OVS - TEORIJA

POSKUS IN DOGODEK

POSKUS: vržemo kovanec.
DOGODEK: pade grb.

VERJETNOST DOGODKA: je verjetnost, da se dogodek pri izbranem poskusu zgodi. Verjetnost je matematični objekt, realno število med 0 in 1.

VERJETNOST, STATISTIČNA DEFINICIJA: poskus G, dogodek D, poskus velikokrat ponovimo, N - št. ponovitev poskusa, k - št. ponovitev poskusa, ki so ugodni za dogodek D. Kvocient: $k/N \rightsquigarrow q \in [0,1]$.
Poskus .. množica G.
Dogodek .. podmnožica $D \subseteq G$.

VERJETNOST DOGODKA, STATISTIČNA DEFINICIJA: $P(A) \sim \frac{k}{N}$.

VERJETNOST DOGODKA - KLASIČNA DEFINICIJA.

$G = \{a_1, a_2, \dots, a_k\}$. $p_i = P(\{a_i\})$.
$\sum_{i=1}^{k} p_i = \sum_{a_i \in G} P(\{a_i\}) = 1$

$A \subseteq G$. $P(A) = \sum_{a_i \in A} P(\{a_i\}) = \sum_{a_i \in A} p_i$

RAČUNANJE Z DOGODKI

A - dogodek.
$\bar{A} = G \setminus A$ - nasprotni dogodek
$A \cap B$ - presek dogodkov (Dogodek, ko se zgodita oba dogodka A in B.)
$A \cup B$ - unija dogodkov (Dogodek, ko se zgodi vsaj eden od A, B.)
$N = \emptyset$ - nemogoč dogodek
G - gotov dogodek

IZREK: $P(\bar{A}) = 1 - P(A)$. Dokaz!
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Dokaz!
Dogodka A in B sta NEZDRUŽLJIVA, če je $A \cap B = N$.

IZREK: $\overline{A \cup B} = \bar{A} \cap \bar{B}$
$\overline{A \cap B} = \bar{A} \cup \bar{B}$

$A = A \cap G = A \cap (B \cup \bar{B}) = (A \cap B) \cup (A \cap \bar{B})$

POGOJNA VERJETNOST

$P(B) > 0$. $P(A|B)$ .. verjetnost, da se zgodi A, pri pogoju, da se je zgodil B.

IZREK: če je $P(B) > 0$ (ni nemogoč dogodek), potem velja: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
Oziroma: $P(A \cap B) = P(B) \cdot P(A|B)$, če $P(B) > 0$.
$P(A \cap B) = P(A) \cdot P(B|A)$, če $P(A) > 0$.

NEODVISNI DOGODKI

Dogodka A in B sta NEODVISNA, če velja: $P(A) = P(A|B)$, $P(B) = P(B|A)$.

IZREK: Dogodka A in B sta neodvisna ntk. je $P(A \cap B) = P(A) \cdot P(B)$. Dokaz!

POPOLN SISTEM DOGODKOV

Družina dogodkov $A_1, A_2, \dots, A_k$ je POPOLN SISTEM DOGODKOV, če je:
- $A_1 \cup A_2 \cup \dots \cup A_k = \Omega$ (vsaj eden izmed njih se gotovo zgodi)
- $A_i \cap A_j = \emptyset$, za vse $i \neq j$. (vsaka 2 dogodka A_i in A_j sta nezdružljiva)

Zgodi se natančno eden izmed $A_1, A_2, \dots, A_k$.

DVOFAZNI POSKUS

1. faza: Nastopi natanko ena od hipotez $H_1, H_2, \dots, H_k$.
2. faza: Opazujemo dogodek A.
Zanima nas: $P(A)$.
Poznamo: $P(H_1), P(H_2), \dots, P(H_k)$ in $P(A|H_1), P(A|H_2), \dots, P(A|H_k)$.

$$P(A) = P(H_1) \cdot P(A|H_1) + P(H_2) \cdot P(A|H_2) + \dots + P(H_k) \cdot P(A|H_k)$$

↓

FORMULA O (PO)POLNI VERJETNOSTI

IZREK: Naj bo $H_1, H_2, \dots, H_k$ popoln sistem dogodkov. Potem je

$$P(A) = P(H_1) \cdot P(A|H_1) + P(H_2) \cdot P(A|H_2) + \dots + P(H_k) \cdot P(A|H_k).$$ Dokaz!

BAYESOV OBRAZEC

IZREK: $P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$

→ Detektor laži, Testiranje redke bolezni, Predavanja in študenti.

ZAPOREDJA NEODVISNIH POSKUSOV

Velikokrat ponovimo isti poskus. Vsakič se zgodi dogodek A z enako verjetnostjo p.
Dogodek A je v VSAKEM poskusu NEODVISEN od prejšnjih izidov.

BERNOULLIJEV OBRAZEC

S $P(n; p; k)$ označimo verjetnost, da se pri n ponovitvah poskusa natanko k-krat zgodi dogodek A, katerega verjetnost v enem poskusu je enaka p.

