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.