3 Parameter estimation - maximum likelihood method

3.1 Estimation of the proportion

Let $x_1, \dots x_n$ be independent realizations of a Bernoulli distributed random variable X. We wish to estimate the parameter p.

- Say that $n = 5$ and that we got the following 5 values: $1, 0, 1, 1, 1$. What would be the probability of this event if $p = 0.2$? What if $p = 0.75$? Plot the curve of these probabilities for various values of p. How would you calculate its peak?

  The probability of this event can be calculated as $0.2^4 0.8^1$, or, in general $p^k(1-p)^{n-k}$, where k is the number of 1s. Denote the event $A = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1\}$. For $p = 0.2$, we have $P(A) = 0.00128$, for $p = 0.75$, we have $P(A) = 0.079$. Plot the curve of probabilities for values of p between 0 and 1:

  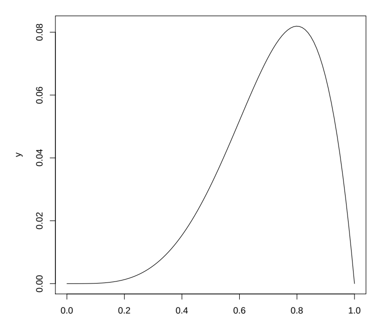

  Figure 1: *The probability of observed event for a given p.*

  The peak of this function can be found by derivation - we take the derivative of $p^k(1-p)^{n-k}$ by p and equal it to 0 (local maximum).
  In our case, the function peaks at $p = 4/5$.

- The data obtained on a sample are denoted as $x_1, \dots, x_n$ (in the above case, $n = 5$, $x_1 = 1$ in $x_2 = 0$). Write the value of $P(X_i = x_i|p)$, i.e.

the probability that the event, we have seen, has happened. Write the likelihood function.

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

The likelihood function is a product of probabilities (we've assumed that the random variables X_i are independent), therefore

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

- Find the estimator of p using the maximum likelihood method

  Since logarithm is a monotone function, the maximum of the logarithm shall be at the same point as the maximum of the function itself:

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

  The derivative of the logarithm equals 0 for $\hat{p}n = \sum_{i=1}^{n} x_i$. The maximum likelihood estimate equals $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. This is the sample proportion.

- Is this estimate unbiased?

  The maximum likelihood method ensures only consistency (unbiasedness, when $n \to \infty$), in our case, we get

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

We see that our estimator is unbiased.

- How can you estimate the standard error?

  The variance of the estimator equals $\frac{1}{n}I(p)^{-1}$, where

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

  Since both formulae give equally complex calculation in our case, we use the first one:

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

  When calculating the expected value, we use that $E(X) = p$. Since X

only appears in the numerator, we get

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

- We wish to estimate the proportion of voters for a certain candidate. In a sample of $n = 500$, he gets 29 % of the votes. Give the 95 % confidence interval for this estimate.

  The sample estimate equals $\hat{p} = 0.29$. Standard error (i.e. the standard deviation of the estimator) on the sample can be estimated using $\hat{p}$, we get

  $$\widehat{SE} = \sqrt{\frac{1}{nI(\hat{p})}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02.$$

  As we know from the theory, the $\frac{p-\hat{p}}{\widehat{SE}}$ is approximately normally distributed. The 95 % confidence interval equals $[0, 25, 0, 33]$.

Understanding the ideas in R:

- Use R to plot the Figure 1:

```
> p <- seq(0,1,length=100)     #for 100 values p between 0 in 1
> y <- p^4*(1-p)               #calculate the probability for each value
> plot(p,y,type="l")           #plot them as a curve
```

- Generate a sample of size 500, in which every individual has the probability 0.3 to vote for a certain candidate. Estimate the probability using the sample proportion. Repeat this procedure 1000x and look at the distribution of sample estimates.

- Add the estimated 95% confidence interval for each sample. What is the proportion of the samples, on which the interval encompasses the true value (0.3)?

3.2 The association of two random variables

We wish to know, how the revenue of a company in a certain branch depends on the number of employees. Assume that the income is randomly distributed with the average $\beta_0 + \beta_1 X$, where X is the logarithm of the number of employees. Say that we have data on a sample of companies and would like to estimate the parameters β_0 and β_1.

- What is the density of the company revenues if we know the variance equals σ^2?

  We assume that $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, therefore

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

- Write the likelihood function. What is the function to maximize?

  We have the data (x_i, y_i), $i = 1, \ldots, n$.

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

  The logarithm of this function equal

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

  Since we are only interested in the parameters β_0 in β_1, the first part of the function can be seen as the constant and we only need to maximize the expression

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

- Estimate β_0 and β_1 using the maximum likelihood method

  First for β_0:

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

Equalling the above expression to 0, we get (the two values of β_0 and β_1 for which the term equals 0 are denoted with a hat)

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

We now take the derivative with respect to β_1:

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

If the above expression equals 0, we get

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

Combining the two expressions (after a bit of algebra), we get

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

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

- Calculate the standard error of both estimates.

  We have to calculate the second derivatives needed for the Fisher information matrix. The logarithm of the likelihood function equals

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

  The first derivatives equal

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

  The second derivatives equal

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

  The terms in the Fisher information matrix are the negative expected values of the second derivatives. Since we do not know the expected value of X or X^2, we estimate them from the data:

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

The inverse of the matrix then equals

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

and hence

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

and

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