	$I_k < 100$
indeks stopnja rasti	$T_k > 0$	$T_k = 0$	$T_k < 0$

Sestavine dinamike v časovnih vrstah: posamezne vrednosti časovnih vrst so rezultat številnih dejavnikov, ki na pojav vplivajo. Iz časovne vrste je moč razbrati skupen učinek dejavnikov, ki imajo širok vpliv na pojav, ki ga proučujemo. Na časovni vrsti opazujemo naslednje vrste sprememb:

- Dolgoročno gibanje ali trend: X_T : podaja dolgoročno smer razvoja. Običajno ga je mogoče izraziti s preprostimi rahlo ukrivljenimi krivuljami.
- Ciklična gibanja - X_C : so oscilirajo okoli trenda. Periode so ponavadi daljše od enega leta in so lahko različno dolge.
- Sezonske oscilacije - X_S : so posledice vzrokov, ki se pojavljajo na stalno razdobje. Periode so krajše od enega leta, po navadi sezonskega značaja.
- Naključne spremembe - X_E : so spremembe, ki jih ne moremo razložiti s sistematičnimi gibanji (1, 2 in 3).

Časovna vrsta ne vsebuje nujno vseh sestavin. Zvezo med sestavinami je mogoče prikazati z nekaj osnovnim modeli. Npr.: $X = X_T + X_C + X_S + X_E$ ali $X = X_T \cdot X_C \cdot X_S \cdot X_E$ ali

Ali je v časovni vrsti trend? Obstaja statistični test, s katerim preverjamo ali trend obstaja v časovni vrsti. Med časom in spremenljivko izračunamo koeficient korelacije rangov: $r_s = 1 -$

$\frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}$, kjer je d_i , razlika med rangoma i tega časa in pripadajoče vrednosti spremenljivke.

Ničelna in osnovna domneva sta: $H_0: \rho_e = 0$ trend ne obstaja IN $H_1: \rho_e \neq 0$ trend obstaja. Če

slučajna spremenljivka R_s spremlja vrednosti indeksa r_s , se $TS = \frac{R_s \cdot \sqrt{n-2}}{\sqrt{1-R_s^2}}$ porazdeljuje približno

po t porazdelitvi z $(n - 2)$ prostostnimi stopnjami.

Metode določanja trenda: prostoročno, metoda drsečih sredin, Metoda najmanjših kvadratov, druge analitične metode.

Drseče sredine: Metoda drsečih sredin lahko pomaga pri določitvi ustreznega tipa krivulje trenda. V tem primeru namesto člena časovne vrste zapišemo povprečje določenega števila sosednjih članov. Če se odločimo za povprečje treh členov, govorimo o tričlanski vrsti drsečih sredin. Tedaj namesto članov v osnovni časovni vrsti X_k : tvorimo tričlenske drseče sredine X : V tem primeru prvega in zadnjega člena časovne vrste moramo izračunati. $X'_k = \frac{X_{k-1} + X_k + X_{k+1}}{3}$.

Včasih se uporablja obtežena aritmetična sredina, včasih celo geometrijska za izračun drsečih sredin. Če so v časovni vrsti le naključni vplivi, dobimo po uporabi drsečih sredin ciklična gibanja (učinek Slutskega). Če so v časovni vrsti stalne periode, lahko drseče sredine zabrišejo oscilacije v celoti. V splošnem so drseče sredine lahko dober približek pravemu trendu.

V primeru linearnega trenda

$$X_T = a + bT,$$

$$\sum_{i=1}^n (X_i - a - bT_i)^2 = \min.$$

dobimo naslednjo **oceno trenda**

$$X_T = \bar{X} + \frac{\sum_{i=1}^n (X_i - \bar{X})(T_i - \bar{T})}{\sum_{i=1}^n (T_i - \bar{T})^2} (T - \bar{T}).$$

Ponavadi je čas T transformiran tako, da je $\bar{T} = 0$. Tedaj je **ocena trenda**

$$X_T = \bar{X} + \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot t_i}{\sum_{i=1}^n t_i^2} t.$$

Analitično določanje trenda: trend lahko obravnavamo kot posebni primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend $X_T = f(T)$, lahko parametre trenda določimo z metoda najmanjših kvadratov

$$\sum_{i=1}^n (X_i - X_{iT})^2 = \min.$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je $s_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - X_{iT})^2}$, kjer je X_{iT} enak X_T v času t_i .

Načrtovanje eksperimentov:

Načrtovanje eksperimentov se pogosto neposredno prevede v uspeh oziroma neuspeh. V primeru parjenja lahko statistik spremeni svojo vlogo iz pasivne v aktivno. Predstavimo samo osnovne ideje, podrobno numerično analizo pa prepustimo statistični programski opreml.

Elementi načrta so eksperimentalne enote ter terapije, ki jih želimo uporabiti na enotah (primer: medicina: bolniki (enote) in zdravila (terapije), optimizacija porabe: taxi-ji (enote) in različne vrste goriva (terapije),...).

Na primeru bomo predstavili tri osnovne principe načrtovanja eksperimentov:

1. Ponavljanje: enake terapije pridružimo različnim enotam, saj ni mogoče oceniti naravno spremenljivost (ang. natural variability) in napake pri merjenju.
2. Lokalna kontrola pomeni vsako metodo, ki zmanjša naravno spremenljivost. En od načinov grupira podobne enote eksperimentov v bloke. V primeru taxijev uporabimo obe vrsti goriva na vsakem avtomobilu in rečemo, da je avto blok.
3. Naključna izbira je bistven korak povsod v statistiki! Terapije za enote izbiramo naključno. Za vsak taksi izberemo vrsto goriva za torek oziroma sredo z metom kovanca. Če tega ne bi storili, bi lahko razlika med torkom in sredo vplivala na rezultate.

















