Statistics Revision Notes for MSc Students

These notes are written for MSc students who wish to revise essential statistics needed for MSc Statistics courses. These notes will cover some of the topics given under the ”Statistics” heading on the Mathematical and Statistics Skills Website. Please be aware that you need to have mastered the relevant parts of the other sections on that website also.

Contents

1 Random Variables and Distributions

1.1 Expectation

The expectation of a random variable (rv) informs us about the location (i.e. mean) of the random variable.

1. If X is a discrete rv taking values $x_1, x_2, \ldots, x_n$ (n possibly infinite) with probability mass function (pmf) f, then
$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i f(x_i).$$
2. If Y is a continuous rv with probability density function (pdf) f, then
$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f(y) dy.$$

Note: the expected value of a rv, X may NOT in fact be one of the possible values of X.

Expectation of a function $g(X)$

Let X be a rv. Then, if we let $Z = g(X)$, Z is also a rv and we have the result that,

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}(g(X)) \\ &= \begin{cases} \sum_x g(x) f(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & X \text{ continuous} \end{cases} \end{aligned}$$

Expectation of XY

Let X and Y be random variables.

$$\mathbb{E}(XY) = \begin{cases} \sum_x \sum_y xy f(x,y) & X \text{ and } Y \text{ discrete} \\ \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xy f(x,y) & X \text{ and } Y \text{ continuous} \end{cases}$$

Properties of Expected Values

1. For any rv X and constants a and b,
$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b;$$
2. For any two rvs X and Y,
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y);$$
(Note this can be extended to n rvs $X_1, \ldots, X_n$ such that $\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i)$).
3. If two rvs X and Y are INDEPENDENT, then
$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

1.2 Variance

The variance of a rv informs us about the variability (or spread) of values the random variable may take.

Let X be a rv with finite mean $\mu \equiv \mathbb{E}(X)$. Then, the variance of X, written Var(X), is defined to be,

$$\text{Var}(X) = \mathbb{E}((X - \mu)^2) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Notes:

- By definition we have that Var$(X) \geq 0$.
- Often, it is easier to calculate Var(X) using,

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

- The standard deviation of a rv X is simply $\sqrt{\text{Var}(X)}$.
- We often denote Var(X) by σ^2, and the standard deviation by σ.
- σ^2 measures the spread of the distribution around its mean (small variance $\Rightarrow$ distribution is highly concentrated around μ; large variance $\Rightarrow$ widespread distribution around μ)

Properties of Variance

1. For any rv X and constants a and b,
$$\text{Var}(aX + b) = a^2\text{Var}(X).$$
2. If the two rvs X and Y are INDEPENDENT then,
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$
This can be extended - if $X_1, \ldots, X_n$ are independent rvs, then,
$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).$$

1.3 Covariance and Correlation

Covariance (and correlation) informs us about the relationship between random variables.

Covariance

The covariance of rvs X and Y is defined to be:

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

Properties of Covariance

1. Cov(X, X) = Var(X);
2. Cov(X, Y) = $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$;
3. If X and Y are independent, Cov(X, Y) = 0 (the converse is not necessarily true);
4. Var$(X + Y)$ = Var(X) + Var(Y) + 2 Cov(X, Y);
(Note: if X and Y are independent, Cov(X, Y) = 0, so that Var$(X + Y)$ = Var(X) + Var(Y)).
5. Cov(aX, bY) = ab Cov(X, Y) for constants a and b.

Correlation

The correlation between the two rvs X and Y, written $\rho(X, Y)$ is given by,

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Notes:

1. Correlation is generally preferred to covariance as it is invariant to the unit of measure;
2. $-1 \leq \rho(X, Y) \leq 1$;
3. If $\rho(X, Y) > 0$ then X and Y are positively correlated;

   If $\rho(X, Y) < 0$ then X and Y are negatively correlated;

   If $\rho(X, Y) = 0$ then X and Y are uncorrelated.

1.4 Normal Distribution

Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The rv Y has a normal distribution with parameters μ and σ^2 and we write $Y \sim N(\mu, \sigma^2)$, if,

1. Y can take any value in $\mathbb{R}$; and
2. Y has pdf,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad -\infty < y < \infty.$$

Notes:

- The parameter μ is the "mean" and σ^2 the "variance" (σ is referred to as the "standard deviation"). The mean describes the location of the distribution and the variance/standard deviation the spread of the distribution.
- The distribution is symmetrical about μ.
- This distribution is also often referred to as a Gaussian distribution or "bell-shaped" distribution.
- The integration of $f(y)$ over a finite interval does not exist in closed form. However computer software can be used to evaluate the probability that a normal rv lies in a given interval.

Special case: $N(0, 1)$

When $\mu = 0$ and $\sigma^2 = 1$, then $Y \sim N(0, 1)$ and this is called the *standard Normal* distribution.

Notes:

- The cumulative distribution function for the standard normal distribution is denoted by Φ. Thus, for $Z \sim N(0, 1)$, $\Phi(z) = \mathbb{P}(Z \leq z)$.
- Tables exist for the standard normal distribution to evaluate $\Phi(z)$ for different values of z - although we can use the computer package R to evaluate such probabilities.

Linear Transformations

If $Y \sim N(\mu, \sigma^2)$, then setting $Z = \frac{Y-a}{b}$ we have that,

$$Z \sim N\left(\frac{\mu - a}{b}, \frac{\sigma^2}{b^2}\right).$$

Letting $a = \mu$ and $b = \sigma$ we have, $Z = \frac{Y-\mu}{\sigma}$ and,

$$Z \sim N(0, 1),$$

i.e. the standard Normal distribution.
This is a very interesting property - it does not hold for other distributions in general!

1.5 Poisson Distribution

Let λ be a real number with $\lambda > 0$. The rv X has a Poisson distribution with parameter λ and we write $X \sim Po(\lambda)$, if,

1. X can take values $0, 1, 2, \ldots$; and
2. X has pmf,

$$f(x) = \begin{cases} \frac{\exp(-\lambda)\lambda^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Note:

- The Poisson distribution is often assumed when considering the number of events that occur within a fixed interval of time (e.g. number of phone call received by a helpline within 24 hours).

2 Estimation

In probability theory we are typically given (or derive) the distribution of a random variable and we wish to calculate certain properties of the random variable. For example, the probability that a discrete random variable takes a particular value, or the probability that it lies in a given region. We now consider statistical inference - we observe a sample set of observations from the distribution, from which we wish to make inference about the population from which they were drawn.

For example, we toss a biased coin n times and observe x heads from which we wish to estimate the probability of obtaining a head in a single toss.

An estimate of the population parameter on its own is difficult to assess. If someone told us that they estimate that 75% of voters in Edinburgh South intend to vote for the SNP in the next election, we would need further information before deciding if the estimate was reliable. In formal terms, we need to know how the estimate was found; that is, we need to know about the estimator, and its sampling properties. If the estimate of 75% was just a guess, we have no statistical means of assessing its reliability. If it was based on a sample survey, we do. If the survey was of 4 people, we would give little credence to the estimate; it would have large variance (if we asked a different 4 people we may get a very different sample estimate). If the survey was of 4,000 people, provided the sample was representative, we would have much more confidence in the estimate.

We know that different samples of 4,000 would give us different answers, but we would expect the precision of the estimator to be high (i.e. its variance would be low), and different samples should yield very similar estimates. The larger the sample we take, the smaller the variance of our estimate.

