Part 0

Introduction and Preliminaries

Contents

1 Introduction **1**
- 1.1 What is Statistical Modelling? 1
- 1.2 What is the aim of this course? 1
  - 1.2.1 Example 1
  - 1.2.2 Notation and terminology 2
  - 1.2.3 Course objectives 3

2 Distributions and inference **4**
- 2.1 Random variables and probability distributions 4
  - 2.1.1 Single random variable 4
  - 2.1.2 Two random variables 5
- 2.2 Statistics of a sample 6
- 2.3 Inferring from a sample 7
  - 2.3.1 The method of moments 8
  - 2.3.2 The method of maximal likelihood 8
  - 2.3.3 The method of least squares 11
- 2.4 Some distribution theory relating to the normal distribution 11
- 2.5 Covariance matrix and some matrix algebra 13

3 Textbooks and online resources **15**

1 Introduction

1.1 What is Statistical Modelling?

Statistical modelling is the process of describing variation in observed data with appropriate probability distributions. For example, suppose we are investigating hospital performance, and we wish to compare mortality rates for a surgical procedure at two hospitals. (Such comparisons are given at `https://drfoster.com`). In hospital A we observe x_A deaths in n_A operations, and in hospital B we observe x_B deaths in n_B operations. A plausible statistical model for these data is

$$x_A \sim Binomial(n_A, \theta_A),$$
$$x_B \sim Binomial(n_B, \theta_B),$$

with θ_A and θ_B the unknown model parameters. Questions of interest regarding the two mortality rates can then be formulated as questions about the model parameters.

For example, the question

"Do the data suggest that a patient undergoing the operation in hospital A is more likely to die than a patient undergoing the operation in hospital B?"

can be formulated as

"Do the data provide evidence that $\theta_A > \theta_B$?"

Note that the role of statistical modelling in comparing mortality rates is crucial here. Simply comparing x_A/n_A with x_B/n_B is not sufficient, as differences in the observed rates could be due to chance alone. The statistical model describes the possible random variation in the data for any given values of θ_A and θ_B, and so will help us understand when differences in observed mortality rates are a cause for concern.

1.2 What is the aim of this course?

In this course, we will learn how to build statistical models to describe relationships between different variables.

1.2.1 Example

A study has investigated the relationship between smoking and lung cancer in males. For 25 different occupational groups, the mean number of cigarettes smoked per day and the rate of deaths from lung cancer have been recorded. From these, a "smoking index" and a "mortality index" have been calculated. The smoking index for each group is the ratio of the mean number of cigarettes smoked per day in that group to the mean number of cigarettes smoked per day by all men. The mortality index for each group is the ratio of the rate of deaths from lung cancer in that group to the rate of deaths from lung cancer among all men. The data are shown in Figure 1.1. (Source: Her Majesty's Stationery Office, London, 1978).

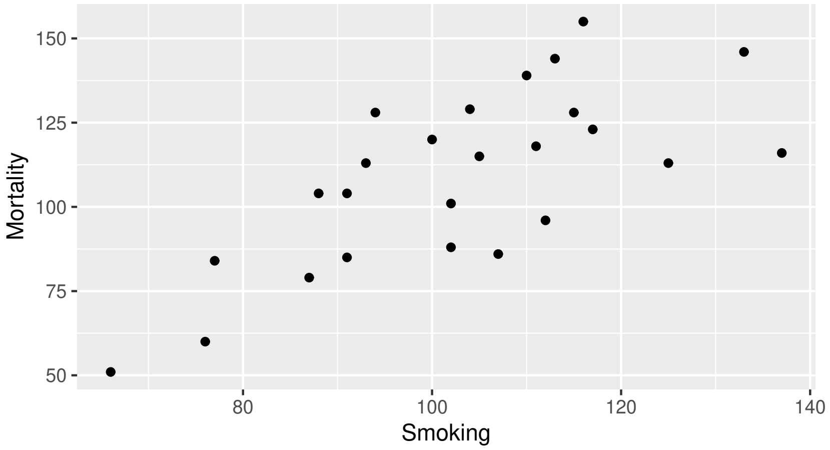

Figure 1.1: Smoking and mortality indices for 25 occupational groups.

In the data, we see that some groups have identical smoking indices, but different mortality indices. Hence given a group's smoking index, we are not able to say with certainty what the group's mortality index will be.

For group i let x_i denote the group's smoking index and y_i denote the group's mortality index. We can then consider a statistical model of the form

$$y_i = f(x_i) + \varepsilon_i,$$

where f is a (deterministic) function of the smoking index x_i. We typically choose f to be some fairly simple function of x such as $f(x) = \beta_0 + \beta_1 x$ or $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. The term ε_i is a **random** "error". Two reasons why an error term would be necessary are as follows:

1. The variable x_i is not sufficient for predicting y_i with certainty. The variable y_i may also depend on other variables which are unknown to us, or there may simply be variation in the population that we are not able to explain with the function f.

