Verjetnost in statistika z nalogami

Maja Pohar Perme

Kazalo

Uvod

Statistika je veda z zelo raznolikimi področji uporabe, vsako področje zaznamuje svoja narava podatkov in z njimi povezani zapleti. Na prvi pogled statistika lahko deluje le kot nepregleden skupek formul in receptov, ključ do njenega razumevanja se skriva v razumevanju skupnih temeljev statističnega sklepanja, ki so osnova vsem metodam in pristopom.

Knjig na temo statističnega sklepanja je veliko, nekatere se skozi osnovne ideje statistike uspejo prebiti skorajda brez formul, druge so precej bolj teoretično zahtevne in za razumevanje potrebujejo veliko matematičnega znanja. Knjiga, ki jo imate v rokah, izbira srednjo pot, za razumevanje ni potrebno znanje teorije mere, še več, večino primerov je moč rešiti z rahlo obogatenim znanjem srednješolske matematike. Navkljub temu je izpeljav veliko in marsikateremu nematematiku se bo snov vsaj na začetku zdela le težko razumljiva.

V nasprotju z večino knjig, ki so namenjene podajanju, izpeljevanju in dokazovanju teorije in so nujen spremljevalec tega gradiva, želi ta knjiga študenta voditi skozi osnovne zakonitosti statistike s pomočjo podrobno rešenih primerov. Knjiga torej ni mišljena kot samostojno gradivo, ki podaja vse definicije in izreke, temveč bolj kot spremljevalec, ki študenta vodi po osnovah teoretične statistike in mu pomaga razumeti posamezne korake izpeljav. Predvsem pa snovi skuša dajati poudarke in študenta opominjati na to, zakaj so na videz zelo teoretične lastnosti tako pomembne za praktično uporabo statistike.

Jedro knjige predstavlja 26 nalog, ki s svojimi rešitvami študentu odkrivajo osnovne zakonitosti in ideje statistike. Pomemben delež nalog najprej zajame nujne osnove verjetnosti, nato sledijo povsem statistične teme. Vsako poglavje se prične s kratkim uvodom, ki pomaga umestiti naloge v snov, vsem rešitvam so dodani povzetki, ki študenta še enkrat opozorijo na bistvene poudarke posamezne naloge. Primeri so namenoma izbrani z različnih področij

in želijo študenta ves čas opominjati tako na široko uporabnost osnovnih statističnih znanj kot tudi na možnost, da za nek problem, na katerega naletijo v okviru podatkov s področja financ, prav lahko že obstaja rešitev na nekem povsem drugem področju, na primer v biostatistiki.
Skoraj vse naloge v besedilu so povsem originalne, a ker so zaradi jasnosti izpeljav in razumevanja izbrani dokaj osnovni primeri, je seveda marsikatero zelo podobno nalogo moč srečati tudi v drugih statističnih knjigah. Podane rešitve nalog niso edine možne in pravilne, študent je vedno vabljen, da najprej poišče svojo rešitev, nato pa s pomočjo danih rešitev razume tudi alternativne možnosti, predvsem pa glavne poudarke, ki iz rešitve sledijo. Za bolj samostojno reševanje pred vpogledom v rešitve so nekaterim nalogam dodani tudi namigi.
Knjiga je namenjena tako matematikom kot tudi tistim z mnogo manj matematičnega znanja. Za slednje prebijanje iz vrstice v vrstico pogosto povzroča precej frustracij, zato je na mestu opozorilo, da naj se ne obremenjujejo preveč s tehničnimi podrobnostmi, temveč skušajo raje razumeti splošne cilje in ugotovitve posamezne naloge. To opozorilo naj pride prav tudi tistim, ki se jim teorija zdi sorazmerno lahka - za pravo razumevanje statistike ne zadostuje le podrobno razumevanje izpeljav, nujno je razumeti, kaj neki dokaz oziroma lastnost pomeni za rešitev nekega realnega problema.
Knjiga po organizaciji snovi sledi knjigi Mathematical Statistics and Data Analysis (Rice, 2009), ki je zaradi podobne matematične zahtevnosti in sosledja snovi vsekakor priporočen vir teorije pri reševanju nalog. Za lažjo orientacijo so pri posameznih snoveh navedena tudi poglavja te knjige, kjer je predstavljena potrebna teorija. Prav tako knjiga vsebuje veliko referenc na dodatno literaturo, s pomočjo katere študenti lahko poglobijo posamezne snovi. Zelo koristna vira sta na primer tudi Casella and Berger (1990) in Shao (2003), vendar je predvsem slednji precej bolj matematično zahteven.
Pomembna lastnost statistične teorije, ki izhaja iz realnih problemov, je, da si lahko vse zakonitosti ogledamo tudi v praksi, z računalniškimi simulacijami. Za resnično razumevanje pojmov, kot so porazdelitev, varianca, nepristranskost, standardna napaka, ipd., je zato neprecenljiva tudi zmožnost, da si jih študent prikaže s pomočjo ustreznega računalniškega programa. Skoraj vsem nalogam so na koncu zato dodani še predlogi, kako teorijo spoznavati s konkretnimi primeri v ustreznem statističnem programu. Za primere je izbran R (R Development Core Team, 2013), statistični program, ki predstavlja eno najbolj uporabljanih orodij med statistiki. Seveda je vse tovrstne naloge mogoče reševati tudi s poljubnim drugim programskim orodjem, ki

omogoča generiranje podatkov. Rešitve teh nalog so prepuščene študentu, dodani so le kosi kode, ki naj bi študentu prihranili nekaj časa. Tako je pri prvih nalogah dodanih nekaj zelo osnovnih nasvetov in komentarjev kode, pri kasnejših nalogah pa je koda podana le, kadar je nekoliko zahtevnejša.

Poglavje 1

Verjetnost

Eden osnovnih ciljev statistične analize je, da na podlagi podatkov, zbranih na nekem, ponavadi sorazmerno majhnem vzorcu, skušamo sklepati o splošnih zakonitostih, ki veljajo v celotni populaciji. Zaradi naključne variabilnosti je na podlagi vzorca o splošnih zakonitostih seveda nemogoče karkoli trditi z gotovostjo, statistika zaupanje v sklepe o populaciji ovrednoti z verjetnostjo. Verjetnost je torej osnovno orodje statističnega sklepanja in prvo poglavje je namenjeno spoznavanju osnovnih pojmov in postopkov, ki jih bomo potrebovali v kasnejših, bolj 'statističnih' delih knjige.

Če želimo govoriti o verjetnosti posameznih dogodkov, moramo poznati porazdelitev slučajne spremenljivke, ki opisuje možne izide. To poglavje se ukvarja z več porazdelitvami, ki se v statistiki pogosto pojavljajo (glej npr. Rice, 2009, razdelka 2.1 in 2.2), in raziskuje njihove osnovne lastnosti. Jedro poglavja predstavljajo štiri teme: izpeljava gostote transformirane spremenljivke (Rice, 2009, razdelek 2.3), formula za vsoto slučajnih spremenljivk (Rice, 2009, poglavje 3), pričakovana vrednost in varianca (Rice, 2009, poglavje 4) ter centralni limitni izrek (Rice, 2009, poglavje 5). Prva naloga raziskuje lastnosti normalne porazdelitve, ki je prav gotovo najpomembnejši primer porazdelitve v statistični teoriji, ter pokaže, da s poljubno linearno transformacijo normalne spremenljivke spet dobimo normalno porazdelitev. Sledi naloga, ki predstavi enakomerno porazdelitev in z njo povezan pomemben statistični izrek ('transformacija integrala verjetnosti'), ki je ključen pri generiranju slučajnih spremenljivk s poljubno drugo porazdelitvijo.
V tretji nalogi izpeljemo formulo za vsoto dveh diskretnih slučajnih spremenljivk ter jo skušamo razumeti na preprostem primeru. Nato v četrti nalogi

pokažemo, da lahko na povsem enak način razumemo tudi formulo za zvezni primer. V peti nalogi formulo za vsoto uporabimo na primeru normalne spremenljivke in pokažemo, da tudi s seštevanjem neodvisnih normalnih spremenljivk ostajamo pri normalni porazdelitvi.
Poglavje se v šesti in sedmi nalogi nadaljuje s spoznavanjem lastnosti pričakovane vrednosti, variance in kovariance, pri tem je izpostavljeno predvsem razumevanje formul za njihove pogojne vrednosti. Uporabnost teh spoznanj je nato prikazana pri izpeljavi lastnosti na videz povsem smiselnega načina uvrščanja, za katerega se izkaže, da deluje drugače, kot bi si želeli.
Poglavje se zaključi s centralnim limitnim izrekom, ki bo igral bistveno vlogo v zakulisju marsikatere metode v nadaljevanju - za kakršnokoli statistično sklepanje bomo morali poznati porazdelitev, kadar ne bo znana oziroma bo njena izpeljava prezahtevna, jo bomo s pomočjo centralnega limitnega izreka skušali aproksimirati z normalno.

1.1 Normalna porazdelitev

Krvni doping je metoda, pri kateri si športnik kri najprej odvzame, nato pa si jo včrpa pred pomembnim nastopom in tako umetno poveča število rdečih krvničk ter si s tem izboljša trenutno počutje in vzdržljivost. Ker ni udeleženih tujih substanc, krvnega dopinga ni mogoče neposredno odkriti. Zato ga skušajo odkrivati s statističnimi metodami - doping naj bi nakazovale vrednosti krvnih parametrov (hemoglobina), ki pretirano narastejo (včrpanje) oz. padejo (odvzem).
Vemo, da je vrednost hemoglobina pri nedopingiranem športniku porazdeljena normalno s povprečjem $\mu = 148$ in varianco $\sigma^2 = 85$. Označimo vrednost hemoglobina z X, torej $X \sim N(148,85)$.

a) Izračunajte verjetnost, da je posameznikova vrednost večja od 166. V ta namen izpeljite formulo:

   - Naj bo $X \sim N(\mu, \sigma^2)$, kako je porazdeljena porazdeljena slučajna spremenljivka $Y = aX + b$, kjer je $a > 0$?
     Namig: Zapišite najprej porazdelitveno funkcijo, nato izrazite gostoto. Ali lahko gostoto zapišete kot gostoto normalne spremenljivke?

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) = P(aX \leq y - b) \\ &= P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a}) \\ f_Y(y) &= \frac{1}{a} f_X(\frac{y-b}{a}) \end{aligned}$$

Za normalno porazdeljeno X torej velja:

$$\begin{aligned} f_Y(y) &= \frac{1}{a \cdot \sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[\frac{y-b}{a} - \mu]^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sqrt{2\pi(a \cdot \sigma)^2}} \exp\left\{ -\frac{[y - (b + a\mu)]^2}{2(a \cdot \sigma)^2} \right\} \end{aligned}$$

Torej, $Y \sim N(a \cdot \mu + b, (a \cdot \sigma)^2)$. Linearna transformacija normalne spremenljivke je še vedno normalna.

- Kaj moramo vzeti kot a in b, da bo Y standardno normalno porazdeljena spremenljivka, torej $Y \sim N(0,1)$?
  a mora biti enak $\frac{1}{\sigma}$, b pa $-\frac{\mu}{\sigma}$. Uporabiti moramo torej transformacijo $Y = \frac{X-\mu}{\sigma}$ in nato verjetnosti odčitati iz tabel za standardno normalno porazdelitev (oz. uporabiti ustrezno numerično metodo). V našem primeru je $X = 166$ in zato $Y = \frac{X-148}{\sqrt{85}} = \frac{166-148}{\sqrt{85}} = 1{,}95$. Iz tabel za standardno normalno porazdelitev (ali pa s pomočjo računalnika) izvemo, da je $P(X \leq 166) = P(Y \leq 1{,}95) = 0{,}974$, zato je verjetnost $P(X > 166) = 0{,}026$.

b) Izračunajte (simetrične) meje, ki jih nedopingiran športnik preseže z verjetnostjo manj kot 0,01.

Naj bo Y standardno normalno porazdeljena spremenljivka, zanimajo nas meje, izven katerih je vrednost te spremenljivke z verjetnostjo 0,01. Če želimo postaviti simetrične meje, pomeni, da nas zanimata tisti vrednosti, izven katerih je v repih na vsaki strani verjetnost 0,005. Iz tabel izvemo, da je $P(Y \geq 2{,}58) = 0{,}995$, ustrezna mejna vrednost standardno normalno porazdeljene spremenljivke je torej $\pm 2{,}58$.

$Y = \frac{X-148}{\sqrt{85}}$, zato

$$\begin{aligned} 0{,}99 &= P(\frac{X-148}{\sqrt{85}} \leq 2{,}58) + P(\frac{X-148}{\sqrt{85}} > -2{,}58) \\ &= P(X \leq 148 + 2{,}58 \cdot \sqrt{85}) + P(X > 148 - 2{,}58 \cdot \sqrt{85}) \\ &= P(X \leq 171{,}8) + P(X > 124{,}2) \end{aligned}$$

c) Naj bodo meje take, kot ste jih izračunali v prejšnji točki. Športnika testiramo 10x na leto. Kolikšna je verjetnost, da vsaj enkrat preseže meje (pri tem predpostavimo, da so meritve narejene v dovolj velikih časovnih presledkih, da so med seboj neodvisne)?

Naj bo U Bernoullijevo porazdeljena spremenljivka $U \sim Ber(0{,}01)$, kjer je $\{U = 1\} = \{\text{vrednost je izven meja}\}$. Imamo 10 neodvisnih realizacij te slučajne spremenljivke, $U_i,\ i = 1, \ldots, 10$, za vsako velja $P(U_i = 1) = 0{,}01$. Ker so neodvisne, velja $P(U_1 = 0, U_2 = 0, \ldots, U_{10} = 0) = \{P(U_1 = 0)\}^{10}$. Verjetnost, da v 10 meritvah ne preseže meja, je $0{,}99^{10}$, verjetnost, da jih preseže vsaj enkrat, je $P = 1 - 0{,}99^{10} = 0{,}096$.

d) Naj bo $Y \sim N(0{,}1)$. Izračunajte porazdelitev slučajne spremenljivke Y^2. Katero znano porazdelitev dobite?
Namig: Porazdelitev gama ima gostoto $f_T(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$.

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(Y^2 \leq z) = P(-\sqrt{z} \leq Y \leq \sqrt{z}) \\ &= F_Y(\sqrt{z}) - F_Y(-\sqrt{z}) \\ f_Z(z) &= \frac{1}{2\sqrt{z}} f_Y(\sqrt{z}) + \frac{1}{2\sqrt{z}} f_Y(-\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}} [f_Y(\sqrt{z}) + f_Y(-\sqrt{z})] \\ &= \frac{1}{2\sqrt{z \cdot 2\pi}} [e^{-z/2} + e^{-z/2}] = \frac{1}{\sqrt{z \cdot 2\pi}} e^{-z/2} \end{aligned}$$

Gama porazdelitev ima gostoto $f_T(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$. Če vzamemo, da je $\alpha = \frac{1}{2}$ in $\lambda = \frac{1}{2}$, ter upoštevamo, da je $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, dobimo natanko gornjo formulo. Torej je $Y^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ (to je hkrati tudi porazdelitev χ^2_1).

Opombi:

- Pri zveznih porazdelitvah je verjetnost, da spremenljivka zavzame točno določeno vrednost $P(Y = y)$, neskončno majhna, zato je vseeno, ali v porazdelitveni funkciji uporabljamo znak $\leq$ ali $<$, torej $P(Y \leq y) = P(Y < y)$.
- Kadar želimo pokazati, da lahko gostoto spremenljivke zapišemo kot gostoto neke že znane porazdelitve, je dovolj pokazati enako funkcijsko obliko argumentov. Konstanta bo potem vedno enaka, saj je integral gostote po celem definicijskem območju vedno enak 1.

e) Raziskovalci na področju športa so dokazali, da je pri biatloncih hemoglobin izven tekmovalnega obdobja porazdeljen kot $N(150, 80)$, med tekmovalnim obdobjem pa kot $N(146, 80)$. Tekmovalno obdobje je pri teh športnikih dolgo približno pol leta. Zanima nas porazdelitev hemoglobina, če ne vemo, kdaj je bil vzorec odvzet. Ali je ta porazdelitev še vedno normalna?
Namig: $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$, kadar velja $P(\bigcap_{i=1}^{n} B_i) = 0$ in $P(\bigcup_{i=1}^{n} B_i) = 1$.

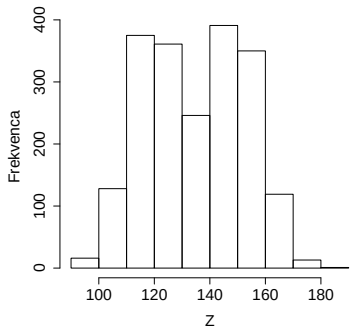
Definiramo Bernoullijevo porazdeljeno spremenljivko Y, ki naj označuje obdobje (0 = izven, 1 = tekme, verjetnost vsakega izida je 0,5). Poznamo pogojni porazdelitvi:
$Z|Y = 0 \sim N(150, 80)$, $Z|Y = 1 \sim N(146, 80)$. Porazdelitev Z je torej (uporabimo namig, kjer je $B_1 = \{Y = 0\}$ in $B_2 = \{Y = 1\}$, namesto z verjetnostmi pišemo z gostotami)

$$\begin{aligned} f_Z(z) &= f_{Z|Y=0}(z)P(Y = 0) + f_{Z|Y=1}(z)P(Y = 1) \\ &= f_{Z|Y=0}(z)\frac{1}{2} + f_{Z|Y=1}(z)\frac{1}{2} \\ &= \frac{1}{2\sqrt{2\pi 80}} e^{-\frac{(z-146)^2}{2\cdot 80}} [1 + e^{-\frac{16-8(z-146)}{2\cdot 80}}] \end{aligned}$$

Ta spremenljivka v splošnem ni normalno porazdeljena, glej sliko 1.1.

Predlogi za vaje v R

- Generirajte 10000 realizacij normalno porazdeljene spremenljivke $X \sim N(148{,}85)$ (`rnorm`). Narišite histogram (`hist`), izračunajte delež vrednosti nad 166 (`sum(x>166)/10000`).



Slika 1.1: *Porazdelitev spremenljivke Z (za ta primer je za povprečje v tekmovalnem obdobju vzeta vrednost 120, da bi bila porazdelitev bolj različna od normalne).*

- Oglejte si funkcijo `qnorm` in z njo poiščite meje, izven katerih je športnik z verjetnostjo 0,01. Primerjajte z deležem v vašem primeru.
- Transformirajte vrednosti spremenljivke X tako, da dobite standardno normalno porazdeljeno spremenljivko (`y=(x-148)/sqrt(85)`). Preverite grafično s histogramom.
- Generirajte po 10 vrednosti za 10000 posameznikov. Izračunajte delež posameznikov, ki imajo vsaj eno vrednost izven intervala $[124{,}2, 171{,}8]$.
- Narišite histogram za vrednosti X^2, primerjajte z rezultatoma funkcij `rgamma` in `rchisq`.
- Narišite še porazdelitev slučajne spremenljivke iz zadnje točke (uporabite bolj različni povprečji, da se prepričate, da porazdelitev zares ni normalna).

```
> set.seed(1)          #dolocimo seme, da bodo rezultati vedno enaki
> z0 <- rnorm(1000,mean=150,sd=sqrt(80)) #1000 vredn.izven tek.obdobja
> z1 <- rnorm(1000,mean=120,sd=sqrt(80)) #1000 vredn. v tek.obdobju
> z <- c(z0,z1)    #obe obdobji enako pogosto zastopani v novi spr.
> hist(z,main="",xlab="Z",ylab="Frekvenca")    #prikaz porazdelitve
```

Povzetek

- Pomembna lastnost, ki precej olajša računanje z normalno porazdelitvijo, je, da je linearna transformacija normalne spremenljivke zopet normalno porazdeljena. To je tudi razlog, zakaj je dovolj, da so statistične tabele vedno podane le za 'standardno normalno porazdelitev´ - z njeno pomočjo lahko namreč preprosto izračunamo verjetnosti za normalno porazdeljeno spremenljivko s poljubnima vrednostima parametrov. Seveda pa navkljub tej lepi lastnosti ne smemo pričakovati, da bo kar vsaka transformacija normalne porazdelitve ostala normalna - protiprimer je podan v točki (e) te naloge.

- V nalogi je prikazan postopek izpeljevanja porazdelitve spremenljivke Y, ki je definirana kot neka transformacija spremenljivke X z znano porazdelitvijo. Kumulativno porazdelitveno funkcijo nove spremenljivke F_Y najprej izrazimo z znano kumulativno porazdelitveno funkcijo F_X, nato pa uporabimo, da je gostota f_Y odvod funkcije F_Y.

- Točka (c) je dodana kot opozorilo, da so vse izračunane verjetnosti veljavne le pri enem preizkusu. Čim bomo nek preizkus uporabili večkrat, ne smemo pozabiti na problem večkratnega preizkušanja , torej dejstva, da se verjetnosti pri večjem številu opravljenih preizkusov močno spremenijo.

- Točka (e) je v to nalogo umeščena predvsem kot vaja kako formalno zapisati nek besedni opis porazdelitve. Dani zapis pri tem seveda ni edini možen, je pa iz njega jasno vidno, da ne gre za normalno porazdelitev. V tej točki smo uporabili pogojno porazdelitev, ki je v knjigi Rice (2009) predstavljena v razdelku 3.5.

- V nalogi smo izpeljali še rezultat, da je kvadrat standardno normalno porazdeljene spremenljivke spremenljivka porazdeljena z gama porazdelitvijo:

$$X \sim N(0,1) \;\Rightarrow\; X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2}).$$

  Gama porazdelitev s tema parametroma je hkrati enaka tudi porazdelitvi χ^2 z eno stopinjo prostosti.

1.2 Generiranje slučajnih spremenljivk s pomočjo enakomerne porazdelitve

Generator (psevdo)slučajnih vrednosti iz enakomerne spremenljivke zgenerira željeno število vrednosti x_i, ki so porazdeljene kot $X \sim U[0,1]$.

a) Kako bi s pomočjo tega generatorja dobili 10 realizacij Bernoullijevo porazdeljene spremenljivke Y, pri kateri je $P(Y = 1) = 0{,}1$?

Generirajmo 10 vrednosti, npr.:

```
> set.seed(4)
> runif(10)
[1]  0.585800305 0.008945796 0.293739612 0.277374958
[5]  0.813574215 0.260427771 0.724405893 0.906092151
[9]  0.949040221 0.073144469
```

Vrednostim, ki so pod 0,1, damo vrednost 1, ostalim pa 0, torej:

```
> set.seed(4)
> (runif(10)<0.1)*1
[1] 0 1 0 0 0 0 0 0 0 1
```

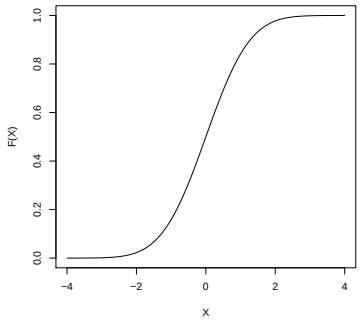
b) Recimo, da imamo spet 10 enot, vendar jim želimo dati različne verjetnosti, da bodo izžrebane. Prvih pet enot želimo izžrebati z verjetnostjo 0,3, drugih pet pa z verjetnostjo 0,1 (kot primer si zamislimo žreb, v katerem želimo dati prednost ženskam. Verjetnost za vsakega posameznika v našem vzorcu določimo glede na spol - prvih pet je žensk, drugih pet je moških). Kako bi iz istim generatorjem zagotovili ustrezno porazdelitev?

```
> set.seed(4)
> (runif(10)<c(0.1,0.1,0.1,0.1,0.1,0.3,0.3,0.3,0.3,0.3))*1
 [1] 0 1 0 0 0 1 0 0 0 1
```

c) Naj bo $Z = F(X)$, kjer je F porazdelitvena funkcija slučajne spremenljivke X.

- Narišite ustrezen graf (na abscisi so vrednosti X, na ordinati pa Z).
- Kakšne vrednosti lahko zavzame spremenljivka Z?

  Med 0 in 1



Slika 1.2: *Kumulativna porazdelitvena funkcija.*

- Naj bo $X \sim N(0,1)$. Pri kateri vrednosti X bo $Z = 0{,}5$? Kolikšna je torej verjetnost, da je $Z \leq 0{,}5$?

  Z bo enak 0,5 pri $X = 0$, verjetnost $P(Z \leq 0{,}5)$, je enaka 0,5.
- Naj bo $X \sim N(0,1)$. Pri kateri vrednosti X bo $Z = 0{,}975$? Kolikšna je torej verjetnost, da je $Z \leq 0{,}975$?

  Verjetnost je enaka 0,975. Vrednosti Z so kvantili porazdelitve X.
- Teoretično izpeljite $F_Z(z)$ za poljuben F (predpostavite, da je F^{-1} definiran za vse vrednosti, ki jih lahko zavzame X).

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(F_X(X) \leq z) = P(X \leq F_X^{-1}(z)) \\ &= F_X(F_X^{-1}(z)) = z \end{aligned}$$

  Spremenljivka Z je enakomerno porazdeljena.

d) Naj bo $U \sim U[0,1]$ in $X = F^{-1}(U)$. Pokažite, da je F porazdelitvena funkcija spremenljivke X.

Vemo, da za enakomerno porazdeljeno spremenljivko U velja $F_U(u) = u$:

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F_U(F(x)) = F(x)$$

F je torej kumulativna porazdelitvena funkcija spremenljivke X.

e) Želimo simulirati vrednosti iz eksponentne porazdelitve ($f(x) = \lambda e^{-\lambda x}$ za $x > 0$). Kako bi jih lahko simulirali z uporabo prej omenjenega generatorja?

Najprej potrebujemo funkcijo F:

$$\begin{aligned} F_Z(z) &= \int_0^z \lambda e^{-\lambda x} dx \\ &= \frac{\lambda}{-\lambda} e^{-\lambda x}\Big|_0^z \\ &= -1[e^{-\lambda z} - 1] = 1 - e^{-\lambda z} \end{aligned}$$

Inverzna porazdelitev F^{-1} je enaka:

$$\begin{aligned} u &= 1 - e^{-\lambda x} \\ 1 - u &= e^{-\lambda x} \\ -\log(1-u) &= \lambda x \\ x &= \frac{-\log(1-u)}{\lambda} \end{aligned}$$

Če so vrednosti u torej realizacije enakomerno porazdeljene slučajne spremenljivke U, so x realizacije eksponentno porazdeljene spremenljivke X. Mimogrede - log tu in v celotnem preostalem besedilu označuje naravni logaritem, torej logaritem z osnovo e.

f) Kako bi hkrati simulirali vrednosti za posameznike z različno vrednostjo λ?

Enako kot zgoraj - le da so vrednosti λ lahko različne.
Generiranje eksponentno porazdeljenih vrednosti bo osnova za generiranje časov pri analizi preživetja - osnovna predpostavka na tem področju so namreč enakomerno porazdeljeni časi od diagnoze neke bolezni do smrti. Parameter eksponentne porazdelitve je pri tem lahko drugačen za vsakega posameznika, odvisen je namreč lahko npr. od njegove starosti, spola, ipd. (glej zadnji primer v R).

