Zapiski pri predmetu Statistika

Minimalni katalog znanja, ki ga bom sproti dopolnjeval. Verjetno bom izpustil kakšen dokaz in pa kakšen zgled.

Ljubljana, 2017 Gregor Vavdi

1 Motivacija

Kako bi "ocenili"verjetnost, da pri metu kovanca pade cifra?
Izvedemo n neodvisnih "enakih"(v istih razmerah, na enak način, pošteno oz.naključno) metov kovanca in iskano verjetnost ocenimo z razmerjem $\frac{\text{število cifer}}{n}$.

Igramo igro, kjer kroglico položimo v eno od treh škatel. Zmešamo škatle med seboj in poskušamo uganiti kje je kroglica. Če uganemo dobimo 10, v nasprotnem primeru pa izgubimo 6.

Kako bi ocenili pričakovano vrednost te igre?
Izvedemo n neodvisnih slučajnih iger in pričakovano vrednost ene igre ocenimo z $\frac{\text{skupni izkupiček}}{n}$.

Zdi se nam, da mora z večjim vzorcem priti boljša ocena.

V 18. stoletju je grof Buffon kovanec vrgel 4040-krat in dobil 2048 cifer. Ocenjena verjetnost cifre je 0.50689.
V 19. stoletju je Pason vrgel kovanec 12000-krat in dobil 6019 cifer. Ocenjena vrejetnost je 0.5016.

Aksiome verjetnosti zgradimo tako, da so naša mnenja glede vprašanj upravičena.

2 Konvergenca slučajnih spremenljivk in limitni izrek

Definicija 2.1. *Naj bodo $X_1, X_2, X_3, \ldots$ slučajne spremenljivke, definirane na skupnem prostoru Ω.*

(1) Pravimo, da zaporedje $\{X_n\}_n$ konvergira k X v porazdelitvi, če

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$$

za vsa tista realna števila x, v katerih je komulativna porazdelitvena funkcija slučajne spremenljivke X zvezna.

(2) Pravimo, da zaporedje $\{X_n\}_n$ konvergira k X ***v verjetnosti****, če velja:*

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$$

za vsak $\varepsilon > 0$.

(3) Pravimo, da zaporedje $\{X_n\}_n$ konvergira k X ***skoraj gotovo****, če je:*

$$P(\{\omega \in \Omega | \exists \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

$$\Longleftrightarrow$$

$$P(\lim_{n\to\infty} Xn = X) = 1$$

TRDITEV 2.2. *Iz konvergence 'skoraj gotovo' sledi konvergenca v verjetnosti.*

TRDITEV 2.3. *(Neenakost Markova)*
Naj bo X slučajna spremenljivka s pričakovano vrednostjo in $a > 0$ pozitivna konstanta. Tedaj je:

$$P(|X| \geq a) \leq \frac{E[|X|]}{a}$$

DOKAZ. Naj bo $a > 0$. Pišemo $A = \{|X| \geq a\} = \{\omega| \quad |X(\omega)| \geq a\}$. Tedaj $|X| \geq a \cdot \mathcal{U}_A$. Sledi $E[|X|] \geq a \cdot P(A)$. ■

POSLEDICA 2.4. *(Neenakost Čebiševa)*
Naj bo X slučajna spremenljivka s (končno) disperzijo. Tedaj velja

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$$

za vsako pozitivno število ε.

DOKAZ.

$$P(|X - E[X]| \geq \varepsilon) = P((|X - E[X]|)^2 \leq \varepsilon^2) < \frac{E((X - E[X])^2)}{\varepsilon^2} = \frac{D(X)}{\varepsilon^2}$$

■

IZREK 2.5. *(Šibki zakon velikih števil)*
Naj bodo $X_1, X_2, \ldots \quad \Omega \to \mathbb{R}$ neodvisne in enako porazdeljene slučajne spremenljivke s pričakovano vrednostjo μ in (končnim) odklonom σ. Tedaj zaporedje "vzorčnih povprečij"

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

*konvergira **v verjetnosti** h konstanti μ.*

DOKAZ. Trdimo, da velja $\lim_{n\to\infty} P(|\frac{X_1+X_2+\ldots+X_n}{n} - \mu| \geq \varepsilon) = 0$ za vsak pozitiven $\varepsilon > 0$. Pišimo $\bar{X} = \frac{X_1+\ldots+X_n}{n}$.

$$P(|\bar{X}-\mu| > \varepsilon) \leq P(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{D(\bar{X})}{\varepsilon^2} = \frac{D(\frac{X_1+\ldots+X_n}{n})}{\varepsilon^2} = \frac{1}{n^2\varepsilon^2} D(X_1)+\ldots+D(X_n) = \frac{\sigma^2}{n\varepsilon^2}$$

Sledi, da rezultat konvergira proti 0, ko gre n v neskončnost. ■

OPOMBA 2.6. *Verjetnost kateregakoli konkretnega neskončnega zaporedja cifer in grbov je 0, ne glede na to, koliko je dejanska verjetnost posameznega meta $p \in (0, 1)$.*

OPOMBA 2.7. *(Česa šibki zakon velikih števil ne trdi.)*
*Denimo, da je $p = \frac{1}{2}$. Beležimo število cifer po n poskusih. **Ne velja**, da je število cifer po n poskusih večje od števila grbov 'približno polovici časa'.*
Zlahka je število cifer ves čas večje od števila grbov.