2. We are not able to observe the quantity we are interested in (e.g. mortality index) with absolute precision. We observe instead the sum of the true value and a random "measurement error" ε_i.

We further assume that $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. (independent and identically distributed) with some probability distribution, so that in our model, two groups with the same smoking index may have different mortality indices ($x_i = x_j \not\Rightarrow y_i = y_j$).

1.2.2 Notation and terminology

Continuing the example, suppose we choose $f(x) = \beta_0 + \beta_1 x$, so that we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

We call the x-variable (smoking index) the **independent** or **predictor** variable. We will always treat the independent variable as a known constant. The y-variable (mortality index) is called the **dependent** variable. The dependent variable is treated as a random variable, as it is expressed as a function of the independent variable plus a random error term. If we only know group i's smoking index and not their mortality index, i.e. we only know the value of x_i, then we denote Y_i as the unknown (random) value of the dependent variable, so that y_i is the observed value of the dependent variable. The dependent variable will always be a scalar, but we will sometimes consider vector independent variables (e.g. smoking index and mean income). We write bold $\boldsymbol{x}$ to denote a vector independent variable.

The **parameters** in this model are β_0, β_1 and σ^2. These model parameters are treated as unknown constants. Note that since β_0 and β_1 are unknown, given x_i and y_i we will still not know the true value of ε_i.

In this course we will always make the same assumption that $\varepsilon_i \sim N(0, \sigma^2)$, with $\varepsilon_1, \varepsilon_2, \dots$ independent. For any dataset we will always need to test if this assumption is reasonable.

1.2.3 Course objectives

By the end of this course you will be able to:

1. Choose an appropriate statistical model to describe the observed data $\{(y_1, \boldsymbol{x}_1), \dots, (y_n, \boldsymbol{x}_n)\}$.

2. Estimate the values of the unknown parameters in your model.

3. Interpret your model to describe the relationship between y and $\boldsymbol{x}$.

4. Use your model to make predictions about the value of y given the value of $\boldsymbol{x}$ only.

2 Distributions and inference

For a good understanding of statistical modelling, you will need to be familiar with the basic ideas and results of the probability theory: random variables, probability distributions and statistical inference. In this section we will briefly survey the necessary knowledge.

2.1 Random variables and probability distributions

Symbolically it is convenient to denote a random variable by an upper-case letter, say Y, and any value observed for it by a lower-case version of the same latter, subscripted if necessary to distinguish separate values. Thus y_1 will denote the value of the first observation of Y, y_2 – the value of the second observation of Y, and so on. A sample of n values will be written $y_1, y_2, \ldots, y_n$ or $\{y_i\}_{i=1\ldots n}$.

2.1.1 Single random variable

A complete description of a random variable is given by specifying all the values it could conceivably take, a **population**, and quoting their their associated probabilities. Such a specification is known as the **probability distribution** of the variable. For a discrete random variable, which takes countably many values, a probability distribution can be specified explicitly by listing all the possible values and their probabilities. This is provided by a **probability mass function**

$$f(y) = \Pr(Y = y)$$

stating the probability of a discrete random variable Y obtaining value y and satisfying the following properties:

(i) $f(y) \geq 0$ for all possible y.

(ii) $\sum_{y \in P} f(y) = 1$, where the sum is over all possible y.

Another important quantity is the **cumulative distribution function** of a random variable. This function is defined by

$$F(y) = \Pr(Y \leq y) = \sum_{x \leq y,\, x \in P} f(x).$$

It is evident from this definition that

(i) $F(-\infty) = 0$.

(ii) $F(\infty) = 1$.

(iii) If $a \leq b$, then $F(a) \leq F(b)$, i.e. F is a non-decreasing function.

(iv) The probability that Y lies in any interval $\Pr(a < Y \leq b) = F(b) - F(a)$.

We will often be interested in the following properties of a discrete random variable Y:

(i) The r-th **central moment** of Y,

$$\mathbb{E}(Y^r) = \sum_{y \in P} y^r f(y).$$

The 1-st central moment $\mathbb{E}(Y)$ is called the **expected value** or **population mean**, and is denoted by μ.

(ii) The r-th **moment about the mean** of Y

$$\mathbb{E}((Y-\mu)^r) = \sum_{y \in P} (y-\mu)^r f(y).$$

The 2-nd moment about the mean is called the **variance** of Y, and is denoted by σ^2. It is easy to show that

$$\sigma^2 = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2.$$

The square root of σ^2 is called the **standard deviation**.

