UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Teja Rupnik

Določitev velikosti vzorca

Delo diplomskega seminarja

Mentorica: prof. dr. Maja Pohar Perme

Ljubljana, 2018

Kazalo

Določitev velikosti vzorca

Povzetek

V svoji diplomski nalogi sem prikazala izpeljave velikosti vzorca pri različnih statističnih raziskavah in različnih pogojih, ki so jim rezultati raziskav morali zadostovati. Prikazala sem, kako formulo za določitev velikost vzorca izpeljemo iz intervala zaupanja oziroma stopnje značilnosti, iz moči testa, kdaj uporabimo aproksimacije in postopek iteracij ter kako ustrezno velikost vzorca poiščemo s tabeliranjem in iskanjem ničel. Izpeljave sem prikazala na asimptotsko normalno porazdeljenih, eksaktno normalno porazdeljenih, nesimetrično porazdeljenih in χ^2-porazdeljenih testnih statistikah za oceno deleža, povprečne vrednosti, Pearsonovega koeficienta korelacije, razmerja obetov in variance. Ugotovila sem, da iz normalno porazdeljenih testnih statistikah, formulo za velikost vzorca izpeljemo direktno iz intervala zaupanja in moči testa, medtem ko pri testnih statistikah, ki niso normalno porazdeljene, uporabimo aproksimacije in iteracijo ali poiščemo iskano velikost vzorca s pomočjo tabeliranja ter z iskanjem ničel.

Sample size determination

Abstract

In my thesis I was solving sample size estimations problems in different statistical analysis with different conditions, witch had to be fulfilled. I demonstrated how to dissolve equation for sample size from confidence interval or significance level α, from power of test, when to use approximation and iteration, how to estimate sample size from generated table and with root finding. I will show derived formulas from asymptotic normal distribution, exact normal distribution, skewed distribution and χ^2-distribution for sample proportion, sample average, Pearson correlation coefficient, odds ration and sample variance. I've discovered that for normal distributed test statistics sample size formula is dissolved directly from confidence interval and power of test, while for non-normal distributed test statistics we use approximations and iteration or we determined sample size based on generated table and with root finding.

Math. Subj. Class. (2010): 62B15, 62D05, 62E17, 62F03, 62H10, 62N02.

Ključne besede: statistična raziskava, velikost vzorca, moč testa, interval zaupanja, Cohenova velikost učinka.

Keywords: statistical analysis, sample size, power of test, confidence interval, Cohen's effect size.

1. Uvod

Statistične raziskave morajo biti temeljito načrtovane. Dobro načrtovanje ima več pomembnih delov, eden izmed njih je pravilna določitev ustrezne velikosti vzorca. S Slednjim problemom se bom tudi sama ukvarjala v svoji diplomski nalogi.

Vzorec mora biti dovolj velik, da bodo rezultati naše raziskave zadoščali vsem zahtevam, ki si jih bomo na začetku zadali, torej stopnji značilnosti, moči testa in velikosti učinka, ter da bomo rezultate raziskave lahko posplošili na populacijo. Velikost vzorca ima tudi ekonomski pomen. Če bomo imeli premajhen vzorec, rezultata ne bomo mogli ustrezno posplošiti na populacijo, medtem ko bomo s prevelikim vzorcem po nepotrebnem zapravili preveč denarja. Kadar so v raziskavah vključeni ljudje ali živali je velikost vzorca pomembna iz etičnih razlogov. Če bomo imeli premajhen vzorec, lahko rezultate napačno interpretiramo in z njimi škodujemo drugim. Z prevelikim vzorcem bomo izpostavili več kot potrebno število ljudi ali živali.

Najpomembnejši viri na tem področju so Mace (1964), Kraemer in Thiemann (1987), Cohen (1988), Desu in Raghavarao (1990), Lipsey (1990), Shuster (1990), Odeh in Fox (1991). Na spletu najdemo veliko programske opreme za preučevanje analize moči in spletnih kalkulatorjev, ki nam izračunajo potrebno velikost vzorca.

Najprej bomo pogledali nekaj osnovne teorije o populaciji, vzorcu, porazdelitvi in testiranju hipotez. Nato bomo podrobneje predstavili vse izpeljave velikosti vzorcev, ki sem jih naredila. V poglavju 3.1. se bomo ukvarjali z asimptotsko porazdeljenimi testnimi statistikami za testiranje hipotez o deležu in eksaktno normalno porazdeljeno testno statistiko za testiranje hipoteze o povprečni vrednosti. Za oceno deleža bomo naredili primere določitev velikosti vzorca iz intervala zaupanja, testiranja hipoteze v enem vzorcu in testiranja hipoteze v dveh vzorcih. Za oceno povprečne vrednosti pa testiranja hipoteze v dveh vzorcih. V naslednjem poglavju 3.2. bomo pogledali testno statistiko s Studentovo $t-$porazdelitvijo za testiranje hipoteze o povprečni vrednosti v enem vzorcu. V poglavju 3.3. bomo izpeljali velikost vzorca za Pearsonov koeficient korelacije in v poglavju 3.4. za razmerje obetov. Nazadnje bomo pokazali, kako velikost vzorca poiščemo pri testu variance, kjer ima testna statistika χ^2-porazdelitev.

2. Teorija

2.1. **Populacija.** Populacija je celotna množica, ki jo želimo statistično preučiti in za katero bodo rezultati študije posplošeno držali.

Osredotočili se bomo na populacijo z neskončno mnogo elementi in vsakega izmed njih označili z indeksom $1, 2, 3, \ldots$. Karakteristike oziroma slučajne spremenljivke dane populacije, ki jih želimo preučiti, bomo označili z veliko črko, na primer Y. Vrednost slučajne spremenljivke za določen element pa z Y_1. Cilj preučevanja vzorca je določiti parametre porazdelitve iz katere izhaja slučajna spremenljivka. [10]

Najpogosteje uporabljeni parametri:

- **Povprečna vrednost populacije:** Povprečno vrednost slučajne spremenljivke X označimo z μ in jo izračunamo s formulo:

$$\mu = \int_{-\infty}^{\infty} x f_X(x) dx,$$

  kjer je $f_X(x)$ funkcijska gostota.

- **Delež populacije:** Naj bo Y dihotomna spremenljivka.

  Dihotomna spremenljivka nam pokaže prisotnost neke lastnosti in je definirana kot:

$$Y_i = \begin{cases} 1 & \text{, če ima } i\text{-ti element lastnost } Y, \\ 0 & \text{, če } i\text{-ti element nima lastnosti } Y. \end{cases}$$

  S p označimo delež elementov v populaciji z iskano lastnostjo:

$$p = \mathbf{P}(Y = 1).$$

- **Varianca:** Z varianco porazdelitve v populaciji merimo razpršenost vrednosti elementov. Označimo jo s σ^2 in jo izračunamo s formulo:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx.$$

- **Standardni odklon:** Označimo ga s σ in je pozitivni koren variance, torej:

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx}.$$

2.2. **Vzorec.** Statistični vzorec je podmnožica populacije, na podlagi katere dobimo ocene parametrov.

Cilj proučevanja vzorca je določiti vrednosti parametrov populacije s pomočjo podatkov iz vzorca. Predpostavimo, da imamo vzorec sestavljen iz n elementov, ki smo jih naključno in neodvisno izbrali iz populacije, in da ima vsak element vrednost neke slučajne spremenljivke Y. Z Y_i označimo vrednost slučajne spremenljivke za i-ti element.

Oceno povprečne vrednosti, deleža in variance lahko dobimo direktno iz vzorčnega povprečja, deleža vzorca in vzorčne variance. Cenilke parametrov v populaciji označimo s ' ˆ ' (hat) nad simbolom za parameter. Vsi rezultati statističnih testov bodo izpeljani iz slučajnih vzorcev velikosti n.

Najpogosteje uporabljene cenilke so:

- **Vzorčno povprečje:** $\bar{X} = \widehat{\mu} = \sum_{i=1}^{n} \frac{X_i}{n}$
- **Delež vzorca:** $\widehat{p} = \sum_{i=1}^{n} \frac{Y_i}{n}$
- **Vzorčna varianca:** $s^2 = \widehat{\sigma}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$
- **Vzorčni standardni odklon:** $s = \widehat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$

Vzorec je **slučajen**, če imajo vsi elementi populacije enako verjetnosti biti izbrani v vzorec.

Ocenjeni parametri praktično nikoli ne bodo identični dejanski vrednosti parametrov populacije. Če bi iz populacije izbrali drug vzorec, bi dobili drugačno oceno iskanega parametra, ki bi bila lahko bodisi bližje bodisi dlje stran od dejanske vrednosti parametra populacije. Ker pa dejanske vrednosti parametra nikoli ne poznamo, ne vemo, kako zelo se naša ocena od le-te razlikuje. Izračunati želimo, za koliko (in s kakšno verjetnostjo) naj bi se naša ocena razlikovala od dejanske vrednosti. Za kar pa je potrebno vedeti nekaj o porazdelitvi slučajnih spremenljivk.[10]

2.3. **Vzorčna porazdelitev.** S simbolom θ označimo parameter, ki ga želimo oceniti, na primer μ, p, ali σ^2. S simbolom $\hat{\theta}$ pa cenilko parametra, na primer $\bar{x}, \widehat{p}$ ali s^2.

Izrek 2.1. ***Krepki zakon velikih števil (KZVŠ).*** *Naj bo* $X_1, X_2, ..., X_n$ *zaporedje neodvisnih in enako porazdeljenih slučajnih spremenljivk, ki imajo pričakovano vrednost* $\mathbf{E}(X_i) = \mu$ *in varianco* $\mathbf{Var}(X_i) = \sigma^2 < \infty$. *Potem velja:*

$$\lim_{n \to \infty} \frac{X_1 + X_2 + ... + X_n}{n} = \mathbf{E}(X).$$

[9]

Izrek 2.2. *Centralni limitni izrek (CLI)*. *Naj bodo $X_1, X_2, ..., X_n$ neodvisne in enako porazdeljene slučajne spremenljivke, za katere je $\mathbf{E}(|X|^2) < \infty$. Tedaj za $\forall z \in \mathbf{R}$ velja:*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\frac{X_1+X_2+...+X_n}{n} - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}} \leq z \right) = \Phi(z),$$

kjer je $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$. [9]

Torej, če ima naključni vzorec končno povprečje μ in varianco σ^2, potem za dovolj velik n porazdelitev vzorčnega povprečja aproksimiramo z gostoto normalne porazdelitve.

Posledica 2.3. *Če poznamo $\mathbf{E}(X)$ in $\mathbf{Var}(X)$, lahko verjetnost $\mathbf{P}(\bar{X} \in (a,b))$ za dovolj velik n ocenimo kot:*

$$\mathbf{P}\left(\frac{\bar{X} - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}} \in \left(\frac{a - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}}, \frac{b - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}} \right) \right) = \Phi\left(\frac{b - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}} \right) - \Phi\left(\frac{a - \mathbf{E}(X)}{\sqrt{\frac{\mathbf{Var}(X)}{n}}} \right).$$

[9]

2.4. **Testiranje hipotez.** Testiranje hipotez je vrsta statističnega testa, pri katerem skušamo zavrniti ničelno hipotezo. Uporabili bomo primer testiranja hipoteze za določitev povprečne vrednosti. Zanimal nas bo populacijski parameter μ in njegova ocena na podlagi vzorca $\bar{X}$. Iz cenilke bomo izpeljali testno statistiko, s katero bomo ničelno hipotezo o populacijskem parametru bodisi zavrnili bodisi sprejeli.