IZREK: $P(n; p; k) = \binom{n}{k} p^k (1-p)^{n-k}$. Dokaz!

permutacije $P(A)\,P(A)\,P(\bar{A})\,P(\bar{A})$

SLUČAJNE SPREMENLJIVKE

SLUČAJNA SPREMENLJIVKA je funkcija, katere funkcijska vrednost ni odvisna od vhodnih podatkov / parametrov, temveč od slučaja.
Privzeli bomo, da slučajne spremenljivke za funkcijske vrednosti vračajo realna št.
Npr.: met poštene igralne kocke, rezultat je št. iz $\{1, 2, 3, 4, 5, 6\}$.

Slučajno spremenljivko X popolnoma opišemo z ZALOGO VREDNOSTI Z_X in za vsak $x_k \in Z_X$ moramo poznati $p_k = P(X = x_k)$.
$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$

Tgornjemu opisu slučajne spremenljivke X pravimo VERJETNOSTNA SHEMA.
Pri tem je $p_i \geq 0$ za vse $i = 1, \dots, n$ in velja $\sum_{i=1}^{n} p_i = 1$.

Dve pomembni porazdelitvi:
- ENAKOMERNA PORAZDELITEV je slučajna spremenljivka, pri kateri imajo vsi izidi enake verjetnosti.
  $X \sim \begin{pmatrix} x_1 & \dots & x_n \\ \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$
- BINOMSKA PORAZDELITEV $b(n, p)$. $X \sim \begin{pmatrix} 0 & 1 & \dots & n \\ P(n;p;0) & P(n;p;1) & \dots & P(n;p;n) \end{pmatrix}$
  $P(n; p; k) = P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$

MATEMATIČNO UPANJE

MATEMATIČNO UPANJE SLUČAJNE SPREMENLJIVKE X, $E(X)$, definiramo kot $E(X) = \sum_{i=1}^{n} x_i p_i$.
→ Matematično upanje spremenljivke $X \sim b(n, p)$? ... $= n \cdot p$?!

DISPERZIJA

DISPERZIJA SLUČAJNE SPREMENLJIVKE X, $D(X)$, meri, kako slučajna spremenljivka X odstopa od matematičnega upanja.

$$D(X) = E((X - E(X))^2) = \sum_{i=1}^{n} (x_i - E(X))^2 p_i$$

IZREK: $D(X) = E(X^2) - E(X)^2$. Dokaz!

STANDARDNI ODKLON

STANDARDNI ODKLON ali STANDARDNA DEVIACIJA slučajne spremenljivke X, $\sigma(X)$, definiramo kot $\sigma(X) = \sqrt{D(X)}$.

IZREK: Naj bo slučajna spremenljivka X porazdeljene z binomsko porazdelitvijo $b(n,p)$. Potem je:
$E(X) = n \cdot p$
$D(X) = n \cdot p \cdot (1-p)$
$\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$

→ Indikatorska slučajna spremenljivka

NEENAKOST MARKOVA

IZREK: Naj velja $X \geq 0$. Za vsak $a > 0$ velja neenakost: $P(X \geq a) \leq \frac{E(X)}{a}$.
Dokaz!

NEENAKOST ČEBIŠEVA

IZREK: Naj bo X slučajna spremenljivka. Za vsak $t > 0$ velja neenakost:
$P(|X - E(X)| \geq t) \leq \frac{D(X)}{t^2}$. Dokaz!

→ UPORABA: 1000-krat vržemo kovanec. Oceni verjetnost, da bo št. grbov med 400 in 600.

LASTNOSTI UPANJA IN DISPERZIJE

IZREK: Če je $a \in \mathbb{R}$ in $P(X = a) = 1$, potem je $E(X) = a$ in $D(X) = 0$.

IZREK: Matematično upanje je linearno. Naj bosta X in Y slučajni spremenljivki, $E(X)$ in $E(Y)$ njuni matematični upanji in a, b poljubni realni števili. Potem je
$E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$. Dokaz!
* Analogna trditev za disperzijo ne velja!

DEF.: Slučajni spremenljivki X in Y sta NEODVISNI, če za vsak $a, b \in \mathbb{R}$ velja, da sta dogodka $(X \geq a)$ in $(Y \geq b)$ neodvisna.
$P(X \geq a \wedge Y \geq b) = P(X \geq a) \cdot P(Y \geq b)$.