Predlogi za vaje v R

- Generirajte podatke za 10000 voznikov, tako da jih je 500 med njimi pijanih, 9500 pa treznih. Naj bo verjetnost, da ima pijani voznik avtomobilsko nesrečo 0,3, verjetnost za zdravega pa 0,003. Izračunajte delež nesreč na simuliranih podatkih in ga primerjajte z dejansko verjetnostjo nesreče.

- Generirajte podatke za 100 posameznikov, tako da bo njihova starost enakomerno porazdeljena med 50 in 80. Preverite s histogramom.

- Vzemimo, da je bila posameznikom iz prejšnje točke postavljena diagnoza hude bolezni. Generirajte čase preživetja z eksponentno porazdelitvijo, tako da bodo imeli starejši posamezniki večjo verjetnost, da umrejo prej.
  Namig: parameter λ naj bo premosorazmeren s starostjo, npr.
  $\lambda = \text{starost}/100$.

Povzetek

- Pomemben izrek (transformacija integrala verjetnosti), ki ga izpeljemo v tej nalogi, pove, da če spremenljivko X preslikamo z njej lastno kumulativno porazdelitveno funkcijo F_X, vedno dobimo enakomerno porazdeljeno spremenljivko. Poleg izpeljave dokaza je dodan še poskus intuitivne razlage. Pokažemo, da velja tudi obratno - če vrednosti neke enakomerno porazdeljene spremenljivke preslikamo z inverzom neke porazdelitvene funkcije F_Y^{-1}, bo F_Y kumulativna porazdelitvena funkcija novo dobljene spremenljivke. Tako lahko s pomočjo generatorja enakomerne porazdeljenih slučajnih spremenljivk dobimo poljubno drugo porazdelitev, za katero lahko zapišemo inverz F^{-1} (ta seveda ne obstaja vedno - protiprimer najdemo že pri normalni porazdelitvi, kjer eksplicitna formula za integral gostote in s tem kumulativno porazdelitveno funkcijo ne obstaja).

- Izrek je v nalogi podan za dva primera porazdelitve - Bernoullijevo in eksponentno. Generiranje poljubne porazdelitve bo prišlo zelo prav v nalogah, ki sledijo - le če znamo generirati podatke iz znanih porazdelitev, lahko teoretično izpeljane lastnosti preverimo tudi s simulacijami.

1.3 Vsota diskretnih slučajnih spremenljivk

Naj bosta X in Y neodvisni Bernoullijevo porazdeljeni spremenljivki, $X, Y \sim Ber(p)$.

a) Kako je porazdeljena njuna vsota?

Označimo $Z = X + Y$. Verjetnost, da je $P(Z = z)$ za nek z, zapišemo kot vsoto verjetnosti vseh kombinacij $X = x$ in $Y = y$, ki dajo vsoto enako z (lahko seštevamo, saj so dogodki nezdružljivi):

$$\begin{aligned} P(Z = z) &= P(X + Y = z) = \sum_y P(X = z - y, Y = y) \\ &= \sum_y P(X = z - y | Y = y) P(Y = y) \end{aligned}$$

Za neodvisni X in Y torej velja

$$P(Z = z) = \sum_y P(X = z - y, Y = y) = \sum_y P(X = z - y) P(Y = y)$$

Uporabimo, da sta spremenljivki Bernoullijevo porazdeljeni, in dobimo:

$$\begin{aligned} P(Z = 0) &= P(X + Y = 0) = P(X = 0)P(Y = 0) = (1 - p)^2 \\ P(Z = 1) &= P(X = 0)P(Y = 1) + P(X = 1)P(Y = 0) \\ &= (1 - p)p + p(1 - p) = 2p(1 - p) \\ P(Z = 2) &= P(X = 1)P(Y = 1) = p^2 \end{aligned}$$

Velja torej

$$P(Z = z) = \binom{2}{z} p^z (1 - p)^{2-z}$$

b) Kako je porazdeljena vsota n neodvisnih enako porazdeljenih Bernoullijevih spremenljivk?

Označimo $Z = \sum_{i=1}^{n} X_i$, $i = 1, \ldots, n$, $X_i \sim Ber(p)$. Za $n = 2$ je Z porazdeljena binomsko, $Z \sim Bin(2, p)$. Dokazati želimo, da to velja za

poljuben n. V ta namen bomo uporabili matematično indukcijo - prvi korak smo že pokazali, sledi še korak $n \to n+1$. Naj bo torej $U \sim Bin(n,p)$ in $X \sim Ber(p)$, pokazati moramo, da velja $U + X \sim Bin(n+1,p)$.
Za $1 \leq z \leq n$ velja

$$\begin{aligned} & P(Z=z) = \\ & = \sum_x P(U=z-x)P(X=x) \\ & = P(U=z)P(X=0) + P(U=z-1)P(X=1) \\ & = \binom{n}{z} p^z (1-p)^{n-z} \cdot (1-p) + \binom{n}{z-1} p^{z-1}(1-p)^{n-z+1} \cdot p \\ & = \left[\binom{n}{z} + \binom{n}{z-1} \right] p^z (1-p)^{n-z+1} \\ & = \binom{n+1}{z} p^z (1-p)^{n+1-z} \end{aligned}$$

Dokaz za $z = 0$ in $z = n+1$ je prepuščen bralcu.

Predlogi za vaje v R

- S pomočjo funkcije `sample` generirajte po 100 realizacij dveh Bernoullijevih spremenljivk in si oglejte porazdelitev njune vsote.

Povzetek

- Naloga služi kot uvod v naslednjo nalogo - na najpreprostejšem primeru diskretnih spremenljivk smo intuitivno skušali razumeti formulo za porazdelitev vsote dveh slučajnih spremenljivk:
  vsota dveh spremenljivk je lahko enaka z na več različnih načinov, in ker slučajna spremenljivka ne more zavzeti dveh različnih vrednosti hkrati, lahko dogodek $X + Y = z$ razbijemo na posamezne nezdružljive dogodke (za vsako vrednost $Y = y$ posebej). Njihove verjetnosti lahko zaradi nezdružljivosti enostavno seštejemo.

  Verjetnost vsote dveh diskretnih slučajnih spremenljivk lahko računamo s formulo

$$P(X+Y=z) = \sum_y P(X=z-y, Y=y)$$

Če sta spremenljivki neodvisni, se formula poenostavi v

$$P(X+Y=z)=\sum_y P(X=z-y)P(Y=y)$$

- V nalogi smo izpeljali še rezultat, da je vsota n neodvisnih in enako porazdeljenih Bernoullijevih spremenljivk porazdeljena z binomsko porazdelitvijo:

$$X_i \sim Ber(p),\ i=1\ldots n,\ X_i \text{ neodvisne } \Rightarrow \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i} \sim \mathrm{Bin(n,p)}$$

1.4 Vsota zveznih slučajnih spremenljivk

Poleg posamičnih vrednosti želijo pri športnikih proučevati tudi zaporedje več meritev. Zanima nas porazdelitev vsote kvadriranih standardiziranih odmikov od povprečja pri ničelni domnevi, da športnik ni kriv. Naj bo torej Z standardizirani odmik od povprečja (po predpostavki normalno porazdeljen), zanima nas $\sum Z^2$ (gledamo vsoto kvadriranih odmikov, saj so vrednosti lahko negativne ali pozitivne). Pri tem predpostavimo, da so bile meritve narejene v dovolj dolgih časovnih presledkih, da so vrednosti med seboj neodvisne.

a) Najprej nas zanima, kako je porazdeljena vsota dveh neodvisnih zveznih spremenljivk (izpeljite formulo za dve zvezni spremenljivki, torej $Z=X+Y$, primerjajte jo s formulo za diskretne).

Zapišemo ustrezno kumulativno porazdelitveno funkcijo kot integral večrazsežne porazdelitve v ustreznih mejah:

$$\begin{aligned} P(Z\le z) &= P(X+Y\le z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_{X,Y}(v-y,y)dvdy \end{aligned}$$

pri čemer smo naredili substitucijo $x=v-y$. Sedaj lahko integrala zamenjamo in odvajamo (privzamemo, da je zunanji integral zvezen v z)

ter dobimo:

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_{X,Y}(v-y,y)dydv \\ f_Z(z) &= \int_{-\infty}^{\infty} f_{X,Y}(z-y,y)dy \\ f_Z(z) &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \end{aligned}$$

Pri tem zadnja vrstica velja, če sta X in Y neodvisni slučajni spremenljivki. Dobili smo rezultat, ki je analogen diskretni verziji.

b) Kako se porazdeljuje $S = Z_1^2 + Z_2^2$, če sta spremenljivki Z_1 in Z_2 porazdeljeni standardno normalno in med seboj neodvisni?
Namig: Pri izpeljavi uporabite rezultat

$$\int_0^a \frac{1}{\sqrt{(a-x)x}}dx = \pi$$

Izpeljali smo že, da je $Z^2 \sim \chi_1^2$ oziroma $Z^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$, torej

$$f_{Z^2}(z) = \frac{1}{\sqrt{2\pi z}} \exp\left\{-\frac{z}{2}\right\};\ z > 0$$

Izračunajmo gostoto vsote $Z_1^2 + Z_2^2$. Za dano vrednost s vemo, da mora biti vrednost z_2 med 0 (z_1 in z_2 ne moreta biti negativni) in s (ker sta obe pozitivni in je njuna vsota enaka s), zato morajo biti meje integracije

med 0 in s.

$$\begin{aligned} f_S(s) &= \int_0^s f_{Z_1^2}(s-z_2) f_{Z_2^2}(z_2) dz_2 \\ &= \frac{1}{2\pi} \int_0^s \frac{1}{\sqrt{s-z_2}} \exp\left\{-\frac{s-z_2}{2}\right\} \frac{1}{\sqrt{z_2}} \exp\left\{-\frac{z_2}{2}\right\} dz_2 \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^s \frac{1}{\sqrt{s-z_2}} \frac{1}{\sqrt{z_2}} dz_2 \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \pi \\ &= \frac{1}{2} \exp\left\{-\frac{s}{2}\right\}, \end{aligned}$$

Za izračun integrala smo pri tem uporabili v namigu podano formulo. Gostota gama porazdelitve je

$$f_X(x) = \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)};\ x > 0, \lambda > 0, a > 0,$$

Dobljeni rezultat je torej porazdelitev gama z $\lambda = \frac{1}{2}$ in $a = 1$.

Na enak način bi lahko izpeljali tudi splošnejši rezultat:

$$X_1 \sim \Gamma(a_1, \lambda),\ X_2 \sim \Gamma(a_2, \lambda)\ \Rightarrow\ X_1 + X_2 \sim \Gamma(a_1 + a_2, \lambda)$$

c) Denimo, da so športnikovi standardizirani odmiki (vrednosti Z) na petih merjenjih naslednji: 1,6; 1,5; −1,6; 1,8; 1,4. Kaj lahko sklepamo?

Uporabimo, da je vsota n neodvisnih enako porazdeljenih spremenljivk $X_i \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ porazdeljena kot $\sum_{i=1}^n X_i \sim \Gamma(\frac{n}{2}, \frac{1}{2})$.

```
> vr <- c(1.6,1.5,-1.6, 1.8, 1.4)
> rez <- sum(vr^2)
> rez
[1] 12.57
> 1-pgamma(rez, 2.5, 0.5)
[1] 0.02775943
```

Naša ničelna domneva je, da športnik ni kriv. Pod to ničelno domnevo se vsota kvadriranih odmikov porazdeljuje po gama porazdelitvi $\Gamma(\frac{5}{2}, \frac{1}{2})$. Verjetnost, da je vsota 12,57 ali več, je približno 0,03.

Predlogi za vaje v R

- Generirajte 10 vrednosti iz porazdelitve $X \sim N(148,85)$, te vrednosti naj predstavljajo 10 meritev pri enem športniku. Vrednosti standardizirajte (tako da dobite spremenljivko porazdljeno kot $N(0,1)$), kvadrirajte in seštejte. To naj bo vrednost za prvega športnika, na enak način generirajte vrednosti za 1000 športnikov. Narišite histogram vrednosti. S pomočjo funkcije `pgamma` poiščite mejo, ki jo porazdelitev $\Gamma(\frac{10}{2}, \frac{1}{2})$ preseže z verjetnostjo manj kot 0,01, in izračunajte delež športnikov na vašem vzorcu, ki presežejo to mejo.

- Recimo, da ima dopingiran športnik enako povprečje, a večjo varianco (vrednosti bolj nihajo, saj manipulira s krvjo). Generirajte 1000 športnikov z večjo varianco in si oglejte, kolikšen delež bo presegel meje iz prejšnje točke.

Povzetek

- Formula za vsoto dveh zveznih slučajnih spremenljivk je analogna formuli za vsoto diskretnih spremenljivk - verjetnosti zamenjajo gostote, vsote pa integrali:
  gostoto vsote dveh zveznih slučajnih spremenljivk lahko zapišemo kot

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)dy$$

  oziroma če sta spremenljivki neodvisni, kot

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy$$

- V nalogi smo izpeljali še rezultat, da je vsota dveh neodvisnih gama porazdeljenih slučajnih spremenljivk (z enakim parametrom λ) zopet

gama porazdeljena slučajna spremenljivka.
Podobno lastnost - da je vsota neodvisnih enako porazdeljenih spremenljivk zopet lahko opisana z isto porazdelitvijo - srečamo tudi pri normalni porazdelitvi (to dokažemo v nalogi 1.5), seveda pa ta lastnost še zdaleč ne velja za poljubno verjetnostno porazdelitev, na primer za vsoto Bernoullijevo ali enakomerno porazdeljenih spremenljivk (glej tudi povzetek naloge 1.9).

1.5 Vsota normalnih spremenljivk

Pokazali smo že, da je linearna transformacija normalno porazdeljene spremenljivke zopet normalno porazdeljena. V tej vaji bomo pokazali, da je vsota dveh neodvisnih (standardno) normalno porazdeljenih spremenljivk zopet normalno porazdeljena. Nato bomo s protiprimerom pokazali, da to ne velja za vsoto poljubnih odvisnih normalno porazdeljenih spremenljivk.

a) Naj bosta spremenljivki $X \sim N(0,1)$ in $Y \sim N(0,1)$ in med seboj neodvisni. Kako je porazdeljena njuna vsota?

Uporabimo formulo za gostoto vsote dveh neodvisnih slučajnih spremenljivk:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(z-y)^2}{2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} \exp\left\{-y^2\right\} \exp\left\{zy\right\} dy \end{aligned}$$

Sedaj dele izraza, v katerih nastopa y, zapišemo kot kvadrat neke vsote:

$$\begin{aligned} y^2 - zy &= y^2 - 2y\frac{z}{2} + \left(\frac{z}{2}\right)^2 - \left(\frac{z}{2}\right)^2 \\ &= \left(y - \frac{z}{2}\right)^2 - \frac{z^2}{4} \end{aligned}$$

Gornji integral torej prepišemo v

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} \exp\left\{-\left(y-\frac{z}{2}\right)^2\right\} \exp\left\{\frac{z^2}{4}\right\} dy \\ &= \frac{1}{2\pi}\exp\left\{-\frac{z^2}{4}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{(y-\frac{z}{2})^2}{2\cdot\frac{1}{2}}\right\} dy \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{4}\right\} \frac{1}{\sqrt{2}} \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{2}}} \exp\left\{-\frac{(y-\frac{z}{2})^2}{2\cdot\frac{1}{2}}\right\} dy\right] \\ &= \frac{1}{\sqrt{2\pi\cdot 2}}\exp\left\{-\frac{z^2}{2\cdot 2}\right\} \end{aligned}$$

V predzadnji vrstici smo pod integralom dobili ravno gostoto normalne porazdelitve $N(\frac{z}{2},\frac{1}{2})$, njen integral je zato 1 (ne glede na vrednost z, ki je znotraj tega integrala konstanta). Spremenljivka Z je normalno porazdeljena, $Z \sim N(0,2)$.

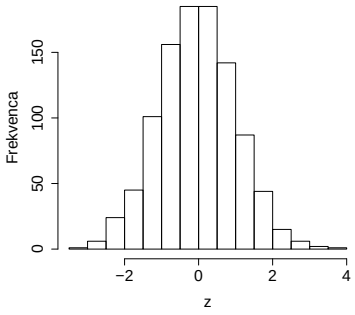
b) Naj bosta X in Y neodvisni standardno normalno porazdeljeni spremenljivki, Z pa enaka $|Y|$, če je $X \geq 0$, in $Z = -|Y|$, če je $X < 0$. Kako je porazdeljena spremenljivka Z?

Najprej narišimo simulirane vrednosti z R:

```
> set.seed(1)
> x <- rnorm(1000,0,1)          #1000 realizacij normalne spr.,
> y <- rnorm(1000,0,1)          #povprecje=0, sd=1
> z <- abs(y)                   #z = |y|
> z[x<0] <- -z[x<0]             #z =-|y|, ce je x<0
> hist(z,main="",ylab="Frekvenca") #histogram z
```

Sedaj še izpeljimo porazdelitveno funkcijo. Naj bo $z < 0$ (torej X negativen), uporabimo, da sta spremenljivki X in Y neodvisni:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X < 0, -|Y| \leq z) \\ &= P(X<0)[P(Y \leq z) + P(Y \geq -z)] \\ &= \frac{1}{2}[2 \cdot P(Y \leq z)] \\ &= P(Y \leq z) = F_Y(z) \end{aligned}$$

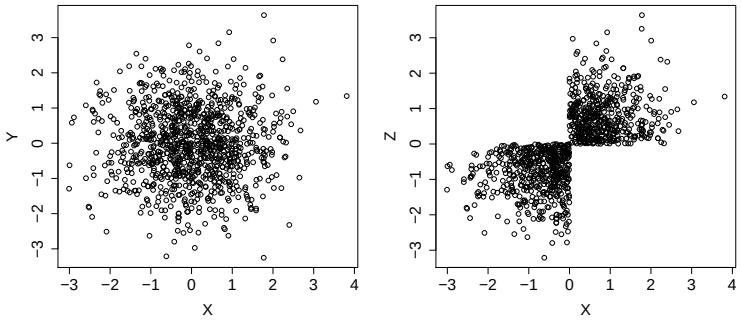


Slika 1.3: *Porazdelitev spremenljivke Z.*

Na enak način izpeljemo še $P(0 \leq Z \leq z) = P(0 \leq Y \leq y)$ za $z > 0$. Pokazali smo, da je porazdelitvena funkcija Z enaka porazdelitveni funkciji Y, Z je torej standardno normalno porazdeljena spremenljivka.

c) Skicirajte skupno porazdelitev spremenljivk X in Z. Ali sta spremenljivki neodvisni?

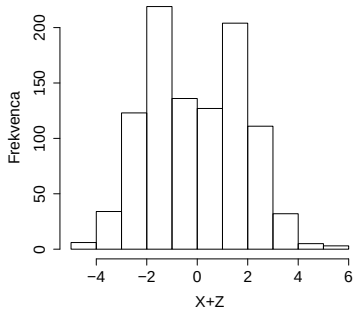
Očitno je (glej sliko 1.4), da spremenljivki nista neodvisni, vedno imata enak predznak.



Slika 1.4: *Razsevni diagram realizacij X in Y ter X in Z.*

d) Ali je vsota $X + Z$ porazdeljena normalno?

Slika 1.5 jasno kaže, da porazdelitev vsote ni normalna. Vsota dveh normalnih spremenljivk torej ni nujno normalna, če spremenljivki nista neodvisni (dana naloga je protiprimer).



Slika 1.5: *Porazdelitev vsote* $X + Z$.

Povzetek

- V nalogi smo najprej izpeljali, da je vsota dveh neodvisnih standardno normalno porazdeljenih spremenljivk zopet normalna spremenljivka. Pri dokazu smo se računanju zapletenega integrala izognili z uporabo dejstva, da je integral gostote neke spremenljivke na celotni realni osi vedno enak 1 ne glede na obliko porazdelitve oziroma vrednost njenih parametrov. Morali smo torej člene preurediti tako, da je pod integralom ostala le gostota neke porazdelitve.

- Rezultat, ki smo ga za dve spremenljivki izpeljali v nalogi, je v resnici še dosti bolj splošen - vsota poljubnega števila neodvisnih normalno porazdeljenih spremenljivk je zopet normalno porazdeljena spremenljivka.

- Naloga je s protiprimerom pokazala, da vsota odvisnih normalno porazdeljenih spremenljivk ni nujno normalna. Definirali smo slučajni spremenljivki X in Z in za obe pokazali, da sta normalni ter med seboj odvisni. Njuna vsota očitno ni normalno porazdeljena. V splošnem lahko o porazdelitvi vsote odvisnih spremenljivk povemo le malo - način odvisnosti bo ključno vplival na obliko dobljene porazdelitve.

1.6 Porazdelitev vzorčnega povprečja

Vrnimo se spet k primeru odkrivanja dopinga. Izkaže se, da ima vsak posameznik sebi lastno povprečje hemoglobina in da se te vrednosti med posamezniki precej razlikujejo. Da bi dosegli večjo občutljivost testa, uvedemo polletno testno obdobje, v katerem vsakega športnika testiramo petkrat. Povprečje teh petih meritev bomo vzeli kot oceno za posameznikovo povprečje pri testih v prihodnosti (meje bomo postavljali glede na to povprečje). Recimo, da vemo, da se vrednosti vsakega športnika okrog njemu lastnega povprečja porazdeljujejo normalno z varianco $\sigma^2 = 5^2$ in da so posamezne vrednosti med seboj neodvisne.

a) Naj bodo $X_i,\ i = 1, \ldots, n$, neodvisne, enako porazdeljene slučajne spremenljivke. Kaj lahko rečete o pričakovani vrednosti in varianci njihovega povprečja? Označite $E(X_i) = \mu$ in $\mathrm{var}(X_i) = \sigma^2$ za vsak i.

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

$$\begin{aligned} E[\bar{X}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \\ \mathrm{var}[\bar{X}] &= \mathrm{var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

   Opomba: neodvisnost smo potrebovali pri izračunu variance, medtem ko bo pričakovana vrednost vsote vedno vsota pričakovanih vrednosti.

b) Izračunajte meje okrog ocenjenega povprečja, znotraj katerih naj bi pri šesti meritvi nedopingiran športnik ostal z verjetnostjo 0,99 (prvih pet meritev uporabimo za oceno športnikovega povprečja).
   Namig: uporabite rezultat, da je vsota neodvisnih normalno porazdeljenih spremenljivk spet normalna.

   Vrednosti posameznika so porazdeljene kot $X \sim N(\mu, \sigma^2)$ ($\sigma = 5$). Vsota $\sum_{i=1}^{5} X_i$ je normalna spremenljivka, prav tako je normalno porazdeljeno tudi povprečje $\bar{X}$, saj vsoto le pomnožimo s konstanto. Zanima nas odstopanje šeste meritve od ocenjenega povprečja v prvih petih meritvah, torej razlika $Z = X_6 - \frac{1}{5}\sum_{i=1}^{5} X_i$. To je torej razlika dveh normalnih

spremenljivk z enakim povprečjem, ena ima varianco σ^2, druga pa σ^2/n. Spremenljivka Z je porazdeljena kot $Z \sim N(0, \sigma^2 + \sigma^2/n) = N(0{,}30)$. Vrednost $z_{0{,}005} = 2{,}57$, meje so torej $\frac{1}{5}\sum_{i=1}^{5} X_i \pm 2{,}57 \cdot \sqrt{30}$.

c) Ali je povprečje neodvisnih enako porazdeljenih slučajnih spremenljivk vedno porazdeljeno z isto porazdelitvijo kot vsaka posamezna spremenljivka?

Ne, to v splošnem ni res. Protiprimer je na primer vsota Bernoullijevo porazdeljenih spremenljivk, ki je porazdeljena binomsko, in prav tako to očitno ne velja za povprečje enakomerno porazdeljenih spremenljivk. Glej tudi nalogo 1.9.

Predlogi za vaje v R

- Predpostavite, da so povprečja športnikov normalno porazdeljena z $N(148, 7{,}5^2)$ in generirajte povprečne vrednosti za 100 športnikov. Nato uporabite normalno porazdelitev $N(0, 5^2)$, ki predstavlja odmike od osebnega povprečja vsakega športnika - generirajte po 6 vrednosti na posameznika. Ocenite osebna povprečja s pomočjo prvih petih vrednosti ter primerjajte varianco teh povprečij s teoretično vrednostjo. Oglejte si porazdelitev odstopanja šeste vrednosti od prvih petih.

Povzetek

- V tej nalogi se prvič srečamo s pojmoma populacije in vzorca, s katerima opišemo cilj statističnega sklepanja: s pomočjo vrednosti na vzorcu želimo opisati populacijo. Pri tem si vzorec zamišljamo kot neodvisne realizacije neke slučajne spremenljivke, ki predstavlja populacijo. Vzorčenju se bomo sicer povsem posvetili v naslednjem poglavju.

- Vzorčno povprečje je slučajna spremenljivka - na vsakem vzorcu iz neke populacije lahko pričakujemo drugačno vrednost. Cilj naloge je bil najti porazdelitev vzorčnega povprečja, če vemo, da je populacija normalno porazdeljena. Vzorčno povprečje je utežena vsota neodvisnih normalnih slučajnih spremenljivk, in zato zopet normalno porazdeljena spremenljivka. Njena pričakovana vrednost je kar populacijsko povprečje μ, njena varianca pa je enaka σ^2/n.

- Varianco vzorčnega povprečja imenujemo standardna napaka - variabilnost vzorčnega povprečja okrog populacijske vrednosti nam namreč

govori o tem, koliko se bomo motili, če bomo z vzorčnim povprečjem skušali ocenjevati povprečje v populaciji.

- Velikost standardne napake je obratno sorazmerna z velikostjo vzorca - večji vzorec imamo, manj se motimo pri oceni populacijskega povprečja. Standardna napaka je odvisna tudi od vrednosti σ - če je variabilnost v populaciji majhna, se pri oceni vzorčnega povprečja ne moremo dosti zmotiti.

1.7 Pogojna pričakovana vrednost in varianca

Raziskovalci na področju športa so dokazali, da je pri kolesarjih hemoglobin izven tekmovalnega obdobja porazdeljen kot $N(150, 7^2)$, med tekmovalnim obdobjem pa kot $N(140, 11^2)$. Vzemimo, da tekmovalno obdobje traja 9 mesecev. Zanimata nas pričakovana vrednost in standardni odklon naključnega odvzetega vzorca.
Namig: zanima nas slučajna spremenljivka Y, vemo $\{Y|X = 0\} \sim N(150, 7^2)$ in $\{Y|X = 1\} \sim N(140, 11^2)$, $P(X = 1) = 0{,}75$

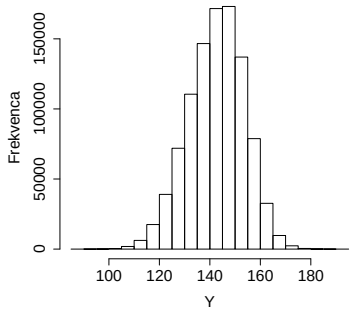
a) Skicirajte porazdelitev Y, kaj lahko rečete o pričakovani vrednosti ter standardnem odklonu?

