VAJE 2: Opisna statistika

Na računalniških vajah se za urejanje in prikazovanje statističnih podatkov uporabi statistični programski paket SPSS in podatkovna datoteka *podatki2.sav*.

NALOGE:

1. Analiza vzorčnih parametrov statistične spremenljivke *TežaO*:

   (a) Vrednosti spremenljivke *TežaO* uredi v ranžirno vrsto po naraščajoči teži (uporabi postopek *Data - Sort Cases - Ascending*) in za vsako vrednost določi pripadajoči rang (uporabi postopek *Transform - Rank Cases* in označi *Smallest Value*).

      i. Katere range imajo vrednosti 45, 56, 67 in 85?

      ii. Katere vrednosti spremenljivke *TežaO* ustrezajo rangom 19, 34, 117 in 181?

   (b) Iz ranžirne vrste določi največjo x_{max} in najmanjšo x_{min} vrednost spremenljivke *TežaO* ter izračunaj vzorčni variacijski razmik vr.

   (c) S pomočjo ranžirne vrste izračunaj vse tri vzorčne kvartile q_1, $q_2 = m$ (vzorčna mediana) in q_3 ter določi še vzorčni kvartilni razmik kr in odklon ko.

   (d) Izdelaj frekvenčno porazdelitev statistične spremenljivke *TežaO* in določi vzorčni modus.

   (e) Z uporabo programa SPSS izračunaj vzorčno povprečje $\overline{x}$, vzorčno disperzijo s^2 in vzorčni standardni odklon s. Uporabi postopek *Analyze - Descriptive Statistics - Frequencies - Statistics* in označi *Mean (povprečje), Variance (disperzija)* ter *Std. Deviation (standardni odklon).*

   (f) S programom SPSS še enkrat izračunaj vrednosti iz primerov (b), (c) in (d). Uporabi postopek *Analyze - Descriptive Statistics - Frequencies - Statistics* in označi *Median (mediana), Mode (modus), Range (variacijski razmik), Minimum (najmanjša vrednost), Maximum (največja vrednost), Quartiles (kvartili).*

2. Izračunaj vzorčne mere srednje vrednosti (modus, mediana, povprečje) in vzorčne mere variabilnosti (variacijski razmik, kvartilni razmik, kvartilni odklon, disperzijo in standardni odklon) za statistično spremenljivko *Starost*. Določi tudi asimetričnost in sploščenost spremenljivke *Starost*.

3. Vrednosti spremenljivke *KoličinaTD* uredi v ranžirno vrsto po naraščajoči količini in za vsako vrednost določi pripadajoči rang.

   (a) S pomočjo ranžirne vrste izračunaj vseh devet vzorčnih decilov d_1, d_2, ... ,d_9 ter določi še vzorčni decilni razmik *dr* in odklon *do*.

   (b) Rezultate iz primera (a) preveri s pomočjo programa SPSS (uporabi postopek *Analyze - Descriptive Statistics - Frequencies - Statistics* in označi *Percentile(s)* in dodaj vrednosti $10, 20, \ldots, 90$).

4. Izračunaj vzorčne mere srednje vrednosti in vzorčne mere variabilnosti za statistično spremenljivko *KoličinaTD* glede na spremenljivko *Šport*.

   (a) Primerjaj vzorčno povprečje popite tekočine na dan pri osebah, ki se oz. se ne ukvarjajo s športom. Kaj ugotoviš?

   (b) V obeh obravnavanih primerih (oseba se oz. se ne ukvarja s športom) glede na vzorčno mediano določi interval v katerem leži osrednjih 50% vzorčnih vrednosti.

   (c) Oceni, koliko odstotkov oseb, ki se ne ukvarja s športom, popije premalo količino tekočine? Rezultat preveri s kontingenčno tabelo (uporabi postopek *Analyze - Descriptive Statistics - Crosstabs - Row (KoličinaPT) - Column (Šport)* in v *Cells* označi *Percentages Column*).

   Navodilo: najprej uporabi postopek *Data - Split File* in označi *Organize output by groups* ter dodaj spremenljivko *Šport*, nato naredi ustrezno analizo.

5. Izračunaj vzorčno povprečje in vzorčni standardni odklon statistične spremenljivke *KoličinaTD* glede na vrednosti spremenljivke *Rasa*. Interpretiraj rezultat!

6. S programom SPSS iz danih podatkov glede na navedene pogoje izberi manjši vzorec za statistični spremenljivki *TežaO* in *Starost* ter oba vzorca ustrezno analiziraj.

   (a) Za statistično spremenljivko *TežaO* izberi naključni vzorec velikosti 50 enot (uporabi postopek *Data - Select Cases* in izberi *Random sample of cases* ter v *Sample...* izberi *Exactly 50 cases from the first 189*). Na svojem vzorcu izračunaj variacijski razmik, kvartile in kvartilni razmik. Dobljene rezultate primerjaj z rezultati na celotnem vzorcu.

