STATISTIKA IN ANALIZA PODATKOV

STATISTIKA je veda, ki proučuje množične pojave in se ukvarja z zbiranjem, predstavitvijo, analizo in interpretacijo podatkov.

ENOTA je posamezni proučevani element (redni študent na Univerzi v Lj v študijskem letu 1994/95)

POPULACIJA je množica vseh proučevanih elementov; pomembna je časovna in prostorska opredelitev populacije (vsi redni študentje na Univerzi v Lj v študijskem letu 1994/95)

VZOREC je podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih cele populacije (slučajni vzorec 300 rednih študentov na Univerzi v Lj v študijskem letu 1994/95)

SPREMENLJIVKA je lastnost enot; označujemo jih z X, Y (spol, uspeh iz matematike v zadnjem letniku srednje šole, izobrazba matere in višina mesečnih dohodkov staršev študenta)

Vrste spremenljivk:

- **glede na tip izražanja vrednosti:** - *opisne, atributivne*; vrednosti lahko opišemo z besedami (poklic, uspeh)
  - *številske, numerične*; vrednosti lahko izrazimo s števili (starost)
- **glede na tip merjenja:** - *nominalne*; vrednosti lahko le razlikujemo med seboj; dve vrednosti sta enaki ali različni (spol)
  - *ordinalne*; vrednosti lahko uredimo od najmanjše do največje (uspeh)
  - *intervalne*; lahko primerjamo razlike med vrednostimi dvojic enot (temperatura)
  - *razmernostne*; lahko primerjamo razmerja med vrednostim dvojic enot (starost)

Statistična analiza je:

- *opisna statistika*; statistična analiza zbranih podatkov brez teženj, da bi iz teh podatkov posploševali čez njihov obseg
- *inferenčna statistika*; statistično sklepanje iz vzorca (dela populacije) na populacijo: - ocenjevanje značilnosti populacije
  - preverjanje domnev
- *univariantna*; analiza ene spremenljivke
- *bivariantna*; analiza dveh spremenljivk
- *multivariantna*; analiza več spremenljivk

Koraki statistične analize:

1. **Določitev vsebine in namena** statističnega proučevanja; opredelitev predmeta opazovanja (enote in populacije) in vsebine opazovanja (spremenljivk)
2. **Statistično opazovanje**; vrste opazovanj: - opazovanje cele populacije (popisi, tekoče registracije)
   - opazovanje vzorca (ankete)
3. **Osnovna obdelava**: - urejanje

- razvrščanje podatkov
- izračun osnovnih statističnih karakteristik

4. Analitična obdelava

Frekvenčna porazdelitev spremenljivke je tabela, ki jo določajo vrednosti ali skupine vrednosti in njihove frekvence.

Razredi so skupine vrednosti številskih spremenljivk.

$x_{i,max}$ = zgornja meja i-tega razreda
$x_{i,min}$ = spodnja meja i-tega razreda

Širina i-tega razreda je

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

Sredina i-tega razreda je

$$\chi_i = \frac{xi,\min + xi,\max}{2}$$

Kumulativa (F_i) je frekvenca do spodnje meje določenega razreda.

f_i% je relativna frekvenca = strukturni odstotki v i-tem razredu

Grafično prikazovanje frekvenčnih porazdelitev:

- *histogram*; drug poleg drugega rišemo stolpce pravokotnika, katerih višina je sorazmerna frekvenci v razredu. Širina pravokotnikov je enaka, ker so razredi enako široki.
- *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 kumulativne frekvenčne porazdelitve s polgonom, kjer v koordinatnem sistem nanašamo točke ($x_{i,min}$,F_i).

Oblika frekvenčnih porazdelitev:

- *normalna*, ki je unimodalna (ima en vrh), simetrična in zvonaste oblike
- *bimodalna*, če ima dva vrha
- *polimodalna* z več vrhovi
- *asimetrična v levo*, če se rep vlečena levo
- *asimetrična v desno*, če se rep vleče v desno
- bolj *koničasta* in *sploščena* od normalne porazdelitve
- *J* in *U* oblike

Za grafično predstavitev s krogi je potrebno izračunati še **stopinje** f_i^0:

Ranžirna vrsta: enote z ustreznimi vrednostnimi spremenljivke uredimo od tiste z najmanjšo vrednostjo do tiste z največjo vrednostjo.

$$f_i^0 = \frac{f_i}{N} 360$$

Rang R: vsaki enoti v ranžirni vrsti priredimo zaporedno mesto.

