Kazalo

1 Verjetnost

1.1 Normalna porazdelitev

Vemo, da je vrednost hemoglobina pri nedopingiranem športniku[1] porazdeljena normalno s povprečjem $\mu = 148$ in varianco $\sigma^2 = 85$.

- 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 $a > 0$?

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

    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, $Y \sim N(a \cdot \mu + b, (a \cdot \sigma)^2)$.
  - Kaj moramo vzeti kot a in b, da bo Y standardizirana normalna spremenljivka.
    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 standardizirano normalno porazdelitev (oz. uporabiti ustrezno numerično

[1] 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).

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 standardizirano normalno porazdelitev (ali pa s pomočjo računalnika) izvemo, da je $P(X \leq 166) = P(Y \leq 1{,}95) = 0{,}974$, zato je verjetnost $P(X > 166) = 0{,}026$.

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

  Naj bo Z standardizirana normalna spremenljivka, zanimajo nas meje, izven katerih je vrednost te spremenljivke z verjetnostjo 0,01. Če želimo postaviti simetrične meje, to pomeni, da nas zanimata tisti vrednosti, izven katerih je v repih na vsaki strani verjetnost 0,005. Iz tabel izvemo, da je $P(Z \geq 2{,}57) = 0{,}995$, ustrezna mejna vrednost standardizirane normalne spremenljivke je torej $\pm 2{,}57$.
  $Z = \frac{X-148}{\sqrt{85}}$, zato

$$\begin{aligned} 0{,}995 &= P(\frac{X-148}{\sqrt{85}} \leq 2{,}57) + P(\frac{X-148}{\sqrt{85}} > -2{,}57) \\ &= P(X \leq 148 + 2{,}57 \cdot \sqrt{85}) + P(X > 148 - 2{,}57 \cdot \sqrt{85}) \\ &= P(X \leq 171{,}1) + P(X > 124{,}3) \end{aligned}$$

- Naj bodo meje take, kot ste jih izračunali v prejšnji točki. Športnika testiramo 10x na leto. Kakšna 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 X Bernoullijevo porazdeljena spremenljivka $X \sim Ber(0{,}01)$, kjer je $\{X = 1\} = \{\text{vrednost je izven meja}\}$. Imamo 10 neodvisnih realizacij te slučajne spremenljivke, X_i, $i = 1, \ldots, 10$, za vsako velja $P(X_i = 1) = 0{,}01$. Ker so neodvisne, velja $P(X_1 = 0, X_2 = 0, \ldots, X_{10} = 0) = \{P(X_1 = 0)\}^{10}$. Verjetnost, da v 10 meritvah ne preseže meja je torej $P = 0{,}99^{10}$, verjetnost, da jih vsaj enkrat preseže, je $P = 1 - 0{,}99^{10} = 0{,}096$.

- Izračunajte porazdelitev slučajne spremenljivke X^2. Katero znano po-

razdelitev dobite?

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(X^2 \le z) = P(-\sqrt{z} \le \sqrt{X} \le \sqrt{z}) \\ &= F_X(\sqrt{z}) - F_X(-\sqrt{z}) \\ f_Z(z) &= \frac{1}{2\sqrt{z}} f_X(\sqrt{z}) + \frac{1}{2\sqrt{z}} f_X(-\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}} [f_X(\sqrt{z}) + f_X(-\sqrt{z})] \\ &= \frac{1}{2\sqrt{z \cdot 2\pi}} [e^{-z/2} + e^{-z/2}] = \frac{1}{\sqrt{z \cdot 2\pi}} e^{-z/2} \end{aligned}$$

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 $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ (to je hkrati tudi porazdelitev χ^2_1).

- 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?

  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.

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 $X \sim U[0,1]$.

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

  Generiramo[2] 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
```

- 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
```

[2]Kot rešitev vseh praktičnih nalog bo v tem gradivu podana koda za statistični paket R (prostodostopen na `http://cran.r-project.org/`), ki je trenutno med statistiki najbolj razširjen.

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

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

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

  F je torej kumulativna porazdelitvena funkcija spremenljivke X.

- 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.
Kako bi hkrati simulirali vrednosti za posameznike z različno vrednostjo λ?
Podobno kot zgoraj - le da so vrednosti λ lahko različne.