Najprej določimo ničelno hipotezo:

- $h_0 : \mu = \mu_0$

Nato določimo alternativno hipotezo, ki je bodisi enostranska bodisi dvostranska:

- $h_a : \mu > \mu_0$
- $h_a : \mu < \mu_0$
- $h_a : \mu \neq \mu_0$

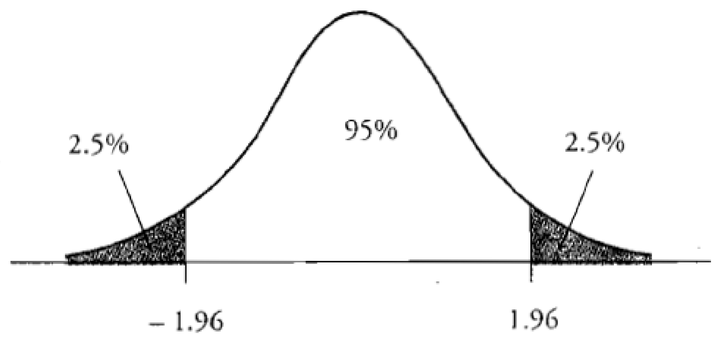
Naj bo torej X slučajna spremenljivka porazdeljena po zakonu $N(\mu, \sigma)$ in naj bo $(X_1, X_2, ..., X_n)$ njen slučajni vzorec, kjer so $(X_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Definiramo standardizirano testno statistiko, tako da naši cenilki odštejemo pričakovano vrednost in jo delimo s standardno napako:

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

Možni izidi oziroma napake pri testiranju hipotez:

- **Napaka I. vrste** se zgodi, kadar zavrnemo ničelno hipotezo, čeprav ničelna hipoteza drži. Verjetnost napake I. vrste označimo z α.
- **Napaka II. vrste** se zgodi, kadar sprejmemo ničelno hipotezo, čeprav ta ne drži. Verjetnost napake II. vrste označimo z β.

Verjetnost napake I. vrste ali stopnja značilnosti α nam pove, kakšna je verjetnost, da napačno zavrnemo ničelno hipotezo. Zato moramo stopnjo značilnosti postaviti čim nižje, najpogosteje kar $\alpha = 0,05$. Ničelno hipotezo torej zavrnemo, ko je $p-$vrednost nižja od α, $p < 0,05$.



SLIKA 1. Dvostranski interval zaupanja pri $\alpha = 0,05$ [10]

Izpeljemo zavrnitveni meji iz napake I. vrste za dvostranski interval zaupanja:

$$
\begin{aligned}
1-\alpha &= \mathbf{P}(\text{sprejmemo } h_0 | h_0 \text{ res}) = \\
&= \mathbf{P}(c_1 < \bar{X} < c_2 | \mu = \mu_0) = \\
&= \mathbf{P}\left(\frac{c_1-\mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{c_2-\mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_0 \right) = \\
&= \mathbf{P}\left(\frac{c_1-\mu_0}{\frac{\sigma}{\sqrt{n}}} < Z < \frac{c_2-\mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_0 \right)
\end{aligned} \tag{1}
$$

Najprej pogledamo za spodnjo mejo c_1. Verjetnost napake je enaka $\frac{\alpha}{2}$, ker gledamo samo desni obarvani del na sliki 1:

$$1-\frac{\alpha}{2} = \mathbf{P}\left(\frac{c_1-\mu_0}{\frac{\sigma}{\sqrt{n}}} < Z \middle| \mu = \mu_0 \right) = 1 - \Phi\left(\frac{c_1-\mu_0}{\frac{\sigma}{\sqrt{n}}} \right) = \Phi\left(\frac{\mu_0-c_1}{\frac{\sigma}{\sqrt{n}}} \right)$$

$$1-\frac{\alpha}{2} = \Phi\left(\frac{\mu_0-c_1}{\frac{\sigma}{\sqrt{n}}} \right)$$

$$\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \frac{\mu_0-c_1}{\frac{\sigma}{\sqrt{n}}}$$

Izrazimo spodnjo zavrnitveno mejo:

$$c_1 = \mu_0 - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Na podoben način izpeljemo zgornjo zavrnitveno mejo c_2:

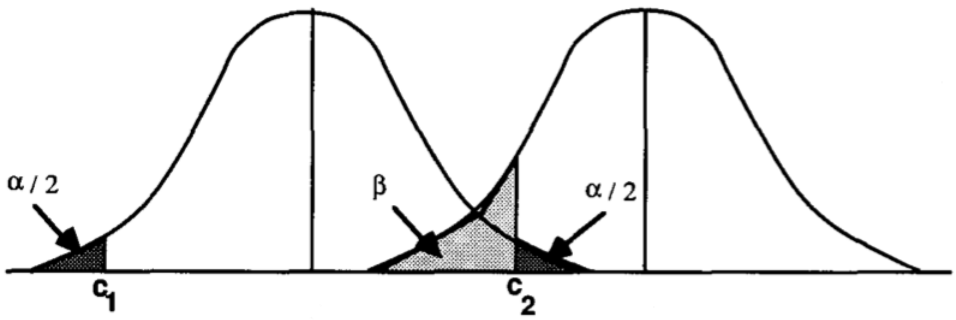
$$1-\frac{\alpha}{2} = \mathbf{P}\left(Z < \frac{c_2-\mu_0}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_0 \right) = \Phi\left(\frac{c_2-\mu_0}{\frac{\sigma}{\sqrt{n}}} \right)$$

$$\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \frac{c_2-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Izrazimo zgornjo zavrnitveno mejo:

$$c_2 = \mu_0 + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Verjetnost napake II. vrste označimo s β, moč testa pa z $1-\beta$. Moč testa želimo čim višjo, na primer $1-\beta = 0,90$.



Slika 2. Porazdelitev pod ničelno in alternativno hipotezo. [10]

Izpeljemo zavrnitveno mejo iz moči testa $1-\beta$ za neko vrednost μ_a:

$$1-\beta = \mathbf{P}(\text{zavrnemo } h_0 | h_0 \text{ ni res}) = \tag{2}$$

$$= \mathbf{P}(c_2 < \bar{X} < c_1 | h_0 \text{ ni res})$$

BŠS: Naj bo $\mu_a > \mu_0$. Zanemarili bomo levi osenčeni del na sliki 2 označen z $\frac{\alpha}{2}$. Ploščina pod funkcijo gostote bo na tem delu zelo majhna in s tem verjetnost, da cenilka pade v ta del, zato jo lahko zanemarimo.

$$\mathbf{P}\left(\frac{c_2-\mu_a}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X}-\mu_a}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_a\right) =$$

$$= \mathbf{P}\left(\frac{c_2-\mu_a}{\frac{\sigma}{\sqrt{n}}} < Z \middle| \mu = \mu_a\right) =$$

$$= 1 - \Phi\left(\frac{c_2-\mu_a}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{\mu_a - c_2}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$\Phi^{-1}(1-\beta) = \frac{\mu_a - c_2}{\frac{\sigma}{\sqrt{n}}}$$

Zavrnitvena meja:

$$c_2 = \mu_a - Z_{1-\beta}\frac{\sigma}{\sqrt{n}}$$

3. Testne statistike

Ogledali si bomo tri primere asimptotsko normalno porazdeljene testne statistike za parameter delež, eksaktno normalno porazdeljeno testno statistiko za primerjavo povprečnih vrednosti v dveh vzorcih, Studentovo $t-$porazdeljeno testno statistiko za povprečno vrednost, Pearsonov koeficient korelacije in razmerje obetov ter njuni asimptotsko normalno porazdeljeni testni statistiki in nazadnje test variance in ustrezno χ^2-porazdeljeno testno statistiko.

3.1. **Normalno porazdeljene testne statistike ($Z-$test).** V naslednjih treh podpoglavjih bomo obravnavali populacijski parameter delež p, kjer bo testna statistika porazdeljena asimptotsko normalno. V prvem primeru si bomo ogledali interval zaupanja, v drugem testiranje hipoteze $p = p_0$ v enem vzorcu in v tretjem testiranje hipoteze $p_1 = p_2$, kjer bomo primerjali deleža dveh vzorcev.

3.1.1. *Določitev deleža.* V dani populaciji bi radi določili delež enot v populaciji, ki imajo neko iskano lastnost. Iskani parameter je populacijski delež, ki ga označimo s p in ga bomo ocenili na podlagi vzorčnega deleža $\widehat{p}$. Velikost vzorca, ki ga potrebujemo za oceno iskanega deleža, bomo izpeljali iz intervala zaupanja.

Naj bo X dihotomna slučajna spremenljivka in $(X_1, X_2, ..., X_n)$ njen slučajni vzorec, kjer so $(X_i)_{i=1}^n$ neodvisne in enko porazdeljene slučajne spremenljivke. Slučajna spremenljivka X za vsak element v populaciji pove, ali ima ta element opazovano lastnost ali ne. Delež elementov v vzorcu, ki imajo iskano lastnost označimo s $\widehat{p}$. [9]

Vsota dihotomnih slučajnih spremenljivk je porazdeljena $\sum_{i=1}^n X_i \sim \mathbf{Bin}(n, p)$.

Izrek 3.1. ***Slutsky.*** *Naj bosta $X_1, X_2, ...$ in $Y_1, Y_2, ...$ zaporedji slučajnih spremenljivki. Naj X_n konvergira v porazdelitvi proti X, ($X_n \xrightarrow{d} X$) in naj Y_n konvergira v verjetnosti proti konstanti c, ($Y_n \xrightarrow{p} c$). Potem velja:*

$$X_n + Y_n \xrightarrow{d} X + c$$

$$Y_n X_n \xrightarrow{d} cX$$

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}, \text{ če } c \neq 0$$

Po definiciji je cenilka $\widehat{p} = \sum_{i=1}^n \frac{X_i}{n}$ povprečna vrednost in po KZVŠ, Izrek (2.6) konvergira proti pričakovani vrednosti $\mathbf{E}(X)$. Po CLI (2.7) bo testna statistika, kjer cenilki $\widehat{p}$ odštejemo pričakovano vrednost in jo delimo s standardno napako, konvergirala proti normalni porazdelitvi $N(0, 1)$.

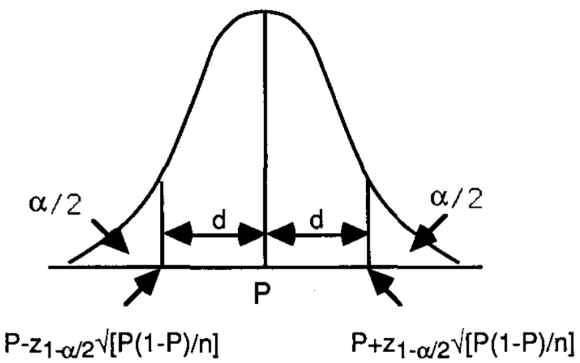
$$\frac{\widehat{p} - \mathbf{E}(X)}{\mathbf{SE}(X)} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0, 1)$$

Vendar pa standardne napake ne poznamo, zato jo bomo ocenili s cenilko $\widehat{\mathbf{SE}}(X) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$. S pomočjo izreka Slutsky (3.1) bo tudi testa statistika, kjer standardno napako nadomestimo z njeno oceno, konvergirala proti normalni porazdelitvi, saj števec v porazdelitvi konvergira proti normalni porazdelitvi, imenovalec pa proti konstanti. Ulomek torej konvergira proti normalni porazdelitvi.