For a continuous random variable Y the probability mass function is replaced by the **probability density function** specified by

$$f(y) = \frac{d}{dy} F(y),$$

and the sums above should be replaced with the integrals

$$\int_P y^r f(y) dy \quad \text{and} \quad \int_P (y-\mu)^r f(y) dy,$$

respectively.

2.1.2 Two random variables

Now let Y_1 and Y_2 be two discrete random variables taking values in the same population P. Then the function

$$f(y_1, y_2) = \Pr(Y_1 = y_1, Y_2 = y_2)$$

is called **joint probability mass function**. It must satisfy

$$\sum_{y_1, y_2 \in P} f(y_1, y_2) = 1,$$

where the summation is over all possible y_1 and y_2. Using this function we can compute probability mass functions of the individual random variables. Observe that

$$\Pr(Y_1 = y_1) = \sum_{y_2 \in P} \Pr(Y_1 = y_1, Y_2 = y_2).$$

Hence the probability mass function of Y_1 is

$$f_{Y_1}(y_1) = \sum_{y_2 \in P} f(y_1, y_2).$$

This is called the **marginal probability mass function** of Y_1. An analogous statement holds for the marginal probability mass function $f_{Y_2}(y_2)$ of Y_2. The random variables Y_1 and Y_2 are **independent** if

$$f(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$$

at all points y_1 and y_2.

We define expectation and moments of the joint distribution in the natural way. For any function $g(Y_1, Y_2)$ we define

$$\mathbb{E}(g(Y_1, Y_2)) = \sum_{y_1, y_2 \in P} g(y_1, y_2) f(y_1, y_2).$$

In particular, the expectation value of the product $Y_1 Y_2$ is

$$\mathbb{E}(Y_1 Y_2) = \sum_{y_1, y_2 \in P} y_1 y_2 f(y_1, y_2).$$

The **covariance** between Y_1 and Y_2 is

$$\mathrm{Cov}(Y_1, Y_2) = \mathbb{E}(Y_1 Y_2) - \mathbb{E}(Y_1)\, \mathbb{E}(Y_2).$$

The random variables Y_1 and Y_2 are **independent** if $\mathrm{Cov}(Y_1, Y_2) = 0$. In this case $\mathbb{E}(Y_1 Y_2) = \mathbb{E}(Y_1)\, \mathbb{E}(Y_2)$.

For continuous random variables Y_1 and Y_2 the joint probability density function is specified by

$$f(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} F(y_1, y_2),$$

and the sum $\sum_{y_2 \in P}$ above should be replaced with the integral $\int_P dy_2$, while the double sum $\sum_{y_1, y_2 \in P}$ should be replaced with the double integral $\int\int_P dy_1 dy_2$.

2.2 Statistics of a sample

We generally deal with an ensemble of n observations, $y_1, y_2, \ldots, y_n$, known as a **sample**. If the data $y_1, y_2, \ldots, y_n$ are regarded as the observed values of random variables $Y_1, Y_2, \ldots, Y_n$, then it follows that the sample and any statistics derived from it might be different from those of the underlying population. However, although we would expect variation over possible sets of data, we would also expect to see systematic patterns induced by the underlying population.

Sample moments are calculated by putting mass $1/n$ on each of the observations. The most commonly used moments of a sample are the **sample average** and the **sample variance**:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \text{and} \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{2.1}$$

You may wonder why the denominators for $\bar{y}$ and s^2 are different. Let us explain why.

Suppose that $Y_1, \ldots, Y_n$ is a sample of independent identically distributed random variables with mean μ and variance σ^2. Then the average $\bar{Y}$ has the expectation

$$\mathbb{E}(\bar{Y}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{n}{n} \mu = \mu$$

and variance

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right)^2\right) = \frac{1}{n^2}\,\mathbb{E}\left(\left(\sum_{i=1}^{n}(Y_i-\mu)\right)^2\right)\\ &= \frac{1}{n^2}\,\mathbb{E}\left(\sum_{i=1}^{n}(Y_i-\mu)^2\right) + \frac{1}{n^2}\,\mathbb{E}\left(\sum_{i,j=1,\, i\neq j}^{n}(Y_i-\mu)(Y_j-\mu)\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{\sigma^2}{n},\end{aligned}$$

We see that the variance of $\bar{Y}$ does not quite represent the true variance of population.

Now let us compute the expectation of the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2.$$

Note that

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}\big(Y_i-\mu-(\bar{Y}-\mu)\big)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2 - 2\sum_{i=1}^{n}(Y_i-\mu)(\bar{Y}-\mu) + \sum_{i=1}^{n}(\bar{Y}-\mu)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2 - 2n(\bar{Y}-\mu)^2 + n(\bar{Y}-\mu)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2 - n(\bar{Y}-\mu)^2.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}((Y_i-\mu)^2) - n\,\mathbb{E}((\bar{Y}-\mu)^2)\right)\\ &= \frac{1}{n-1}(n\sigma^2 - n\sigma^2/n)\\ &= \sigma^2.\end{aligned}$$

This explains the use of the denominator $n-1$ when defining the sample variance s^2 in (2.1).