2.1 Unbiased and Consistent Estimators

Mathematically, suppose that we have random variables (rvs) $X_1, \ldots, X_n$ and we wish to estimate some property of the underlying distribution from which the rvs are drawn. (We assume that each X_i are independent and identically distributed). Let θ denote the population level parameter of interest, such as the mean or variance.

The true value of the parameter, θ, is an unknown constant that can be ascertained correctly only by an exhaustive study of the population. The objective of estimation is to provide an estimate of the unknown true value of the parameter (along with an indication of the precision of the estimate) following observing values for the rvs $X_1, \ldots, X_n$, which we denote by the dataset $\mathbf{x} = \{x_1, \ldots, x_n\}$.

Definition: An estimator $T = f(X_1, \ldots, X_n)$ is and *unbiased* estimator of the parameter θ if $\mathbb{E} = \theta$.
Definition: An estimator T is a **consistent** estimator of the parameter θ if,

1. $\mathbb{E}(T) \to \theta$ as the sample size $n \to \infty$ (i.e. it is an unbiased estimator of θ in the limiting case); and
2. $Var(T) \to 0$ as $n \to \infty$.

A statistic $T = f(X_1, \ldots, X_n)$ is an estimator and is itself a random variable (as it is a function of random variables) and hence has some distribution (completely determined by the distribution of the X_i's and function f).

A realisation of T, denoted $T = f(X_1, \ldots, X_n)$ is an *estimate* of the unknown population parameter, after observing sampled values $\mathbf{x} = \{x_1, \ldots, x_n\}$. It has a numerical value.

Unbiased estimators are generally desirable, since they provide an accurate estimate of the parameter of interest ”on average”. Similarly, precise estimators (i.e. estimators with small variability) are also desirable.

In general an unbiased estimator is preferred to a biased estimator. However, a biased estimator, T_1, may sometimes be preferred to an unbiased estimator, T_2, if the bias is very small and the variance of T_1 is significantly smaller than that for T_2.

2.2 Sample Mean and Variance

Let $X_1, \dots, X_n$ be independent rvs each with mean μ and variance σ^2. The sample mean is defined by $\bar{X}$, where,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample variance is denoted by S^2 and given by,

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

Aside: Note that we can write,

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right)$$

These expressions for the sample variance are typically easier to calculate than the original formula. The proof that these formula are equivalent to each other is simply an exercise of algebraic manipulation.

Theorems

1. The sample mean $\bar{X}$ is an unbiased and consistent estimator of the population mean μ.
2. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

Example

Let $X_1, \dots, X_n$ be independent rvs each with mean μ. Let T be the estimator of μ such that $T = \frac{1}{n-1}\sum_{i=1}^{n} X_i$. Then,

1. T is a biased estimator of μ; but
2. T is a consistent estimator of μ.

For simplicity, we note that $T = \frac{n}{n-1}\bar{X}$. Then, we have that

$$\mathbb{E}(T) = \frac{n}{n-1}\mathbb{E}(\bar{X}) = \frac{n}{n-1}\mu \quad (\neq \mu).$$

Thus T is a biased estimator of μ. However, $\mathbb{E}(T) \to \mu$ as $n \to \infty$.
Similarly, we have that,

$$\begin{aligned} \text{Var}(T) &= \text{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{(n-1)^2}\sum_{i=1}^{n}\text{Var}(X_i) \qquad \text{since } X_i\text{'s are independent} \\ &= \frac{n}{(n-1)^2}\sigma^2, \end{aligned}$$

So that Var$(T) \to 0$, as $n \to \infty$.
Thus we have $\mathbb{E}(T) \to \mu$ as $n \to \infty$ and Var$(T) \to 0$, as $n \to \infty$ so that T is a consistent estimator of μ.

Example

Suppose that we wish to estimate the proportion p of Mathematics students whose favourite branch of Mathematics is Statistics. Then, we could carry out a survey and ask n students at random. Of these n students, let X denote the number that answered that Statistics was their favourite branch of Mathematics. Then, $X \sim Bin(n, p)$, where we wish to estimate p.
We can estimate p by $T = X/n$, where,

$$\mathbb{E}(T) = \mathbb{E}(X/n) = \frac{1}{n}\mathbb{E}(X) = \frac{1}{n}.np = p.$$

Thus X/n is an unbiased estimator of p.
Further,

$$\text{Var}(X/n) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2}.np(1-p) = \frac{p(1-p)}{n}.$$

Thus Var$(T) \to 0$ as $n \to \infty$, so that T is also a consistent estimator of p.

Sample Covariance and Correlation

Recall that Cov$(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ and the correlation is given by $\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$. They both provide information on the relationship between the random variables X and Y (but the correlation is typically preferred due to its invariance to scale property).
Suppose that you have paired random variables X and Y: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The covariance of X and Y can be estimated using the *sample covariance*:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right].$$

This is again an unbiased and consistent estimator of Cov(X,Y). (Proofs follows similarly to that for the sample variance).

Notes:

- S_{xx} and S_{yy} are the *sample variances* for X and Y, respectively - this is immediately seen by comparing with the formula for the sample variance - see §2.2 (remember that Cov(X,X) = Var(X)).
- An estimate for the sample correlation coefficient, $\rho(X,Y)$, is simply

  $$R_{xy} = \frac{S_{xy}}{S_xS_y}$$

  where S_x is the sample standard deviation of the X_i (i.e. $\sqrt{S_{xx}}$), and similarly for S_y.
- Once more note the distinction between the population covariance and correlation (Cov(X,Y) and $\rho(X,Y)$) which are the properties of the rvs X and Y; and the sample covariance and correlation (S_{xy} and R_{xy}) which are estimators of Cov(X,Y) and $\rho(X,Y)$ based on observations of X and Y, and so subject to random variation.

2.3 Maximum Likelihood Estimators

In many cases it can be obvious what a sensible estimator may be for a given parameter. For example, for the population mean, μ, the sample mean, $\bar{X}$ (that is unbiased and consistent), or for a proportion p, the sample proportion, X/n. However, in other cases it is not so obvious how to estimate parameters, and a general method is needed.
We will consider the method of maximum likelihood. We note that maximum likelihood estimators (MLEs) are consistent estimators (but may be biased - though the bias is usually small). Maximum likelihood is a simple idea,

most easily described by a motivating example.

Example: Consider rolling a loaded die, which has probability p of coming up with a six. The die is rolled a total of n times. Let X denote the total number of sixes that are obtained. Then $X \sim Bin(n, p)$. You roll the die $n = 10$ times and get $x = 4$ sixes. The probability of this happening is

$$\mathbb{P}(\text{4 sixes in 10 rolls}) = P(X = 4) = \binom{10}{4} p^4 (1-p)^6$$

.
Consider the value of p which will make this probability as high as possible (i.e. find the maxima by differentiating the probability with respect to p and set the value equal to zero):

$$\frac{d}{dp}\left(\binom{10}{4} p^4 (1-p)^6\right) = 0 \quad \Rightarrow \quad 4 = 10p \quad \Rightarrow \quad \hat{p} = \frac{4}{10}$$

Note that, in general, you should check that you have obtained a maxima rather than a minima. This can be easily done by taking the second derivative and showing that this is negative.

This estimate is known as the **maximum likelihood estimate** (MLE) of p. We typically denote the MLE of the parameter by using a ”hat” on the given parameter, $\hat{p}$. The idea is that the most likely value for the parameter is the one which makes the data appear most probable. Maximum likelihood estimation is the process of finding the parameters which make a set of data look as probable as possible.

