

Verjetnost in Statistika - odgovori na odprta vprašanja

Jernej Vivod

September 22, 2018

1 Kombinatorika in verjetnost

1. Definiraj pojem permutacije n elementov ($n!$ različnih jih je) ter pojem permutacije s ponavljanjem (kako so povezane s kombinacijami).

rešitev:

Permutacija je z medsebojnimi zamenjavami preurejeno zaporedje znanega končnega števila elementov (pri tem pa število elementov ostane enako). V matematiki je to bijekcija množice elementov samega v sebe, ki se jo zapiše kot:

$$\pi : X_n \rightarrow X_n$$

2. Definiraj Γ funkcijo in podaj čimveč njenih vrednosti (vsaj eno neceloštevilsko).

rešitev: Gama porazdelitev je dvo-parameterska družina zveznih verjetnostnih porazdelitev in je posplošitev eksponentne porazdelitve. Exponentna, Erlangova in hi-kvadrat porazdelitev so vse posebni primeri gama porazdelitve.

Naj bosta $b, c > 0$. Tedaj ima Gama porazdelitev, oznaka $\Gamma(b, c)$, gostoto:

$$p(x) = \begin{cases} \frac{c^b}{\Gamma(b)} x^{b-1} e^{-cx} & x > 0 \\ 0 & \text{sicer} \end{cases} \quad (1)$$

kjer je b parameter oblike, c pa parameter raztega. Na sliki je $k = b$ in $\theta = \frac{1}{c}$

3. Definiraj nezdružljiva dogodka ter popoln sistem dogodkov in pojasni, kako lahko slednjega uporabimo za klasično definicijo verjetnosti.

rešitev:

Nezdružljiva dogodka sta dogodka, ki se ne morate zgoditi hkrati. Torej je verjetnost njunega produkta vedno enaka 0. Popoln sistem dogodkov je množica dogodkov, ki so med sabo nezdružljivi in se pri vsaki ponovitvi poskusa zgodi natanko eden izmed njih.

Naj je verjetnostni poskus, ki ima n med seboj enakovrednih izidov (enakovrednost izidov pomeni, da se vsi izidi pojavijo približno enako pogosto, če se poskus ponovi večkrat). Opazuje se dogodek A , za katerega je ugodnih m izidov. Po klasični definiciji je verjetnost dogodka A razmerje med številom ugodnih izidov in številom vseh možnih izidov. Med seboj enakovredni izidi, ki tvorijo dogodke, predstavljajo popoln sistem dogodkov.

4. Definiraj vsoto in produkt dogodkov ter utemelji zvezo med verjetnostmi dveh dogodkov, njuno vsoto ter njunim produktom.

rešitev:

Vsota dogodkov je dogodek, ki se zgodi, če se zgodi vsaj eden od dogodkov v vsoti. Produkt dogodkov je dogodek, ki se zgodi, če se zgodijo vsi dogodki, ki predstavljajo produkt.

Velja:

$$P(A + B) = P(A) + P(B) - P(AB)$$

5. Opiši polinomsko porazdelitev (ali spada med večrazsežne diskretne ali zvezne porazdelitve) ali pa definiraj mediano za slučajno spremenljivko X (to ni mediana zaporedja!).

rešitev: Polinomska porazdelitev $P(n; p_1, \dots, p_r)$, $\sum p_i = 1$, $\sum k_i = n$ je določena s predpisom (Spremenljivke X_i opisujejo število pojavitev rezultata i) :

$$P(X_1 = k_1, \dots, X_r = k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r}$$

In je posplošitev binomske porazdelitve. Število možnih izidov katerega koli poskusa je lahko večje kot 2.

Primer: poznamo porazdelitve krvnih skupin. Želimo izračunati verjetnost, da jih v vzorcu desetih ljudi ima 6 krvni tip O, 2 krvni tip A, 1 krvni tip B in 1 krvni tip AB.

Koeficient šteje permutacije s ponavljanjem. Za $r = 1$ dobimo binomsko porazdelitev, tj. $B(n, p) = P(n; p, q)$.

Mediana je vrednost, od katere je polovica vrednosti v populaciji manjših, polovica pa večjih. primer: medianska plača...

2 Verjetnost in računanje z dogodki

Denimo, da je verjetnost, da študent reši pravilno prvo vprašanje štirikrat manjša od verjetnosti, da pravilno reši drugo vprašanje in za eno polovico manjša od verjetnosti, da pravilno reši vsaj eno od obeh vprašanj. , Na naši fakulteti so si študentje izbirne predmete izbrali takole: 55 študentov statistiko, 80 študentov kriptografijo, 75 študentov verjetnost, 25 študentov kriptografijo in verjetnost, 20 študentov statistiko in verjetnost, 10 statistiko in kriptografijo, 5 študentov vse tri predmete.

1. Definiraj nezdružljiva dogodka ter popoln sistem dogodkov in pojasni kako lahko slednjega uporabimo za klasično definicijo verjetnosti.

rešitev: Nezdružljiva dogodka sta dogodka, ki se ne morate zgoditi hkrati. Torej je verjetnost njunega produkta vedno enaka 0. Popoln sistem dogodkov je množica dogodkov, ki so med sabo nezdružljivi in se pri vsaki ponovitvi poskusa zgodi natanko eden izmed njih.

Naj je verjetnostni poskus, ki ima n med seboj enakovrednih izidov (enakovrednost izidov pomeni, da se vsi izidi pojavijo približno enako pogosto, če se poskus ponovi večkrat). Opazuje se dogodek A, za katerega je ugodnih m izidov. Po klasični definiciji je verjetnost dogodka A razmerje med številom ugodnih izidov in številom vseh možnih izidov. Med seboj enakovredni izidi, ki tvorijo dogodke, predstavljajo popoln sistem dogodkov.

2. Definiraj vsoto in produkt dogodkov ter utemelji zvezo med verjetnostmi dveh dogodkov, njuno vsoto ter njunim produktom.

rešitev: Vsota dogodkov je dogodek, ki se zgodi, če se zgodi vsaj eden od dogodkov v vsoti. Produkt dogodkov je dogodek, ki se zgodi, če se zgodijo vsi dogodki, ki predstavljajo produkt.

Velja:

$$P(A + B) = P(A) + P(B) - P(AB)$$

3. Določi verjetnost, da je študent pravilno rešil obe vprašanji, če veš, da je ta verjetnost, za tretino manjša od razlike verjetnosti, da je pravilno rešil drugo vprašanje in da je pravilno rešil prvo vprašanje.

rešitev:

Velja $4P(A) = P(B)$ oziroma za $P(A) = x$ je $P(B) = 4x$ in $P(A) + \frac{1}{2} = P(A+B)$ oziroma $P(A+B) = x + \frac{1}{2}$ ter $P(AB) + \frac{1}{3} = P(B) - P(A)$ oziroma $P(AB) = 3x - \frac{1}{3}$. Končna zveza $P(A) + P(B) = P(A+B) + P(AB)$ preide v $x + 4x = x + \frac{1}{2} + (3x - \frac{1}{3})$ oziroma $x = \frac{1}{6}$

4. Koliko študentov je v tem letniku?

rešitev: TODO

5. Koliko študentov je izbralo dva predmeta od vseh treh naštetih?

rešitev: TODO

6. Koliko študentov je izbralo samo en predmet od vseh treh naštetih?

rešitev: TODO

7. Koliko študentov je izbralo kriptografijo in ne verjetnosti?

rešitev: TODO

8. Izračunaj verjetnost dogodka, da je slučajno izbran študent izbral statistiko.

rešitev: TODO

9. Izračunaj verjetnost dogodka, da je slučajno izbran študent izbral verjetnost, pri pogoju, da je izbral statistiko.

rešitev: TODO

10. Napiši pravilo o vključitvi in izključitvi ter na osnovi tega posploši zvezo iz zgornjega vprašanja na štiri dogodke.

rešitev: TODO

3 Pogojna verjetnost

1. Definiraj oziroma pojasni, kaj pomeni $P(A|B)$ in kako jo izračunamo.

rešitev: $P(A|B)$ pomeni pogojno verjetnost, da se zgodi dogodek A, če vemo, da se je zgodil dogodek B.

2. Verjetnost, da študent opravi izpit, če se je pripravljal, je 90%. Verjetnost, da se je naključno izbran študent pripravljal in opravil izpit, je 50%. Kolikšna je verjetnost, da se je naključno izbran študent pripravljal na izpit?

rešitev:

$$P(O|P) = 0.9$$

$$P(OP) = 0.5$$

Iz tega sledi:

$$P(P) = \frac{P(OP)}{P(O|P)} = \frac{0.5}{0.9} = 0.5556$$

3. Kako izračunamo pogojno verjetnost dveh neodvisnih dogodkov?

rešitev: Pogojna verjetnost dveh neodvisnih dogodkov je kar verjetnost dogodkov samih. Za neodvisna dogodka A in B velja $P(A|B) = P(A)$ in $P(B|A) = P(B)$.

4. Utemelji zvezo iz točke (a) bodisi s klasično ali statistično definicijo verjetnosti.

rešitev: Če vemo, da se je zgodil dogodek A , ki je sestavljen iz nekega števila dogodkov iz popolnega sistema dogodkov, lahko to vpliva na verjetnost, da se bo zgodil dogodek B , ki je npr. sestavljen iz n dogodkov iz popolnega sistema dogodkov in je presek teh dogodkov z dogodki, ki sestavljajo B neprazen.

5. Napiši in utemelji Bayesovo formulo.

rešitev: Odgovorjeno v enem od sledečih razdelkov.

4 Dvofazni poskusi, formula za popolno verjetnost in Bayesov obrazec

1. Definiraj pogojno verjetnost in podaj formulo za njen izračun.


rešitev: Odgovorjeno v enem od sledečih razdelkov

2. Definiraj popoln sistem dogodkov ter podaj formulo za popolno verjetnost.

rešitev:

4. POPOLNA VERJETNOST IN BAYESOV OBRAZEC

Naj tvorijo dogodki H_1, H_2, \dots, H_n popoln sistem dogodkov [$H_i \cap H_j = \emptyset$, $i \neq j$ in $\bigcup_{i=1}^n H_i = G$ (vsi možni izidi)].