2.3 Inferring from a sample

We are usually concerned with making statements about the underlying population by inferring statistical data from a sample. Suppose we wish to make an “informed guess” at the value of a population parameter θ. Since we have to use only sample data, and we need to produce a single number as the value of θ, it is natural to use some statistic T to deliver this value. Such a statistic is called an **estimator** of θ, whilst the value t of the statistic observed for a particular sample is called an **estimate** of θ. The estimator is called **unbiased** if its estimate coincides with its expectation value. There are three popular methods of estimation: the method of **moments**, **maximal likelihood**, and **least squares**.

2.3.1 The method of moments

The method of moments chooses the estimate of any general unknown parameter to be that value which makes the population mean equal to the sample mean. For any particular probability model model, the required estimator can be derived as follows. First, the population mean is found in terms of the unknown parameter. Next, this expression is set equal to the sample mean. Finally, the resulting equation is solved for the unknown parameter. We illustrate this with an example.

Example 2.1. Suppose we have a sample $y_1, y_2, \ldots, y_n$ drawn from a distribution with probability density function $f(y) = \lambda e^{-\lambda y}$ for $0 \leq y < \infty$. The mean of such a distribution is

$$\mu = \int_0^\infty y f(y) dy = \int_0^\infty y \lambda e^{-\lambda y} dy = \frac{1}{\lambda}.$$

Thus if $\bar{y} = \frac{1}{n}\sum_{i-1}^n y_i$ is the mean of the $y_1, y_2, \ldots, y_n$, we set $\bar{y} = 1/\lambda$ and solve for λ, to obtain $\hat{\lambda} = 1/\bar{y}$ as the method of moments estimate of λ. We put a hat above λ to distinguish the estimate $\hat{\lambda}$ from the parameter λ in question. The estimator of λ is $1/\bar{Y}$. □

If there are several unknown parameters in the model, we need more than one equation to solve in this case. For two unknown parameters we need to use the first two moments to obtain the estimator, i.e., we set $\mathbb{E}(Y)$ equal to $\frac{1}{n}\sum_{i=1}^n y_i$ and $\mathbb{E}(Y^2)$ equal to $\frac{1}{n}\sum_{i=1}^n y_i^2$ and solve the resulting pair of simultaneous equations. If there are three unknown parameters, we need to use first three moments in this way, and so on.

2.3.2 The method of maximal likelihood

Let $y_1, y_2, \ldots, y_n$ be a sample from a distribution with probability density function $f(y;\theta)$ that depends on an unknown parameter θ. The **likelihood** of the sample is the joint probability density function $f(y_1, y_2, \ldots, y_n; \theta)$ of the sample, treated as a function of θ. Writing the likelihood as $L(\theta)$, and noting that in a random sample the individuals are independent of each other (so that the joint probability density is a product of individual densities), we have

$$L(\theta) = \prod_{i=1}^n f(y_i;\theta).$$

The function $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$ that maximizes $L(\theta)$ with respect to θ is the **maximum likelihood estimator** of θ. Its actual value for a given sample is the **maximum likelihood estimate** (m.l.e.) of θ for that sample. Roughly speaking, the m.l.e. can be interpreted as the value of θ that ascribes the highest possible probability of the sample that was actually obtained, which is an intuitive justification for its use. Note also that the value maximizing $L(\theta)$ maximizes the natural logarithm of $L(\theta)$ as well, and the latter maximization is often easier to effect in practice.

Writing $l(\theta) = \log(L(\theta))$ the m.l.e. is a root of the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

at which

$$\frac{\partial^2 l(\theta)}{\partial^2 \theta} < 0.$$

If more than one such root exists, we take the root at which the value of $L(\theta)$ is the largest.

Example 2.2. Consider again a sample of n observations $y_1, y_2, \ldots, y_n$ each drawn from a distribution with probability density function $f(y) = \lambda e^{-\lambda y}$ for $0 \leq y < \infty$. Then

$$L(\lambda) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}.$$

Hence

$$l(\lambda) = \log(L(\lambda)) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

implying

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

Requiring the r.h.s. equal to zero and solving for λ yields

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\bar{y}}.$$

Moreover,

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n}{\lambda^2},$$

which must be negative at the estimated value $\hat{\lambda}$. This is indeed true since $\lambda^2 > 0$. The maximal likelihood estimate $\hat{\lambda}$ is thus $1/\bar{y}$, which agrees with what we found using the method of moments. □