(b) Iz celotnega vzorca izberi le tiste osebe, ki so starejše od 23 let (uporabi postopek *Data - Select Cases* in izberi *If condition is satisfied* ter v *If...* za spremenljivko *Starost* določi *Starost* >= 23). Na manjšem vzorcu izračunaj vzorčne mere srednje vrednosti (modus, mediana, povprečje) in vzorčne mere variabilnosti (variacijski razmik, kvartilni razmik, kvartilni odklon, disperzijo in standardni odklon). Prav tak določi še asimetričnost in sploščenost manjšega vzorca. Kaj ugotoviš, če dobljene rezultate primerjaš z rezultati na celotnem vzorcu?

Teoretično ozadje

Srednje vrednosti

Pod pojmom *srednja vrednost* razumemo vrednost, ki pokaže osrednjo tendenco (težnjo) vrednosti statistične spremenljivke na neki populaciji. To je vrednost okoli katere se na populaciji zgostijo vrednosti statistične spremenljivke. Poznamo več meril za srednjo vrednost. Najpomembnejše med njimi so *aritmetična sredina* ali povprečje, *mediana* in *modus*. Srednje vrednosti populacije velikokrat tudi ne poznamo in jih praviloma ocenjujemo iz srednjih vrednosti vzorca.

Posebej velja poudariti, da je srednja vrednost (povprečje, mediana, modus) parameter, če se uporablja za opis populacije, ker je ta vrednost na populaciji fiksna. V primeru, ko govorimo o srednji vrednosti na vzorcu neke populacije, pa je srednja vrednost statistika, torej statistična spremenljivka, ki se od vzorca do vzorca spreminja. Zato v tem primeru ustrezne srednje vrednosti vzorca imenujemo tudi *vzorčna aritmetična sredina* ali *vzorčno povprečje*, *vzorčna mediana* in *vzorčni modus*.

V nadaljevanju naj bo X številska statistična spremenljivka na populaciji G. Naj bo H vzorec te populacije, ki ga sestavlja n enot. Z

$$x_1, x_2, \ldots, x_n$$

po vrsti označimo vrednosti statistične spremenljivke X, ki jih je le-ta zavzela na vzorcu H. V vzorcu so torej vrednosti zapisane po vrstnem redu izbiranja oz. merjenja. Realizacijo vzorca lahko prikažemo tudi s t.i. *ranžirno vrsto*, kjer rezultate razvrstimo po velikosti v naraščajočem zaporedju

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \quad \text{t.p.}$$

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}\,.$$

Vsaki enoti v ranžirni vrsti predpišemo zaporedno številko (od 1 do n), ki jo imenujemo rang. Kadar se v ranžirni vrsti pojavi več enot z isto vrednostjo, seštejemo

range, ki bi jih enote dobile, in vsoto delimo s številom teh enot. Povprečen, na ta način dobljeni rang, potem damo vsem enotam, ki imajo to isto vrednost.

Vzorčni modus

Najpreprostejša in zato tudi najmanj uporabna mera srednje vrednosti je modus ali najpogostejša vrednost, t.j. vrednost, ki se pojavi največkrat. Posebej velja omeniti, da modus ni nujno enolično določen, lahko obstaja več modusov. V primeru, da obstaja več modusov, je to znak, da statistična spremenljivka na vzorcu ali populaciji ni homogena. Vzorčni modus dobro predstavlja vrednosti vzorca, saj je to vrednost, okoli katere so vrednosti najbolj goste.

Vzorčna mediana

Vzorčna mediana, ki jo označimo z m, je srednja vrednost po položaju, je v sredini ranžirne vrste vrednosti statistične spremenljivke. Za mediano je torej značilno, da je polovica vrednosti statistične spremenljivke večjih ali enakih, polovica pa manjših ali enakih od nje. Iz ranžirne vrste izračunamo vzorčno mediano kot

$$m = \begin{cases} \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & ; \quad n = 2k \\ x_{(k+1)} & ; \quad n = 2k+1 \end{cases} .$$

Opomniti velja, da v primeru sode velikosti vzorca vzorčna mediana v splošnem ne leži v vzorcu. Dobra lastnost mediane je, da na njeno vrednost ne vplivajo ekstremne vrednosti, njena slabost pa, da ne izčrpamo vse informacije, ki nam jo dajejo podatki.

Vzorčno povprečje

Vzorčno povprečje, ki ga označimo z $\overline{x}$, statistične spremenljivke X na vzorcu je defnirano s predpisom

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} .$$

Praviloma, je vzorčno povprečje najboljša mera za srednjo vrednost, ker pri izračunu povprečja upoštevamo vse vrednosti in njihove velikosti. Na povprečje imajo velik vpliv tudi ekstremne vrednosti. Ravno zaradi te lastnosti je povprečje mogoče manj primerna srednja vrednost za tiste statistične spremenljivke, ki so zelo nehomogene.

Mere variabilnosti (razpršenosti)

Pri analizi statističnih spremenljivk nas poleg srednje vrednosti spremenljivke zanima tudi koliko so posamezne vrednosti razpršene oz. varirajo okoli te vrednosti. Poznamo več vrst mer za variabilnost, delimo jih na razmike in odklone.