1.3 Pričakovana vrednost in varianca

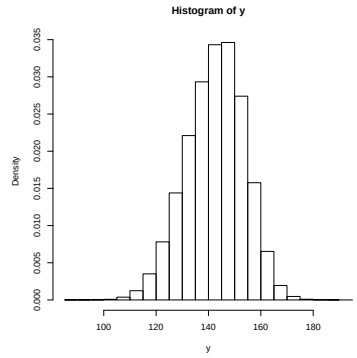
Raziskovalci na področju športa so dokazali, da je pri kolesarjih hemoglobin izven tekmovalnega obdobja porazdeljen kot $N(150, 7^2)$, med tekmovalnim obdobjem pa kot $N(140, 11^2)$. Vzemimo, da tekmovalno obdobje traja 9 mesecev. Zanimata nas pričakovana vrednost in standardni odklon za naključno odvzeti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

- Skicirajte porazdelitev Y in skušajte oceniti pričakovano vrednost ter standardni odklon

  Zgenerirajmo veliko število vrednosti in si oglejmo njihovo porazdelitev

```
> set.seed(1)
> a <- rnorm(1000000,mean=140,sd=11)                        #generiram 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,replace=T,prob=c(0.25,0.75)) #porazdelitev X - obdobje
> y <- a*x+b*(1-x)                                          #slucajna spremenljivka Y
> hist(y,prob=T)                                            #narisemo spremenljivko
> mean(y)                                                   #ocena povprecja
> var(y)                                                    #ocena standardnega odklona
```

  Dobljena ocena povprečja je 142,5, variance pa 121,87.



Slika 1: *Porazdelitev nove spremenljivke Y.*

- 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

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

  Torej

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

- 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 spre-

menljivke $Z = E(Y|X)$:

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

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 $\text{var}(Y) = 121{,}75$, $\text{sd}(Y) = 11{,}03$. Vrednosti izven tekmovalnega obdobja torej le malo povečajo variabilnost rezultatov.

- 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

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

Varianca slučajne spremenljivke Z je enaka

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

Izrazimo še drugi del, slučajna spremenljivka $\mathrm{var}(Y|X=0)$ je enaka

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

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

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

- Izračunajte kovarianco X in Y. Izrazite je 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?

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

Zanima nas torej pričakovana vrednost glede na skupno porazdelitev X in Y (lahko bi pisali $E_{X,Y}$). Pri izračunu uporabimo, da velja $f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)$, in najprej izračunamo integral po y

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

V integralu $\int E(Y) f_{Y|X}(y|x)dy$ 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 $\int f_{Y|X}(y|x)dy = 1$, torej

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

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

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

Vrednost $E(Y|X=0) = \mu_0$ je pričakovana vrednost Y pri $X = 0$, torej izven tekmovalnega obdobja, podobno 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 $E(X) = P(X=1) = p$. Zato

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

in

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

V našem primeru $cov(X,Y) = 0{,}75 * 0{,}25 * (140 - 150) = -1{,}875$, $cor(X,Y) = -\frac{1{,}875}{11{,}03 * \sqrt{0{,}75 * 0{,}25}} = 0{,}392$.

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.

- 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 nista 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.

1.4 Vsota diskretnih slučajnih spremenljivk

Naj bosta X in Y neodvisni Bernoullijevo porazdeljeni spremenljivki, $B(p)$.

- 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):

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

  Za neodvisni X in Y torej velja

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

V našem primeru:

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

Velja torej

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

- Kako pravimo porazdelitvi vsote n i.i.d. Bernoullijevih spremenljivk?

  Označimo $Z = \sum_{i=1}^{n} X_i$, $i = 1, \ldots, n$, $X_i \sim B(p)$. Z je porazdeljena binomsko, $Z \sim Bin(n, p)$.

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

1.5 Vsota zveznih slučajnih spremenljivk

Poleg posamičnih vrednosti, želijo pri športnikih proučevati tudi zaporedje večih 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).

- 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:

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

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

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

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

- Kako se porazdeljuje $S = Z_1^2 + Z_2^2$, če sta Z_1 in Z_2 porazdeljena standardno normalno?

  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.

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

Naredimo substitucijo $z_2 = sv$, $dz_2 = sdv$, meje so torej od 0 do 1:

$$\begin{aligned} f_S(s) &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^1 \frac{1}{\sqrt{s-sv}} \frac{1}{\sqrt{sv}} z dv \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^1 \frac{1}{\sqrt{1-v}} \frac{1}{\sqrt{v}} dv \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \pi \\ &= \frac{1}{2} \exp\left\{-\frac{s}{2}\right\} \end{aligned}$$

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 $a = 1$.