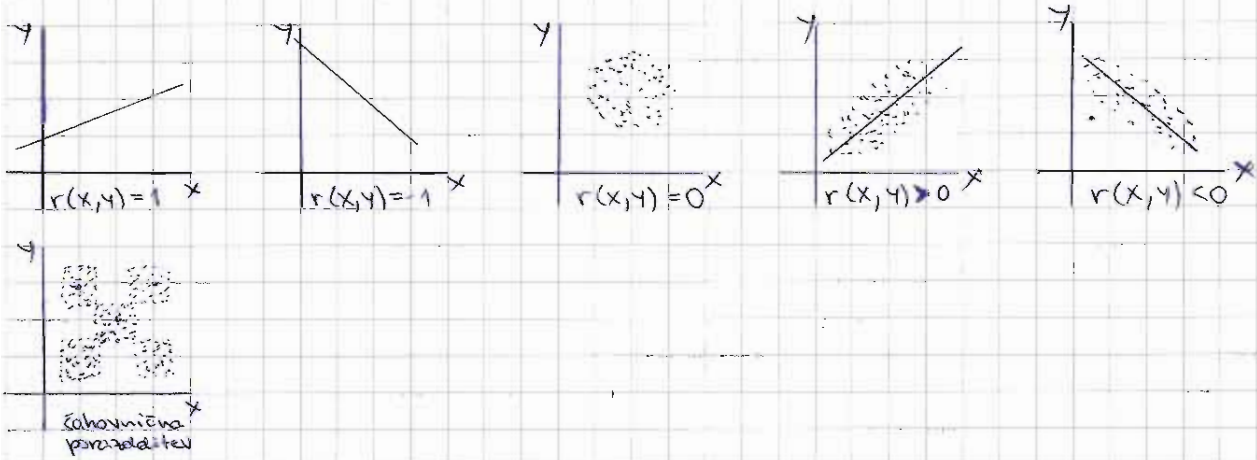
KOVARIANCA

Naj bosta X in Y slučajni spremenljivki. Količini $K(X,Y) = E((X - E(X))(Y - E(Y)))$ pravimo KOVARIANCA SLUČAJNIH SPREMENLJIVK X in Y.
Velja: $K(X,Y) = E(XY) - E(X) \cdot E(Y)$ in $|K(X,Y)| \leq \sigma(X) \cdot \sigma(Y)$. Dokaži!

KORELACIJA

Naj bosta X in Y slučajni spremenljivki. Izraz $r(X,Y) = \frac{K(X,Y)}{\sigma(X)\cdot\sigma(Y)}$ je KORELACIJSKI KOEFICIENT SLUČAJNIH SPREMENLJIVK X in Y. Velja: $-1 \leq r(X,Y) \leq 1$.

Kdaj je in kaj pomeni:
- $r(X,Y) = 1$ ... $Y = a\cdot X + b$, $a > 0$
- $r(X,Y) = -1$ ... $Y = -a\cdot X + b$, $a > 0$
- $r(X,Y) = 0$ ... sl. spr. sta NEKORELIRANI, sta neodvisni
- $r(X,Y) > 0$ ... POZITIVNO KORELIRANI
- $r(X,Y) < 0$ ... NEGATIVNO KORELIRANI



LASTNOSTI DISPERZIJE

IZREK: Za slučajni spremenljivki X in Y velja: $D(X+Y) = D(X) + D(Y) + 2\cdot K(X,Y)$.
Dokaz!

IZREK: Za neodvisni slučajni spremenljivki X in Y velja:

$$D(X+Y) = D(X) + D(Y)$$
$$D(a\cdot X) = a^2\cdot D(X)$$
$$\sigma(a\cdot X) = \sqrt{D(a\cdot X)} = \sqrt{a^2 D(X)} = |a|\cdot\sigma(X)$$

POSLEDICA: Slučajne spremenljivke $X_1, X_2, \ldots, X_n$ naj bodo enako porazdeljene, neodvisne, z matematičnim upanjem $E(X_i) = a$ in disperzijo $D(X) = \sigma^2$. Za slučajno spremenljivko $Y = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ velja:

$$E(Y) = \frac{1}{n}\cdot\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n\cdot a = a$$

$$D(Y) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

SLUČAJNE SPREMENLJIVKE Z NESKONČNO ZALOGO

ZVEZNO PORAZDELJENE SLUČAJNE SPREMENLJIVKE

Iščemo slučajno spremenljivko X, z zalogo vrednosti $[0,1]$, ki ima naslednjo lastnost: če je $J \subseteq [0,1]$ interval, potem je $P(X \in J)$ enaka dolžini intervala J.
Kako opisati X? $X_n \sim \begin{pmatrix} \frac{1}{n} & \frac{2}{n} & \frac{3}{n} & \dots & \frac{n}{n} \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$, $n = 100$.

$$P[X_n = [a,b]] = \frac{1}{100} \cdot (\#\text{ vrednosti oblike } \tfrac{k}{n} \text{ iz intervala } [a,b]) \approx$$
$$\approx \frac{1}{100} \cdot (b-a) \cdot 100 = (b-a)$$

$$P(X \in [a,b]) = \int_a^b g(x)\,dx = \int_a^b 1\,dx = (b-a). \qquad g_X(x) \geq 0, \quad \int_{-\infty}^{\infty} g_X(x)\,dx = 1$$

GOSTOTA VERJETNOSTI

Funkcija $g: \mathbb{R} \to \mathbb{R}$ je GOSTOTA VERJETNOSTI neke slučajne spremenljivke, če:
(G1) $g(x) \geq 0$ za vse $x \in \mathbb{R}$,
(G2) $\int_{-\infty}^{\infty} g(x)\,dx = 1$ in
(G3) $g(x)$ je odsekoma zvezna funkcija.

Zanimajo nas dogodki $P(X \in I)$. ← netrivialen integral

Kako izračunamo verjetnost dogodka: $P(a < X \leq b) = \int_a^b g_X(x)\,dx$
$a \in \mathbb{R}$, $P(X = a) = \int_a^a g_X(x)\,dx = 0$.