   Porazdelitev je na sliki 1.6. Pričakujemo, da bo povprečna vrednost nekje med vrednostima v obeh obdobjih. Ker je verjetnost, da je meritev izšla iz tekmovalnega obdobja, večja, pričakujemo, da bo vrednost nekoliko bližja povprečju v tem obdobju. Varianca bo nekoliko večja od posameznih varianc, odvisna bo tudi od razdalje med povprečjema.

b) Na danem primeru razložite formulo $E(Y) = E[E(Y|X)]$. Je $E(Y|X)$ slučajna spremenljivka ali konstanta? Izračunajte pričakovano vrednost spremenljivke Y.

   $Z = E(Y|X)$ je slučajna spremenljivka, ki lahko zavzame dve vrednosti: $P(Z = 140) = 0{,}75$, $P(Z = 150) = 0{,}25$. Pričakovana vrednost te spremenljivke je torej

   $$E(Z) = 140 \cdot P(Z = 140) + 150 \cdot P(Z = 150) = 140 \cdot 0{,}75 + 150 \cdot 0{,}25 = 142{,}5$$



Slika 1.6: *Porazdelitev nove spremenljivke Y.*

Torej

$$E[E(Y|X)] = \sum_x E(Y|X = x) \cdot P(X = x)$$

c) Izračunajte varianco Y.

Pri izračunu variance si bomo pomagali s formulo

$$\text{var}(Y) = \text{var}[E(Y|X)] + E[\text{var}(Y|X)]$$

V prvem delu gornje formule nas torej zanima varianca slučajne spremenljivke $Z = E(Y|X)$:

$$\begin{aligned} \text{var}(Z) &= E([Z - E(Z)]^2) \\ &= 7{,}5^2 \cdot P[(Z - E(Z)) = 7{,}5] + 2{,}5^2 \cdot P[(Z - E(Z)) = 2{,}5] \\ &= 56{,}25 \cdot 0{,}25 + 6{,}25 \cdot 0{,}75 = 18{,}75 \end{aligned}$$

Standardni odklon povprečij v različnih obdobjih okrog robnega povprečja je torej 4,33.
Člen $E[\text{var}(Y|X)]$ je pričakovana vrednost za varianco Y pri znanem X. Vemo, da je $\text{var}(Y|X = 0) = 49$ in $\text{var}(Y|X = 1) = 121$. Pričakovana vrednost je

$$E[\text{var}(Y|X)] = 49 \cdot 0{,}25 + 121 \cdot 0{,}75 = 103$$

Sestavimo oba dela skupaj in dobimo var(Y) = 121,75, sd(Y) = 11,03. Vrednosti izven tekmovalnega obdobja torej le malo povečajo variabilnost rezultatov.

d) Izrazite varianco v splošnem: $\{Y|X=0\} \sim N(\mu_0, \sigma_0^2)$, $\{Y|X=1\} \sim N(\mu_1, \sigma_1^2)$, $P(X=1)=p$.

Vrednost $E(Y|X=0)=\mu_0$ je pričakovana vrednost Y pri $X=0$, torej izven tekmovalnega obdobja, podobno z $\mu_1 = E(Y|X=1)$ označimo pričakovano vrednost Y med tekmovalnim obdobjem. Verjetnost, da je športnik v tekmovalnem obdobju, označimo s p. Ker je X Bernoullijevo porazdeljena spremenljivka, velja $E(X) = P(X=1) = p$. Funkcijo

$$E(Y|X) = \begin{cases} \mu_0; & X=0 \\ \mu_1; & X=1 \end{cases}$$

zapišemo kot $E(Y|X) = \mu_0(1-X) + \mu_1 X$. Pričakovana vrednost slučajne spremenljivke $Z = E(Y|X)$ je

$$E(Z) = E(E(Y|X)) = \sum_{X=x} E(Y|X=x)P(X=x) = \mu_0(1-p) + \mu_1 p$$

Pričakovana vrednost Y je torej uteženo povprečje pogojnih pričakovanih vrednosti v posameznih obdobjih, bližja je tistemu obdobju, iz katerega smo bolj verjetno dobili meritev.

Varianca slučajne spremenljivke Z je enaka

$$\begin{aligned}
\text{var}(Z) &= \sum_{X=x} \left[E(Y|X=x) - E(Y)\right]^2 P(X=x) \\
&= [\mu_0 - \mu_0(1-p) - \mu_1 p]^2(1-p) + [\mu_1 - \mu_0(1-p) - \mu_1 p]^2 p \\
&= [-p(\mu_0 - \mu_1)]^2(1-p) + [(1-p)(\mu_1 - \mu_0)]^2 p \\
&= [\mu_1 - \mu_0]^2 p^2(1-p) + [\mu_1 - \mu_0]^2(1-p)^2 p \\
&= [\mu_1 - \mu_0]^2 p(1-p)(p+1-p) \\
&= [\mu_1 - \mu_0]^2 p(1-p)
\end{aligned}$$

Izrazimo še drugi del, slučajna spremenljivka var$(Y|X)$ je enaka

$$\text{var}(Y|X) = \begin{cases} \sigma_0^2; & \text{z verjetnostjo } (1-p) \\ \sigma_1^2; & \text{z verjetnostjo } p \end{cases}$$

Spremenljivka $\text{var}(Y|X)$ je torej Bernoullijevo porazdeljena, njena pričakovana vrednost je $E(\text{var}(Y|X)) = \sigma_0^2(1-p) + \sigma_1^2 p$.
Združimo oba dela skupaj in dobimo

$$\text{var}(Y) = [\mu_1 - \mu_0]^2 p(1-p) + \sigma_0^2(1-p) + \sigma_1^2 p$$

Varianca slučajne spremenljivke Y je torej seštevek uteženega povprečja varianc v posameznih obdobjih in člena, ki je odvisen od razmika med povprečjema.

e) Izračunajte kovarianco X in Y. Izrazite jo splošno ($\{Y|X=0\} \sim N(\mu_0, \sigma_0^2)$, $\{Y|X=1\} \sim N(\mu_1, \sigma_1^2)$, $P(X=1) = p$).
Kako je kovarianca odvisna od parametrov? Kaj pa korelacija?

Po definiciji je kovarianca enaka

$$\begin{aligned}
&\text{cov}(X,Y) = \\
&\quad = E[(X - E(X))(Y - E(Y))] \\
&\quad = \int_{\mathbb{R}} \int_{\mathbb{R}} (x - E(X))(y - E(Y)) f_{X,Y}(x,y) dx dy
\end{aligned}$$

Zanima nas torej pričakovana vrednost glede na skupno porazdelitev X in Y (lahko bi pisali $E_{X,Y}$). Gostote skupne porazdelitve ne poznamo, zato poizkusimo izraz preurediti tako, da bo v njem nastopala pogojna gostota. Uporabimo torej enakost $f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)$ in najprej izračunajmo integral po y

$$\begin{aligned}
&\text{cov}(X,Y) = \\
&\quad = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} (x - E(X))(y - E(Y)) f_{Y|X}(y|x) dy \right] f_X(x) dx \\
&\quad = \int_{\mathbb{R}} (x - E(X)) \left[\int_{\mathbb{R}} (y - E(Y)) f_{Y|X}(y|x) dy \right] f_X(x) dx
\end{aligned}$$

V integralu $\int_{\mathbb{R}} E(Y) f_{Y|X}(y|x) dy$ lahko vrednost $E(Y)$ izpostavimo, saj je konstanta. Funkcija $f_{Y|X}(y|x)$ predstavlja pogojno gostoto - pri vsaki vrednosti x imamo torej neko slučajno spremenljivko $U = Y|_{X=x}$ z gostoto $f_U(u) = f_{Y|X}(y|x)$. Zato je integral pri dani vrednosti x enak

$\int_{\mathbb{R}} f_{Y|X}(y|x)dy = 1$, torej

$$\begin{aligned}\operatorname{cov}(X,Y) &= \\ &= \int_{\mathbb{R}} (x - E(X)) \left[\int_{\mathbb{R}} y f_{Y|X}(y)dy - E(Y)\right] f_X(x)dx \\ &= \int_{\mathbb{R}} (x - E(X)) \left[E(Y|X) - E(Y)\right] f_X(x)dx\end{aligned}$$

V našem primeru je X diskretna spremenljivka, integral po X lahko torej zamenjamo z vsoto dveh členov:

$$\begin{aligned}\operatorname{cov}(X,Y) &= \\ &= (0 - E(X))[E(Y|X=0) - E(Y)]P(X=0) + \\ &\quad (1 - E(X))[E(Y|X=1) - E(Y)]P(X=1)\end{aligned}$$

Vrednost $E(Y|X=0) = \mu_0$ je pričakovana vrednost Y pri $X = 0$, torej izven tekmovalnega obdobja, podobno smo z $\mu_1 = E(Y|X=1)$ označili pričakovano vrednost Y med tekmovalnim obdobjem. Označimo še $E(Y) = \mu$. Verjetnost, da je športnik v tekmovalnem obdobju, označimo s p. Ker je X Bernoullijevo porazdeljena spremenljivka, velja $E(X) = P(X=1) = p$. Zato

$$\begin{aligned}\operatorname{cov}(X,Y) &= \\ &= -p[\mu_0 - \mu](1-p) + (1-p)[\mu_1 - \mu]p \\ &= p(1-p)(-\mu_0 + \mu_1)\end{aligned}$$

in

$$\operatorname{cor}(X,Y) = \frac{p(1-p)(\mu_1 - \mu_0)}{\sqrt{\operatorname{var} Y}\sqrt{p(1-p)}}$$

V našem primeru

$$\begin{aligned}\operatorname{cov}(X,Y) &= 0{,}75 \cdot 0{,}25 \cdot (140 - 150) = -1{,}875 \\ \operatorname{cor}(X,Y) &= -\frac{1{,}875}{11{,}03 \cdot \sqrt{0{,}75 \cdot 0{,}25}} = -0{,}392\end{aligned}$$

Kovarianca in korelacija sta odvisni od razlike med povprečjema - večja kot je razlika, večji sta (po absolutni vrednosti). Če bi bila razlika 0, torej

povprečje neodvisno od obdobja, bi bili tudi korelacija oz. kovarianca enaki 0. Vrednosti sta negativni, kadar večji X pomeni manjši Y. Odvisni sta tudi od p - največji sta, kadar je $p = 0{,}5$, torej kadar obe obdobji enako prispevata k skupnemu povprečju (če bi bilo vrednosti v enem obdobju zelo malo, bi bila korelacija majhna). Dodatno je korelacija odvisna tudi od variabilnosti v enem in drugem obdobju. Če bi bila ta variabilnost velika v primerjavi z razliko med povprečjema, spremenljivki ne bi bili močno povezani.

f) Kolikšne so vrednosti variance, kovariance in korelacije, če sta povprečji v tekmovalnem in izven tekmovalnega obdobja enaki? Ali sta spremenljivki X in Y tedaj neodvisni?

Če je razlika enaka 0, torej povprečje neodvisno od obdobja, je varianca Y enaka $\text{var}(Y) = \sigma_0^2(1-p) + \sigma_1^2 p$, korelacija in kovarianca pa sta enaki 0. Vendar pa to ne pomeni, da sta spremenljivki X in Y neodvisni - od vrednosti X je odvisna varianca Y. Porazdelitev Y je torej odvisna od X, četudi X ne vpliva na povprečje. Torej: vemo, da je korelacija enaka 0, če sta spremenljivki neodvisni, vendar obratno ni nujno res.

Predlogi za vaje v R

- Generirajte veliko število vrednosti in si oglejte njihovo porazdelitev:

```
> set.seed(1)
> a <- rnorm(1000000,mean=140,sd=11) #vrednosti Y|X za tek. obd.
> b <- rnorm(1000000,mean=150,sd=7)  #vrednosti Y|X za ne-tek. obd.
> x <- sample(0:1,size=1000000,      #obdobje - porazdelitev X
+ replace=T,prob=c(0.25,0.75))
> y <- a*x+b*(1-x)                   #slucajna spremenljivka Y
> hist(y,prob=T)                     #narisemo spremenljivko
> mean(y)                            #ocena povprecja
> var(y)                             #ocena variance
```

- Poizkusite preveriti vsakega od rezultatov še z R. Primerjajte teoretične vrednosti z njihovimi ocenami.

Povzetek

- Pogojna pričakovana vrednost $E(Y|X)$ je slučajna spremenljivka. Pričakovana vrednost te slučajne spremenljivke je $E[E(Y|X)] = E(Y)$.

- Če poznamo pogojno porazdelitev $Y|X$, lahko varianco spremenljivke Y izrazimo s formulo

$$\text{var}(Y) = \text{var}[E(Y|X)] + E[\text{var}(Y|X)]$$

- Kovarianca in korelacija dveh neodvisnih spremenljivk sta enaki 0, obratno pa ni nujno res: četudi je kovarianca oziroma korelacija enaka 0, spremenljivki nista nujno neodvisni. Protiprimer je razložen v točki f).

1.8 Uvrščanje

Na podlagi nekega kazalnika želimo ocenjevati kreditno sposobnost posameznikov, želimo jih uvrstiti v dve skupini - tiste, ki bodo kredit odplačali, in tiste, ki ga ne bodo. Kot učni vzorec imamo na razpolago vrednosti tega kazalnika za posameznike, ki so lansko leto najeli enoletni kredit, in podatke o tem, ali je letos kredit odplačan ali ne (naloga je povzeta po Blagus, 2011). Predpostavimo, da so vrednosti kazalnika porazdeljene pri obeh skupinah posameznikov približno normalno. Na podlagi letošnjih podatkov ocenimo povprečno vrednost kazalnika za posameznike, ki so kredit odplačali ($\bar{X}_d$), in za posameznike, ki ga niso ($\bar{X}_s$). Ti dve oceni sedaj uporabimo za uvrščanje strank: napovemo, da bo nek posameznik uspel odplačati kredit, če je trenutna vrednost njegovega kazalnika bližja $\bar{X}_d$ kot $\bar{X}_s$.
Raziskati želimo lastnosti takega uvrščanja. Recimo, da smo v učni vzorec zajeli $n_d = 750$ posameznikov, ki so kredit odplačali, in $n_s = 250$, ki ga niso. Recimo, da je kazalnik povsem neuporaben za naš namen, torej da je njegova porazdelitev enaka pri "dobrih" kot pri "slabih" strankah: X_s in X_d sta enako porazdeljena, označimo kar z X: $X \sim N(50, 15^2)$. Ugotoviti želimo, kolikšna bo verjetnost, da neko naključno stranko na podlagi današnje vrednosti njenega kazalnika uvrstimo med "dobre".

a) Kakšna je porazdelitev $\bar{X}_s$ (v splošnem, torej za neko varianco σ^2 in neko število slabih n_s v učnem vzorcu)?

   $X_s \sim N(\mu, \sigma^2)$. Vemo, da je povprečje neodvisnih normalnih spremenljivk normalno porazdeljeno ter da velja $\bar{X}_s = \frac{1}{n_s} \sum X_{s,i} \sim N(\mu, \frac{\sigma^2}{n_s})$.

b) Kakšna je porazdelitev $X - \bar{X}_s$?

Vemo že (glej nalogo 1.6), da velja $X - \bar{X}_s \sim N(0, \sigma^2 + \frac{\sigma^2}{n_s})$.

c) Označimo $Z = X - \bar{X}_s$ in $Y = X - \bar{X}_d$. Izračunajte kovarianco in korelacijo. Interpretirajte izračunano formulo za korelacijo za $n_s = n_d$. Kako je odvisna od velikosti vzorcev?

Nova vrednost X je neodvisna od vzorčnih povprečij, povprečji pa sta med seboj prav tako neodvisni. Zato velja

$$\mathrm{cov}[X - \bar{X}_s, X - \bar{X}_d] = \mathrm{cov}[X, X] = \mathrm{var}(X) = \sigma^2$$

Korelacija je enaka

$$\begin{aligned} \mathrm{cor}[X - \bar{X}_s, X - \bar{X}_d] &= \frac{\sigma^2}{\sqrt{\sigma^2(1 + \frac{1}{n_s})}\sqrt{\sigma^2(1 + \frac{1}{n_d})}} \\ &= \frac{1}{\sqrt{(1 + \frac{1}{n_s})}\sqrt{(1 + \frac{1}{n_d})}} \end{aligned}$$

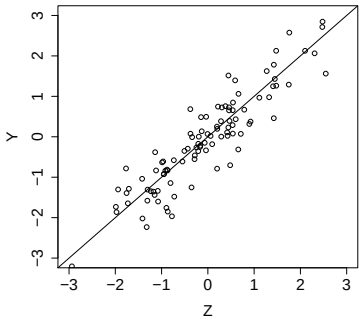
Če sta velikosti vzorca med seboj enaki, velja

$$\mathrm{cor}[X - \bar{X}_s, X - \bar{X}_d] = \frac{n_s}{1 + n_s}$$

Ko vzorca naraščata, gre korelacija proti 1. To je intuitivno jasno, saj z naraščanjem vzorca oceni povprečij postaneta zelo natančni (v primerjavi s posamezno vrednostjo sta skoraj konstanti).

d) Izrazite formulo za gostoto $f_{Z,Y}(z, y)$.
Namig: uporabimo rezultat, da je skupna gostota v tem primeru normalna (Rice (2009), stran 102), kar pa v splošnem ni nujno vedno res.

Za zapis dvorazsežne normalne porazdelitve potrebujemo robna povprečja in variance (izpeljali smo jih v drugi točki) ter korelacijo ρ med spremenljivkama (izpeljali smo jo v prejšnji točki). Skupno gostoto torej zapišemo



Slika 1.7: *Razsevni diagram razlik za 100 realizacij slučajne spremenljivke.*

kot

$$\begin{aligned} f_{Z,Y}(z,y) &= \frac{1}{2\pi\sigma_Z\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(z-\mu_Z)^2}{\sigma_Z^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right.\right. \\ &\quad \left.\left.-\frac{2\rho(z-\mu_z)(y-\mu_y)}{\sigma_Z\sigma_Y}\right]\right) \\ &= \frac{1}{2\pi\sigma^2 sd\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}\left[\frac{z^2}{s^2}+\frac{y^2}{d^2}-\frac{2\rho zy}{sd}\right]\right) \end{aligned}$$

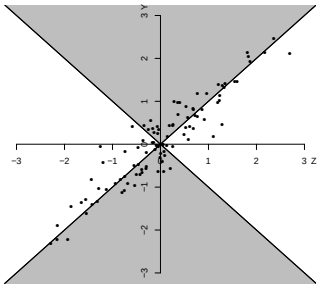
Pri tem smo označili $s=\sqrt{1+\frac{1}{n_s}}$ in $d=\sqrt{1+\frac{1}{n_d}}$, torej $\rho^2=1/(sd)$. Definirajmo še $c=\sqrt{\frac{n_s n_d}{1+n_s+n_d}}$ in dobimo

$$f_{Z,Y}(z,y)=\frac{c}{2\pi\sigma^2}e^{-\frac{c^2}{2\sigma^2}(y^2\frac{1+n_s}{n_s}+z^2\frac{1+n_d}{n_d}-2yz)}$$

e) Zanima nas verjetnost $P(|X-\bar{X}_s|<|X-\bar{X}_d|)$. Kako bi jo izračunali? Skicirajte območje, ki nas zanima, nastavite integral in meje.

Slika 1.8 prikazuje območje integracije. Integral, ki ga moramo izračunati, je enak

$$\begin{aligned} P(|Y|>|Z|) &= \int\int_{|Y|>|Z|} f_{Z,Y}(z,x)dxdz \\ &= \frac{c}{2\pi\sigma^2}\int\int_{|Y|>|Z|} e^{-\frac{c^2}{2\sigma^2}(y^2\frac{1+n_s}{n_s}+z^2\frac{1+n_d}{n_d}-2yz)}dzdy \end{aligned}$$



Slika 1.8: *Območje* $|Z| < |Y|$*, za* $n_s = 10$*,* $n_d = 20$*.*

Zaradi simetrije je integral za pozitivne y (glej sliko 1.8) enak kot za negativne, zato je dovolj, da integriramo le gornji del (in množimo z 2). Gornja kraka dodatno razdelimo na negativne in pozitivne vrednosti z:

$$\begin{aligned} P(|Y| > |Z|) &= \frac{c}{\pi\sigma^2} \int_0^\infty \int_0^y e^{-\frac{c^2}{2\sigma^2}(y^2 \frac{1+n_s}{n_s} + z^2 \frac{1+n_d}{n_d} - 2yz)} dz dy + \\ &\quad + \frac{c}{\pi\sigma^2} \int_0^\infty \int_{-y}^0 e^{-\frac{c^2}{2\sigma^2}(y^2 \frac{1+n_s}{n_s} + z^2 \frac{1+n_d}{n_d} - 2yz)} dz dy. \end{aligned}$$

Izračunamo zgornje integrale in dobimo

$$\begin{aligned} P(|Y| > |Z|) &= \frac{1}{\pi} \left(\arctan\left(\frac{1}{n_d} \sqrt{\frac{n_d n_s}{1 + n_d + n_s}} \right) + \right. \\ &\quad \left. + \arctan\left(\frac{1 + 2n_d}{n_d} \sqrt{\frac{n_d n_s}{1 + n_d + n_s}} \right) \right). \end{aligned}$$

Za $n_s = n_d$ je rezultat 0,5, za $n_s = 250, n_d = 750$ pa 0,49. Če sta vzorca enako velika, bo ta način uvrščanja enote v vsako skupino uvrstil z verjetnostjo 0,5, kar je smiselno (saj kazalnik nič ne pove o kvaliteti). Čim pa vzorca nista enako velika, verjetnost ne bo več enaka 0,5 niti ne bo proporcionalna velikosti vzorca. Za neenake velikosti vzorcev (neuravnotežene podatke), to uvrščanje torej ne daje želenih oz. pričakovanih rezultatov (je pa res, da bodo morali biti podatki zelo neuravnoteženi, da bo verjetnost bistveno različna od 0,5).

Predlogi za vaje v R

- Generirajte vrednosti obeh slučajnih spremenljivk in narišite dobljene razlike. Pomagajte si s spodnjo kodo.

```
>  set.seed(1)
>  slabi <- rnorm(10)          #10 vrednosti kazalnika za slabe
>  dobri <- rnorm(10)          #10 vrednosti kazalnika za dobre
>  nov <- rnorm(1)             #ena vrednost za novega posameznika
>  plot(nov-mean(slabi),nov-mean(dobri),
+  ylim=c(-3,3),xlim=c(-3,3))  #razdalja vrednosti od obeh povprecij
>  abline(0,1)                 #simetrala
>  for(it in 1:99){            #ponovim 100x, vsakic znova simuliram
+  slabi <- rnorm(10)          #10 slabih in  10 dobrih
+  dobri <- rnorm(10)          # ter enega novega, ki ga na podlagi
+  nov <- rnorm(1)             # dobljenega kriterija uvrstim
+ points(nov-mean(slabi),nov-mean(dobri))
+ }
```

- Spreminjajte velikosti vzorcev in opazujte, kako se spreminja verjetnost uvrstitve v posamezno skupino. Preštejte vrednosti, ki so uvrščene v vsako skupino, in ocenite verjetnosti.

- Podatke generirajte še na primeru bolj smiselnega indeksa (ki ima različni povprečji v skupinah). Preverite, kako dobro sedaj uvrščate vrednosti.

Povzetek

- Naloga uporabi predhodno pridobljeno znanje za izpeljavo skupne porazdelitve dveh slučajnih spremenljivk. Verjetnost, ki nas zanima, nato izračunamo kot integral po ustreznem delu tega prostora, torej območju, kjer sta slučajni spremenljivki v želeni zvezi.

- V nalogi pokažemo, da pri neuravnoteženih podatkih (vzorca različne velikosti) sicer na videz smiselno uvrščanje ne daje pričakovanih rezultatov. Smiselnost rezultatov smo preverili za primer, ko kazalnik ne nosi nobene informacije o skupini - ker uvrščanje že v tem primeru ni smiselno, bo rezultate uvrščanja nemogoče interpretirati v primeru, ko bo kazalnik pomemben.

1.9 Centralni limitni izrek

V raziskavo o pojavih alergije pri dojenčkih so vključili vse otroke iz lokalnega zdravstvenega doma, ki so bili na nek datum stari manj kot eno leto, in jih nato spremljali pol leta. Kakšno povprečno starost otrok ob vključitvi v raziskavo pričakujete? Določite meje, v katerih bo s 95% verjetnostjo povprečna starost otrok na vzorcu, če so bili otroci v raziskavo zares izbrani naključno (predpostavimo, da so vsi rojstni dnevi v letu enako verjetni).

a) Kakšna je približna porazdelitev vzorčnega povprečja $\bar{X}$ za nek n?

Zaradi preprostosti vzemimo, da je starost X kar zvezna spremenljivka, torej enakomerno porazdeljena na intervalu [0,1]. Vsota enakomerno porazdeljenih spremenljivk očitno ni enakomerno porazdeljena, za izpeljevanje porazdelitve vsote bi lahko uporabili formulo za vsoto slučajnih spremenljivk (glej nalogo 1.4). Ker nam zadostuje približek, bomo ubrali lažjo pot in uporabili centralni limitni izrek, ki pravi, da bo porazdelitev vsote za dovolj velik n približno normalna. Izračunati moramo torej parametra te normalne porazdelitve, ker vemo, da velja $E(\bar{X}) = E(X)$ in $\mathrm{var}(\bar{X}) = \mathrm{var}X/n$, potrebujemo le pričakovano vrednost in varianco spremenljivke X. Uporabimo, da je gostota $f_X(x)$ enakomerne porazdelitve enaka 1 na intervalu [0,1] in 0 sicer, in dobimo

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} x f_X(x)dx = \int_0^1 x dx = \frac{x^2}{2}|_0^1 = \frac{1}{2} \\ \mathrm{var}(X) &= \int_{\mathbb{R}} (x - E(X))^2 f_X(x)dx = \int_0^1 (x - 0{,}5)^2 dx \\ &= \int_0^1 (x^2 - x + \frac{1}{4})dx = (\frac{x^3}{3} - \frac{x^2}{2} + \frac{x}{4})|_0^1 = \frac{1}{12} \end{aligned}$$

Porazdelitev $\bar{X}$ torej lahko aproksimiramo z $N(\frac{1}{2}, \frac{1}{12n})$.

b) Zapišite meje, v katerih bo s 95% verjetnostjo vzorčno povprečje, če v vzorec zajamemo 100 otrok.