Likelihood for Discrete Distributions

Consider a set of observations $x_1, \ldots, x_n$ which are modelled as observations of independent discrete random variables with probability mass function $f(x_i; \theta)$ depending on some (vector of) parameters θ. Recall that since the random variables are discrete $f(x_i; \theta) = P(X_i = x_i)$. Assuming that the data are independent, the joint probability mass function of the observations $\mathbf{x} = \{x_1, \ldots, x_n\}$, denoted $f(x_1, \ldots, x_n; \theta)$ is the product of the probability mass functions for each observation. Mathematically,

$$f(x_1, \ldots, x_n; \theta) = \Pi_{i=1}^{n} f(x_i, \theta)$$

.
We seek the parameters of the model that make the data look most probable, so we seek to maximise $f(x_1, \ldots, x_n; \theta)$ with respect to θ. When $f(x_1, \ldots, x_n; \theta)$ is considered as a function of the parameters in this way, it is known as the **likelihood function of the parameters** (rather than the probability of the data) and we use the alternative notation $L(\theta; x_1, \ldots, x_n)$.

Note that the logarithm of a function is maximised at the same set of parameters as the function itself. Very often it is easier to maximise the **log-likelihood**

$$l(\theta; x_1, \ldots, x_n) = \log L(\theta; x_1, \ldots, x_n)$$

.
Recall that (assuming the data are independent observations), the likelihood $L(\theta; x_1, \ldots, x_n) = \Pi_{i=1}^{n} f(x_i, \theta)$, so we have that,

$$l(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} f(x_i, \theta)$$

.

Example

Consider the above example relating to the loaded dice, where $X \sim Bin(10, p)$, where p is unknown and to be estimated. We have already maximised the likelihood function to find the MLE for p. Now consider the log-likelihood function:

$$l(; x_1, \ldots, x_n) = \log[\mathbb{P}(X = 4)] = \log \binom{10}{4} + 4 \log p + 6 \log(1 - p)$$

so that

$$\frac{\mathrm{d}l}{\mathrm{d}p} = \frac{4}{p} - \frac{6}{(1 - p)}.$$

Equating this to 0 gives $4(1 - p) = 6p$, from which $\hat{p} = 4/(6 + 4) = 0.4$. However, the differentiation is much easier using the log-likelihood function compared to the likelihood function.
To show that the turning point is indeed a maximum, taking the second derivative we obtain,

$$\frac{\mathrm{d}^2 l}{\mathrm{d}p^2} = -\frac{4}{p^2} - \frac{6}{(1 - p)^2} < 0.$$

Likelihood for Continuous Distributions

Maximum likelihood can also be used to estimate parameter(s) from continuous distributions. The only difference is that the likelihood function is now equal to the joint probability density function of the observed data. Consider a set of observations $x_1, \ldots, x_2$ which are modelled as observations of independent continuous random variables with probability density function $f(x_i; \theta)$ depending on some (vector of) parameters θ. Assuming that the data are independent, we can write the likelihood in the form,

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) = \Pi_{i=1}^n f(x_i, \theta)$$

This is still a likelihood, but can no longer be interpreted as a probability of getting the observed data, given θ; instead, it is the joint probability density function of the observed data. This makes no difference to the actual calculations. We still maximise the likelihood with respect to the parameters,and it is still easier to use the log-likelihood in most cases.

The maximum likelihood estimator $\hat{\sigma}^2$ of the variance σ^2 is not the sample variance S^2 (which is the usual unbiased estimator of σ^2). In general MLEs are biased estimators (although the bias is usually fairly small). MLEs are consistent estimators.

Invariance Property of MLEs

Let $\hat{\theta}$ denote the maximum likelihood estimator of θ and g be any function of θ. Then, the maximum likelihood estimator of $g(\theta)$ is $g(\hat{\theta})$.

Example: let $x_1, \ldots, x_k$ be independent observations of binomial rvs, each with n trials and unknown probability p. The likelihood function is

$$L(p; x_1, \ldots, x_k) = \prod_{i=1}^{k} \binom{n}{x_i} p^{x_i} (1 - p)^{n - x_i}.$$

The MLE of p is calculated by finding the value of p that maximises the likelihood function. This is equivalent to finding the value of p that maximises the log-likelihood function:

$$l(p; x_1, \ldots, x_k) = \sum_{i=1}^{k} \left[\log \left(\frac{n!}{(n - x_i)! x_i!} \right) + x_i \log(p) + (n - x_i) \log(1 - p) \right]$$

Differentiating the log-likelihood, and setting it equal to zero, we have:

$$\begin{aligned} \frac{\mathrm{d}l}{\mathrm{d}p} &= 0 \\ \Rightarrow \quad \frac{\sum_{i=1}^k x_i}{p} - \frac{\sum_{i=1}^k (n-x_i)}{1-p} &= 0 \\ \Rightarrow \quad p\sum_{i=1}^k (n-x_i) &= (1-p)\sum_{i=1}^k x_i \\ \Rightarrow \quad \hat{p} &= \frac{\sum_{i=1}^k x_i}{kn} = \bar{x}/n. \end{aligned}$$

Using the invariance property, we can immediately deduce the maximum likelihood estimates of the mean and variance of the Bin(n,p) distribution. Recall that mean $= \mu = np$ and variance $= \sigma^2 = np(1-p)$. Thus, using the invariance property:

$$\hat{\mu} = n\hat{p} = \bar{x} \qquad \text{and} \qquad \hat{\sigma}^2 = n\hat{p}(1-\hat{p}) = \bar{x}(1-\bar{x}/n).$$

2.4 Interval Estimation

The previous point estimates provide an indication of the location of the parameter value. However, they provide no indication of the uncertainty associated with the estimate. Confidence intervals can be seen to fill this gap, providing a measure of how *precise* the estimate is: a small confidence interval means that there is little uncertainty regarding the estimate; whereas a large confidence interval means that there is a great deal of uncertainty regarding the estimate.

Definition: A $100(1-\alpha)\%$ **Confidence Interval (CI)** for a parameter θ is an interval that contains the true value of θ with some (typically high) probability. For example, let (θ_1, θ_2) be a 95% CI for θ (so $\alpha = 0.05$). If we were to repeat our experiment 100 times, we would expect that 95 of the derived Confidence Intervals contain the true value of θ.

Normal Distribution with Known Variance

Suppose that we have independent rvs $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is *known*. We wish to (i) estimate μ; and (ii) obtain a $100(1-\alpha)\%$ CI for μ.

(i) From §2.2 (Theorem 1), we know that $\bar{X}$ is a consistent and unbiased estimator of μ.