MATEMATIČNO UPANJE IN DISPERZIJA

Slučajna spremenljivka X naj bo opisana z gostoto verjetnosti $g_X(x)$.
$E(X) = \int_{-\infty}^{\infty} t \cdot g_X(t)\,dt$

$$D(X) = \int_{-\infty}^{\infty} (t - E(X))^2 g_X(t)\,dt = \int_{-\infty}^{\infty} t^2 g_X(t)\,dt - E(X)^2.$$

Integrala morata biti konvergentna.

ENAKOMERNA PORAZDELITEV

Zvezna slučajna spremenljivka X je razporejena ENAKOMERNO na intervalu $[a,b]$, $X \sim U[a,b]$, če je njena gostota verjetnosti g_X enaka

$$g_X(x) = \begin{cases} \frac{1}{b-a}; & x \in [a,b] \\ 0; & \text{sicer} \end{cases}$$

$E(X) = \frac{b+a}{2}$ $\qquad D(X) = \frac{1}{12}(b-a)^2$ → izpeljave!

„LIMITA" BINOMSKE PORAZDELITVE

Naj velja $X \sim b(n,p)$.

$$P(X=k) = \binom{n}{k} p^k (1-p)^k \approx \frac{1}{\sqrt{np(1-p)} \cdot \sqrt{2\pi}} \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right) \approx$$

$$P(X=k) = \binom{n}{k} \cdot p^k (1-p)^{n-k} \doteq \frac{1}{\sqrt{2\pi} \cdot \sigma(x)} \cdot \exp\left(-\frac{(k-E(x))^2}{2\sigma^2(x)}\right)$$

Opazujmo vse „binomske" porazdelitve z matematičnim upanjem a in disperzijo σ^2. Te porazdelitve se približujejo zvezno porazdeljeni slučajni spremenljivki X z gostoto verjetnosti enako

NORMALNA PORAZDELITEV

Slučajna spremenljivka X je porazdeljena NORMALNO z matematičnim upanjem a in disperzijo σ^2, $X \sim N(a,\sigma)$, če ima gostoto porazdelitve enako

$$g_X(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Slučajna spremenljivka Z je porazdeljena STANDARDNO NORMALNO, $Z \sim N(0,1)$, če ima gostoto

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$E(N(a,\sigma)) = a$ in $D(N(a,\sigma)) = \sigma^2$. Dokaži!

TRDITEV: Če $X \sim N(a,\sigma)$ in $Z = \frac{X-a}{\sigma}$, potem je $Z \sim N(0,1)$.

Če vemo, da je X porazdeljena normalno, lahko za računanje verjetnosti problem prevedemo na standardno normalno slučajno spremenljivko $Z \sim N(0,1)$.

$$\Phi(x) = \int_0^x \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt = P(0 \le Z \le x)$$

Φ je liha funkcija.

STANDARDIZACIJA SLUČAJNE SPREMENLJIVKE

Naj bo Y slučajna spremenljivka z upanjem a in disperzijo σ^2. Potem obstajajo takšni konstanti $c,d \in \mathbb{R}$, da ima slučajna spremenljivka $Z = cY + d$ matematično upanje enako 0 in disperzijo (in standardni odklon) enako 1.

PORAZDELITVENA FUNKCIJA

Naj bo X zvezno porazdeljena slučajna spremenljivka. PORAZDELITVENA FUNKCIJA $F_X(x)$ sl. spr. X je definirana s predpisom $F_X(x) = P(X \le x) = \int_{-\infty}^{x} g_X(t)\,dt$, pri čemer je $g_X(x)$ gostota verjetnosti spremenljivke X.

CENTRALNI LIMITNI IZREK

IZREK: Naj bo $X_1, X_2, X_3, X_4, \ldots$ zaporedje NEODVISNIH slučajnih spremenljivk, ki so vse enako porazdeljene (zato imajo isto mat. upanje in isto disperzijo) z matematičnim upanjem $E(X_i) = a$ in $D(X_i) = \sigma^2$.
Opazujemo slučajno spremenljivko $S_n = X_1 + X_2 + \ldots + X_n$.
Velja $E(S_n) = n \cdot a$ (ker je E linearna) ter $D(S_n) = n \cdot \sigma^2$ in $\sigma(S_n) = \sigma\sqrt{n}$ (ker so X_i neodvisne).
Za S_n velja:

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

$\frac{S_n - E(S_n)}{\sigma(S_n)}$ je razporejena PRIBLIŽNO kot $N(0,1)$.
S_n je razporejena PRIBLIŽNO kot $N(na, \sigma\sqrt{n})$. ... $n \geq 30$

→ CLI velja tudi, če niso enako porazdeljene, če X_i nimajo vse iste disperzije in/ali mat. upanja, če za X_i velja $E(X_i) = n$ in $D(X_i) = \sigma^2$.
→ če so X_i primerljive po velikosti, potem $\frac{S_n - E(S_n)}{\sigma(S_n)} \sim N(E(S_n), \sigma(S_n)) = N(0,1)$.

