VJEROJATNOST I STATISTIKA
1. KOLOKVIJ

1. Srednje vrijednosti.

Srednje vrijednosti nazivaju se još i mjerama centralne tendencije. Srednja vrijednost izračunava se uvijek kao prosječna vrijednost obilježja elemenata iz kojih se izračunava. Srednja vrijednost uvijek se nalazi između najmanje i najveće vrijednosti obilježja.
Najčešće su u upotrebi slijedeće srednje vrijednosti:
a) Aritmetička sredina
b) Harmonijska sredina
c) Geometrijska sredina
d) Medijan
e) Mod
Aritmetička sredina je onaj jednaki dio vrijednosti numeričkoga obilježja koji otpada na jedan element skupa. Također može se definirati kao omjer zbroja brojeva i broja brojeva. Aritmetiča sredina računa se iz vrijednosti obilježja svih elemenata statističkoga skupa i broja elemenata u statističkome skupu.

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

Harmonijska sredina predstavlja recipročnu vrijednost aritmetičke sredine recipročnih vrijednosti iz kojih se ona izračunava.

$$H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}}$$

Geometrijska sredina najbolju primjenu ima kao srednja stopa promjene vremenskih nizova, ako je ta stopa približno konstanta. Kod numeričkih nizova ona nema logičnu interpretaciju. Geometrijska sredina je manja od aritmetičke sredine. Za negrupirane i grupirane nizove:

$$G = \sqrt[N]{\prod_{i=1}^{N} x_i}\ ,\ G = \sqrt[\sum_{i=1}^{k} x_i]{\prod_{i=1}^{k} x_i^{x_i}}$$

Medijan je ona srednja vrijednost koja frekvencije statističkog niza dijeli na dva jednaka dijela. Kod negrupiranoga niza medijan je vrijednost obilježja koja pripada članu statističkog niza koji se nalazi u sredini niza. Ukoliko je broj članova paran, onda se uzima aritmetička sredina vrijednosti obilježja dvaju članova koji se nalaze u sredini aritmetičkoga niza. Kod grupiranog statističkog niza treba najprije izračunati frekvencije kumulativnoga niza "manje od" ili "više od", te u tako formiranome nizu pronaći srednji član. Ovdje će se pokazati računanje medijana preko kumulativnoga niza "manje od".

$$M = L_1 + \frac{\frac{N}{2} - \sum_{i=1}^{m} f_i}{f_{med.}} \cdot i$$

Mod je ona vrijednost obilježja koja se najčešće pojavljuje. Mod ima smisla računati samo kod tzv. "unimodalnih distribucija". Kod bimodalne distribucije (koja ima dva vrha) postoje glavni mod i lokalni mod.

$$M_0 = L_1 + \frac{(b-a)}{(b-a)+(b-c)} \cdot i$$

2. Mjere disperzije.

Pod pojmom disperzije podrazumijevamo raspršenost vrijednosti numeričkoga obilježja. Mjere disperzije služe za ocjenjivanje reprezentativnosti srednje vrijednosti obilježja. Najčešće u upotrebi su sljedeće mjere disperzije.
a) Apsolutne mjere disperzije:
- raspon varijance obilježja

- srednje apsolutno odstupanje
- varijanca i standardna devijacija
- interkvartil

b) Relativne mjere disperzije:
- koeficijent varijacije
- koeficijent kvartilne devijacije

Raspon varijance obilježja je razlika između najveće i najmanje vrijednosti numeričkoga obilježja. Ta mjera predstavlja grubu mjeru disperzije obilježja.

$R = X_{max} - X_{min}$

Srednje apsolutno odstupanje (MAD) dobiva se kao aritmetička sredina apsolutnih odstupanja od aritmetičke sredine vrijednosti obilježja.

$MAD = \frac{\sum_{i=1}^{N} |X_i - \bar{X}|}{N}$

Varijanca je srednje kvadratno odstupanje numeričkih vrijednosti obilježja od aritmetičke sredine. Standardna devijacija je pozitivni korijen iz varijance i predstavlja apsolutnu mjeru disperzije u prvome stupnju.

Varijanca za negrupirane podatke se računa preko sljedećega izraza:

$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}$

Za negrupirane nizove može se primijeniti sljedeći izraz:

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

Standardna devijacija se izračunava kao pozitivni korijen iz varijance:

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

Interkvartil predstavlja mjeru raspona obilježja srednjih 50% jedinica u distribuciji, odnosno razliku između gornjega i donjega kvartila:

$I_q = Q_3 - Q_1$

Kvartili zajedno sa medijanom dijele distribuciju na četiri jednaka dijela.

Koeficijent varijacije je relativna mjera disperzije, a služi za mjerenje i uspoređivanje disperzije u različitim distribucijama. Pomoću standardne devijacije ne može se uspoređivati intenzitet disperzije u različitim distribucijama, naročito ako su jedinice mjere različite. Koeficijent varijacije je omjer između standardne devijacije i aritmetičke sredine. Taj omjer se može pomnožiti sa 100, i onda predstavlja postotak standardne devijacije od aritmetičke sredine. U određenim slučajevima može biti i veći od 100%.

$V = \frac{\sigma}{\bar{x}} \cdot (100)$

Koeficijent kvartilne devijacije je relativna mjera disperzije srednjih 50% jedinica u statističkome nizu:

$V = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Njegova vrijednost kreće se između 0 i 1.

3. Pojam vjerojatnosti. Adicijski i multiplikacijski teorem. Bernoullijev zakon velikih brojeva.

Slučajni događaj je takav događaj koji se može, ali ne mora realizirati, odnosno realizira se uz određenu vrijednost. Prema klasičnoj definiciji vjerojatnost realizacije slučajnoga događaja A jednaka je omjeru broja povoljnih ishoda i svih mogućih ishoda:

$P(A) = \frac{m(A)}{n}$

S obzirom na činjenicu da broj povoljnih ishoda ne može premašiti broj svih mogućih ishoda, zaključujemo da je vjerojatnost mjera slučaja koja se kreće između 0 i 1, tj. $0 \leq P \leq 1$. Sigurnome događaju odgovara vjerojatnost jedan.
Ukoliko vjerojatnost realizacije slučajnoga događaja A nije poznata unaprijed, može se izračunati tzv. "vjerojatnost a posteriori".
$P(A) = p \cdot \lim_{n\to\infty} \frac{m(A)}{n}$
Izraz p lim se čita "granična vrijednost po vjerojatnosti" da se razlikuje od limesa u linearnoj algebri. Gornji izraz predstavlja Bernoullijev zakon velikih brojeva. Vjerojatnost "a posteriori" jednaka je graničnoj vrijednosti relativne frekvencije kada broj pokusa teži u beskonačnost.
Ako dva slučajna događaja ne mogu nastupiti istodobno, kažemo da se ti događaji međusobno isključuju tj. da su pripadni skupovi elementarnih događaja disjunktni. Vjerojatnost realizacije jednoga ili drugoga događaja jednaka je zbroju realizacije jednoga i vjerojatnosti realizacije drugoga događaja.
$P(ili\ A\ ili\ B) = P(A) + P(B)$
Ukoliko se slučajni događaji A i B ne isključuju, tada se vjerojatnost realizacije slučajnoga događaja A ili događaja B dobiva na sljedeći način.
$P(ili\ A\ ili\ B) = P(A) + P(B) - P(AB)$
P(AB) predstavlja vjerojatnost da slučajni događaji A i B nastupe istovremeno.
Ukoliko se događaji A i B međusobno ne isključuju, i ako vjerojatnost realizacije jednoga događaja ne zavisi o vjerojatnosti realizacije drugoga događaja, tada je vjerojatnost istovremene realizacije događaja A i događaja B jednaka umnošku vjerojatnosti realizacije događaja A i događaja B.
$P(A\ i\ B) = P(A) \cdot P(B)$
Za takve događaje kažemo da su (stohastički) nezavisni.

4. Uvjetna vjerojatnost. Bayesov teorem.

Ukoliko je realizacija događaja A uvjetovana prethodnom realizacijom događaja B, radi se o uvjetnoj ili kondicionalnoj vjerojatnosti događaja A.
$P\left(\frac{A}{B}\right) = \frac{P(AB)}{P(B)}, P(B) > 0$
odnosno
$P\left(\frac{B}{A}\right) = \frac{P(AB)}{P(B)}, P(A) > 0$
Kod nezavisnih događaja vrijedi sljedeće:
$P\left(\frac{A}{B}\right) = P(A)$
Ako se događaj B realizira istovremeno kada nastupi jedan od n diskunktnih događaja A_1, A_2, ..., A_n za koje je $\sum_{i=1}^{n} P(A_i) = 1$, onda se vjerojatnost događaja B dobiva po formuli potpune vjerojatnosti:
$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$
$P\left(\frac{A_k}{B}\right) = \frac{P(A_k)\cdot P(B/A_k)}{\sum_{i=1}^{n} P(A_i)\cdot P(B/A_k)} \quad \sum_{i=1}^{n} P(A_i) = 1$
Ovaj teorem naziva se Bayesov teorem. Primjenjuje se u medicini i ostalim bioznanostima jer pruža izračun vjerojatnosti za ispravno postavljanje dijagnoze odnosno testa, kako bi se što više smanjila mogućnost za nepotrebne medicinske zahvate. Ovdje se pita kolika je vjerojatnost da će se dogoditi događaj A ako se dogodi događaj B, drugim riječima, vjerojatnost da posljedica A ima za uzrok B.

5. Diskontinuirana slučajna varijabla. Svojstva i teorijske distribucije diskontinuirane slučajne varijable.

Diskontinuirana slučajna varijabla je takva varijabla koja može poprimiti najviše prebrojivo beskonačno mnogo vrijednosti s određenom vjerojatnošću:
$P(x_i), P(x_2), \ldots$ pri čemu mora biti
$\sum_{i=1}^{\infty} P(x_i) = 1$ i $P(x_i) \geq 0$ za svaki i
Uređeni skup parova $\{x_i, P(x_i)\}$, i=1,2,3,... naziva se distribucija slučajne varijable X. Zakon po kojem svakoj vrijednosti slučajne varijable X pripada vjerojatnost $P(x_i)$ naziva se zakon vjerojatnosti slučajne varijable X.
Funkcija distribucije slučajne varijable X predstavlja vjerojatnost da slučajna varijabla X ne premaši neku određenu vrijednost $x_k \in R$
$F(x_k) = P(X \leq x_k) = \sum_{i=1}^{k} P(x_i)$
Funkcija distribucije slučajne varijable X je monotono nepadajuća funkcija.
Očekivanje slučajne varijable jednako je zbroju umnožaka vrijednosti varijable X i odgovarajućih vrijednosti $P(x_i)$.
$E(X) = \sum_{i=1}^{\infty} x_i \cdot p(x_i) = \mu$
Varijanca slučajne varijable X je očekivanje kvadratnoga odstupanja vrijednosti varijable X od njenog očekivanja:
$V(X) = E(X - \mu)^2$
Teorijske distribucije diskontinuirane slučajne varijable X:
1. Binomna distribucija
Ako je vjerojatnost da nastupi neki slučajni događaj poznata i uvijek ista tijekom izvođenja pokusa može se izračunati vjerojatnost da se slučajna varijabla X realizira x puta u n pokusa. U tome slučaju kažemo da se diskontinuirana slučajna varijabla X ravna prema tzv. binomnoj distribuciji koja ima sljedeći zakon vjerojatnosti:
$P(X = x) = \binom{n}{x} \cdot p^x \cdot q^{n-x}$ x=0,1,2,3,...,n
2. Poissonova distribucija
Ako je vjerojatnost slučajnoga događaja veoma malena i konstantna tijekom izvođenja pokusa, umjesto binomne distribucije može se koristiti tzv. Poissonova distribucija. U tom slučaju broj pokusa raste u beskonačnost, ali očekivana vrijednost μ=np ostaje konstanta. Za praktične primjene uzimamo da se binomna distribucija može aproksimirati Poissonovom distribucijom ako je $n \geq 50$ i ako je $p \leq 0.10$. Što je v rijednost n veća, a vrijednost od p manja, to je aproksimacija bolja. Izraz za Poissonovu distribuciju je sljedeći:
$P(X = x) \frac{(np)^x \cdot e^{-np}}{x!}$ x=0,1,2,3,...,∞
odnosno, s obzirom da je μ=np funkcija poprima oblik:
$P(X = x) \frac{\mu^x \cdot e^{-\mu}}{x!}$ x=0,1,2,3,...,∞
3. Hipergeometrijska distribucija
Osnovni skup se sastoji od dva dijela: skup elemenata koji imaju neko obilježje A i skup elemenata koji to obilježje nemaju. U uzorak se bira određeni broj elemenata osnovnoga skupa. Slučajna varijabla je broj jedinica u uzorku koje imaju određeno obilježje A. Ukoliko se jedinice osnovnoga skupa, koje su jednom uzete u uzorak ne vraćaju ponovno u osnovni skup (dakle, nemaju šansu da budu ponovno birane), radi se o tzv. hipergeometrijskoj distribuciji. Izraz za hipergeometrijsku distribuciju je sljedeći:
$P(X = x) = \frac{\binom{M}{n} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$ x=0,1,2,...,n
4. Jednolika distribucija
Ako slučajna varijabla X poprima s istom vjerojatnošću bilo koju od n vrijednosti, kažemo da X ima jednoliku distribuciju:

$P(X = x_i) = \frac{1}{n} \quad x_i = R_x = \{x_1, x_2, \dots, x_n\}$
Najbolji primjer je bacanje kocke gdje se vjerojatnosti da kocka padne na broj 1,2,3,4,5,6 iste i iznose 1/6.