(ii) In order to obtain a corresponding CI for μ, we need to know the distribution of the estimator $\bar{X}$. We also know that (from §2.2),

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Finally, we will use a standard result for Normal distribution that a linear sum of normally distributed rvs is also a normally distributed rv, i.e. for $X_1, \ldots, X_n \sim N$ then $T = \sum_{i=1}^n X_i \sim N$.
Combining these results we have,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

We specify the linear transform, $Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$, and use the standard result (for the normal distribution) that,

$$Z \sim N(0,1).$$

The final piece of the jigsaw puzzle we need before we put everything together is that for $Z \sim N(0,1)$, then

$$\mathbb{P}(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

where $z_{\alpha/2}$ is such that $\mathbb{P}(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$. However, the distribution is symmetrical about 0, so that $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$) and we can write

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Putting all the pieces together we have that,

$$\begin{aligned}
\mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} < z_{\alpha/2}\right) &= 1 - \alpha \\
\Rightarrow \mathbb{P}\left(\mu - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
\Rightarrow \mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right) &= 1 - \alpha.
\end{aligned}$$

(Note that the random quantity here is $\bar{X}$ and not μ). The probability indicates that *before observing the data*, the random interval $(\bar{X} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}})$ will include the unknown parameter μ with probability $(1 - \alpha)$. Once the data are observed, the $100(1 - \alpha)\%$ CI for μ is given by,

$$\left(\bar{x} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right);$$

or equivalently,

$$\bar{x} \pm \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}.$$

Note that this is not a probability interval (the probability interval relates to the rv $\bar{X}$). The data have been observed and $\bar{x}$ is an observed number.

Notes:

- Before observing any data, a confidence interval $\left(\bar{X} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right)$ is a random interval for the true value of the mean, μ.
- The probability
$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right) = 1 - \alpha$$
is interpreted such that the long-term relative frequency that the intervals include μ is equal to $1 - \alpha$. (I.e. if we could repeat the experiment a large number of times then of the constructed confidence intervals, $1 - \alpha$ of them would include μ).
- Once we observe the data, and hence $\bar{x}$, the interval,
$$\left(\bar{x} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right)$$
is described as a $100(1 - \alpha)\%$ *confidence interval* for μ. It is no longer appropriate to speak of a probability that it will include a fixed value, μ. (It either includes μ or does not).
- For any particular application it is never known if the calculated 95% CI contains the unknown mean μ.

Normal Distribution with Unknown Variance

Suppose that we have independent rvs $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is *unknown*. As before, we wish to (i) estimate μ and (ii) obtain a $100(1 - \alpha)\%$ CI for μ.

(i) Once more we consider the unbiased and consistent estimator $\bar{X}$ for μ.

(ii) An issue arises when we wish to obtain the corresponding CI for μ. As before, we consider the transformation,

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}.$$

However, we no longer know the variance σ^2. We can estimate σ^2 by the sample variable S^2 and consider the transformation,

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}.$$

Now we need to derive the corresponding distribution for T (it is no longer normal when we estimate the variance, σ^2 by the sample variance S^2). In order to do this we need some additional distributional results relating to the normal distribution

χ^2 Distribution

Let $Z_1, \dots, Z_n$ be *independent* $N(0,1)$ rvs and let $X = \sum_{i=1}^n Z_i^2$. Then the rv X has a chi-squared distribution with n degrees of freedom, and we write $X \sim \chi^2_n$.

Notes:

- X is a continuous rv and can take values $x \geq 0$.
- Let $Z \sim N(0,1)$ and let $Y = Z^2$. Then $Y \sim \chi^2_1$ (set $n = 1$ in the above definition).
- Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$, *independently*. Then, $X + Y \sim \chi^2_{n+m}$.
- If $X \sim \chi^2_n$, then $\mathbb{E}(X) = n$ and Var$(X) = 2n$.

t Distribution

Let Z and Y be *independent* rvs, such that $Z \sim N(0,1)$ and $Y \sim \chi^2_n$. Let

$$T = \frac{Z}{\sqrt{Y/n}}.$$

Then, T has a t-distribution with n degrees of freedom (*df*), and we write $T \sim t_n$.

Notes:

- T is a continuous rv and can take any value $t \in \mathbb{R}$.
- As $n \to \infty$, $t_n \to N(0,1)$. In practice, we often use the approximation $t_n N(0,1)$ for $n \geq 30$.
- If $T \sim t_n$, then $\mathbb{E}(T) = 0$ and Var$(T) = n/(n-2)$ for $n > 2$.
- We use the notation $t_{n;\alpha}$ to denote the upper α quantile of the t_n distribution, i.e. $\mathbb{P}(T \geq t_{n;\alpha}) = \alpha$. As the distribution is symmetrical about 0, $\mathbb{P}(T \geq t) = \mathbb{P}(T \leq -t)$ and $t_{n;\alpha} = -t_{n;1-\alpha}$.

Properties of the Sample Mean and Variance for the Normal Distribution

Let $X_1, \dots, X_n$ be independent $N(\mu, \sigma^2)$ rvs, and $\bar{X}$ and S^2 the corresponding sample mean and variance. Then,

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \tag{1}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2}$$

$$\bar{X} \text{ and } S^2 \text{ are independent.} \tag{3}$$

Note that (3) is quite remarkable. It only holds for normally distributed random variables.

Confidence Interval

We can now combine all the above results to derive the desired confidence interval for μ, for the case where we have independent rvs $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, such that σ^2 is unknown. Consider the statistic,

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \\ &= \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \\ &= \frac{Z}{\sqrt{Y/(n-1)}}, \end{aligned}$$

where,

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \qquad \text{and} \qquad Y = \frac{(n-1)S^2}{\sigma^2}.$$

From (1) we have that $Z \sim N(0,1)$; from (2), $Y \sim \chi^2_{n-1}$ and from (3) Z and Y are independent (as they are simply linear transformations of $\bar{X}$ and S^2). Hence, from the definition of the t distribution, $T \sim t_{n-1}$. In other words,

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}.$$

Notationally, we let $t_{n-1;\alpha}$ denote the value such that $\mathbb{P}(T \geq t_{n-1;\alpha}) = \alpha$, for $T \sim t_{n-1}$. Values of $t_{n-1;\alpha/2}$ can be obtained in `R` using the inverse cdf command `qt`. We note that the t distribution is symmetrical about 0, so that $t_{n-1;1-\alpha/2} = -t_{n-1;\alpha/2}$.
Finally, for a $100(1-\alpha)\%$ CI for μ, we use,

$$\mathbb{P}\left(-t_{n-1;\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{n-1;\alpha/2}\right) = 1 - \alpha.$$

Rearranging the expression, we obtain,

$$\mathbb{P}\left(\bar{X} - \frac{t_{n-1;\alpha/2} S}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{t_{n-1;\alpha/2} S}{\sqrt{n}}\right) = 1 - \alpha.$$

In other words, a $100(1-\alpha)\%$ CI for μ is,

$$\left(\bar{x} - \frac{t_{n-1;\alpha/2} s}{\sqrt{n}}, \bar{x} + \frac{t_{n-1;\alpha/2} s}{\sqrt{n}}\right),$$

or equivalently,

$$\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}.$$

2.5 Exercises

1. Let $X_1, \ldots, X_n$ denote a set of independent and identically distributed random variables. Let $T = f(X_1, \ldots, X_n)$ be a consistent estimator of some parameter θ. Which of the following statements is always true?

   (a) T is an unbiased estimator of θ.

   (b) T is a biased estimator of θ.

   (c) $\mathbb{E}(T) \to \theta$ as $n \to \infty$.

   (d) $\mathbb{E}(T) \to \theta$ and Var$(T) \to \theta$ as $n \to \infty$.

2. Let $X \sim Bin(n, p)$ with probability mass function,

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n; \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbb{E}(X) = np$ and $\text{Var}(X) = np(1-p)$. Which of the following statements is false?

(a) The maximum likelihood estimate of p is $\frac{x}{n}$.

(b) The estimator $\frac{X}{n}$ is an unbiased estimator of p.

(c) The maximum likelihood estimate of p^2 is $\left(\frac{x}{n}\right)^2$.