Izračunali smo, da je približna porazdelitev povprečnega časa enaka $\bar{X} \sim N(0{,}5, 0{,}0008)$. Pri tej porazdelitvi se vrednosti spremenljivke s 95% verjetnostjo nahajajo v mejah $0{,}5 \pm 1{,}96 \cdot \sqrt{0{,}0008}$, torej $[0{,}44, 0{,}56]$.

Predlogi za vaje v R

- Generirajte velik vzorec iz enakomerne porazdelitve (`runif`) in na njem preverite izračun variance.
- Oglejte si porazdelitev vzorčnega povprečja za vzorce velikosti $n = 5, 20, 50$ in grafično s histogramom preverite, kako dobro se jim prilega ustrezna normalna porazdelitev.
- Simulirajte vrednosti iz različnih porazdelitev (npr. `rgamma`, `rchisq`, `rf`, `rt` ...) in izračunajte njihovo povprečje. Postopek ponovite velikokrat in si oglejte porazdelitev vzorčnega povprečja in na ta način s simulacijami preverite centralni limitni izrek.
- Poizkusite si izmisliti porazdelitev, pri kateri bo pri vzorcih velikosti $n = 50$ normalna porazdelitev še vedno slaba aproksimacija za porazdelitev vzorčnega povprečja.

Povzetek

- Centralni limitni izrek nam poda asimptotsko aproksimacijo porazdelitve vsote neodvisnih enako porazdeljenih slučajnih spremenljivk. Izrek je ključnega pomena v statistični teoriji, saj pomaga izpeljati porazdelitev neke vzorčne statistike, ne da bi morali poznati porazdelitev populacije, edini pogoj je, da se dà statistika zapisati kot vsota neodvisnih slučajnih realizacij iz te populacije. V tem primeru bomo porazdelitev vzorčne statistike lahko aproksimirali z normalno in to je tudi razlog, zakaj se normalna porazdelitev v statistiki tako pogosto pojavlja.
- Pri uporabi centralnega limitnega izreka se moramo seveda zavedati, da gre vedno le za približek, ki bo tem boljši, čim večji bo vzorec. Kako dobra je aproksimacija na majhnem vzorcu, bi lahko preverili s simulacijami.

- V nalogi 1.4 smo pokazali primer porazdelitve, pri kateri je vsota neodvisnih enako porazdeljenih slučajnih spremenljivk zopet porazdeljena z isto porazdelitvijo (spremenijo se le parametri). Po drugi strani nam centralni limitni izrek pove, da vsota neodvisnih enako porazdeljenih slučajnih spremenljivk konvergira proti normalni porazdelitvi. Z vsoto torej lahko ostajamo znotraj iste porazdelitve le pri tistih porazdelitvah, ki so pri določenih vrednostih parametrov zelo podobne normalni porazdelitvi oziroma konvergirajo proti normalni porazdelitvi.

1.10 Normalna aproksimacija binomske porazdelitve

V nekem kraju želijo zgraditi obvoznico, zanima jih delež ljudi, ki to gradnjo podpirajo. V ta namen izvedejo anketo. Recimo, da je verjetnost, da se posameznik strinja, enaka 0,65. Kolikšna je verjetnost, da bo med 6 posamezniki večina za gradnjo?

a) Naj bo $X = I\{\text{posameznik se strinja}\}$, X je Bernoullijevo porazdeljena spremenljivka. Kako je porazdeljena vsota $S_6 = \sum_{i=1}^{6} X_i$? Izračunajte pričakovano vrednost in standardni odklon.

Pričakovana vrednost Bernoullijeve spremenljivke je

$$\begin{aligned} E(X) &= 0 \cdot P(X=0) + 1 \cdot P(X=1) = p \\ \text{var}(X) &= E(X^2) - E(X)^2 = 1^2 \cdot P(X=1) - p^2 = p - p^2 = p(1-p) \end{aligned}$$

za vsoto pa velja

$$\begin{aligned} E(S_n) &= E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = np \\ \text{var}(S_n) &= \text{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p) \end{aligned}$$

Vsota neodvisnih enako Bernoullijevo porazdeljenih spremenljivk je porazdeljena binomsko (glej nalogo 1.3), verjetnost posameznega izida je

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

b) Izračunajte verjetnost, da so vsaj 4 posamezniki za gradnjo.

Z uporabo formule za binomsko porazdelitev:

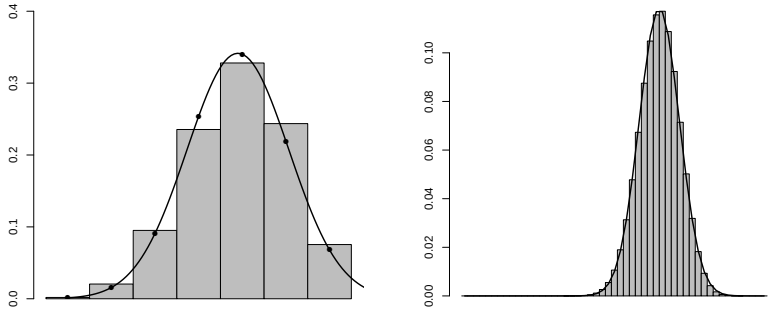
$$P(S_n \geq 4) = \sum_{k=4}^{6} \binom{6}{k} (0{,}65)^k (0{,}35)^{6-k} = 0{,}647$$

c) Aproksimirajte to verjetnost še s pomočjo centralnega limitnega izreka.

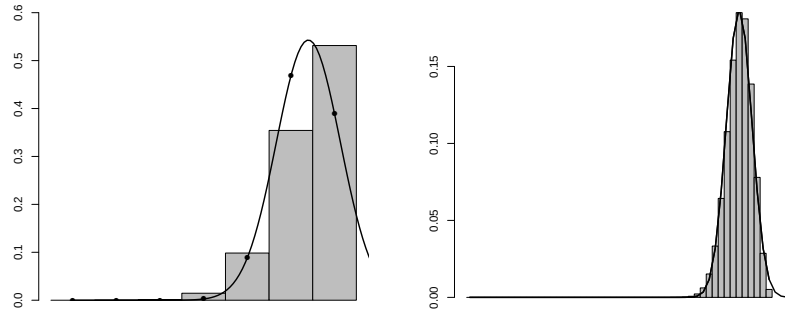
Vemo, da $\frac{S_n - np}{\sqrt{np(1-p)}}$ konvergira (v porazdelitvi) proti standardno normalno porazdeljeni spremenljvki Z. Poglejmo, kako dobra bo aproksimacija z normalno porazdelitvijo pri $n = 6$:

$$\begin{aligned} P(S_n > 3{,}5) &= P(\frac{S_n - np}{\sqrt{np(1-p)}} > \frac{3{,}5 - np}{\sqrt{np(1-p)}}) \\ &= P(Z > \frac{3{,}5 - 3{,}9}{1{,}17}) = 0{,}634 \end{aligned}$$

Na slikah 1.9 in 1.10 je prikazana kvaliteta aproksimacije za dve vrednosti p. Vidimo, da je aproksimacija nekoliko slabša, če je porazdelitev bolj asimetrična, a še vedno zelo dobra.



Slika 1.9: *Aproksimacija binomske porazdelitve z normalno za* $p = 0{,}65$ *in (a)*$n = 6$*, (b)*$n = 500$.



Slika 1.10: *Aproksimacija binomske porazdelitve z normalno za* $p = 0{,}9$ *in (a)*$n = 6$*, (b)*$n = 500$.

Predlogi za vaje v R

- Oglejte si, kako dobra je aproksimacija za različne velikosti vzorca in različne vrednosti p:

```
> win.graph(height=6,width=12)  #pripravimo okno za risanje
> par(mfrow=c(1,2))             #narisali bomo dva grafa na isto sliko
> p <- 0.65                     #verjetnost, da je posameznik za
> n <- 6                        #velikost vzorca
> verb <- dbinom(0:n,n,p)       #verjetnosti posameznih izidov
> barplot(verb,space=0,width=1,ylim=c(0,.4)) #stolpicni diagram
> sd <- sqrt(n*p*(1-p))         #standardni odklon
> e <- n*p                      #pricakovana vrednost
> vern <- dnorm(0:n,e,sd)       #vrednost gostote v posameznih tockah
> points(0:n+.5,vern,pch=16)    #dorisemo vrednosti na graf
> sek <- seq(0,n+1,length=100)  #izberemo tocke za racunanje gostote
> vern <- dnorm(sek,e,sd)       #vrednost gostote v izbranih tockah
> lines(sek+.5,vern,lwd=2)      #dorisemo krivuljo na graf
##################
> n <- 50                       #se enkrat za vecji vzorec
> verb <- dbinom(0:n,n,p)       #verjetnosti posameznih izidov
> barplot(verb,space=0,width=1) #stolpicni diagram
> sd <- sqrt(n*p*(1-p))         #standardni odklon
> e <- n*p                      #pricakovana vrednost
> vern <- dnorm(0:n,e,sd)       #vrednost gostote v posameznih tockah
> lines(0:n+.5,vern,lwd=2)      #dorisemo vrednosti na graf
```

Povzetek

- Centralni limitni izrek nam poda asimptotsko aproksimacijo porazdelitve neke spremenljivke, ki se jo dà zapisati kot vsoto neodvisnih enako porazdeljenih slučajnih spremenljivk. V tej nalogi smo porazdelitev vsote poznali in smo izrek uporabili le kot pripomoček (približek) za lažje oziroma hitrejše računanje.

- Da računanje z binomsko porazdelitvijo ne bi bilo prezamudno, smo uporabili zelo majhen n. Navkljub temu je bil približek z normalno porazdelitvijo že kar soliden, seveda bi bil za večje vrednosti n še toliko boljši.

Poglavje 2

Vzorčenje

Statistični del knjige pričenjamo s poglavjem o vzorčenju (Rice, 2009, poglavje 7), ki bo služilo kot osnova za razumevanje lastnosti cenilk. Z izrazom cenilka označujemo funkcijo slučajnih spremenljivk (vrednosti na vzorcu), s katero želimo oceniti populacijsko količino, ki nas zanima. Pri vsaki cenilki nas bo zanimalo predvsem dvoje - ali je cenilka nepristranska in kolikšna je njena varianca. Poiskali bomo nepristranske cenilke za različne primere populacij in načine vzorčenja - začeli bomo z neskončno populacijo, saj slučajno vzorčenje iz take populacije da neodvisne enako porazdeljene slučajne spremenljivke. Nato si bomo ogledali, kako se izrazi za cenilke in njihove variance spremenijo, če se med spremenljivkami pojavi odvisnost bodisi zaradi končne populacije bodisi zaradi dejstva, da so enote vzorčene iz posameznih skupin s specifičnimi lastnostmi.
V večini nalog se bomo osredotočili na ocenjevanje povprečja - proučevali bomo lastnosti intuitivno smiselnih cenilk populacijskega povprečja. Naša prva naloga bo pokazati, da je cenilka nepristranska, torej da je njena pričakovana vrednost enaka povprečju v populaciji. Naslednji korak je poiskati standardno napako te ocene, skušali jo bomo izraziti z varianco populacije ter velikostjo vzorca. Tako izraženo varianco lahko potem primerjamo med alternativnimi cenilkami in ugotavljamo, katera ima najmanjšo. Zadnji korak večine nalog je poiskati še cenilko variance, torej ugotoviti, kako njeno velikost lahko ocenimo s pomočjo vzorca.

2.1 Vzorčenje - neskončna populacija

Oceniti želimo znižanje vrednosti pritiska pri pacientih z esencialno hipertenzijo po treh mesecih jemanja nekega zdravila. V ta namen smo zbrali vzorec 25 bolnikov, naj bo X_i vrednost razlike pri i-tem bolniku našega vzorca. Predpostavimo, da so slučajne spremenljivke X_i neodvisne in enako porazdeljene.

a) Pokažite, da je povprečje našega naključnega vzorca nepristranska ocena povprečnega znižanja v populaciji bolnikov (to označimo z μ).

   Povprečje vzorca je $\frac{1}{n}\sum_{i=1}^{n} X_i$, povprečje populacije označimo z μ. Predpostavljamo, da je vzorec naključen, torej da so vrednosti X_i enako porazdeljene, za vse velja $E(X_i) = \mu$.

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

b) Kaj lahko rečemo o $cov(X_i, X_j)$, če je $i \neq j$?

   Ker so vrednosti bolnikov med seboj neodvisne, je kovarianca enaka 0.

c) Naj bo varianca v populaciji enaka σ^2. Kolikšna je varianca (oz. standardna napaka) naše cenilke?

$$\begin{aligned}
\operatorname{var}\left(\bar{X}\right) &= \operatorname{cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{cov}\left(X_i, \sum_{j=1}^{n} X_j\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left[\operatorname{cov}\left(X_i, X_i\right) + \sum_{j=1, j\neq i}^{n} \operatorname{cov}\left(X_i, X_j\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left[\operatorname{cov}\left(X_i, X_i\right) + (n-1)\operatorname{cov}\left(X_i, X_j\right)\right]
\end{aligned}$$

Uporabimo, da so vrednosti med seboj neodvisne, torej da je $\text{cov}[X_i, X_j] = 0$ za vsak $i \neq j$.

$$\begin{aligned}
\text{var}\left(\bar{X}\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\text{cov}(X_i, X_i)\right] \\
&= \frac{1}{n^2}\cdot n\text{cov}(X_i, X_i) = \frac{1}{n}\text{var}(X_i) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Standardna napaka je enaka $SE = \frac{\sigma}{\sqrt{n}}$.

d) Na podlagi našega vzorca želimo oceniti σ^2. Naj bo naša cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$. Kolikšna mora biti vrednost konstante c, da bo naša cenilka nepristranska?

Vemo, da velja $\sigma^2 = E(X^2) - E(X)^2$, torej $E(X^2) = \sigma^2 + \mu^2$ in podobno tudi $SE^2 = \text{var}(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - E(\bar{X})^2$, torej $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$. Pri prehodu iz prve v drugo vrstico upoštevamo, da velja $\sum_i X_i\bar{X} = \bar{X}\sum_i X_i = n\bar{X}^2 = \sum_i \bar{X}^2$:

$$\begin{aligned}
E\left[c\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= cE\left[\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)\right] \\
&= cE\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right] \\
&= cE\left[\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)\right] \\
&= c\sum_{i=1}^{n}\left[E(X_i^2) - E(\bar{X}^2)\right] \\
&= c\sum_{i=1}^{n}\left[\left(\mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= cn\left[\sigma^2\left(1 - \frac{1}{n}\right)\right] \\
&= \sigma^2(n-1)c
\end{aligned}$$

Ker želimo, da velja $E(\widehat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1}$.

e) Na vzorcu smo dobili naslednje rezultate: $\bar{x} = 4$, $\hat{\sigma} = 20$. Ocenite standardno napako (torej standardni odklon vzorčnega povprečja). Ali boste na podlagi podatkov lahko trdili, da se tlak zniža tudi v populaciji?

Ocena standardne napake je enaka $\widehat{SE} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{20}{5} = 4$. Znižanje pritiska je enako standardni napaki - prave razlike ne moremo razločiti od naključne variabilnosti. V ta namen bi potrebovali precej večji vzorec.

f) Kako bi s pomočjo dobljenih podatkov zapisali 95% interval zaupanja za populacijsko povprečje?

Interval zaupanja je verjetnostni interval, za njegov izračun moramo poznati porazdelitev cenilke. Če bi bile vrednosti pritiska v populaciji normalno porazdeljene, bi bilo tudi vzorčno povprečje normalno porazdeljeno (naloga 1.6). Ker smo varianco cenilke ocenili iz podatkov, bi za meje intervala zaupanja morali uporabiti porazdelitev t (Rice, 2009, poglavje 6). Če pa porazdelitev v populaciji ni znana, si pomagamo s centralnim limitnim izrekom, ki nam pove, da lahko porazdelitev vzorčnega povprečja aproksimiramo z normalno. Dobimo torej (glej npr. Rice, 2009, razdelek 7.3.3):

$$\bar{X} \pm z_{\alpha/2} \cdot \widehat{SE} = 4 \pm 1{,}96 \cdot 4 = [-3{,}84, 11{,}84]$$

Predlogi za vaje v R

- Generirajte vzorce velikosti 30 iz enakomerne porazdelitve, vsakič izračunajte povprečje ter si oglejte porazdelitev tega povprečja. Ocenite pričakovano vrednost za povprečje in standardno napako ter oceni primerjajte s teoretičnimi vrednostmi.
- Postopek ponovite še za druge (morda bolj asimetrične) porazdelitve.

Povzetek

- Medsebojna neodvisnost enot v primeru neskončne populacije močno pripomore k preprostosti izpeljav.
- Nepristransko cenilko za varianco smo izpeljali tako, da smo najprej zapisali intuitivno smiseln izraz, nato pa določili še konstanto, pri kateri je ta cenilka nepristranska.

- Če je populacija neskončna in enote v vzorcu neodvisne, velja:

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n};\ \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Da bi na podlagi vzorca lahko napovedovali vrednost povprečja v populaciji, bi seveda morali poznati celotno porazdelitev in ne le pričakovane vrednosti ter variance vzorčnega povprečja. Če so vrednosti X normalno porazdeljene, je vzorčno povprečje prav tako normalno porazdeljeno, da bi lahko računali poljubne verjetnosti, moramo zato izračunati le parametra te porazdelitve. A tudi v splošnem lahko porazdelitev vzorčnega povprečja aproksimiramo z normalno porazdelitvijo - tu ključno vlogo odigra centralni limitni izrek.

2.2 Vzorčenje - končna populacija

Oceniti želimo povprečno število zaposlenih v podjetjih neke panoge ob začetku letošnjega leta. Panoga je razdeljena na podskupine, v eni izmed skupin je le 11 podjetij. Uspeli smo pridobiti podatke za naključen vzorec šestih izmed teh podjetij. Naj bo X_i število zaposlenih v i-tem podjetju našega vzorca, μ naj označuje njihovo povprečje, σ pa standardni odklon.

a) Naj X_1 in X_2 označujeta vrednosti prvih dveh naključno izbranih podjetij. Kaj lahko rečemo o $\text{cov}(X_1, X_2)$? Kaj pa za splošen $i \neq j$?

Populacija je končna, označimo vrednosti v populaciji z x_k, $k = 1, \ldots, 11$. Po nekem vrstnem redu izberimo vseh 11 podjetij, prvih 6 naj predstavlja vzorec, X_i označuje število zaposlenih v i-tem izbranem podjetju. Ker je vsak izmed X_i ena od vrednosti x_k in imajo vse enako verjetnost, je $\text{cov}(X_1, X_2) = \text{cov}(X_i, X_j)$ za poljubna različna i in j. Vendar pa sedaj X_i in X_j nista neodvisni slučajni spremenljivki - če je $X_i = x_k$, X_j ne more zavzeti k-te vrednosti.

$$\text{cov}\left(X_i, \sum_{j=1}^{N} X_j\right) = \text{cov}\left(X_i, \sum_{k=1}^{N} x_k\right) = 0$$

Ker je vsota vseh vrednosti konstanta, je zgornji izraz enak 0, torej

$$\text{cov}\left(X_i, \sum_{j=1}^{N} X_j\right) = \text{cov}\,(X_i, X_i) + (N-1)\text{cov}\,(X_i, X_j) = 0$$

in zato (za $i \neq j$)

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Spremenljivki sta torej negativno povezani.

b) Izračunajte še korelacijo $\text{cor}(X_i, X_j)$ za neka $i \neq j$. Od česa je odvisna?

Korelacija med spremenljivkama je enaka

$$\text{cor}(X_i, X_j) = -\frac{\sigma^2}{(N-1)\sigma^2} = -\frac{1}{N-1}$$

Korelacija je torej odvisna izključno od velikosti populacije in z večanjem populacije hitro postane zelo majhna.

c) Izračunajte standardno napako vzorca velikosti n.

$$\begin{aligned}
\text{var}[\bar{X}] &= \text{cov}[\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \cdot \text{cov}[X_i, \frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\{\text{cov}[X_i, X_i] + (n-1)\text{cov}[X_i, X_j]\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\{\sigma^2 - (n-1)\frac{\sigma^2}{N-1}\} = \frac{\sigma^2}{n}\frac{N-n}{(N-1)}
\end{aligned}$$

Standardna napaka je torej enaka $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

d) V drugi panogi imamo 100 podjetij. Kako velik vzorec moramo vzeti iz te panoge, da bomo dobili približno enako veliko standardno napako (privzemimo, da je varianca tudi v tej panogi enaka σ^2)? Kaj pa če bi imeli panogo z zelo velikim številom podjetij?

Za $N = 11$ in $n = 6$ dobimo $SE^2 = \frac{\sigma^2}{6}\frac{11-6}{(10)} = \frac{\sigma^2}{12}$. Pri populaciji velikosti 100 nam vzorec velikosti 10 da standardno napako $SE^2 = \frac{\sigma^2}{11}$, vzorec velikosti 11 pa standardno napako $SE^2 = \frac{\sigma^2}{12,2}$.
Če je populacija večja, bomo za enako standardno napako torej potrebovali več enot. Če bi imeli skupino z zelo velikim številom podjetij, bi bil izraz $\frac{N-n}{(N-1)}$ približno enak 1, zato bi potrebovali 12 podjetij. Velikost potrebnega vzorca se z večanjem populacije veča, vendar pa to večanje kmalu ni več bistveno.

e) Kolikšna mora biti vrednost konstante c, da bo cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$ nepristransko ocenila vrednost σ^2?

Ponovimo izračun iz zadnje točke prejšnje naloge, upoštevamo, da je $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1}$:

$$\begin{aligned} E[c\sum_{i=1}^{n}(X_i - \bar{X})^2] &= cE\left[\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)\right] \\ &= c\sum_{i=1}^{n}\left[\left(\mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1}\right)\right] \\ &= cn\left[\sigma^2(1 - \frac{N-n}{n(N-1)})\right] \\ &= \sigma^2 c\frac{N(n-1)}{N-1} \end{aligned}$$

Ker želimo, da velja $E(\widehat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1}\frac{N-1}{N}$, torej

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

f) Zapišite še nepristransko cenilko za varianco povprečja.

Združimo dosedanje rezultate in dobimo

$$\begin{aligned} \widehat{SE}^2 &= \frac{\widehat{\sigma}^2(N-n)}{N-1} = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2\frac{(N-n)}{N-1} \\ &= \frac{\widehat{\sigma}^2}{n}\frac{N-n}{N} \end{aligned}$$

kjer je $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Predlogi za vaje v R

- Izmislite si 11 vrednosti ter nato generirajte vzorce velikosti 6. Oglejte si porazdelitev vzorčnih povprečij. Pokažite, da gornja formula predstavlja nepristransko oceno populacijske variance.

Povzetek

- Če je populacija končna, vrednosti enot izbranih v vzorec navkljub naključnemu izbiranju med seboj niso neodvisne, kovarianca med enotami je enaka:

$$\mathrm{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

- Varianca vzorčnega povprečja in nepristranska cenilka za varianco v populaciji sta enaki

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{(N-1)};\ \widehat{\sigma}^2 = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- V končni populaciji bo na izbiro velikosti vzorca vplivala tudi velikost populacije. Vendar pa bo ta vpliv pomemben le pri sorazmerno majhnih populacijah oziroma takrat, ko velikost vzorca predstavlja nezanemarljiv delež velikosti populacije.

2.3 Določitev načrta vzorčenja

Zanima nas povprečna teža bolnikov (μ) s hipertenzijo v starostni skupini 60 do 80 let. Težo bi radi ocenili na podlagi vzorca, jasno je, da bo teža precej različna pri moških (μ_1) kot pri ženskah (μ_2). Čas in denar, ki ju imamo na voljo za raziskavo, nam dopuščata vzorec velikosti 100. Vemo, da se v populaciji deleža moških in žensk s hipertenzijo razlikujeta, delež moških označimo z d. Zanima nas, kolikšen delež moških in kolikšen delež žensk naj naberemo v vzorec, da bo standardna napaka naše ocene najmanjša možna. Pri tem predpostavimo, da je standardni odklon teže moških k-krat večji od standardnega odklona teže žensk.

a) Zapišite nepristransko cenilko za populacijsko povprečje.

Populacijsko povprečje μ lahko zapišemo kot $\mu = d\mu_1 + (1-d)\mu_2$. Ker velja $E(\bar{X}_1) = \mu_1$ in $E(\bar{X}_2) = \mu_2$, je nepristranska cenilka enaka

$$M = d\bar{X}_1 + (1-d)\bar{X}_2$$

b) Standardno napako ocene izrazite z velikostima podvzorcev (n_1 je število moških, n_2 število žensk).

Ker je povprečje pri moških neodvisno od povprečja pri ženskah, velja

$$\text{var}(M) = d^2 var(\bar{X}_1) + (1-d)^2 var(\bar{X}_2) = d^2 \frac{\sigma_1^2}{n_1} + (1-d)^2 \frac{\sigma_2^2}{n_2}$$

c) Naj velja $\sigma_1 = k\sigma_2$. Pri kakšni razdelitvi vzorca je standardna napaka najmanjša? Izračunajte n_1 za primera $k = 1$ in $k = 2$, vzemite, da je delež moških enak 0,7.

Varianco vzorčnega povprečja zapišemo kot

$$var(M) = \sigma^2 \left(\frac{d^2 k^2}{n_1} + \frac{(1-d)^2}{n - n_1} \right)$$

Minimizirati moramo torej izraz v oklepaju. Odvajamo po n_1 in izenačimo z 0

$$\begin{aligned} -\frac{d^2k^2}{n_1^2} + \frac{(1-d)^2}{(n-n_1)^2} &= 0 \\ -(n-n_1)^2 d^2 k^2 + n_1^2 (1-d)^2 &= 0 \\ (d^2k^2 - (1-d)^2) n_1^2 - 2nd^2k^2 n_1 + n^2 d^2 k^2 &= 0 \end{aligned}$$

Rešimo kvadratno enačbo in dobimo rešitev (druga rešitev je večja od n in tako nesmiselna):

$$n_1 = \frac{ndk}{dk + 1 - d}$$

Za $k = 1$ dobimo rešitev $n_1 = nd$, v konkretnem primeru torej $n_1 = 70$ in $n_2 = 30$. Če sta deleža moških in žensk na vzorcu enaka deležema v

populaciji, je standardna napaka najmanjša možna.
Če je standardni odklon teže moških 2x večji od standardnega odklona žensk, dobimo $n_1 = 82{,}4$, vzeti moramo torej ustrezno več posameznikov iz bolj variabilne skupine.

d) Ali porazdelitev cenilke M konvergira proti normalni?

V splošnem ne, to smo že pokazali v nalogi 1.1e). Če sta povprečji stratumov različni, bo porazdelitev cenilke odvisna od deleža d.

Predlogi za vaje v R

- Izmislite si smiselne vrednosti za parametre v nalogi ter generirajte podatke. Grafično prikažite, kako se pri različnih izbirah velikosti vzorcev spreminja kvaliteta vaše ocene.
- Poglejte si kaj se dogaja s porazdelitvijo cenilke za povprečje, ko se vzorec veča.