→ CLI NE VELJA, če spremenljivke NISO NEODVISNE !!

VZORČENJE

POPULACIJA ... velika količina podatkov .. N
VZOREC ... majhna podmnožica podatkov ... n
SLUČAJNI VZOREC ... vsak vzorec iste moči ima isto verjetnost, da bo izbran
SLUČAJNI VZOREC S PONAVLJANJEM ... dovolimo, da se podatki v vzorcu ponavljajo
SLUČAJNI VZOREC BREZ PONAVLJANJA ...

CILJ STATISTIČNIH METOD: na podlagi analize vzorca želimo sklepati na celotno populacijo

PONAVLJANJE, DA ALI NE

N .. velikost populacije, n... velikost vzorca.

TRDITEV: Če je N velik v primerjavi z n, potem je število vzorcev brez ponavljanja približno enako kot število vzorcev s ponavljanjem.

Če je populacija dovolj velika, je slučajno izbrani vzorec s ponavljanjem brez ponavljanja.
Dokaz!

Število vzorcev brez ponavljanja: $\binom{N}{n}$
Število vzorcev s ponavljanjem: $\binom{N+n-1}{n}$

PONAVLJANJE, NAČELOMA DA, V PRAKSI NE

DEF.: STATISTIKA je vsaka vrednost, ki jo lahko izračunamo iz podatkov vzorca; je slučajna ~~vred~~ spremenljivka ... izberemo slučajni vzorec, na njem izračunamo statistiko.

KVANTILI, MEDIANA, KVARTILI, PERCENTILI

Naj X določa slučajno spremenljivko - vrednost parametra na populaciji.
Izberimo $q \in (0,1)$. Vrednosti a pravimo q-kvantil za X, če je $P(X \leq a) \geq q$ in $P(a \leq X) \geq 1-q$.

MEDIANA je 0.5 - kvantil. Mediana je precej bolj robustna od E.

0.25, 0.5 in 0.75 - kvantilom pravimo KVARTILI.

0.01, ..., 0.99 - kvantili so PERCENTILI.

VZORČNO POVPREČJE

Naj bo Y spremenljivka na populaciji. Izberemo vzorec velikosti n, pridobimo vrednosti $Y_1, Y_2, \dots, Y_n$. VZORČNO POVPREČJE je $\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \dots + Y_n)$.

IZREK: Matematično upanje vzorčnega povprečja je enako povprečni vrednosti (matematičnemu upanju) na celotni populaciji: $E(\bar{Y}) = E(Y)$. Dokaz!

MANJKALA!!

IZREK: Naj bo Y spremenljivka na populaciji velikosti N, $\mu = E(Y)$, $\sigma^2 = D(Y)$. Vzorčimo brez ponavljanja, vzorci velikosti n. Velja:

(a) $E(\bar{Y}) = E(Y) = \mu$.

(b) $D(\bar{Y}) = D(Y) \cdot \frac{N-n}{n(N-1)} \xrightarrow{n \to \infty} \frac{1}{n} \cdot D(Y) = \frac{1}{n} \cdot \sigma^2$ in $\sigma(\bar{Y}) \approx \frac{1}{\sqrt{n}}\sigma$

(c) Če je Y na populaciji normalno porazdeljena z $N(\mu, \sigma)$, potem je $\bar{Y}$ na vzorcih tudi normalno porazdeljena z $N(\mu, \frac{\sigma}{\sqrt{n}})$.

(d) Če tudi Y na populaciji ni normalno porazdeljena, je $\bar{Y}$ na populaciji približno normalno porazdeljena z $N(\mu, \frac{\sigma}{\sqrt{n}})$.

DISPERZIJA VZORCA

Vzorec velikosti n z vrednostmi $X_1, \dots, X_n$ in vzorčnim povprečjem $\bar{X}$. VZORČNO DISPERZIJO s^2 in POPRAVLJENO VZORČNO DISPERZIJO $\hat{s}^2$ definiramo kot:

$$s^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n}, \qquad \hat{s}^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}.$$

IZREK: $E(s^2) = \frac{N}{N-1} \cdot \frac{n-1}{n} \cdot \sigma^2$ in $E(\hat{s}^2) = \frac{N}{N-1} \cdot \sigma^2$.

PORAZDELITEV VZORČNEGA POVPREČJA

Naj bo μ matematično upanje na populaciji. Potem je: $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $T = \frac{\bar{X} - \mu}{\hat{s}/\sqrt{n}}$.

Če σ ne poznamo, jo nadomestimo s $\hat{s}$. Če je vzorec dovolj velik ($n \geq 30$), je tudi $T \approx N(0,1)$.

Če je $n \geq 30$, potem je potrebno uporabiti Studentovo porazdelitev.

OCENJEVANJE PARAMETROV

Naj bo λ parameter, odvisen od slučajne spremenljivke X. I je INTERVAL ZAUPANJA za vrednost parametra λ pri stopnji zaupanja γ, če velja naslednje:

- verjetnost, da parameter λ pripada intervalu I, je (ne glede na porazdelitev X) vsaj γ;
- interval I je najmanjši možen.