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \xrightarrow{d} N(0, 1)$$

[9, 15]

Z izbrano stopnjo značilnosti α oziroma verjetnostjo napake I. vrste in testno statistiko Z določimo želen interval zaupanja.



SLIKA 3. Dvostranski interval zaupanja. Na sliki je cenilka $\widehat{p}$ označena s P. [10]

Ker imamo v tem primeru dvostranski interval zaupanja, izrazimo napako I. vrste z $\frac{\alpha}{2}$ podobno, kot smo to naredili v izpeljavi (1). Pričakovana vrednost je enaka $\mathbf{E}(\widehat{p}) = p$, ocenjena standardna napaka pa $\widehat{\mathbf{SE}}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$.
Dobimo zavrnitveni meji:

$$c = \widehat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Interval zaupanja:

$$IZ = \left[\widehat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right]$$

Z d označimo oddaljenost ocene populacijskega deleža od meje zavrnitve:

$$d = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Širina intervala zaupanja je enaka $2d$, torej je z verjetnostjo $1-\alpha = 95\%$ cenilka od prave vrednosti oddaljena za manj kot d.
Iz zgornje enačbe izrazimo velikost vzorca n pri želeni širini intervala $2d$:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \widehat{p}(1-\widehat{p})}{d^2}$$

Zanima nas še, kako izbrati vrednost $\widehat{p}$, če je ne poznamo oziroma ne znamo določiti. Priporočena izbira za $\widehat{p}$ je $0,5$, ker nam ta da največjo velikost vzorca in s tem konzervativno oceno, saj je pri tej izbiri produkt $\widehat{p}(1-\widehat{p})$ najvišji in z večanjem razlika med $\widehat{p}$ in $(1-\widehat{p})$ počasi pada. Vzorec bo pri tej izbiri kvečjemu prevelik, vendar se bomo zagotovo izognili morebitnemu premajhnemu vzorcu. [10]

Primer 3.2. Zanima nas delež otrok na nekem območju, ki so bili cepljeni za neko bolezen. Koliko otrok moremo preučiti, da bo ocena deleža odstopala od dejanskega deleža populacije za manj kot 0.1 v vsako smer z verjetnostjo $1 - \alpha = 95\%$?

$$\mathbf{P}(|\widehat{p} - p| < d) = 0,95$$

Podatki:
$d = 0,1$; $\alpha = 0,05$; $Z_{1-\frac{\alpha}{2}} = 1,960$; $\widehat{p} = 0,5$
Podatke vstavimo v enačbo:

$$n = \frac{1,960^2 * 0,25}{0,10^2} = 96,04$$

Potrebujemo $n = 97$ otrok, da bomo z verjetnostjo $1 - \alpha = 95\%$ dobili oceno, ki bo od dejanske vrednosti odstopala za manj kot $0,1$. ♢[10]

Opomba 3.3. Ker pa vemo, da je dejanski populacijski delež bližje vrednosti $0,3$, bomo naredili enak izračun za $\widehat{p} = 0,3$. Izračunamo, da je velikost vzorca enaka $n = 81$.

Včasih namesto, da ocenjena vrednost $\widehat{p}$ pade znotraj 10 procentih točk okoli dejanske vrednosti p, želimo, da pade znotraj 10% od dejanske vrednosti p. Torej, če bi bila prava vrednost deleža $p = 0,20$, bi v zgornjem primeru ocenjena vrednost padla med $0,10$ in $0,30$ v 95 od 100 vzorcev. Namesto tega zahtevamo, da pade ocenjena vrednost v 10% od $0,20$. S tem bi v 95 od 100 vzorcev dobili oceno, ki bi padla znotraj $0,20 + 0,10 * 0,20 = 0,22$ in $0,20 - 0,10 * 0,20 = 0,18$.

Definicija 3.4. Naj bo θ neznani populacijski parameter in naj bo $\hat{\theta}$ ocena za θ. Definiramo **natančnost** ϵ kot:

$$\epsilon = \frac{|\hat{\theta} - \theta|}{\theta}$$

Posledica 3.5. *Če prej definirano enačbo za razdaljo med dejansko vrednostjo in ocenjeno vrednostjo $d = |p - \widehat{p}|$, delimo z ocenjeno vrednostjo $\widehat{p}$, dobimo:*

$$\epsilon = \frac{|p - \widehat{p}|}{\widehat{p}} = Z_{1-\frac{\alpha}{2}} \frac{\sqrt{1 - \widehat{p}}}{\sqrt{n\widehat{p}}}$$

Iz te enačbe izrazimo n in dobimo formulo za potrebno velikost vzorca:

$$n = Z^2_{1-\frac{\alpha}{2}} \frac{(1 - \widehat{p})}{\epsilon^2 \widehat{p}}$$

[10]

Opomba 3.6. Ker v zgornji enačbi produkt $\widehat{p}(1 - \widehat{p})$ več ne nastopa, za konzervativno oceno velikost vzorca ne moremo uporabiti $\widehat{p} = 0,5$.

Primer 3.7. Vzamemo primer 3.2, le da tokrat zahtevamo oceno znotraj 10% od dejanske vrednosti p. Ponovno predvidevamo, da je dejanski populacijski delež enak $0,3$.

Podatki:
$\widehat{p} = 0,3$; $\alpha = 0,05$; $Z_{1-\frac{\alpha}{2}} = 1,960$; $\epsilon = 0,1$
Podatke vstavimo v enačbo:

$$n = \frac{1,960^2 * 0,7}{0,1^2 * 0,3} = 896,37$$

Potrebujemo $n = 897$ otrok, da bomo z verjetnostjo $1 - \alpha = 95\%$ dobili oceno znotraj 10% dejanske populacijske vrednosti. $\diamondsuit$[10]

3.1.2. *Testiranje hipoteze: populacijski delež.* Opazujemo dihotomno slučajno spremenljivko X in njen slučajni vzorec $(X_1, X_2, ..., X_n)$, kjer so $(X_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Ponovno nas bo zanimal delež elementov v populaciji z neko iskano lastnostjo. Definiramo testno statistiko podobno kot prej:

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1),$$

kjer je $\mathbf{SE}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$ standardna napaka, parameter p delež elementov v populaciji z iskano lastnostjo in $\widehat{p}$ ocena deleža elementov z iskano lastnostjo. Pričakovana vrednost je enaka $\mathbf{E}(\widehat{p}) = p$. Po CLI $(2,7)$ testna statistika Z konvergira k normalni porazdelitvi. S testiranjem hipoteze bi radi preverili, če je delež v populaciji enak p_0.
Definiramo ničelno hipotezo:

$$h_0 : p = p_0$$

Definiramo enostransko alternativno hipotezo:

$$h_a : p > p_0$$

Velikost vzorca bomo izpeljali iz intervala zaupanja, za kar poleg vrednosti testne statistike Z, potrebujemo še željeni velikosti napak I. in II. vrste, α in β. S tem določimo želeno moč testa $1 - \beta$. Označimo dejansko vrednost p v populaciji z oznako p_a.
Uporabimo izpeljavo (1) in izrazimo zgornjo zavrnitveno mejo c iz napake I. vrste, torej kadar h_0 zavrnemo, čeprav h_0 drži:

$$c = p_0 + Z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$$

Uporabimo še izpeljavo (2) in izrazimo zavrnitveno mejo c iz moči testa $1 - \beta$ za neko vrednost p_a:

$$c = p_a - Z_{1-\beta}\sqrt{\frac{p_a(1-p_a)}{n}}$$

Enačbi izenačimo in izrazimo velikost vzorca n:

$$p_0 + Z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} = p_a - Z_{1-\beta}\sqrt{\frac{p_a(1-p_a)}{n}}$$

$$p_a - p_0 = \frac{Z_{1-\alpha}\sqrt{p_0(1-p_0)} + Z_{1-\beta}\sqrt{p_a(1-p_a)}}{\sqrt{n}}$$

$$n = \frac{\left(Z_{1-\alpha}\sqrt{p_0(1-p_0)} + Z_{1-\beta}\sqrt{p_a(1-p_a)}\right)^2}{(p_a-p_0)^2}$$

Uporaba zgornje formule za določitev velikosti vzorca, nam omogoča, da dobimo zahtevano moč testa. [10]

Primer 3.8. V preteklosti je bilo ugotovljeno, da je delež ptic okuženih s tetanusom enak 15%, vendar se nam dozdeva, da je ta številka padla na 10%. Kolikšen vzorec potrebujemo, da bomo pri stopnji značilnosti $\alpha = 0,05$ in z močjo testa $1-\beta = 0,90$ pravilno zavrnili ničelno hipotezo $h_0 : p = 0,15$?

Podatki:
$h_0 : p = 0,15$; $h_a : p = 0,10$; $\alpha = 0,05$; $\beta = 0,10$; $1-\beta = 0,90$
Podatke vstavimo v enačbo:

$$n = \frac{\left(1,645\sqrt{0,15 * 0,85} + 1,282\sqrt{0,10 * 0,90}\right)^2}{(0,10-0,15)^2} = 377,90$$

Potrebujemo $n = 378$ ptic. ♢ [10]

Opomba 3.9. Podoben pristop bi uporabili, če bi imeli dvostransko alternativno hipotezo $h_a : p \neq p_0$, le da bi v enačbi $Z_{1-\alpha}$ nadomestili z $Z_{1-\frac{\alpha}{2}}$.

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}}\sqrt{p_0(1-p_0)} + Z_{1-\beta}\sqrt{p_a(1-p_a)}\right)^2}{(p_a-p_0)^2}$$

[10]

3.1.3. *Testiranje hipoteze: primerjava deležev.* Naj bosta X in Y slučajni spremenljivki ter $(X_1, X_2, ..., X_n)$ in $(Y_1, Y_2, ..., Y_n)$ njuna slučajna vzorca, kjer so $(X_i)_{i=1}^n$ in $(Y_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. X in Y predstavljata dihotomni slučajni spremenljivki, ki povesta kateri elementi v populacijah imajo iskano lastnost. Zanimata nas deleža elementov z iskano lastnostjo v obeh populacijah, ki ju označimo s p_1 in p_2. Njuna vzorčna deleža označimo s $\widehat{p_1}$ in $\widehat{p_2}$. S testiranjem hipoteze želimo preveriti, če sta deleža enaka. Za razliko med dvema populacijskima deležema definiramo nov parameter $p_1 - p_2$, ki ga ocenimo z razliko vzorčnih deležev $\widehat{p_1} - \widehat{p_2}$. Povprečna vrednost vzorčne porazdelitve $\widehat{p_1} - \widehat{p_2}$ je enaka:

$$\mathbf{E}(\widehat{p_1} - \widehat{p_2}) = p_1 - p_2$$

Varianca pa:

$$\mathbf{Var}(\widehat{p_1} - \widehat{p_2}) = \mathbf{Var}(\widehat{p_1}) + \mathbf{Var}(\widehat{p_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Testirati želimo ničelno hipotezo $h_0 : p_1 = p_2$ proti enostranski alternativni hipotezi $h_a : p_1 > p_2$. V vzorčni porazdelitvi pod ničelno hipotezo lahko p_1 in p_2 nadomestimo s skupno oznako p:

$$\mathbf{Var}(\widehat{p_1} - \widehat{p_2}) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Naj bosta oba vzorca enako velika $n_1 = n_2 = n$, potem je $\mathbf{Var}(\widehat{p_1} - \widehat{p_2}) = 2\left(\frac{p(1-p)}{n}\right)$. Oceniti moramo neznani parameter p. Izračunamo ga kot povprečje vrednosti iz obeh vzorcev: $\bar{p} = \frac{p_1+p_2}{2}$.

