Q1). It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%.

a) [10pt] If an email is detected as spam, then what is the probability that it is in fact a non-spam email?

b) [5pt] And what is the probability that it is really a spam email?

Solution:

a).

Define events

$$A = \text{event that an email is detected as spam,}$$
$$B = \text{event that an email is spam,}$$
$$B^c = \text{event that an email is not spam.}$$

We know $P(B) = P(B^c) = .5, \qquad P(A \mid B) = 0.99, \qquad P(A \mid B^c) = 0.05.$

Hence by the Bayes's formula we have

$$\begin{aligned} P(B^c \mid A) &= \frac{P(A \mid B^c)P(B^c)}{P(A \mid B)P(B) + P(A|B^c)P(B^c)} \\ &= \frac{0.05 \times 0.5}{0.05 \times 0.5 + 0.99 \times 0.5} \\ &= \frac{5}{104}. \end{aligned}$$

b). probability of spam email

$$P(B \mid A) : 1 - \frac{5}{104}$$

Q2). Let X be a single observation from a Binom(n, p), where p is an unknown parameter. (In this case, we will consider n known.)

a) [10pt] Find the maximum likelihood estimator (MLE) of p.

b) [5pt] Suppose you roll a 6-sided die 40 times and observe eight rolls of a 6. What is the maximum likelihood estimate of the probability of observing a 6?

Solution:

a).

We just have *one observation*, so the likelihood is just the pmf:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 < p < 1, \quad x = 0, 1, \dots n$$

The log-likelihood is:

$$\log L(p) = \log \left\{ \binom{n}{x} \right\} + x \log(p) + (n-x) \log(1-p).$$

The derivative of the log-likelihood is:

$$\frac{d}{dp} \log L(p) = \frac{x}{p} - \frac{n-x}{1-p}.$$

Setting this to be 0, we solve:

$$\frac{x}{p} - \frac{n-x}{1-p} = 0 \iff x - px = np - px \iff p = \frac{x}{n}.$$

b).

Here, we can let X be the number of sixes in 40 (independent) rolls of the die: $X \sim \text{Binom}(40, p)$, where p is the probability of rolling a 6 on this die.

Then

$$\boxed{\hat{p} = \frac{8}{40} = 0.2}$$

is the maximum likelihood **estimate** for p.

Q3). Let $(x_1, y_1),\ldots (x_n, y_n)$ be n points.

a) [10pt]. Derive the least squares regression line for the usual simple linear regression: $y = \beta_0 + \beta_1 x + \varepsilon$.

b) [5pt] Derive the least squares regression line for simple linear regression through the origin.

Solution:

a).

To minimize

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

$$\begin{cases} \frac{\partial SS}{\partial \beta_0} = 0 \\ \frac{\partial SS}{\partial \beta_1} = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \sum_{i=1}^{n} x_i(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the least squares regression line.

b).

To minimize

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

$$\frac{\partial SS}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 x_i) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$\hat{y} = \hat{\beta}_1 x$ is the least squares regression line through the origin.