Izrek 2.8. *(Krepki zakon velikih števil)*
Naj bo $X_1, X_2, \ldots$ zaporedje neodvisnih in enako porazdeljenih slučajnih spremenljivk s končno pričakovano vrednostjo $E(X_i) \in \mathbb{R}$. Tedaj zaporedje "vzorčnih povprečij"

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

*konvergira k $E[X_i] =: \mu$ **skoraj gotovo**.*

Opomba 2.9. *Končna pričakovana vrednost pomeni $E[|X_i|] < \infty$*

Zgled 2.10. *Ponavljamo Bernoullijev poskus z verjetnostjo enice p. Tedaj skoraj gotovo velja:*

$$\lim_{n\to\infty} \frac{\text{št. enic v n poskusih}}{n} = p \tag{1}$$

To pomeni: verjetnost tistih neskončnih zaporedij $(\omega_1, \omega_2, \ldots)$ za katere (1) velja, je 1.

Opomba 2.11. *Krepki zakon velikih števil je uzakonitev frekvenistične definicije (intuicije) v verjetnosti.*

Opomba 2.12. *Iz izreka 2.8 sledi izrek 2.5*

2.1 Centralni limitni izrek

Izrek 2.13. *Naj bodo $X_1, X_2, \ldots$ neodvisno enako porazedeljene Bernoullijevke ($B(1,p)$). Tedaj zaporedje **standardiziranih povprečij***

$$\frac{\frac{X_1 + X_2 + \ldots + X_n}{n} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \frac{\sqrt{n}}{\sqrt{p(1-p)}} \left(\frac{X_1 + X_2 + \ldots + X_n}{n} - p \right)$$

konvergira k standardni normalni porazdelitvi v porazdelitvi.

Z drugimi besedami: Če velja $Y_n \sim Bin(n,p)$ sledi:

$$\frac{\sqrt{n}}{\sqrt{p(1-p)}} \left(\frac{Y_n}{n} - p \right) \xrightarrow[n\to\infty]{\text{v porazdelitvi}} \mathcal{N}(0,1)$$

Opomba 2.14. *Dokaz bomo izpustili.*
Za $p = \frac{1}{2}$ je dokazal leta 1733 De Moivre.
Za splošen p ga je dokazal Laplace.
Uporabljamo ga za aproksimacijo binomskih porazdelitev za velike n z normalnimi porazdelitvami.
Ohlapno lahko rečemo:

$$Bin(n,p) \sim \mathcal{N}(np, np(1-p))$$

za velike n-je.

Izrek 2.15. *(Centralni limitni izrek)*
Naj bodo $X_1, X_2, \ldots$ neodvisne, enako porazdeljene slučajne spremenljivke s končno disperzijo σ^2 in pričakovano vrednostjo μ. Tedaj zaporedje standardiziranih vzorčnih povprečij:

$$\frac{\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

konvergira v porazdelitvi k $\mathcal{N}(0,1)$.

Opomba 2.16. *V statistiki izrek* 2.15 *uporabljamo tipično v primerih, ko so $X_1, X_2, \ldots$ neodvisne replikacije preučevane slučajne spremenljivke X.*

Zgled 2.17. Ljubljanske mlekarne proizvajajo litrsko plastenko jogurta Mu 3, 2. 'Jamčijo', da ima taka plastenka 'v povprečju' 32g maščob. Privzamemo tudi, da Ljubljanske mlekarne 'jamčijo', da je odlklon vsebnosti maščob $1,5$g.

(1) Ali znamo izračunati (ali oceniti) $P(X \in (31g, 33g))$, če je X zvezna spremenljivka, ki predstavlja maso maščob v slučajno izbrani plastenki?

V splošnem ne znamo odgovoriti, saj ne poznamo porazdelitve.

(2) Naključno izberemo 100 takih plastenk in označimo X_i maso maščob v i-ti plastenki. Ali znamo izračunati (ali oceniti)?

Lahko ocenimo s pomočjo izreka 2.15. Praktične izkušnje kažejo, da je $n = 100$ že dovolj veliko

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{100} \Rightarrow P\left(\frac{\bar{X} - 32}{\frac{3}{2\sqrt{100}}} \in \left(\frac{31-32}{\frac{3}{2\sqrt{100}}}, \frac{33-32}{\frac{3}{2\sqrt{100}}}\right)\right) = \phi\left(\frac{20}{3}\right) - \phi\left(\frac{-20}{3}\right) = 1$$

(3) Kaj pa verjetnost $P(\bar{X} \in (31{,}9; 32{,}1))$?

$$P(\bar{X} \in (31{,}9; 32{,}1)) = \phi\left(\frac{2}{3}\right) \; - \; \phi\left(\frac{-2}{3}\right) = 0,7486 - 0,2514 = 0,4972$$

3 Deskriptivna statistika

Deskriptivna (opisna) statistika poskuša povzeti oziroma predstaviti značilnosti danega nabora podatkov, ki ga razumemo kot populacijo. Beseda 'statistika' v naslovu pomeni število, o katerem predpostavljamo značilnost, ki nas zanima. Formalneje je statistika funkcija, ki naboru podatkov priredi smiselno število, s katerim povzamemo določeno lastnost.