(d) The estimator $\left(\frac{X}{n}\right)^2$ is an unbiased estimator of p^2.

3. The random variable X is such that $\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$ for each of a series of binary trials (i.e. $X \sim Bernoulli(p)$). Let the random variable Y be the number of trials up to and including the first success (so that $Y \sim Geom(p)$).

(a) Show that the probability mass function of Y is $f(y) = pq^{y-1}$, for $y = 1, 2, \ldots$, where $q = 1 - p$.

(b) Given a single observation y, find the maximum likelihood estimate of p.

(c) What is the maximum likelihood estimate of the log odds defined by $\log\left(\frac{p}{q}\right)$?

3 Hypothesis Testing

Note: *hypothesis testing is also referred to as "significance testing" by some authors.*

Often we wish to test a "hypothesis" about a parameter:
In a study to test whether a coin is fair, we may toss the coin 10 times and record the number of heads we obtain.

$$\text{null hypotheses} \rightarrow H_0\text{: coin is fair}$$
$$\text{alternative hypotheses} \rightarrow H_1\text{: coin is fair}$$

To formulate this example statistically. Let p be the probability of a head when we toss the coin; and let T be the number of successes. Then, $T \sim Bin(10, p)$. We consider two steps. Firstly, we consider the hypotheses:

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

If we reject H_0 in favour of H_1, we would conclude that there is evidence against $p = 0.5$ in favour of the alternative $p \neq 0.5$. We could then consider a secondary step (if this was of interest) and see whether the bias was in favour of a head or a tail (i.e. which "tail" the parameter, p, was in).

3.1 Underlying Concepts

Suppose a data set is drawn from a distribution which depends on a parameter $\theta \in \Theta$. A statistical hypothesis is a statement about the value of θ (a subset of Θ). A **simple hypothesis** specifies θ completely (a single point in Θ), i.e. $\theta = \theta_0$. A **composite hypothesis** specifies a set of more than one value of θ, e.g. $\theta \in [0, 0.5]$.
To formulate a hypothesis test, two hypotheses are needed: the null hypothesis, H_0, which may be rejected in favour of the alternative hypothesis H_1 (sometimes H_A is used). H_0 is typically the hypothesis of no change and is usually (but not always) a simple hypothesis and H_1 is typically a composite hypothesis. We observe some data, $\mathbf{x}$, and calculate the observed **test statistic**, t, which is a function of $\mathbf{x}$.
A hypothesis test is a rule of the form:

$$\text{Reject } H_0 \text{ in favour of } H_1 \text{ if } t \in C.$$

C is called the **critical region** of the test. An associated idea (and which can be used to determine the critical region) is the p-value. This is defined as the probability of observing a result at least as extreme as the observed data statistic, t, given that the null hypothesis is true. In other words, the p-value is the probability of observing the data we observed (as summarized by the test statistic) or less likely data, given that H_0 is true.

- If the p-value is "small", it is unlikely that the null hypothesis is true (as the observed data do not appear to be consistent with H_0), so we would reject H_0.
- If the p-value is "large", then the data observed are consistent with H_0, and there is no evidence to reject H_0.

In practice, statisticians are pragmatic and do not simply "reject" or "not reject" a hypothesis. Instead we typically interpret p-values as follows:

p**-value**	**Interpretation**
> 0.1	No evidence against H_0 (not significant)
$0.05 - 0.1$	Weak evidence against H_0
$0.01 - 0.05$	Moderate evidence against H_0
< 0.01	Strong evidence against H_0

Note: "No evidence against H_0" $\neq$ "Evidence for H_0".

3.2 Significance Level and Power

Within hypothesis testing there are two possible errors:

$$\textbf{Type 1 error} : \text{Reject } H_0 \text{ when } H_0 \text{ is true.}$$
$$\textbf{Type 2 error} : \text{Fail to reject } H_0 \text{ when } H_0 \text{ is false.}$$

In the classical theory of testing, a Type I error is more serious. We let $\mathbb{P}(\text{Type I error}) = \alpha$ and is called the **significance level** or size of the test. A test result is significant (at level α) and we reject H_0 if the p-value $\leq \alpha$. For example, if $\alpha = 0.05$ (often denoted 5%), then we reject H_0 if the probability of observing at least as extreme value as the observed value x (i.e. the p-value) is less than 0.05.
Alternatively, the **power** of the test is the probability of rejecting H_0, when it is false, and is denoted by β. Then, $\mathbb{P}(\text{Type II error}) = 1 - \beta$. Ideally we would like $\alpha = 1 - \beta$ but this is only achieved in trivial cases.
Finally, the **power function** is the power expressed as a function of the parameter θ and is denoted by, $\beta(\theta^*) = \mathbb{P}(\text{reject } H_0 : \theta = \theta_0 \text{ when the true value is } \theta^*)$. Since β is a probability, $0 \leq \beta(\theta^*) \leq 1$ for all $\theta^* \in \Theta$.

3.3 Conducting a Hypothesis Test

The same steps are undertaken when conducting any form of statistical hypothesis:

1. State the null and alternative hypotheses, H_0 and H_1.
2. Select an appropriate test statistic, T, and evaluate this at the observed data, $\mathbf{x}$, giving an observed test statistic, t.
3. Determine the distribution of the test statistic, T, assuming the null hypothesis is true.
4. Calculate the critical region for the test statistic for a given significance level α, denoted C (The set of values of the test statistic for which we would reject H_0 at the 100α% level) **or** Calculate the p-value of the observed test statistic (The probability of observing at least an extreme result as the observed test statistic assuming the null hypothesis is true).
5. Reject H_0 in favour of H_1 if $t \in C$ **or** the p-value is less than α. Fail to reject H_0 if $t \notin C$ **or** the p-value is greater than α.

3.4 Coin Toss Example

Recall that, we wish to test whether a coin is fair or not. The coin is tossed 10 times. Let T be the number of successes. Then, $T \sim Bin(10, p)$. Then, we can rewrite the hypotheses in the form:

$$H_0 : p = 0.5,$$
$$H_1 : p = 0.5.$$

Then, to test whether the coin is fair, we need to calculate the p-value of the observed test statistic. Recall the p-value is defined as the probability of observing a result at least as extreme as the observed data, given the null hypothesis is true, i.e. $p = 0.5$. But what is ”at least as extreme” in this case?

Calculating the p-value

Under the null hypothesis the test statistic $T \sim Bin(10, 0.5)$. Consider an observed test statistic of $t <$ median(T). Then the probability of the extreme ”left-tail” is $\mathbb{P}(T \leq t)$. Recall that this is simply evaluating the cumulative density function (cdf). We assume that ”at least as extreme” corresponds to being in the same upper quantile of the distribution. So that the p-value is $2 \times \mathbb{P}(T \leq t)$.
Alternatively (and symmetrically) if $t >$ median(T), the probability of the extreme ”right-tail” is $\mathbb{P}(T \geq t)$. We assume that ”at least as extreme” corresponds to being in the same lower quantile of the distribution. So that the p-value is $2 \times \mathbb{P}(T \geq t)$. To calculate these probabilities we use R and the command ”pbinom” that evaluates the cdf

of a binomial distribution. Note that if $t = \text{median}(T)$ then this is clearly not an extreme value as it is the middle value of the distribution and we define the associated p-value to be equal to 1.
For our coin tossing example we have that $\text{median}(T) = 5$, so that for $t = 5$ the associated p-value is 1. To calculate $\mathbb{P}(T \leq t)$, for $t = 0, 1, \dots, 4$ use the R command:

```
> pbinom(t,10,0.5)
```