Slika 4.1: Dogodek A v popolnem sistemu dogodkov prostora G

Za vsak dogodek A velja t. l. **obrazec za popolno verjetnost** dogodka A :

$$P(A) = \sum_{i=1}^n P(H_i)P(A/H_i)$$

S pomočjo enakosti $P(A \cap H_i) = P(A) P(H_i/A) = P(H_i) P(A/H_i)$ dobimo z razreševanjem enačbe na $P(H_i/A)$ t. l. **Bayesov obrazec**

$$P(H_i/A) = \frac{P(H_i) P(A/H_i)}{P(A)} = \frac{P(H_i) P(A/H_i)}{\sum_{k=1}^n P(H_k) P(A/H_k)}$$

3. Kako lahko izračunamo verjetnost, da je dogodek A nastopil skupaj z določenim dogodkom (hipotezo) H_i iz prve faze?

rešitev:

$$P(AH_i) = P(A|H_i) \cdot P(H_i)$$

4. Oglejmo si preprost primer poligrafskih testov. Študije so pokazale, da v primeru, ko testirani pri preiskusi laže, poligraf le-to (laž) zazna v 88% primerov; v primeru, ko testirani govorijo resnico, pa je pri odkrivanju resnice uspešen v 86% primerov. Denimo, da 99% vseh testirancev na testu govorijo resnico. Vzemimo neko osebo, ki ji je poligraf pripisal laž. Kolikšna je verjetnost, da dejansko laže?

rešitev: To je tipičen primer uporabe Bayesovega obrazca... Glej rešeno podobno nalogo o testu za bolezen.

5. Izpelj Bayesov obrazec.

rešitev:

Trditev 3.4. (**Bayesov obrazec**) Za popoln sistem dogodkov H_i , $i \in I$ in poljuben dogodek A velja

$$P(H_k/A) = \frac{P(H_k) \cdot P(A/H_k)}{\sum_{i \in I} P(H_i) \cdot P(A/H_i)}. \quad \square$$

5 Bernoullijevo zaporedje neodvisnih poskusov

1. Definiraj zaporedje neodvisnih poskusov.

rešitev: Zaporedje neodvisnih poskusov je zaporedje poskusov, ki so med sabo neodvisni. Torej rezultat naslednjega poskusa ni odvisen od prejšnjega.

2. Kaj je to Bernoullijevo zaporedje neodvisnih poskusov?

rešitev: Zaporedje neodvisnih poskusov se imenuje Bernoullijevo zaporedje, če se more zgoditi v vsakem poskusu iz zaporedja neodvisnih poskusov le dogodek A z verjetnostjo $P(A) = p$ ali dogodek \bar{A} z verjetnostjo $P(\bar{A}) = 1 - P(A) = 1 - p = q$.

3. Kaj so to kombinacije in kako izračunamo njihovo število?

rešitev: Kombinacije so izbori k elementov iz množice n elementov, kjer vrstni red ni pomemben. Število možnih kombinacij velikosti k iz množice velikosti n izračunamo kot $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$.

4. Izpelj Bernoullijev obrazec in predstavi vsaj dva načina/metodi za njegovo računanje.

rešitev:

Bernoullijev obrazec pravimo enačbi:

$$p_n(k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

Število načinov izbora ugodnih dogodkov množimo z verjetnostjo, da se je zgodilo k ugodnih dogodkov ter verjetnostjo, da so bili ostali dogodki neugodni.

Če je n velik, lahko uporabimo De Moivrov obrazec. Če je n velik in $p \approx \frac{1}{2}$, lahko uporabimo Laplaceov obrazec (aproximacija z naravno porazdelitvijo).

Izrek 4.3. (**De Moivrov točkovni obrazec**)

Za velike n velja

$$P_n(k) \approx \frac{1}{\sqrt{\pi n/2}} e^{-\frac{(k-n/2)^2}{n/2}}.$$



De Moivrov točkovni obrazec je poseben primer **Laplaceovega točkovnega obrazca**. Slednjega smemo uporabljati, ko je n velik in p blizu $1/2$:

$$P_n(k) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}.$$

5. Zapiši Stirlingov obrazec

rešitev:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n,$$

ko gre $n \rightarrow \infty$

6. Podaj Bernoullijev zakon velikih števil.

rešitev:

Za vsak $\epsilon > 0$ velja:

$$\lim_{n \rightarrow \infty} P(|\frac{k}{n} - p| < \epsilon) = 1$$

6 Slučajne spremenljivke

1. Kaj je to slučajna spremenljivka in kaj moramo vedeti o njej, da lahko rečemo, da jo poznamo?

rešitev: Slučajna spremenljivka je spremenljivka, katere možne vrednosti so rezultati naključnega fenomena.

Vedeti moramo, kakšne vrednosti lahko zavzame (torej kakšna je njena domena) ter kakšne so verjetnosti, da zavzame posamezno vrednost iz domene (torej kako je porazdeljena).

2. Navedi osnovne lastnosti porazdelitvenih funkcij.

rešitev: Odgovor se nahaja v enem od sledečih razdelkov.

3. Kako delimo slučajne spremenljivke (naštej po tri najbolj pomembne predstavnike vsake vrste) in pojasni zakaj je razlikovanje pomembno?

rešitev: Glede na vrsto vrednosti delimo slučaje spremenljivke na:

- številske (ali numerične) spremenljivke - vrednosti lahko izrazimo s števili. Za podatke rečemo, da so kvalitativni (kategorični) kadar jih delimo v kategorije in zanje ni kvantitativnih interpretacij.
- opisne (ali atributivne) - vrednosti lahko opišemo z imeni razredov (npr. poklic, uspeh, spol); Za podatke rečemo, da so kvalitativni (kategorični) kadar jih delimo v kategorije in zanje ni kvantitativnih interpretacij.

Glede na vrsto merske lestvice, gledimo slučajne spremenljivke na:

- imenske (ali nominalne) spremenljivke
- urejenostne (ali ordinalne) spremenljivke
- razmične (ali intervalne) spremenljivke
- razmerostne spremenljivke
- absolutne spremenljivke - štetja (npr. število prebivalcev)

Sparkly Oats

IURRA ali Indecent Unicorns Rammmed Radioactive Astronauts

Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi v obratnem vrstem redu, kot smo jih našli.

4. Naj bo X število dvojok, ki padejo v dvanajstih neodvisnih metih standardne kocke. Zapišite in poimenujte porazdelitev te slučajne spremenljivke.

rešitev: Slučajna spremenljivka je porazdeljena binomsko $X \sim B(12, \frac{1}{6})$.

5. Študent dobi na izpitu pozitivno oceno z verjetnostjo $\frac{1}{4}$. Na izpit hodi, dokler ga prvič ne opravi. Naj X označuje število opravljanj izpita. Zapišite in poimenujte porazdelitev te slučajne spremenljivke.

rešitev: Iščemo porazdelitev, ki opisuje število potrebnih poskusov do prvega uspeha. Takšna je geometrijska porazdelitev.

6. Koliko je verjetnost, da bo študent šel na vsaj 8 izpitov?

rešitev: Seštejemo verjetnosti, da ga bo opravil v 1., 2., ..., 7. izpitu in to vsoto odštejemo od 1.

$$p = \left(\frac{3}{4}\right)^7$$

Rezultat je 0.1335

7. Kolikokrat bo v povprečju moral na izpit?

rešitev: Zanima nas pričakovana vrednost. Računamo: $E(X) = \frac{1}{p} = 4$.

8. Definiraj slučajni vektor ter robno porazdelitveno funkcijo in pojasni, kdaj so slučajne spremenljivke med seboj neodvisne. Podaj kakšen primer.

rešitev: Slučajni vektor je n -terica slučajnih spremenljivk. Neodvisni sta, ko velja $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ za $\forall x, y$ v domeni X in Y .

Robna porazdelitev slučajnega vektorja je porazdelitev poljubnega dela vektorja. Robna porazdelitev slučajnega vektorja X, Y je porazdelitev ene izmed slučajnih spremenljivk, ki sestavljata slučajni vektor. Robni verjetnostni funkciji slučajnega vektorja X, Y sta $p_X(x_i)$ in $p_Y(y_j)$, robni porazdelitveni funkciji pa sta $F_X(x)$ in $F_Y(y)$.

Pri slučajnih vektorjih pravimo funkciji $F_i(x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty)$ robna porazdelitvena funkcija spremenljivke X_i .

7 Diskretne slučajne spremenljivke

1. Definiraj slučajno spremenljivko in pojasni kdaj je le-ta diskretna. Kako lahko podaš/predstaviš diskretno slučajno spremenljivko?

rešitev: Slučajna spremenljivka je spremenljivka, katere vrednosti so rezultati nekega naključnega pojava. Diskretna je, ko lahko zavzame le končno mnogo vrednosti. Podamo jo lahko z porazdelitveno shemo.

2. Kaj je porazdelitvena funkcija in kako izgleda v primeru diskretne slučajne spremenljivke?

rešitev: Porazdelitvena funkcija je funkcija, ki zbira komulativno verjetnost, da je vrednost s.s. manjša od parametra funkcije. V primeru diskretne slučajne spremenljivke ima stopničasto obliko.

3. Naštej dva primera diskretnih slučajnih spremenljivk, ter pri vsaki opiši tudi njeno verjetnostno tabelo in zgled uporabe.

rešitev: TODO

4. Kaj je to slučajni vektor in kako opišemo v primeru diskretne dvorazsežne porazdelitve njegovo verjetnostno funkcijo? Ali se da iz nje ugotoviti kdaj sta njegovi komponenti neodvisni (pojasni kako oziroma zakaj ne?).

rešitev: Slučajni vektor je n-terica slučajnih spremenljivk. V primeru diskretne dvorazsežnostne porazdelitve ga opišemo s kontingenčno tabelo. Njegovi komponenti sta neodvisni, ko je verjetnost v vsaki celici enaka produktu komponente robnih porazdelitev, ki ustrezajo tisti celici.

5. Razloži kaj si predstavljamo pod pogojno porazdelitvijo (lahko se omejiš na diskretni primer) in kaj je to pogojna verjetnostna funkcija.

rešitev: Pogojna porazdelitev je porazdelitev neke slučajne spremenljivke, če vemo, da je neka druga slučajna spremenljivka zavzela neko vrednost. Označimo jo kot npr. $P(X|Y = y)$. Pogojna verjetnostna funkcija je funkcija, ki prireja verjetnosti vsaki vrednosti v domeni neke spremenljivke ob vedenju, kakšno vrednost je zavzela neka druga slučajna spremenljivka.