- 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 $\sum_{i=1}^n X_i \sim \Gamma(\frac{n}{2}, \frac{1}{2})$ (dokazali smo le za $n = 2$).

```
> 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
```

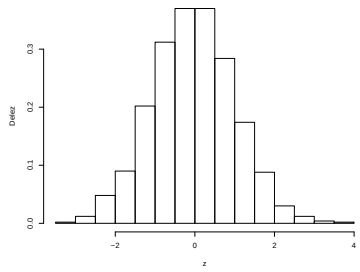
Naša ničelna domneva je, da športnik ni kriv. Pod to ničelno domnevo se vsota kvadriranih odmikov porazdeljuje po gama porazdelitvi $\Gamma(\frac{5}{2}, \frac{1}{2})$. Verjetnost, da je vsota 12,57 ali več, je približno 0,03.

1.6 Vsota dveh odvisnih spremenljivk

Naj bosta X in Y neodvisni standardizirani normalni spremenljivki, Z pa enaka $|Y|$, če je $X \geq 0$, in $-|Y|$ če je $X < 0$.

- Kako je porazdeljena spremenljivka Z?
  Najprej narišimo simulirane vrednosti z R-om:

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



Slika 2: *Porazdelitev spremenljivke Z.*

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

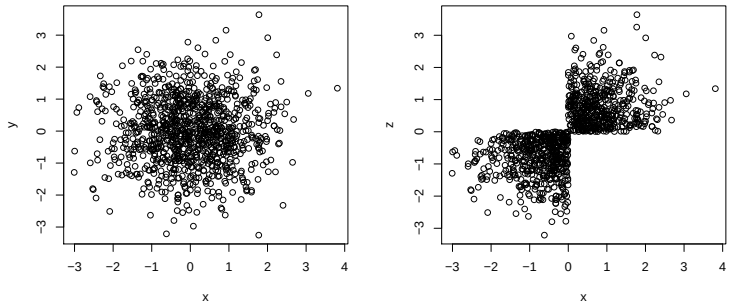
$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X < 0, -|Y| \leq z) \\ &= P(X < 0)[P(Y \leq z) + P(Y \geq -z)] \\ &= \frac{1}{2}[2 \cdot P(Y \leq z)] \\ &= P(Y \leq z) = F_Y(z) \end{aligned}$$

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 standardizirana normalna spremenljivka.

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

```
> plot(x,y)
> plot(x,z)
```

Očitno je, da spremenljivki nista neodvisni, vedno imata enak predznak.

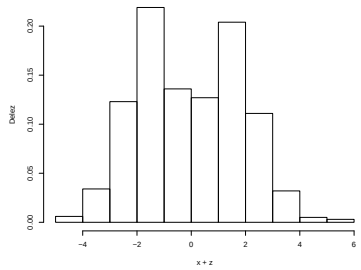


Slika 3: *Razsevni diagram realizacij X in Y ter X in Z.*

- Ali je vsota $X + Z$ porazdeljena normalno?

Iz slike je očitno, da porazdelitev vsote ni normalna. Vsota dveh normalnih spremenljivk torej ni nujno normalna, če spremenljivki nista normalni (dana naloga je protiprimer).

```
> hist(x+z,main="",prob=T,ylab="Delez")
```



Slika 4: *Porazdelitev vsote $X + Z$.*

1.7 Porazdelitev 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, zato 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 $\sigma^2 = 5^2$.

- Pokažite, da je vsota dveh neodvisnih standardiziranih normalnih spremenljivk spet normalna. Za vsoto dveh splošnih normalnih spremenljivk izpeljite le pričakovano vrednost in varianco

  Naj bosta $X \sim N(0,1)$ in $Y \sim N(0,1)$, 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

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

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

Naj bosta $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ in med seboj neodvisni:

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) = \mu_1 + \mu_2 \\ \mathrm{var}(X+Y) &= \mathrm{var}(X) + \mathrm{var}(Y) = \sigma_1^2 + \sigma_2^2 \end{aligned}$$

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

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

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

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

1.8 Razvrščanje

Na podlagi nekega kazalnika želimo ocenjevati kreditno sposobnost posameznika, želimo jih razvrstiti 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. 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 razvrščanje 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 razvrščanja. 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 (torej X_s in X_d enako porazdeljena, označimo kar z X: $X \sim N(50, 15^2)$). Ugotoviti želimo, kakšna bo verjetnost, da neko naključno stranko na podlagi današnje vrednosti njenega kazalnika razvrstimo med "dobre".

- 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 $\bar{X}_s = \frac{1}{n_s} \sum X_{s,i} \sim N(\mu, \frac{\sigma^2}{n_s})$.

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

  Iz prejšnje naloge vemo, da velja $X - \bar{X}_s \sim N(0, \sigma^2 + \frac{\sigma^2}{n_s})$.