3.1 Kvantili

Opomba 3.1. *Te poznamo že od prej.*

3.2 Aritmetična sredina

Definicija 3.2. *Naj bodo $X_1, \ldots, X_N$ številske spremenljivke. Aritmetična sredina je:*

$$\frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{j=1}^{N} f_j \cdot x_j = \frac{f_1 \cdot x_1 + \ldots + f_N \cdot x_N}{f_1 + \ldots + f_N}$$

Opomba 3.3. *Zadnja enakost zgoraj je ravno $E[X]$, če na množici $\{1, 2, 3, \ldots, N\}$ vzamemo* ***enakomerno*** *verjetnost pri kateri je $P(X = x_j) = \frac{f_j}{N}$.*

3.3 Modus

Definicija 3.4. *Modus je vrednost z največjo frekvenco, če obstaja. Če je taka ena sama, govorimo o* ***unimodalni*** *porazdelitvi. (tipično za unimodalnost zahtevamo še kaj več)*

Opomba 3.5. *Modus ima bistveno večji pomen pri zveznih porazdelitvah oz. številskih spremenljivkah, pri katerih so načeloma možne vse vrednosti iz nekega intervala. Pri zveznih porazdelitvah bi za unimodalnost zahtevali en lokalni maksimum porazdelitvene gostote, pri splošnješih pa en geometrijsko definiran prevoj komulativne porazdelitvene funkcije F.*

3.4 Razmiki

Definicija 3.6. ***Variacijski razmik*** *je razlika med maksimalno in minimalno vrednostjo, pri katerih maksimalno razumemo kot zadnjo vrednost v ranžirni vrsti, minimalno pa kot prvo vrednost v ranžirni vrsti.*

$$X_{max} - X_{min} = X_(N) - X_(1)$$

Opomba 3.7. *Pomankljivost: občutljivost na ekstremne vrednosti.*

Interkvartilni razmik : $Q_{\frac{3}{4}} - Q_{\frac{1}{4}}$

Seminterkvartilni razmik : $\frac{Q_{\frac{3}{4}} - Q_{\frac{1}{4}}}{2}$

3.5 Odstopanje od srednjih vrednosti

Povprečni absolutni odklon od aritmetične sredine

$$\frac{1}{N}\sum_{i=1}^{N} |X_i - \bar{X}| = \frac{1}{N}\sum_{j=1}^{r} |f_j \cdot x_j - \bar{X}|$$

Povprečni absolutni odklon od mediane

$$\frac{1}{N}\sum_{i=1}^{N}|X_i - Me| = \frac{1}{N}\sum_{j=1}^{r}|f_j \cdot x_j - Me|$$

Povprečno kvadratno odstopanje od aritmetične sredine

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 = D(X) = Var(X)$$

Standardni odklon

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{1}{\sqrt{N}}||(X_1, \ldots, X_N) - (\bar{X}, \ldots, \bar{X})||$$

OPOMBA 3.8. *Evklidska razdalja med* $(X_1, \ldots, X_N)$ *in njegovo pravokotno projekcijo na premico* $\{t \cdot (1, 1, \ldots, 1) | t \in \mathbb{R}\}$

Kvadratni odklon je ugoden za računanje (tako praktično in teoretično), ker je ustrezna razdalja porojena z skalarnim produktom.

TRDITEV 3.9. *Naj bo* σ_1 *povprečni absolutni odklon in* $\sigma_2 = \sigma$ *standardni odklon. Potem velja:*

$$\sigma_1 \leq \sigma \leq \sqrt{N} \cdot \sigma_1$$

DOKAZ. Ocena $\sigma \leq \sqrt{N} \cdot \sigma_1$ sledi iz neenakosti $a_1^2 + \ldots + a_N^2 \leq (a_1 + \ldots + a_N)^2$ za pozitivna števila $a_i = |X_i - \bar{X}| \quad i = 1, 2, \ldots, N$
Ocena $\sigma_1 \leq \sigma$ je posledica Cauchy-Schwarzeve neenakosti. Naj bo $u = (a_1, \ldots, a_N) \quad a_i = |X_i - \bar{X}|$ in $v = (1, \ldots, 1) \in \mathbb{R}^N$. Iz neenakosti sledi: $| < u, v > | \leq ||u|| \cdot ||v||$. ■

3.6 Povezanost dveh številskih spremenljivk

DEFINICIJA 3.10. *Naj bosta* $X_1, X_2, \ldots, X_N$ *in* $Y_1, Y_2, \ldots, Y_N$ *številski spremenljivki, ki sta definirani na istem naboru podatkov oziroma populaciji. Kovarianca je:*

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) = \frac{1}{N}\sum_{i=1}^{N}x_i \cdot y_i - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}y_i\right)$$

Kot mera za jakost linearne povezanosti je kovarianca odvisna od variance posameznih spremenljivk. 'Pravo' (relativno) mero dobimo z normiranjem:

$$\varphi(X, Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sigma_X - \sigma_Y}$$

To je t.i. Pearsonov korelacijski koeficient. Iz Cauchy-Schwarzove neenakosti sledi $|\varphi(X, Y)| \leq 1$.

4 Sklepna statistika

Osnovno vprašanje statistike?
Preučujemo slučajno spremenljivko $X : \Omega \to \mathbb{R}$. Kakšna je njena porazdelitev?
Tipično nas v praksi zanimajo samo nekatere lastnosti porazdelitve, npr. pričakovana vrednost ali disperzija (za razliko od celotne komulativne funkcije $f_X : \mathbb{R} \to [0, 1]$).