6. Dvodimenzionalna distribucija vjerojatnosti. Marginalna distribucija vjerojatnosti.

Diskontinuirana slučajna varijabla X može poprimiti vrijednosti x_1, x_2, x_3,...,x_k, dok diskontinuirana slučajna varijabla Y može istovremeno poprimiti vrijednosti y_1, y_2, y_3,...,x_m. Vjerojatnost da slučajna varijabla X poprimi vrijednost x_i, a istovremeno slučajna varijabla Y poprimi vrijednosti y_j označava se ovako:
$P(X = x_i, Y = y_j) = P(x_i, y_j)$
Budući da se radi o distribuciji vjerojatnosti, mora biti zadovoljen sljedeći uvjet:
$\sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) = 1$
Skup svih uređenih parova {(x_i, y_j); P(x_i, y_j)} sačinjava dvodimenzionalnu distribuciju slučajne varijable (X, Y).
Kod marginalnih distribucija primjenjuje se adiocioni teorem iz reorije vjerojatnosti. Traži se vjerojatnost da slučajna varijabla X poprimi neku vrijednost X_i bez obzira na to koju će vrijednost poprimiti slučajna varijabla Y_j. Slično tomu, marginalna distribucija slučajne varijable Y predstavlja vjerojatnost da varijabla Y poprimi vjerojatnost Y_j bez obzira koju vrijednost poprima slučajna varijabla X.
$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j)$
$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j)$

7. Kontinuirana slučajna varijabla. Svojstva i teorijske distribucije kontinuirane slučajne varijable.

Kontinuirana slučajna varijabla X može poprimiti neprebrojivo beskonačno mnogo vrijednosti. Zbog toga se kod kontinuiranih slučajnih varijabli ne računa vjerojatnost u određenoj točki, nego nad određenim intervalom vrijednosti slučajne varijable X. Vjerojatnost da će neka kontinuirana slučajna varijabla X poprimiti određenu vrijednost x jednaka je nuli, ali to ne znači i nemogući događaj.
Funkcija vjerojatnosti kontinuirane slučajne varijable X ima sljedeća svojstva:
a) $f(x) \geq 0$
b) $\int_{-\infty}^{+\infty} f(x)dx = 1$, površina ispod krivulje vjerojatnosti jednaka je 1
c) $\int_{x_1}^{x_2} f(x)dx = P\{x_1 < X \leq x_2\}\, x_2 > x_1$
Funkcija distribucije kontinuirane slučajne varijable X je sljedeća:
$F(x) = \int_{-\infty}^{x} f(x)dx \frac{dF(x)}{dx} = f(x)$
F(x) predstavlja vjerojatnost da slučajna varijabla X ne premaši neku unaprijed zadanu vrijednost x.
Očekivana vrijednost kontinuirane slučajne varijable X jednaka je:
$E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \mu$ ako integral konvergira
Varijanca kontinuirane slučajne varijable X dana je izrazom:
$V(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x)dx = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - \mu^2$
Teorijske distribucije kontinuirane slučajne varijable:
1. Normalna distribucija
Normalna distribucija je najvažnija distribucija u statističkoj teoriji. Funkcija gustoće vjerojatnosti normalne distribucije je sljedeća:

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})} \quad x \in <-\infty, +\infty>$

Parametri normalne distribucije su njeno očekivanje μ i varijanca σ^2.

Normalna distribucija je potpuno simetrična distribucija, pa svi koeficijenti asimetrije iznose 0, dok je vrijednost koeficijenata zaobljenosti jednaka 3.

Ako uvedemo standardiziranu varijablu Z oblika:

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

dobivamo sljedeći oblik normalne distribucije:

$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} \cdot z^2}$

2. Studentova distribucija

Studentova distribucija je česta u primjenama kod procjene parametara i kod testiranja hipoteza na osnovu uzorka.

Područje vrijednosti varijable "t" je interval (-∞;+∞). Studentova distribucija je simetrična s obzirom na t=0. Spljoštenija je od normalne distribucije. Ukoliko $v \mapsto \infty$, Studentova distribucija teži jediničnoj normalnoj distribuciji.

Ako je Z varijabla jedinične normalne distribucije N(0;1), a χ^2 varijabla gama distribucije s brojem stupnjeva slobode v, onda je:

$t = \frac{z}{\sqrt{\frac{\chi^2}{v}}}$ varijabla Studentove distribucije s brojem stupnjeva slobode v.

$\lim_{k\to\infty} f(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}t^2}$

3. Hi-kvadrat distribucija

Ako su X_1, X_2, X_3,...,X_n nezavisne normalne varijable koje imaju jednaka očekivanja, $E(X_1)=...=E(X_n)=\mu$ i jednake varijance $V(X_1)=...=V(X_n)=\sigma^2$, tada je:

$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i-\mu}{\sigma}\right)$ gama varijabla sa stupnjem slobode v=n

Područje vrijednosti varijable χ^2 je (0;+∞)

Varijabla

$\chi^2 = \sum_{i=1}^{r} \frac{(f_i - f_{ti})^2}{f_{ti}}$ pripada hi-kvadrat distribuciji s brojem stupnjeva slobode koji je jednak:

v=r – broj neopoznatih parametara u pretpostavljenoj distribuciji – 1.

Za v≥3 distribucija je pozitivno asimetrična. Što je broj stupnjeva slobode veći, distribucija je bliža obliku normalne distribucije. Kod v>30 upotrebljava se aproksimacija normaknom distribucijom:

$\chi^2 = \frac{1}{2}(z + \sqrt{2 \cdot v - 1})^2$

gdje je "z" varijabla iz jedinične normalne distribucije.

4. F-distribucija

F-distribucija je određena s dva parametra v_1 i v_2 koji predstavljaju stupnjeve slobode. Ako su χ_1^2 i χ_2^2 nezavisne χ^2 razdiobe sa stupnjevima slobode v_1 i v_2 tada varijabla:

$F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$ pripada F-distribuciji sa stupnjevima slobode v_1 i v_2

Koristi se najviše pri testiranju značajnosti omjera varijanci.

8. Metoda uzoraka. Pristranost procjene na osnovu uzoraka. Nabrojite sve vrste grješaka koje se javljaju kod rada s uzorcima.