- Označimo $Z = X - \bar{X}_s$ in $Y = X - \bar{X}_d$. Zapisati želimo formulo za gostoto $f_{Z,Y}(z, y)$.
  V ta namen izračunajte kovarianco spremenljivk $X - \bar{X}_s$ in $X - \bar{X}_d$. Poizkusite skicirati nekaj realizacij teh dveh slučajnih spremenljivk za $n_s = n_d = 10$ (z razsevnim diagramom). Izračunajte korelacijo med spremenljivkama za $n_s = n_d$. Kako je korelacija odvisna od velikosti učnega vzorca?

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

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

  Korelacija je enaka

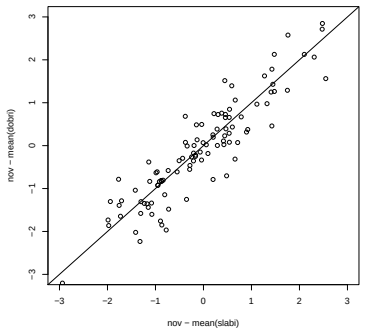
  $$\begin{aligned} cor[X - \bar{X}_s, X - \bar{X}_d] &= \frac{\sigma^2}{\sqrt{\sigma^2(1 + \frac{1}{n_s})}\sqrt{\sigma^2(1 + \frac{1}{n_d})}} \\ &= \frac{1}{\sqrt{(1 + \frac{1}{n_s})}\sqrt{(1 + \frac{1}{n_d})}} \end{aligned}$$

  Če sta velikosti vzorca med seboj enaki, velja

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

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).

```
>  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 10 dobrih
+  slabi <- rnorm(10)              # in 10 slabih, ter enega novega, ki ga na podlagi
+  dobri <- rnorm(10)               # dobljenega kriterija uvrstim
+  nov <- rnorm(1)
+ points(nov-mean(slabi),nov-mean(dobri))
+ }
```



Slika 5: *Razsevni diagram razlik za 100 realizacij slučajne spremenljivke.*

Zapišimo še skupno gostoto (gostota bivariatne normalne spremenljivke):

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

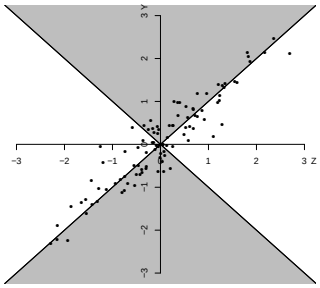
kjer je $s=\sqrt{1+\frac{1}{n_s}}$ in $d=\sqrt{1+\frac{1}{n_d}}$. Označimo še $c=\sqrt{\frac{n_s n_d}{1+n_s+n_d}}$ in

dobimo

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

- 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 6 prikazuje območje integracije. Integral, ki ga moramo izračunati



Slika 6: *Območje* $|Z| < |Y|$*, za* $n_s = 10$*,* $n_d = 20$*.*

je enak:

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

Zaradi simetrije je integral za pozitivne y (glej sliko 6) 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:

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

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 razvrščanja enote v vsako skupino razvrstil 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 razvrščanje torej ne daje želenih oz. pričakovanih rezultatov (bodo pa podatki morali biti zelo neuravnoteženi, da bo verjetnost bistveno različna od 0,5).

1.9 Centralni limitni izrek

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. Kakšna je verjetnost, da bo med 6 posamezniki večina za gradnjo?

- 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, verjetnost posameznega izida je

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

- Izračunajte verjetnost, da so vsaj 4 posamezniki za gradnjo.

  Z uporabo formule za binomsko porazdelitev:

$$P(S_n > 3) = \sum_{k=4}^{6} \binom{6}{k} (0{,}65)^k (0{,}35)^{6-k} = 0{,}647$$

- Aproksimirajte to verjetnost še s pomočjo centralnega limitnega izreka.

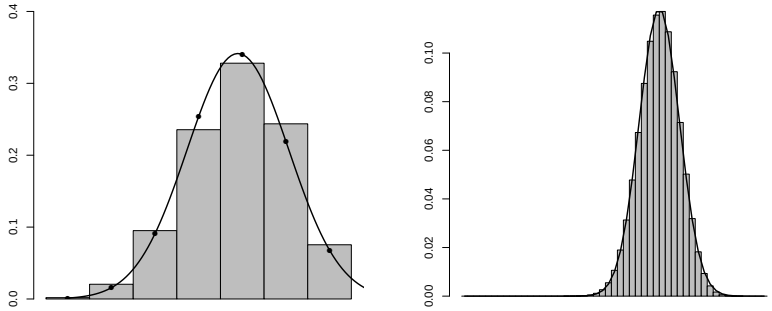
  Vemo, da $\frac{S_n - np}{\sqrt{np(1-p)}}$ konvergira (v porazdelitvi) proti standardni normalni spremenljvki Z. Poglejmo, kako dobra bo aproksimacija z normalno porazdelitvijo pri $n = 6$:

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

- 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)  #dodamo se kup tock, v katerih nas zanima gostota
> 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
```



Slika 7: *Aproksimacija binomske porazdelitve z normalno za* $p = 0{,}65$ *in (a)n=6, (b)n=500.*

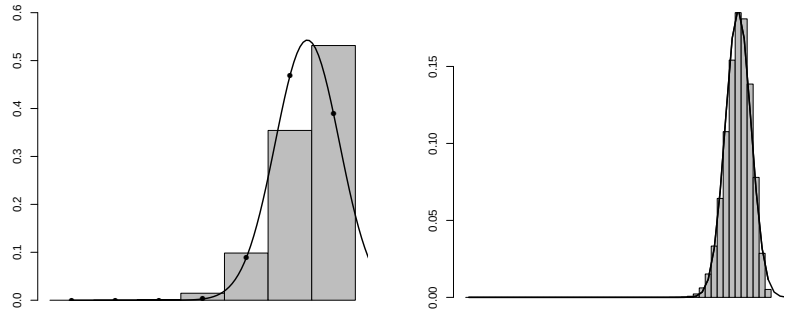
Sliko ponovimo še za druge vrednosti p, vidimo, da je aproksimacija nekoliko slabša, če je porazdelitev bolj asimetrična, a še vedno zelo dobra.

2 Vzorčenje

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.

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



Slika 8: *Aproksimacija binomske porazdelitve z normalno za* $p = 0{,}9$ *in (a)n=6, (b)n=500.*

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 $E(X_i) = \mu$.

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

- 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

- Naj bo varianca v populaciji enaka σ^2. Kakšna je standardna napaka

naše ocene?

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

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

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

- Na podlagi našega vzorca želimo oceniti σ^2. Naj bo naša cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$. Kakšna mora biti vrednost konstante c, da bo naša ocena 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 $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$:

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

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

- 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.

Povzemimo rezultate naloge: če je populacija neskončna in enote v vzorcu neodvisne, velja:

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

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.

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

  Populacija končna, označimo vrednosti v populaciji z x_k, $i = k, \ldots, 11$. Izberimo po nekem vrstnem redu vseh 11 podjetij, prvih 6 naj jih 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 $cov(X_1, X_2) = cov(X_i, X_j)$ 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.

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

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

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

  In zato (za $i \neq j$)

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

  Spremenljivki sta torej negativno povezani. Ta povezanost je seveda šibka, korelacija je enaka

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

  Korelacija je torej odvisna izključno od velikosti populacije.

- Izračunajte standardno napako našega vzorca (privzemite da poznate σ^2). V drugi skupini imamo 100 podjetij. Kako velik vzorec moramo vzeti iz te skupine, da bomo dobili približno enako veliko standardno napako (privzemimo da je varianca tudi v tej skupini enaka σ^2)?

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

  Za $N = 11$ in $n = 6$ dobimo $SE^2 = \frac{\sigma^2}{6}\frac{11-6}{(10)} = \frac{\sigma^2}{12}$. Če je populacija večja, bomo za enako standardno napako potrebovali več enot. Pri $N = 100$, dobimo željeno velikost napake z vzorcem velikosti 11.

- Kakšna 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 $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1}$:

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

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

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

- Zapišite še nepristransko cenilko za standardno napako

  Združimo dosedanje rezultate in dobimo

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

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

Povzemimo rezultate naloge: če je populacija končna, velja:
Vrednosti izbrane v vzorec navkljub naključnemu izbiranju med seboj niso neodvisne, kovarianca med enotami je enaka:

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

varianca povprečja in nepristranska cenilka za varianco v populaciji pa sta enaki

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

2.3 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_j) in rezultatom na testu (Y_j).
Za vsakega posameznika iz vzorca imamo torej par slučajnih spremenljivk (X_i, Y_i), $i = 1 \ldots n$.

- Utemeljite, da je količina $cov(X_i, Y_j)$ 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_i) imajo tako enako porazdelitev in zato je enaka tudi kovarianca X_i in Y_j.

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

  Vsota vseh vrednosti iz populacije je konstanta, zato velja

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

  Velja torej (za $i \neq j$)

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

- Kovarianca med spremenljivkama X in Y je definirana kot

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

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

  Na vzorcu 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

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

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

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

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

(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

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

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 (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 $\frac{1}{n-1}\frac{N-1}{N}$.

- Kako bi ocenili korelacijo? Ali je taka ocena korelacije nepristranska? Preverite s simulacijo.

  Uporabimo izpeljane formule za oceno kovariance in varianc:

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

  Nepristranskosti te ocene še nismo dokazali, saj pričakovana vrednost kvocienta ni enaka kvocientu pričakovanih vrednosti. Ker ne vemo, ali je ocena 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)