S pomočjo CLI (2.7) in izreka Slutsky (3.1), bo standardizirana testna statistika Z, kjer cenilki odštejem pričakovano vrednost in jo delimo z oceno standardne napake, konvergirala k normalni porazdelitvi.

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - \mathbf{E}(\widehat{p_1} - \widehat{p_2})}{\widehat{\mathbf{SE}}(\widehat{p_1} - \widehat{p_2})} \xrightarrow{d} N(0,1)$$

Pod ničelno hipotezo ima razlika deležev $\widehat{p_1} - \widehat{p_2}$ pričakovano vrednost:

$$\mathbf{E}(\widehat{p_1} - \widehat{p_2}) = p_1 - p_2 = 0$$

Standardno napako ocenimo z:

$$\widehat{\mathbf{SE}}(\widehat{p_1} - \widehat{p_2}) = \sqrt{\frac{2\bar{p}(1-\bar{p})}{n}}$$

Podobno kot prej iz verjetnosti napake I. vrste, izpeljava (1), izrazimo zavrnitveno mejo c:

$$c = Z_{1-\alpha}\sqrt{\frac{2\bar{p}(1-\bar{p})}{n}}$$

Pod alternativno hipotezo ima ocena razlike deležev $\widehat{p_1} - \widehat{p_2}$ pričakovano vrednost $\mathbf{E}(\widehat{p_1} - \widehat{p_2}) = p_1 - p_2$ in standardno napako $\mathbf{SE}(\widehat{p_1} - \widehat{p_2}) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$.
Iz moči testa $1 - \beta$ za neko vrednost $p_1 - p_2$, izpeljava (2), izrazimo zavrnitveno mejo c:

$$c = (p_1 - p_2) - Z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

Enačbi izenačimo in izpostavimo velikost vzorca n:

$$Z_{1-\alpha}\sqrt{\frac{2\bar{p}(1-\bar{p})}{n}} = (p_1 - p_2) - Z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

$$p_1 - p_2 = \frac{1}{\sqrt{n}}\left(Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right)$$

$$n = \frac{\left(Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right)^2}{(p_1 - p_2)^2}$$

[9, 10]

Primer 3.10. V neki raziskavi, kjer so bili zbrani podatki dveh skupin velikosti 50, so preučevali deleža zaposlenih v industriji z nevarno kemikalijo. V prvi skupini so bili opazovani ljudje, ki so zboleli za neko opazovano boleznijo, v drugi ta bolezen ni bila prisotna. Ugotovili so, da je delež zaposlenih v industriji z nevarno kemikalijo v prvi skupini $p_1 = 0,60$ v drugi pa $p_2 = 0,50$.

Testirati želimo ničelno hipotezo $h_0 : p_1 = p_2$ proti enostranski alternativni hipotezi $h_a : p_1 > p_2$. Kolikšen vzorec potrebujemo, da bomo z močjo testa $1 - \beta = 0,90$ zavrnili ničelno hipotezo pri stopnji značilnosti $\alpha = 0,05$?

Podatki:
$p_1 = 0,60$; $p_2 = 0,50$; $\bar{p} = 0,55$; $\alpha = 0,05$; $\beta = 0,10$
Podatke vstavimo v enačbo:

$$n = \frac{(1,645\sqrt{2 * 0,55 * 0,45} + 1,282\sqrt{0,6 * 0,4 + 0,5 * 0,5})^2}{(0,6 - 0,5)^2} = 422,20$$

Potrebujemo vzorca velikosti $n = 423$. ♢[10]

V naslednjem podpoglavju si bomo ogledali primer, kjer nas bo zanimal parameter povprečna vrednost μ. Primerjali bomo povprečno vrednost v dveh vzorcih $h_0 : \mu_1 = \mu_2$. Testna statistika bo porazdeljena eksaktno normalno.

3.1.4. *Testiranje hipoteze: primerjava povprečnih vrednosti.* Naj bo X slučajna spremenljivka porazdeljena po zakonu $N(\mu, \sigma)$ in $(X_1, X_2, ..., X_n)$ njen slučajni vzorec, kjer so $(X_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Parameter μ predstavlja populacijsko povprečno vrednost, ki jo želimo oceniti z vzorčno povprečno vrednostjo, definirano kot:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i)$$

Pričakovana vrednost je enaka $\mathbf{E}(\bar{X}) = \mathbf{E}\big(\frac{1}{n}\sum_{i=1}^n (X_i)\big) = \frac{1}{n} n\mathbf{E}(X_i) = \mu$
in varianca $\mathbf{Var}(\bar{X}) = \mathbf{Var}\big(\frac{1}{n}\sum_{i=1}^n (X_i)\big) = \frac{1}{n^2} n\mathbf{Var}(X_i) = \frac{\sigma^2}{n}$. [12]

Primerjati želimo povprečni vrednosti iz dveh vzorcev in izračunati željeno velikost vzorca. Izpeljali jo bomo iz intervala zaupanja in moči testa za neko vrednost $\mu_1 - \mu_2$ ter vključili koeficient velikosti učinka. Naj bosta X in Y slučajni spremenljivki porazdeljeni po zakonu $N(\mu_1, \sigma_1)$ in $N(\mu_2, \sigma_2)$. Naj bosta $(X_1, X_2, ..., X_n)$ in $(Y_1, Y_2, ..., Y_n)$ njuna slučajna vzorca, kjer so $(X_i)_{i=1}^n$ in $(Y_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Definiramo nov populacijski parameter $\mu_1 - \mu_2$, ki ga bomo ocenili s pomočjo razlike vzorčnih povprečij $\bar{X}_1 - \bar{X}_2$.

Vzorčna pričakovana vrednost: $\mathbf{E}(\bar{X}_1 - \bar{X}_2) = \mathbf{E}(\bar{X}_1) - \mathbf{E}(\bar{X}_2) = \mu_1 - \mu_2$
Vzorčna varianca: $\mathbf{Var}(\bar{X}_1 - \bar{X}_2) = \mathbf{Var}(\bar{X}_1) + \mathbf{Var}(\bar{X}_2) = \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$

Predpostavili bomo homoskedastičnost varianc: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in enako velikost vzorcev: $n_1 = n_2 = n$
Vzorčna varianca: $\mathbf{Var}(\bar{X}_1 - \bar{X}_2) = \frac{2\sigma^2}{n}$

Testirati želimo hipotezo, ki pravi da sta povprečni vrednosti vzorcev enaki:
$h_0 : \mu_1 = \mu_2$, $h_a : \mu_1 > \mu_2$
Standardizirana testna statistika Z po CLI (2.7) konvergira k normalni porazdelitvi:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mathbf{E}(\bar{X}_1 - \bar{X}_2)}{\mathbf{SE}(\bar{X}_1 - \bar{X}_2)} \sim N(0,1)$$

Podobno kot v izpeljavi (1), zavrnitveno mejo c izrazimo iz verjetnosti napake I. vrste α, kjer sta $\mathbf{E}(\bar{X}_1 - \bar{X}_2) = 0$ in $\mathbf{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{2\sigma^2}{n}}$ pod ničelno hipotezo.
Dobimo torej interval zaupanja $IZ = \left[-\infty, Z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n}}\right]$ in zavrnitveno mejo:

$$c = Z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n}}$$

Prav tako uporabimo izpeljavo (2) za določitev zavrnitvene meje c iz moči testa $1-\beta$ za neko vrednost $\mu_1 - \mu_2$, kjer sta $\mathbf{E}(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ in $\mathbf{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{2\sigma^2}{n}}$ pod alternativno hipotezo.
Dobimo interval zaupanja $IZ = \left[(\mu_1 - \mu_2) - Z_{1-\beta}\sqrt{\frac{2\sigma^2}{n}}, \infty\right]$ in zavrnitveno mejo:

$$c = (\mu_1 - \mu_2) - Z_{1-\beta}\sqrt{\frac{2\sigma^2}{n}}$$

Opomba 3.11. Tudi tu smo zanemarili levi del pri izpeljavi moči testa za $\mu_0 - \mu_a$, kot v izpeljavi (2).

Enačbi izenačimo in izrazimo velikost vzorca n:

$$Z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n}} = (\mu_1 - \mu_2) - Z_{1-\beta}\sqrt{\frac{2\sigma^2}{n}}$$

$$\mu_1 - \mu_2 = Z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n}} - Z_{1-\beta}\sqrt{\frac{2\sigma^2}{n}}$$

$$\mu_1 - \mu_2 = \sqrt{\frac{2\sigma^2}{n}}(Z_{1-\alpha} - Z_{1-\beta})$$

$$n = \frac{2\sigma^2(Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}$$

[10]

Cohen [7] je definiral velikost učinka z enačbo:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Standardiziral ga je v tri razrede:

- $d = 0,2$ majhna velikost učinka
- $d = 0,5$ srednja velikost učinka
- $d = 0,8$ velika velikost učinka

$$n = 2\left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{d}\right)^2$$

Velikost vzorca je odvisna od:

- stopnje značilnosti α
- moči testa $1 - \beta$
- velikosti učinka d

[7, 10]
Vsaka testna statistika ima svoj indeks velikosti učinka. K temu se bomo vrnili v nadaljnjih primerih.

Primer 3.12. V klinični raziskavi bi radi ugotovili ali novo zdravilo vpliva na zmanjšanje krvnega tlaka. Če bo novo zdravilo pokazalo 5 enot spremembe v povprečni vrednosti krvnega tlaka, bo to pomenilo, da zdravilo učinkuje. Kolikšen vzorec potrebujemo, da bomo dobili zahtevane rezultate s stopnjo značilnosti $\alpha = 0,05$ in močjo testa $1 - \beta = 0,80$?
V preteklih študijah je bilo ugotovljeno, da je standardni odklon krvnega tlaka enak 19.

Podatki:
$h_0 : \mu_1 = \mu_2$, $h_a : \mu_1 > \mu_2$
$|\mu_1 - \mu_2| = 5$; $\sigma = 19$; $\alpha = 0,05$; $1 - \beta = 0,80$
Podatke vstavimo v enačbo:

$$d = \frac{5}{19} = 0,26$$

$$n = 2\left(\frac{1,645 + 0,84}{0,26}\right)^2 = 182,70$$

Potrebujemo vzorca velikosti $n = 183$. ♢ [18]

Opomba 3.13. V primeru smo uporabili standardni odklon, ki smo ga privzeli iz drugih študij. Namesto tega bi lahko uporabila eno izmed standardiziranih vrednosti, npr. $d = 0,2$ in dobili rezultat:

$$n = 2\left(\frac{1,645 + 0,84}{0,20}\right)^2 = 308,76$$

[1, 18]

V naslednjem razdelku si bomo ogledali primer, v katerem nas bo zanimal parameter povprečna vrednost μ. Predpostavili bomo, da vrednosti σ ne poznamo in testirali hipotezo $\mu = \mu_0$ v enem vzorcu. Testna statistika bom imela Studentovo $t-$porazdelitev.