Latinski kvadrati: latinski kvadrat reda v je $v \times v$ -razsežna matrika, v kateri vsi simboli iz množice $\{1, \dots, v\}$ nastopajo v vsaki vrstici in v vsakem stolpcu.

Trije paroma ortogonalni latinski kvadrati reda 4, tj. vsak par znak-črka ali črka-barva ali barva-znak se pojavi natanko enkrat.

Projektivni prostor $PG(d, q)$ (razseženosti d nad q) dobimo iz vektorskega prostora $[GF(q)]^{d+1}$, tako da naredimo kvocient po 1-razsežnih podprostorih.

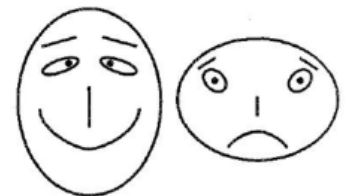
Projektivna ravnina $PG(2, q)$ je incidenčna struktura z 1- in 2-dim. podprostorimi prostora $[GF(q)]^3$ kot točkami in premicami, kjer je " \subset " incidenčna relacija. To je $2-(q^2 + q + 1, q + 1, 1)$ -design, tj.,

- $v = q^2 + q + 1$ je število točk (in število premic b)
- vsaka premica ima $k = q + 1$ točk (in skozi vsako točko gre $r = q + 1$ premic)
- vsak par točk leži na $\lambda = 1$ premicah (in vsaki premici se sekata v natanko eno točki)

Zaključki:

Kako pa predstavimo več kot dve spremenljivki na ravnem listu papirja? Med številnimi možnostmi moramo omeniti idejo Hermana Chernoffa, ki je uporabil človeški obraz, pri čemer je vsako lastnost povezal z eno spremenljivko. X = naklon obrvi, Y = velikost oči, Z = dolžina nosu, T = dolžina ust, U = višino obraza.



Multivariantna analiza: Širok izbor multivariantnih modelov nam omogoča analizo in ponazoritev n -razsežnih podatkov. Združevalna/grozdna tehnika (ang. cluster technique): Iskanje delitve populacije na homogene podskupine, npr. z analizo vzorcev senatorskih glasovanj v ZDA zaključimo, da jug in zahod tvorita dva različna grozda.

Diskriminacijska analiza: je obraten proces. Npr. odbor/komisija za sprejem novih študentov bi rad našel podatke, ki bi že vnaprej opozorili ali bodo prijavljeni kandidati nekega dne uspešno zaključili program (in finančno pomagali šoli - npr. z dobrodelnimi prispevki) ali pa ne bodo uspešni (gre delati dobro po svetu in šola nikoli več ne sliši zanj(o)).

Analiza faktorjev: išče poenostavljeno razlago večrazsežnih podatkov z manjšo skupino spremenljivk. Rezultate testa lahko potem povzamemo le z nekaterimi sestavljenimi rezultati v ustreznih dimenzijah. (primer: Npr. Psihiater lahko postavi 100 vprašanj, skrivoma pa pričakuje, da so odgovori odvisni samo od nekaterih faktorjev: ekstravertiranost, avtoritativnost, alutarizem,...)

Naključni sprehodi: pričnejo z metom kovanca, recimo, da se pomaknemo korak nazaj, če pade grb, in korak naprej, če pade cifra. (z dvema kovancema se lahko gibljemo v 2-razsežnemu

prostoru - tj. ravnini). Če postopek ponavljamo, pridemo do stohastičnega procesa, ki ga imenujemo naključni sprehod (ang. random walk). Modeli na osnovi naključnih sprehodov se uporabljajo za nakup/prodajo delnic in portfolio management.

Vizualizacija in analiza slik: Slike lahko sestavlja 1000×1000 pikslov, ki so predstavljeni z eno izmed 16,7 milijonov barv. Statistična analiza slik želi najti nek pomen iz "informacije" kot je ta.

Ponovno vzorčenje: Pogosto ne moremo izračunati standardne napake in limite zaupanja. Takrat uporabimo tehniko ponovnega vzorčenja, ki tretira vzorec, kot bi bila celotna populacija. Za takšne tehnike uporabljamo pod imeni: randomization Jackknife, in Bootstrapping.

Kvaliteta podatkov: navidezno majhne napake pri vzorčenju, merjenju, zapisovanju podatkov, lahko povzročijo katastrofalne učinke na vsako analizo. R. A. Fisher, genetik in ustanovitelj moderne statistike ni samo načrtoval in analiziral eksperimentalno rejo, pač pa je tudi čistil kletke in pazil na živali. Zavedal se je namreč, da bi izguba živali vplivala na rezultat. Moderni statistiki, z njihovimi računalniki in podatkovnimi bazami ter vladnimi projekti (beri denarjem) si pogosto ne umažejo rok.

Inovacija: Najboljše rešitve niso vedno v knjigah (no vsaj najti jih ni kar tako). Npr. Mestni odpad je najel strokovnjake, da ocenijo kaj sestavljajo odpadki, le-ti pa so se znašli pred zanimivimi problemi, ki se jih ni dalo najti v standardnih učbenikih.

Komunikacija: Še tako uspešna in bistroumna analiza je zelo malo vredna, če je ne znamo jasno predstaviti, vključujoč stopnjo statistične značilnosti? v zaključku. Npr. V medijih danes veliko bolj natančno poročajo o velikosti napake pri svojih anketah.

Timsko delo: V današnji kompleksni družbi. Reševanje številnih problemov zahteva timsko delo. Inženirji, statistiki in delavci sodelujejo, da bi izboljšali kvaliteto produktov.