```
> set.seed(1)
> xi <- runif(300)*40+25                #300 posameznikov, starosti 25-65 let
> yi <- 100 - xi + rnorm(300)*20        #rezultat na testu 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 simulacije
> for(it in 1:runs){                    #simulacija po korakih
+ inx <- sample(1:length(xi),size=10,replace=F)     #izberemo vzorec 10-ih
+ 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))   #izracunamo korelacijo
+ }

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

Vidimo, da sta obe 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.

```
> (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.

2.4 Enostavni slučajni vzorec, 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 $C = \frac{1}{n}\sum_{i=1}^{N} I_i x_i$.

- Koliko je vsota $\sum_{i=1}^{N} I_i$? Kakšna 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:
  Jemljem 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:

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

- 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

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

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

  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 $cov(I_1, I_1 + \ldots I_N) = cov(I_1, n) = 0$ in $cov(I_i, I_j)$ je enaka za vsak $i \neq j$:

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

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

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

Populacijska varianca definirana kot:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 \\
\sum_{i=1}^{N} x_i^2 &= N(\sigma^2 + \mu^2)
\end{aligned}$$

Izpeljimo varianco:

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

2.5 Vzorčenje po skupinah

Oceniti želimo dosežek ljubljanskih sedmošolcev na nekem testu znanja, ki ga izvajajo v večih 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) 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.

- 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

$$\mu = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{K} N_i \cdot \mu_i \tag{2}$$

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

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

- 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

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

- 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.

- Zapišite varianco cenilke s pomočjo varianc in kovarianc

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

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

  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 $E(\bar{X}_i^2) = \mathrm{var}(\bar{X}_i) + E(\bar{X}_i)^2 = \frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1} + \mu_i^2$ in dobimo

$$\mathrm{var}(I_i\bar{X}_i) = \frac{k}{K}\left(\frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1} + \mu_i^2\right) - \left(\frac{k}{K}\right)^2\mu_i^2 = \mu_i^2\frac{k(K-k)}{K^2} + \frac{k}{K}\frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1}$$

  Sedaj izrazimo še kovarianco:

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

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

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

- 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.

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

Velja:

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

in zato

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

- Kaj bi se razlikovalo v naših izračunih če bi šole vzorčili proporcionalno glede na njihovo velikost, tako da bi bila verjetnost, da je izbrana šola

i enaka $\frac{kN_i}{N}$?

Če bi bile vse šole enako velike, bi se spremenil le izračun kovariance. Ker ne vemo, kakšna bo velikost vzorca, vsota I_i ni več konstanta, zato ne moremo kovariance izračunati z istim "trikom" - potrebno jo bo izračunati po definiciji.

3 Metoda največjega verjetja

3.1 Ocenjevanje deleža

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

- Recimo, da je $n = 5$ in da smo dobili naslednjih 5 vrednosti: 1,0,1,1,1. Kakšna 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?

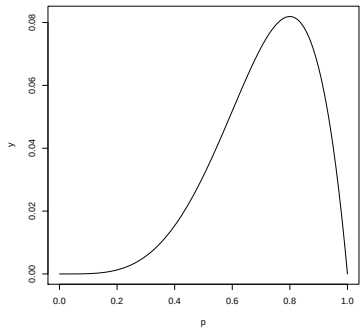
  Verjetnost dogodka izračunamo kot $0{,}2^4 0{,}8^1$, torej $p^k(1-p)^{n-k}$, kjer je k število enk. Označimo z A dogodek $A = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1\}$. Za $p = 0{,}2$ dobimo $P(A) = 0{,}00128$, za $p = 0{,}75$ dobimo $P(A) = 0{,}079$. Narišemo krivuljo za vrednosti p med 0 in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

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

  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.

- Podatke, ki jih dobimo na nekem vzorcu, označimo z $x_1, \dots, x_n$ (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.

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



Slika 9: *Verjetnost opaženega dogodka glede na p.*

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

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

- 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 enk v vzorcu.

- 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.

- Zapišite oceno standardne napake

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

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

  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 imenovalcu, dobimo

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

- 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 $p - \hat{p}$ približno normalno porazdeljena, zato je kvocient $\frac{p-\hat{p}}{\widehat{SE}}$ porazdeljen po t porazdelitvi z 499 prostostnimi stopnjami (ocenili smo 1 parameter). Ustrezna mejna vrednost t je torej 1,965. 95 % interval zaupanja je enak $[0{,}25, 0{,}33]$.