The p-value is then given by $2 \times \mathbb{P}(T \leq t)$. To calculate $\mathbb{P}(T \leq t)$, for $t = 6, \dots, 10$, recall that as T is a discrete random variable, so that,

$$\mathbb{P}(T \geq t) = 1 - \mathbb{P}(T < t) = 1 - \mathbb{P}(T \leq t - 1).$$

So, for a given value of t, in R this can be obtained by using:

```
> 1 - pbinom(t-1,10,0.5)
```

The p-value is then given by $2 \times \mathbb{P}(T \geq t)$. We obtain the following p-values:

t	0	1	2	3	4	5	6	7	8	9	10
p-**value**	0.002	0.021	0.109	0.344	0.754	1.000	0.754	0.344	0.109	0.021	0.002

Assuming a significance level $\alpha = 0.05$, we reject the null hypothesis if the p-value is $\leq \alpha = 0.05$. Thus we reject H_0 for $t \leq 1$ and $t \geq 9$ (i.e. the critical region is $\{t : t \leq 1, t \geq 9\}$). For all these values we would interpret this as evidence in favour of the coin being biased.

Calculating the Critical Region

Alternatively, we may be interested in calculating the set of values of t for which we would reject H_0 at the 5% significance level, i.e. the critical region C, such that,

$$\mathbb{P}(T \in C) = 0.05.$$

This is most easily done in R and the inverse cdf command ”qbinom”. At the 5% significance level we want 2.5% in each tail. To calculate the right-hand tail we want the value of t such that $\mathbb{P}(T \geq t) = 0.025$. R uses the inverse cdf and calculates the value of t such that $\mathbb{P}(T \leq t)$ is equal to some value. Thus, we want to find the value of t, such that $\mathbb{P}(T \leq t) \leq 0.975$. In R:

```
> qbinom(0.975,10,0.5)
[1] 8
```

Note that in this case, since T is discrete, we cannot find a value of t such that $\mathbb{P}(T \leq t) = 0.975$ exactly. For such discrete distributions, the above command in R gives the smallest value of t such that

$$\mathbb{P}(T \leq t) \geq 0.975 \Rightarrow P(T > t) \leq 0.025.$$

In other words in this case, R tells us that $\mathbb{P}(T \leq 8) \geq 0.975$ and $\mathbb{P}(T < 8) = \mathbb{P}(T \leq 7) < 0.975$. Thus (using the honesty condition), $\mathbb{P}(T \geq 9) \leq 0.025$ and $\mathbb{P}(T \geq 8) > 0.025$. Thus $t = 9$ is the smallest values such that $\mathbb{P}(T \geq 9) < 0.025$.
Using the R command for the cdf ”pbinom” we can calculate the probabilities exactly (and check our above answer). To calculate $\mathbb{P}(T \geq 9) = 1 - \mathbb{P}(T \leq 8)$; and $\mathbb{P}(T \geq 8) = 1 - \mathbb{P}(T \leq 7)$ we can use:

```
> 1 - pbinom(8,10,0.5)
[1] 0.011

> 1 - pbinom(7,10,0.5)
[1] 0.055
```

For the left hand-tail, following a similar (but simpler procedure), we find the largest value of t such that $\mathbb{P}(T \leq t) \leq 0.025$ using:

```
> qbinom(0.025,10,0.5)
[1] 2
```

Thus $t = 2$ is the smallest value such that $\mathbb{P}(T \le t) \ge 0.025$. Thus $t = 1$ is the largest values such that $\mathbb{P}(T \le t) \le 0.025$. We can again check by calculating $\mathbb{P}(T \le 2)$ and $\mathbb{P}(T \le 1)$:

```
> pbinom(2,10,0.5)
[1] 0.055

> pbinom(1,10,0.5)
[1] 0.011
```

Thus, the critical region is $C = \{t : t \le 1, t \ge 9\}$ or $C = \{0, 1, 9, 10\}$.
Note: since T is a discrete random variable we cannot choose a critical region such that the size of the test is exactly α in general. We choose the region that has at most $\alpha/2$ in each tail.

Power Function

The power function is the probability that we reject H_0. In other words, $\mathbb{P}(T \le 1|p) + \mathbb{P}(T \ge 9|p)$. These can be easily calculated using for example R and the command:

```
> pbinom(1,10,p)+1-pbinom(8,10,p)
```

Substituting in different values of p we obtain:

p	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
$\beta(p)$	0.376	0.244	0.149	0.086	0.048	0.028	0.021	0.028	0.048	0.086	0.129	0.244	0.376

3.5 Hypothesis Tests for Normal Distribution

3.5.1 z-test - Known Variance

Let $X_1, \dots, X_n$ be independent $N(\mu, \sigma^2)$ rvs, where σ^2 is known. Suppose that we wish to test,

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

We define the test statistic $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$. Then, under H_0, $T \sim N(0, 1)$, so that the critical region at the α significance level (i.e. values at which we would reject H_0) is given by,

$$|T| = \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \ge z_{\alpha/2}.$$

Example

Let $X_1, \dots, X_{100}$ be scores on an exam. We assume that $X_i \sim N(\mu, 144)$ independently of each other, so that $\bar{X} \sim N(\mu, 1.44)$. We wish to test whether or not the mean μ is equal to 65, so that we specify

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_1 : \mu \neq 65.$$

We observe $\bar{x} = 63.5$ and wish to perform a hypothesis test at the 5% significance level. To perform the hypothesis test we calculate the corresponding p-value. We consider the statistic,

$$T = \frac{\bar{X} - 65}{\sqrt{1.44}}$$

so that if H_0 is true $T \sim N(0, 1)$. The observed test statistic is $(63.5 - 65)/1.2 = -1.25$.

Calculating the p-value

Note that T is symmetrical around 0. The p-value is calculated as,

$$\mathbb{P}(|T| \ge 1.25) \quad = \quad 2\mathbb{P}(T \ge 1.25) = 2\mathbb{P}(T \le -1.25) = 0.21.$$

(using `2*pnorm(-1.25,0,1)`). We do not reject H_0 as the p-value is $\ge \alpha = 0.05$.

Calculating the Critical Region

An alternative to calculating the p-value is to calculate the critical region, and see whether the observed test statistic lies within this region. For size $\alpha = 0.05$, by definition,

$$\begin{aligned} \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) &= 0.05 \\ \Rightarrow \mathbb{P}(\text{do not reject } H_0 | H_0 \text{ true}) &= 0.95 \\ \Rightarrow \mathbb{P}(-z_{\alpha/2} < T < z_{\alpha/2} | H_0 \text{ true}) &= 0.95 \end{aligned}$$

We need to calculate the critical value of $z_{\alpha/2}$. If H_0 is true, using R we find that,

$$\mathbb{P}(-1.96 < T < 1.96 | H_0 \text{ true}) = 0.95.$$

For example, we could find this in R using the command:

```
> qnorm(0.025, 0, 1)
[1] -1.96
```

So the test is to reject H_0 iff,

$$\begin{aligned} t \leq -1.96 \quad &\text{or} \quad t \geq 1.96 \\ \Rightarrow \frac{\bar{x} - 65}{1.2} \leq -1.96 \quad &\text{or} \quad \frac{\bar{x} - 65}{1.2} \geq 1.96, \end{aligned}$$

i.e. if $\bar{x} \leq 62.65$ or $\bar{x} \geq 67.35$. Thus, since we observe $\bar{x} = 63.5$ we do not reject the null hypothesis.

