1.6 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 $\sigma^2 = 5^2$, in da so posamezne vrednosti med seboj neodvisne.

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

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

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

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

- Izračunajte meje okrog ocenjenega povprečja, znotraj katerih naj bi pri šesti meritvi nedopingiran športnik ostal z verjetnostjo 0,99 (prvih pet meritev uporabimo za oceno športnikovega povprečja).
  Namig: uporabite rezultat, da je vsota neodvisnih normalno porazdeljenih spremenljivk spet normalna.
  Vrednosti posameznika so porazdeljene kot $X \sim N(\mu, \sigma^2)$ ($\sigma = 5$). Vsota $\sum_{i=1}^{5} X_i$ je normalna spremenljivka, prav tako je normalno porazdeljeno tudi povprečje, saj vsoto le pomnožimo s konstanto. Zanima nas odstopanje šeste meritve od ocenjenega povprečja v prvih petih meritvah, torej razlika $Z = X_6 - \frac{1}{5}\sum_{i=1}^{5} X_i$. To je torej razlika dveh normalnih spremenljivk z enakim povprečjem, ena ima varianco σ^2, druga pa σ^2/n. Spremenljivka Z je torej porazdeljena kot

$Z \sim N(0, \sigma^2 + \sigma^2/n) = N(0{,}30)$. Vrednost $z_{0{,}005} = 2{,}57$, meje so torej $\frac{1}{5}\sum_{i=1}^{5} X_i \pm 2{,}57 \cdot \sqrt{30}$.

- Ali je povprečje neodvisnih enako porazdeljenih slučajnih spremenljivk vedno znova porazdeljeno z isto porazdelitvijo?

  Ne, to v splošnem ni res. Protiprimer je na primer vsota Bernoullijevo porazdeljenih spremenljivk, ki ni porazdeljena po Bernoullijevi porazdelitvi in prav tako to očitno ne velja za povprečje.

Predlogi za vaje v R-u:

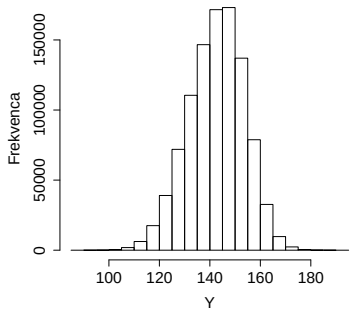
- Predpostavite, da so povprečja športnikov normalno porazdeljena z $N(148, 7{,}5^2)$ in zgenerirajte povprečne vrednosti za 100 športnikov. Nato uporabite normalno porazdelitev $N(0, 5^2)$, ki predstavlja odmike od osebnega povprečja vsakega športnika - zgenerirajte po 6 vrednosti na posameznika. Ocenite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

1.7 Pogojna pričakovana vrednost in varianca

Raziskovalci na področju športa so dokazali, da je pri kolesarjih hemoglobin izven tekmovalnega obdobja porazdeljen kot $N(150, 7^2)$, med tekmovalnim obdobjem pa kot $N(140, 11^2)$. Vzemimo, da tekmovalno obdobje traja 9 mesecev. Zanimata nas pričakovana vrednost in standardni odklon 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, kaj lahko rečete o pričakovani vrednosti ter standardnem odklonu?

  Porazdelitev je na sliki 1. Pričakujemo, da bo povprečna vrednost nekje med vrednostima v obeh obdobjih. Ker je verjetnost, da je meritev izšla iz tekmovalnega obdobja večja, pričakujemo, da vrednost nekoliko bližja povprečju v tem obdobju.



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

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

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

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

  Sestavimo oba dela skupaj in dobimo $\text{var}(Y) = 121{,}75$, $\text{sd}(Y) = 11{,}03$. 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

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

Izrazimo še drugi del, slučajna spremenljivka $\mathrm{var}(Y|X)$ 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?

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

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

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

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

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

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

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

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

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

in

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

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

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 neodvisni - od vrednosti X je odvisna varianca Y. Porazdelitev Y je torej odvisna od X, četudi X ne vpliva na povprečje. Torej, vemo, da je korelacija enaka 0, če sta spremenljivki neodvisni, vendar obratno ni nujno res.

Predlogi za vaje v R-u:

- Zgenerirajte veliko število vrednosti in si oglejte njihovo porazdelitev:

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

```
> hist(y,prob=T)                    #narisemo spremenljivko
> mean(y)                           #ocena povprecja
> var(y)                            #ocena variance
```

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