6. Definiraj pogojno pričakovano vrednost in izračunaj pričakovano vrednost slučajne spremenljivke $E(X|Y)$.

rešitev:

Pogojna pričakovana vrednost je dolgoročna povprečna vrednost, ki jo zavzame slučajna spremenljivka, če vemo, kakšno vrednost je zavzela neka druga slučajna spremenljivka (fiksiramo njeno vrednost).

$$E(X|Y = y_k) = \sum_{i=1}^{\infty} x_i p_{i|k} = \frac{1}{q_k} \sum_{i=1}^{\infty} x_i p_{ik}.$$

8 Zvezne slučajne spremenljivke

1. Definiraj slučajno spremenljivko in pojasni, kdaj je le-ta zvezna. Kako z gostoto porazdelitve predstaviš verjetnostno porazdelitev zvezne slučajne spremenljivke?

rešitev:

Slučajna spremenljivka je spremenljivka, katere možne vrednosti so rezultati nekega naključnega pojava. Zvezna je, ko lahko zavzame neštevno mnogo vrednosti. Njena verjetnostna funkcija je zvezna. Gostota verjetnosti je enaka seštevek verjetnosti vseh možnih vrednosti manjših od x , ki jih s.s. lahko zavzame. Torej je enaka določenemu integralu $\int_{-\infty}^x p(t) dt$.

Slučajna spremenljivka X je **zvezno porazdeljena**, če obstaja taka **integrabilna** funkcija p ,⁵ imenovana **gostota verjetnosti**, da za vsak $x \in \mathbb{R}$ velja:

$$F(x) = P(X \leq x) = \int_{-\infty}^x p(t) dt,$$

2. Napiši vsaj tri lastnosti porazdelitvene funkcije. Kako izgleda le-ta v primeru diskretne slučajne spremenljivke (lahko poveš tudi samo, kako izgleda graf le-te)?

rešitev: Porazdelitvena funkcija v primeru diskretne slučajne spremenljivke ima stopničasto obliko. Nekaj lastnosti:

- Je nenegativna.
- Je desno zvezna.
- V limiti, ko x narašča proti neskončnosti, se približuje 1.

3. Naštej dva primera zveznih slučajnih spremenljivk, ter pri vsaki podaj tudi njeno gostoto verjetnosti.

rešitev: Primera sta slučajna spremenljivka, porazdeljena s Poissonovo porazdelitvijo ter slučajna spremenljivka, porazdeljena z enakomerno zvezno porazdelitvijo.

Poissonova porazdelitev : $p_k(n) = \lambda^k \frac{e^{-\lambda}}{k!}$ Enakomerna zvezna porazdelitev:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{sicer} \end{cases}$$

4. Kaj je to slučajni vektor? Kako z gostoto verjetnosti opišemo v primeru zvezne dvorazsežne porazdelitve njegovo verjetnostno funkcijo? Kakšnemu pogoju mora ustrezati verjetnostna funkcija, da so komponente slučajnega vektorja neodvisne?

rešitev: Definicija slučajnega vektorja - glej nekaj vprašanj nazaj.

V primeru zvezne dvorazsežne porazdelitve je porazdelitvena funkcija dvojni integral, ki meri prostornino pod funkcijo. Ta prostornina se v limiti akumulira do 1. Da so komponente slučajnega vektorja neodvisne, mora za gostoto verjetnosti veljati:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

5. Pri dvorazsežni diskretni slučajni spremenljivki definiraj $P(X = x_k | Y = y_h)$. Razloži s pomočjo verjetnostne tabele, kaj si (v tem primeru) predstavljamo pod pogojno porazdelitvijo (glede na pogoj $Y = y_h$).

rešitev: TODO

6. Definiraj pogojno pričakovano vrednost in izračunaj pričakovano vrednost slučajne spremenljivke $E(X|Y)$, tj. $E(E(X|Y))$ izrazi z $E(X)$.

rešitev: Ta vrednost funkcije nam predstavlja relativno verjetnost, da s.s. X zavzame vrednost x , če je s.s. Y zavzela vrednost y . Pri verjetnostni tabeli, kjer vsak stopec predstavlja vrednost, ki jo lahko zavzame s.s. X in vsaka vrstica vrednost, ki jo lahko zavzame Y , gledamo stolpec označen z x_k in vrstico označeno z y_k . Vrednost, ki se nahaja v celici pri teh indeksih, je pogojna verjetnost, ki nas zanima.

9 Slučajne spremenljivke in neodvisnost

1. Kaj moramo vedeti o slučajni spremenljivki, da lahko rečemo, da jo poznamo (bodite bolj konkretni v primeru diskretne in zvezne slučajne spremenljivke)

rešitev:

Slučajna spremenljivka je spremenljivka, katere možne vrednosti so rezultati naključnega fenomena. Moramo vedeti kakšna je njena domena in kakšna je verjetnost, da zavzame posamezno vrednost iz domene. Za zvezno slučajno spremenljivko potrebujemo neko funkcijo, ki nam opisuje porazdelitev verjetnosti. Porazdelitev poznamo, če imamo funkcijo $F(x) = P(X \leq x)$, ki opisuje verjetnost, da s.s. X zavzame vrednost manjšo od x .

2. Definiraj slučajni vektor, njegovo porazdelitveno funkcijo in jo opiši v primeru diskretne dvorazsežne porazdelitve.

rešitev: Slučajni vektor je n -terica slučajnih spremenljivk. Neodvisni sta, ko velja $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ za $\forall x, y$ v domeni X in Y .

Robna porazdelitev slučajnega vektorja je porazdelitev poljubnega dela vektorja. Robna porazdelitev slučajnega vektorja X, Y je porazdelitev ene izmed slučajnih spremenljivk, ki sestavljata slučajni vektor. Robni verjetnostni funkciji slučajnega vektorja X, Y sta $p_X(x_i)$ in $p_Y(y_j)$, robni porazdelitveni funkciji pa sta $F_X(x)$ in $F_Y(y)$.

Pri slučajnih vektorjih pravimo funkciji $F_i(x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty)$ robna porazdelitvena funkcija spremenljivke X_i .

3. Opiši kdaj rečemo, da je slučajni vektor zvezno porazdeljen. Kako v tem primeru z gostoto verjetnosti opišemo njegovo porazdelitveno funkcijo in obratno, kako iz porazdelitvene funkcije dobimo gostoto (lahko se omejite na dvorazsežni primer)?

rešitev:

Slučajni vektor $X = (X_1, X_2, \dots, X_n)$ je zvezno porazdeljen, če obstaja integrabilna funkcija (gostota verjetnosti) $p(x_1, x_2, \dots, x_n) \geq 0$ z lastnostjo

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} p(t_1, \dots, t_n) dt_1 dt_2 \dots dt_n$$

$$F(\infty, \dots, \infty) = 1.$$

4. V primeru zvezne porazdelitve definiraj robno porazdelitveno funkcijo in robno verjetnostno gostoto ter pojasni kdaj so slučajne spremenljivke med seboj neodvisne.

rešitev:

V primeru zvezne dvorazsežne porazdelitve imamo robni verjetnostni gostoti, ki sta enaki:

$$p_X(x) = F'_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

$$p_Y(y) = F'_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx$$

5. Ali se da iz verjetnostne funkcije diskretnega dvorazsežnega slučajnega vektorja ugotoviti neodvisnost njegovih komponent (pojasni kako oziroma zakaj ne)?

rešitev: Da. Če sta neodvisni, za vsak par vrednosti s.s. $X = x$ in $Y = y$ velja $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$.

6. Pri diskretni dvorazsežni porazdelitvi definiraj $P(X = x_k | Y = y_k)$. Razloži s pomočjo verjetnostne tabele, kaj si (v tem primeru) predstavljamo pod pogojno porazdelitvijo (glede na pogoj $Y = y_h$).

rešitev: To je verjetnost, da s.s. X zavzame vrednost x_k , če vemo, da je Y zavzela vrednost y_k . Če gledamo verjetnostno tabelo, kjer so vrednosti s.s. X na vrhu in vrednosti Y na levi, potem je ta pogojna verjetnost enaka vrednosti v celici, ki jo določata x_k in y_k . Pogojna porazdelitev spremenljivke X , ob pogoju, da je $Y = y_k$ pa je vrstica verjetnosti, kjer je $Y = y_k$.

10 Pričakovana vrednost slučajne spremenljivke

1. Definiraj pričakovano vrednost (tj. matematično upanje) slučajne spremenljivke.

rešitev: Glej odgovor v enem od sledečih razdelkov.

2. Vsaka srečka stane 2 EUR. Prodanih je bilo 400 srečk, med njimi pa ena dobi 250 EUR, drugi dve pa vsaka po 25 EUR. Izračunaj kakšen povprečen izkupiček lahko pričakuješ z eno srečko?

rešitev: Verjetnost, da zadanemo 250 EUR je $\frac{1}{400}$. Verjetnost, da zadanemo 25 EUR je $\frac{2}{400}$. Računajmo pričakovano vrednost:

$$E(X) = -2 + \frac{1}{400} \cdot 250 + \frac{2}{400} \cdot 25 = -1.25$$

3. Naj bosta X in Y slučajni spremenljivki. Izeplji, koliko je pričakovana vrednost spremenljivke $aX + bY$, kjer sta a in b realni števili.

rešitev: Upoštevamo latnosti aditivnosti in homogenosti (linearnost)

$$E(aX + bY) = E(aX) + E(bY) = aE(X) + bE(Y)$$

4. Izračunaj pričakovano vrednost slučajne spremenljivke W , ki je porazdeljena (a) Normalno $N(3, 14)$.

rešitev: Pričakovana vrednost je poleg standardnega odklona ena od vrednosti, ki zelo dobro definira normalno porazdelitev. Je prva vrednost v zgornji notaciji. Torej velja: $\mu = 3$.

5. Izračunaj pričakovano vrednost slučajne spremenljivke W , ki je porazdeljena binomsko $B(100, 0.25)$.

rešitev: Formula za izračun pričakovane vrednosti je $E(X) = np$, kar si ni težko intuitivno predstavljati. Velja torej $E(X) = 100 \cdot 0.25 = 25$

6. Naj bo spremenljivka porazdeljena po shemi $\begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{10} & \frac{2}{10} & \frac{3}{10} & ? \end{pmatrix}$

Določi pričakovano vrednost $E(X)$.

rešitev: Mankajočo vrednost poiščemo tako, da

7. Spremenljivka Y ima gostoto verjetnosti $f(y) = cy(1 - y)$ za $0 < y < 1$, določi pričakovano vrednost $E(Y)$.

rešitev: rešimo enačno:

$$F(y) = c \int_0^1 y(1 - y)dy = 1$$

po trivialnem postopku za c in dobimo rezultat $c = 4$.