Interpretation of Significant Results

Suppose our particular concern is whether the exam was too difficult. This means that we are particularly interested in one tail of the distribution - in this case whether the test statistic falls in the left-hand tail (i.e. exam marks were too low). We would conduct the above hypothesis test and conclude the exam is too difficult if $\bar{X} \leq 62.65$ (i.e. we would not conclude this if $\bar{X} \geq 67.35$, although we would reject H_0).

Power Function: β

The power function is the probability that we reject H_0. In other words, $\mathbb{P}(|T| \geq 1.96 | \mu)$. These can be easily calculated, (and plotted) using for example R. If the true mean is $\mu = \mu_0$ then $\bar{X} \sim N(\mu_0, 1.44)$. Set $T = \frac{\bar{X} - 65}{1.2}$ so that $T \sim N\left(\frac{\mu_0 - 65}{1.2}, 1\right)$. Thus, to calculate the power function, we could type in R,

```
> pnorm(-1.96,(mu-65)/1.2,1) + (1-pnorm(1.96,(mu-65)/1.2,1))
```

for different values of μ to calculate $\mathbb{P}(T \leq -1.96) + \mathbb{P}(T > 1.96)$, for $T \sim N\left(\frac{\mu_0 - 65}{1.2}, 1\right)$.
We obtain:

μ	60	61	62	63	64	65	66	67	68	69	70
$\beta(\mu)$	0.986	0.915	0.705	0.385	0.133	0.050	0.133	0.385	0.705	0.915	0.986

Increasing Sample Size

Now, suppose that the number of students taking the course doubles to 200. Let $X_1, \ldots, X_{200}$ be the corresponding scores exam. As before, we assume that $X_i \sim N(\mu, 144)$ independently of each other, so that now $\bar{X} \sim N(\mu, 0.72)$. What is the critical region at the 5% significance level?
We consider the test statistic

$$T = \frac{\bar{X} - 65}{\sqrt{0.72}}.$$

If H_0 is true $T \sim N(0,1)$. Following the same steps as above, we can calculate the critical region. In particular we have that we reject H_0 iff,

$$\begin{aligned} |T| &\geq 1.96 \\ \Rightarrow \left| \frac{\bar{X} - 65}{\sqrt{0.72}} \right| &\geq 1.96 \\ \Rightarrow |\bar{X} - 65| &\geq 1.663 \end{aligned}$$

i.e. if $\bar{X} \leq 63.34$ or $\bar{X} \geq 66.66$.
The power function is calculated as $\mathbb{P}(|T| \geq 1.96|\mu)$. Note that if $\mu = \mu_0$, then,

$$T \sim N\left(\frac{\mu_0 - 65}{\sqrt{0.72}}, 1\right).$$

This can be easily checked as before.
So to calculate the power function, in R we could type,

```
> pnorm(-1.96,(mu-65)/sqrt(0.72),1) + (1-pnorm(1.96,(mu-65)/sqrt(0.72),1))
```

for different values of mu. Thus, considering the values mu $= 60, \ldots, 70$, we obtain the values of the power function:

μ	60	61	62	63	64	65	66	67	68	69	70
$\beta(\mu)$	1.000	0.997	0.942	0.654	0.218	0.050	0.218	0.654	0.942	0.997	1.000

Notes

- Increasing the sample size (i.e. increasing the number of students taking the exam) means that we are more likely to reject the null hypothesis if it is false.
- Increasing the same size also means that it is possible to identify smaller deviations of the value of the parameter from the null hypothesis.

3.5.2 One Sample t-tests (unknown variance)

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables, where σ^2 is unknown. Suppose that we wish to test,

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

We consider the statistic,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

From §2.4 we have that $T \sim t_{n-1}$. Thus, we reject H_0 if $|T| \geq t_0$.
To find t_0, we require,

$$\begin{aligned} \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) &= \alpha \\ \Rightarrow \qquad \mathbb{P}(-t_0 \leq T \leq t_0 | T \sim t_{n-1}) &= 1 - \alpha. \end{aligned}$$

In other words $t_0 = t_{n-1;\alpha/2}$.

Example

A set of 39 observations on pulse rates (in heart beats per minute) has sample mean, $\bar{x} = 70.31$ and sample variance, $s^2 = 90.219$. We wish to test the following hypothesis at the 1% level,

$$H_0 : \mu = 75 \qquad \text{vs} \qquad H_1 : \mu \neq 75.$$

We define the statistic,

$$T = \frac{\bar{X} - 75}{S/\sqrt{n}}.$$

Then, under H_0, $T \sim t_{38}$. Note that the observed test statistic is $t = -3.08$. We reject H_0 if $|t| \geq t_{38,0.005}$. To calculate $t_{38,0.005}$ in R:

```
> qt(0.995,38)
[1] 2.711558
```

Thus we reject H_0 in favour of H_1 at the 1% level.

Alternatively, we could perform the hypothesis test by calculating the p-value. The p-value is $\mathbb{P}(|T| \geq 3.08) = 2\mathbb{P}(T \leq -3.08) = 0.003799$ (using`2*pt(-3.083593,38)` - more decimal places are used here for numerical accuracy). Thus again we reject at the 1% level. We would reject at all significance levels $\geq 0.38\%$.

Finally, suppose that we are particularly interested in whether the mean pulse rate is higher than 75. Although we reject H_0 that the mean is equal to 75, we would NOT conclude that there is evidence of a higher mean pulse rate since the observed data suggests a lower mean pulse rate (i.e. the data is observed in the left tail, whereas a higher mean pulse rate corresponds to the right-tail).

3.5.3 Paired t-tests

In many circumstances we may have paired data of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the two measurements are dependent for an individual i. In this situation, we consider the difference between the two measurements on each unit. In particular, we set $D_i = Y_i - X_i$ for $i = 1, \ldots, n$, and assume that each $D_i \sim N(\mu, \sigma^2)$, independently. Thus, the problem reduces to a one-sample t-test (σ^2 is unknown).

Example

The following table provides the corneal thickness in microns of both eyes of patients who have glaucoma in one eye:

Healthy	484	478	492	444	436	398	464	476
Glaucoma	488	478	480	426	440	410	458	460
Difference	4	0	-12	-18	4	12	-6	-16

The corneal thickness is likely to be similar in the two eyes of any single patient, so that the two observations on the same patient cannot be assumed to be independent. We consider the difference between each pair of observations, denoted by d_i. We wish to test,

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0.$$

Under H_0,

$$T = \frac{\bar{D}}{S/\sqrt{n}} \sim t_{n-1}.$$

Note that $\sum_{i=1}^{8} d_i = -32$ and $\sum_{i=1}^{8} d_i^2 = 936$. Thus, we have $\bar{d} = -4$ and $s^2 = 808/7$. Thus, the observed test statistic is -1.05. To calculate the p-value, we calculate $2 \times \mathbb{P}(T \leq -1.05)$, where $T \sim t_7$. Using R,

```
> 2*pt(-1.05,7)
[1] 0.3286108
```

Thus, we do not reject H_0 at any reasonable significance level.

Alternatively, the critical region at a 5% significance level is $|T| \geq t_{7;0.025} = 2.3646$.