Variacijski razmik

Variacijski razmik ali *variacijski razpon* je mera variabilnosti, ki jo izračunamo kot razliko med največjo in najmanjšo vrednostjo, ki jo zavzame statistična spremenljivka:

$$vr = x_{max} - x_{min}\,,$$

to je razlika med najveljo $x_{(n)}$ in najmanjšo $x_{(1)}$ vrednostjo v ranžirni vrsti. Variacijski razmik je groba in zelo nestabilna mera, ki jo določata samo dve skrajni vrednosti statistične spremenljivke, zato ni primerna za nadaljne analitične obravnave.

Kvantilni razmiki

Naj bo naravno število r določeno s predpisom

$$r = \begin{cases} np & ; \text{ če je } np \text{ naravno število} \\ [np]+1 & ; \text{ če } np \text{ ni naravno število} \end{cases}.$$

Potem vrednost

$$q_p = x_{(r)}$$

imenujemo *p-ti vzorčni kvantil*. Določeni kvantili nosijo tudi posebna imena. Tako kvantilom

$$q_1 = q_{\frac{1}{4}}, q_2 = q_{\frac{2}{4}}, q_3 = q_{\frac{3}{4}}$$

pravimo *kvartili*. Kvartili, kot že ime pove, ranžirno vrsto razdelijo na štiri dele. V literaturi srečamo tudi *decile*, ki ranžirno vrsto razdelijo na 10 delov in *centile*, ki ranžirno vrsto razdelijo na 100 delov.

Kvartilni razmik ali tudi *interkvartilni razmik* je mera variabilnosti, ki jo izračunamo kot razliko med tretjim in prvim kvartilom

$$kr = q_3 - q_1\,.$$

Kvartilni razmik teoretično obsega 50% vrednosti (rezultatov) populacije ali vzorca, 25% manjših in 25% večjih vrednosti je zunaj njega. Kvartilni razmik je zanesljivejša

mera variabilnosti od variacijskega razmika, ker nanj ne vplivajo skrajne vrednosti. V literaturi srečamo še decilni in centilni razmik, ki sta definirana kot

$$dr = d_9 - d_1 \quad \text{in} \quad cr = c_{99} - c_1 \,.$$

Kvantilni odkloni

Glede na mediano m poznamo tako imenovani:

- *kvartilni odklon*, ki je definiran kot

$$ko = \frac{q_3 - q_1}{2} \,;$$

- *decilni odklon*, določen s predpisom

$$do = \frac{d_9 - d_1}{2} \,;$$

- *centilni odklon*, določen kot

$$co = \frac{c_{99} - c_1}{2} \,.$$

Varianca (disperzija) in standardni odklon

Varianca ali *disperzija* je najpomembnejša mera variabilnosti. Disperzija vzorca je definirana kot povprečje kvadratov odklonov od vzorčnega povprečja t.j.

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \,.$$

Korenu vzorčne deisperzije

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

rečemo *standardni odklon vzorca*. Na populaciji praviloma za standardni odklon uporabimo grško črko σ in disperzijo označujemo s σ^2.

V zvezi z disperzijo vzorca je za nadaljno analitično obravnavo (npr. za oceno populacijske disperzije iz vzorčne disperzije) najbolj pomembna naslednja statistika

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \,,$$

imenovana vzorčna disperzija. Pod vzorčno disperzijo bomo imeli v mislih statistiko in pod vzorčnim standardnim odklonom bomo razumeli

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} \,.$$

Omenimo še, da je v praksi potrebno pogosto primerjati variranje različnih statističnih spremenljivk, ki so med seboj v kakšni povezavi. V ta namen uporabjamo *variacijski koeficient*, ki je defniran z deležem

$$kv = \frac{s}{\overline{x}} \,.$$

Asimetričnost in sploščenost

Asimetričnost (ang. *Skewness*) definiramo kot

$$A = \frac{m_3}{m_2^{\frac{3}{2}}} \,,$$

sploščenost (ang. *Kurtosis*) pa s formulo

$$K = \frac{m_4}{m_2^2} \,,$$

kjer je m_k k-ti centralni moment. V teoriji verjenosti je centralni moment definiran kot $m_k = E((X - E(X))^k)$, kjer je $E(X)$ matematično upanje, ki mu pogosto pravimo kar povprečje. S koeficientom asimetrije merimo asimetrijo vzorca. Kritične vrednosti:

- $A > 0$ - asimetrija v desno,
- $A = 0$ - porazdelitev je simetrična,
- $A < 0$ - asimetrija v levo.

S koeficientom sploščenosti merimo stopnjo sploščenosti vzorca. Kritične vrednosti:

- $K > 0$ - koničasta porazdelitev,
- $K = 0$ - spremenljivka se porazdeljuje normalno,
- $K < 0$ - sploščena porazdelitev.

Literatura

[1] D. Benkovič, Vaje iz biostatistike, Medicinska fakulteta Univerze v Mariboru.

[2] R. Jamnik, Matematična statistika, Državna založba Slovenije, Ljubljana 1980.

[3] J. Sagadin, Statistične metode za pedagoge, Obzorja, Maribor 2003.