VAJE 4: Ocenjevanje parametrov

Na računalniških vajah se za urejanje in prikazovanje statističnih podatkov uporabi statistični programski paket SPSS in podatkovna datoteka *podatki4.sav*.

NALOGE:

1. Ocena parametrov statistične spremenljivke *KolicinaTD*.

   (a) Izračunaj vzorčno povprečje $\bar{X}$ in vzorčni standardni odklon S ter standardno napako SE ocene vzorčnega povprečja. Uporabi postopek *Analyze - Descriptive Statistics - Descriptives - Options* in označi *Mean*, *Std. Deviation*, *S.E. Mean*.

   (b) S pomočjo točke (a) določi interval zaupanja na stopnji zaupanja 0.95 in 0.99 za populacijsko povprečje popite tekočine v *ml* na dan. Kateri od intervalov je širši?

   (c) S programom SPSS preveri vrednosti iz (b). Uporabi postopek *Analyze - Descriptive Statistics - Explore* v *Dependent List* vstavi *KolicinaTD* in v *Statistics* vstavi vrednosti 95 oz. 99 za *Confidence Interval for Mean.*

2. S programom SPSS iz danih podatkov naključno izberi vzorec velikosti 20 enot (uporabi postopek *Data - Select Cases* in izberi *Random sample of cases* ter v *Sample...* izberi *Exactly* 20 *cases from the first* 189). Za statistično spremenljivko *KolicinaTD* na svojem vzorcu

   (a) ponovi 1. nalogo;

   (b) določi interval zaupanja za populacijsko disperzijo σ^2 na stopnji zaupanja 0.95.

3. Na stopnji zaupanja 0.95 določi interval zaupanja za delež oseb, ki se ukvarjajo s športom.

4. Predpostavimo, da je količina popite tekočine v *ml* na dan pri osebah, ki se ukvarjajo s športom na populaciji porazdeljena normalno $N(\mu, \sigma)$ in da je tudi količina popite tekočine v *ml* na dan pri osebah, ki se ne ukvarjajo s športom porazdeljena normalno $N(\nu, \sigma)$. Na stopnji zaupanja 0.95 določi interval zaupanja za razliko povprečne količine popite tekočine v *ml* na dan,

pri osebah, ki se oz. se ne ukvarjajo s športom. Upoštevaj, da sta dana vzorca velika!

Rezultat lahko preveriš tudi s programom SPSS. Uporabi postopek *Analyze - Compare Means - Independent - Samples T Test* vstavi *KolicinaTD* v *Variable(s)* in *Šport* v *Grouping Variable* ter v *Define Groups* vstavi vrednosti 0 in 1.

Teoretično ozadje

Točkovno ocenjevanje parametra

Pri točkovnem ocenjevanju ocenimo neznani parameter q z vrednostjo slučajne spremenljivke U, ki jo imenujemo *cenilka* parametra q. V statistiki imamo največkrat opravka z ocenjevanjem populacijskega povprečja, disperzije in standardnega odklona ter deleža (verjetnosti).

Naj bo X statistična spremenljivka in naj bo

$$(X_1, X_2, \cdots, X_n)$$

njen vzorec velikosti n. Same vrednosti X_i so sedaj tudi statistične spremenljivke, ker se od vzorca do vzorca spreminjajo, zato jih pišemo z velikimi črkami.

- Cenilka za populacijsko povprečje μ statistične spremenljivke X je (vzorčno povprečje)
$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
- Cenilka za populacijsko disperzijo σ^2 spremenljivke X je (vzorčna disperzija)
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$
- Cenilka za standardni odklon σ spremenljivke X je (vzorčni standardni odklon)
$$S = \sqrt{\frac{1}{n-1}\left(X_i - \bar{X}\right)^2}.$$

Intervalno ocenjevanje parametrov

Pri točkovnem ocenjevanju ocenimo dani parameter z neko fiksno vrednostjo, zato je ocenjevanje bolj ali manj zanesljivo. Zato je za ocenjevanje parametrov primernejše t.i. *intervalno ocenjevanje*. Pri intervalnem ocenjevanju parameter q ocenimo z *intervalom zaupanja* $[D, G]$ in *stopnjo zaupanja* $1 - \alpha$. Pri tem je stopnja zaupanja običajno 0.95 ali 0.99. To pomeni, da lahko z verjetnostjo $1 - \alpha$ (npr. 0.95) trdimo, da parameter q na populaciji leži med vrednostima D in G. Intervale zaupanja računamo neposredno iz cenilk parametra q in njihove porazdelitve.