Example 2.3. Suppose that $y_1, y_2, \ldots, y_n$ is a sample drawn from a distribution with mean μ and variance σ^2, two parameters we want to estimate. Then

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2}$$

so that

$$\begin{aligned}
l(\mu, \sigma^2) &= \log(L(\mu, \sigma^2)) \\
&= \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \\
&= -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial l(\mu, \sigma)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu), \\
\frac{\partial l(\mu, \sigma)}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2.
\end{aligned}$$

Equating the first equation to zero and solving for μ yields the m.l.e. $\hat{\mu}$:

$$\sum_{i=1}^{n}(y_i-\hat{\mu})=0 \qquad \Longrightarrow \qquad \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i=\bar{y}. \tag{2.2}$$

Equating the second equation to zero and solving for σ yields the estimate $\hat{\sigma}$:

$$\frac{n}{\hat{\sigma}}=\frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(y_i-\mu)^2 \qquad \Longrightarrow \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu})^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Note that the m.l.e. $\hat{\sigma}^2$ has the divisor n, as opposite to the divisor $n-1$ in the definition of the sample variance.

It only remains to verify that the log-likelihood $l(\mu,\sigma)$ indeed attains a maximum at $\mu=\hat{\mu}$ and $\sigma=\hat{\sigma}$. We need to calculate second derivatives of $l(\mu,\sigma)$ and evaluate the Hessian determinant at $\mu=\hat{\mu}$ and $\sigma=\hat{\sigma}$,

$$H=\begin{vmatrix}\dfrac{\partial^2 l(\mu,\sigma)}{\partial\mu^2} & \dfrac{\partial^2 l(\mu,\sigma)}{\partial\mu\partial\sigma}\\ \dfrac{\partial^2 l(\mu,\sigma)}{\partial\mu\partial\sigma} & \dfrac{\partial^2 l(\mu,\sigma)}{\partial\sigma^2}\end{vmatrix}.$$

We calculate

$$\frac{\partial^2 l(\mu,\sigma)}{\partial\mu^2}=\frac{\partial}{\partial\mu}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\right)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(-1)=-\frac{n}{\sigma^2}\overset{\mu=\hat{\mu},\,\sigma=\hat{\sigma}}{=}-\frac{1}{\hat{\sigma}^2}$$

and

$$\frac{\partial^2 l(\mu,\sigma)}{\partial\mu\partial\sigma}=\frac{\partial}{\partial\sigma}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\right)=-\frac{2}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)\overset{\mu=\hat{\mu},\,\sigma=\hat{\sigma}}{=}0,$$

where the last equality follows from (2.2), i.e., after evaluating $\mu=\hat{\mu}$ and $\sigma=\hat{\sigma}$. In a similar way we find

$$\begin{aligned}\frac{\partial^2 l(\mu,\sigma)}{\partial\sigma^2}&=\frac{\partial}{\partial\sigma}\left(-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2\right)\\&=\frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2\\&=\frac{n}{\sigma^2}-\frac{3}{\sigma}\left(\frac{n}{\sigma}-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2\right)\overset{\mu=\hat{\mu},\,\sigma=\hat{\sigma}}{=}\frac{n}{\hat{\sigma}^2}-\frac{3n}{\hat{\sigma}^2}=-\frac{2n}{\hat{\sigma}^2}.\end{aligned}$$

Hence

$$H=\begin{vmatrix}-\dfrac{n}{\sigma^2} & 0\\ 0 & -\dfrac{2n}{\sigma^2}\end{vmatrix}=\frac{2n^2}{\sigma^4}>0, \qquad \frac{\partial^2 l(\mu,\sigma)}{\partial\mu^2}<0, \qquad \frac{\partial^2 l(\mu,\sigma)}{\partial\sigma^2}<0$$

and thus $l(\mu,\sigma)$ indeed attains a maximum at $\mu=\hat{\mu}$ and $\sigma=\hat{\sigma}$. We conclude that

$$\hat{\mu}=\bar{y}, \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

are the maximal likelihood estimates of μ and σ, respectively. □

2.3.3 The method of least squares

Let $Y_1, Y_2, \ldots, Y_n$ be such that

$$Y_i = g(\beta_1, \beta_2, \ldots, \beta_k) + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, where $g(\beta_1, \beta_2, \ldots, \beta_k)$ is a constant function of k scalar parameters β_i and the ε_i are independent and identically distributed random variables with zero mean and a common variance σ^2. Let $y_1, y_2, \ldots, y_n$ be observations of $Y_1, Y_2, \ldots, Y_n$. The **least squares estimates** (l.s.e.) of the parameters β_i are the values $\hat{\beta}_i$ which minimize

$$V = \sum_{i=1}^{n} (y_i - g(\beta_1, \beta_2, \ldots, \beta_k))^2.$$

A standard calculus can be employed to find these values. The least squares estimators coincide with with the maximal likelihood ones if the distribution of the ε_i is normal, but not necessarily otherwise.