3.2. $t-$**test.** Naj bo X slučajna spremenljivka porazdeljena po zakonu $N(\mu, \sigma)$. Izberemo slučajni vzorec oblike $(X_1, X_2, ..., X_n)$, kjer so $(X_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Zanima nas populacijski parameter μ, ki predstavlja aritmetično sredino ali povprečno vrednost. Ocenimo ga z vzorčnim povprečjem, ki ga izračunamo s formulo:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i)$$

Sledi, da je pričakovana vrednost vzorčnega povprečja enaka matematičnemu upanju $\mathbf{E}(\bar{X}) = \mathbf{E}\big(\frac{1}{n}\sum_{i=1}^{n}(X_i)\big) = \mu$ in varianca vzorčnega povprečja enaka $\mathbf{Var}(\bar{X}) = \mathbf{Var}\big(\frac{1}{n}\sum_{i=1}^{n}(X_i)\big) = \frac{\sigma^2}{n}$, ker pa σ v tem primeru ne poznamo, bomo definirali še ustrezno cenilko $\widehat{\mathbf{Var}}(\bar{X}) = \frac{s^2}{n}$, kjer je $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ cenilka standardnega odklona.

Standardizirano testno statistiko izpeljemo, tako da ji odštejemo upanje in jo delimo z oceno standardne napake:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1},$$

Testna statistika ima po CLI (2.7) in izreku Slutsky (3.1) Studentovo $t-$porazdelitev z $(n-1)$ stopnjami prostosti.
[9, 11]

Ogledali si bomo primer izpeljave velikosti vzorca n, kadar standardni odklon σ ni znan in uporabimo oceno s.

3.2.1. *Testiranje hipoteze: povprečna vrednost.* Testirali bomo hipotezo:

$$h_0 : \mu = \mu_0$$

$$h_a : \mu \neq \mu_0$$

Uporabili bomo zgoraj definirano testno statistiko:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

[9]
Podobno kot v izpeljavi (1) iz verjetnosti napake I. vrste izpeljemo zgornjo zavrnitveno mejo c_2:

$$1 - \alpha = \mathbf{P}(\text{sprejmemo } h_0 | h_0 \text{ drži}) = \mathbf{P}(c_1 < \bar{X} < c_2 | h_0 \text{ res}) =$$

$$= \mathbf{P}\left(\frac{c_1 - \mu_0}{\frac{s}{\sqrt{n}}} < \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} < \frac{c_2 - \mu_0}{\frac{s}{\sqrt{n}}} \middle| \mu = \mu_0\right) = \mathbf{P}\left(\frac{c_1 - \mu_0}{\frac{s}{\sqrt{n}}} < T < \frac{c_2 - \mu_0}{\frac{s}{\sqrt{n}}} \middle| \mu = \mu_0\right)$$

Zgornja zavrnitvena meja:

$$c_2 = \mu_0 + t_{1-\frac{\alpha}{2}, n-1}\frac{s}{\sqrt{n}}$$

Izrazimo še zavrnitveno mejo c_2 iz moči testa $1-\beta$ za neko vrednost μ_a, kot v izpeljavi (2):

$$1-\beta = \mathbf{P}(\text{zavrnemo } h_0|h_a \text{ drži}) = \mathbf{P}\left(\frac{c_2-\mu_a}{\frac{s}{\sqrt{n}}} < T \middle| \mu = \mu_a\right)$$

Zavrnitvena meja:

$$c_2 = \mu_a - t_{1-\beta,n-1}\frac{s}{\sqrt{n}}$$

Opomba 3.14. Ponovno zanemarimo levi del pri izpeljavi moči za dan μ_a, kot v izpeljavi (2).

Izenačimo zgornjo zavrnitveno mejo intervala zaupanja pri stopnji značilnosti α in zavrnitveno mejo za moč testa $1-\beta$ pri dani vrednosti μ_a in izpostavimo n:

$$\mu_0 + t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} = \mu_a - t_{1-\beta,n-1}\frac{s}{\sqrt{n}}$$

$$\mu_a - \mu_0 = (t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})\frac{s}{\sqrt{n}}$$

$$n = \frac{s^2(t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{(\mu_a-\mu_0)^2}$$

Ker potrebujemo parameter n za določitev t-vrednosti iz tabele Studentove porazdelitve, ga ne bomo mogli preprosto izpostaviti iz enačbe. Za večje vzorce bi lahko $t-$vrednost nadomestili z $Z-$vrednostjo, ki je asimtotsko normalno porazdeljena.

$$n = \frac{s^2(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{(\mu_a-\mu_0)^2}$$

Pri manjših vzorcih pa za določitev velikosti vzorca potrebujemo še proces iteracij. [10, 16]

Primer 3.15. Veljalo naj bi, da je povprečna teža moškega s 55 leti in novo postavljeno diagnozo bolezni srca enaka $90kg$. Kako velik vzorec potrebujemo, da bi s stopnjo značilnosti $\alpha = 0,05$ in močjo testa $1-\beta = 0,90$ zavrnili ničelno hipotezo proti alternativni, da je povprečna teža padla iz $90kg$ na $85kg$? Pogledali bomo enostransko alternativno hipotezo $h_o : \mu_0 > \mu_a$, zato bomo v enačbi za velikost vzorca $Z_{1-\frac{\alpha}{2}}$ nadomestili z $Z_{1-\alpha}$. Predpostavimo, da je standardni odklon σ omenjene povprečne vrednosti enak $20kg$.

Podatki:
$h_0 : \mu = 90$; $h_a : \mu = 85$; $\sigma = 20$; $\alpha = 0,05$; $1-\beta = 0,90$
Podatke vstavimo v enačbo, kjer $t-$vrednosti nadomestimo z $Z-$vrednostmi:

$$n = \frac{\sigma^2(Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_a-\mu_0)^2} = \frac{20^2(1,645 + 1.282)^2}{(90-85)^2} = 137,08$$

Potrebujemo torej vzorec velikosti $n = 138$. [10] ◊

S simulacijo v programskem jeziku R bomo preverili, če nam izračunana velikost vzorca dejansko da zahtevano moč testa $1-\beta$ za $\mu_0 - \mu_a = 5$. Pri danih populacijskih parametrih bomo naredili 10000 ponovitev izračuna $t-$vrednosti in pogledali v koliko primerih smo ničelno hipotezo zavrnili, torej moč testa. Pričakujemo, da bomo dobili rezultat okoli $1-\beta = 0,90$.

Algorithm 1: Izračun moči testa za $t-$test.

input : n, μ_0, μ_a in σ.
output: moč testa $1-\beta$.

Function $t-$`test`$(n,\ \mu_a,\ \sigma,\ \mu_0)$**:**
  $x \leftarrow rnorm(n, \mu_a, \sigma)$;
  $t.stat \leftarrow \frac{mean(x)-\mu_0}{sd(x)/\sqrt{n}}$;
  return 1 - pt(abs(t.stat), n-1);

$n \leftarrow 138$;
$\mu_0 \leftarrow 90$;
$\mu_a \leftarrow 85$;
$\sigma \leftarrow 20$;

for $i \leftarrow 1$ **to** 10000 **do**
  $pvalues[i] \leftarrow t - test(n, \mu_a, \sigma, \mu_0)$;
return $mean(pvalues < 0,05)$

S simulacijo smo potrdili, da nam izračunana velikost vzorca pri danih populacijski parametrih prinese zahtevano moč testa. Naš algoritem (1) je izračunal moč $1-\beta = 0,9003$.

Poglejmo si še primer, kjer imamo majhen vzorec in bomo poleg aproksimacije uporabili še iteracijo.

Primer 3.16. V neki raziskavi so ugotovili, da je povprečna vrednost časa raztopitve nekega koncentrata v vodo enaka $45s$. Radi bi pokazali, da se dejanska povprečna vrednost od predpostavljene razlikuje za $2s$. Kolikšen vzorec potrebujemo, da bomo s stopnjo značilnosti $\alpha = 0,05$ in močjo testa $1-\beta = 0,90$ zavrnili ničelno hipotezo? Standardni odklon je enak $3s$.

Podatki:
$h_0 : \mu = 45;\ |\mu_0 - \mu_a| = 2;\ \sigma = 3;\ \alpha = 0,05;\ 1-\beta = 0,90$
Podatke vstavimo v enačbo z $Z-$vrednostmi:

$$n = \frac{\sigma^2(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{(\mu_a - \mu_0)^2} = \frac{3^2(1,960 + 1,282)^2}{2^2} = 23,65$$

Podatek $n = 24$ uporabimo v enačbi s t-vrednostmi in izračunamo novo velikost vzorca:

$$n = \frac{\sigma^2(t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{(\mu_a - \mu_0)^2} = \frac{3^2(2,069 + 1,319)^2}{2^2} = 25,83$$

Naslednji korak iteracije nam da rezultat $n = 25.64$, kar pomeni, da se bomo na tem koraku ustavili. Potrebna velikost vzorca je torej $n = 26$. ♢ [16]

Dobljen rezultat bomo preverili s simulacijo v programskem jeziku R. Glede na dobljeno velikost vzorca, stopnjo značilnosti, povprečno vrednost pod ničelno in alternativno porazdelitvijo ter vzorčnim standardnim odklonom bomo pogledali, ali dobimo zahtevano moč testa.

Algorithm 2: Izračun moči testa za $t-$test.

input : n, μ_0, μ_a in σ.
output: moč testa $1-\beta$.

Function $t-$`test`$(n, \mu_a, \sigma, \mu_0)$:
    $x \leftarrow rnorm(n, \mu_a, \sigma)$;
    $t.stat \leftarrow \frac{mean(x)-\mu_0}{sd(x)/\sqrt{n}}$;
    return (2*(1 - pt(abs(t.stat), n-1));

$n \leftarrow 26$;
$\mu_0 \leftarrow 45$;
$\mu_a \leftarrow 47$;
$sigma \leftarrow 3$;

for $i \leftarrow 1$ **to** 10000 **do**
    $pvalues[i] \leftarrow t - test(n, \mu_a, \sigma, \mu_0)$;

return $mean(pvalues < 0,05)$

Uporabili smo enak algoritem kot pri prejšnji nalogi in ponovno pokazali, da nam izračunana velikost vzorca da zahtevano moč testa. Algoritem (2) je pokazal moč testa $1 - \beta = 0,8996$.

V naslednjem poglavju si bomo ogledali Pearsonov koeficient korelacije in testiranje hipoteze $\rho = 0$. Uporabili bomo Fisherjevo transformacijo, s katero bomo dobili asimptotsko normalno porazdelitev. Velikost vzorca bomo izrazili iz intervala zaupanja in moči testa.

3.3. **Pearsonov koeficient korelacije.** Pearsonov koeficient korelacije je mera, ki predstavlja linearno povezanost med dvema slučajnima spremenljivkama. Naj bo (X, Y) slučajni vektor, kjer je X porazdeljen po zakonu $N(\mu_1, \sigma)$ in Y po zakonu $N(\mu_2, \tau)$. Kovarianca med slučajnima spremenljivkama je definirana kot:

$$\mathbf{cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

Njun korelacijski koeficient pa kot:

$$\rho(X,Y) = \frac{\mathbf{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- $\rho = -1$ opazovani spremenljivki sta močno negativno povezani.
- $\rho = 0$ opazovani spremenljivki sta neodvisni
- $\rho = 1$ opazovani spremenljivki sta močno pozitivno povezani.

[12]
Zanima nas torej, kako sta ti dve slučajni spremenljivki med seboj povezani. Formulo za velikost vzorca bomo izpeljali iz intervala zaupanja in moči testa. Populacijski parameter, ki nas bo v tem odlomku zanimal, je korelacijski koeficient ρ.