Zgled 4.1. *Preučujemo učinek nekega statina na krvni holesterol (LDL). Razliko nivojev LDL holesterola p zdravljenjem in pred zdravljenjem prglasimo za slučajno spremenljivko, recimo X. Mera za učinek statina bo pričakovana vrednost.*

Zgled 4.2. *Preučujemo učinek kemoterapije za neko rakavo bolezen. Terapija je uspešna, če raven teles, ki so značilna za to bolezen, pade pod predpisani prag. Učinek meri Bernulijeva slučajna spremenljvika. katere vrednost je 1, če je terapija uspešna in 0, sicer.*

Prostor Ω je v praksi prevelik, predrag ali drugače nedostopen v celoti, zato želimo lastnost, ki nas zanima, oceniti s pomočjo 'vzorca', natnačneje s pomočjo več (zaporednih, neodvisnih) ponovitev slučajnega eksperimenta, ki je zakodiran v slučajni spremenljivki X. Kaj lahko ugotovimo iz vzorca, je odvisno od dejanske porazdelitve slučajne spremenljivke X (ki je ne poznamo) in od velikosti vzorca:

- Če je vzorec 'dovolj velik' si lahko pomagamo z limitnimi izreki.
- Če je vzorec majhen, moramo vsaj nekaj vedeti o porazelitvi slučajne spremenljivke X.

Zgled 4.3. *Imamo 100 pokritih števil. Naključno izberemo 10 števil.*

$$13, 13, 10, 11, 17, 25, 12, 19, 18, 10$$

Ugotovitev: Povprečje naključno izbranih števil je $14, 8$.
Ali lahko kaj smislenega povemo o povprečju teh 100 števil? NE.

'Nekaj vedeti' o porazdelitvi slučajne spremenljivke X v praksi pomeni **določiti primeren nabor možnih (dopustnih) porazdelitev** za X in 'izbrati' samo med njimi.

Zgled 4.4. *Privzamemo, da je* $X \sim \mathcal{N}(\mu, \sigma)$ *za neka neznan parametra* $\mu \in \mathbb{R}$ *in* $\sigma \in (0, \infty)$. *Tedaj je porazdelitev določena z dvema parametra. Vzamemo tisti par, ki najbolj (med vsemi) ustreza vzorcu podatkov.*

Zgled 4.5. *Privzamemo, da je* $X \sim B(1, p)$. *Tu je porazdelitev določena z enim samim parametrom p. Izberemo tistega, ki se najbolj sklada s karakternimi podatki.*

Nabor dopustnih porazdelitev za slučajno spremenljivko X pravimo **statistični model**, njihovi izbiri oziroma (tukaj nekaj manjka Blaž) pa **modeliranje**. Pri modeliranju delamo napake (včasih velike). Po izboru modela ima statistično sklepanje preciznost matematike.

Definicija 4.6. ***Slučajni vzorec*** *(prirejen s slučajno spremenljivko X)* ***velikosti n*** *je n-terica slučajnih spremenljivk $X_1, \ldots, X_n$, definiranih na prostoru vseh vzorcev velikosti n, kjer vrednosti slučajnih spremenljivk X_i na danem vzorcu dobimo tako, da X uporabimo na i-tem elementu vzorca:*

$$X_i(\text{vzorec velikosti n}) = X_i(\omega_1, \ldots, \omega_n) = X(\omega_i)$$

To je formalizacija odeje ponavljanja danega slučajnega eksponenta X. Pri vzorčenju s ponavljanjem (rečemo tudi neodvisno vzorčenje) so komponente X_i tudi **neodvisne**. V tem primeru je prostor vzorcev velikosti n kar kartezični produkt.

5 Intervali zaupanja

Recimo, da ocenjujemo $E(X)$ na podlagi vzorca. Intervalska ocena je metoda oziroma napoved oblike: *Predvidevamo, da je $E(X) \in [L(\text{vzorec}), U(\text{vzorec})]$, kjer sta L in U odvisna od vzorca.*

V splošnem lahko zapišemo takole:

Definicija 5.1. *Recimo, da je c karakteristika (parameter) slučajne spremenljivke, ki nas zanima (na primer pričakovana vrednost, disperzija, ...) Tedaj je intervalska ocena za c metoda, ki vzorcu priredi napoved (predvidevanje):*

$$c \in [L(\text{vzorec}), U(\text{vzorec})]$$

Definicija 5.2. *Naj bo* ***c*** *karakteristika, ki nas zanima in* ***n*** *velikost vzorca. Interval zaupanja za c* ***stopnje zaupanja*** *β za vzorec velikosti nsestoji iz funkcij, odvisnih od (dejanskega) vzorca velikosti n,imenujemo ju L in U, za kateri je verjetnost tistih vzorcev, za katere interval [L(vzorec),U(vzorec)] vsebuje c, večja ali enaka β.*

$$P(L(X1, X2, ..., Xn) \leq c \leq U(X1, X2, ..., Xn)) \geq \beta$$

5.1 Clopper-Pearsonov eksaktni interval zaupanja

Definicija 5.3. *Naj bo n velikost vzorca in beta $\in (0, 1)$ stopnja zaupanja. Pišimo $\alpha = 1 - \beta$. Naj k označuje število pozitivnih vzorcev.*

$$L(k) = \begin{cases} Beta(k, n-k+1)_{-\alpha/2} & k \geq 1 \\ 0, & k = 0. \end{cases}$$

$$U(k) = \begin{cases} Beta(k+1, n-k)_{\alpha/2} & k \leq n \\ 1 & k = n \end{cases}$$

Tu sta $Beta(a, b)_{\alpha/2}$ oziroma $Beta(a, b)_{-\alpha/2}$ zgornji oziroma spodnji $\alpha/2$-precentil porazdelitve $Beta(a, b)$.