Povzetek

- Velikost vzorca ne vpliva na pristranskost ocene, je pa ključnega pomena pri njeni standardni napaki. Pri stratificiranem vzorčenju na standardno napako vpliva velikost vzorcev in variabilnost v posameznih stratumih. Če je variabilnost v stratumih enaka, si najmanjšo standardno napako zagotovimo s proporcionalnim vzorčenjem, torej takim, pri katerem je delež enot izbranih v podvzorec enak deležu enot podskupine v populaciji.
  Če variabilnost v stratumih ni enaka, moramo iz bolj variabilnih podskupin vzeti ustrezno večji vzorec.
- Če je stratumov več, je izpeljava nekoliko težja - minimizirati je potrebno po več spremenljivkah hkrati. Izpeljava z vezanim ekstremom (omejitev $n_1 + \ldots + n_k = n$) je podana v knjigi Rice (2009), razdelek 7.5.3.
- Opisani vzorčni načrt je primer stratificiranega vzorčenja: populacija je razdeljena na več skupin (stratumov), vsak vzorec je sestavljen tako, da je vanj izbranih nekaj elementov vsakega stratuma. Vzorčna shema, pri kateri najprej naključno izberemo nekaj skupin in nato vzorčimo le iz njih, je primer večstopenjskega vzorčenja, srečali jo bomo v nalogi 2.6.

2.4 Ocena kovariance

V nekem podjetju velikosti N so izvedli izobraževanje za naključen vzorec n zaposlenih. Ob koncu izobraževanja so novo znanje preverili s testom. Podjetje se želi odločiti, ali je smiselno uvesti izobraževanje za vse zaposlene, zato jih zanima povezanost med starostjo zaposlenega (X_i) in rezultatom na testu (Y_i).
Za vsakega posameznika iz vzorca imamo torej par slučajnih spremenljivk (X_i, Y_i), $i = 1 \dots n$.

a) Utemeljite, da je količina $\mathrm{cov}(X_i, Y_j)$ za poljubna $i \neq j$ enaka.

   Vzorčenje si lahko predstavljamo tako, da smo populacijo naključno uredili, nato pa v vzorec zajeli prvih n posameznikov. Ker imajo vsi vrstni redi enako verjetnost, bo na i-tem mestu z enako verjetnostjo katerikoli posameznik. Vsi pari (X_i, Y_j) imajo tako enako porazdelitev in zato je enaka tudi kovarianca X_i in Y_j.

b) Naj bo $\gamma = \mathrm{cov}(X_i, Y_i)$. Izračunajte kovarianco $\mathrm{cov}(X_i, Y_j)$ za $i \neq j$.

   Uporabimo isto idejo kot pri prejšnji nalogi - vsota vseh vrednosti iz populacije je konstanta, zato velja

$$\mathrm{cov}(X_i, \sum_{j=1}^{N} Y_j) = \mathrm{cov}(X_i, Y_i) + (N-1)\mathrm{cov}(X_i, Y_j) = 0$$

   Velja torej (za $i \neq j$)

$$\mathrm{cov}(X_i, Y_j) = -\frac{\gamma}{N-1}$$

c) Kovarianca med spremenljivkama X in Y je definirana kot

$$\mathrm{cov}(X, Y) = \mathrm{cov}(X_1, Y_1) = \frac{1}{N}\sum_{i=1}^{N}[(x_i - \mu)(y_i - \nu)] = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \mu\nu$$

   kjer smo z μ in ν označili povprečji $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ in $\nu = \frac{1}{N}\sum_{i=1}^{N} y_i$.

S pomočjo vzorca bi kovarianco radi ocenili s cenilko

$$\widehat{\gamma} = c\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right]$$

Določite vrednost konstante c, da bo cenilka nepristranska.

Pričakovana vrednost cenilke je

$$E(\widehat{\gamma}) = c\left[\sum_{i=1}^{n} E(X_i Y_i) - nE(\bar{X}\bar{Y})\right] \tag{2.1}$$

Zaradi simetrije je $E(X_i Y_i) = E(X_j Y_j)$ za poljubna i in j. Vemo, da velja

$$\mathrm{cov}(X_i, Y_i) = E(X_i Y_i) - E(X_i)E(Y_i) = E(X_i Y_i) - \mu\nu$$

Torej je $E(X_i Y_i) = \mu\nu + \gamma$. Oglejmo si še drugi člen na desni strani (2.1):

$$\begin{aligned}
E(\bar{X}\bar{Y}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{j=1}^{n} Y_j\right] \\
&= \frac{1}{n^2} E\sum_{i=1}^{n}\left[X_i Y_i + X_i \sum_{j=1, i\neq j}^{n} Y_j\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E(X_i Y_i) + \sum_{j=1, i\neq j}^{n} E(X_i Y_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E(X_i Y_i) + (n-1)E(X_i Y_j)\right]
\end{aligned}$$

Uporabimo rezultat

$$\begin{aligned}
\mathrm{cov}(X_i, Y_j) &= E(X_i Y_j) - E(X_i)E(Y_i) \\
E(X_i Y_j) &= \mu\nu - \frac{\gamma}{N-1}
\end{aligned}$$

in zato

$$\begin{aligned} E(\bar{X}\bar{Y}) &= \frac{1}{n^2} n \left[\mu\nu + \gamma + (n-1)(\mu\nu + \frac{-\gamma}{N-1}) \right] \\ &= \frac{1}{n} \left[n\mu\nu + \gamma(1 - \frac{(n-1)}{N-1}) \right] \\ &= \frac{1}{n} \left[n\mu\nu + \gamma \frac{N-n}{N-1} \right] \end{aligned}$$

To vstavimo v enačbo (2.1)

$$\begin{aligned} E(\widehat{\gamma}) &= c \left[\sum_{i=1}^{n} (\mu\nu + \gamma) - n \frac{1}{n} \left[n\mu\nu + \gamma \frac{N-n}{N-1} \right] \right] \\ &= c \left[n\mu\nu + n\gamma - n\mu\nu - \gamma \frac{N-n}{N-1} \right] \\ &= c \left[n\gamma - \gamma \frac{N-n}{N-1} \right] \\ &= c\gamma \left[\frac{nN-n}{N-1} - \frac{N-n}{N-1} \right] \\ &= c\gamma \frac{N(n-1)}{N-1} \end{aligned}$$

c mora biti torej enak $\frac{1}{n-1}\frac{N-1}{N}$.

d) Kako bi ocenili korelacijo? Kaj vemo o nepristranskosti te cenilke?

Uporabimo izpeljane formule za oceno kovariance in varianc:

$$\begin{aligned} \widehat{\rho} &= \frac{\widehat{\mathrm{cov}}(X,Y)}{\widehat{\sigma}_X \widehat{\sigma}_Y} \\ &= \frac{\frac{1}{n-1}\frac{N-1}{N} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\frac{N-1}{N} \sum_{i=1}^{n} (X_i - \bar{X})^2 \frac{1}{n-1}\frac{N-1}{N} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \\ &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \end{aligned}$$

Nepristranskosti te ocene nismo dokazali, saj pričakovana vrednost kvocienta ni enaka kvocientu pričakovanih vrednosti.

Predlogi za vaje v R

- Ker ne vemo, ali je ocena korelacije pristranska ali ne, pristranskost preverimo s simulacijo.
  Vzamemo populacijo velikosti $N = 300$, vzorci naj bodo velikosti $n = 10$. Naj bo X starost porazdeljena enakomerno med 25 in 65, uspeh na testu pa negativno povezan s starostjo, tako da je v povprečju enak $100 - starost$ (predpostavimo, da so odstopanja od tega povprečja razpršena s standardnim odklonom 20 in normalno porazdeljena).

```
> set.seed(1)
> xi <- runif(300)*40+25            #300 oseb starosti 25-65 let
> yi <- 100 - xi + rnorm(300)*20    #rezultat testa za populacijo
> cov(xi,yi)                        #kovarianca v populaciji
[1] -136.8110
> cor(xi,yi)                        #korelacija v populaciji
[1] -0.5207052

> runs <- 10000                     #stevilo korakov simulacije
> cova <- cora <- rep(NA,runs)      #sem bomo zapisali rezultate
> for(it in 1:runs){                #simulacija po korakih
+ inx <- sample(1:length(xi),size=10,replace=F) #vzorec n=10
+ xa <- xi[inx]                     #pogledamo njihove starosti
+ ya <- yi[inx]                     #pogledamo njihove rezultate
+ cova[it] <- 1/9*299/300*
+   sum( (xa-mean(xa))*(ya-mean(ya))) #izracunamo kovarianco
+ cora[it] <-  sum( (xa-mean(xa))*(ya-mean(ya)))/
+   sqrt(sum( (xa-mean(xa))^2)*sum((ya-mean(ya))^2)) #korelacija
+ }

> mean(cova)                        #povprecna kovarianca
[1] -135.4745
> mean(cora)                        #povprecna korelacija
[1] -0.5034081
```

- Vidimo, da sta obe vrednosti po absolutni vrednosti nekoliko manjši od populacijskih, preverimo, ali je odstopanje veliko glede na standardno napako, ki jo lahko pričakujemo pri takem številu simulacij.
  Zanima nas, ali povprečna kovarianca (`mean(cova)`) bistveno odstopa od prave vrednosti (`cov(xi,yi)`). Povprečna kovarianca je slučajna spremenljivka, če bomo vnovič pognali simulacijo (vseh 10000 korakov), bomo dobili drugo vrednost. Predpostavimo, da je približno normalno porazdeljena, ocenimo njeno varianco (varianca povprečja n i.i.d

spremenljivk je varianca spremenljivk deljeno z n, pri nas je n število korakov simulacije). Ničelna domneva, ki jo preverjamo, je: H_0: povprečna kovarianca je enaka populacijski vrednosti. Odstopanje od te ničelne domneve preverjamo s testom t.

```
> (mean(cova)-cov(xi,yi))/sqrt(var(cova)/runs)
[1] 1.509540
```

Ta rezultat je v okviru pričakovanj, saj smo teoretično pokazali, da je ocena kovariance nepristranska. Enako ponovimo za korelacijo:

```
> (mean(cora)-cor(xi,yi))/sqrt(var(cora)/runs)
[1] 6.66459
```

Odstopanje pri korelaciji je bistveno večje, verjamemo, da se v naši simulaciji ni zgodilo po naključju, temveč je ocena dejansko pristranska.

Povzetek

- Cilj naloge je bil oceniti povezanost med vrednostima dveh slučajnih spremenljivk na isti enoti neke populacije. Ker je populacija v našem primeru končna, je potrebno pri tej oceni upoštevati tudi kovarianco med vrednostmi na različnih enotah. Pokazali smo nepristranskost ocene kovariance, ne pa tudi nepristranskosti ocene korelacije. Za slednjo se izkaže, da je pristranska, in sicer nekoliko podceni pravo vrednost (je bližje 0).

2.5 Enostavni vzorec iz končne populacije, še enkrat

Vzemimo še enkrat enostavni slučajni vzorec velikosti n iz populacije N, vrednosti v populaciji označimo z x_i; $i = 1, \ldots, N$, populacijsko vrednost povprečja označimo z μ, variance pa z σ^2. Definirajmo slučajno spremenljivko $I_i = I_{[i \text{ je izbran v vzorec}]}$ in zapišimo cenilko populacijskega povprečja μ kot $C = \frac{1}{n} \sum_{i=1}^{N} I_i x_i$.

a) Koliko je vsota $\sum_{i=1}^{N} I_i$? Kolikšna je verjetnost $P(I_i = 1)$?

Vsota $\sum_{i=1}^{N} I_i = n$, saj smo vzeli vzorec velikosti n. Izračunajmo še verjetnost, da bo izbran element i:
izbiramo vzorce velikosti n in iz populacije velikosti N. Vseh možnih kombinacij je $\binom{N}{n}$, kakšno je število tistih vzorcev, v katerih je element i? Pri teh vzorcih en element že poznamo, izmed ostalih $N-1$ smo jih izbrali $n-1$. Torej je iskana verjetnost enaka:

$$P(I_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

b) Pokažite, da je cenilka nepristranska.

Radi bi ocenili $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$E(C) = \frac{1}{n}\sum_{i=1}^{N} E(I_i)x_i$$

Ker lahko I_i zavzame le vrednosti 0 in 1, je $E(I_i) = P(I_i = 1) = \frac{n}{N}$ (vzorec je slučajen, zato so verjetnosti za vse i enake), zato dobimo

$$E(C) = \frac{1}{n}\sum_{i=1}^{N} \frac{n}{N}x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$$

c) Izračunajte $\mathrm{var}(I_i)$ in $\mathrm{cov}(I_i, I_j)$.

Spremenljivka I_i je Bernoullijeva z verjetnostjo $P(I_i = 1) = \frac{n}{N}$. Njena varianca je zato enaka

$$\mathrm{var}(I_i) = \frac{n}{N}(1 - \frac{n}{N}) = \frac{n}{N}\frac{N-n}{N}$$

Kovarianco izračunamo tako, da upoštevamo, da je $\mathrm{cov}(I_1, I_1 + \ldots I_N) = \mathrm{cov}(I_1, n) = 0$ in $\mathrm{cov}(I_i, I_j)$ je enaka za vsak $i \neq j$:

$$\mathrm{cov}(I_i, I_j) = -\frac{\frac{n}{N}\frac{N-n}{N}}{N-1} = -\frac{n(N-n)}{N^2(N-1)}$$

d) Pokažite še, da je varianca tako zapisane cenilke enaka $\text{var}(C) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$.

$$\begin{aligned}
\text{var}(C) &= \frac{1}{n^2}\text{cov}(\sum_{i=1}^{N} I_i x_i, \sum_{j=1}^{N} I_j x_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{N} cov(I_i x_i, \sum_{j=1}^{N} I_j x_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{N}\left[\text{cov}(I_i x_i, I_i x_i) + \sum_{j=1, j\neq i}^{N} cov(x_i I_i, I_j x_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N}\left[x_i^2\text{cov}(I_i, I_i) + \sum_{j=1, j\neq i}^{N} x_i x_j \text{cov}(I_i, I_j)\right]
\end{aligned}$$

Populacijska varianca je definirana kot:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

zato velja $\sum_{i=1}^{N} x_i^2 = N(\sigma^2 + \mu^2)$. V izpeljavi variance bomo potrebovali še naslednji rezultat:

$$(\sum_{i=1}^{N} x_i)^2 = \sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} x_i x_j \;\Rightarrow\; \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} x_i x_j = N^2\mu^2 - \sum_{i=1}^{N} x_i^2$$

Sedaj to uporabimo in dobimo

$$\begin{aligned}
\mathrm{var}(C) &= \frac{1}{n^2}\sum_{i=1}^{N}\left[x_i^2\mathrm{var}(I_i) - \sum_{j=1,j\neq i}^{N} x_i x_j \frac{\mathrm{var}(I_i)}{N-1}\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\left[(N-1)\sum_{i=1}^{N} x_i^2 - (N^2\mu^2 - \sum_{i=1}^{N} x_i^2)\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\left[N\sum_{i=1}^{N} x_i^2 - N^2\mu^2\right] \\
&= \frac{N^2\mathrm{var}(I_i)}{n^2(N-1)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] \\
&= \frac{N^2\mathrm{var}(I_i)}{n^2(N-1)}\sigma^2 \\
&= \frac{N^2\sigma^2}{n^2(N-1)}\frac{n}{N}\frac{N-n}{N} \\
&= \frac{\sigma^2}{n}\frac{N-n}{N-1}
\end{aligned}$$

Povzetek

- V tej nalogi smo ponovili rezultate iz naloge 2.2, vendar z nekoliko drugačnim zapisom vzorčnega povprečja. Medtem ko so bile izpeljave na tem primeru zaradi tega nekoliko zapletenejše, se bo ta pristop izkazal za ključnega v naslednji nalogi, ko bo populacija razdeljena na več skupin. Ta naloga tako služi predvsem kot nekoliko bolj pregleden uvod v naslednjo.

2.6 Večstopenjsko vzorčenje

Oceniti želimo dosežek ljubljanskih sedmošolcev na nekem testu znanja, ki ga izvajajo v več državah. Populacijo $N = 2800$ učencev te starosti bomo vzorčili po šolah ($K = 46$). V vzorec bomo najprej slučajno (in neodvisno od števila N_i sedmošolcev na šoli i) vzorčili $k = 10$ šol, nato pa bomo na vsaki šoli izbrali vzorec $n = 15$ učencev. Naj μ označuje populacijsko povprečje dosežka na testu, μ_i pa naj bo povprečje za vsako šolo posebej. Vzorčenje znotraj šol je neodvisno od vzorčenja na prvem koraku.

a) Zapišite nepristransko cenilko za μ.

Najprej izrazimo μ s povprečji šol, torej μ_i. Naj bo x_{ij} vrednost j-tega učenca na i-ti šoli. Velja

$$\mu = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{K} N_i \cdot \mu_i \tag{2.2}$$

Označimo ocenjeno povprečje vsake šole z $\bar{X}_i$, I_i pa naj bo indikatorska spremenljivka, ki je enaka 1, če je šola izbrana v vzorec. Naša cenilka naj bo enaka

$$\bar{X} = \sum_{i=1}^{K} c_i I_i \bar{X}_i$$

Določiti moramo vrednost konstante c_i, tako da bo cenilka nepristranska. Upoštevamo, da smo na vsaki šoli vzeli naključni vzorec, in zato velja $E(\bar{X}_i) = \mu_i$. Ker je vzorčenje na drugem koraku neodvisno od vzorčenja na prvem, velja $E(I_i\bar{X}_i) = E(I_i)E(\bar{X}_i)$. Ker smo na prvem koraku vzorčili vse šole z enako verjetnostjo, je $E(I_i) = \frac{k}{K}$ za vsak i. Uporabimo vse našteto in dobimo

$$\begin{aligned} E(\bar{X}) &= \sum_{i=1}^{K} c_i E(I_i\bar{X}_i) = \sum_{i=1}^{K} c_i E(I_i)E(\bar{X}_i) \\ &= \sum_{i=1}^{K} c_i \frac{k}{K}\mu_i \end{aligned}$$

Zaradi (2.2) mora veljati $c_i\frac{k}{K} = \frac{N_i}{N}$, zato je naša cenilka enaka

$$\bar{X} = \frac{K}{N}\frac{1}{k}\sum_{i=1}^{K} N_i I_i \bar{X}_i$$

b) Kako bi ocenili populacijsko povprečje, če bi imele vse šole enako število učencev L?

Ker velja $N = \sum_{i=1}^{K} N_i$, za enake $N_i = L$ velja $N = KL$, in zato

$$\bar{X} = \frac{1}{L}\frac{1}{k}\sum_{i=1}^{K} L I_i \bar{X}_i = \frac{1}{k}\sum_{i=1}^{K} I_i \bar{X}_i$$

c) Ali je za nepristranskost pomembno, koliko učencev z vsake šole vzamete?

Ne, $\bar{X}_i$ je nepristranska cenilka μ_i ne glede na velikost vzorca. Seveda pa velikost vzorca vpliva na standardno napako te cenilke.

d) Zapišite varianco cenilke s pomočjo varianc in kovarianc

$$\begin{aligned}\mathrm{var}(\bar{X}) &= \mathrm{var}(\frac{K}{N}\frac{1}{k}\sum_{i=1}^{K} N_i I_i \bar{X}_i) \\ &= \left(\frac{K}{Nk}\right)^2 \sum_{i=1}^{K}\left[N_i^2 \mathrm{var}(I_i\bar{X}_i) + \sum_{j=1, i\neq j}^{K} N_i N_j \mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j)\right]\end{aligned}$$

e) Označimo varianco znotraj vsake šole z $\sigma^2_{wi} = \frac{1}{N_i}\sum_{j=1}^{N_i}(x_{ij}-\mu_i)^2$. Kaj je $\mathrm{var}(I_i\bar{X}_i)$ in kaj $\mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j)$?

Uporabimo, da je vzorčenje na drugem koraku neodvisno od vzorčenja na prvem in da je $I_i^2 = I_i$ ($1^2 = 1,\ 0^2 = 0$):

$$\begin{aligned}\mathrm{var}(I_i\bar{X}_i) &= E(I_i^2\bar{X}_i^2) - E(I_i\bar{X}_i)^2 = E(I_i)E(\bar{X}_i^2) - E(I_i)^2E(\bar{X}_i)^2 \\ &= \frac{k}{K}E(\bar{X}_i^2) - \left(\frac{k}{K}\right)^2\mu_i^2\end{aligned}$$

Upoštevamo še, da je $E(\bar{X}_i^2) = \mathrm{var}(\bar{X}_i) + E(\bar{X}_i)^2 = \frac{\sigma^2_{wi}}{n}\frac{N_i-n}{N_i-1} + \mu_i^2$, in dobimo

$$\begin{aligned}\mathrm{var}(I_i\bar{X}_i) &= \frac{k}{K}\left(\frac{\sigma^2_{wi}}{n}\frac{N_i-n}{N_i-1} + \mu_i^2\right) - \left(\frac{k}{K}\right)^2\mu_i^2 \\ &= \mu_i^2\frac{k(K-k)}{K^2} + \frac{k}{K}\frac{\sigma^2_{wi}}{n}\frac{N_i-n}{N_i-1}\end{aligned}$$

Sedaj izrazimo še kovarianco:

$$\mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j) = E(I_iI_j\bar{X}_i\bar{X}_j) - E(I_i\bar{X}_i)E(I_j\bar{X}_j)$$

Upoštevamo neodvisnost vzorčenja na prvem in drugem koraku in dejstvo, da je povprečje na eni šoli neodvisno od povprečja druge šole:

$$\begin{aligned}\mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j) &= E(I_iI_j)\mu_i\mu_j - E(I_i)E(I_j)\mu_i\mu_j \\ &= \mu_i\mu_j cov(I_i, I_j) = -\mu_i\mu_j\frac{k(K-k)}{K^2(K-1)}\end{aligned}$$

f) Izpeljite formulo za varianco cenilke v primeru, ko so vse vrednosti N_i enake L in je varianca znotraj šole enaka za vse šole, varianco med šolami označite z σ_b^2.

$$\begin{aligned}\mathrm{var}(\bar{X}) &= \left(\frac{1}{Lk}\right)^2\sum_{i=1}^{K}\left[L^2\mathrm{var}(I_i\bar{X}_i) + \sum_{i=1,i\neq j}^{K}L^2\mathrm{cov}(I_i\bar{X}_i, I_j\bar{X}_j)\right] \\ &= \left(\frac{1}{k}\right)^2\sum_{i=1}^{K}\left[\mu_i^2\frac{k(K-k)}{K^2} + \frac{k}{K}\frac{\sigma_w^2}{n}\frac{L-n}{L-1}\right. \\ &\quad \left. - \sum_{j=1,i\neq j}^{K}\mu_i\mu_j\frac{k(K-k)}{K^2(K-1)}\right]\end{aligned}$$

Uporabimo izraz za varianco med šolami

$$\sigma_b^2 = \frac{1}{K}\sum_{i=1}^{K}\left[\mu_i - \mu\right]^2 = \frac{1}{K}\sum_{i=1}^{K}\left[\mu_i^2 - 2\mu\mu_i - \mu^2\right] = \frac{1}{K}\sum_{i=1}^{K}\mu_i^2 - \mu^2$$

in podobno kot pri nalogi 2.5

$$\sum_{i=1}^{K}\sum_{j=1,i\neq j}^{K}\mu_i\mu_j = \mu^2K^2 - \sum_{i=1}^{K}\mu_i^2$$

$$\begin{aligned}
&\sum_{i=1}^{K}\mu_i^2\frac{k(K-k)}{K^2}-\sum_{i=1}^{K}\sum_{j=1,i\neq j}^{K}\mu_i\mu_j\frac{k(K-k)}{K^2(K-1)}\\
&= \frac{k(K-k)}{K^2(K-1)}\left[(K-1)\sum_{i=1}^{K}\mu_i^2-\sum_{i=1}^{K}\sum_{j=1,i\neq j}^{K}\mu_i\mu_j\right]\\
&= \frac{k(K-k)}{K^2(K-1)}\left[(K-1)\sum_{i=1}^{K}\mu_i^2-(\mu^2K^2-\sum_{i=1}^{K}\mu_i^2)\right]\\
&= \frac{k(K-k)}{K^2(K-1)}\left[K\sum_{i=1}^{K}\mu_i^2-\mu^2K^2\right]\\
&= \frac{k(K-k)}{K-1}\left[\frac{1}{K}\sum_{i=1}^{K}\mu_i^2-\mu^2\right]\\
&= \frac{k(K-k)}{K-1}\sigma_b^2
\end{aligned}$$

in zato

$$\begin{aligned}
\mathrm{var}(\bar{X}) &= \frac{1}{k^2}\left[\frac{k(K-k)}{(K-1)}\sigma_b^2+\sum_{i=1}^{K}\left(\frac{k}{K}\frac{\sigma_w^2}{n}\frac{L-n}{L-1}\right)\right]\\
&= \frac{1}{k^2}\frac{k(K-k)}{(K-1)}\sigma_b^2+\frac{K}{k^2}\frac{k}{K}\frac{\sigma_w^2}{n}\frac{L-n}{L-1}\\
&= \frac{\sigma_b^2}{k}\frac{K-k}{(K-1)}+\frac{1}{k}\frac{\sigma_w^2}{n}\frac{L-n}{L-1}
\end{aligned}$$

g) Ali lahko porazdelitev cenilke za povprečje aproksimiramo z normalno porazdelitvijo?

Da bi porazdelitev cenilke konvergirali proti neskončnosti ni dovolj le naraščanje velikosti vzorca - nujen pogoj je, da gre proti neskončnosti tudi število izbranih skupin.

Predlogi za vaje v R

- Preverite rezultate naloge z R.

- S simulacijami si oglejte, kaj se dogaja s porazdelitvijo cenilke povprečja, ko gre velikost vzorca proti neskončnosti.

Povzetek

- Ko je populacija razdeljena na skupine, ki imajo različna povprečja, povprečje enot, zajetih v vzorec, ni več nujno nepristranska cenilka populacijskega povprečja - v ta namen je povprečja posameznih skupin potrebno ustrezno utežiti.

- Možne so različne sheme vzorčenja, v katerih imajo lahko nepristranske cenilke pri enakem številu enot različno standardno napako. Pogosto bo tako cilj poiskati shemo vzorčenja, ki nam pri enaki skupni velikosti vzorca da najbolj natančen rezultat, torej najmanjšo standardno napako.