Uzorak predstavlja podskup osnovnoga statističkoga skupa koji se uzima u svrhu ispitivanja obilježja elemenata osnovnoga skupa. Potrebno je da bude reprezentativan i da je izbor jedinica izvršen na slučajan način.

Postoji mogućnost odabira jedinica u uzorak s ponavljanjem i bez ponavljanja.

Frakcija odabiranja predstavlja omjer broja jedinica u uzorku i broja jedinica u osnovnome skupu,

$f = \frac{n}{N}$

Recipročna vrijednost frakcije odabiranja naziva se korakom izbora, a upotrebljava se kod sistematskoga izbora jedinica u uzorak.

Broj svih mogućih uzoraka bez ponavljanja veličine n iz osnovnoga skupa veličine N jednak je broju kombinacija bez ponavljanja n-tog razreda od N elemenata:

$K = \frac{N}{n}$

Pomoću uzorka procjenjuju se određeni parametri osnovnoga skupa ili se testiraju hipoteze o nepoznatim parametrima osnovnoga skupa:

Prednost rada sa uzorcima:

a) osnovni skup (populacija) iz kojega se uzorak bira može biti veoma velik, čak i beskonačan, što je npr. U bioznanostima ili tehničkim znanostima gotovo pravilo

b) troškovi i vrijeme istraživanja čitavoga osnovnoga skupa su ogromni. Ponekad je vrijeme koje je potrebnoza istraživanje veliki protivnik jer se očekuje rezultat što prije npr. kod pronalaska novoga lijeka ili novoga cjepiva za nepoznatu bolest. Predugačko vrijeme istraživanja može stajati mnogih ljudskih života

c) obrada uzorka često je kvalitetnija od obrade čitavog osnovnog skupa upravo zbog veličine samoga istraživanja. Istraživač će temeljitije i kvalitetnije obraditi jedinice u uzorku negoli jedinice u osnovnome skupu, čime se greške mjerenja smanjuju

d) uzorci, ako su dovoljno veliki, daju veoma pouzdane rezultate procjena ili testiranja hipoteza pa nije ni potrebno ići na obradu čitavoga osnovnoga skupa

9. Procjena aritmetičke sredine na osnovu uzorka. Procjena totala na osnovu uzorka.

Distribucija aritmetičkih sredina svih mogućih uzoraka naziva se sampling distribucija. Sampling distribucija aritmetičkih sredina teži normalnom obliku kada veličina uzorka teži k beskonačnosti. Kod praktičnih primjena uzima se da je uzorak mali ako je n≤30 jedinica. U tome slučaju sampling distribucija ima oblik Studentove t-distribucije.

Očekivana vrijednost sampling distribucije aritmetičkih sredina jednaka je očekivanoj vrijednosti osnovnoga skupa. To svojstvo naziva se nepristranost.

$E\left(\hat{\bar{X}}\right) = \sum P_i \hat{\bar{X}}_i$

Očekivana vrijednost sampling distribucije može, ali ne mora biti jednaka očekivanoj vrijednosti osnovnoga skupa. Razlika između očekivane vrijednosti parametra u osnovnome skupu i u sampling distribuciji naziva se pristranost.

$E\left(\hat{\theta}\right) = \theta$ onda kažemo da je procjena $\hat{\theta}$ nepristrana, a ako vrijedi:

$E\left(\hat{\theta}\right) \neq 0$ onda kažemo da je procjena $\hat{\theta}$ pristrana.

Pristranost procjene na bazi uzorka predstavlja razliku između očekivane vrijednosti parametra u sampling distribuciji i vrijednosti populacijskog parametra. Ocjena aritmetičke sredine na osnovu uzorka je nepristrana ako se uzima jednostavni slučajni uzorak. Vrijedi:

$E\left(\hat{\bar{X}}\right) = \bar{X}$

Standardna devijacija sampling distribucije naziva se standardna greška. Izračunava se na sljedeći način:

$Se\left(\hat{\bar{X}}\right) = \frac{\sigma}{\sqrt{n}}$ ako je f<0.05, odnosno

$Se\left(\hat{\bar{X}}\right) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ ako je f>0.05

Ukoliko standardna devijacija osnovnoga skupa σ nije poznata, upotrebljavamo procjenu na osnovu uzorka. Procjena standardne devijacije na osnovu uzorka može se dobiti na sljedeći način:

$s^2 = \hat{\sigma} \cdot \left(\frac{n}{n-1}\right)$ gdje s^2 predstavlja nepristranu procjenu varijance.

Određivanje veličine uzorka uz zadani nivo povjerenja i maksimalnu grešku može se izvršiti na sljedeći način:

$n' = \left[\frac{z \cdot \sigma}{gre\check{s}ka}\right]^2$

Ako je frakcija odabiranja manja od 0.05, tada se konačna veličina uzorka određuje kao n'. Ako je frakcija izbora veća od 0.05, onda se pristupa korekciji na sljedeći način:

$n = \frac{n'}{1+\frac{n'}{N}}$

Procjena aritmetičke sredine na osnovu uzorka naziva se točkastom procjenom aritmetičke sredine osnovnoga skupa. Da bi se dobila intervalna procjena aritmetičke sredine, potrebno je u račun uzeti varijabilitet jedinica u usnovnome skupu (izražen preko varijance osnovnoga skupa ili njene ocjene pomoću uzorka) i veličinu uzorka.

Intervalna procjena aritmetičke sredine osnovnoga skupa dobiva se ovako:

$Pr\left\{\hat{\bar{X}} - z \cdot Se(\bar{x}) < \bar{X} < \hat{\bar{X}} + z \cdot Se(\bar{x})\right\} = 1 - \alpha$

Koeficijent "z" dobija se iz tablice površina ispod normalne krivulje kada veličina uzorka premaši 30 jedinica. Nivo pouzdanosti procjene jednak je 1-α. Ukoliko je n≤30 jedinica, upotrebljava se "t" iz Studentove distribucije umjesto koeficijenta "z".

Interval povjerenja procjene totala osnovnoga skupa glasi:

$Pr\left\{N \cdot \hat{\bar{X}} - z \cdot N \cdot Se(\bar{x}) < \sum_{i=1}^{N} x_i < N \cdot \hat{\bar{X}} + z \cdot N \cdot Se(\bar{x})\right\} = 1 - \alpha$

U gornjem izrazu za procjenu totala standardna greška procjene totala osnovnoga skupa jednaka je:

$Se(\sum x_i) = N \cdot Se(\bar{x})$

Točkasta procjena totala na osnovu uzorka jednaka je:

$\sum_{i=1}^{N} x_i = N \cdot \hat{\bar{X}}$

Procjena totala na osnovu uzorka također je nepristrana jer je procjena aritmetičke sredine nepristrana.

