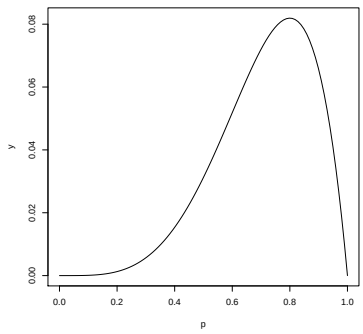
3 Ocenjevanje parametrov - metoda največjega verjetja

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:



Slika 1: *Verjetnost opaženega dogodka glede na p.*

  Vrh funkcije lahko poiščemo z odvajanjem - odvajamo funkcijo $p^k(1-p)^{n-k}$ po p in izenačimo z 0 (lokalni maksimum). Vrh ni odvisen od vrstnih redov.
  V našem primeru je vrh funkcije dosežen pri $p = 4/5$.

- 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.

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

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

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

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

Predlogi za vaje v R-u:

- Z R-om narišite sliko 1:

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

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

- Na vsakem vzorcu ocenjenemu deležu dodajte še 95% interval zaupanja. Kakšen je delež vzorcev, pri katerih interval zaupanja zajema pravo vrednost (0,3)?

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

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

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

- 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:

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

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:

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

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

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

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

- Izračunajte standardno napako za obe oceni.

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

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

  Prva odvoda sta enaka

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

  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

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