- Oglejmo si še enkrat izraz za varianco cenilke in ga skušajmo interpretirati. Denimo, da imamo veliko število šol in veliko število učencev na vsaki šoli (popravka za končno populacijo postaneta zanemarljiva), in upoštevajmo, da so vse šole enako velike (enako prispevajo k skupnemu povprečju) in da je varianca znotraj vseh šol enaka. Dobljeni izraz za varianco cenilke poenostavi v

$$\mathrm{var}(\bar{X}) = \frac{\sigma_b^2}{k} + \frac{\sigma_w^2}{kn}$$

  Vidimo, da za varianco cenilke pomembnejšo vlogo igra varianca med povprečji šol, medtem ko del, ki izhaja iz variabilnosti znotraj šol, hitro postane zelo majhen.
  Pri dani velikosti vzorca bo cenilka torej imela najmanjšo varianco takrat, ko bomo vzeli le po enega učenca z vsake šole. V tem primeru je vzorec sestavljen iz neodvisnih slučajnih spremenljivk in tako se noben del informacije ne podvaja zaradi odvisnosti. Je pa iz takega vzorca seveda nemogoče ocenjevati varianco znotraj skupin, pa tudi o povprečjih skupin vemo zelo malo.

Poglavje 3

Ocenjevanje parametrov - metoda največjega verjetja

V prejšnjem poglavju smo intuitivno smiselno cenilko za povprečje ali varianco določili kar neposredno, tako da smo sledili definiciji teh količin v populaciji. V splošnem bo smiselno cenilko pogosto precej težje določiti. Še pomembnejše kot le imeti ustrezno cenilko je tudi poznavanje njene porazdelitve. V prejšnjem poglavju smo za aproksimacijo porazdelitve uporabili centralni limitni izrek - cenilke, ki so nas zanimale, se je namreč dalo izraziti kot vsote neodvisnih enako porazdeljenih slučajnih spremenljivk, kar pa v splošnem seveda ne bo vedno res.
Metoda največjega verjetja (Rice, 2009, razdelek 8.5) je generična metoda, ki pomaga izpeljati smiselno cenilko, katere asimptotska porazdelitev je znana. Z večanjem velikosti vzorca n proti neskončnosti se porazdelitev cenilke po metodi največjega verjetja približuje normalni porazdelitvi. Vrednost te cenilke z večanjem vzorca konvergira k populacijski vrednosti, ki jo ocenjujemo, cenilka po metodi največjega verjetja je torej dosledna (vendar na majhnih vzorcih ne nujno nepristranska). Hkrati je z metodo podan tudi postopek izpeljave variance, s čimer je asimptotska porazdelitev določena.
Ključno vlogo v teoritični podlagi te metode igrata prvi in drugi odvod logaritma verjetja (Rice, 2009, razdelek 8.5.2). Prvi odvod imenujemo funkcija zbira ('score function'), z njegovo pomočjo izpeljemo cenilko, negativno pričakovano vrednost drugega odvoda imenujemo Fisherjeva informacija in je del formule za varianco.
Poglavje pričnemo z iskanjem cenilke za populacijski delež. Na tem primeru pokažemo, da metoda največjega verjetja predlaga enostavno cenilko, do ka-

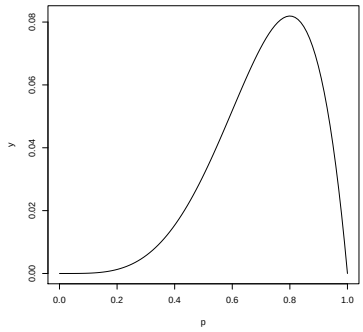
tere bi tudi sicer prišli po intuitivni poti - na ta način na primeru skušamo razumeti smiselnost metode. Druga naloga v tem poglavju obravnava ravno tako zelo pogosto uporabljan problem - po metodi največjega verjetja izpelje cenilki za parametra v linearni regresiji (Rice, 2009, poglavje 14). Ta naloga tako služi za uvod v naslednja poglavja kot tudi za primer, v katerem hkrati ocenjujemo več kot en parameter.

3.1 Ocenjevanje deleža

Naj bodo $x_1, \dots x_n$ neodvisne realizacije Bernoullijevo porazdeljene slučajne spremenljivke X. Radi bi ocenili parameter p.

a) Recimo, da je $n = 5$ in da smo dobili naslednjih 5 vrednosti: 1,0,1,1,1. Kolikšna bi bila verjetnost tega dogodka, če bi bil $p = 0{,}2$? Kaj pa za $p = 0{,}75$? Narišite krivuljo verjetnosti tega dogodka glede na p. Kako bi izračunali njen vrh?

   Verjetnost dogodka izračunamo kot $0{,}2^4 0{,}8^1$, torej $p^k(1-p)^{n-k}$, kjer je k število enic. Označimo z A dogodek $A = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1\}$. Za $p = 0{,}2$ dobimo $P(A) = 0{,}00128$, za $p = 0{,}75$ dobimo $P(A) = 0{,}079$. Narišemo krivuljo za vrednosti p med 0 in 1:



Slika 3.1: *Verjetnost opaženega dogodka glede na p.*

   Vrh funkcije lahko poiščemo z odvajanjem - odvajamo funkcijo $p^k(1-p)^{n-k}$ po p in izenačimo z 0 (lokalni maksimum). Vrh ni odvisen od vrstnih

redov.
V našem primeru je vrh funkcije dosežen pri $p = \frac{4}{5}$.

b) Podatke, ki jih dobimo na nekem vzorcu, označimo z $x_1, \ldots, x_n$ (v zgornjem primeru je bil $n = 5$, $x_1 = 1$ in $x_2 = 0$). Za vsako enoto zapišite $P(X_i = x_i|p)$, torej verjetnost, da se je zgodil dogodek, ki smo ga videli. Zapišite funkcijo verjetja.

$$P(X_i = x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

Funkcija verjetja je produkt posameznih verjetnosti (predpostavili smo, da so slučajne spremenljivke X_i neodvisne), torej

$$\begin{aligned} L(p,x) = P(X_1 = x_1, \ldots, X_n = x_n|p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

c) Poiščite oceno za p po metodi največjega verjetja.

Ker je logaritem monotona funkcija, lahko namesto lokalnega maksimuma te funkcije gledamo raje maksimum logaritma:

$$\begin{aligned} \log L(p,x) &= \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p) \\ \frac{\partial \log L(p,x)}{\partial p} &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\ &= \frac{\sum_{i=1}^n x_i - p\sum_{i=1}^n x_i - p(n - \sum_{i=1}^n x_i)}{p(1-p)} \\ &= \frac{\sum_{i=1}^n x_i - pn}{p(1-p)} \end{aligned}$$

Odvod logaritma verjetja bo enak 0 pri $\hat{p}n = \sum_{i=1}^n x_i$. Ocena po metodi največjega verjetja je torej $\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i$. Ocena je ravno delež enic v vzorcu.

d) Ali je ocena nepristranska?

Metoda največjega verjetja zagotavlja le doslednost (nepristranost, ko gre $n \to \infty$), v našemo primeru dobimo

$$E(\hat{p}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} p = p$$

V našem primeru je torej ocena nepristranska.

e) Zapišite oceno standardne napake

Varianca ocene je enaka $\frac{1}{n}I(p)^{-1}$, kjer $I(p)$ označuje Fisherjevo informacijo:

$$I(p) = -E\left[\frac{\partial^2}{\partial p^2}\log(f(X,p))\right] = E\left[\frac{\partial}{\partial p}\log(f(X,p))\right]^2$$

V našem primeru sta izračuna po obeh formulah enako težka, uporabimo prvo formulo:

$$\begin{aligned}
f(X|p) &= p^X(1-p)^{1-X} \\
I(p) &= -E\left[\frac{\partial^2}{\partial p^2}\log(f(X|p))\right] \\
&= -E\left[\frac{\partial^2}{\partial p^2}\left(X\log p + (1-X)\log(1-p)\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{X}{p} - \frac{1-X}{1-p}\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{(1-p)X - (1-X)p}{p(1-p)}\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{X-p}{p(1-p)}\right)\right] \\
&= -E\left[\frac{p(1-p)(-1) - (1-2p)(X-p)}{p^2(1-p)^2}\right] \\
&= -E\left[\frac{-p+p^2-X+2pX+p-2p^2}{p^2(1-p)^2}\right] \\
&= -E\left[\frac{-p^2-X+2pX}{p^2(1-p)^2}\right]
\end{aligned}$$

Pri računanju pričakovane vrednosti upoštevamo, da je $E(X) = p$, ker je X le v števcu, dobimo

$$\begin{aligned} I(p) &= -E\left[\frac{-p^2 - X + 2pX}{p^2(1-p)^2}\right] \\ &= -\left[\frac{-p + p^2}{p^2(1-p)^2}\right] \\ &= \frac{1}{p(1-p)} \end{aligned}$$

f) Oceniti želimo delež volilcev nekega kandidata. Na vzorcu $n = 500$ zanj glasuje 29 % volilcev. Podajte 95% interval zaupanja za to oceno.

Vzorčna ocena je $\hat{p} = 0{,}29$. Standardno napako (torej standardni odklon cenilke) na vzorcu ocenimo s pomočjo $\hat{p}$, ocena standardne napake je torej enaka

$$\widehat{SE} = \sqrt{\frac{1}{nI(\hat{p})}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}02$$

Teorija nam pove, da je lahko porazdelitev kvocienta $\frac{p-\hat{p}}{\widehat{SE}}$ aproksimiramo z normalno porazdelitvijo, 95% interval zaupanja je enak $[0{,}25, 0{,}33]$.

Predlogi za vaje v R

- Z R narišite sliko 3.1:

```
> p <- seq(0,1,length=100)  #za 100 vrednosi p med 0 in 1
> y <- p^4*(1-p)            #za vsako vrednost izracunam verjetnost
> plot(p,y,type="l")        #narisem in povezem s krivuljo
```

- Generirajte vzorec velikosti 500, v katerem ima vsak posameznik verjetnost 0,3, da glasuje za nekega kandidata. Ocenite verjetnost z deležem na vzorcu. Ponovite poskus 1000x in si oglejte porazdelitev vzorčnih ocen.

- Na vsakem vzorcu ocenjenemu deležu dodajte še 95% interval zaupanja. Kolikšen je delež vzorcev, pri katerih interval zaupanja zajema pravo vrednost (0,3)? Temu deležu pravimo pokritje (angl. 'coverage'). Pozor - ta delež ni nujno enak 95%, saj nam teorija da le približek porazdelitvi, na manjših vzorcih bo zato ta odstotek lahko precej odstopal.

Povzetek

- V nalogi smo za preprost primer narisali funkcijo verjetja in intuitivno utemeljili, zakaj je točka, v kateri ta funkcija doseže vrh, smiselna ocena parametra.
- Ocena deleža v populaciji po metodi največjega verjetja je kar delež na vzorcu. Pokazali smo, da je ta ocena nepristranska.
- Varianca ocene deleža je enaka $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Odvisna je od velikosti vzorca in od dejanske vrednosti deleža, ki ga ocenjujemo - napaka bo največja pri ocenjevanju deležev blizu polovice.

3.2 Povezanost dveh spremenljivk

Zanima nas, kako je prihodek podjetja v neki panogi odvisen od števila zaposlenih. Predpostavimo, da je prihodek podjetja normalno porazdeljen s povprečjem $\beta_0 + \beta_1 X$, kjer je X logaritem števila zaposlenih. Denimo, da imamo podatke o številu zaposlenih in prihodku za vzorec podjetij, radi bi ocenili parametra β_0 in β_1.

a) Zapišite gostoto porazdelitve prihodka podjetja, če vemo, da je varianca enaka σ^2.

Predpostavljamo, da je $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, torej

$$f(Y, X|\beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y-\beta_0-\beta_1 X)^2}{2\sigma^2}}$$

b) Zapišite funkcijo verjetja. Kaj je funkcija, ki jo moramo maksimizirati?

Dani so podatki (x_i, y_i), $i = 1, \ldots, n$.

$$\begin{aligned} L(y, x, \beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \end{aligned}$$

Logaritem te funkcije je

$$\log L(y, x, \beta_0, \beta_1, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

Ker nas zanimata le parametra β_0 in β_1, je prvi del funkcije konstanta, maksimizirati je potrebno le izraz

$$-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Izraz $y_i - (\beta_0 + \beta_1 x_i)$ predstavlja razdaljo med točkama $T(x_i, y_i)$ in $T(x_i, \beta_0 + \beta_1 x_i)$, to vrednost imenujemo ostanek (razdalja točke od premice). Oceni za parametra β_0 in β_1 določata premico, ki se najbolje prilega podatkom v smislu, da je vsota kvadriranih ostankov točk od premice najmanjša možna. To oceno zato imenujemo ocena po metodi najmanjših kvadratov (Rice, 2009, razdelek 14.1).

c) Izračunajte oceni β_0 in β_1 po metodi največjega verjetja.

Najprej za β_0:

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\ = \quad -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo (izraz je enak nič za posebni vrednosti β_0 in β_1, ki ju označimo s strešico)

$$\begin{aligned}-2\left(\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1\sum_{i=1}^{n} x_i\right) &= 0 \\ \widehat{\beta}_0 &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \widehat{\beta}_1\sum_{i=1}^{n} x_i\right)\end{aligned}$$

Sedaj odvajamo še po β_1:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\\
&= -2\sum_{i=1}^{n}x_i\left(y_i-\beta_0-\beta_1x_i\right)\\
&= -2\left(\sum_{i=1}^{n}x_iy_i-\beta_0\sum_{i=1}^{n}x_i-\beta_1\sum_{i=1}^{n}x_i^2\right)
\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_0\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}$$

Združimo obe izpeljavi in (po malce premetavanja členov) dobimo

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

d) Izračunajte standardno napako za obe oceni.

Za Fisherjevo matriko informacije moramo izračunati druge odvode. Logaritem funkcije verjetja je enak

$$\log f(Y,X|\beta_0,\beta_1,\sigma) = -\frac{1}{2}\log(2\pi\sigma^2)-\frac{(Y-\beta_0-\beta_1X)^2}{2\sigma^2}$$

Prva odvoda sta enaka

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\log f(Y,X|\beta_0,\beta_1,\sigma) &= \frac{1}{\sigma^2}(Y-\beta_0-\beta_1X)\\ \frac{\partial}{\partial\beta_1}\log f(Y,X|\beta_0,\beta_1,\sigma) &= \frac{X}{\sigma^2}(Y-\beta_0-\beta_1X)\end{aligned}$$

Drugi odvodi so potem

$$\begin{aligned}\frac{\partial^2}{\partial\beta_0^2}\log f(Y,X|\beta_0,\beta_1,\sigma) &= -\frac{1}{\sigma^2}\\ \frac{\partial^2}{\partial\beta_1^2}\log f(Y,X|\beta_0,\beta_1,\sigma) &= -\frac{X^2}{\sigma^2}\\ \frac{\partial^2}{\partial\beta_1\beta_0}\log f(Y,X|\beta_0,\beta_1,\sigma) &= -\frac{X}{\sigma^2}\end{aligned}$$

Členi Fisherjeve matrike informacije so negativne pričakovane vrednosti drugih odvodov. Ker pričakovane vrednosti X oziroma X^2 ne poznamo, ju ocenimo iz podatkov:

$$I(\beta_0,\beta_1)=\frac{1}{\sigma^2}\begin{bmatrix}1 & \bar{x}\\ \bar{x} & \frac{1}{n}\sum_{i=1}^{n}x_i^2\end{bmatrix}$$

Inverz te matrike je potem

$$I^{-1}(\beta_0,\beta_1)=\frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}$$

in zato

$$\begin{aligned}\mathrm{var}(\widehat{\beta}_0) &= \frac{I_{11}^{-1}}{n}=\frac{1}{n}\frac{\sigma^2\frac{1}{n}\sum_{i=1}^{n}x_i^2}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}\\ &= \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\end{aligned}$$

ter

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_1) &= \frac{I_{22}^{-1}}{n} = \frac{1}{n}\frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{aligned}$$

Povzetek

- Če je pogojna porazdelitev spremenljivke Y glede na X normalna, je ocena parametrov v linearni regresiji po metodi največjega verjetja enaka oceni po metodi najmanjših kvadratov.
- Metoda najmanjših kvadratov je intuitiven pristop, ki zagotavlja smiselno premico. Da bi lahko uporabili to metodo, moramo definirati model in s tem ostanke, ni pa nam potrebno narediti nobenih predpostavk o porazdelitvi ostankov. Metoda nam da cenilke koeficientov, ne izvemo pa ničesar o njihovi standardni napaki oziroma porazdelitvi. V posameznih primerih bo navkljub temu porazdelitev dokaj preprosto najti, saj so cenilke koeficientov linearne kombinacije spremenljivke Y. Nasprotno metoda največjega verjetja zahteva predpostavke o porazdelitvi ostankov ter nato tudi poda asimptotsko porazdelitev teh ocen glede na pravo vrednost parametrov v populaciji.

Poglavje 4

Preizkušanje domnev

Statistično sklepanje, opisano v prejšnjih poglavjih je bilo osredotočeno, na ocenjevanje vrednosti v populaciji s pomočjo vzorca. V tem razdelku bomo s pomočjo vzorca preverjali resničnost trditev o populaciji.
Frekventistična statistika pri tovrstnem preizkušanju domnev temelji na Neyman-Pearsonovi paradigmi (Rice, 2009, razdelek 9.2), ki predstavlja način odločanja med domnevami. Pri tem osrednjo vlogo igra ena izmed domnev, imenujemo jo ničelna domneva, ki naj bi predstavljala naše dosedanje znanje. Od nje smo se pripravljeni oddaljiti le, če je verjetnost, da vzorec izhaja iz populacije, v kateri ta domneva drži, zelo majhna. Da bi o tem lahko presojali, definiramo smiselno testno statistiko, ki predstavlja nek povzetek informacije o ničelni domnevi, ki nam jo ponuja vzorec. Če poznamo porazdelitev te testne statistike pod ničelno domnevo, lahko potem presojamo o 'nenavadnosti' njene vrednosti na našem vzorcu.

Porazdelitev testne statistike pod ničelno domnevo nam bo omogočila postavitev mej, v katerih se z veliko verjetnostjo nahaja vrednost testne statistike za vzorce, ki izhajajo iz populacije, v kateri ničelna domneva velja. Če bodo meje na vzorcu presežene, bomo govorili o statistično značilnem rezultatu in ničelno domnevo zavrnili v korist alternativne. Pri tovrstnem odločanju lahko zagrešimo dve vrsti napak. Napaka prve vrste je verjetnost, da zavrnemo ničelno domnevo, četudi v populaciji drži. Velikost te napake določimo sami, glede na to kako veliko verjetnost te napake si bomo dovolili, bomo postavili meje oziroma določili območje zavrnitve. Največjo dovoljeno verjetnost napake bomo imenovali stopnja značilnosti in jo označili z α. Napaka druge vrste (označili jo bomo z β) je verjetnost, da ničelne domneve na

podlagi vzorca ne bomo zavrnili, četudi bi jo morali. O velikosti te napake lahko sklepamo, če poznamo porazdelitev testne statistike pod alternativno domnevo. Na primerih si bomo ogledali, kaj vse vpliva na njeno velikost. Pogosto bomo namesto o napaki druge vrste govorili o nasprotni verjetnosti, torej verjetnosti, da uspemo zavrniti ničelno domnevo, kadar vzorec dejansko ne izhaja iz populacije, v kateri bi ničelna domneva veljala. To verjetnost bomo imenovali moč testa.

Poglavje pričnemo s preprostim primeom s pomočjo katerega uvedemo osnovne pojme in ideje statističnega preizkušanja domnev. Sledi konkreten primer iz prakse, ki nam pokaže, kako pomemben je lahko izračun moči pred začetkom raziskave. Nato se vrnemo na nekoliko splošnejši primer, pri katerem se na podlagi vzorca želimo odločati med dvema domnevama o populaciji, a imamo vsaj za začetek domnevi za enakovredni. Odločanje poteka tako, da na vzorcu definiramo smiselno testno statistiko, katere vrednosti bodo 'dokazi' v prid eni ali drugi domnevi. Odločali se bomo glede na verjetnost, torej je nujno poznati porazdelitev te testne statistike, ki pa bo pod različnima domnevama seveda različna. Nalogo bomo nato še posplošili in si pogledali primer uporabe centralnega limitnega izreka pri iskanju porazdelitve testne statistike.
Poglavje bomo zaključili s posplošenim testom razmerja verjetij (Rice, 2009, razdelek 9.4). Podobno kot je metoda največjega verjetja pomenila generičen pristop k iskanju smiselne cenilke, ki hkrati podaja tudi njeno asimptotsko porazdelitev, tudi posplošeni test razmerja verjetij pomeni generičen pristop, ki nam predlaga smiselno testno statistiko in poda njeno asimptotsko porazdelitev.

4.1 Osnovni pojmi pri statističnem preizkušanju domnev

Želimo preveriti, ali je kovanec pošten. Naredili smo poizkus, kjer smo 10-krat vrgli kovanec in pri tem 7-krat dobili grb.

a) Zapišite ničelno domnevo za vaš primer. Ali je ničelna domneva enostavna ali sestavljena? Zapišite testno statistiko, označite jo z X - kakšna je njena porazdelitev pod ničelno domnevo?

$H_0 : p = 0{,}5$. S tem je porazdelitev slučajne spremenljivke pod ničelno domnevo natanko določena, zato pravimo, da je ničelna domneva enostavna. Testna statistika X je število grbov: porazdelitev testne statistike pod ničelno domnevo je binomska $B(10, 0{,}5)$.

b) Denimo, da je vaša alternativna domneva $H_A : p > 0{,}5$. Ali je ta domneva enostavna ali sestavljena? Pri kakšnih vrednostih X boste zavrnili ničelno domnevo v prid alternativni? Ali je domneva enostranska ali dvostranska?

Domneva je sestavljena, saj zajema več vrednosti istega parametra - tudi če vemo, da domneva drži, še ne poznamo porazdelitve. Zavračali bomo pri velikih vrednostih X. Domneva je enostranska, saj nas bo zanimal le desni rep porazdelitve.

c) V našem primeru je $X = 7$. Kolikšna je verjetnost, da se na vzorcu zgodi ta dogodek, če ničelna domneva drži?

Če ničelna domneva drži ($p = 0{,}5$), je $P_0(X = 7) = 0{,}117$.

d) Denimo, da je območje zavrnitve sestavljeno iz vrednosti $\{10\}$. Kolikšna je stopnja značilnosti α v tem primeru? Kolikšna je stopnja značilnosti, če je območje zavrnitve sestavljeno iz vrednosti $\{6{,}7{,}8{,}9{,}10\}$?

Če ničelna domneva drži ($p = 0{,}5$), je $P_0(X = 10) = 0{,}001$.
Če ničelna domneva drži ($p = 0{,}5$), je $P_0(X \geq 6) = 0{,}377$.

e) Določite območje zavrnitve pri stopnji značilnosti $\alpha = 0{,}05$. Ali lahko na podlagi dobljenih podatkov zavrnete ničelno domnevo pri stopnji značilnosti $\alpha = 0{,}05$?

Največja vrednost k, tako da velja $P_0(X \geq k) < 0{,}05$, je enaka 9. Stopnja značilnosti je v tem primeru enaka 0,01. Ničelne domneve seveda ne moremo zavrniti.

f) Kolikšna je moč testa pri tej vrednosti α, če predpostavimo, da je prava vrednost parametra $p = 0{,}6$? Kaj pa pri $p = 0{,}7$? Kolikšna je v teh primerih napaka druge vrste?

Moč testa je majhna - 0,046 oziroma 0,149. Pri tako majhnem vzorcu

in majhni stopnji značilnosti bomo ničelno domnevo le redko uspešno zavrnili. Napako druge vrste izračunamo kot 1 - moč.

g) Predpostavite, da je vaša alternativna domneva enaka $H_A : p \neq 0{,}5$. Ali je ta domneva enostavna ali sestavljena? Ali je domneva enostranska ali dvostranska?

Alternativna domneva je še vedno sestavljena, zdaj je tudi dvostranska.

h) Kakšno bo sedaj območje zavrnitve, če želite, da je $\alpha \leq 0{,}05$? Kolikšna natanko bo stopnja značilnosti za to območje?

Območje zavrnitve bo sestavljeno iz vrednosti $\{0{,}1{,}9{,}10\}$. Stopnja značilnosti za to območje je enaka $\alpha = 0{,}02$.

i) Izračunajte še moč testa v tem primeru.

Pri moči testa moramo upoštevati, da bomo sedaj ničelno domnevo zavrnili tudi če bo X enak 0 ali 1. Ker pa je verjetnost teh dveh vrednosti za $p = 0{,}6$ oz. $p = 0{,}7$ zelo majhna, se moč praktično ne spremeni.

Tabela 4.1: Kumulativne verjetnosti $P(X \leq k|p)$ za binomsko porazdelitev pri $n = 10$.

p\k	0	1	2	3	4	5	6	7	8	9	10
0,5	0,001	0,011	0,055	0,172	0,377	0,623	0,828	0,945	0,989	0,999	1
0,6	0,000	0,002	0,012	0,055	0,166	0,367	0,618	0,833	0,954	0,994	1
0,7	0,000	0,000	0,002	0,011	0,047	0,150	0,350	0,617	0,851	0,972	1

Predlogi za vaje v R

- 1000x ponovite poskus, v katerem po 10x mečete kovanec. Oglejte si verjetnost zavrnitve za posamezno območje.
- Spremenite verjetnost, s katero pade grb, in si oglejte moč testa.
- Povečajte vzorec in si oglejte, kako se spreminja moč testa.

Povzetek

- V nalogi smo se z enostavnim primerom sprehodili čez osnovne pojme preizkušanja domnev, kot so: ničelna domneva, testna statistika, enostavna oziroma sestavljena domneva, stopnja značilnosti, zavrnitveno območje, enostranska oziroma dvostranska alternativna domneva, moč testa.

- Neyman-Pearsonova lema (Rice, 2009, stran 332) pove, da ima razmerje verjetij pri dani stopnji značilnosti α največjo moč izmed vseh testnih statistik. Pogoj za Neyman-Pearsonovo lemo je, da sta ničelna in alternativna domneva enostavni, torej natanko določata porazdelitev.

4.2 Moč testa

Iz literature lahko povzamemo, da se športnikovo povprečje hemoglobina ob vsaj 14-dnevnem bivanju na višini nad 1500 m zviša za 2 g/l, medtem ko višinski treningi ne vplivajo na varianco njegovih vrednosti. Ob običajnih treningih se posameznikove vrednosti porazdeljujejo normalno, $X \sim N(\mu_1, 5^2)$, kjer je μ_1 športnikovo povprečje.
Športnik pogosto opravlja višinske treninge, vendar v krajših intervalih. Zanima ga, ali se njegovo povprečje hemoglobina v obdobju višinskih treningov kljub temu zviša. V sezoni opravi 12 meritev, 8 med obdobjem višinskih priprav in 4 sicer. Cilj naloge je ugotoviti, kolikšna bo moč njegovega testa, če bo pri sklepanju uporabil stopnjo značilnosti $\alpha = 0{,}05$?

a) Kaj je športnikova ničelna in kaj alternativna domneva?

   Zapišimo porazdelitev hemoglobina v obdobju višinskih priprav z $N(\mu_2,5^2)$.
   Ničelna domneva je:
   H_0: povprečje hemoglobina v obeh obdobjih je enako, $\mu_1 = \mu_2$.
   Alternativna domneva je, da je $\mu_2 > \mu_1$, zanima ga torej le enostranski test.

b) Predlagajte testno statistiko. Izračunajte njeno porazdelitev pod ničelno domnevo.