8. Vzorčno povprečje je

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

Kjer je množica $\{X_i\}$ slučajen vzorec slučajne spremenljivke X . Izračunaj $E(\bar{X})$, če veš, da je $E(X) = \mu$

rešitev: Odgovor se nahaja v enem od sledečih razdelkov.

11 Disperzija slučajne spremenljivke

- Definiraj oziroma pojasni, kaj je razpršenost (disperzija) slučajne spremenljivke.

rešitev: Disperzija (tudi statistična variabilnost, variacija) je v opisni statistiki variiranje statističnih spremenljivk v populaciji okoli srednje vrednosti.

Pod pojmom razpršenost razumemo variranje (odklanjanje) vseh ali dela individualnih vrednosti neke populacije (npt. variranje telesne teže 100 naključno izbranih ljudi) okoli mere sredine. Sredina se lahko pojavlja v več oblikah, med drugim mediana, aritmetična sredina ali modus. Če so vsi podatki identični in se ne odklanjajo od mere sredine, je razpršenost enaka nič. Velik povprečni odklon individualnih vrednosti od povprečne vrednosti populacije pomeni veliko razpršenost.

Ena bolj pomembnih mer razpršenosti je varianca. Standardni odklon je drugi koren variance in je bolj razširjena oblika za ponazoritev variiranja kot varianca.

Druge mere razpršenosti so različni razmiki (variacijski razmik, kvartilni razmik, decilni razmik) in odkloni (varianca, standardni odklon, kvartilni odklon in povprečni absolutni odklon). Nobena izmed mer ne more biti negativna, najbolj redka vrednost pa je nič.

- Vržemo kovanec za 20 centov. Kar pade je slučajna spremenljivka X (če pokaže 20, je njena vrednost 20, sicer 0). Koliko je disperzija od X ?

rešitev: Uporabimo formulo $D(X) = E(X^2) - (E(X))^2$.

Izračunajmo vrednosti $E(X^2)$ in $E(X)^2$.

$$E(X^2) = \begin{bmatrix} 20^2 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = 200$$

$$E(X)^2 = \left(\begin{bmatrix} 20 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \right)^2 = 100$$

Iz tega sledi, da je disperzija enaka

$$D(X) = 200 - 100 = 100.$$

- Naj bodo X , Y in Z slučajne spremenljivke. Koliko je disperzija od aX , kjer je a realno število in koliko je disperzija od $Y + Z$, če sta slučajni spremenljivki Y in Z neodvisni.

rešitev: Velja:

$$D(aX) = a^2 D(X)$$

in

$$D(Y + Z) = D(Y) + D(Z)$$

Če sta slučajni spremenljivki Y in Z neodvisni. Ker sta neodvisni, sta tudi nekorelirani in člen korelacije je enak 0 ter posledično velja linearnost.

12 Normalna porazdelitev

1. Definiraj standardni odklon in pričakovano vrednost slučajne spremenljivke.

rešitev: Odgovorjeno v enem od naslednjih razdelkov.

2. Opiši postopek za standardizacijo slučajne spremenljivke in izračunaj njeno pričakovano vrednost.

rešitev: Odgovorjeno v enem od naslednjih razdelkov.

3. Opiši normalno porazdelitev in razloži, zakaj je pomembna.

rešitev: Normalna porazdelitev je ena od najpomembnejših porazdelitev, ki se zelo veliko pojavlja v naravi in drugod. Zelo dobro jo opisujeta povprečna vrednost in standardni odklon.

Njena uporabnost v veliki meri izvira iz Centralnega limitnega izreka. Zaradi tega je zanjo razvitih veliko učinkovitih načinov analize.

Omenimo še pravila 68-95-99.7, komulativno distributivno funkcijo, standardizacijo.

4. Naj bo X normalna slučajna spremenljivka s parametroma $\mu = 19$ in $\sigma^2 = 9$. Izračunaj naslednji vrednosti: $P(X > 21)$, $P(X \leq 20)$.

rešitev:

Vrednost normaliziramo, da dobimo oddaljenost od pričakovane vrednosti v standardnih deviacijah. Nato izračunamo verjetnost, da standardizirana normalno porazdeljena slučajna spremenljivka zavzame to vrednost.

$$P(X > 21) = 1 - P(X \leq 21) = 1 - P\left(Z < \frac{21 - 19}{3}\right) = 1 - P(Z < 0.67) = 1 - 0.7486 = 0.2514$$

$$P(X \leq 20) = P\left(Z \leq \frac{20 - 19}{3}\right) = P(Z \leq 0.333) = 0.6304$$

5. Opiši večrazsežno normalno porazdelitev

rešitev: Odgovorjeno v enem od naslednjih razdelkov.

13 Porazdelitvene funkcije vektorjev in srednje vrednosti

1. Definiraj porazdelitveno funkcijo slučajnega vektorja (X, Y) in naštej vsaj tri njene lastnosti.

rešitev:

Slučajni vektor je n -terica slučajnih spremenljivk $\mathbf{X} = (X_1, \dots, X_n)$. Tudi za slučajni vektor opišemo porazdelitveni zakon s porazdelitveno funkcijo ($x_i \in \mathbb{R}$):

$$F(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n),$$

(pri čemer slednja oznaka pomeni $P(\{X_1 \leq x_1\} \cap \dots \cap \{X_n \leq x_n\})$) in za katero velja: $0 \leq F(x_1, \dots, x_n) \leq 1$. Funkcija F je za vsako spremenljivko naraščajoča in z desne zvezna, veljati pa mora tudi

$$F(-\infty, \dots, -\infty) = 0 \quad \text{in} \quad F(\infty, \dots, \infty) = 1.$$

Funkciji $F_i(x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty)$ pravimo **robna porazdelitvena funkcija** spremenljivke X_i .

Primer: Katere od naslednjih funkcij so lahko porazdelitvene funkcije nekega slučajnega vektorja (X, Y) :

- (a) $F(x, y)$ je enaka $1 - e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0,
- (b) $F(x, y)$ je enaka $1 + e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0,
- (c) $F(x, y) = x^2$,
- (d) $F(x, y) = x^2 - y^2$,
- (e) $F(x, y) = 0$.

2. Utemelji katere od naslednjih funkcij so porazdelitvene funkcije nekega slučajnega vektorja (X, Y) in katere ne morejo biti: (i) $F(x, y)$ je enaka $1 - e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0, (ii) $F(x, y)$ je enaka $1 + e^{-x-y} - e^{-x} - e^{-y}$, ko sta x in y oba nenegativna, sicer pa je 0, (iii) $F(x, y) = x^2$, (iv) $F(x, y) = x^2 - y^2$, (v) $F(x, y) = 0$.

Funkcija iz (a) nima vseh vrednosti na intervalu $[0, 1]$, npr. $F(0, 0) = 1 - 1 - 1 - 1 = -2 < 0$, zato ne more biti porazdelitvena funkcija. Podobno je tudi v primerih (c): $F(2, 0) = 4 \notin [0, 1]$ in (d): $F(0, 1) = -1 \notin [0, 1]$. V primeru (e) pa velja $F(\infty, \infty) = 0 \neq 1$, kar pomeni, da nam ostane le še možnost (b). V tem primeru lahko zapišemo $F(x, y) = (1 - e^{-x})(1 - e^{-y})$ od koder vidimo, da za $x \geq 0$ in $y \geq 0$ velja $F(x, y) \in [0, 1]$. Preverimo še $F(0, 0) = 0$ in $F(\infty, \infty) = 1$. \diamond

rešitev:

3. Iz prejšnje točke izberi eno porazdelitveno funkcijo in zanjo izračunaj mediano robne porazdelitve.

rešitev:

$$p_X(x) = F'_X(x) = \int_{-\infty}^{\infty} p(x, y) dy, \quad \text{in} \quad p_Y(y) = F'_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx.$$

Mediano izračunaš tako, da izračunaš točko, pri kateri ima komulativna distirbucijska (porazdelitvena) funkcija vrednost 0.5.

4. Naj bosta X in Y diskretni slučajni spremenljivki. Dokažite, da sta neodvisni natanko takrat, ko za vsak par celih števil k, l velja

$$P(X \geq k, Y \leq l) = P(X \geq k) \cdot P(Y \leq l)$$

rešitev:

$$P(X \leq x_k, Y \leq y_l) = \sum_{i \leq k} \sum_{j \leq l} p_{ij} = \sum_{i \leq k} \sum_{j \leq l} p_i \cdot q_j = \sum_{i \leq k} p_i \cdot \sum_{j \leq l} q_j = P(X \leq x_k) \cdot P(Y \leq y_l).$$

14 Kovarianca

1. Definiraj pričakovano vrednost in disperzijo slučajne spremenljivke, nato pa še kovarianco slučajnih spremenljivk.

rešitev: Odgovorjeno v naslednjih razdelkih.

2. Naj bodo X , Y in Z slučajne spremenljivke. Izračunaj kovarianco $K(aX + bY, Z)$, če poznaš $K(X, Z)$ in $K(Y, Z)$. Koliko je kovarianca $K(Y, Z)$, če poznaš $K(Z, Y)$? Kaj lahko poveš o kovarianci $K(X, Y)$, če veš, da velja $D(X + Y) = D(X) + D(Y)$?

rešitev: Ker za kovarianco velja bilinearnost, lahko zapišemo:

$$K(aX + bY, Z) = aK(X, Z) + bK(Y, Z)$$

Ker je kovarianca simetrična funkcija, velja $K(Y, Z) = K(Z, Y)$.

Če velja, $D(X + Y) = D(X) + D(Y)$, potem lahko iz tega potegnemo implikacijo, da je kovarianca enaka nič. Spremenljivki sta nekorelirani, ampak ne nujno neodvisni.

3. X in Y naj bosta slučajni spremenljivki z $D(X) = 2$, $D(Y) = 3$ in $K(X, Y) = -1$. Koliko je a , da velja $K(2X + Y, 3Y - aX) = 0$? Kaj to pomeni za slučajni spremenljivki $2X + Y$ in $3Y - aX$? Kaj pa, če dodatno veš, da sta X in Y porazdeljeni normalno?

rešitev:

Za izračun a upoštevamo lastnost bilinearnosti in zapišemo:

$$K(2X + Y, 3Y - aX) = 0$$

$$2K(X, 3Y - aX) + K(Y, 3Y - aX) = 0$$

$$2K(3Y - aX, X) + K(3Y - aX, Y) = 0$$

$$6K(Y, X) - 2aK(X, X) + 3K(Y, Y) - aK(X, Y)$$

Upoštevamo, da velja $K(X, X) = D(X)$ ter uporabimo znane vrednosti.