OCENJEVANJE μ

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

Torej je $\bar{X}$ na vzorcih porazdeljen približno kot $N(\mu, \frac{\sigma}{\sqrt{n}})$.
Recept:
1. Izberi stopnjo zaupanja γ (tipično 0.9, 0.95, 0.99 ali 0.999).
2. Dobi c, za katerega velja $\Phi(c) = \gamma/2$.
3. Izračunaj $\bar{X}$ in $k = \frac{c \cdot \sigma}{\sqrt{n}}$.
4. Interval zaupanja je $i = [\bar{X} - k, \bar{X} + k]$.

→ Kaj, če σ ne poznamo?
→ Kaj, če je vzorec majhen ($n \leq 30$)?

PORAZDELITEV χ^2

Naj bodo slučajne spremenljivke $X_1, X_2, \ldots, X_n$ porazdeljene standardno normalno ($X_i \sim N(0,1)$) in neodvisne. $\chi^2(n) := X_1^2 + X_2^2 + \ldots + X_n^2$.
Slučajna spremenljivka $\chi^2(n)$ Hi-KVADRAT z n-PROSTOSTNIMI STOPNJAMI je vsota n kvadratov neodvisnih standardno normalno porazdeljenih slučajnih spremenljivk.
Če je št. prostostnih stopenj nedvoumno, namesto $\chi^2(n)$ pišemo samo χ^2.

$$P(\chi^2 \leq a) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \int_0^a x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx \text{, če je } a \geq 0.$$

Za izračun verjetnosti slučajne spremenljivke χ^2 uporabimo tabele. Fiksirajmo število prostostnih stopenj. χ^2_p je vrednost, pri kateri je $P(\chi^2 \leq \chi^2_p) = p$.
→ p-ti kvantil za χ^2

df ... degree of freedom - prostostna stopnja.

INTERVAL ZAUPANJA ZA STANDARDNO DEVIACIJO

IZREK: $\frac{n \cdot s^2}{\sigma^2} = \frac{(n-1) \cdot \hat{s}^2}{\sigma^2} \sim \chi^2(n-1)$

Radi bi določili interval $[c_1, c_2]$, za katerega velja, da je $P(c_1 \leq \sigma \leq c_2) \geq \gamma$, kjer je γ izbrana stopnja zaupanja. Tipično 0.09, 0.95 ali 0.99.

Iščemo interval, za katerega je $P(\chi^2 \leq \chi^2_{p_1})$ za $p_1 = \frac{1+\gamma}{2}$, $P(\chi^2 \leq \chi^2_{p_2})$ za $p_2 = \frac{1-\gamma}{2}$. Potem je:

$$P(\chi^2_{p_2} \leq \chi^2 \leq \chi^2_{p_1}) = \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma.$$

Dokaz!

$$P\left(\frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_{\frac{1+\gamma}{2}}}} \leq \sigma \leq \frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_{\frac{1-\gamma}{2}}}}\right) = \gamma \text{ . izpeljava!}$$

Pri stopnji zaupanja γ je interval zaupanja za σ enak: $\left[\frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_{\frac{1+\gamma}{2}}(n-1)}}, \frac{\sqrt{n-1} \cdot \hat{s}}{\sqrt{\chi^2_{\frac{1-\gamma}{2}}(n-1)}}\right]$.
χ^2 lahko izračunamo, lahko ocenimo z lin. interpolacijo, ali s CAS.

STATISTIČNE HIPOTEZE

STATISTIČNA HIPOTEZA je hipoteza o porazdelitvi slučajne spremenljivke.
STATISTIČNI TEST je bodisi:
- PARAMETRIČEN - znan je tip porazdelitve, hipoteza govori o parametru.
- NEPARAMETRIČEN - porazdelitev je neznana, hipoteza govori o vrsti porazdelitve.

Parametrični testi so ENOSTRANSKI in DVOSTRANSKI.

H_0 NIČELNA HIPOTEZA o porazdelitvi slučajne spremenljivke. → točka je postavna
... Slučajna spr. je porazdeljena normalno.
... Slučajna spr. ima mat. upanje enako 2.

H_{alt} ALTERNATIVNA HIPOTEZA o porazdelitvi slučajne spremenljivke.
... Slučajna spr. Ni porazdeljena normalno.
... Slučajna spr. ima mat. upanje RAZLIČNO od 2.

S statističnim testom testiramo ničelno hipotezo H_0 PROTI alternativni hipotezi H_{alt}.

NAPAKE, KI JIH LAHKO NAREDIMO

NAPAKA 1. VRSTE: ničelna hipoteza je pravilna, s testom jo zavrnemo.
Verjetnost, da naredimo napako 1. vrste je merljiva in jo lahko poljubno zmanjšamo.

NAPAKA 2. VRSTE: ničelna hipoteza je napačna, s testom jo sprejmemo.
Verjetnosti, da naredimo napako 2. vrste, NI MOGOČE oceniti. Zato napak 2. vrste NE DELAMO. To pomeni, da NIČELNIH HIPOTEZ NIKOLI NE SPREJMEMO.