Naj bo $(X_1,Y_1),...,(X_n,Y_n)$ slučajni vzorec vektorja (X,Y), kjer so $(X_i,Y_i)_{i=1}^n$ med seboj neodvisni slučajni vektorji, vsi porazdeljeni kot vektor (X,Y). Vzorčni korelacijski koeficient označimo z r in lahko zavzema vrednosti med -1 in 1. Izračunamo ga s pomočjo vzorčne kovariance in vzorčnih standardnih odklonov:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y},$$

kjer sta vzorčna standardna odklona:

$$s_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

$$s_Y = \sqrt{\frac{\sum_{i=1}^n (Y_1 - \bar{Y})^2}{n}}$$

in vzorčna kovarianca:

$$s_{XY} = \frac{1}{n}\left(\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}\right)$$

S testiranjem hipoteze želimo preveriti, kakšna je korelacija med opazovanima spremenljivkama oziroma ali sta spremenljivki povezani:

$$h_o : \rho = 0$$

$$h_a : \rho \neq 0$$

Testna statistika pod ničelno hipotezo je definirana kot:

$$T = \frac{r-\rho}{\widehat{\mathbf{SE}}(r)} = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}\ ,$$

kjer je $\widehat{\mathbf{SE}}(r) = \sqrt{\frac{1-r^2}{n-2}}$ ocena standardne napake [7, 8]. Prava vrednost variance je $\mathbf{Var}(r) = \frac{(1-\rho^2)^2}{n-1}$ in s tem standardna napaka $\mathbf{SE}(r) = \frac{1-\rho^2}{\sqrt{n-1}}$. Vendar bi dobili previsoko vrednost testne statistike in $p-$vrednost, ker vrednosti ρ ne poznamo, r pa ni vedno natančna ocena za ρ. Zato moramo uporabiti oceno standardne napak $\widehat{\mathbf{SE}}(r) = \sqrt{\frac{1-r^2}{n-2}}$. Testna statistika je asimptotsko normalno porazdeljena le za ničelno hipotezo $h_0 : \rho = 0$, saj je tu $\mathbf{E}(r) = \rho$, za hipoteze o $\rho \neq 0$ bi dobili asimetrično vzorčno porazdelitev. Po CLI (2.7) in izreku Slutsky (3.1) testna statistika T konvergira k $t-$porazdelitev z $(n-2)$ stopnjami prostosti.

Fisher [2, 8] je razvil transformacijo, kjer z naraščanjem velikosti vzorca hitreje dobimo normalno porazdelitev.

$$A = arctanh(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Transformiran A ima varianco $\mathbf{Var}(A) = \frac{1}{n-3}$, ki je neodvisna od vrednosti korelacije in je konstantna ter pričakovano vrednost $\mathbf{E}(A) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$. Tako dobimo novo testno statistiko [8]:

$$Z = \frac{A - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\sqrt{\frac{1}{n-3}}} \xrightarrow{d} N(0,1),$$

ki po CLI (2.7) konvergira k normalni porazdelitvi [2].

Izberemo stopnjo značilnosti α in izrazimo zgornjo zavrnitveno mejo c_2 iz verjetnosti napake I. vrste, kot v izpeljavi (1):

$$c_2 = Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}$$

Izberemo moč testa $1-\beta$ in izrazimo zavrnitveno mejo c_2 iz moči testa, kot v izpeljavi (2):

$$c_2 = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - Z_{1-\beta}\sqrt{\frac{1}{n-3}}$$

Enačbi izenačimo in izrazimo velikost vzorca n:

$$\begin{aligned}
Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}} &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - Z_{1-\beta}\sqrt{\frac{1}{n-3}} \\
\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) &= \sqrt{\frac{1}{n-3}}(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}) \\
\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\right)^2 &= \frac{1}{n-3}(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \\
n-3 &= \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\frac{1}{4}\ln\left(\frac{1+r}{1-r}\right)^2}
\end{aligned}$$

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\frac{1}{4}\ln\left(\frac{1+r}{1-r}\right)^2} + 3$$

Vrnimo se k velikosti učinka, ki smo ga omenili v podpoglavju (3.1.4.). Tu je velikost učinka parameter r, ki je standardiziran v tri razrede:

- $r = 0,10$: majhna velikost učinka
- $r = 0,30$: srednja velikost učinka
- $r = 0,50$: velika velikost učinka

Za določitev velikosti vzorca pri testiranju Pearsonovega koeficienta korelacije potrebujemo stopnjo značilnosti α, moč testa $1-\beta$ in velikost učinka oziroma Pearsonov koeficient korelacije r. [7, 19]

Primer 3.17. Študije so pokazale, da je korelacija med količino užite soli in krvnim pritiskom enaka $0,30$. Kolikšen vzorec potrebujemo, da zavrnemo ničelno hipotezo $h_0 : \rho = 0$ s stopnjo značilnosti $\alpha = 0,01$ in močjo testa $1-\beta = 0,90$?

Podatki:
$r = 0,30$; $\alpha = 0,01$; $1-\beta = 0,9$
Podatke vstavimo v enačbo:

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\frac{1}{4}\ln\left(\frac{1+0,3}{1-0,3}\right)^2} + 3 = \frac{(2,5758 + 1,282)^2}{\frac{1}{4}\ln\left(\frac{1+0,3}{1-0,3}\right)^2} + 3 = 158,35$$

Za testiranje hipoteze potrebujemo vzorec velikosti $n = 159$. ♢

V naslednjem poglavju si bomo ogledali parameter razmerje verjetij. Uporabili bomo logaritemsko transformacijo, s katero se bomo približali normalni porazdelitvi in nato izrazili velikost vzorca n iz intervala zaupanja in moči testa.

3.4. **Razmerje obetov (OR).** Razmerje obetov uporabljamo, kadar želimo opazovati vplive nekega zdravila na določeno bolezen. Za takšne študije uporabljamo dva vzorca, in sicer na prvem uporabimo neko zdravilo na drugem pa ne. Označimo verjetnosti:

- p_1 : verjetnost bolezni ob prisotnosti preučevanega faktorja
- $(1 - p_1)$: verjetnost, da bolezni ni ob prisotnosti nekega faktorja
- p_2 : verjetnost bolezni brez prisotnosti faktorja
- $(1 - p_2)$: verjetnost, da bolezni ni, ko ne uporabimo faktorja

Razmerje obetov izračunamo kot:

$$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

Ker je razmerje obetov (OR) strogo pozitivno in redko večje od 10, po navadi pa dosti manjše, sledi, da vzorčna porazdelitev ocene OR ni normalno porazdeljena s strogo pozitivnim koeficientom asimetrije (ang. skewness). Transformacija z naravni logaritmom bi zato izboljšala lastnosti porazdelitve in jo približala normalni porazdelitvi. Običajno torej za določanje intervala zaupanja razmerja obetov in testiranje hipotez o razmerju obetov uporabimo naravni logaritem razmerja obetov $\ln OR$. Kasneje rezultate pretvorimo nazaj $e^{\ln(OR)} = OR$. [10]

S $\widehat{p_1}$ in $\widehat{p_2}$ označimo ocene verjetnosti prisotnosti bolezni v obeh vzorcih. Naj bosta oba vzorca enako velika $n_1 = n_2 = n$, potem je aproksimacija variance vzorčne porazdelitve $\ln OR$ enaka:

$$\mathbf{Var}(\ln \widehat{OR}) = \mathbf{Var}\left(\ln\left(\frac{\widehat{p_1}(1-\widehat{p_2})}{\widehat{p_2}(1-\widehat{p_2})}\right)\right) =$$

$$= \mathbf{Var}(\ln(\widehat{p_1}) + \ln(1-\widehat{p_2}) - \ln(\widehat{p_2}) - \ln(1-\widehat{p_1})) =$$

$$= \mathbf{Var}(\ln(\widehat{p_1}) - \ln(1-\widehat{p_1})) + \mathbf{Var}(\ln(1-\widehat{p_2}) - \ln(\widehat{p_2})) =$$

$$= \mathbf{Var}\left(\ln\left(\frac{\widehat{p_1}}{(1-\widehat{p_1})}\right)\right) + \mathbf{Var}\left(\ln\left(\frac{\widehat{p_2}}{(1-\widehat{p_2})}\right)\right) =$$

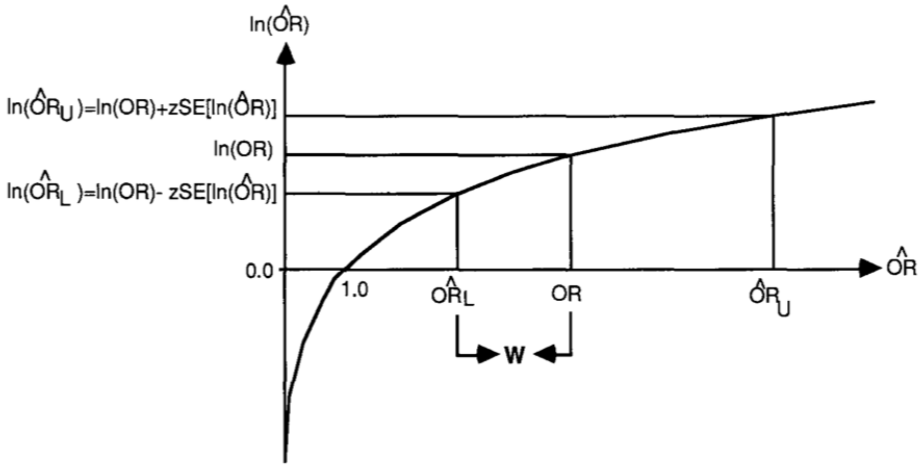
$$= \frac{1}{np_1(1-p_1)} + \frac{1}{np_2(1-p_2)} = \frac{1}{n}\left(\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}\right)$$

Iz predzadnje v zadnjo vrstico uporabimo metodo delta za iskanje aproksimacije variance $\mathbf{Var}(g(\widehat{\theta})) \doteq g'(\theta)\mathbf{Var}(\widehat{\theta})$ [4]

Zanima nas kakšno velikost vzorca n potrebujemo za oceno OR znotraj ϵ od prave vrednosti OR s stopnjo značilnosti $1 - \alpha$. Velikost vzorca bomo izrazili iz intervala zaupanja. Standardno napako ocenimo z: $\mathbf{SE}(\ln \widehat{OR}) = \sqrt{\frac{1}{n}\left(\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}\right)}$

Definiramo zgornjo in spodnjo mejo intervala zaupanje porazdelitve $\ln OR$:

- $\ln \widehat{OR}_L = \ln OR - z\mathbf{SE}(\ln \widehat{OR})$
- $\ln \widehat{OR}_U = \ln OR + z\mathbf{SE}(\ln \widehat{OR})$



SLIKA 4. Transformacija intervala zaupanja okoli OR v interval zaupanja okoli $\ln OR$ [10]

Z w označimo razdaljo med $\widehat{OR}_L$ in OR. $w = \epsilon OR$, kjer je ϵ enak

$$\epsilon = \frac{|\widehat{OR}_L - OR|}{OR}$$

Opomba 3.18. Ogledali si bomo razdaljo med $\widehat{OR_L}$ in OR, ker je razdalja med $\widehat{OR_U}$ in OR veliko večja in s tem tudi izpeljani vzorec. Smiselna izbira je torej razdalja med spodnjo mejo in OR, saj prinese zadostno velikost vzorca, ki nam prinese želene rezultate.