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.

- Zapišite porazdelitev prihodka podjetja, če vemo, da je varianca enaka σ^2.

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

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

- 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

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

- Izračunajte oceni β_0 in β_1 po metodi največjega verjetja

  Najprej za β_0:

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

  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

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

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

- 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

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

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

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

3.3 Moč testa

Iz literature lahko povzamemo, da se športnikovo povprečje hemoglobina ob vsaj 14-dnevnem bivanju na višini nad 1500m 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. Kakšna bo moč njegovega testa, če bo pri sklepanju uporabil stopnjo značilnosti $\alpha = 0{,}05$?

- Kaj je športnikova ničelna in kaj alternativna domneva?

  Predpostavljamo, da se hemoglobin v obdobju višinskih priprav porazdeljuje kot $N(\mu_2,5^2)$. 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.

- Kakšno testno statistiko bo uporabil za preverjanje ničelne domneve?

  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 $Z > 1{,}64$.

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

  Zanima nas $P(Z > 1{,}64)$, torej $P(R > 1{,}64 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$, torej v našem primeru $P(R > 5{,}02)$. Pod alternativno domnevo je $R \sim N\left(2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$ in zato

  $$P\left(R > 1{,}64 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = P\left(\frac{\bar{X}_2 - \bar{X}_1 - 2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} > 1{,}64 - \frac{2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}\right)$$

  $$P\left(U > 1{,}64 - \frac{2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right),$$

  kjer je U standardna normalna spremenljivka. V našem primeru:

  $$P\left(U > 1{,}64 - \frac{2}{5\sqrt{\frac{1}{8} + \frac{1}{4}}}\right) = P(U > 0\ 99) = 0{,}16 \qquad (3)$$

  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).

- 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

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

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

  Iz enačbe (3) 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.

4 Test razmerja verjetij

4.1 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.

- Zapišite ničelno in alternativno domnevo

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

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

- 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:

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

Pa še za varianco:

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

- Utemeljite, da so pod alternativno domnevo ocene parametrov enake

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

Funkcija verjetja pod alternativno domnevo je enaka

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

Funkcijo logaritmiramo, del v katerem nastopata parametra za nek i je neodvisen od ostalih delov in zato enak tistemu, kar bi dobili, če bi imeli le posameznika i. Zato dobimo gornji oceni.

- 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 pod alternativno domnevo.

- Kakšna je ocena variance pod ničelno domnevo?
  Del logaritma funkcije verjetja, ki nas zanima, je enak

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

  Odvod po σ izenačimo z 0 in dobimo

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

- Kako bi ničelno domnevo preverili s testom razmerja verjetij?

  Zapišemo Wilksov Λ (zgoraj je funkcija verjetja pod alternativno do-

mnevo, spodaj pod ničelno):

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

Vstavimo ocene za variance v eksponent in tako v števcu kot tudi 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 alternativno domnevo ocenili $k-1$ parametrov več kot pod ničelno.

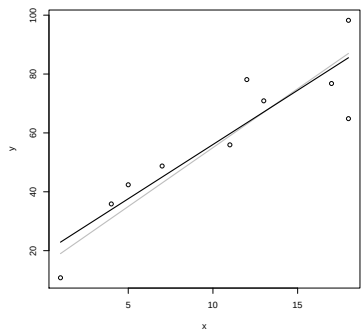
5 Linearna regresija

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: $Y|X \sim N(\beta_0+\beta_1 X, \sigma^2)$.

- 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 iz normalne porazdelitve
> plot(x,y)                                      #narisemo tocke
> popul <- beta0 + beta1*x                       #populacijska vrednost premice
> lines(x,popul,col="grey",lwd=2)                #dodamo populacijsko vrednost premice v sivi barvi
> 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
```



Slika 10: *Točke na vzorcu, ocenjena premica (črna) in populacijska premica (siva).*

- 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. Metoda največjega verjetja nam pove, da se ocene parametrov okrog prave vrednosti porazdeljujejo približno normalno (za dovolj velik n). Standardna napaka je ocenjena iz podatkov, uporabimo test t:

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

Vemo, da je slučajna spremenljivka T porazdeljena približno kot t z 8 stopinjami prostosti (pri ocenjevanju SE porabimo dve stopinji prostosti). Ta test se imenuje Waldov test.