10. Procjena proporcije (relativne frekvencije) na osnovu uzorka.

Veličina uzorka određuje se sljedećom formulom:

$n' = \left[\frac{z \cdot \sqrt{P \cdot Q}}{gre\check{s}ka}\right]^2$ PQ=varijanca osnovnog skupa

Ukoliko je umjesto varijance osnovnoga skupa poznat koeficijent varijacije osnovnoga skupa, tada formula poprima sljedeći oblik:

$n' = \left[\frac{z \cdot \sqrt{\frac{Q}{P}}}{rel.gre\check{s}ka}\right]^2$

Ukoliko je frakcija izbora sada veća od 0.05, onda se pristupa korekciji kao i pri procjeni aritmetičke sredine osnovnoga skupa.
Intervalna procjena proporcije (odnosno relativne frekvencije) je sljedeća:
$Pr\{\hat{p} - z \cdot Se(p) < P < \hat{p} + z \cdot Se(p)\} = 1 - \alpha$
Pri tome se standardna greška procjene proporcije dobiva na sljedeći način:
$Se(p) = \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$ ako je n>30 jedinica, odnosno:
$Se(p) = \sqrt{\frac{\hat{p} \cdot \hat{q}}{n-1}}$ ako je n≤30 jedinica
Ukoliko je frakcija izbora veća od 0.05, tada se formule za standardnu grešku korigiraju (množe) s korektivnim faktorom
$\sqrt{\frac{N-n}{N-1}}$ kao i pri procjeni aritmetičke sredine osnovnoga skupa.

11. Procjena varijance na osnovu uzorka.

Sampling distribucija varijanci ima oblik Hi-kvadrat distribucije. Ukoliko uzorak raste u beskonačnost, sampling distribucija teži k normalnome obliku. Intervalna procjena varijance može se dobiti na sljedeći način:
$Pr\left\{\frac{n \cdot s^2}{\chi^2_{\alpha/2}} < \frac{n \cdot s^2}{\chi^2_{1-\alpha/2}}\right\} = 1 - \alpha$
Vrijednost Hi-kvadrata očitavamo iz odgovarajućih tablica Hi-kvadrat distribucije na određenom nivou α i uz određeni broj stupnjeva slobode df=v=n-1.
Ukoliko veličina uzorka prelazi 30 jedinica Hi-kvadrat distribucija može se aproksimirati pomoću normalne distribucije. Tada se vrijednosti Hi-kvadrata mogu dobiti pomoću sljedeće formule.
$\chi^2_{\alpha/2} = \frac{1}{2} \cdot (z_{\alpha/2} + \sqrt{2 \cdot v - 1})^2$
$\chi^2_{1-\alpha/2} = \frac{1}{2} \cdot (-z_{\alpha/2} + \sqrt{2 \cdot v - 1})^2$
Intervalna procjena standarde devijacije dobiva se tako da se izračuna pozitivni korijen iz donje i gornje granice intervalne procjene varijance osnovnoga skupa.

2. KOLOKVIJ

12. Testiranje hipoteze o nepoznatoj aritmetičkoj sredini. Dvosmjerni i jednosmjerni testovi.

Postavljaju se sljedeće hipoteze:
$H_0 = \ldots \bar{X} = \bar{X}_0$
$H_1 = \ldots \bar{X} \neq \bar{X}_0$
Nultom hipotezom pretpostavlja se da je aritmetička sredina osnovnoga skupa jednaka nekoj pretpostavljenoj vrijednosti, a alternativnom suprotno.
Interval prihvaćanja nulte hipoteze H_0 glasi:
$\bar{X}_0 \pm z \cdot \mathrm{Se}(\bar{x})$
Ako se aritmetička sredina nalazi u navedenom intervalu, prihvaćamo nultu hipotezu H_0. Ukoliko se aritmetička sredina uzorka nalazi izvan navedenog intervala, prihvaća se alternativna hipoteza H_1.
Testiranje se može izvesti tzv. "t" testom:
$t^* = \frac{\bar{X} - \bar{X}_0}{Se(\bar{x})}$

Ako je n>30 jedinica onda se upotrebljava se z^*.
Zaključci se izvode na sljedeći način:
ako je $-t_{\frac{\alpha}{2}} < t^* < +t_{\frac{\alpha}{2}}$ prihvaća se hipoteza H_0
ako je $-t_{\frac{\alpha}{2}} > t^* ili \ + t_{\frac{\alpha}{2}} < t^*$ prihvaća se hipoteza H_1
Testiranje hipoteze o nepoznatoj aritmetičkoj sredini osnovnoga skupa može se postaviti i jednosmjerno.
Pri testiranju na donju granicu postavlja se alternativna hipoteza da je aritmetička sredina osnovnoga skupa manja od neke pretpostavljene vrijednosti. Hipoteza H_0 glasi da je aritmetička sredina osnovnoga skupa veća ili jednaka nekoj pretpostavljenoj vrijednosti.
$H_0 = \ldots \bar{X} \geq \bar{X}_0$
$H_1 = \ldots \bar{X} < \bar{X}_0$
Donja granica prihvaćanja H_0 glasi:
$DG = \bar{X}_o - z \cdot Se(\bar{x})$
Ako aritmetička sredina uzorka ima vrijednost manju od donje granice prihvaćanja nulte hipoteze, prihvaća se alternativna hipoteza da je aritmetička sredina manja od neke pretpostavljene vrijednosti.
Zaključci se izvode na sljedeći način:
ako je $t^* < -t_\alpha$ prihvaća se alternativna hipoteza H_1
ako je $t^* > -t_\alpha$ prihvaća se hipoteza H_0
Testiranje na gornju granicu vrši se:
$H_0 = \ldots \bar{X} \leq \bar{X}_0$
$H_1 = \ldots \bar{X} > \bar{X}_0$
Gornja granica prihvaćanja glasi:
$GG = \bar{X}_o + z \cdot Se(\bar{x})$
Ukoliko aritmetička sredina uzorka premaši gornju granicu prihvaćanja nulte hipoteze prihvaća se alternativna hipoteza.