Dobimo, da je $a = 1$.

V splošnem sta $2X + Y$ in $3Y - X$ nekorelirani, če sta X in Y normalni, pa sta tudi neodvisni.

Pomni: Nekorelirani normalno porazdeljeni spremenljivki sta tudi neodvisni.

4. Za slučajni spremenljivki X in Y definiraj korelacijski koeficient $r(X, Y)$ ter pojasni kaj veš v primeru, ko je $|r(X, Y)| = 1$.

rešitev: Korelacijski (Pearsonov) koeficient definiramo kot kvocient $r(X, Y) = \frac{K(X, Y)}{\sigma_X \sigma_Y}$

Če je absolutna vrednost korelacijskega koeficienta enaka 1, potem obstaja med s.s. X in Y linearna zveza z verjetnostjo 1.

5. Zvezo, ki si jo uporabil v zadnji točki pri (b), utemelji in sploši na slučajne spremenljivke X_1, \dots, X_n .

rešitev:

Po definiciji disperzije velja $D(X, Y) = E(X + Y - E(X, Y))^2 - E(X - E(X) + Y - E(Y))^2$

Oziroma velja:

$$D(X + Y) = D(X) + D(Y) + 2K(X, Y)$$

n

To lahko posplošimo na poljubno število s.s. kot:

$$D\left(\sum_i X_i\right) = \sum_i D(X_i) + \sum_{i \neq j} K(X_i, X_j)$$

15 Momenti, centralni limitni izrek in neenakost Čebiševa

1. Vpelji moment slučajne spremenljivke X reda k glede na točko a (kjer je k naravno število, a pa realno), pojasni tudi kdaj obstaja.

rešitev: Višji momenti so posplošitev pojmov pričakovana vrednost in disperzija. Moment reda $k \in \mathbb{N}$ glede na točko $a \in \mathbb{R}$ imenujemo količino

$$m_k(a) = E((X - a)^k)$$

Moment obstaja, če obstaja pričakovana vrednost $E(|X - a|^k) < \infty$. Za $a = 0$ dobimo začetni moment $z_k = m_k(0)$; za $a = E(X)$ pa centralni moment $m_k = m_k(E(X))$.

primer: $E(X) = z_1$ in $D(X) = m_2$.

trditev: Če obstaja moment $m_n(a)$, potem obstajajo tudi vsi momenti $m_k(a)$ za $k < n$. Če obstaja moment z_n , obstaja tudi moment $m_n(a)$ za vse $a \in \mathbb{R}$ ter velja

$$m_n(a) = E((X - a)^n) = \sum_{k=0}^n \binom{n}{k} (-a)^{n-k} z_k$$

Posebej za centralni moment velja $m_0 = 1$, $m_1 = 0$, $m_2 = z_2 - z_1^2$, $m_3 = z_3 - 3z_2z_1 + 2z_1^3$, ...

$$m_n = m_n(z_1) = \sum_{k=0}^n \binom{n}{k} (-z_1)^k z_{n-k}$$

2. Definiraj še začetni in centralni moment in nato z njimi še pričakovano vrednost in razpršenost ali asimetrijo spremenljivke X .

rešitev: Glej prejšno vprašanje.

3. Vpelji standardizacijo Z slučajne spremenljivke X in nato izračunaj pričakovano vrednost in razpršenost standardizirane spremenljivke Z .

rešitev: Odgovorjeno v sledečih razdelkih.

4. Predstavi centralni limitni izrek

rešitev: Odgovorjeno v sledečih razdelkih.

5. Zapiši neenakost Čebiševa, nato pa z besedami pojasni njen pomen (ali pa podaj kakšno njeno posledico). Dokaži Čebiševo neenakost!

rešitev:

Izrek 9.2. [Neenakost Čebiševa] Če ima slučajna spremenljivka X končno disperzijo, tj. $D(X) < \infty$, potem za vsak $\varepsilon > 0$ velja

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



Dokaz: Pokažimo jo za zvezne spremenljivke

$$\begin{aligned} P(|X - E(X)| \geq \varepsilon) &= \int_{|x - E(X)| \geq \varepsilon} p(x) dx = \frac{1}{\varepsilon^2} \int_{|x - E(X)| \geq \varepsilon} \varepsilon^2 p(x) dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x - E(X))^2 p(x) dx = \frac{D(X)}{\varepsilon^2}. \quad \square \end{aligned}$$

Posledica 9.3. [Markov] Če gre za zaporedje slučajnih spremenljivk X_i izraz $\frac{D(S_n)}{n^2} \rightarrow 0$, ko gre $n \rightarrow \infty$, velja za zaporedje šibki zakon velikih števil.

Posledica 9.4. [Čebišev] Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C , tj. $D(X_i) < C$ za vsak i , velja za zaporedje šibki zakon velikih števil.

Za Bernoullijevo zaporedje X_i so spremenljivke paroma neodvisne, $D(X_i) = pq$, $S_n = k$. Pogoji izreka Čebiševa so izpolnjeni in zopet smo prišli do Bernoullijevega zakona velikih števil, tj. izreka 5.5.

16 Binomska in normalna porazdelitev ter CLI in vzorčenje

1. Definiraj pričakovano vrednost in standardni odklon slučajne spremenljivke.

rešitev: Glej naslednji razdelek.

2. Opiši postopek za standardnizacijo slučajne spremenljivke in zapiši njeno pričakovano vrednost ter njen odklon.

rešitev: Glej naslednji razdelek.

3. Opiši binomsko porazdelitev ter Laplaceov obrazec.

rešitev: Binomska porazdelitev opisuje verjetnost, da se bo v zaporedju n poskusov neodvisnih eksperimentov, ki imajo dva možna izida, zgodilo k uspešnih izidov.

Binomska porazdelitev ima zalogo vrenosti $0, 1, 2, \dots, n$ in verjetnosti, ki jih računamo po Bernoullijevem obrazcu:

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$k = 0, 1, 2, \dots, n$.

Binomska porazdelitev je natanko določena z dvema podatkom - parametroma $n \in \mathbb{N}$ in $p \in [0, 1]$. Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p , zapišemo:

$$X \sim B(n, p)$$

Laplaceov intervalski obrazec

Iz Laplaceovega točkovnega obrazca izhaja, da za p blizu $1/2$ in velike n velja:

$$B(n, p) \approx N(np, \sqrt{npq}).$$

Sedaj pa nas zanima še, kolikšna je verjetnost $P_n(k_1, k_2)$, da se v Bernoullijevem zaporedju neodvisnih poskusov v n zaporednih poskush zgodi dogodek A vsaj k_1 -krat in manj kot k_2 -krat. Označimo

$$x_k = \frac{k - np}{\sqrt{npq}} \quad \text{in} \quad \Delta x_k = x_{k+1} - x_k = \frac{1}{\sqrt{npq}}.$$

Tedaj je, če upoštevamo Laplaceov točkovni obrazec,

$$P_n(k_1, k_2) = \sum_{k=k_1}^{k_2-1} P_n(k) = \frac{1}{\sqrt{2\pi}} \sum_{k=k_1}^{k_2-1} e^{-\frac{1}{2}x_k^2} \Delta x_k.$$

Za (zelo) velike n lahko vsoto zamenjamo z integralom

$$P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx.$$

4. Predstavi centralni limitni izrek (CLI)

rešitev: Glej video od Khan Academy +

Leta 1810 je Pierre Laplace (1749-1827) študiral anomalije orbit Jupitera in Saturna, ko je izpeljal razširitev De Moivrevega limitnega izreka.



(Osnovni CLI) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo ter je $S_n = X_1 + \dots + X_n$, potem zanje velja **centralni limitni zakon**, tj. porazdelitvene funkcije za standardizirane spremenljivke

$$Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)}$$

gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve oziroma simbolično za vsak $x \in \mathbb{R}$ velja

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n - E(S_n)}{\sigma(S_n)} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Opomba: v praksi uporabimo $n > 30$.

5. Opiši večrazsežno normalno porazdelitev

rešitev:**Večrazsežna normalna porazdelitev**

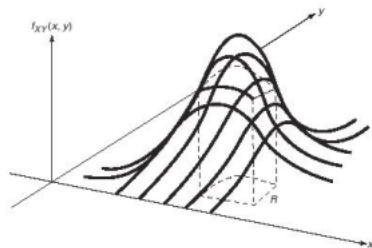
V dveh razsežnostih označimo normalno porazdelitev z $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ in ima gostoto

$$p(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}.$$

V splošnem pa jo zapišemo v matrični obliki

$$p(\mathbf{x}) = \sqrt{\frac{\det A}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T A (\mathbf{x}-\boldsymbol{\mu})},$$

kjer je A simetrična pozitivno definitna matrika, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$, 'eksponent' T pa pomeni transponiranje (tj. zrcaljenje matrike preko glavne diagonale). Obe robni porazdelitvi sta normalni.



6. Naj bo X slučajna spremenljivka z $E(X) = \mu$ in $D(X) = \sigma^2$. Za njen slučajen vzorec $X_{i=1}^n$ definiraj vzorčno povprečje \bar{X} in napiši, kaj se dogaja z vzorčnim povprečjem $E(\bar{X})$ in s standardno napako $D(\bar{X})$ z naraščanjem velikosti vzorca.

rešitev: Glej naslednji razdelek.

17 Opisna statistika

1. Definiraj kvantil, centil, mediano, modus in komulativo.

rešitev:

Kvantili so točke, ki delijo domeno verjetnostne porazdelitve na intervale, ki so enako verjetni. Kvantilov je eden manj kot intervalov porazdelitvene funkcije, ki z delitvijo nastanejo.

Centil (ali perentil) je mera, ki kaže, pod katero pade dan procent opazovanj v skupini opazovanj. 20. percentil je na primer vrednost, pod katero najdemo 20% opazovanj.

Mediana je vrednost, ki loči zgornjo polovico vrednosti podatkov od spodnje. Je vrednost, od katere je enako število ostalih vrednosti manjših kot večjih. Prednost mediane pred povprečjem je ta, da je bolj "odporna" na zelo ekstremne vrednosti in lahko da boljšo idejo kaj je to tipična vrednost kot pa povprečje.