Interval zaupanja za povprečje

Veliki vzorci (velikosti $n > 30$).

Naj bo statistična spremenljivka X na populaciji porazdeljena kakorkoli, ne nujno normalno. Določimo interval zaupanja za povprečje μ pri stopnji zaupanja $1 - \alpha$. Ločimo dva primera:

- Naj bo standardni odklon σ statistične spremenljivke znan. Potem se izkaže, da je njeno vzorčno povprečje $\bar{X}$ porazdeljeno približno normalno $N(\mu, \sigma/\sqrt{n})$ z matematičnim upanjem μ in standardnim odklonom $\sigma/\sqrt{n}$. Za vrednost

  $$SE = \sigma/\sqrt{n}$$

  standardnega odklona vzorčne porazdelitve $\bar{X}$ v literaturi zasledimo tudi ime *standardna napaka vzorčnega povprečja*. Zato je statistika

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

  porazdeljena približno standardizirano normalno. Pri danem α lahko s pomočjo tabele A izračunamo tak z_α, da je

  $$P(|Z| \leq z_\alpha) = 1 - \alpha.$$

  Npr. pri $\alpha = 0.05$ je $z_{0.05} = 1.96$, pri $\alpha = 0.01$ je $z_{0.01} = 2.58$. To pomeni, da z verjetnostjo $1 - \alpha$ velja

  $$-z_\alpha \leq \frac{\bar{X} - \mu}{SE} \leq z_\alpha \quad \text{oziroma}$$
  $$\bar{X} - z_\alpha SE \leq \mu \leq \bar{X} + z_\alpha SE.$$

  Zato je

  $$\left[\bar{X} - z_\alpha SE, \bar{X} + z_\alpha SE\right]$$

interval zaupanja za povprečje μ na stopnji zaupanja $1-\alpha$. Vidimo, da je to simetričen interval glede na vzorčno povprečje $\bar{X}$ in največjo oddaljenostjo $z_\alpha SE = z_\alpha \sigma/\sqrt{n}$.

- Naj bo sedaj standardni odklon σ statistične spremenljivke X neznan. Za oceno standardnega odklona vzamemo vzorčni standardni odklon S. Potem se izkaže, da je vzorčno povprečje $\bar{X}$ porazdeljeno približno normalno $N\left(\mu, S/\sqrt{n}\right)$ z matematičnim upanjem μ in standardnim odklonom $SE = S/\sqrt{n}$. Zato je statistika
$$Z = \frac{\bar{X}-\mu}{SE} = \frac{\bar{X}-\mu}{S}\sqrt{n} \approx N(0,1)$$
porazdeljena približno standardizirano normalno. Kot v prejšnjem primeru naj bo z_α tak, da je $P(|Z| \leq z_\alpha) = 1-\alpha$. Potem je
$$\left[\bar{X} - z_\alpha SE, \bar{X} + z_\alpha SE\right]$$
interval zaupanja za povprečje μ na stopnji zaupanja $1-\alpha$. To je simetričen interval glede na $\bar{X}$ in oddaljenostjo $z_\alpha SE = z_\alpha S/\sqrt{n}$.

Majhni vzorci (velikosti $n \leq 30$).

Naj bo statistična spremenljivka X na populaciji porazdeljena normalno $N(\mu, \sigma)$. Spet ločimo dva primera, ali je standardni odklon σ znan ali neznan.

- Naj bo standardni odklon σ znan. Potem je vzorčno povprečje $\bar{X}$ porazdeljeno normalno $N\left(\mu, \sigma/\sqrt{n}\right)$ z matematičnim upanjem μ in standardnim odklonom $SE = \sigma/\sqrt{n}$. Kot pri velikih vzorcih je sedaj
$$\left[\bar{X} - z_\alpha SE, \bar{X} + z_\alpha SE\right]$$
interval zaupanja za povprečje μ na stopnji zaupanja $1-\alpha$.