13. Testiranje hipoteze o nepoznatoj proporciji (relativnoj frekvenciji). Dvosmjerni i jednosmjerni testovi.

Nultom hipotezom pretpostavlja se da je proporcija osnovnoga skupa jednaka nekoj pretpostavljenoj vrijednosti:
$H_0 = \ldots P = P_0$
$H_1 = \ldots P \neq P_0$
Interval prihvaćanja nulte hipoteze glasi:
$P_0 \neq z \cdot Se(p)$
Standardna greška računa se:
$Se(p) = \sqrt{\frac{P_0 \cdot Q_0}{n}}$ ako je n>30 jedinica
$Se(p) = \sqrt{\frac{P_0 \cdot Q_0}{n-1}}$ ako je n≤30 jedinica
Ako je uzorak mali onda se umjesto "z" upotrebljava vrijednost "t" iz Studentove distribucije s n-1 stupnjeva slobode.
Ako se proporcija iz uzorka nalazi u gornjem intervalu prihvaćanja, tada prihvaćamo H_0 hipotezu kao istinitu. Ako se proporcija uzorka nalazi izvan intervala prihvaćanja tada prihvaćamo kao moguću alternativnu hipotezu.
Nultom hipotezom pretpostavlja se da nema rezlike u proporcijama dvaju osnovnih skupova, odnosno da uzorci koje smo odabrali potječu iz istih populacija. Hipoteze glase:
$H_0 = \ldots P_1 = P_2$

$H_1 = \ldots P_1 \neq P_2$
Interval prihvaćanja nulte hipoteze glasi:
$0 \pm z \cdot Se(p_1 - p_2)$
Standardna greška može se izračunati:
$$Se(p_1 - p_2) = \sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}$$

14. Testiranje hipoteze o razlici aritmetičkih sredina dvaju nezavisnih osnovnih skupova.

Postavlja se nulta hipoteza da nema značajne razlike između aritmetičkih sredina dvaju nezavisnih osnovnih skupova. Nezavisni uzorci potječu iz različitih osnovnih skupova te među njima ne postoji povezanost. Ukoliko se na istome uzorku vrše različita mjerenja pod različitim uvjetima, onda govorimo o zavisnim uzorcima. Kod zavisnih uzoraka postoji korelacija između rezultata prije i poslije eksperimenata. Kod nezavisnih uzoraka standardna greška je veća negoli kod zavisnih.
$H_0 = \ldots \bar{X}_1 - \bar{X}_2 = 0$
$H_1 = \ldots \bar{X}_1 - \bar{X}_2 \neq 0$
Interval prihvaćanja nulte hipoteze glasi:
$0 \pm z \cdot Se(\bar{x}_1 - \bar{x}_2)$
Standardna greška u navedenome testiranju računa se ovako:
$$Se(\bar{x}_1 - \overline{x_2}) = \sigma \cdot (\frac{1}{n_1} + \frac{1}{n_2})$$
Pretpostavlja se da su varijance osnovnih skupova jednake, ali su nepoznate, te se ocjena njihove vrijednosti dobiva iz uzoraka, pa se izraz za standardnu grešku svodi na:
$$Se(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{n_1 \cdot \bar{\sigma}_1^2 + n_2 \cdot \bar{\sigma}_2^2}{n_1 + n_2 - 2}\right) \cdot (\frac{n_1 + n_2}{n_1 \cdot n_2})}$$
Ako se radi o velikome uzorku:
$$Se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
Treba naći apsolutnu vrijednost razlike aritmetičkih sredina uzoraka. Ukoliko ta razlika premaši interval prihvaćanja nulte hipoteze, tada prihvaćamo alternativnu hipotezu kao istinitu, tj. zaključujemo da se aritmetičke sredine dvaju osnovnih skupova značajno razlikuju.
Pomoću "z" testa:
$$z^* = \frac{|\bar{X}_1 - \bar{X}_2|}{Se(\bar{x}_1 - \bar{x}_2)}$$
ako je $z^* < z_{\frac{\alpha}{2}}$ prihvaćamo hipotezu H_0
ako je $z^* > z_{\frac{\alpha}{2}}$ prihvaćamo hipotezu H_1
Vrijednost "t" se upotrebljava umjesto "z" ukoliko zbroj veličina dvaju uzoraka ne prelazi 32. Naime stupnjevi slobode se određuju kod ovog testiranja na sljedeći način:
$df = n_1 - n_2 - 2$

15. Testiranje hipoteze o razlici aritmetičkih sredina dvaju zavisnih osnovnih skupova.

Postavljaju se sljedeće hipoteze:
$H_0 = \ldots \bar{X}_1 - \bar{X}_2 = 0$
$H_1 = \ldots \bar{X}_1 - \bar{X}_2 \neq 0$
Interval prihvaćanja H_0 glasi:
$0 \pm z \cdot Se(\bar{x}_1 - \bar{x}_2)$

Standardna greška u navedenome testiranju računa se:

$$Se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2 \cdot r_{1,2} \cdot Se(\bar{x}_1) \cdot Se(\bar{x}_2)}$$

gdje $r_{1,2}$ predstavlja Pearsonov koeficijent linearne korelacije između dvaju mjerenja iste slučajne varijable na istome uzorku. Što je jača korelacija među mjerenjima standardna greška je manja.

16. Testiranje hipoteze o nezavisnosti obilježja.

Nulta hipoteza glasi da ne postoji zavisnost dvaju obilježja. Hipoteze glase:

$H_0 = \ldots P_{ij} = P_{i\cdot} \cdot P_{\cdot j}$ za svaki i i j

$H_1 = \ldots \exists P_{ij} \neq P_{i\cdot} \cdot P_{\cdot j}$

Test veličina je sljedeća:

$$\chi^{*2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(m_{ij} - e_{ij})^2}{e_{ij}}$$

m_{ij} – očekivanje frekvencije (empirijske)

e_{ij} - očekivane (teorijske) frekvencije

Očekivane frekvencije izračunavaju se na sljedeći način:

$$e_{ij} = \frac{m_{i\cdot} \cdot m_{\cdot j}}{e_{ij}}$$

$m_{i\cdot}$ - marginalna frekvecija i-tog retka

$m_{\cdot j}$ - marginalna frekvencija j-tog stupca

n – veličina uzorka

Navedena veličina uspoređuje se s kritičnom vrijednošću Hi-kvadrata iz tablica. Broj stupnjeva slobode jednak je $df = (r-1) \cdot (c-1)$, gdje oznaka r označava broj redaka, a oznaka c broj stupaca u tabeli kontigence. Ukoliko empirijska vrijednost Hi-kvadrat premaši kritičnu vrijednost Hi-kvadrata iz tablica, prihvaćamo nultu hipotezu kao istinitu, tj. zaključujemo da nema ovisnosti obilježja osnovnoga skupa. U drugom slučaju, ako vrijednost empirijskog Hi-kvadrata premaši kritičnu vrijednost prihvaćamo alternativnu hipotezu, tj. zaključujemo da postoji zavisnost dvaju obilježja elemenata osnovnoga skupa. Tabela iz koje se izračunava vrijednost Hi-kvadrata naziva se tabelom kontigence. Hi-kvadrat testom ustanovljava se samo vjerojatnost povezanosti dviju varijabli, ali ne i visina povezanosti. Aproksimativnu visinu povezanosti možemo ustanoviti pomoću Peasnonovoga koeficijenta kontigence:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Najmanja vrijednost koeficijenta kontigence iznosi nula.