5.1.1 Družina porazdelitev beta

DEFINICIJA 5.4. *Naj bosta a in b pozitivni realni števili.* ***Porazdelitev beta s parametroma a in b*** *je zvezna porazdelitev, ki jo označimo Beta(a,b) s porazdelitevno gostoto:*

$$f(x) = f(x; a, b) = \begin{cases} \frac{1}{Beta(a,b)} \cdot x^{a-1}(1-x)^{b-1} & x \in (0,1) \\ 0 & x \notin (0,1) \end{cases}$$

OPOMBA 5.5. *Velja*

$$Beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}dx$$

5.2 Studentove porazdelitve

IZREK 5.6. *Naj bodo $X_1, X_2, \ldots, X_n$ neodvisne repliakcije normalnega slučajnega eksperimenta $X \sim \mathcal{N}(\mu, \sigma)$. Tedaj sta slučajni spremenljivki $\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ in $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$* ***neodvsini*** *(v verjetnostnem smislu).*

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

To pomeni, da lahko $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ zapišemno kot

$$t_{n-1} = \frac{Z}{\sqrt{\frac{H}{n-1}}}$$

pri čemer je $Z \sim \mathcal{N}(0,1)$ in $H \sim \chi^2_{n-1}$

V taki sitaciji pravimo da ima t_{n-1} ***Studnetovo porazdelitev z n-1 prostostnimi stopnjami****.*

5.3 Ocenjevanje pričakovane vrednosti za normalno porazdeljene slučajne spremenljivke

Slučajni vzorec velikosti n je sestavljen iz n-terice $X_1, \cdots X_n$ neodvisnih slučajnih spremenljivk, ki so vse porazdeljene enako kot preučevana spremenljiva X.

Privzemimo, da je $X \sim \mathcal{N}(\mu, \sigma)$. To v praksi pogosto privzamemo za zvezne slučajne spremenljivke, ki so dovolj simetrične in dovolj 'lepe' koncentrirane okrog μ. Spoimnimo se, da za standardno cenilko μ velja $\bar{X} = \frac{X_1 + \cdots X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

5.3.1 Z znano disperzijo

Privzamimo, da standardni odklon poznamo. (Sicer zelo nepraktično).

Izberimo tak pra realnih števil $a < b$, da zanj velja:

$$P(\mathcal{N}(0,1) \in [a,b]) = \beta$$

kjer je $\beta \in (0,1)$ (blizu 1) vnaprej predpisana stopnja zaupanja.

Ker je $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ velja:

$$P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \in [a,b]\right) = \beta$$

$$P\left(a \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) = \beta$$

$$P(-a \cdot \frac{\sigma}{\sqrt{n}} \geq \mu - \bar{X} \geq -b \cdot \frac{\sigma}{\sqrt{n}}) =$$

$$P(\bar{X} - b \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - a \cdot \frac{\sigma}{\sqrt{n}}) = \beta$$

Verjetnost tistih vzorcev velikosti n, za katere μ pripada intervalu $[\bar{X} - b \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} - a \cdot \frac{\sigma}{\sqrt{n}}]$, je enak β. **To pomeni, da je slučajni interval** $[\bar{X} - b \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} - a \cdot \frac{\sigma}{\sqrt{n}}]$ **intreval zaupanja za** μ **stopnje zaupanja** β**.**

V praksi imamo najraje *najinformativnejši* interval zaupanja. Izkaže se, da je razlika $b - a$ (pri pogoju $\phi(b) - \phi(a) = \beta$) najmanjša, če je

$$b = -a = \phi^{-1}\left(\frac{1+\beta}{2}\right)$$

5.3.2 Z neznano disperzijo

Privzamimo (kar je bolj praktično), da odklona σ ne poznamo.

σ lahko ocenimo s vzorčno disperzijo.

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

ki jo imenujemo tudi standardna nepristranska cenilka za disperzijo.

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

Podobno kot prej bi želeli slučajno spremenljivko transformirati na standardno normalno..

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

Izkaže se, da je porazdelitev slučajne spremenljivke $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ **ni odvisna** niti od μ niti od σ.

DEFINICIJA 5.7. *Tako porazdelitev imenujemo* ***Studentova t- porazdelitev z*** $n-1$ ***prostostnimi stopnjami***

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Privzeli bomo, da so vse porazdelitve t_{n-1} *simetrične.*

Izberimo tak par realnih števil $a < b$, da je zanj velja:

$$P(t_{n-1} \in [a,b]) = \beta$$

$$P(\bar{X} - b \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} - a \cdot \frac{S}{\sqrt{n}}) = \beta$$

V praksi imamo najraje *najinformativnejši* interval zaupanja. Izkaže se, da je razlika $b-a$ (pri pogoju $\phi(b) - \phi(a) = \beta$) najmanjša, če je:

$$b = -a = t_{n-1,(1-\beta)/2} = F^{-1}_{t_{n-1}}\left(\frac{1+\beta}{2}\right)$$

5.4 Ocenjevanje disperzije normalno porazdeljene slučajne porazdelitve

DEFINICIJA 5.8. *Pravimo, da ima Y porazdelitev* ***gama*** *s parametroma* α *in* β*, če ima Y gostoto:*

$$f_y(x) = f_{\Gamma(\alpha,\beta)}(x) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x} \quad x > 0$$

POSLEDICA 5.9. *Naj bo* $Y \sim \Gamma(\alpha,\beta)$*. Velja:*

$$E[Y] = \frac{\alpha}{\beta^2}$$

DEFINICIJA 5.10. *Poddružina gama porazdelitve so tako imenovane* ***Hi-kvadrat*** *porazdelitve. Za definicijo vzamemo tole:*

$$\chi^2_k = \Gamma(\frac{k}{2}, \frac{1}{2})$$

s k prostostnimi stopnjami.