Example 2.4. Consider again n independent normal random variables Y_1, Y_2, $\ldots$, Y_n. We may write them as $Y_i = \mu + \varepsilon_i$, where ε_i are normal random variables with zero mean, $\varepsilon_i \sim N(0, \sigma^2)$. We may thus say that ε_i measure the departure of Y_i from the population mean μ. Our goal is to find the l.s.e. of μ. We have found above that the log-likelihood is

$$l(\mu, \sigma^2) = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Then, regarding σ^2 as an irrelevant constant, maximizing the log-likelihood with respect to μ is the same as minimizing $\sum_{i=1}^{n} (y_i - \mu)^2$ with respect to μ. In other words, the l.s.e. of μ is the value which minimizes the sum of squared departures between the same values and the parameter, $\hat{\mu} = \bar{y}$. □

2.4 Some distribution theory relating to the normal distribution

The normal distribution is the most commonly used distribution in statistics. It is very conveniently parametrized in terms of its mean and its variance, and these two moments completely determine the whole distribution. If Y has a normal distribution with mean μ and variance σ^2, then it is defined over the whole real line and its probability density function is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \quad \text{for} \quad -\infty < y < \infty.$$

To standardize a normal variable Y, we subtract its mean and divide by its standard deviation. This converts $Y \sim N(\mu, \sigma^2)$ into the standardised variable $Z = (Y - \mu)/\sigma \sim N(0, 1)$. The standard normal probability density function $\varphi(y)$ and its distribution function $\Phi(y)$ are

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}, \quad \Phi(y) = \int_{-\infty}^{y} \varphi(x) dx \quad \text{for} \quad -\infty < y < \infty.$$

Let $Y_1, Y_2, \ldots, Y_n$ denote normally distributed random variables with $Y_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ and let the covariance of Y_i and Y_j be denoted by $\mathrm{Cov}(Y_i, Y_j) = \sigma_{ij}^2$. Then the joint distribution of the Y's is the **multivariate normal distribution** $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ and variance-covariance matrix V with elements σ_{ij}^2. We write this as

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, V).$$

Now suppose that the random variable Z is a linear combination of the Y_i's,

$$Z = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n,$$

where a_i's are constants. Then the mean and variance of Z are

$$\mathbb{E}(Z) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n,$$
$$\text{Var}(Z) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

Furthermore, Z is normally distributed,

$$Z = \sum_{i=1}^{n} a_iY_i \sim N\Bigg(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\Bigg). \tag{2.3}$$

There are three distributions which can be derived from the normal distribution and which occur very frequently in all branches of statistics. Their definitions are fairly straightforward and their cumulative distribution functions are extensively tabulated, but other aspects such as formulae for probability density functions and moments are of relatively minor importance. We therefore simply give their definitions here, and introduce their uses as necessary in succeeding sections.

Definition 2.5.

(i) *If Y_i for $i = 1, \dots, n$ are mutually independent standard normal random variables, $Y_i \sim N(0,1)$, then the sum of their squares is distributed according to the* **chi-squared** *distribution with n degrees of freedom. This is usually denoted as*

$$Z = Y_1^2 + Y_2^2 + \ldots + Y_n^2 \sim \chi_n^2.$$

(ii) *If $Y \sim N(0,1)$, then $Y^2 \sim \chi_1^2$.*

(iii) *If $Y \sim N(0,1)$ and $Z \sim \chi_n^2$, and Y is independent of Z, then*

$$W = \frac{Y}{\sqrt{Z/n}} \sim t_n, \tag{2.4}$$

i.e. the random variable W is said to have the Student's t-distribution with n degrees of freedom.

(iv) *If $Z \sim \chi_p^2$ and $V \sim \chi_q^2$, and Z is independent of V, then*

$$W = \frac{Z/p}{V/q} \sim F_{p,q}, \tag{2.5}$$

i.e. the random variable W is said to have the F-distribution with p and q degrees of freedom.

We also recall the Central Limit Theorem.