Športnik bo izračunal razliko povprečij na vzorcih, ki je porazdeljena kot

$$R = \bar{X}_2 - \bar{X}_1 \sim N\left(\mu_2 - \mu_1, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

Testna statistika

$$Z = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

je torej pod ničelno domnevo porazdeljena standardno normalno. Ker ga zanima le enostranska alternativna domneva, bo ničelno domnevo zavrnil, kadar bo $Z > z_\alpha$, torej $Z > 1{,}64$.

c) Izračunajte moč testa, torej verjetnost, da bo ničelno domnevo uspel zavrniti, če se mu povprečje hemoglobina v obdobju višinskih priprav zares poveča za 2 g/l?

Zanima nas $P_A(Z > 1{,}64)$. Označimo $SE = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, pod alternativno domnevo je pričakovana vrednost Z enaka

$$E_A(Z) = E\left(\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}\right) = \frac{\mu_2 - \mu_1}{SE}$$

torej velja $Z \sim N(\frac{2}{SE},1)$, in zato

$$P_A(Z > 1{,}64) = P(U > 1{,}64 - \frac{2}{SE})$$

kjer je U standardna normalna spremenljivka. V našem primeru

$$P\left(U > 1{,}64 - \frac{2}{5\sqrt{\frac{1}{8} + \frac{1}{4}}}\right) = P(U > 0{,}99) = 0{,}16 \tag{4.1}$$

Moč testa je zelo majhna - pri tako majhnem številu meritev je le majhna verjetnost, da bo športnik zavrnil ničelno domnevo (četudi se mu povprečje dejansko zares zviša za 2 g/l).

d) Kako bi se moč testa spremenila, če bi imel na voljo enako število meritev v vsakem obdobju?

Če bi imel po 6 meritev v vsakem obdobju, bi bila moč enaka

$$P\left(U > 1{,}64 - \frac{2}{5\sqrt{\frac{1}{6}+\frac{1}{6}}}\right) = P(U > 0{,}95) = 0{,}17$$

e) Kako je moč testa odvisna od variance posameznikovih meritev in kako od dejanske velikosti razlike v populaciji?

Iz enačbe (4.1) je očitno, da večja razlika pomeni večjo moč - če je dejanska razlika med obdobjema večja, jo bomo lažje opazili na podatkih.
Če bi bila varianca posameznikovih meritev manjša, bi imeli manjšo standardno napako, in zato večjo moč testa.

Povzetek

- V nalogi smo si ogledali izračun moči testa pri primerjavi povprečij dveh skupin. Cilj opisane raziskave je dokazati, da obstaja razlika med skupinama, a smo z izračunom moči ugotovili, da je izvedba opisane raziskave precej nesmiselna, saj ima športnik le zelo majhno verjetnost, da bo razliko tudi dejansko dokazal. Za zanesljive trditve o razlikah bi potreboval precej večji vzorec.

- Izračun moči testa je pomemben korak pri načrtovanju raziskave in ključ do izbire primerno velikega vzorca. Ponavadi pravimo, da raziskave ni smiselno izvesti, če moč testa ni vsaj 0,8, torej če raziskovalec nima vsaj verjetnosti 0,8, da bo ničelno domnevo uspel zavrniti. Za izračun moči moramo vedno nekaj vedeti o razpršenosti podatkov, hkrati pa moramo vedeti tudi, kako veliko razliko bi radi dokazali. To vprašanje ni vprašanje za statistika, temveč za raziskovalca - že pri načrtovanju raziskave mora vedeti, kako velike razlike bi rad dokazal oziroma kolikšna je najmanjša razlika, ki bo zanj tudi strokovno pomembna.

- Če smo na podlagi vzorca s tveganjem, manjšim od α uspeli dokazati, da obstaja razlika med skupinama v populaciji, pravimo, da je razlika

statistično značilna. Vendar pa statistična značilnost še ne pomeni, da je razlika tudi dejansko strokovno pomembna oziroma da je smiselno razmišljati o razlogih zanjo. V populaciji bosta dve skupini le redko imeli natanko enako povprečje, zato bomo z zelo velikimi vzorci lahko skoraj vedno dobili statistično značilne rezultate. Kot primer si zamislimo raziskavo, s katero želimo po višini primerjati prebivalce dveh mest. Povprečji obeh populacij zagotovo nista natanko enaki, lahko se na primer razlikujeta za 2 mm, kar pa je povsem nepomembno. Če bomo imeli zares velik vzorec, bomo na vzorcih torej lahko dobili statistično značilno razliko, ki pa bo v praksi povsem nepomembna.

- S statističnim testom nikoli ne bomo mogli pokazati, da v populaciji ni razlik med skupinama. A če navkljub veliki moči testa ne bomo uspeli zavrniti ničelne domneve, bo dobljeni interval zaupanja za razliko verjetno dovolj ozek, da bomo lahko morebitno razliko v populaciji omejili na vrednosti, ki niso strokovno pomembne.

- Testna statistika, ki smo jo uporabili v nalogi, je zelo podobna statistiki, ki jo uporablja test t. Razlika je le, da v naši nalogi naredimo precej predpostavk glede porazdelitve populacije (vemo, da je normalna, poznamo varianco), in zato lahko preprosto izračunamo tudi porazdelitev testne statistike. Kadar teh predpostavk ne bomo naredili, si bomo do približne porazdelitve pomagali s centralnim limitnim izrekom, upoštevati pa bomo morali tudi, da ne poznamo variance, temveč jo ocenjujemo s pomočjo vzorca.

- V nalogi smo si ogledali še, kako na moč vpliva velikost skupin. Izkaže se, da je v smislu moči optimalen načrt raziskave tak, da sta skupini enako veliki.

4.3 Enostavni domnevi

Računalnik skuša razlikovati med dvema viroma signalov. Prvi vir oddaja signale, katerih jakost je normalno porazdeljena z $N(0,1)$, drugi vir ima enako porazdelitev, a višjo povprečno jakost - $N(2,1)$. Računalnik prejme 10 signalov in se mora odločiti, iz katerega vira so prišli. Posamezni signali iz istega vira so med seboj neodvisni.

a) Računalnik se odloča med domnevama

H_1: signal prihaja iz vira 1 in H_2: signal prihaja iz vira 2

Zapišite testno statistiko, ob pomoči katere naj se odloča računalnik. (Izrazite malo bolj splošno - naj imata porazdelitvi obeh virov varianco σ^2, povprečna jakost drugega vira naj bo a, $a > 0$, velikost vzorca naj bo n.) *Namig:* pri tem, kaj je bolj verjetno, si pomagajte z gostotami.

Kot testno statistiko uporabimo razmerje verjetij - bolj ko bo razmerje različno od 1, bolj bomo prepričani v eno izmed domnev:

$$\begin{aligned}
\prod_{i=1}^{n} \frac{f_2(x_i)}{f_1(x_i)} &= \prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x_i-a)^2}{2\sigma^2}\}}{\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x_i)^2}{2\sigma^2}\}} \\
&= \prod_{i=1}^{n} \frac{\exp\{-\frac{(x_i-a)^2}{2\sigma^2}\}}{\exp\{-\frac{(x_i)^2}{2\sigma^2}\}} \\
&= \exp\{-\sum_{i=1}^{n} \frac{(x_i-a)^2 - (x_i)^2}{2\sigma^2}\} \\
&= \exp\{-\sum_{i=1}^{n} \frac{-2ax_i + a^2}{2\sigma^2}\} \\
&= \exp\{\sum_{i=1}^{n} \frac{2ax_i - a^2}{2\sigma^2}\} \\
&= \exp\{\frac{na}{\sigma^2}\left(\bar{x} - \frac{a}{2}\right)\}
\end{aligned}$$

b) Denimo, da želimo, da računalnik reagira le, če je precej prepričan, da signal ne prihaja iz vira 1. Domnevo H_1 proglasimo za ničelno domnevo, domnevo H_2 pa za alternativno. Odločitveno pravilo postavimo tako, da bo verjetnost zmote, kadar je ničelna domneva resnična, največ $\alpha = 0{,}05$.

- Testna statistika je slučajna spremenljivka (označite jo z Y). Kaj lahko rečemo o njeni porazdelitvi pod ničelno domnevo?
  Namig: da bo porazdelitev preprostejša, uporabite logaritem.

  Označimo $Y = \frac{na}{\sigma^2}\left(\bar{X} - \frac{a}{2}\right)$. Večja kot bo vrednost Y, bolj bomo prepričani v domnevo H_2. Da bi vedeli, katere vrednosti so 'velike', moramo poznati porazdelitev slučajne spremenljivke Y.

Pod ničelno domnevo so vrednosti X_i standardno normalno porazdeljene. Ker so a in σ^2 konstante (znane vrednosti), je Y linearna kombinacija neodvisnih normalnih spremenljivk, in zato normalno porazdeljena. Poiščemo povprečje in standardni odklon spremenljivke Y:

$$\begin{aligned} E_0(Y) &= E_0\left(\frac{na}{\sigma^2}\bar{X} - \frac{na^2}{2\sigma^2}\right) \\ &= \frac{na}{\sigma^2}E_0(\bar{X}) - \frac{na^2}{2\sigma^2} = -\frac{na^2}{2\sigma^2} \\ \mathrm{var}_0(Y) &= \mathrm{var}_0\left(\frac{na}{\sigma^2}\bar{X} - \frac{na^2}{2\sigma^2}\right) = \\ &= \frac{n^2a^2}{\sigma^4}\mathrm{var}_0\bar{X} = \frac{n^2a^2\sigma^2}{\sigma^4 n} = n\frac{a^2}{\sigma^2} \\ sd_0(Y) &= \frac{a\sqrt{n}}{\sigma} \end{aligned}$$

- Izrazite mejno vrednost, pri kateri naj računalnik reagira.

  Ničelno domnevo bomo zavrnili pri velikih vrednostih Y, zanima nas torej vrednost c, tako da bo $P_0(Y \geq c) = 0{,}05$, pri čemer P_0 označuje verjetnost pri predpostavki, da ničelna domneva drži.
  Y je normalno porazdeljena slučajna spremenljivka, zato velja $\alpha = P_0(\frac{Y-E(Y)}{sd(Y)} \geq z_\alpha) = \alpha$ oz. $P_0(Y \geq E(Y) + z_\alpha sd(Y))$. Za primer, ko je $\alpha = 0{,}05$, je mejna vrednost $c = -\frac{na^2}{2\sigma^2} + 1{,}64\frac{a\sqrt{n}}{\sigma}$. Če je $n = 10$, $a = 2$ in $\sigma = 1$, je mejna vrednost za $c = -\frac{10\cdot 4}{2} + 1{,}64 \cdot 2\sqrt{10} = -9{,}63$.

c) Kolikšna je verjetnost, da bo računalnik reagiral, če signal v resnici prihaja iz drugega vira? (Tej verjetnosti pravimo moč testa.)
Namig: izračunajte porazdelitev testne statistike pod alternativno domnevo.

Pod alternativno domnevo se Y prav tako porazdeljuje normalno (linearna kombinacija normalnih), povprečje je enako

$$\begin{aligned} E_A(Y) &= E_A\left(\frac{na}{\sigma^2}\bar{X} - \frac{na^2}{2\sigma^2}\right) \\ &= \frac{na}{\sigma^2}a - \frac{na^2}{2\sigma^2} = \frac{na^2}{\sigma^2 a} \end{aligned}$$

Ker je varianca pri obeh hipotezah enaka, je enaka tudi varianca vzorčnega povprečja, in zato je tudi varianca pod alternativno domnevo enaka

$$\mathrm{var}_A(Y) = n\frac{a^2}{\sigma^2}$$

Naj bo Z standardno normalno porazdeljena spremenljivka. Moč testa je

$$\begin{aligned} P_A(Y > c) &= P_A(Y > -\frac{na^2}{2\sigma^2} + z_\alpha \frac{a\sqrt{n}}{\sigma}) \\ &= P_A(\frac{Y - \frac{na^2}{2\sigma^2}}{\frac{a\sqrt{n}}{\sigma}} > \frac{-\frac{na^2}{2\sigma^2} + z_\alpha \frac{a\sqrt{n}}{\sigma} - \frac{na^2}{2\sigma^2}}{\frac{a\sqrt{n}}{\sigma}}) \\ &= P(Z > \frac{z_\alpha\sqrt{n} - \frac{na}{\sigma}}{\sqrt{n}}) \\ &= P(Z > z_\alpha - \frac{\sqrt{n}a}{\sigma}) \end{aligned}$$

Vidimo, da je moč testa odvisna od velikosti vzorca (večji vzorec, večja moč), variance (če podatki bolj variirajo, imamo manjšo moč) ter seveda od povprečja a pod alternativno domnevo. V našem primeru je povprečje pod alternativno domnevo kar za dva standardna odklona proč od povprečja pod ničelno domnevo, zato je moč zelo velika navkljub majhnemu vzorcu:

$$P(Z > z_\alpha - \frac{\sqrt{n}a}{\sigma}) = P(Z > 1{,}64 - 2\sqrt{10}) = P(Z > -4{,}68)$$

Moč testa je skoraj enaka 1.

Vidimo tudi, da je spodnja meja moči (a je večji od 0) enaka α, kar je verjetnost, da zavrnemo ničelno domnevo, kadar je $a = 0$.

d) Testno statistiko Y transformirajte tako, da bo pod ničelno domnevo standarna normalna spremenljivka.

Standardizirajmo Y:

$$\begin{aligned} Z &= \frac{Y + \frac{na^2}{2\sigma^2}}{\frac{a\sqrt{n}}{\sigma}} \\ &= \frac{\frac{na}{\sigma^2}\bar{X} - \frac{na^2}{2\sigma^2} + \frac{na^2}{2\sigma^2}}{\frac{a\sqrt{n}}{\sigma}} \\ &= \frac{\bar{X}}{\sigma/\sqrt{n}} \end{aligned}$$

e) Povzemite: ali je mejna vrednost testne statistike odvisna od a? Intuitivno razložite. Je vrednost a torej sploh pomembna?

V prejšnji točki smo testno statistiko standardizirali, ker gre za bijektivno preslikavo (linearno transformacijo), je Z testna statistika, ki je ekvivalentna Y (oz. prvotno definirani testni statistiki). S tem smo pokazali, da mejna vrednost testne statistike ni odvisna od vrednosti a. To je intuitivno smiselno - mejna vrednost pod ničelno domnevo je postavljena glede na porazdelitev pod ničelno domnevo - poznavanje alternativne domneve za določitev mejne vrednosti ni potrebno (je pa pomembno, da vemo, da je enostranska). So pa vrednosti pod alternativno domnevo seveda ključne za moč testa.

Predlogi za vaje v R

- Generirajte podatke v dveh korakih. Najprej z enako verjetnostjo izberite vrednost a (0 ali 2), nato generirajte 10 vrednosti iz porazdelitve $N(a,1)$. Izračunajte vrednost testne statistike iz prve točke in se odločite za eno izmed domnev glede na to, ali je vrednost testne statistike večja ali manjša od 1. Postopek velikokrat ponovite in izračunajte delež poskusov, v katerih se odločite za vsako od domnev, ter delež poskusov, ko je ta odločitev pravilna.

- Zamenjajte verjetnost v prvem koraku (naj bo npr. bolj verjetna izbira $a = 2$). Kako se spreminjajo deleži iz prejšnje točke?

- Generirajte podatke, tako da je a ves čas enak 0, in preverite, da je vrednost α pri izračunani mejni vrednosti c zares enaka 0,05.

- Generirajte podatke še tako, da je $a = 2$, in preverite moč testa.

Povzetek

- V nalogi smo se želeli odločati med dvema domnevama o populaciji, v ta namen smo najprej definirali testno statistiko, ki povzame ustrezno informacijo iz vzorca in katere vrednosti kažejo na eno ali drugo domnevo. Glede na to katera od domnev drži v populaciji, obstaja določena verjetnost, da testna statistika zavzame vrednost, ki smo jo opazili. Kriterij za odločanje med domnevama temelji na teh verjetnostih. V tej nalogi eno izmed domnev določimo za našo osnovno, za drugo se odločimo le, če obstaja le majhna verjetnost, da vzorec izhaja iz populacije, v kateri drži osnovna domneva.

- V nalogi smo izpeljali porazdelitev testne statistike na vzorcih iz populacije, v kateri drži ničelna domneva. Na podlagi te porazdelitve smo določili kriterij za odločanje glede na željeno napako prve vrste. Nasprotno nam porazdelitev testne statistike pod alternativno domnevo omogoča izračun moči našega testa.

- V nalogi smo pokazali, da je moč testa odvisna od:

  - velikosti vzorca: večji kot je naš vzorec, bolj smo prepričani v svoje zaključke in manjša je verjetnost, da zagrešimo napako;
  - velikosti razlike v populaciji: večja bo razlika, lažje bomo razločili med skupinama tudi že pri manjših vzorcih;
  - variance v populaciji: če so razlike med enotami v populaciji majhne, tudi z majhnim vzorcem ne moremo pretirano zgrešiti prave ocene povprečja. Če pa je razpršenost v populaciji velika, bodo tudi vzorčne vrednosti zelo razpršene, naše ocene precej nenatančne, in zato naši zaključki manj gotovi.

- Opisani test imenujemo test razmerja verjetij. Seveda bi se lahko v danem primeru domislili tudi kake druge testne statistike, a uporabljena ima zelo pomembno lastnost. Neyman-Pearsonova lema (Rice, 2009, stran 332) namreč pove, da ima test razmerja verjetij pri odločanju med dvema enostavnima domnevama največjo moč med vsemi možnimi testi, ki bi jih lahko definirali pri enaki velikosti α.

4.4 Enostavni domnevi, posplošitev

Prejšnjo nalogo zapišimo v splošnem (Perman, 2013).
Predpostavljamo, da so opazovane vrednosti neodvisne, enako porazdeljene slučajne spremenljivke $X_1, X_2, ..., X_n$. Predpostavite, da sta samo dve možnosti: ali je gostota spremenljivk enaka $f(x)$ ali pa $g(x)$, kjer sta $f(x)$ in $g(x)$ znani pozitivni gostoti. Formalno postavimo:

$$H_0\text{: gostota je } f(x) \quad \text{in} \quad H_1\text{: gostota je } g(x)$$

a) Predlagajte testno statistiko za preizkušanje zgornje domneve, če imate opazovane vrednosti $x_1, ..., x_n$.

Kot testno statistiko uporabimo razmerje

$$L = \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$$

Velike vrednosti L so 'dokazno gradivo' proti ničelni domnevi.

b) Kdaj bi zavrnili ničelno domnevo pri stopnji značilnosti α? Izrazite aproksimativno kritično mejo s količinama

$$a = \int_{\mathbb{R}} \log\left(\frac{g(x)}{f(x)}\right) f(x)dx \text{ in } b = \int_{\mathbb{R}} \log\left(\frac{g(x)}{f(x)}\right)^2 f(x)dx$$

Zanima nas porazdelitev testne statistike pod ničelno domnevo, za testno statistiko vzemimo

$$W = \sum_{i=1}^{n} \log \frac{g(X_i)}{f(X_i)}$$

Opazimo, da je testna statistika vsota slučajnih spremenljivk $Y_i = \log \frac{g(X_i)}{f(X_i)}$, ki so neodvisne in enako porazdeljene, saj so take tudi spremenljivke X_i. Zato lahko uporabimo centralni limitni izrek - spremenljivka W je približno normalno porazdeljena (ne glede na porazdelitev X_i). Da bi lahko izračunali poljubno verjetnost, moramo poznati parametra te normalne porazdelitve. Pod ničelno domnevo lahko pričakovano vrednost Y_i izračunamo kot

$$E_0(Y_i) = \int_{\mathbb{R}} \log\left(\frac{g(x)}{f(x)}\right) f(x)dx$$

Pričakovana vrednost spremenljivke W je torej na. Podobno lahko izpeljemo, da je varianca spremenljivke W pod ničelno domnevo enaka

$$\text{var}_0(W) = n\text{var}_0(Y_i) = n(b - a^2)$$

Aproksimativno torej velja $P(l > na + z_\alpha\sqrt{n(b-a^2)}) \approx \alpha$.

Povzetek

- Naloga predstavlja posplošitev prejšnje - medtem ko sta v prejšnji nalogi obe domnevi predpostavljali normalno porazdelitev, smo tu vzeli povsem splošni porazdelitvi. V obeh primerih je razmerje gostot smiselna testna statistika, vendar pa v splošnem njene porazdelitve seveda ne poznamo.

- V nalogi smo približek za porazdelitev testne statistike poiskali s pomočjo centralnega limitnega izreka. Testno statistiko smo zapisali kot vsoto neodvisnih enako porazdeljenih slučajnih spremenljivk (funkcij slučajnih spremenljivk), zato konvergira proti normalni porazdelitvi. Izpeljali smo tudi parametra te normalne porazdelitve (povprečje in std. odklon) in tako lahko določili ustrezne meje za testno statistiko.

- Za nekatere porazdelitve f in g (npr. normalno, eksponentno ipd.) lahko izpeljemo natančno porazdelitev testne statistike. Kadar to ni mogoče, si pomagamo s približkom, ki velja asimptotsko. Vendar pa to ne zagotavlja, da bo test zadovoljivo deloval na majhnih vzorcih in pod ničelno domnevo dejansko zavračal z želeno napako α. Kadarkoli imamo torej le asimptotske približke testne statistike, moramo veljavnost testa na majhnih vzorcih preveriti s simulacijami.

4.5 Posplošeni test razmerja verjetij

Zanima nas, ali imajo zares vsi športniki enako variabilnost hemoglobina. Primerjati želimo meritve k športnikov, naj bodo vrednosti i-tega športnika ($i = 1, \ldots, k$) porazdeljene normalno, torej $X_{ij} \sim N(\mu_i, \sigma_i^2)$, kjer $j = 1, \ldots, n_i$ označujejo meritve pri posamezniku. Predpostavimo, da so vse meritve med seboj neodvisne.

a) Zapišite ničelno in alternativno domnevo ter testno statistiko (posplošeni test razmerja verjetij).

Ničelna domneva: H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
Alternativna domneva:

$$H_1 : \sigma_i^2 \text{ niso vse enake}$$

Testna statistika posplošenega testa razmerja verjetij je enaka:

$$\Lambda = \frac{\max\limits_{\theta \in \omega_A \cup \omega_0} L(x, \theta)}{\max\limits_{\theta \in \omega_0} L(x, \theta)} = \frac{L(x, \widehat{\theta}_{A \cup 0})}{L(x, \widehat{\theta}_0)}$$

kjer sta ω_0 in ω_A prostora parametrov, kot jih določa posamezna domneva, $\widehat{\theta}_0$ in $\widehat{\theta}_A$ pa oceni po metodi največjega verjetja glede na posamezno domnevo. Ker se števec podatkom lahko bolj 'prilagodi' kot imenovalec, bo vrednost ulomka vedno večja od 1. Večja kot je, bolj je to dokaz v prid alternativne domneve, zanima nas, ali je vrednost Λ pod ničelno domnevo nenavadno velika. Da bi preverili ničelno domnevo, moramo torej narediti naslednje: zapisati funkcijo verjetja pod ničelno in alternativno domnevo, oceniti parametre po metodi največjega verjetja za obe domnevi, vstaviti cenilke za parametre v izraz za Λ in izračunati izraz na svojih podatkih.

b) Najprej vzemimo, da imamo le enega športnika in n njegovih meritev. Kako bi ocenili njegova parametra μ in σ^2 z metodo največjega verjetja? Funkcija verjetja je enaka

$$L(x, \mu, \sigma) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x_j - \mu)^2}{2\sigma^2}\}$$

del njenega logaritma, v katerem nastopata parametra, ki ju želimo oceniti, pa je enak

$$\log L(x, \mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2$$

Poiščimo maksimum po μ:

$$\begin{aligned}
\frac{\partial \log L(x, \hat{\mu}, \hat{\sigma})}{\partial \mu} &= 0 \\
-\frac{1}{2\widehat{\sigma}^2} \sum_{j=1}^{n} (x_j - \widehat{\mu})(-2) &= 0 \\
\sum_{j=1}^{n} (x_j - \widehat{\mu}) &= 0 \\
\widehat{\mu} &= \frac{1}{n} \sum_{j=1}^{n} x_j
\end{aligned}$$

Pa še za varianco:

$$\begin{aligned}
\frac{\partial \log L(x, \hat{\mu}, \hat{\sigma})}{\partial \sigma} &= 0 \\
-\frac{n}{\widehat{\sigma}} - \frac{1}{2} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 \frac{-2}{\widehat{\sigma}^3} &= 0 \\
-\widehat{\sigma}^2 n + \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 &= 0 \\
\widehat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2
\end{aligned}$$

c) Vrnimo se h k športnikom. Utemeljite, da so, kadar prostora možnih parametrov ne omejujemo z ničelno domnevo, ocene parametrov enake

$$\begin{aligned}
\hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \\
\hat{\sigma}_i^2 &= \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)^2
\end{aligned}$$

Funkcija verjetja je enaka

$$L(x, \mu, \sigma) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{(x_{ij} - \mu_i)^2}{2\sigma_i^2}$$

Vsak člen vsote, ki jo dobimo po logaritmiranju gornje funkcije, je sestavljen le iz parametrov enega posameznika, ko odvajamo po tistem parametru, ostanejo torej le členi, ki so vezani na tistega posameznika. Za ocenjevanje parametrov za nekega posameznika i torej potrebujemo izključno njegove vrednosti, parametre posameznikov ocenimo povsem neodvisno drug od drugega.

d) Kakšna je ocena povprečij pod ničelno domnevo?

Pod ničelno domnevo je σ_i enak za vse i, zato ga v logaritmu funkcije verjetja lahko izpostavimo in ne vpliva na našo oceno posameznih povprečij. Ocena posameznih povprečij je zato enaka kot, če dovolimo različne variance.

e) Kakšna je ocena variance pod ničelno domnevo?

Del logaritma funkcije verjetja, ki nas zanima, je enak

$$\log L(x, \mu, \sigma_0) = -\sum_{i=1}^{k} n_i \log \sigma_0 - \frac{1}{2\sigma_0^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2.$$

Odvod po σ izenačimo z 0 in dobimo

$$\hat{\sigma}_0^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)^2$$

f) Kako bi ničelno domnevo preverili s testom razmerja verjetij?