TRDITEV 5.11. *Naj bodo $Z_1, Z_2, \dots Z_n$ neodvisno porazdeljene standardno normalne slučajne spremenljivke. Tedaj ima slučajna spremenljivka*

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

TRDITEV 5.12. *Naj bodo $X_1, X_2, \dots, X_n$ neodvisno porazdeljene standardno normalne slučajne spremenljivke. Tedaj velja:*

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

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

DOKAZ. Dokaz trditve napravimo s pomočjo 'konvolucijskih' formul za porazdelitev vsote. Velja:
Če $Y_1 \sim \Gamma(\alpha_1, \beta)$ in $Y_2 \sim \Gamma(\alpha_2, \beta)$ in sta Y_1 in Y_2 neodvisna, je vsota

$$Y_1 + Y_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$$

Posledično: $\chi_k^2 + \chi_l^2 \sim X_{k+l}^2$ (χ_k^2 in χ_l^2 sta neodvisni) ■

Prva porazdelitev sledi neposredno iz prejšne trditve, pri drugi pa odštejemo $\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n$ povzroči "vez"v sistemu z n prostostnimi stopnjami, ker zmanjša število prostostnih stopenj za 1.

OPOMBA 5.13. *'Sistem' s n prostosntnimi stopnjami ima n količin, ki se lahko spreminjajo neodvisno (prosto) druga od druge. Če je vsota $X_1 + X_2 + \ldots + X_n$ določena, lahko prosto spreminjamo npr. $X_1, X_2, \dots, X_{n-1}$, X_n pa je določena z "vezjo".*

5.4.1 Z znano pričakovano vrednostjo μ

Privzemimo, da so $X_1, X_2, \dots, X_n$ neodvisne replikacije preučevanega slučajnega eksperimenta. $X \sim \mathcal{N}(\mu, \sigma)$. Velja:

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

Izberimo realni števili a in b, za katere velja $P(\chi_n^2 \in [a, b]) = \beta$, pri čemer je β stopnja zaupanja.

$$P(a \leq \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \leq b) = \beta \Rightarrow$$

$$P\left(\frac{1}{b} \sum_{i=1}^{n} (X_i - \mu)^2 \leq \sigma^2 \leq \frac{1}{a} \sum_{i=1}^{n} (X_i - \mu)^2 \right) = \beta$$

Za a lahko vzamemo: $a = \chi^2_{n,\frac{1-\beta}{2}} = F^{-1}_{\chi^2_n}\left(\frac{1-\beta}{2}\right)$

Za b lahko vzamemo: $b = \chi^2_{n,\frac{1+\beta}{2}} = F^{-1}_{\chi^2_n}\left(\frac{1+\beta}{2}\right)$

Tako dobimo enakorepi interval zaupanja za σ^2 pri znanem μ:

$$\left[\frac{1}{F^{-1}_{\chi^2_n}\left(\frac{1+\beta}{2}\right)}\sum_{i=1}^{n}(X_i-\mu)^2, \frac{1}{F^{-1}_{\chi^2_n}\left(\frac{1-\beta}{2}\right)}\sum_{i=1}^{n}(X_i-\mu)^2\right]$$

Opomba 5.14. *Ta interval zaupanja ni najkrajši (v pričakovani vrednosti dolžine intervala). Najkrajši interval lahko dobimo s numeričnim računananjem zahtevnih integralskih enačb. V praski še dandanes uporabljamo enakorepei interval.*

5.4.2 Z neznano pričakovano vrednostjo μ

Privzemimo, da so $X_1, X_2, \ldots, X_n$ neodvisne replikacije preučevanega slučajnega eksperimenta. $X \sim \mathcal{N}(\mu, \sigma)$. Velja:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1}$$

Izberemo realni števili a in b za katere velja $P(\chi^2_{n-1} \in [a,b]) = \beta$, pri čemer je β stopnja zaupanja.

$$P\left(\frac{1}{b}\sum_{i=1}^{n}(X_i-\bar{X})^2 \le \sigma^2 \le \frac{1}{a}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \beta$$

Tedaj dobimo **enakorepi interval**:

$$\left[\frac{1}{F^{-1}_{\chi^2_n}\left(\frac{1+\beta}{2}\right)}\sum_{i=1}^{n}(X_i-\bar{X})^2, \frac{1}{F^{-1}_{\chi^2_n}\left(\frac{1-\beta}{2}\right)}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$$

Opomba 5.15. *Ta interval ni najkrajši. V praksi vzamemo pričakovano vrednost tega intervala. ($\bar{X}$ je tukaj slučajna spremenljivka).*

Opomba 5.16.

$$0 \le L(vzorec) \le \sigma^2 \le U(vzrorec) \iff \sqrt{L(vzorec)} \le \sigma^2 \le \sqrt{U(vzorec)}$$

To pomeni, da interval zaupanja za σ enake stopnje zaupanja (β) dobimo tako, da korenimo intervalski meji za σ^2.