STOPNJA ZNAČILNOSTI testa je verjetnost, da zavrnemo pravilno hipotezo (naredimo napako 1. vrste). Stopnja značilnosti α je tipično 0.1, 0.05 ali 0.01.

PARAMETRIČNI TEST – DVOSTRANSKI

... H_0 verjetnost rojstva dečka je enaka $\frac{1}{2}$.
... H_{alt} verjetnost rojstva dečka je RAZLIČNA od $\frac{1}{2}$.

PARAMETRIČNI TEST – ENOSTRANSKI

... H_0 verjetnost rojstva dečka je večja ali enaka $\frac{1}{2}$.
... H_{alt} verjetnost rojstva dečka je MANJŠA od $\frac{1}{2}$.

SPREJEMANJE NIČELNIH HIPOTEZ

Lahko s finto. Nove hipoteze. $H_{alt} = H_0$. V primeru dvostranskega testa takšna finta ne deluje.

NEPARAMETRIČNI TEST χ^2

Naj bo X_{ref} znana slučajna spremenljivka. Poznamo njeno verjetnostno shemo.
$X_{ref} \sim \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & & p_k \end{pmatrix}$.
Naj bo X opazovana slučajna spremenljivka. Z neparametričnim testom χ^2 testiramo ničelno hipotezo. H_0 - slučajna spr. X je porazdeljena ENAKO kot slučajna spr. X_{ref}. proti alternativni hipotezi H_{alt} - slučajna spr. X Ni porazdeljena enako kot sl. spr. X_{ref}.

Izberemo X-vzorec velikosti n in sestavimo tabelico:

Dogodek	$X = x_1$	$X = x_2$	...	$X = x_k$
Izmerjena frekvenca	X_1	X_2	...	X_k
Pričakovana frekvenca	$n \cdot p_1$	$n \cdot p_2$	...	$n \cdot p_k$

Izračunamo statistiko:

$$\chi^2 = \frac{(X_1 - n \cdot p_1)^2}{n \cdot p_1} + \frac{(X_2 - np_2)^2}{n \cdot p_2} + \dots + \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{i=1}^{k} \frac{(X_i - n \cdot p_i)^2}{n \cdot p_i},$$

pri čemer je $X_1 + X_2 + \dots + X_k = n$.

$$\chi^2 = \left(\sum_{i=1}^{k} \frac{X_i^2}{n \cdot p_i}\right) - n$$

Kaj pomeni, da je $\chi^2 = 0$? Ni odstopanja med izmerjenimi in pričakovanimi frekvencami.

χ^2 je enostranski test.

IZREK: Pri veljavni ničelni hipotezi, dovolj velikem vzorcu ($n \geq 30$ - zaradi CLI) in če je $n \cdot p_i \geq 5$ za vsak i, je statistika χ^2 porazdeljena približno kot hi-kvadrat s $(k-1)$ prostostnimi stopnjami $\chi^2(k-1)$.
Izberemo stopnjo značilnosti α (tipično 0.01 ali 0.05).
Če je $\chi^2 > \chi^2_{1-\alpha}(k-1)$, potem hipotezo ZAVRNEMO.
Če je $\chi^2 \leq \chi^2_{1-\alpha}(k-1)$, potem hipoteze NE ZAVRNEMO.

$P(\chi^2 \geq \chi^2_{1-\alpha}(k-1)) = \alpha$
$P(\chi^2 \leq \chi^2_{1-\alpha}(k-1)) = 1-\alpha$

TEST χ^2 ZA TESTIRANJE NEODVISNOSTI

Izračunamo količino $\chi^2 = \sum_{\text{po vseh celicah}} \frac{(\text{izmerjena vrednost} - \text{pričakovana vrednost})^2}{\text{pričakovana vrednost}}$.

IZREK: Pri veljavni ničelni hipotezi, dovolj velikem vzorcu ($n \geq 30$) in če je vsaka pričakovana vrednost ≥ 5, je statistika χ^2 porazdeljena približno kot hi-kvadrat s $(r-1)(s-1)$ prostostnimi stopnjami.

Če so kakšne pričakovane vrednosti premajhne, lahko združujemo razrede ali pa spremenimo starostne meje.

Naj bosta X in Y slučajni spremenljivki. Test χ^2 uporabljamo za testiranje ničelne hipoteze H_0 - X in Y sta NEODVISNI. proti alternativni hipotezi H_{alt} - X in Y sta odvisni.

TEST Z ZNAKI

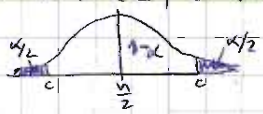
Naj bosta X in Y slučajni spremenljivki. TEST Z ZNAKI uporabljamo za testiranje ničelne hipoteze H_0 - $P(X > Y) = 0.5$ (lahko tudi $P(X > Y) \geq 0.5$) proti alternativni hipotezi H_{alt} - $P(X > Y) \neq 0.5$ (oz. $P(X > Y) < 0.5$).
Potrebujemo isto število meritev $X_1, X_2, \dots, X_n$ in $Y_1, Y_2, \dots, Y_n$.