17. Testiranje hipoteze da distribucija ima neki određeni oblik.

Može se testirati hipoteza da distribucija slučajne varijable X ima oblik neke teorijske distribucije. Kod svih testiranja test-veličina je vrijednost empirijskoga Hi-kvadrata:

$$\chi^{*2} = \sum_{j=1}^{k} \frac{(f_i - f_{ti})^2}{f_{ti}}$$

f_i - originalne frekvencije

f_{ti} - teorijske frekvencije koje se izračunavaju pod pretpostavkom da distribucija ima oblik neke teorijske distribucije. Teorijske frekvencije za diskontinuirane slučajne varijable dobivaju se ovako:

$$f_{ti} = \left(\sum_{i=1}^{k} f_i\right) \cdot p(x_i)$$

gdje $p(x_i)$ predstavlja vjerojatnost da slučajna varijabla X poprimi vrijednost x_i prema zakonu vjerojatnosti slučajne varijable koje smo pretpostavili nultom hipotezom.
Empirijska vrijednost Hi-kvadrata se uspoređuje s kritičnom vrijednossću Hi-kvadrata iz tablica na određenom nivou signifikantnosti testa α i uz određeni broj stupnjeva slobode df=v.
Broj stupnjeva slobode izračunava se na sljedeći način:
- za jednoliku distribuciju df=k-1
- za binomnu distribuciju df=k-2
- za Poissonovu distribuciju df=k-2
- za normalnu distribuciju df=k-3

Veličina "k" predstavlja broj frekvencija.
Ukoliko empirijska vrijednost Hi-kvadrata ne premaši kritičnu vrijednost iz tablica, prihvaćamo hipotezu da empirijska distribucija ima pretpostavljeni oblik. Ukoliko je empirijska vrijednost Hi-kvadrata veća od kritične vrijednosti iz tablica, prihvaćamo alternativnu hipotezu da distribucija nema pretpostavljeni oblik.

18. Analiza varijance s jednim i s dva promjenjiva faktora. Analiza varijance na osnovu rangova.

U svrhu djelovanja faktora A na vrijednost slučajne varijable X potrebno je uzeti onoliko uzoraka koliko ima varijacija faktora A (A_1, A_2, ..., A_k). Imamo k uzoraka od kojih svaki ima n_j elemenata. Navedeni uzorci mogu, ali ne moraju imati isti broj elemenata. Ovo testiranje svodi se na testiranje značajnosti razlike aritmetičkih sredina osnovnih skupova (populacija), odnosno testiramo hipotezu da svi uzorci potječu iz iste populacije. U postupku testiranja značajnosti djelovanja promjenjivog faktora A na slučajnu varijablu X postavljaju se sljedeće hipoteze:
$H_0 \dots \sigma_A^2 = 0$
$H_1 \dots \sigma_B^2 \neq 0$
Testiranje se vrši F-testom. F-omjer predstavlja omjer varijance koja se pripisuje djelovanju promjenjivog faktora A i varijance unutar uzoraka, koja nije uzrokovana djelovanjem promjenjivog faktora A. Unutar svakog pojedinog uzorka djeluje samo jedna varijacija faktora A, tj. A_i. Pretpostavke za primjenu analize varijacije s jednim promjenjivim faktorom su da uzorci potječu iz normalnih populacija i da imaju jednake varijance. No, ti uvjeti ne moraju biti ispunjeni ako su uzorci jednake ili slične veličine i ako su populacije međusobno slične u odstupanju od normalne distribucije. Ocjene varijance dobivaju se iz odgovarajućih uzoraka na čije vrijednosti varijable X su djelovale varijacije faktora A.
Zbroj kvadrata ukupnih odstupanja vrijednosti varijable X od očekivane vrijednosti (aritmetičke sredine) može se raščlaniti na sljedeći način:
$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{.j})^2 + \sum_{j=1}^{k} n_j \cdot (\bar{X}_{.j} - \bar{X}_{..})$
(ukuono) (unutar uzorka) (između uzorka)
Izraz sa lijeve strane predstavlja ukupan zbroj kvadrata odstupanja vrijednosti slučajne varijable X od aritmetičke sredinesvih uzoraka. Prvi dio izraza s desne strane predstavlja zbroj kvadrata odstupanja vrijednosti varijable X od aritmetičke sredine unutar uzoraka. Drugi dio izraza s desne strane predstavlja zbroj kvadrata odstupanja aritmetičkih sredina uzoraka od zajedničke aritmetičke sredine, odnosno zbroj kvadrata odstupanja između uzoraka. Zbroj kvadrata odstupanja vrijednosti numeričke varijable između uzoraka predstavlja u stvari efekat djelovanja faktora A.
Analiza varijance sa dva promjenjiva faktora – na vrijednost neke slučajne varijable X može djelovati više od jednoga promjenjivog faktor. Ovdje će biti prikazano statističko testiranje značajnosti djelovanja dvaju promjenjivih faktora na vrijednost slučajne varijable X. Promjenjive faktore čiji utjecaj testiramo nazivamo A i B. Analiza varijance sa