- Standardni odklon σ ni znan. Potem je matematično upanje vzorčne statistike $\bar{X}$ enako μ in njen standardni odklon je $SE = S/\sqrt{n}$. Izkaže se, da je statistika
$$T = \frac{\bar{X}-\mu}{SE} = \frac{\bar{X}-\mu}{S}\sqrt{n} \sim S(n-1)$$
porazdeljena po Studentovem zakonu z $n-1$ prostostnimi stopnjami. Pri danem α iz tabele B izbereno tak t_α, da je
$$P(|T| \leq t_\alpha) = 1-\alpha.$$

Npr. pri $n-1=10$, $\alpha=0.05$ je $t_{0.05}=2.3$. To pomeni, da z verjetnostjo $1-\alpha$ velja

$$-t_\alpha \le \frac{\bar{X}-\mu}{SE} \le t_\alpha \quad \text{oziroma}$$
$$\bar{X}-t_\alpha SE \le \mu \le \bar{X}+t_\alpha SE.$$

Zato je

$$\left[\bar{X}-t_\alpha SE, \bar{X}+t_\alpha SE\right]$$

interval zaupanja za povprečje μ na stopnji zaupanja $1-\alpha$.

Interval zaupanja za disperzijo

Naj bo statistična spremenljivka X na populaciji porazdeljena normalno $N(\mu,\sigma)$ z neznanim σ.

Majhni vzorci (velikosti $n \le 30$).

V tem primeru dobimo interval zaupanja na stopnji zaupanja $1-\alpha$ za disperzijo σ^2 tako, da uporabimo statistiko

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2\,(n-1)\,,$$

ki je porazdeljena po zakonu hi kvadrat z $n-1$ prostostnimi stopnjami. V tabelah za to porazdelitev najprej določimo taki vrednosti χ_1^2 in χ_2^2, da velja

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \frac{\alpha}{2}.$$

V tabeli C je $\chi_1^2 = \chi^2_{\frac{\alpha}{2}}\,(n-1)$ in $\chi_2^2 = \chi^2_{1-\frac{\alpha}{2}}\,(n-1)$. Potem je

$$P\left(\chi_1^2 \le \frac{(n-1)\,S^2}{\sigma^2} \le \chi_2^2\right) = 1-\alpha$$

in od tod izpeljemo, da je

$$\left[\frac{(n-1)\,S^2}{\chi_2^2}, \frac{(n-1)\,S^2}{\chi_1^2}\right]$$

interval zaupanja za disperzijo σ^2 s stopnjo zaupanja $1-\alpha$. Lahko rečemo tudi, da je

$$\left[\frac{\sqrt{n-1}S}{\chi_2}, \frac{\sqrt{n-1}S}{\chi_1}\right]$$

interval zaupanja za standardni odklon σ s stopnjo zaupanja $1-\alpha$.

Veliki vzorci (velikosti $n > 30$).

Za veliki vzorec velikosti n lahko statistiko

$$Z = \frac{S}{\sigma}\sqrt{2(n-1)} - \sqrt{2n-3} \approx N(0,1)$$

aproksimiramo s standardizirano normalno porazdelitvijo. Naj bo z_α tako število, da je $P(|Z| \leq z_\alpha) = 1-\alpha$. Potem je

$$P\left(-z_\alpha \leq \frac{S}{\sigma}\sqrt{2(n-1)} - \sqrt{2n-3} \leq z_\alpha\right) = 1-\alpha$$

in izpeljemo, da z verjetnostjo $1-\alpha$ velja

$$\frac{\sqrt{2(n-1)}S}{\sqrt{2n-3}+z_\alpha} \leq \sigma \leq \frac{\sqrt{2(n-1)}S}{\sqrt{2n-3}-z_\alpha}.$$

Interval zaupanja za verjetnost (delež

Veliki vzorci (velikosti $n > 30$).