Modus je vrednost, ki se pojavi največkrat.

Komulativa je funkcija, ki sešteva verjetnosti do neke meje. Npr. funkcija $F(x) = P(X \leq x)$ vrne verjetnost, da je slučajna spremenljivka X zavzela vrednost manjšo ali enako x . Če je porazdelitev diskretna, potem preprosto seštejemo verjetnosti za vse vrednosti manjše ali enake x , v primeru zvezne porazdelitve pa naredimo neskončno (Riemmanovo) vsoto, kar pomeni, da integriramo od $-\infty$ do x .

2. Definiraj srednjo vrednost in odklon.

rešitev: Pričakovana vrednost $E(X)$ (oz. matematično upanje) za končne diskretne slučajne spremenljivke X smo vpeljali kot posplošitev povprečne vrednosti. Splošna diskretna slučajna spremenljivka X z verjetnostno funkcijo p_k ima pričakovano vrednost:

$$E(X) = \sum_{i=1}^{\infty} x_i p_i$$

Če je

$$\sum_{i=1}^{\infty} |x_i| p_i < \infty.$$

Zvezna slučajna spremenljivka X z gostoto $p(x)$ ima pričakovano vrednost:

$$E(X) = \int_{-\infty}^{\infty} x p_x dx$$

če je

$$\int_{-\infty}^{\infty} |x| p(x) dx < \infty$$

Standardni odklon je način, kako izrazimo varianco v enakih enotah kot je povprečje. Je mera, ki pove, kako je porazdelitvena funkcija stisnjena ali raztegnjena. Vzorčno standardno deviacijo izračunamo kot:

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

Kjer so $\{x_1, x_2, \dots, x_N\}$ opazovane vrednosti vzorčnih elementov, \bar{x} je vzorčno povprečje in n je velikost vzorca.

3. Vpelji standardizacijo Z slučajne spremenljivke X in izračunaj srednjo vrednost in odklon od Z (posamezne korake utemelji).

rešitev: Vsaki vrednosti x_i spremenljivke X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ .

$$z_i = \frac{x_i - \mu}{\sigma}$$

Za novo spremenljivko Z bomo rekli, da je standardizirana, z_i pa je standardizirana vrednost.

Potem je $\mu(Z) = 0$ in $\sigma(Z) = 1$.

$$E\left(\frac{C - \mu}{\sigma}\right) = \frac{1}{\sigma} E(C - \mu) = \frac{1}{\sigma} E(\mu - \mu) = 0$$

$$V\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2} V(X - \mu) = \frac{1}{\sigma^2} V(X) = \frac{\sigma^2}{\sigma^2} = 1$$

4. Ali obstaja kakšna zveza med ogivo in porazdelitveno funkcijo?

rešitev: Pri statistiki je ogiva prostoročno narisani graf, ki prikazuje krivuljo kumulativne distributivne funkcije. Ime je dobila po tem, da ogiva normalne porazdelitve izgleda kot ogiva iz arhitekture.

18 Vzorčne statistike

1. Kaj je to enostavni slučajni vzorec?

rešitev: Recimo, da merimo spremenljivko X , tako da n -krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X . Postopku ustreza slučajni vektor.

$$(X_1, \dots, X_n)$$

vrednostim meritev (x_1, \dots, x_n) pa rečemo vzorec. Število n je velikost vzorca.

Ker v vzorcu merimo isto spremenljivko in posamezna meritev ne sme vplivati na ostale, lahko postavimo:

- vsi členi X_i vektorja imajo isto porazdelitev, kot spremenljivka X ,
- členi vektorja X_i so med seboj neodvisni.

Takemu vzorcu rečemo enostavni slučajni vzorec. Večina statistične teorije temelji na predpostavki, da imamo opravka z enostavnim slučajnim vzorcem. Če je populacija končna, lahko dobimo enostavni slučajni vzorec tako, da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo. Z vprašanjem, kako sestaviti dobre vzorce v praksi, se ukvarja posebno področje statistike - *teorija vzorčenja*.

2. Pojasni pojem vzorčna statistika.

rešitev: Glej naslednji razdelek prvo vprašanje.

3. Vzorčno povprečje je določeno z zvezo

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

Recimo, da ima spremenljivka X parametra $E(X) = \mu$ in $D(X) = \sigma^2$. Izračunaj $E(\bar{X})$ in $D(\bar{X})$.

rešitev:

Naj so X_1, X_2, \dots, X_n neodvisne enote iz distirbucije s pričakovano vrednostjo μ in varianco σ^2 .

Torej velja $E(X_i) = \mu$ in $Var(X_i) = \sigma^2$.

Naj je \bar{X} povprečje teh n neodvisnih observacij:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Računajmo:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \left(\frac{1}{n}\right)E(X_1 + X_2 + \dots + X_n) = \\ &= \left(\frac{1}{n}\right)(E(X_1) + E(X_2) + \dots + E(X_n)) = \\ &= \frac{1}{n}(\mu + \mu + \dots + \mu) = \\ &= \frac{1}{n} \cdot n\mu = \mu \\ Var(\bar{X}) &= Var\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \left(\frac{1}{n}\right)^2 Var(X_1 + X_2 + \dots + X_n) = \\ &= \left(\frac{1}{n}\right)^2 (var(X_1) + var(X_2) + \dots + var(X_n)) = \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

4. Iz populacije, porazdeljene normalno $N(\mu, 6)$ vzamemo slučajni vzorec in dobimo:

$$[115, 101, 110, 103, 111, 114, 107, 118, a]$$

Kolikšna naj bo vrednost zadnje meritve a , tako da bo vzorčna disperzija čim manjša?

rešitev: Vzorčna disperzija bo najmanjša, če bo a enak povprečju prejšnjih meritev. Lahko tudi zapišemo formulo za vzorčno disperzijo in jo minimiziramo glede na a .

5. Kaj se dogaja z vzorčnim povprečjem in s standardno napako z naraščanjem velikosti vzorca (to nam zagotavlja tudi krepki zakon velikih števil).

rešitev: Standardna napaka porazdelitve vzorčnega povprečja se zmanjšuje, saj je obratno sorazmerna kvadratnemu korenu velikosti vzorca. Posledično se tudi intervali zaupanja, s katerimi z neko gotovostjo podamo, kje se nahaja dejansko populacijsko povprečje, ožajo.

Naj bo X_1, \dots, X_n zaporedje spremenljivk, ki imajo matematično upanje. Označimo $S_n = \sum_{k=1}^n X_k$ in

$$Y_n = \frac{S_n - E(S_n)}{n} = \frac{1}{n} \sum_{k=1}^n (X_k - E(X_k)) = \frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n E(X_k).$$

Pravimo, da za zaporedje slučajnih spremenljivk X_k velja:

- Šibki zakon velikih števil, če gre z verjetnostjo $Y_n \rightarrow 0$, tj. če $\forall \epsilon > 0$ velja

$$\lim_{n \rightarrow \infty} P(|\frac{S_n - E(S_n)}{n}| < \epsilon) = 1$$

- Kreпки zakon velikih števil velja, če gre skoraj gotovo $Y_n \rightarrow 0$, tj. če velja

$$P(\lim_{n \rightarrow \infty} \frac{S_n - E(S_n)}{n} = 0) = 1$$

Če za zaporedje X_1, \dots, X_n velja kreпки zakon velikih števil, velja tudi šibki zakon velikih števil.

19 Vzorčne statistike in porazdelitve

1. Kakšna je razlika med cenilko in vzorčno statistiko (a najprej definiraj obe)?

rešitev: Za definicijo cenilke glej naslednji razdelek.

Vzorčna statistika je poljubna simetrična funkcija (tj. njena vrednost je neodvisna od permutacije argumentov vzorca

$$Y = g(X_1, X_2, \dots, X_n)$$

Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca. Najzanimivejši sta značilni vrednosti

- pričakovana vrednost $E(Y)$, za katero uporabimo vzorčno povprečje, in
- standardni odklon ρ_Y , ki mu pravimo tudi standardna napaka statistike Y (angl. standard error - zato oznaka $SE(Y)$), za katerega uporabimo vzorčni odklon.

Statistika je funkcija vzorca. Cenilka pa je funkcija vzorca povezana z neko kvantiteto porazdelitve.

Cenilka je statistika, ki je uporabljena za inferenco vrednosti neznanega parametra v statističnem modelu.

2. Kako pridemo do porazdelitve $\chi^2(n)$ (najlažje je to opisati v dveh korakih)? Kakšno porazdelitev dobimo, če je n zelo velik.

rešitev: Naj je

$$X_1 \sim N(0, 1) \Rightarrow E(X_1) = 0, \quad Var(x_1) = 1$$

In naj je

$$Q = X_1^2$$

Potem velja, da je Q porazdeljena kot $Q \sim \chi_1^2$.

Če vzamemo še eno standardno normalno porazdeljeno slučajno spremenljivko (recimo ji X_2), je njuna vsota kvadratov porazdeljena kot:

$$X_1^2 + X_2^2 \sim \chi_2^2$$

Podobno lahko nadaljujemo za poljubne vsote kvadratov neodvisnih standardno normalno porazdeljenih slučajnih spremenljivk.

3. Kdaj rečemo, da je slučajna spremenljivka X porazdeljena s Fisherjevo (Snedecorjevo) porazdelitvijo? Če je X porazdeljena z $F(3, 6)$, kako je porazdeljena slučajna spremenljivka $\frac{1}{X}$?

rešitev: Kadar se jo da napisati kot kvocient dveh neodvisnih slučajnih spremenljivk, ki sta porazdeljeni s hi-kvadrat

$F(6, 3)$

Minilo je malo več kot 100 let od kar je Gosset objavil članek o Studentovi t -porazdelitvi z $n - 1$ prostostnimi stopnjami. Kako pridemo do nje? Kakšno porazdelitev dobimo, če je (a) n zelo velik, in kakšno, če je (b) $n - 1 = 1$?

rešitev: Studentova t -porazdelitev je podobna normalni porazdelitvi, le da ima debelejša repe. Če je n zelo velik, se porazdelitev približuje standardni normalni porazdelitvi. Če velja $n - 1 = 1$, potem dobimo Cauchyjevo porazdelitev.

20 Statistično sklepanje o korelacijski povezanosti in regresija

1. Kaj je cenilka in kdaj je nepristranska?

rešitev:

Točkovna cenilka je pravilo ali formula, ki nam pove, kako izračunati numerično oceno parametra populacije na osnovi merjenj vzorca.