6 Preizkuševanje domnev

6.1 Uvod

Spomnimo se na 'problem' zaposlenega v podpori, ki nudi storitve strankam. Predpostavljamo, da slučajno izberemo storitev opravi zadovoljivo z verjetnostjo p. Tej verjetnosti pravimo 'notranji' oziroma dejanski p. Direktor se odloča za povišico. Pravi, da je povišica smiselna, če je p zaposlenega večji od 0.7.

Opomba 6.1. *To bi lahko rešili z intervali zaupanja. Na primer: zahtevali bi, da je spodnja meja intervala zaupanja enaka 0.7.*

Opomba 6.2. *Odločitev ali podpremo ali zavrnemo domnevo $p > 0.7$,bomo napravili na podlagi vzorca storitev.*

Jasno je, da imamo samo dve opciji:

(1) podpremo $p > 0.7$

(2) ne podpremo $p > 0.7$ (podpremo $p \leq 0.7$)

Preizkus domneve (oz. test hipoteze) $p \leq 0.7$ nasporti domnevi $p > 0.7$ temelji na **odločitvenem pravilu**, ki pove, kako se bomo odločili glede na rezultat n neodvisnih ponovitev poskusa. V našem primeru bomo izbrali n storitev, k od njih bo ocenjenih zadovoljivo. Narediti moramo odločitevno pravilo. Recimo, da je $n = 20$.
Vidimo, da je smiselno, da je naše odločitveno pravilo naslednje oblike:

$$\begin{cases} \text{podpremo} & p > 0.7 & \text{če} & k > C \\ \text{ne podpremo} & p \leq 0.7 & \text{če} & k \leq C \end{cases}$$

kjer je C primerno število med 0 in n.

Pri odločitvi bomo naredili napako. Kakšne so možne napake?

Dejansko stanje Odločitev	$p \leq 0.7$	$p > 0.7$
Podpremo ($p \leq 0.7$)	OK	Napaka
Ne podpremo ($p > 0.7$)	Napaka	OK

Izračunajmo verjetnost, da podpremo $p > 0.7$, pri pogoju, da je v resnici $p \leq 0.7$.

$$P(k > C|\text{dejanski} \quad p \leq 0.7)$$

Vrednost napake ne število, ampak funkcija dejanskega p, za katerega privzamemo, da je $p \leq 0.7$. Če želimo, da je ta verjetnost manjša ali enaka 0.1 za vse $p \leq 0.7$ pri $n = 20$ bi izbrali $C = 17$. Se pravi, odločitev $p > 0.7$ bomo podprli le za $k = 18, 19, 20$.

Kaj pa 'druga napaka' (v prvi vrstici)?
To je napaka, ki jo storimo, če podpremo $p \leq 0.7$, kljub temu, da je $p > 0.7$. Njena verjetnost je

$$P(k \leq C|\text{dejanski} \quad p > 0.7) = \sum_{k=0}^{C} \binom{n}{k} p^k (1-p)^{n-k}$$

To je funkcija dejanskega p, za katerega privzamemo $p > 0.7$.

6.2 Preizkušanje domnev v splošnem

Definicija 6.3. ***Domneva*** *ali hipoteza je izjava o porazdelitvi preučevane slučajne spremenljivke. (V uvodnem primeru $X \sim Bin(1, p)$; p je lasnost porazdelitve)*

Nasporti ji stoji tako imenovana ***alternativna domneva****, ki jo v osnovnem primeru negacija prvotne domneve.*

Preizkus ali test domneve je odločitveno pravilo, po katerem glede na vse možne rezultate n neodvisnih ponovitev eksperimenta X odločimo ali potrdimo hipotezo H ali njeno alternativo A.

Pri odločitvi lahko naredimo napako. Ustrezni diagram:

Dejansko stanje Odločitev	Drži H	Drži A
Podpremo H	OK	Napaka 2.vrste
Podpremo A	Napaka 1.vrste	OK

Obeh napak hkrati ne moremo minimizirati. V najboljšem primeru sta si napaki komplementarni. Eno od njiju si izberemo kot pomembno in jo zmanjšujemo pod vnaprej predpisano raven $\alpha \in (0, 1)$, ki ji pravimo **stopnja zančilnosti**.

Po potrebi vlogi H in A zamenjamo, tako da je napaka katere verjetnost $< \alpha$ tista, pri kateri podpremo A, čeprav drži H. Tej napaki pravimo **napak prve vrste**. Napaki, ki jo storimo, če pa podpremo H, čeprav drži A pa pravimo **napaka druge vrste**.

Statistični test (preizkus) domneve H proti domnevi A je odločitveno pravilo oblike:

$$\begin{cases} \text{podpremo A (zavrnemo H)} & \text{če izpolnjuje neki pogoj glede na vzorec } X_1, \ldots X_n \\ \text{podpremo H (ne zavrnemo H)} & \text{sicer} \end{cases}$$

Paziti moramo, saj gre v naprej določeno velikost vzorca n.

V zgledu z zaposlenim v podpori smo imeli $X \sim Bin(1,p)$ ter domnevi $H : p \leq 0.7$ in $A : p > 0.7$. Test je oblike:

$$\begin{cases} \text{zavrnemo H} & \text{če je } X_1 + X_2 + \ldots + X_n > C \\ \text{ne zavrnemo H} & \text{če je } X_1 + X_2 + \ldots + X_n \leq C \end{cases}$$

kjer so $X_1, X_2, \ldots X_n$ neodvisne replikacije slučajne spremenljivke X.

Slučajni spremenljivki $T = X_1 + X_2 + \ldots + X_n$ iz zgleda pravimo **testna statistika**, kjer se o tem, katero od obeh hipotez v testu podpremo, odločimo na podlagi vrednosti slučajne spremenljivke T.