Denimo, da želimo na populaciji oceniti delež p enot z določeno lastnostjo L. Frekvenca X lastnosti L v vzorcu velikosti n je spremenljivka, ki je porazdeljena po binomskem zakonu $b(n,p)$. Za velike n lahko omenjeno porazdelitev aproksimiramo z normalno porazdelitvijo $N\left(np, \sqrt{np(1-p)}\right)$ oziroma lahko zapišemo

$$Z = \frac{X-np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1).$$

Naj bo $\bar{p} = \frac{X}{n}$ vzorčni delež. Izkaže se, da lahko za velike n tudi porazdelitev vzorčnih deležev aproksimiramo z normalno porazdelitvijo $N\left(p, \sqrt{\bar{p}(1-\bar{p})/n}\right)$ z matematičnim upanjem p in standardnim odklonom $SE(\bar{p}) = \sqrt{\bar{p}(1-\bar{p})/n}$. Zato je statistika

$$Z = \frac{\bar{p}-p}{SE(\bar{p})} = \frac{\bar{p}-p}{\sqrt{\bar{p}(1-\bar{p})}}\sqrt{n} \approx N(0,1)$$

porazdeljena približno standardizirano normalno. Zato je

$$[\bar{p} - z_\alpha SE(\bar{p}), \bar{p} + z_\alpha SE(\bar{p})]$$

interval zaupanja za verjetnost p na stopnji zaupanja $1-\alpha$, kjer je $P\left(|Z| \leq z_\alpha\right) = 1-\alpha$.

Interval zaupanja za razliko povprečij

Majhni vzorci (velikosti $n \leq 30$).

Naj bo statistična spremenljivka X porazdeljena po zakonu $N(\mu, \sigma)$, spremenljivka Y pa po zakonu $N(\nu, \sigma)$, kjer so vsi populacijski parametri μ, ν, σ neznani. Zanima nas interval zaupanja za razliko $\mu - \nu$ populacijskih povprečij s stopnjo zaupanja $1-\alpha$. Naj bosta $(X_1, X_2, \cdots, X_m)$ in $(Y_1, Y_2, \cdots, Y_n)$ neodvisna vzorca spremenljivk X in Y. Za vsak vzorec določimo cenilke

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \ , \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \ , \quad \text{in}$$

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2 \ , \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

ter vpeljemo

$$S^2 = \frac{(m-1)\,S_X^2 + (n-1)\,S_Y^2}{m+n-2}$$

Izkaže se, da je matematično upanje spremenljivke $\bar{X} - \bar{Y}$ enako populacijski razliki povprečij $\mu - \nu$ in njena disperzija je enaka

$$S^2 \frac{m+n}{nm},$$

torej je standardna napaka vzorčne razlike $\bar{X} - \bar{Y}$ enaka

$$S\sqrt{\frac{m+n}{nm}}$$

Statistika

$$T = \frac{\bar{X} - \bar{Y} - (\mu - \nu)}{S}\sqrt{\frac{nm}{m+n}} \sim S\,(m+n-2)$$

je porazdeljena po Studentovem zakonu z $m+n-2$ prostostnimi stopnjami. Če izberemo t_α, da je $P\left(|T| \leq t_\alpha\right) = 1-\alpha$, je interval zaupanja za populacijsko razliko $\mu - \nu$ na stopnji zaupanja $1-\alpha$ enak

$$\left[\bar{X} - \bar{Y} - t_\alpha S\sqrt{\frac{m+n}{nm}}, \bar{X} - \bar{Y} + t_\alpha S\sqrt{\frac{m+n}{nm}}\right].$$

Veliki vzorci (velikosti $n > 30$).

Če sta v zgornjem obravnavanem primeru oba vzorca velika, potem lahko porazdelitev spremenljivke $\bar{X} - \bar{Y}$ aproksimiramo z normalno porazdelitvijo

$$N\left(\mu - \nu, \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\right).$$

Naj bo z_α tak, da je $P\left(|Z| \leq z_\alpha\right) = 1 - \alpha$, kjer je Z porazdeljena standardizirano normalno. Potem je

$$\left[\bar{X} - \bar{Y} - z_\alpha\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}, \bar{X} - \bar{Y} + z_\alpha\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\right]$$

interval zaupanja za populacijsko razliko $\mu - \nu$ s stopnjo zaupanja $1 - \alpha$.

Literatura

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

[2] Š. Adamič: Temelji biostatistike, Medicinska fakulteta Univerze v Ljubljani, Ljubljana 1995.

[3] R. Jamnik: Verjetnostni račun in statistika, DMFA, Ljubljana 1995.

[4] B. R. Kirkwood, J. A. C. Sterne: Essential medical statistics, Blackwell Publishing company, Malden 2004.

[5] B. Sluban: Uporaba statističnih metod v tekstilstvu, Fakulteta za strojništvo Univerze v Mariboru, Maribor 2004.