Kvantilni rang P pove, na katerem delu celotnega ranžirnega razmika leži določena enota oziroma koliki del celotnega razmika ima manjše vrednosti od dane vrednosti.

$$P = \frac{R - 0.5}{N}$$

Kvantil je vrednost spremenljivke, ki pripada določenemu kvantilnemu rangu.
Običajni kvantili:

- **mediana** $M_e(P=0.5)$
- **kvartili** $Q_1(P=0.25)$, $Q_2(P=0.50)$, $Q_3(P=0.75)$
- **decili** $D_1(P=0.1)$, $D_2(P=0.2)$,… $D_9(P=0.9)$
- **centili** $C_1(P=0.01)$, $C_2(P=0.02)$,… $C_{99}(P=0.99)$

Linearna interpolacija:

$$\frac{R - ¿R0}{R1 - R0} = \frac{x - x0}{x1 - x0} ¿$$

Srednje vrednosti:

1. **Mediana** (M_e) je tista vrednost spremenljivke, od katere je ravno toliko manjših vrednosti od nje, kolikor jih je večjih od nje. Zato je mediana vrednost spremenljivke, ki pripada kvantilnemu rangu 0.5.
   - Če je liho število enot N=2m+1, je mediana (m+1) vrednost v ranžirni vrstici.
   - Če pa je sodo število enot N=2m, je mediana

     $$M_e = \frac{xm + xm+1}{2}$$

     Iz frekvenčne porazdelitve pa izračunamo mediano tako, da izračunamo vrednost spremenljivke, ki pripada kvantilnemu rangu P=0.5. Ranžirno vrsto s pripadajočimi rangi predstavljajo spodnje meje razredov in ustrezne kumulative.

2. **Modus** (M_0) je vrednost spremenljivke, ki se v populaciji najpogosteje pojavlja. Modus lahko razumemo kot vrednost spremenljivke, okoli katere se vrednosti najbolj gostijo.
3. **Aritmetična sredina** ali povprečje je vsota vseh vrednosti deljena s številom enot v populaciji.
4. **Geometrijska sredina** je enaka N-temu korenu iz produkta N vrednosti številske spremenljivke (pogoj $x_i>0$).
5. **Harmonična sredina** je enaka recipročni vrednosti aritmetične sredine.

MERE RAZPRŠENOSTI:

1. **Variacijski razmik:** $R = x_{max} - x_{min}$
2. **Kvartilni odklon:**

$$Q = \frac{Q_3 - Q_1}{2}$$

3. **Povrečni absolutni odklon:**

Negrupirani podatki:

$$AD_{\mu} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$$

$$AD_{Me} = \frac{1}{N}\sum_{i=1}^{N} |x_i - M_e|$$

Grupirani podatki:

$$AD_{\mu} = \frac{1}{N}\sum_{i=1}^{k} f_i |x_i - \mu|$$

$$AD_{Me} = \frac{1}{N}\sum_{i=1}^{k} f_i |x_i - M_e|$$

4. **Varianca:**

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

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

5. **Standardni odklon**

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

$$\sigma = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \mu)^2}$$

Sheppardov popravek:

$$\sigma^2_{pop} = \sigma^2 - \frac{d^2}{12}$$

→ <u>*Relativne mere razpršenosti*</u> so absolutne mere deljene z ustrezno srednjo vrednostjo:

- relativni variacijski razmik:

$$\frac{(x_{max} - x_{min}) \cdot 2}{x_{max} + x_{min}}$$

- relativni kavrtilni odklon:

$$\frac{Q_3 - Q_1}{2 \cdot M_e}$$

- relativni povprečni absolutni odklon:

$$\frac{AD_{Me}}{M_e}$$

- relativni standardni odklon – koeficient variacije:

$$KV = \frac{\sigma}{\mu}$$

→ *Normalna porazdelitev*: Denimo, da se spremenljivka X porazdeljuje normalno. Zanjo je izračunana aritmetična sredina μ in standardni odklon σ. Tedaj velja, da v razmiku:
◆[μ-σ; μ+σ] leži 68.3% enot
◆[μ-2σ; μ+2σ] leži 95.4% enot
◆[μ-3σ; μ+3σ] leži 99.7% enot

→ *Meri asimetrije*:

$$KA_{M0}=\frac{\mu-M_0}{\sigma}\begin{cases}<0 & \text{as. v desno}\\ >0 & \text{sim.}\end{cases}$$

→ *Mere sploščenosti*:

$$KS=1.9\frac{Q_3-Q_1}{D_9-D_1}\begin{cases}<1 & \text{konicasta}\\ >1 & \text{normalna}\end{cases}$$

→ *Meri asimetrije in sploščenosti s centralnimi momenti:*
L-ti centralni moment je:

◆Koeficienti asimetrije:

$$m_l=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^l$$

$$g_1=\frac{m_3}{\sqrt{m_2^3}}\begin{cases}<0 & \text{as. v desno}\\ =0 & m_1=0\\ & m_2=\sigma^2\\ >0 & \text{sim.}\end{cases}$$

◆Koeficienti sploščenosti:

$$g_2=\frac{m_4}{m_2^2}-3\begin{cases}<0 & \text{konicasta}\\ >0 & \text{normalna}\end{cases}$$

Permutacija je vsaka preureditev n elementov.
Variacija reda r iz n elementov je, če iz množice n elementov vzamemo r elementov in jih na nek način razporedimo.