TRDITEV 6.4. *Naj bo $T : \mathbb{R}^n \to \mathbb{R}$ (ali splošneje $\mathbb{R}^n \to \mathbb{R}^m$) primerna funkcija. Tedaj je T testna statistika za test dane domneve H proti alternativi A, če se v testu odločimo (izključno) glede na vrednosti $T(X_1, \ldots, X_n)$.*

Spomnimo se, da je napak prve vrste napaka, ki jo storimo, če zavzamemo H (podpremo A) kljub temu da H drži. Test ima značilnost α, če je verjetnost napake prve vrste navzgor omejena z α.
Napaka druge vrste je napaka, ki jo storimo če zavzamemo H (podpremo H), kljub temu da H ne drži (drži A).

OPOMBA 6.5. *Napako prve vrste lahko naredimo le, če drži H.*

text in

OPOMBA 6.6. *Napako druge vrste lahko naredimo le, če drži A.*

Če ima naš test zančilnosti α (kjer je α blizu 0), potem udobno zavrnemo domnevo H (če so izpolnjeni pogoji), ker je verjetnost da se to po nesreči zgodi, majhna.

Ne rečemo pa radi, da H podpremo (če so izpolnjene predpostavke), ker je verjentost napake druge vrste lahko velika.

6.3 Preizkušanje v zvezi z Bernoullijevim p

Enostranski test ničelne hipoteze $H_0 : p \leq p_0$

Testna statistika za vzorce velikosti n:

$$T = X_1 + X_2 + \ldots + X_n$$

Testiranje $H_0 : p \leq p_0$ proti $p > p_0$ stopnje zančilnosti α

$$\begin{cases} H_0 \text{ zavrnemo} & \text{če } T > C \\ H_0 \text{ ne zavrnemo} & \text{če } T \leq C \end{cases}$$

kjer je C **najmanjše** tako celo število, za katero je

$$P(B(n,p_0) > C) = 1 - P(B(n,p_0) \leq C) \leq \alpha$$

ZGLED 6.7. *Tokrat direktor pravi, da bo zaposlene s* $p < 0.7$ *odpustil. Napaka, da po nesreči podpremo* $p < 0.7$ *(zavrnemo* $p \geq 0.7$*), je kritična in jo želimo omejiti navzgor z* $\alpha = 0.1$*.*

Lahko gledamo C *od 1 do 14 (ne do 20)*
$C = 11$*, zavrnemo, če* $T < 11$

Zanimiva je primerjava med T *in* $E[T|p = p_0]$
Efekt pravičnega vzorca:

$$\lim_{n\to\infty} \frac{C_n}{E[T|p = p_0]} = 1$$

(zakon o velikih številih)

Enostranski test ničelne hipoteze $H_0 : p \geq p_0$

Testna statistika za vzorce velikosti n:

$$T = X_1 + X_2 + \ldots + X_n$$

Testiranje $H_0 : p \geq p_0$ proti $p < p_0$ stopnje zančilnosti α

$$\begin{cases} H_0 \text{ zavrnemo} & \text{če } T < C \\ H_0 \text{ ne zavrnemo} & \text{če } T \geq C \end{cases}$$

kjer je C **največje** tako celo število, za katero je

$$P(B(n, p_0) < C) = P(B(n, p_0) \leq C - 1) \leq \alpha$$

ZGLED 6.8. *Test $H_0 : p = p_0$ proti $A : p \neq p_0$*

$$\begin{cases} H_0 \text{ zavrnemo} & \text{če } T < C_1 \quad \text{ali} \quad T > C_2 \\ H_0 \text{ ne zavrnemo} & \text{sicer} \end{cases}$$

Določanje konstant C_1 in C_2 je algoritmično zahtevno. Izkaže se, da je za primerna C_1 in C_2 zgornji test ekvivalenten test oblike:

Dvostranski test ničelne hipoteze $H_0 : p = p_0$

Testna statistika za vzorce velikosti n:

$$T = X_1 + X_2 + \ldots + X_n$$

Testiranje $H_0 : p = p_0$ proti $p \neq p_0$ stopnje zančilnosti α
Privzamimo, da imamo interval zaupanja za p za vzorec velikosti n
stopnje zaupanja $\geq 1 - \alpha$.

$$\begin{cases} H_0 \text{ zavrnemo} & \text{če } p_0 \text{ ne pripada intervalu zaupanja}(T) \\ H_0 \text{ ne zavrnemo} & \text{če } p_0 \text{ pripada intervalu zaupanja}(T) \end{cases}$$

Vzamemo lahko Clopper-Pearsonov eksaktni interval:

$$L(k) = \begin{cases} Beta(k, n-k+1)_{-\alpha/2} & k \geq 1 \\ 0, & k = 0. \end{cases}$$

$$U(k) = \begin{cases} Beta(k+1, n-k)_{\alpha/2} & k \leq n \\ 1 & k = n \end{cases}$$

Tu sta $Beta(a,b)_{\alpha/2}$ oziroma $Beta(a,b)_{-\alpha/2}$ zgornji oziroma spodnji $\alpha/2$-precentil porazdelitve $Beta(a,b)$.

TRDITEV 6.9. *Če ima interval zaupanja za p stopnjo zaupanja $1 - \alpha$ ima zgornji test stopnjo značilnosti α.*

6.4 Preizkušanje E(X) v normalnih populacijah

6.4.1 Sigma poznan