1 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 ($N(0,(\sqrt{\frac{1}{2}})^2)$), njen integral je zato 1. 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:

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

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?

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

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

- Izračunajte meje okrog ocenjenega povprečja, znotraj katerih naj bi pri šesti meritvi nedopingiran športnik ostal z verjetnostjo 0,99
  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 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 $\frac{1}{5}\sum_{i=1}^{5} X_i \pm 2{,}57 * \sqrt{30}$.

2 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 ($X \sim N(50, 15^2)$). 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 $\bar{X}_s = \frac{1}{n_s} X_{s,i} \sim N(\mu, \frac{\sigma^2}{n_s})$.

- 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 $Y = X - \bar{X}_d$. Zapišati 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

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

  Korelacija je enaka

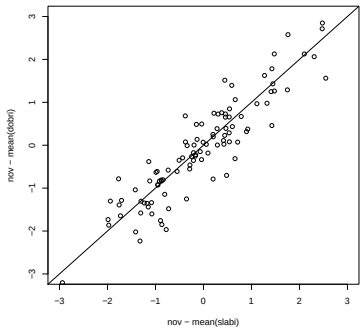
$$\begin{aligned} cor[X - \bar{X}_s, X - \bar{X}_d] &= \frac{\sigma^2}{\sqrt{\sigma^2(1 + \frac{1}{n_s})}\sqrt{\sigma^2(1 + \frac{1}{n_d})}} \\ &= \frac{1}{\sqrt{(1 + \frac{1}{n_s})}\sqrt{(1 + \frac{1}{n_d})}} \end{aligned}$$

  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 povpre"cij
>  abline(0,1)                      #simetrala
>  for(it in 1:99){                #ponovim 100x
+  slabi <- rnorm(10)
+  dobri <- rnorm(10)
+  nov <- rnorm(1)
+ points(nov-mean(slabi),nov-mean(dobri))
+ }
```



Slika 1: *Razsevni diagram razlik za 100 realizacij slučajne spremenljivke.*

Zapišimo še skupno gostoto:

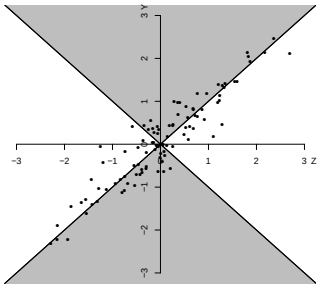
$$\begin{aligned} f_{Z,Y}(z,y) &= \frac{1}{2\pi\sigma_Z\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(z-\mu_Z)^2}{\sigma_Z^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right.\right. \\ &\quad \left.\left.-\frac{2\rho(z-\mu_z)(y-\mu_y)}{\sigma_Z\sigma_Y}\right]\right) \\ &= \frac{1}{2\pi\sigma^2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

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

  Slika 2 prikazuje območje integracije. Integral, ki ga moramo izračunati



Slika 2: *Območje* $|Z| < |Y|$*, za* $n_s = 10$*,* $n_d = 20$*.*

je enak:

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

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.

3 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:

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

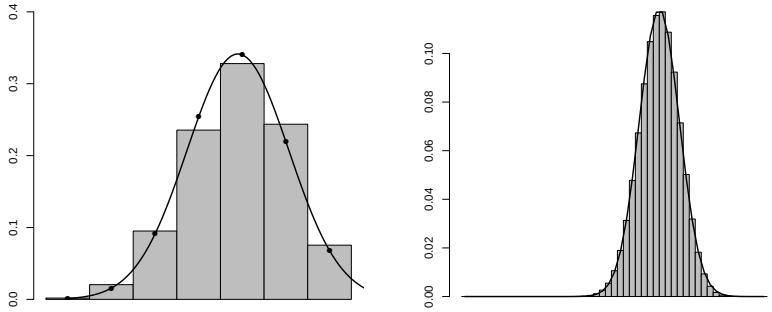
- Aproksimirajte to verjetnost še s pomočjo centralnega limitnega izreka.

  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 $n = 6$:

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

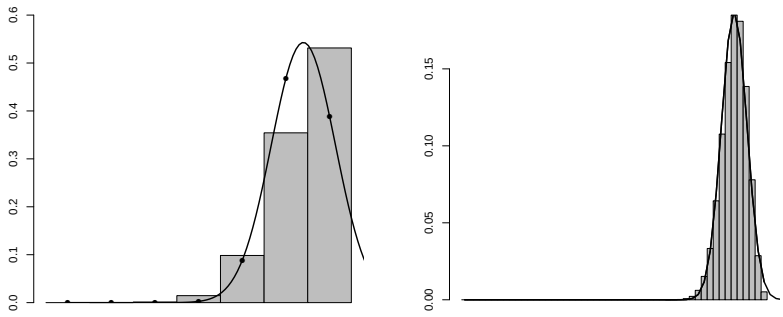
- Oglejte si, kako dobra je aproksimacija za različne velikosti vzorca in različne vrednosti p.

```
> 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 3: *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.



Slika 4: *Aproksimacija binomske porazdelitve z normalno za* $p = 0{,}9$ *in (a)n=6, (b)n=500.*