Theorem 2.6. *Let $\{X_1, X_2, \dots, X_n\}$ be a sequence of i.i.d. random variables with mean μ and finite variance σ^2 and let $S_n = \sum_{i=1}^n X_i/n$. Then as n approaches infinity, the random variable $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$, that is*

$$\lim_{n\to\infty} \sqrt{n}(S_n - \mu) \sim N(0, \sigma^2).$$

2.5 Covariance matrix and some matrix algebra

We recall some basis facts that will be useful in further sections. Let $\boldsymbol{z}$ be a random $r \times 1$ vector and let $\boldsymbol{w}$ be a random $p \times 1$ vector. Then:

1. The cross covariance of $\boldsymbol{z}$ and $\boldsymbol{w}$ is an $r \times p$ matrix is defined by

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{z},\boldsymbol{w}) &= \mathbb{E}\big((\boldsymbol{z}-\mathbb{E}(\boldsymbol{z}))(\boldsymbol{w}-\mathbb{E}(\boldsymbol{w}))^T\big)\\ &= \mathbb{E}\big(\boldsymbol{z}\boldsymbol{w}^T - \boldsymbol{z}\,\mathbb{E}(\boldsymbol{w})^T - \mathbb{E}(\boldsymbol{z})\boldsymbol{w}^T + \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T\big)\\ &= \mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) - \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T - \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T + \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T\\ &= \mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) - \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T.\end{aligned}$$

2. If $\boldsymbol{z}$ and $\boldsymbol{w}$ are independent variables, $\mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) = \mathbb{E}(\boldsymbol{z})\mathbb{E}(\boldsymbol{w})^T$, their covariance is zero.

3. The diagonal entries of the matrix $\operatorname{Cov}(\boldsymbol{z}) = \operatorname{Cov}(\boldsymbol{z},\boldsymbol{z})$ are the variances of each element of the vector $\boldsymbol{z}$. We will denote the diagonal matrix of variances of $\boldsymbol{z}$ by $\operatorname{Var}(\boldsymbol{z})$.

4. Assume that $\boldsymbol{z} \sim N_r(\boldsymbol{\mu}, V)$, where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z})$ and V is the $r \times r$ variance-covariance matrix. Let $\boldsymbol{c} = \boldsymbol{a} + B\boldsymbol{z}$ for any $p \times 1$ vector $\boldsymbol{a}$ and any $p \times r$ matrix B. Then

$$\boldsymbol{c} \sim N_p\big(\boldsymbol{a} + B\boldsymbol{\mu}, BVB^T\big). \tag{2.6}$$

5. Let M be an $r \times p$ matrix and let N be a $p \times r$ matrix. Then

$$(MN)^T = N^T M^T.$$

6. The trace of an $r \times r$ square matrix M, written $\operatorname{tr}(M)$ is defined as the sum of its diagonal elements. If S is an $r \times p$ matrix, then

$$\operatorname{tr}(S^T M S) = \operatorname{tr}(M S S^T). \tag{2.7}$$

7. The diagonal part of a matrix M is denoted by $\operatorname{diag}(M)$.

Example 2.7.

1. Let $Y_1 \sim N(1,2)$ and $Y_2 \sim N(2,3)$ be independent random variables. What is the distribution of $W_1 = Y_1 - Y_2$ and $W_2 = 2Y_1 + 3Y_2$?

   Using (2.3) we find

$$W_1 = 1 \cdot Y_1 + (-1) \cdot Y_2 \sim N(1 \cdot 1 + (-1) \cdot 2,\ 1^2 \cdot 2 + (-1)^2 \cdot 3) = N(-1,5)$$

   and

$$W_2 = 2 \cdot Y_1 + 3 \cdot Y_2 \sim N(2 \cdot 1 + 3 \cdot 2,\ 2^2 \cdot 2 + 3^2 \cdot 3) = N(8,35)$$

2. Let $Y_1 \sim N(0,1)$ and $Y_2 \sim N(2,4)$ be independent random variables. What is the distribution of $W = Y_1^2 + (Y_2-2)^2/4$?

   First, notice that $(Y_2-2)^2/4 \sim N(0,1)$. Then using Definition 2.5 (i) we find

$$W = Y_1^2 + (Y_2-2)^2/4 \sim \chi^2_2.$$

3. Let $Y_1 \sim N(1,2)$ and $Y_2 \sim N(0,1)$ be independent random variables. Set $\boldsymbol{Y} = \begin{pmatrix}(Y_1-1)/2\\ Y_2\end{pmatrix}$. What is the distribution of $\boldsymbol{Y}^T\boldsymbol{Y}$?