Zapišemo Wilksov Λ (zgoraj je funkcija verjetja na celotnem prostoru

parametrov, spodaj pod ničelno domnevo):

$$\begin{aligned}\Lambda &= \frac{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{\sqrt{2\pi}\widehat{\sigma}_i}\exp\{-\frac{(x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\}\right)}{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{\sqrt{2\pi}\widehat{\sigma}_0}\exp\{-\frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\}\right)} \\ &= \frac{\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\widehat{\sigma}_i}\right)\prod_{i=1}^{k}\exp\{-\frac{\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\}}{\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\widehat{\sigma}_0}\right)\exp\{-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\}}\end{aligned}$$

Vstavimo ocene za variance v eksponent in tako v števcu kot v imenovalcu dobimo $\exp\{-\frac{1}{2}\sum_{i=1}^{k} n_i\}$, ki se zato pokrajša. Logaritem Λ je enak

$$\begin{aligned}\log\Lambda &= -\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(\widehat{\sigma}_i)\right)+\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(\widehat{\sigma}_0)\right) \\ &= \left(\sum_{i=1}^{k} n_i\log(\widehat{\sigma}_0)\right)-\left(\sum_{i=1}^{k} n_i\log(\widehat{\sigma}_i)\right) \\ &= \sum_{i=1}^{k} n_i\left[\log(\widehat{\sigma}_0)-\log(\widehat{\sigma}_i)\right]\end{aligned}$$

Dvakratna vrednost logaritma verjetij je porazdeljena kot χ^2_{k-1}, saj smo pod ničelno domnevo ocenili $k-1$ parametrov manj.

Predlogi za vaje v R

- Generirajte primer s k športniki in n_i meritvami pri vsakem športniku. Individualna povprečja športnikov naj bodo normalno porazdeljena okrog 148 s standardnim odklonom 7. Vzemite, da ničelna domneva drži in da so standardni odkloni okrog individualnih povprečij za vse športnike enaki (npr. 5). Na vsakem generiranem vzorcu ocenite potrebne parametre in izračunajte vrednost testne statistike. Porazdelitev vzorčnih vrednosti testne statistike primerjajte s teoretično (asimptotsko) porazdelitvijo.

- Pri delu si lahko pomagate s spodnjo kodo. Oglejte si, kako spreminjanje parametrov vpliva na to, ali bo asimptotska porazdelitev uporabna pri vašem primeru - predvsem spreminjajte velikost vzorca, torej število športnikov in število meritev na športnika. Opazili boste, da je pri majhnih vzorcih asimptotska porazdelitev dokaj slab približek. Ali to v vašem primeru pomeni, da se vrednost α poveča ali zmanjša?
- Spremenite še vrednosti varianc in izračunajte moč testa za nekaj primerov.

```
> runs <- 100 #stevilo simulacij
> rez<- rep(NA,runs) #pripravimo za rezultate

> k <- 30 #stevilo sportnikov
> ni <- 10 #stevilo meritev na sportnika
> mi <- rnorm(k,148,7) #povprecna vrednost za vsakega sportnika
> si <- rep(5,k) #variance za vsakega sportnika

> for(jt in 1:runs){ #simulacija po korakih

> x <- NULL #pripravimo za vpisovanje vrednosti
> for(it in 1:k){ #za vsakega sportnika posebej
> xij <- rnorm(ni,mi[it],si[it]) #generiramo vrednosti
> x <- c(x,xij) #zdruzimi s prejsnjimi
> }
> data <- data.frame(id=rep(1:k,each=ni),x=x)  #podatki

#sedaj pa ocenimo parametre s pomocjo podatkov
> mihat <- sihat <- rep(NA,k)
> for(it in 1:k){ #za vsakega sportnika
> mihat[it] <- 1/ni*sum(data$x[data$id==it]) #ocena povprecja
#ocena standardnega odklona:
> sihat[it] <- sqrt(1/ni * sum((data$x[data$id==it] - mihat[it])^2))
> }

> mihat <- rep(mihat,each=ni)
#ocena standardnega odklona pod nicelno domnevo
> s0hat <- sqrt(1/(ni*k)* sum((data$x - mihat)^2))

#log Lambda:
> logL <- ni*sum(log(s0hat) - log(sihat))

> rez[jt] <- 2*logL #vrednost testne statistike na tem vzorcu
> }
```

```
> plot(ecdf(rez)) #empiricna porazdelitvena funkcija
> x <- seq(min(rez),max(rez),length=100)
> lines(x,pchisq(x,k-1),col=4) #teoreticna porazdelitvena funkcija
```

Povzetek

- Posplošeni test razmerja verjetij predstavlja generičen pristop, ki nam poda smiselno testno statistiko in njeno porazdelitev. Testna statistika je enaka kvocientu maksimumov dveh verjetij - na celotnem prostoru parametrov in na podprostoru, ki ga določa ničelna domneva. V našem primeru je celoten prostor parametrov dovoljeval k vrednosti za varianco, medtem ko smo pod ničelno domnevo zahtevali, da so vse te vrednosti enake. Nasprotno smo v primeru iz naloge 2.2 v številcu vzeli le vrednost pod alternativno domnevo in ne vseh možnih vrednosti parametra (torej pod ničelno in pod alternativno domnevo).

- Porazdelitev testne statistike pri posplošenem testu razmerja verjetij je seveda le približna in na majhnih vzorcih je lahko približek precej slab. To pa pomeni, da bomo ob uporabi te porazdelitve na majhnih vzorcih zavračali z verjetnostjo, ki ne bo povsem enaka α. To smo raziskali s primerom v R in pokazali, da na majhnih vzorcih (pri majhnem številu meritev na športnika) ničelno domnevo zavračamo prepogosto.

- Mimogrede opazimo tudi, da ocena σ po metodi največjega verjetja ni nepristranska - metoda največjega verjetja zagotavlja le doslednost (cenilka konvergira k pravi vrednosti).

Poglavje 5

Linearna regresija

To poglavje predstavlja le kratek uvod v linearno regresijo (Rice, 2009, poglavje 14). V prvi nalogi je preizkušanje domnev v linearni regresiji predstavljeno le kot poseben primer, pri katerem lahko uporabimo različne testne statistike z različnimi lastnostmi. Ogledali si bomo Waldov test ter izpeljali testno stastistiko za posplošeni test razmerja verjetij. Druga naloga je posvečena matričnemu računanju ter izpeljavi cenilk in njihovih lastnosti v linearni regresiji. Tovrstni pristop je neprimerno bolj pregleden od klasičnega, in zato nujen za vsakršne zahtevnejše izpeljave. Za konec sledi še naloga, ki se dotakne predpostavk linearne regresije in nakaže težave, ki se pojavijo, če so posamezne predpostavke kršene.

5.1 Linearna regresija

Zanima nas povezanost števila ur učenja na teden z rezultatom na izpitu iz statistike. Vzemimo, da vemo, da se rezultat na izpitu v populaciji porazdeljuje pogojno normalno: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

a) Iz spodnjega izpisa preberite ocene populacijskih parametrov. Interpretirajte rezultate, katere ničelne domneve so testirane in kako?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-20.683  -4.746   2.844   4.512  14.693
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2049     7.5172   2.555 0.033921 *
x             3.6850     0.6217   5.927 0.000351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 8 degrees of freedom
Multiple R-squared: 0.8145,     Adjusted R-squared: 0.7913
F-statistic: 35.13 on 1 and 8 DF,  p-value: 0.0003508
```

Ocene parametrov so $\widehat{\beta}_0 = 19{,}2$, $\widehat{\beta}_1 = 3{,}7$, $\widehat{\sigma} = 11{,}4$. Testirani sta dve ničelni domnevi: H_{0int}: $\beta_0 = 0$ in H_0: $\beta_1 = 0$. Pri linearni regresiji nas ponavadi zanima le druga - saj ta govori o povezanosti med spremenljivkama v populaciji. Pri iskanju porazdelitve cenilke $\widehat{\beta}_1$ bi se lahko oprli na teorijo metode največjega verjetja, vendar pa v tem primeru aproksimacija ni potrebna. Cenilka je namreč linearna kombinacija vrednosti Y (glej nalogo 3.2), zato je normalno porazdeljena. Njena varianca (standardna napaka) je ocenjena iz podatkov, zato je standardizirana vrednost cenilke porazdeljena kot t. Testna statistika

$$T = \frac{\widehat{\beta}_1}{\widehat{SE_{\beta_1}}} = \frac{3{,}7}{0{,}6} = 5{,}9$$

je torej porazdeljena kot t z 8 stopinjami prostosti (pri ocenjevanju SE porabimo dve stopinji prostosti). Ta test se imenuje Waldov test.

b) Kako bi ničelno domnevo H_0: $\beta_1 = 0$ preverili s posplošenim testom razmerja verjetij?
Namig: kjer je le mogoče, uporabite rezultate iz prejšnje naloge.

Začnimo z ocenami pod ničelno domnevo. Pod ničelno domnevo je povprečje za vse posameznike enako, neposredno torej lahko uporabimo rezultate iz naloge 4.5, le da namesto μ pišemo β_0, zato je maksimum funkcije

verjetja pod ničelno domnevo enak

$$\begin{aligned} L_0(y,x,\widehat{\beta}_0,\widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{\sum_{i=1}^n(y_i-\widehat{\beta}_0)^2}{2\widehat{\sigma}^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n\sum_{i=1}^n(y_i-\widehat{\beta}_0)^2}{2\sum_{i=1}^n(y_i-\widehat{\beta}_0)^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n}{2}\right\} \end{aligned}$$

Za števec bomo uporabili rezultat, da je ocena $\widehat{\sigma}$ enaka

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n(y_i-\widehat{\beta}_0-x_i\widehat{\beta}_1)^2}{n}$$

Funkcija verjetja je enaka:

$$L(y,x,\beta_0,\beta_1,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\{-\frac{\sum_{i=1}^n(y_i-\beta_0-\beta_1x_i)^2}{2\sigma^2}\}$$

in tako je maksimum funkcije verjetja enak

$$\begin{aligned} L_{A\cup 0}(y,x,\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{2\widehat{\sigma}^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{2\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n}{2}\right\} \end{aligned}$$

Wilksov Λ je enak

$$\begin{aligned} \Lambda &= \frac{L_{A\cup 0}}{L_0} = \frac{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_A)^n}\exp\left\{-\frac{n}{2}\right\}}{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_0)^n}\exp\left\{-\frac{n}{2}\right\}} \\ &= \frac{\widehat{\sigma}_0^n}{\widehat{\sigma}_A^n} \\ &= \left(\frac{\sum_{i=1}^n(y_i-\widehat{\beta}_{00})^2}{\sum_{i=1}^n(y_i-\widehat{\beta}_{0A}-x_i\widehat{\beta}_{1A})^2}\right)^n \end{aligned}$$

Vrednost maksimuma pri neomejenem prostoru parametrov izračunamo tako, da vstavimo ocenjene $\widehat{\beta_0}$ in $\widehat{\beta_1}$, za izračun vrednosti pod ničelno domnevo moramo oceniti še β_0 v ničelnem modelu. Vrednost $2 \log \Lambda$ se porazdeljuje kot χ^2_1.

Predlogi za vaje v R

- Naj bo X enakomerno porazdeljena spremenljivka (med 0 in 20, zaokrožena navzdol), $\beta_0 = 15$, $\beta_1 = 4$, $\sigma = 10$. Generirajte vzorec velikosti 10, narišite podatke in vrišite populacijsko ter ocenjeno vrednost premice.

```
> set.seed(1)
> n <- 10                       #velikost vzorca
> beta0 <- 15
> beta1 <- 4
> sigma <- 10
> x <- floor(runif(n)*20)       #navzdol zaokrozene vrednosti x
> x <- sort(x)                  #uredimo podatke po velikosti x
> y <- rnorm(n,mean=beta0+beta1*x,sd=sigma)  #generiramo y-one
> plot(x,y)                     #narisemo tocke
> popul <- beta0 + beta1*x      #populacijska vrednost premice
> lines(x,popul,col="grey",lwd=2) #dodamo popul. vrednost premice
> fit <- lm(y~x)                #ocenimo premico na podatkih
> summary(fit)                  #ogledamo si ocene koeficientov
> beta0h <- fit$coef[1]         #ocenjena beta0
> beta1h <- fit$coef[2]         #ocenjena beta1
> napoved <- beta0h + beta1h*x
> lines(x,napoved,lwd=2)        #vrisemo ocenjeno premico na sliko
```

- Izračunajte posplošeni test razmerja verjetij v R

```
> fit0 <- lm(y~1)                  #pod nic. domnevo - le konstanta
> res0 <- y - fit0$coef            #ostanki pod nicelno domnevo
> resA <- y - beta0h - beta1h*x    #ostanki pod alternativno domnevo
#zanima nas razlika log verjetij - konstanto lahko izpustimo:
> logl0 <- -.5*n*log(sum(res0^2))  #loglik pod nicelno
> loglA <- -.5*n*log(sum(resA^2))  #loglik pod alternativno
> Lambda <- 2*(loglA-logl0)        #Wilksov lambda
> 1-pchisq(Lambda,1)               #likelihood ratio test
[1] 4.048e-05
```

Povzetek

- V nalogi smo si ogledali osnovni primer linearne regresije z eno neodvisno spremenljivko. Najprej smo si ogledali izpis programa R in ga interpretirali. Omenili smo dve možnosti preizkušanja domneve o regresijskem koeficientu: Waldov test in posplošeni test razmerja verjetij.
- Pri obeh omenjenih testih je znana le asimptotska porazdelitev testne statistike, ničesar pa nismo povedali o kvaliteti aproksimacije na majhnih vzorcih. Izkaže se, da ima posplošeni test razmerja verjetij na majhnih vzorcih pogosto boljše lastnosti od Waldovega testa, ki je avtomatično poročan v vseh statističnih programih.

5.2 Matrično računanje

Vrednosti neodvisnih spremenljivk združimo v matriko X (design matrix), vrednosti odvisne spremenljivke ter koeficientov predstavljajo vektorja Y in β:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}; \; Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrika X je dimenzije $n \times (p+1)$, kjer je p število spremenljivk. Če naš model ne bi vseboval konstante, bi prvi stolpec X izpustili.

a) Zapišite vsoto vrednosti $\sum_{i=1}^{n} Y_i^2$ v matrični obliki.

$$Y^T Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = Y_1^2 + Y_2^2 + \ldots + Y_n^2$$

b) Kaj dobimo, če matrično pomnožimo $X\beta$ (da bo manj pisanja, vzemite $p = 1$)?

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = E(Y)$$

c) V matrični obliki oceno koeficientov po metodi najmanjših kvadratov (= po metodi največjega verjetja) zapišemo kot $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Pokažite, da za $p = 1$ dobite oceni:

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Izračunajmo najprej $X^T X$:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Inverz te 2×2 matrike je enak:

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

Izračunajmo še $X^T Y$:

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}$$

Velja torej

$$\begin{aligned}(X^TX)^{-1}X^TY &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix}\sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n\end{bmatrix}\begin{bmatrix}\sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_iY_i\end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix}\sum_{i=1}^n x_i^2\sum_{i=1}^n Y_i - \sum_{i=1}^n x_i\sum_{i=1}^n x_iY_i \\ -\sum_{i=1}^n x_i\sum_{i=1}^n Y_i + n\sum_{i=1}^n x_iY_i\end{bmatrix}\end{aligned}$$

Zgornja vrstica pri tem predstavlja oceno $\widehat{\beta}_0$, spodnja pa $\widehat{\beta}_1$.

d) Izpeljite oceno po metodi najmanjših kvadratov še v matrični obliki. Pri tem boste potrebovali naslednje formule za matrično računanje:

$$\begin{aligned}(A+B)^T &= A^T + B^T;\ \ (A^T)^T = A;\ \ (AB)^T = B^TA^T; \\ \frac{\partial\beta^TA}{\partial\beta} &= A;\ \ \frac{\partial\beta^TA^TA\beta}{\partial\beta} = 2A^TA\beta\end{aligned}$$

Namig: kaj minimiziramo? Kako zapišemo vsoto kvadriranih ostankov v matrični obliki?

Če je v modelu ena spremenljivka, iščemo minimum funkcije

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1x_i)^2$$

V matrični obliki iščemo minimum funkcije

$$\begin{aligned}(Y - X\beta)^T(Y - X\beta) &= (Y^T - \beta^TX^T)(Y - X\beta) \\ &= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta\end{aligned}$$

Pri tem smo v zadnji vrstici uporabili, da je $(\beta^TX^TY)^T = \beta^TX^TY$, saj je matrika dimenzije 1×1. Sedaj odvajamo po β

$$\frac{\partial}{\partial\beta}(Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta) = -2X^TY + 2X^TX\beta$$

in izenačimo z 0 (ter predpostavimo, da X^TX ni singularna):

$$\begin{aligned} -2X^TY + 2X^TX\widehat{\beta} &= 0 \\ X^TX\widehat{\beta} &= X^TY \\ \widehat{\beta} &= (X^TX)^{-1}X^TY \end{aligned}$$

e) Pokažite, da je ocena koeficientov nepristranska (vzemite, da so vrednosti x-ov dane in ne slučajne).

$$E(\widehat{\beta}) = (X^TX)^{-1}X^TE(Y) = (X^TX)^{-1}X^TX\beta = \beta$$

f) Izpeljite formulo za standardno napako ocenjenih koeficientov v matrični obliki. Intuitivno razložite, od česa je odvisna standardna napaka koeficienta β_1 (za $p = 1$). Uporabite formulo: $\mathrm{var}(cY) = c\,\mathrm{var}Y\,c^T$.

$$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= \mathrm{var}[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\mathrm{var}Y[(X^TX)^{-1}X^T]^T \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \end{aligned}$$

Izpeljali smo variančno-kovariančno matriko, variance so na diagonali. Standardna napaka SE_{β_1} je torej enaka

$$\begin{aligned} SE_{\beta_1} &= \sqrt{\frac{\sigma^2}{(X^TX)^{-1}_{22}}} \\ &= \sqrt{\frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \\ &= \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= \sqrt{\frac{\sigma^2}{n\sigma_x^2}} = \frac{\sigma}{\sigma_x\sqrt{n}} \end{aligned}$$

Standardna napaka koeficienta je tako kot vedno odvisna od velikosti vzorca n ter razpršenosti podatkov. Vrednost σ je standardni odklon

ostankov okrog premice - večja kot je, bolj se lahko zmotimo pri oceni premice. Vendar tu ni pomembna le absolutna velikost variance ostankov, zanima nas variabilnost ostankov glede na variabilnost neodvisne spremenljivke. Če je razpon x-ov majhen, je naša ocena pri isti variabilnosti ostankov manj natančna. Razložimo to na našem primeru - če bi v vzorec zajeli le posameznike, ki so se učili 3-5 ur, bi bila povezanost med spremenljivkama (npr. merjena s korelacijskim koeficientom) pri istih regresijskih koeficientih dosti manjša, in zato bi bila možna večja odstopanja pri ocenjevanju.

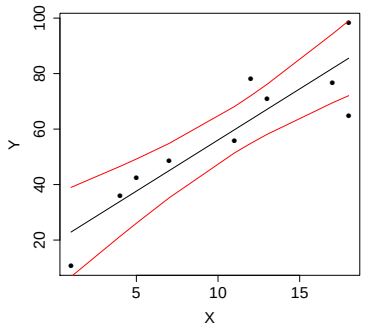
g) Kako bi izračunali interval zaupanja za napovedano premico v našem primeru ($p = 1$)? Kako bo tak interval izgledal na sliki?

Izračunamo standardno napako za vsako točko posebej.

$$\text{var}(\widehat{y_i}) = \text{var}(\widehat{\beta}_0 + \widehat{\beta}_0 x_i) = \text{var}(\widehat{\beta}_0) + x_i^2 \text{var}(\widehat{\beta}_1) + 2x_i cov(\widehat{\beta}_0, \widehat{\beta}_1)$$

Matrični izračun (za vse točke naenkrat):

$$\text{var}(\widehat{Y}) = \text{var}(X\widehat{\beta}) = X\text{var}(\widehat{\beta})X^T = \sigma^2 X(X^T X)^{-1} X^T$$



Slika 5.1: *Točke na vzorcu in ocenjena premica z intervalom zaupanja.*

h) Recimo, da nas zanima, kako sta število ur učenja in spol (0 = ženski, 1 = moški) povezana z rezultatom na izpitu iz statistike. Predpostavimo model, ki vključuje interakcijo. Kako bi preverili, ali je število ur učenja pri moških povezano z rezultatom na izpitu?

Model, ki ga prilagodimo podatkom, zapišemo kot:

$$Y = \beta_0 + \beta_1 \text{spol} + \beta_2 \text{ure} + \beta_3 (\text{ure} * \text{spol})$$

Ničelna domneva, ki jo želimo preveriti, je torej $H_0 : \beta_2 + \beta_3 = 0$.
Zapišemo v matrični obliki. Naj bo vektor $a^T = [0,0,1,1]$, zanima nas H_0: $a^T\beta = 0$. Varianca $a^T\beta$ je enaka $\text{var}(a^T\beta) = a^T \text{var}\beta a$, za preizkušanje ničelne domneve uporabimo Waldov test.

Predlogi za vaje v R

- V R v matrični obliki zapišite podatke, izračunajte oceno po metodi najmanjših kvadratov ter jo primerjajte z izpisom funkcije `lm`.
- Ocenite tudi standardno napako ter jo primerjajte z izpisom
- Narišite sliko napovedane premice ter ji dodajte interval zaupanja.

```
> X <- cbind(1,x)
> sigma <- summary(fit)$sigma
> inv <- solve(t(X)%*%X)
> mat <-  X%*%inv%*%t(X)
> se <- sigma*sqrt(diag(mat))
> betah <- c(beta0h,beta1h)
> plot(x,y)
> lines(x,X%*%betah)
> t8 <- qt(.975,8)
> lines(x,X%*%betah - t8*se,col=2)
> lines(x,X%*%betah + t8*se,col=2)
```

Povzetek

- Predstavitev podatkov v linearni regresiji z vektorji in matrikami se izkaže za zelo učinkovit ter pregleden način izpeljevanja cenilk ter dokazovanja njihovih lastnosti. Poleg tega so izpeljave splošnejše, saj niso odvisne od števila neodvisnih spremenljivk. V tej nalogi smo si zato ogledali osnove matričnega računanja ter jih uporabili za izpeljavo cenilk in lastnosti v linearni regresiji.
- Ogledali smo si tudi, kako lahko Waldov test zapišemo splošneje in z njim preverimo poljubno domnevo o koeficientih v modelu, ki jo lahko zapišemo kot linearno kombinacijo parametrov.
- Program R podatke vedno obravnava kot vektorje oziroma matrike, zato lahko cenilke v matrični obliki tudi neposredno uporabimo v kodi.

5.3 Predpostavke linearne regresije

Z osnovnim modelom linearne regresije naredimo štiri predpostavke:

- ostanki so okrog premice porazdeljeni normalno
- varianca ostankov ni odvisna od vrednosti neodvisne spremenljivke (homoskedastičnost)
- ostanki so med seboj neodvisni
- povezanost med X in Y je linearna

Kaj se zgodi z ocenami koeficientov, njihovo pričakovano vrednostjo, standardno napako in intervali zaupanja, če je katera izmed prvih treh predpostavk kršena?

a) Kaj se spremeni v izpeljavah, če ostanki okrog premice niso porazdeljeni normalno?

   V tem primeru ocena koeficientov po metodi največjega verjetja ni enaka oceni po metodi najmanjših kvadratov. Ocena po metodi najmanjših kvadratov bo identična kot do sedaj, enaka bo tudi ocena standardne napake. Prav tako bo cenilka po metodi najmanjših kvadratov nepristranska ocena populacijskih vrednosti. Da pa bi izračunali kak interval zaupanja, moramo izpeljati porazdelitev cenilke. Vemo, da je cenilka po metodi najmanjših kvadratov linearna kombinacija vrednosti Y, zato smo pri normalni porazdelitvi (pri danih x-ih) ostajali znotraj normalne porazdelitve. V splošnem to seveda ni res.
   Druga možnost je, da izpeljemo raje oceno po metodi največjega verjetja - v tem primeru se cenilke spremenijo, vemo pa, da je njihova asimptotska porazdelitev normalna.

b) Recimo, da je varianca ostankov odvisna od x.

   Če varianca ostankov ni enaka za vsak x, moramo varianco pisati v ma-

trični obliki, npr.:

$$\Sigma = \sigma \begin{bmatrix} w_1 & 0 & \dots & 0 & 0 \\ 0 & w_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & w_n \end{bmatrix}$$

Varianca sicer ne vpliva na oceno koeficientov po metodi najmanjših kvadratov, vendar pa se spremeni ocena po metodi največjega verjetja, saj moramo maksimizirati funkcijo $-2(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$:

$$\begin{aligned} &(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) = \\ &= (Y^T - \beta^T X^T)\Sigma^{-1}(Y - X\beta) \\ &= Y^T\Sigma^{-1}Y - \beta^T X^T\Sigma^{-1}Y - Y^T\Sigma^{-1}X\beta + \beta^T X^T\Sigma^{-1}X\beta \\ &= Y^T\Sigma^{-1}Y - 2\beta^T X^T\Sigma^{-1}Y + \beta^T X^T\Sigma^{-1}X\beta \end{aligned}$$

in zato

$$\begin{aligned} -2X^T\Sigma^{-1}Y + 2X^T\Sigma^{-1}X\widehat{\beta} &= 0 \\ X^T\Sigma^{-1}X\widehat{\beta} &= X^T\Sigma^{-1}Y \\ \widehat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \end{aligned}$$

Ustrezno se spremeni tudi varianca ocene:

$$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= \mathrm{var}[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] \\ &= (X^T\Sigma^{-1}X)^{-1}X^T\mathrm{var}Y[(X^T\Sigma^{-1}X)^{-1}X^T]^T \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

Če so vrednosti w_i znane, se v ocenjevanje koeficientov in standardne napake le vrine diagonalna matrika. Statistično sklepanje je enako kot do sedaj.

c) Kaj pa če ostanki med seboj niso neodvisni?

Potem variančna matrika Σ ni več diagonalna (je npr. bločno diagonalna). Rezultati bodo podobni tistim v prejšnji točki, bo pa seveda ocenjevanje odvisno od tega, kaj vemo o elementih Σ.

Povzetek

- Naloga predstavlja le uvod v zelo široko področje posplošitev linearne regresije. Osnovni primer linearne regresije namreč zahteva precej stroge predpostavke, ki so v praksi seveda pogosto kršene. V nalogi smo si le površno ogledali, kaj se v teoretičnih izpeljavah spremeni, kadar je kršena posamezna predpostavka, medtem ko se z možnimi rešitvami oziroma njihovimi lastnostmi nismo ukvarjali.

Literatura

Blagus, R. (2011). *Razvrščanje visoko-razsežnih neuravnoteženih podatkov : doktorsko delo.* Medicinska fakulteta, Univerza v Ljubljani.

Casella, G. and Berger, R. L. (1990). *Statistical inference.* Wadsworth, Belmont, CA.

Perman, M. (2013). Metodologija statističnega raziskovanja, izpitne naloge. URL: http://valjhun.fmf.uni-lj.si/ mihael/ul/vs/izpiti.html.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rice, J. A. (2009). *Mathematical Statistics and Data Analysis.* Duxbury advanced series, Cengage learning.

Shao, J. (2003). *Mathematical Statistics, second edition.* Springer-Verlag, New York.

Stvarno kazalo