dvapromjenjiva faktora također se svodina hipotezu o razlici aritmetičkih stredina više osnovnih skupova, ali sada po 2 različita faktora koji djeluju na istu numeričku varijablu. Pretpostavke koje bi trebale biti ispunjene su prigodom testiranja razlika u aritmetičkim sredinama vrijede kao i kad analize varijance s jednim promjenjivim faktorom. U postupku testiranja značajnosti djelovanja varijacija u faktorima A i B postavljaju se sljedeće hipoteze:
$H_0 \dots \sigma_A^2 = 0$ odnosno $H_0 \dots \sigma_B^2 = 0$
$H_1 \dots \sigma_B^2 \neq 0$ odnosno $H_1 \dots \sigma_B^2 \neq 0$
Testiranje gornjih hipoteza vrši se standardnim F-testom. Omjer dvaju varijanci uvijek pripada F-distribuciji s određenim stupnjevima slobode u brojniku i nazivniku i uz određenu signifikantnost α. Dekompozicija ukupnog zbroja kvadrata odstupanja varijable X od sredine može se prikazati ovako:
(ukupan zbroj kvadrata)=(zbroj kvadrata između redaka)+(zbroj kvadrata između stupaca)+(rezidualni zbroj kvadrata)
Analiza varijance na osnovu rangova – u situaciji kada neki od uvjeta za provedbu tih testova nisu ispunjeni (npr. ako su uzorci uzeti iz populacija koje nisu normalno distribuirane ili nisu sa jednakom varijancom ili kada se podaci za statističku analizu sastoje samo od rangova) upotrebljavamo neparametarski Kruskal-Wallis test. Test veličina je:
$H = \frac{12}{n(n+1)} \cdot \sum_{i=1}^{k} \frac{R_j^2}{n_i} - 3(n+1)$
Postupak testiranja provodi se ovim koracima:
1. Opservacije iz uzoraka n_1, n_2, ...n_k iz k uzoraka kombiniraju se u jedinstveni niz veličine n i poredaju se prema veličini od najmanjega prema najvećemu. Zatim se opservaciju zamjenjuju njihovim rangovima, počevši od 1, koji se daje najmanjoj opservaciji do n, koji se daje najvećoj opservaciji. Ukoliko dvije opservacije imaju istu vrijednost, onda im se daje aritmetička sredina rangova.
2. Rangovi koji su dodijeljeni opservacijama u svakoj od k grupa zbrajaju se posebno te dobivamo k sumu rangova
3. Izračunava se test statistika prema prije navedenoj formuli
4. Ukoliko se radi o 3 uzorka ili 5 ili manje opservacija, signifikantnost navedene test-veličine određuje se iz posebnih tablica. Ukoliko ima više od 5 opservacija i jednome ili više uzoraka, vrijednost H se uspoređuje s tabličnim vrijednsotima χ^2 s k-1 stupnjeva slobode.

19. Linearna korelacija. Testiranje značajnosti koeficijenta linearne korelacije.

Najvžnija mjera linearne korelacije među slučajnim varijablama naziva se Pearsonov koeficijent linearne korelacije. Ovako se izračunava:
$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$
Izraz u brojniku podijeljen s n naziva se kovarijancom, ili prvim mješovitim momentom oko vrijednosti aritmetičkih sredina slučajnih varijabli X i Y. Koeficijent linearne korelacije može se izračunati i ovako:
$r = \frac{\sum_{i=1}^{n} X_i \cdot Y_i - n \cdot \bar{X} \cdot \bar{Y}}{n \cdot \sigma_X \cdot \sigma_Y}$
Vrijednost koeficijenta nalazi se između -1 i +1. Ukoliko vrijednost koeficijenta korelacije iznosi 0, korelacija uopće ne postoji. Ako vrijednost koeficijenta korelacije iznosi +1, radi se o pozitivnoj funkcionalnoj povezanosti, dok vrijednost koeficijenta korelacije od -1 znači funkcionalnu negativnu povezanost između slučajnih varijabli. Vrijednost koeficijenta korelacije između tih vrijednosti označava stohastičku (ili statističku) vezu.
$-1 \leq r \geq +1$
Testiranje značajnosti koeficijenta linearne korelacije provodi se t-testom

$H_0 \ldots r = 0$
$H_1 \ldots r \neq 0$
Ukoliko je koeficijent linearne korelacije osnovnoga skupa jednak nuli, onda sampling distribucija koeficijenta linearne korelacije teži normalnome obliku ako broj parova vrijednosti slučajnih varijabli X i Y teži k beskonačnosti. Ukoliko je broj parova mali (n<30) može se koristiti Studentova t-distribucija. Test veličine je:
$t^* = \frac{\hat{r}}{Se(r)}$
Za velike uzorke standardna greška se računa ovako:
$Se(r) = \sqrt{\frac{1}{r-1}}$
dok se za male uzorke izračunava ovako:
$Se(r) = \sqrt{\frac{1-\hat{r}^2}{n-2}}$
Koeficjent linearne korelacije koji je statistički značajan ne mora biti i prekatično značajan. Događa se da veoma mali koeficijent linearne korelacije bude statistički značajan zbog velikoga uzorka, odnosno male standardne greške. U tome slučaju vrijednost empirijskoga t-omjera je velika. Primjena velikoga uzorka ide u korist odbacivanja nulte hipoteze, što rađa opasnost da se trivijalni rezultati prihvate kao značajni. Obično se uzima ovo:
$r > 0.80$ radi se o jakoj pozitivnoj korelaciji
$0.50 < r \leq 0.80$ radi se o srednje jakoj pozitivnoj korelaciji
$0 < r \leq 0.50$ radi se o slaboj pozitivnoj korelaciji
Na isti način zaključuje se ako je korelacija negativna.

20. Korelacija ranga. Testiranje značajnosti koeficijenta korelacije ranga.

Od svih koeficijenata koji koriste rangove umjesto originalnih vrijednosti najviše u uporabi je Spearmanov koeficijent korelacije ranga. Njegova vrijednost kreće se u istom rasponu kao i koeficijent linearne korelacije između redoslijednih obilježja i obilježja ranga, ali se može koristiti i kod numeričkih varijabli ukoliko nisu zadovoljene pretpostavke za primjenu Pearsonovoga koeficijenta linearne korelacije. U tome slučaju ne koriste se numeričke vrijednosti varijable X i Y nego njihovi rangovi. Računa se ovako:
$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d^2}{n^3 - n}$
gdje n predstavlja broj parova vrijednosti X i Y, a d_i predstavlja razliku rangova vrijednosti X i Y. Broj 6 je stalan i ne ovisi o broju parova vrijednosti X i Y.

21. Linearna regresija.

Kod regresijske analize treba odrediti koja će varijabla biti regresand. Nezavisne varijable u modelu obično se nazivaju regresorskim varijablama. U velikom broju primjena veza između varijable je linearna, ali postoje također i slučajevi nelinearne regresije.
Regresijski model izgleda ovako:
$Y = \beta_0 + \beta_1 \cdot X + e$
gdje je "e" slučajna greška koja mora imati sljedeće uvjete:
1. $E(e) = 0$ za svaki i
2. $E(e_i, e_j) = \sigma_e^2 < +\infty$ za i=j
3. $E(e_i, e_j) = 0$ za svaki i≠j
4. $E(e_i, X_i) = 0$
Ocjene parametara dobiju se metodom najmanjih kvadrata:

$\min \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Standardna greška regresije dobija se ovako:

$\hat{\sigma}_{\hat{Y}}^2 = \sqrt{\frac{SR}{n-2}}$

dok se koeficijent varijacije regresije dobiva kao omjer standardne greške regresije i aritmetičke sredine varijable Y

$\hat{V}_{\hat{Y}} = \frac{\hat{\sigma}}{\bar{Y}} \cdot (100)$

Ukoliko se za regresand varijablu postavi varijabla X umjesto varijable Y, dobiva se regresijski pravac koji će se sa pravcem Y poklapati samo u slučaju perfektne zavisnosti između X i Y.