   Since $(Y_1-1)/2 \sim N(0,1)$, we have that $\boldsymbol{Y}^T\boldsymbol{Y} = ((Y_1-1)/2)^2 + Y_2^2 \sim \chi_2^2$. □

Example 2.8. Suppose that $\boldsymbol{y} \sim N_3(\boldsymbol{\mu}, V)$, a random 3×1 vector, i.e., $\boldsymbol{y} = (y_1, y_2, y_3)^T$, described by a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix V:

$$\boldsymbol{\mu} = \begin{pmatrix}1\\-1\\2\end{pmatrix}, \qquad V = \begin{pmatrix}2 & -1 & 0\\ -1 & 3 & 1\\ 0 & 1 & 4\end{pmatrix}.$$

1. Find the individuals distributions of y_1, y_2 and y_3.

   The answer is
   $$y_1 \sim N(1,2), \quad y_2 \sim N(-1,3), \quad y_3 \sim N(2,4).$$

2. Find the distribution of $z = y_1 - 2y_2 + y_3$.

   We need to use formula (2.6) (with $\boldsymbol{a} = 0$). Notice that z can be written as $z = B\boldsymbol{y}$ with $B = (1\ {-2}\ 1)$, namely
   $$z = (1\ {-2}\ 1)\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix} = y_1 - 2y_2 + y_3.$$

   Then
   $$B\boldsymbol{\mu} = (1\ {-2}\ 1)\begin{pmatrix}1\\-1\\2\end{pmatrix} = 1+2+2 = 5$$

   and
   $$BVB^T = (1\ {-2}\ 1)\begin{pmatrix}2 & -1 & 0\\ -1 & 3 & 1\\ 0 & 1 & 4\end{pmatrix}\begin{pmatrix}1\\-2\\1\end{pmatrix} = 18.$$

   giving
   $$z \sim N_3\left(B\boldsymbol{\mu}, BVB^T\right) = N_3\left(5, 18\right).$$

3. Find the joint distribution of y_1 and y_2, i.e. the multivariate normal dist. of $\begin{pmatrix}y_1\\y_2\end{pmatrix}$.

   To find the multivariate normal dist. of $\begin{pmatrix}y_1\\y_2\end{pmatrix}$ we simply restrict to the first and second rows and columns of $N_3(\boldsymbol{\mu}, V)$:
   $$\begin{pmatrix}y_1\\y_2\end{pmatrix} \sim N_2\left(\begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}2 & -1\\ -1 & 3\end{pmatrix}\right).$$

4. Find the joint distribution of y_1 and $\frac{1}{2}(y_2+y_3)$.

   Write $\boldsymbol{z} = \begin{pmatrix} y_1 \\ \frac{1}{2}(y_2+y_3) \end{pmatrix}$ and notice that $\boldsymbol{z} = B\boldsymbol{y}$ with $B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}$. Then since

   $$B\boldsymbol{\mu} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$$

   and

   $$BVB^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 2 & -1/2 \\ -1/2 & 9/4 \end{pmatrix}$$

   giving

   $$\boldsymbol{z} \sim N_2(B\boldsymbol{\mu}, BVB^T) = N_2\left(\begin{pmatrix} 1 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 2 & -1/2 \\ -1/2 & 9/4 \end{pmatrix} \right).$$

5. Which pairs of random variables are independent, i.e., are y_1 and y_2 independent? are y_1 and y_3 independent? and so on.

   Recall that random variables y_i and y_j are independent if $\text{Cov}(y_i, y_j) = 0$. Thus the only pair of independent random variables is (y_1, y_3), since $V_{13} = V_{31} = 0$. □

3 Textbooks and online resources

The recommended textbooks for this course:

- Douglas C. Montgomery, Elizabeth A. Peck, Geoffrey G. Vining, *Introduction to Linear Regression Analysis*, 5th Edition, Wiley (2012). [Sections 1–3 (complete) and Sections 4, 7–9 (partially)].
- Simon J. Sheather, *A Modern Approach to Regression with R*, Springer (2009). [Sections 1, 2, 5 (complete), 3, 6, 7 (partially)].
- Douglas C. Montgomery, *Design and Analysis of Experiments*, 9th Edition, Wiley (2017). [Sections 3, 5, 13, 14 (partially)].

Other suggested textbooks:

- Krzanowski, W. J. *An Introduction to Statistical Modelling*. Wiley (2002). [Sections 1–4].
- Nicolas H. Bingham, John M. Fry, *Regression: Linear Models in Statistics*, Springer (2010). [Sections 2.6–2.8 and 3.1–3.4].
- Sanford Weisberg, *Applied Linear Regression*, 4th Edition, Wiley (2014). [Sections 1–3].
- Michael H. Kutner, Christopher J. Nachtsheim, John Neter, *Applied Linear Statistical Models*, 4th Edition, McGraw-Hill Irwin (2005).

Recommended free on-line course and e-books:

- DataCamp.com course *Introduction to R*,

  https://www.datacamp.com/courses/free-introduction-to-r

  Please use your university email (@herts.ac.uk) to register to DataCamp. The registration link will be given once the module starts.

- ReliaSoft's Experiment Design and Analysis Reference [Sections 2-6],

  http://reliawiki.org/index.php/Experiment_Design_and_Analysis_Reference

- Grolemund, G. and Wickham, H., *R for Data Science*, https://r4ds.had.co.nz