Osnovni izrek kombinatorike: Imejmo izbor, sestavljen iz k delnih izborov. Prvič izbiramo med n_1 možnostmi, drugič med n_2 možnostmi,… in k-tič med n_k možnostmi. Pri tako sestavljenem izboru je vseh možnosti:

$$n=n_1*n_2*\ldots*n_k$$

Število variacij, permutacij in kombinacij:

1. Število variacij reda r iz n elementov s ponavljanjem:

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

2. Število variacij reda r iz n elementov brez ponavljanja:

$$V_n^r = n\cdot(n-1)\cdot¿\cdot(n-r+1)$$

3. Število premutacij:
4. Število kombinacij:

$$P_n = V_n^n = n\cdot(n-1)\cdot(n-2)\cdot¿\cdot 2\cdot 1 = n!$$

$$C_n^r = \frac{n!}{r!(n-r)!} = ¿\frac{(n¿)¿¿¿}{n¿(n-r)!¿\cdot(n¿-r+1)}$$

$$C_n^r = \frac{V_n^r}{P_r} = \frac{n\cdot(n-1)\cdot(n-2)\cdot¿\cdot(n-r+1)}{r\cdot(r-1)\cdot(r-2)\cdot¿\cdot 1}$$

VERJETNOSTNI RAČUN

→ Poskus je realizacija neke množice skupaj nastopajočih dejstev.

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

Dogodek je lahko:

- gotov dogodek G; ob vsaki ponovitvi poskusa se zgodi
- nemogoč dogodek N; nikoli se ne zgodi
- slučajen dogodek; včasih se zgodi, včasih ne

Računanje z dogodki:

1. Dogodek A je **način** dogodka B ($A \subset B$), če se vsakič, ko se zgodi dogodek A, zagotovo zgodi tudi dogodek B.
2. Če je dogodek A način dogodka B in sočasno dogodek B način dogodka A, sta dogodka **enaka**: $A \subset B \wedge B \subset A \Leftrightarrow A = B$
3. **Vsota dogodkov** A in B ($A \cup B$) je, če se zgodi vsaj eden od dogodkov A in B.
   Velja: $A \cup B = B \cup A$; $A \cup N = A$;
   $A \cup G = G$, $A \cup A = A$
4. **Produkt dogodkov** A in B ($A \cap B$) se imenuje dogodek, če se zgodita A in B hkrati.
   Velja: $A \cap B = B \cap A$; $A \cap N = N$;
   $A \cap G = A$; $A \cap A = A$
5. Dogodku A **nasproten dogodek** $\bar{A}$ imenujemo negacija dogodka.
   Velja: $A \cap \bar{A} = N$; $A \cup \bar{A} = G$;
   $\bar{N} = G$; $\bar{\bar{A}} = A$
6. Dogodka A in B sta **nezdružljiva**, če se ne moreta zgoditi hkrati, njun produkt je torej nemogoč dogodek, $A \cap B = N$
   Velja: $A \cap \bar{A} = N \wedge A \cup \bar{A} = G$
7. Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A **sestavljen** dogodek. Dogodek, ki ni sestavljen, imenujemo **elementaren** dogodek.
8. Množico dogodkov $S = \{A_1, A_2, \ldots, A_n\}$ imenujemo **popoln sistem dogodkov**, če se v vsaki ponovitvi poskusa zgodi natanko eden od dogodkov iz množice S.

→ Verjetnost

Statistična definicija verjetnosti: 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.

Osnovne lastnosti verjetnosti:

1. Ker je relativna frekvenca vedno negativna, je verjetnost $P(A) \geq =$
2. $P(G) = 1$
3. Naj bosta dogodka A in B nezdružljiva. Pokaže se lahko, da velja:
   $P(A \cup B)=P(A)+P(B)$
4. Za združljiva dogodka A in B ($A \cap B \neq N$) velja:
   $P(A \cup B)=P(A) + P(B) – P(A \cap B)$
5. $P(\bar{}A)=1 – P(A)$

→ Pogojna verjetnost:

$P'(A)=P(A/B)$

$$P(A/B)=\frac{P(A \cap B)}{P(B)}$$

Za neodvisna dogodka velja: $P(A \cap B) = P(A)*P(B)$