Število, ki je rezultat izračuna, se imenuje točkova ocena (in mu ne moremo zaupati v smislu verjetnosti). Cenilka parametra ζ je vzorčna statistika $C = C(X_1, \dots, X_n)$, katere porazdelitveni zakon je odvisen le od parametra ζ , njene vrednosti pa ležijo v prostoru parametrov. Seveda je odvisna tudi od velikosti vzorca n .

Primer: vzorčna mediana \tilde{X} in vzorčno povprečje \bar{X} sta cenilki za populacijsko povprečje μ . Popravljen vzorčna disperzija s^2 pa je cenilka za populacijsko disperzijo σ^2 .

Vzorčno povprečje \bar{X} je dosledna cenilka za populacijsko povprečje μ . Tudi vsi vzorčni začetni momenti

$$Z_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

So dosledne cenilke ustreznih začetnih populacijskih momentov $z_k = E(X^k)$, če le-ti obstajajo.

Primer: Vzorčna mediana \tilde{X} je dosledna cenilka za populacijsko mediano.

Cenilka C_n parametra ζ je nepristranska, če je $E(C_n) = \zeta$ (za vsak n); in je asimptotično nepristranska, če je $\lim_{n \rightarrow \infty} E(C_n) = \zeta$. Količino $B(C_n) = E(C_n) - \zeta$ imenujemo pristranost (angl. bias) cenilke C_n .

Primer: Vzorčno povprečje \bar{X} je nepristranska cenilka za populacijsko povprečje μ ; vzorčna disperzija s_0^2 je samo asimptotično nepristranska cenilka za σ^2 . Popravljen vzorčna disperzija s^2 pa je nepristranska cenilka za σ^2 .

2. Definiraj (Pearsonov) koeficient korelacije ρ za dve številske slučajni spremenljivki. Kakšne vrednosti lahko zavzame (morda veš zakaj) in kaj se zgodi v primeru, če sta slučajni spremenljivki neodvisni (je to potreben pogoj)?

rešitev: TODO

3. Kaj lahko poveš v primeru, ko ρ doseže največjo oziroma najmanjšo možno vrednost?

rešitev: TODO

4. Kaj sta prva in druga regresijska funkcija in kje se srečata?

rešitev: TODO

5. Opiši preverjanje domneve o linearni povezanosti (če se ne spomniš konkretne testne statistike, napiši vsaj kako se porazdeljuje ter opiši omenjeno porazdelitev).

rešitev: TODO

6. Kaj so časovne vrste? Opiši določanje trenda z metodo najmanjših kvadratov.

rešitev: TODO

21 Intervali zaupanja

1. Pojasni razliko med točkovno in intervalno oceno.

rešitev:

Točkovna ocena:

Točkovna ocena populacijskega parametra je ena sama vrednost statistike. Na primer, vzorčno povprečje je točkovna ocena populacijskega povprečja. Podobno je vzorčni delež \hat{p} točkovna ocena populacijskega deleža P .

Intervalna ocena:

Intervalno oceno definirata dve vrednosti, med katerima lahko z neko gotovostjo trdimo, da leži populacijski parameter. Na primer, $a < \bar{X} < b$ je intervalna ocena populacijskega povprečja μ . Pomeni, da lahko z neko gotovostjo trdimo, da je populacijsko povprečje večje od a in manjše od b .

2. Opiši postopek intervalskega ocenjevanja parametravo in pojasni, kaj nam pove koeficient zaupanja $(1 - \alpha)$ (teoretična interpretacija)

rešitev:

S slučajnim vzorcem ocenjujemo parameter γ . Poskušamo najti:

- (a) (1) statistiko g , ki je nepristranska (tj. $E(g) = \gamma$ in se na vseh možnih vzorcih vsaj približno normalno porazdeljuje s standardno napako $SE(g)$) in
- (b) (2) t.i. interval, v katerem se bo z dano gotovostjo $(1 - \alpha)$ nahajal ocenjevani parameter. Natančno povedano, nam ta gotovost pove, da se bo z verjetnostjo $(1 - \alpha)$ v intervalu zaupanja, ki ga z znanim postopkom zgradimo z vzorčno statistiko nekega naključnega slučajnega vzorca, nahajala dejanska vrednost parametra γ , ki ga želimo oceniti.

3. Na vzorcu velikosti $n = 100$ podjetnikov v majhnih podjetjih v Sloveniji, ki je bil izveden v okviru ankete 'Drobno gospodarstvo v Sloveniji', so izračunali, da je povprečna starost anketiranih podjetnikov $\bar{X} = 40.4$ let in standardni odklon $s = 10.2$ let. Pri 5% tveganju želimo z intervalom zaupanja oceniti povprečno starost podjetnikov v majhnih podjetjih v Sloveniji.

rešitev: Konstruirajmo dvostrani interval zaupanja po znanem postopku. Ker je velikost vzorca velika, lahko uporabimo z-statistiko.

$$I = 40.4 \pm 1.96 \frac{10.2}{\sqrt{1000}} \equiv (39.7678, 41.0322)$$

Interpretacija rezultata: Trdimo lahko, da bo 95% intervalov, ki jih zgradimo po tej metodi iz slučajnih naključnih vzorcev, vsebovalo pravilno vrednost populacijskega parametra. Torej velja, da lahko z gotovostjo 95% trdimo, da se resnični parameter populacije, ki nas zanima (v našem primeru populacijsko povprečje) nahaja na izračunanem intervalu.

4. Opiši ocenjevanje parametrov z majhnimi vzorci (čim več možnosti).

rešitev: Zanimajo nas 3 ključne lastnosti:

- Ali poznamo standardni odklon populacije?
- Ali je vzorec velik?

(a) **DA/DA**

Uporabimo z statistiko. Standardni odklon vzorčne statistike ocenimo z $\frac{\sigma}{\sqrt{n}}$

(b) **DA/NE**

Uporabimo t -statistiko. Standardni odklon vzorčne statistike ocenimo z $\frac{\sigma}{\sqrt{n}}$

(c) **NE/DA**

Če je vzorec velik (≥ 30), je standardni odklon vzorca dobra ocena populacijskega standardnega odklona. Uporabimo lahko z -statistiko, kjer standardni odklon porazdelitve vzorčne statistike izračunamo kot $\frac{s}{\sqrt{n}}$, kjer je s standardni odklon vzorca.

(d) **NE/NE** Ker ne poznamo standardnega odklona populacije in ker imamo opravka z majhnim vzorcem, bomo uporabili t -statistiko. Standardni odklon vzorčne statistike bomo ocenili kot $\frac{s}{\sqrt{n}}$.

5. Podaj osnovni izrek statistike.

rešitev:

Spremenljivka X ima na populaciji porazdelitev $F(x) = P(X \leq x)$. Toda tudi vsakemu vzorcu ustreza neka porazdelitev. Za realizacijo vzorca (x_1, \dots, x_n) in $x \in R$ postavimo

$$K(x) = |\{x_i \mid x_i < x, i = 1, \dots, n\}|$$

in

$$V_n(x) = \frac{K(x)}{n}$$

(komentar: $K(x)$ je funkcija, ki vrne število realizacij slučajne spremenljivke X , katerih vrednost je manjša od x . $V_n(x)$ pa je funkcija, ki vrne delež realizacij, ki so manjše od x .)

Vzorčna porazdelitvena funkcija ima tudi tako kot $K(x)$, $n + 1$ možnih vrednosti.

Torej je njena verjetnostna funkcija $B(n, F(x))$.

(Komentar: Ta funkcija je porazdeljena binomsko, kjer je verjetnost P enaka verjetnosti, da je $X \leq x$. To verjetnost vrača funkcija F)

Če vzamemo n neodvisnih Bernoullijevih (indikatorskih) spremenljivk

$$Y_i(x) \sim \begin{pmatrix} 1 & 0 \\ F(x) & 1 - F(x) \end{pmatrix}$$

Potem velja

$$V_n(x) = \frac{1}{n} \sum_{i=1}^n Y_i(x)$$

(Komentar: i -ta indikatorska slučajna spremenljivka Y_i zavzame vrednost 1, če je $X \leq x$. To stori z verjetnostjo, ki jo vrača funkcija F . Torej velja, da je $V_n(x)$ enaka pričakovani vrednosti te slučajne spremenljivke deljeni z velikostjo vzorca.)

Krepki zakon števil nam tedaj zagotavlja, da za vsak x velja

$$P(\lim_{x \rightarrow \infty} V_n(x) = F(x)) = 1$$

(komentar: V limiti, ko n raste prek vseh mej, se vrednost funkcije $V_n(x)$ približuje funkciji $F(x)$. $V_n(x)$ si lahko predstavljamo kot porazdelitveno funkcijo za vzorec. Ko velikost vzorca raste, odstopanje te funkcije od porazdelitvene funkcije populacije pada.)

Odstopanje med $V_n(X)$ in $F(x)$ lahko izmerimo s slučajno spremenljivko

$$D_n = \sup_{x \in R} |V_n(x) - F(x)|.$$

Velja $P(\lim_{n \rightarrow \infty} D_n = 0) = 1$

(komentar: razkorak med $V_n(x)$ in $F(x)$ z naraščanjem n pada)

22 Preverjanje domnev

1. Opiši splošni postopek preverjanja domneve.

rešitev:

Splošni postopek preverjanja domneve:

- (1) Postavi domnevo o parametrih: ničelno H_0 in osnovno/alternativno H_1/H_a .
- (2) Za parameter poiščemo kar se da dobro cenilko (npr. nepristransko) in njeno porazdelitev ali porazdelitev ustrezne statistike (izraz, v katerem nastopa cenilka).
- (3) Določi odločitveno pravilo: izberemo stopnjo značilnosti α in na osnovi nje ter porazdelitve statistike določimo kritično območje;
- (4) Zberi/manipuliraj podatke ter na vzorčnih podatkih izračunaj (eksperimentalno) vrednost testne statistike.
- (5) Primerjaj in naredi zaključek: če kritično območje eksperimentalno vrednost
 - (A) vsebuje, ničelno domnevo zavrne in sprejmi osnovno domnevo ob stopnji značilnosti α ;
 - (B) ne vsebuje, pa pravimo da vzorčni podatki kažejo na statistično neznačilne razlike med parametrom in vzorčno oceno.