Razdaljo w zapišemo kot:

$$w = \epsilon OR = e^{\ln OR} - e^{\ln OR - z\mathbf{SE}(\ln \widehat{OR})}$$

Enačbo preuredimo:

$$\begin{aligned}
\epsilon OR &= OR - ORe^{-Z\mathbf{SE}(\ln(\widehat{OR}))} \\
\epsilon &= 1 - e^{-Z\mathbf{SE}(\ln(\widehat{OR}))} \\
1 - \epsilon &= e^{-Z\mathbf{SE}(\ln(\widehat{OR}))} \\
\ln(1-\epsilon) &= -(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\mathbf{SE}(\ln(\widehat{OR})) \\
\ln(1-\epsilon) &= -(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{n}\left(\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}\right)}
\end{aligned}$$

Izrazimo n:

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2\left(\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}\right)}{\ln(1-\epsilon)^2}$$

Kadar želimo določiti razmerje verjetij vpliva zdravila na neko bolezen, moramo določiti stopnjo značilnosti α, verjetnost napake II. vrste β, parametre p_1, p_2 in OR ter dovoljeno odstopanje ϵ izražen v procentih. [5, 10]

Opomba 3.19. V resnici potrebujemo le dva izmed treh parametrov p_1, p_2 in OR. Tretjega pa lahko izrazimo s pomočjo formule:

$$p_1 = \frac{OR * p_2}{OR * p_2 + (1-p_2)}$$

Primer 3.20. Kako velik vzorec bi potrebovali v študiji s kontrolno skupino, če bi želeli z močjo testa $1-\beta = 0.90$ oceniti razmerje verjetij znotraj 25% prave vrednosti, če predvidevamo, da je prava vrednost 2 in je izpostavljenost v kontrolni skupini enaka $0,30$?

$p_2 = 0,30$; $OR = 2$
Izračunamo p_1: $p_1 = \frac{OR*p_2}{OR*p_2+(1-p_2)} = \frac{2*0,30}{2*0,30+0,70} = 0,46$

$\epsilon = 0,25$; $\alpha = 0,05$; $\beta = 0,90$
Podatke vstavimo v enačbo:

$$n = \frac{(1,960+1,282)^2\left(\frac{1}{0,46*0,54}+\frac{1}{0,3*0,7}\right)}{\ln(1-0,25)^2} = 1116,03$$

Potrebujemo torej skupini velikosti $n = 1117$. ♢ [10]

V naslednjem poglavju si bomo ogledali test variance in kako izpeljati ustrezno velikost vzorca, kadar je preučevani parameter varianca σ in ima testna statistika χ^2-porazdelitev.

3.5. **Test variance.** S testom preverjamo hipotezo o varianci s χ^2-porazdeljeno testno statistiko. Naj bo slučajna spremenljivka X porazdeljena po zakonu $N(\mu,\sigma)$ in $(X_1, X_2, ..., X_n)$ njen slučajni vzorec, kjer so $(X_i)_{i=1}^n$ neodvisne in enako porazdeljene slučajne spremenljivke. Zanima nas populacijski parameter varianca σ^2. Oceniti ga želimo s pomočjo vzorčne variance oz. popravljene vzorčne variance:

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_1-\bar{X})^2 \ , \ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_1-\bar{X})^2$$

Izpeljemo testno statistiko iz popravljene vzorčne variance:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Pokazali bomo, zakaj je testna statistika χ^2 porazdeljena χ^2_{n-1}:

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n(X_i-\bar{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{(X_i-\mu)-(\bar{x}-\mu)}{\sigma}\right)^2 =$$

$$= \sum_{i=1}^{n}\left(\frac{(X_i-\mu)^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2} - 2\frac{(X_i-\mu)(\bar{X}-\mu)}{\sigma^2}\right) =$$

$$= \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} + n\frac{(\bar{X}-\mu)^2}{\sigma^2} - 2\frac{(\bar{X}-\mu)}{\sigma}\sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^2}$$

Zadnjo vsoto preoblikujemo:

$$\sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma} = \frac{\sum_{i=1}^n X_I - n\mu}{\sigma} = \frac{n\bar{X}-n\mu}{\sigma} = \frac{n(\bar{X}-\mu)}{\sigma}$$

In dobimo:

$$= \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} + n\frac{(\bar{X}-\mu)^2}{\sigma^2} - 2n\frac{(\bar{X}-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^2} - \left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

Vsota n kvadratov standardiziranih normalno porazdeljenih slučajnih spremenljivk je porazdeljena:

$$\sum_{i=1}^{n} \frac{(X_i - \mu)}{\sigma^2} \sim \chi_n^2$$

Kvadrat standardizirane normalno porazdeljene slučajne spremenljivke pa:

$$\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2 \sim \chi_1^2$$

Izkaže se, da je razlika porazdeljena kot χ_{n-1}^2:

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu)}{\sigma^2} - \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2 \sim \chi_{n-1}^2 \ [13]$$

Testirati želimo hipotezo:

$$h_0 : \sigma^2 = \sigma_0^2$$

$$h_a : \sigma^2 \neq \sigma_0^2$$

Ponovno lahko izpeljemo interval zaupanja iz napake I. vrste, kot v izpeljavi (1):

$$\alpha = \mathbf{P}(c_1 < \widehat{s} < c_2 | h_0 \text{ res}) = \mathbf{P}\left(\frac{c_1(n-1)}{\sigma_0^2} < \frac{\widehat{s}(n-1)}{\sigma_0^2} < \frac{c_2(n-1)}{\sigma_0^2} \middle| \sigma^2 = \sigma_0^2 \right) =$$

$$= \mathbf{P}\left(\frac{c_1(n-1)}{\sigma_0^2} < \chi^2 < \frac{c_2(n-1)}{\sigma_0^2} \middle| \sigma^2 = \sigma_0^2 \right)$$

Poglejmo si najprej spodnjo zavrnitveno mejo c_1:

$$\frac{\alpha}{2} = \mathbf{P}\left(\frac{c_1(n-1)}{\sigma_0^2} < \chi^2 \middle| \sigma^2 = \sigma_0^2 \right) = 1 - F\left(\frac{c_1(n-1)}{\sigma_0^2} \right)$$

$$\chi_{1-\frac{\alpha}{2}, n-1}^2 = \frac{c_1(n-1)}{\sigma_0^2}$$

$$c_1 = \frac{\chi_{1-\frac{\alpha}{2}, n-1}^2 \sigma_0^2}{n-1}$$

Izpeljemo še zgornjo zavrnitveno mejo c_2:

$$\frac{\alpha}{2} = \mathbf{P}\left(\chi^2 < \frac{c_2(n-1)}{\sigma_0^2} \middle| \sigma^2 = \sigma_0^2 \right) = F\left(\frac{c_2(n-1)}{\sigma_0^2} \right)$$

$$\chi_{\frac{\alpha}{2}, n-1}^2 = \frac{c_2(n-1)}{\sigma_0^2}$$

$$c_2 = \frac{\chi_{\frac{\alpha}{2}, n-1} \sigma_0^2}{n-1}$$

Funkcija F je kumulativna porazdelitvena funkcija χ^2-porazdelitve.

Interval zaupanja:

$$IZ = \left[\frac{\chi_{1-\frac{\alpha}{2}, n-1} \sigma_0^2}{n-1}, \frac{\chi_{\frac{\alpha}{2}, n-1}^2 \sigma_0^2}{n-1} \right]$$

[21]

Vendar tudi v tem primeru ne moremo preprosti izpostaviti velikost vzorca n, saj n nastopa v stopnjah prostosti χ^2-porazdelitve. Zato bomo uporabili drugačen pristop, ki si ga bomo ogledali na primeru testiranja hipoteze:

Ničelna hipoteza: $h_0 : \sigma^2 = \sigma_0^2$
Enostranska alternativna hipoteza: $h_a : \sigma^2 \geq \sigma_0^2 + \delta$

Poiskali bomo velikost vzorca, ki bo pri željeni stopnji značilnosti α in moči testa $1-\beta$ zavrnila ničelno hipotezo in pokazala razliko $> \delta$. Pogledali bomo dva primera. Najprej bomo poiskali velikost vzorca s tabeliranjem in iz tabele razbrali katera velikost vzorca nam da želeno moč testa, nato bom poiskali velikost vzorca z iskanjem ničle funkcije, ki jo bomo definirali.

Podobno kot v izpeljavi (2), izpeljemo zavrnitveno mejo iz moči testa za neko vrednost $\sigma_0^2 + \delta$:

$$1-\beta = \mathbf{P}(\text{zavrnemo } h_0 | h_a \text{ drži}) = \mathbf{P}(c_1 < \widehat{s} | h_a \text{ drži}) =$$

$$= \mathbf{P}\left(\frac{c_1(n-1)}{\sigma_0^2+\delta} < \frac{\widehat{s}(n-1)}{\sigma_0^2+\delta} \middle| \sigma^2 = \sigma_0^2 + \delta \right) =$$

$$= \mathbf{P}\left(\frac{c_1(n-1)}{\sigma_0^2+\delta} < \chi^2 \middle| \sigma^2 = \sigma_0^2 + \delta \right) = 1 - F\left(\frac{c_1(n-1)}{\sigma_0^2+\delta} \right)$$

$$\chi^2_{\beta,n-1} = \frac{c_1(n-1)}{\sigma_0^2+\delta}$$

Zavrnitvena meja:

$$c_1 = \frac{\chi^2_{\beta,n-1}(\sigma_0^2+\delta)}{n-1} = \frac{\chi^2_{\beta,n-1}\sigma_0^2(1+\frac{\delta}{\sigma_0^2})}{n-1} = \frac{\chi^2_{\beta,n-1}\sigma_0^2 R}{n-1},$$

kjer je $R = 1 - \frac{\delta}{\sigma_0^2}$.

Zavrnitveno mejo izenačimo z zavrnitveno mejo pri enostranskem intervalu zaupanja:

$$\frac{\chi^2_{\beta,n-1}\sigma_0^2 R}{n-1} = \frac{\chi^2_{\alpha,n-1}\sigma_0^2}{n-1}$$

$$\chi^2_{\beta,n-1} = \frac{\chi^2_{\alpha,n-1}}{R}$$

Generiramo tabelo, kjer za vrednosti stopnji značilnost $\nu = n-1$, na primer med 1 in 200, izračunamo kvantil $\chi^2_{\beta,\nu}$ in kumulativno porazdelitveno funkcijo $F(\chi^2|\nu)$, ki jo označimo s C_ν.

$$C_\nu = \mathbf{P}(\chi^2 > \chi^2_{\beta,\nu})$$

Kvantil $\chi^2_{\alpha,\nu}$ je kritična vrednosti χ^2 porazdelitve pri stopnji prostosti ν in stopnji značilnosti $\alpha = 0,05$. Funkcija C_ν izračuna verjetnost napake II. vrste β pri dani stopnji prostosti ν in kvantilu $\chi^2_{\beta,\nu}$ za testno statistiko χ^2. Vrednost ν, ki nam vrne C_ν najbližje zahtevani napaki II. vrste β, je pravilna stopnja prostosti ν in dobimo velikost vzorca $n = \nu + 1$.

Drug način je iskanje ničle spodaj definirane funkcije. Za določitev stopnje prostosti ν, in s tem velikost vzorca n, bomo poiskali ničlo funkcije C_ν.

$$C_\nu = F\left(\frac{F^{-1}(\alpha,\nu)}{R},\nu\right) - \beta,$$

kjer je F^{-1} inverzna funkcija χ^2, torej nam vrne kvantil $\chi^2_{\alpha,\nu}$ pri dane stopnji značilnosti α in stopnji prostosti ν. Porazdelitvena funkcija F nam vrne verjetnost napake II. vrste β, pri katerem je kvantil $\chi^2_{\beta,\nu}$ zavrnitvena meja pri dani stopnji prostosti ν. Ničlo funkcije C_ν bomo dobili za stopnjo prostosti ν, ki bo dala verjetnost napake II. vrste enako želeni vrednosti β. [17]

Za oba primera tabeliranja in iskanje ničle funkcije, bomo uporabil programski jezik R.