$X_1 - Y_1$
$X_2 - Y_2$
$\vdots$
$X_n - Y_n$
(vse različne od 0)

K^+ ... # pozitivnih razlik
K^- ... # negativnih razlik
α ... stopnja značilnosti

$\Phi\left(\frac{c - \frac{n}{2}}{\frac{1}{2}\cdot\sqrt{n}}\right) = \frac{1}{2} - \frac{\alpha}{2}$

Če H_0 ($P(X > Y) = 0.5$) velja, da sta: $K^+, K^- \sim N\left(\frac{n}{2}, \frac{1}{2}\cdot\sqrt{n}\right)$.



HANN - WHITNEYEV TEST

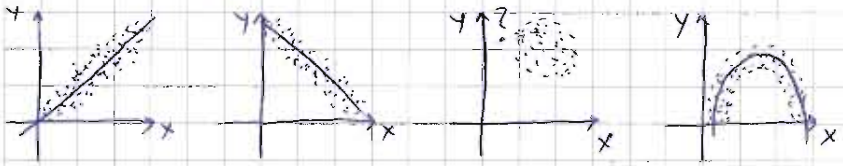
Naj bosta X in Y slučajni spremenljivki (neznani). HANN - WHITNEYEV TEST uporabljamo za testiranje ničelne hipoteze H_0 - X in Y sta ENAKO porazdeljeni proti alternativni hipotezi H_{alt} - X in Y nista enako porazdeljeni. Pri tem pa ne vemo, kako je katera od X oz. Y porazdeljena.

$X_1, X_2, \dots, X_m$
$Y_1, Y_2, \dots, Y_n$
} združimo in uredimo po velikosti $\rightarrow z_1 < z_2 < z_3 < z_4 < \dots < z_{m+n}$

$U = \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} [1, \text{če } X_i > Y_j \,;\, 0, \text{sicer}]$. $U \in \{0, 1, \dots, mn\}$

IZREK: Če $m+n \geq 20$, $m \geq 4$, $n \geq 4$, potem je pri veljavni H_0 $U \sim N\left(\frac{m\cdot n}{2}, \sqrt{\frac{m\cdot n\cdot(m+n+1)}{12}}\right)$.

RAZSEVNI GRAFIKON



METODA NAJMANJŠIH KVADRATOV

— graf X, Y. $Y_1^{\wedge} = a + bX_1$. ...

LINEARNA REGRESIJA

Regresijska krivulja je premica. V REGRESIJSKI PREMICI $y' = a + bx$ parametra a in b ocenimo po metodi najmanjših kvadratov. Želimo minimitirati izraz $F(a,b) = \sum_{i=1}^{N} (y_i - y_i')^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2$.
Odvajamo po obeh parametrih in rešujemo sistem $\frac{\partial F}{\partial a} = 0$, $\frac{\partial F}{\partial b} = 0$. Pridelamo sistem LINEARNIH enačb, ki je ENOLIČNO rešljiv; parametra a in b sta enolično določena.

PRVA REGRESIJSKA PREMICA: $y' = E(Y) + \frac{K(X,Y)}{D(X)} (x - E(X))$.

DRUGA REGRESIJSKA PREMICA: $x' = E(X) + \frac{K(X,Y)}{D(Y)} (y - E(Y))$.

TRDITEV: Regresijski premici se sekata v točki $(E(X), E(Y))$.

Tooda mi delamo z vzorci. Torej vzorčni regresijski premici:

$$y' = \bar{y} + \frac{\hat{k}(x,y)}{(\hat{s}(x))^2} (x - \bar{x}), \quad x' = \bar{x} + \frac{\hat{k}(x,y)}{(\hat{s}(y))^2} (y - \bar{y}).$$

Pri tem so $\hat{s}(x)$, $\hat{k}(x,y)$ in $r(x,y)$ (popravljena) vzorčne ustreznice $\sigma(x)$, $K(x,y)$ in $\rho(x,y)$.

GENERATORJI PSEVDONAKLJUČNIH ŠTEVIL

Generirati želimo zaporedje (celih) števil med 1 in m, ki so izbrana tako, da izgleda, kot da smo jih dobili slučajno. Generatorji so zelo hitri.
- Linearna kongruenčna metoda: $x_{n+1} = (a \cdot x_n + c) \bmod m$
- Kvadratna metoda: $x_{n+1} = (a x_n^2 + b x_n + c) \bmod m$
- Fibonaccijeva metoda: $x_{n+1} = (a x_n + b x_{n-1})$

DEJANSKO NAKLJUČNA ŠTEVILA

www.random.org, www.lavarand.org.
Kakšne so prednosti in slabosti SLUČAJNIH števil (za razliko od psevdoslučajnih)?

KAJ ŽELIMO OD NAŠEGA PSEVDOSLUČAJNEGA GENERATORJA

Kriptografsko ustrezen generator. Kaotični sistem. ZIP algoritem.
Kaj, če iz mentev dobimo več ničel kot enic? 00001100000011001101

STATISTIČNI TEST ZA PREVERJANJE SLUČAJNOSTI

χ^2 test. Test Kolmogorova. Testiranje parov. Poker test. Spektralni testi.