- Kako bi ničelno domnevo $H_0 : \beta_1 = 0$ preverili s testom razmerja verjetij?
  Uporabite rezultat, da ocena $\widehat{\sigma}$ po metodi največjega verjetja ni nepristranska in je enaka

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

  Izračunati moramo vrednost maksimuma funkcije verjetja pod ničelno in alternativno domnevo. Funkcija verjetja je enaka:

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

Maksimum funkcije verjetja je enak

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

Logaritem funkcije verjetja v ocenjenih vrednostih je zato enak

$$\begin{aligned} &\log L(y, x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}) = \\ &= -\frac{n}{2}\left(\log(2\pi) - \log[\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2] + \log n - 1 \right) \end{aligned}$$

Vrednost maksimuma pod alternativno domnevo 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. Dobljeni Wilksov Λ se porazdeljuje kot χ^2_1.

```
> fit0 <- lm(y~1)                          #ocenimo premico pod nicelno domnevo - le konstanta
> res0 <- y - fit0$coef                    #ostanki pod nicelno domnevo
> resA <- y - beta0h - beta1h*x            #ostanki pod alternativno domnevo
> logl0 <- .5*n*(-log(2*pi) - log(sum(res0^2))+log(n) - 1 )   #loglik pod nicelno
> loglA <- .5*n*(-log(2*pi) - log(sum(resA^2))+log(n) - 1 )   #loglik pod alternativno
> Lambda <- 2*(loglA-logl0)                #Wilksov lambda
> 1-pchisq(Lambda,1)                       #likelihood ratio test
[1] 4.048e-05
```

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. V našem primeru imamo le eno neodvisno spremenljivko, matrika X je enaka:

```
> X <- cbind(1,x)                          #zlepimo dva stolpca
> X
         x
 [1,] 1  1
 [2,] 1  3
 [3,] 1  5
 [4,] 1  7
 [5,] 1  7
 [6,] 1  8
 [7,] 1 11
 [8,] 1 16
 [9,] 1 19
[10,] 1 19
> round(y)              #zaokrozimo za vecjo preglednost
 [1] 11 36 42 49 56 78 71 77 65 98
```

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

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

- Kaj dobimo, če matrično pomnožimo $X\beta$?

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

- 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

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

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

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

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

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

Iščemo minimum funkcije

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

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

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

in izenačimo z 0:

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

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

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

- 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, da velja $\mathrm{var}(cY) = c\ \mathrm{var}Y c^T$.

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

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 standardna napaka ostankov okrog premice - večja kot je, bolj se lahko zmotimo pri oceni premice. Vendar tu ni pomembna 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 možna večja odstopanja pri ocenjevanju.

- Kako bi izračunali interval zaupanja za napovedano premico v našem primeru ($p = 1$)? Dodajte ga na sliko

  Izračunamo standardno napako za vsako točko posebej.

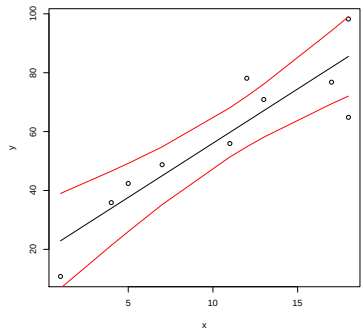
  $$\mathrm{var}(\widehat{y_i}) = \mathrm{var}(\widehat{\beta}_0 + \widehat{\beta}_0 x_i) = \mathrm{var}(\widehat{\beta}_0) + x_i^2\mathrm{var}(\widehat{\beta}_1) + 2x_i cov(\widehat{\beta}_0, \widehat{\beta}_1)$$

Matrični izračun (za vse točke naenkrat):

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

Izračun v R-u za naš primer:

```
> 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)
```



Slika 11: *Točke na vzorcu in ocenjena premica z intervalom zaupanja.*

- 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 preverjanje ničelne domneve uporabimo test t.

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 z intervali zaupanja, če je katera izmed prvih treh predpostavk kršena?

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

  Najprej vzemimo, da 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 ocena po metodi najmanjših kvadratov nepristranska ocena populacijskih vrednosti. Ne moremo pa o populacijskih vrednostih sklepati ničesar več, saj ne poznamo porazdelitve ocene okrog prave vrednosti.

- Recimo, da je varianca ostankov odvisna od x.

  Če varianca ostankov ni enaka za vsak x, moramo varianco pisati v matrični 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:

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

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

- 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.