Primer 3.21. Varianca pri meritvah upornosti silicijevih diod naj bi bila enaka $100(ohm * cm)^2$. Kupec bo sprejel ponudbo, če je dejanska vrednost varianca za 55 večja od omenjene, torej $155(ohm * cm)^2$. Kolikšen vzorec potrebujemo, da bomo s stopnjo značilnosti $\alpha = 0,05$ in močjo testa $1-\beta = 0,90$ zavrnili ničelno hipotezo?

Podatki:
$h_0 : \sigma^2 = 100$, $h_a : \sigma^2 = 100 + 55 = 155$
$\sigma_0^2 = 100$; $\delta = 55$; $\alpha = 0,05$; $1-\beta = 0,90$
Izračunamo R:

$$R = 1 + \frac{\delta}{\sigma_0^2} = 1 + \frac{55}{100} = 1,55$$

Poiščemo ničlo prej definirane funkcije:

$$C_\nu = F\left(\frac{F^{-1}(\alpha,\nu)}{R},\nu\right) - \beta$$

Iščemo vrednost ν za katero bo kvantil $\chi^2_{\beta,\nu}$ zavrnitvena meja pri želeni moči testa $1-\beta = 0,90$ za vrednost $\sigma_0^2+\delta = 155$. Inverzna funkcija F^{-1} nam vrne kvantil $\chi^2_{\alpha,\nu}$, ki je zgornja zavrnitvena meja intervala zaupanja pri stopnji značilnosti $\alpha = 0,05$ in stopnji prostosti ν. Porazdelitvena funkcija F nam vrne vrednost napake II. vrste β pri danem kvantilu $\chi^2_{\beta,\nu}$ in pri dani stopnji prostosti ν. Ničlo bomo dobili za tisto vrednost ν, kjer bo porazdelitvena funkcija F vrnila vrednost napake II. vrste enako želeni vrednosti $\beta = 0,10$.

Algoritem (3) nam je izračunal, da je $C_\nu = 0$ pri $\nu = 169,3$, torej je potrebna velikost vzorca enaka $n = 170$. ♢ [17]

Prikazana sta oba algoritma (3) in (4) za izračun velikost vzorca pri testu variance v programskem jeziku R. V prvem iščemo ničlo funkcije C_ν, to je torej ν, ki nam da verjetnost napake II. vrste β. V drugem pa izpišemo tabelo s tremi stolpci: ν, $\chi^2_{\beta,\nu}$ in C_ν za vrednost ν med 1 in 200 in pogledamo kateri ν nam da C_ν najbližje želeni verjetnosti napake II. vrste β.

Algorithm 3: Iskanje stopnje prostosti ν [17].

input : δ, σ_0^2, α in β.
output: stopnja prostosti ν.

$\delta \leftarrow 55;$
$\sigma_0^2 \leftarrow 100;$
$R \leftarrow 1 + \frac{\delta}{\sigma_0^2};$
$\alpha \leftarrow 0,05;$
$\beta \leftarrow 0,01;$

$C_\nu = F\left(\frac{F^{-1}(\alpha,\nu)}{R}, \nu\right) - \beta;$
$size = uniroot(C_\nu, c(1, 200);$
$size\$root;$

Algorithm 4: Iskanje stopnje prostosti ν [17].

input : δ, σ_0^2 in α.
output: stopnja prostosti ν.

$\delta \leftarrow 55;$
$\sigma_0^2 \leftarrow 100;$
$R \leftarrow 1 + \frac{\delta}{\sigma_0^2};$
$\alpha \leftarrow 0,05;$

$x \leftarrow matrix(nrow = 200, ncol = 3);$
for $\nu \leftarrow 1$ **to** 200 **do**
  $\chi^2_{\beta,\nu} = \frac{qchisq(1-\alpha,\nu)}{R};$
  $C_\nu = pchisq(\chi^2_{\beta,\nu}, \nu);$
  $x[\nu, 1] = \nu;$
  $x[\nu, 2] = \chi^2_{\beta,\nu};$
  $x[\nu, 3] = C_\nu;$
return $x[165 : 175,]$

4. Zaključek

V svoji nalogi sem ugotovila, da lahko za normalno ali asimptotsko normalno porazdeljene testne statistike izpeljemo formulo za velikost vzorca iz intervala zaupanja oziroma stopnje značilnosti α. Običajno sicer zahtevamo, da bo naša raziskava imela neko določeno moč testa $1-\beta$. Takrat izpeljemo formulo za velikost vzorca, tako da izenačimo zgornjo zavrnitveno mejo pri intervalu zaupanja in zavrnitveno mejo iz moči testa. Pri testnih statistikah, ki imajo Studentovo $t-$porazdelitve, izračunamo velikost vzorca, tako da $t-$vrednosti nadomestimo z $Z-$vrednostmi in nato uporabimo postopek iteracij, dokler dva koraka iteracije ne pokažeta enakega rezultata. Pri Pearsonovem koeficientu korelacije si pomagamo s Fisherjevo transformacije, da dobimo asimptotsko normalno porazdeljeno testno statistiko, ki zavzeva vrednosti na realni osi. Tako lahko velikost vzorca ponovno izpeljemo iz intervala zaupanja in moči testa. Tudi ocena razmerja obetov ni normalno porazdeljena. Normalno porazdelitev dobimo s transformacijo naravnega logaritma ter izrazimo velikost vzorca iz intervala zaupanja in moči testa. Testa statistika pri testu variance ima χ^2-porazdelitev. Velikost vzorca poiščemo s tabeliranjem, tako da pogledamo pri kateri stopnji prostosti bo izračunana verjetnost napake II. vrste, pri dani zavrnitveni meji $\chi^2_{\beta,\nu}$, enaka zahtevani verjetnosti napake II. vrste.

Slovar strokovnih izrazov

approximation približek – približna vrednost
attribute atributivna spremneljivka – katero vrednost opišemo z besedami
binary variable dihotomna spremenljivka – spremenljivka, ki zavzame dve vrednosti
coefficient of skewness koeficient asimetrije – merilo, ki meri asimetrijo verjetnostne porazdelitve realne slučajne spremenljivke
confidence level stopnja zaupanja – verjetnost s katero bo prava vrednost parametra padla znotraj itervala zaupanja
critical value meja zavrnitve – meja znotraj katere sprejmemo ničelno hipotezo
degrees of freedom stopnja prostosti – je število neodvisnih vrednosti slučajne spremenljike, ki se lahko v statističnih izračunih spreminjajo
effect size velikost učinka – razlika med aritmetičnima sredinama pod ničelno in pod alternativno hipotezo
Fisher's transformation Fisherjeva transformacija – transformacija, ki asimetrično vzorčno porazdelitev spremeni v približno normalno
homoscedasticity homoskedastičnost – varianca enaka za vse opazovane vrednosti spremenljivke
iteration iteracija – ponavljanje
non-parametric test neparametrični test – test, kjer preverjamo atributivne lastnosti slučajne spremenljivke
odds ratio razmerje obetov – pove, za koliko se verjetnost za določeni pojav poveča ali zmanjša ob prisotnosti nekega faktorja
Pearson's coefficient of correlation Pearsonov koeficient korelacije – velikost linearne povezanosti med dve spremenljivkama
power of a test moč testa – verjetno zavrnitve ničelne hipoteze, če ta ne drži
precision natančnost – stopnja, za katero nadaljnje meritve ali izračuni kažejo enake ali podobne rezultate
random sample sličajni vzorec – vsi elementi v populaciji imajo enako verjetnost, da bodo izbrani v vzorec
sample size velikost vzorca
Slutsky theorem Slutskyov izrek
standardisation standardizacija – postopek s katerim verdnosti spremenljivke transformiramo, tako da odštevejmo povprečno vrednost in delimo s standardnim odklonom
test statistic testna statistika
variance varianca – razpršenost

Literatura

[1] C. L. Chuan in J. Penyelidikan, *Sample size estimation using Krejcie and Morgan and Cohen statistical power analysis: A comparison*, Jurnal Penyelidikan IPBL (2006) strani 88–81.

[2] R. A. Fisher, *On a Distribution Yielding the Error Functions of Several Well Known Statistics*, Proceedings of the International Congress of Mathematics, Toronto, 1924, strani 805?813.

[3] R. V. Lenth, *Some Practical Guidelines for Effective Sample Size Determination*, The American Statistician, Vol. 55, No. 3 (2001) strani 187–193.

[4] R. Dorfman, *A Note on the ?-Method for Finding Variance Formulae*, The Biometric Bulletin, 1, 1938, 129?137.

[5] H. Wang, S. Chow in G. Li, *On sample size calculation based on odds ratio in clinical trials*, Journal of biopharmaceutical statistics Vol. 12, No. 4, (2002) strani 471?483.

[6] C. R. Wilson VanVoorhis in B. L. Morgan, *Understanding Power and Rules of Thumb for Determining Sample Sizes*, Tutorials in Quantitative Methods for Psychology vol. 3 (2), (2007) strani 43–50.

[7] J. Cohen, *Statistical power analysis for the behavioral sciences - 2nd ed.*, Lawrence Erlbaum Associates, New York, 1988.

[8] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, London, 1936.

[9] R. Jamnik, *Matematična statistika*, Matematika-fizika **12**, Državna založba Slovenije, Ljubljana, 1979.

[10] S. Lemeshow, D. W. Hosmer Jr, J. Klar in S. K. Lwang, *Adequacy of sample size in health studies*, John Wiley & Sons Ltd., Baffins Lane Chichester England, 1990.

[11] S. A. Nolan in T. E. Heinzen, *Statistics for theBehavioral Sciences - 2nd ed.*, Worth Publishers, New York, 2012.

[12] M. Pohar Perme, *Verjetnost in statistika z nalogami*, Ekonomska fakulteta v Ljubljani, Založništvo , Ljubljana, 2014.

[13] J. A. Rice, *Mathematical Statistics and Data Analysis, Third Edition*, Thomson Brooks/Cole , Duxbury, 2007.

[14] J. Mauko, *Testiranje statističnih hipotez*, diplomsko delo, Fakulteta za naravoslovje in matematiko, Univerza v Mariboru, 2010.

[15] S. Babič, *Interval zaupanja za populacijsko povpre?je, ko porazdelitev preu?evane spremenljivke ni normalna*, diplomsko delo, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Univerza na Primorskem, 2015.

[16] *Engineering statistics handbook*, [ogled 25. 8. 2018], dostopno na `https://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm`.

[17] *Engineering statistics handbook*, [ogled 20. 8. 2018], dostopno na `https://www.itl.nist.gov/div898/handbook/prc/section2/prc232.htm`.

[18] *Power and Sample Size Determination*, [ogled 25. 5. 2018], dostopno na `http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_power/BS704_Power_print.html`.

[19] *Statsdirect*, [ogled 2. 6. 2018], dostopno na `https://www.statsdirect.com/help/sample_size/correlation.htm`.

[20] *StatLect*, [ogled 20. 8. 2018], dostopno na `https://www.statlect.com/fundamentals-of-statistics/hypothesis-testing-variance`.

[21] *xycoon*, [ogled 31. 8. 2018], dostopno na `https://www.xycoon.com/ht_variance.htm`.