2. Kaj je to zavrnitveni kriterij?

rešitev: Izberemo stopnjo tveganja α (občajno 10%, 5% ali 1%). Glede na to, ali gre za enostranski ali dvostranski test določimo kritično območje: enostranski test $(-\infty, -z_\alpha)$ ali (z_α, ∞) ter dvostranski test $(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$

Ničelno hipotezo lahko zavrnemo, če pade verjetnost, da bi, če bi veljala ničelna hipoteza dobili takšno vzorčno statistiko, v kritično območje. Torej če je verjetnost, da se to zgodi manj kot stopnja tveganja α .

3. Kaj je stopnja značilnosti?

rešitev:

Stopnja značilnosti testa (signifikantnosti) je največji α , ki ga je vodja eksperimenta pripravljen sprejeti (zgornja meja za napako 1. vrste). Torej največja meja verjetnosti, ko bomo, če bomo dobili vzorčno statistiko, ki bi se v primeru, da bi ničelna hipoteza držala zgodila s to verjetnostjo, še sprejeli ničelno hipotezo. P-vrednost (ali ugotovljena bistvena stopnja za določen statistični test) je verjetnost (ob predpostavki, da drži H_0), da ugotovimo vrednost testne statistike, ki je vsaj toliko v protislovju s H_0 in podpira H_a kot tisto, ki je izračunana iz vzorčnih podatkov. Razlaga P-vrednosti:

- Izberi največjo vrednost za α , ki smo jo pripravljeni tolerirati.
- Če je P-vrednost testa manjša kot maksimalna vrednost parametra α , potem zavrne ničelno hipotezo.

4. Pojasni razliko med napako 1. in 2. vrste

rešitev: Napako prve vrste napravimo, če zavrnemo ničelno hipotezo, čeprav ta v resnici drži. Verjetnost, da se to zgodi je enaka stopnji tveganja α . Torej je verjetnost te napake sorazmerna z velikostjo α , ki si jo izberemo.

Napako druge vrste pa napravimo, če ne zavrnemo ničelne hipoteze, čeprav ta v resnici ne drži. Verjetnost, da se to zgodi označimo z β . Če povečamo α , se verjetnost, da bomo napravili napako drugega tipa zmanjša.

5. Slučajna spremenljivka je porazdeljena normalno z $\sigma = 0.3$. Vzorec je:

[5.93 6.08 5.86 5.91 6.12]

Testiraj hipotezo, da je $\mu = 6.15$ proti $\mu < 6.15$ z $\alpha = 0.05$.

rešitev: Najprej izračunamo vzorčno povprečje.

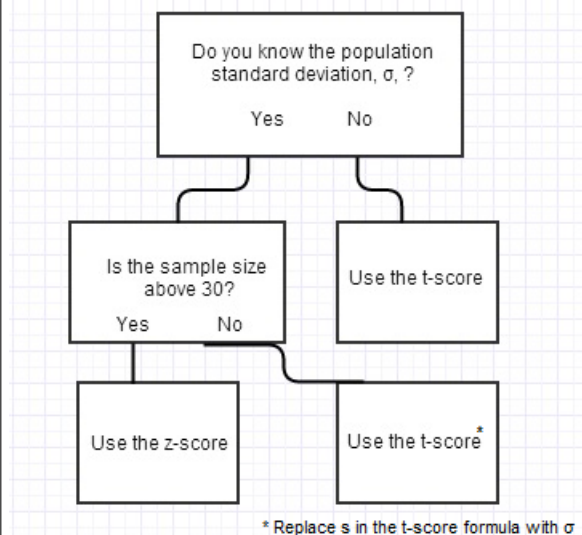
Vzorčno povprečje izračunamo na naslednji način:

$$\mu_X = \frac{\sum x_i}{n} = 5.98$$

Standardni odklon populacije je poznan. Standardni odklon vzorčnih povprečij iz njega ocenimo tako, da ga delimo s kvadratnim korenom velikosti vzorcev.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{5}}$$

Poiščimo kritično vrednost za t -statistiko pri stopnji tveganja 0.05. Pogledamo v tabelo in dobimo vrednost $t = 2.132$. To pomeni, da če dobimo vrednost testne statistike, za katero je verjetnost, da bi jo, če bi ničelna hipoteza držala, dobili po naključju, manjša od 5%, bo padla standardizirana vrednost v kritično območje, ki je manjše ali enako tej vrednosti. Takrat lahko z neko mero gotovosti trdimo, da je povprečje res manjše, kot ga postaja ničelna hipoteza.



Računajmo interval zaupanja.

$$I = 5.98 - 2.132 \cdot \frac{0.3}{\sqrt{5}} \equiv (5.69396, \infty)$$

Vidimo, da naše vzorčno povprečje ne pade v kritično območje. Ničelne hipoteze tako ne moremo zavrniti.

6. Slučajna spremenljivka je porazdeljena normalno z $\sigma = 6$. Vzorec je

[137.4 140.1 134.9 135.9 140.9 138.4 136.5 137.6 140.3]

Testiraj hipotezo, da je $\mu = 142$ proti $\mu < 142$ z $\alpha = 0.01$.

rešitev:

Najprej izračunajmo vzorčno povprečje. Skupaj z informacijama velikost vzorca ter populacijski standardni odklon, ki ju poznamo, bomo lahko nato opredelili porazdelitveni zakon vzorčnih povprečij in preverili ničelno hipotezo.

$$\mu = \frac{\sum x_i}{n} = 138$$

Ker je vzorec majhen (9), uporabimo t-statistiko z 8 stopnjami prostosti.

Računamo interval zaupanja:

$$I = 138 - 1.860 \cdot \frac{6}{\sqrt{9}} \equiv (134.28, \infty)$$

Vzorčno povprečje, ki smo ga izračunali, ne pade v kritično območje. Ničelne hipoteze torej ne moremo zavrniti.

23 Regresija

1. Definiraj (Pearsonov) koeficient korelacije za dve številski spremenljivki.

rešitev: Pearsonov korelacijski koeficient je merilo linearne korelacije med dvema spremenljivkama X in Y . Zavzame lahko vrednosti na intervalu $[-1, 1]$, kjer vrednost 1 predstavlja popolno pozitivno linearno korelacijo, Vrednost -1 pa popolno negativno linearno korelacijo. Vrednost 0 pa predstavlja odsotnost linearne korelacije.

Definiran je z izrazom:

$$r(X, Y) = \frac{K(X, Y)}{\sigma_X \sigma_Y}$$

Kot smo že omenili velja, da če je $|r(X, Y)| = 1$, obstaja med X in Y linearna zveza z verjetnostjo 1.

2. Kakšne vrednosti lahko zavzame in kaj lahko poveš v primeru, ko doseže največjo oziroma najmanjšo možno vrednost?

rešitev: Odgovorjeno v prejšnjem vprašanju.

3. Kaj sta prva in druga regresijska funkcija in kje se srečata?

rešitev: Naj je $Y = \alpha + \beta X$ linearna funkcija, ki smo si jo izbrali za naš model. Ker ne vemo kakšna sta α in β ju moramo nekako oceniti iz vzorca.

Premica se bo najboljše prilegala našim podatkom, če bo minimizirala vsoto kvadratov razdalj (odklonov) točk od premice. Želimo poiskati takšni vrednosti α in β , da bo premica zadostovala tej lastnosti.

Z nekaj računanja ugotovimo, da sta najboljši cenilki za parametra α in β

$$B = \frac{C_{xy}}{C_x^2}$$

za smerni koeficient β in

$$A = \bar{Y} - B\bar{X}$$

za začetno vrednost α .

Brez poznavanja porazdelitve Y in U lahko pokažemo, da velja

$$E(A) = \alpha \text{ in } D(A) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{C_x^2}\right), E(B) = \beta \text{ in } D(B) = \frac{\sigma^2}{C_x^2}, K(A, B) = -\sigma \frac{\bar{X}}{C_x^2}.$$

Izkaže se, da sta A in B najboljši linearni nepristranski cenilki za α in β .

Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

$$Y = \mu_Y + \frac{K(X, Y)}{\sigma_x^2}(X - \mu_x)$$

To funkcijo imenujemo tudi **prva regresijska premica**.

Podobno lahko ocenimo linearno regresijsko funkcijo

$$X = a * + b * Y$$

Če z metodo najmanjših kvadratov podobno ocenimo parametra $a*$ in $b*$, dobimo:

$$X = \mu_X + \frac{K(X, Y)}{\sigma_Y^2}(Y - \mu_Y).$$

4. Vzemimo spremenljivki X - število ur gledanja TV na teden in Y - število obiskov konopredstav na mesec. Podatki za 6 oseb so

X	10	15	6	7	20	8
Y	2	1	2	4	1	2

Z linearno regresijsko funkcijo ocenimo, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden TV.

rešitev:

Izračunamo regresijsko premico za odzivno s.s. Y , ki se podatkom najboljše prilega po metodi najmanjših kvadratov.

Izračunajmo μ_x , $K(X, Y)$, μ_Y in σ_x^2 .

Dobimo premico

$$Y = 2 + \frac{-4.2}{29.6}(X - 11)$$

$$Y = 3.561 - 0.142X$$

Če vstavimo v premico za X vrednost 18, dobimo rezultat 1.

5. Kaj so to časovne vrste? Opiši določanje trenda z metodo najmanjših kvadratov.

rešitev: Časovna vrsta je niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke. Osnovni namen časovnih vrst je:

- opazovati časovni razvoj pojavov,
- iskati njihove zakonitosti,
- predvidevati nadaljni razvoj.

Trend lahko obravnavamo kot poseben primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend $X_T = f(T)$, lahko parametre trenda določimo z metodo najmanjših kvadratov $\sigma_{i=1}^n (X_i - X_{iT})^2 = \min$. V primeru linearnega trenda:

$$X_T = a + bT, \quad \sigma_{i=1}^n (X_i - a - bT_i)^2 = \min$$

dobimo naslednjo oceno trenda:

$$X_T = \bar{X} + \frac{\sigma_{i=1}^n (X_i - \bar{X})(T_i - \bar{T})}{\sigma_{i=1}^n (T_i^2)} T.$$

Ponavadi je čas T transformiran tako, da je $\bar{T} = 0$. Tedaj je ocena trenda

$$X_T = \bar{X} + \frac{\sigma_{i=1}^n (X_i - \bar{X})(T_i - \bar{T})}{\sigma_{i=1}^n T_i^2} T$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je

$$\sigma_e = \sqrt{\frac{1}{n} \sigma_{i=1}^n (X_i - X_{iT})^2},$$

kjer je X_{iT} enak X_T v času T_i .