3.5.4 Two Sample t-tests

Suppose that we have two sets of independent rvs, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, where we assume,

$$X_i \sim N(\mu_X, \sigma^2) \qquad \text{and} \qquad Y_j \sim N(\mu_Y, \sigma^2).$$

We wish to test:

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_1 : \mu_X \neq \mu_Y.$$

Now we know that,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right).$$

Using the standard result for (independent) normally distributed rvs:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

If H_0 is true,

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1).$$

However, we typically will not know σ^2, and need to estimate it.
We define the pooled sample variance:

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{m+n-2} \\ &= \frac{(n-1)S_X^2 + (m-1)S_Y^2}{(m+n-2)}. \end{aligned}$$

This is an unbiased estimator of σ^2 - exercise - check.
Now we have that,

$$\begin{aligned} \frac{(n-1)S_X^2}{\sigma^2} &\sim \chi^2_{n-1}; \\ \frac{(m-1)S_Y^2}{\sigma^2} &\sim \chi^2_{m-1}; \\ \Rightarrow \quad \frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} &\sim \chi^2_{m+n-2}. \end{aligned}$$

since S_X^2 and S_Y^2 are independent (since the random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent). (See §2.4 for stated distributional results).
In addition, since we assume that the rvs are normally distributed, the sample variances, S_X^2 and S_Y^2 are independent of $\bar{X}$ and $\bar{Y}$. Using the usual argument, under H_0,

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}.$$

And we can conduct a hypothesis test, using this statistic.

Example

Suppose that we observe two samples from two normal populations (with common variance), such that $\bar{x} = 80.02$, $s_X^2 = 0.024$, with $n = 13$ and $\bar{y} = 79.98$, $s_Y^2 = 0.031$, with $m = 8$. Then $\bar{x} - \bar{y} = 0.04$ and the pooled sample variance is $s_p^2 = \frac{12\times 0.024 + 7\times 0.031}{19} = 0.027$. The observed test statistic is then $\frac{0.04}{\sqrt{0.027}\sqrt{1/13+1/8}} = 0.542$.
The critical region is $|T| \geq t_{19;0.025} = 2.093$. Thus we do not reject H_0 at the 5% level.
Alternatively, the p-value can be calculated, using $\mathbb{P}(|T| \geq 0.542) = 2 \times \mathbb{P}(T \geq 0.542)$

3.5.5 One-sided and Two-sided Tests

Suppose θ is the parameter of interest. A two-tailed test is of the form:

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta \neq \theta_0.$$

In some cases, it may only be possible for the parameter to take values on one-side of θ_0 or there may only be evidence against H_0 in one tail of the distribution. In these cases, one-sided tests are appropriate and are specified as:

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta < \theta_0.$$

Evidence against H_0 in favour of the alternative hypothesis then only considers one tail of the distribution, dependent on the sign of the inequality in H_1 (the left tail for $H_1 : \theta < \theta_0$ and the right tail for $H_1 : \theta > \theta_0$). The nature of the problem typically dictates whether a one-sided or two-sided test is most appropriate.

3.5.6 Confidence Intervals and Hypothesis Tests

Hypothesis testing and confidence intervals are related! In particular, consider the null hypothesis $H_0 : p = p_0$ against the alternative hypothesis $H_1 : p \neq p_0$. The set of values of p_0 for which H_0 is not rejected at significance level α, given the observed data, is equivalent to the $100(1-\alpha)\%$ confidence interval for p. Alternatively, the set of values of p_0 for which the null hypothesis is rejected is equivalent to the complement of the equivalent $100(1-\alpha)\%$ confidence interval for p. This means that if a $100(1-\alpha)\%$ confidence interval is calculated for p, we can immediately deduce (without any further calculations) the outcome of a hypothesis test for p at the $100\alpha\%$ significance level. Suppose that the $100(1-\alpha)\%$ confidence interval for p is (a, b), then for the equivalent hypothesis test at the $100\alpha\%$ significance level:

- If $p_0 \in (a, b)$ - do not reject $H_0 : p = p_0$ in favour of $H_1 : p \neq p_0$;
- If $p_0 \notin (a, b)$ - reject $H_0 : p = p_0$ in favour of $H_1 : p \neq p_0$.

However, confidence intervals provide more information than the equivalent hypothesis tests. For example, suppose that a study found no evidence of an increase in burglaries in Edinburgh in the last 10 years. This conclusion could be reached by failing to reject a null hypothesis that the change in burglary rate was zero at the 5% level. However, the 95% CI for the change in burglary rate may be $(-5\%, 200\%)$. The CI actually provides further information regarding the sort of change which could have been detected in the study (i.e. something about power).

3.6 Exercises

1. Let $X_1, X_2, \dots, X_n$ denote random variables. The estimator $T = f(X_1, \dots, X_n)$ is used to calculate a confidence interval for a parameter $\mu \in \mathbb{R}$. Considering the continuous distribution of T, a statistician calculates that $\mathbb{P}\left(\frac{3T}{\mu} - 3 \leq 12\right) = 0.95$ and $\mathbb{P}\left(\frac{3T}{\mu} + 1 \geq 6\right) = 0.95$. Data are subsequently collected and an observed value of $t = 5$ is obtained. Calculate a 90% confidence interval for μ, given the observed data.

2. Consider a 95% confidence interval $(3, 6)$ for a parameter θ. Which one the following interpretations is correct?

   (a) If we were to repeat our experiment 100 times, we would expect the true value of θ to lie within $(3, 6)$ 95 times.

   (b) The true value of θ lies within $(3, 6)$ with probability 95%.

   (c) If we were to repeat our experiment 100 times, we would expect the true value of θ to lie within the derived confidence interval (not necessarily $(3, 6)$) 95 times.

   (d) We are 95% certain that the true value of θ lies within $(3, 6)$.

3. Let $X_1, \dots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$ random variables, where μ is known and σ^2 is unknown.

   (a) State the distribution of $\sum_{i=1}^n X_i$, clearly justifying your answer.

   (b) Hence, or otherwise, evaluate $\mathbb{P}\left(\sum_{i=1}^n X_i > n\mu\right)$.

   (c) State the distributions of $\frac{X_i - \mu}{\sigma}$ and $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$.

   (d) By considering the distribution of $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$, or otherwise, show that the interval given by,

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:\alpha/2}}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:1-\alpha/2}} \right),$$

is a $100(1-\alpha)\%$ confidence interval for σ^2, where $\chi^2_{n:\alpha}$ corresponds to the upper α quantile of the χ^2_n distribution.

(e) Suppose that we observe the following data $x_1, \ldots, x_{10}$ from a $N(\mu, \sigma^2)$ distribution, where $\mu = 4$ is known, and σ^2 is unknown.

$$2.4 \quad 3.2 \quad 5.4 \quad 4.5 \quad 4.3 \quad 3.6 \quad 2.5 \quad 4.3 \quad 4.9 \quad 4.9$$

Calculate a 95% confidence interval for σ^2.
(Note that $\sum_{i=1}^{10} x_i = 40$ and $\sum_{i=1}^{10} x_i^2 = 169.62$).

4 Linear Regression

Many statistical investigations are concerned with relationships between two (or more) variables, for example: height and weight or rainfall and pressure. In many important cases, one variable (known as the **explanatory variable**) can be measured without error (or with negligible error), whereas the other variable (known as the **response variable**) is random. We let the response random variable be denoted by Y and the associated explanatory variable by x. We assume that the distribution for Y depends on the explanatory variable x. The relationship between the response variable and the explanatory variable is known (for historical reasons) as regression.
Simple linear regression involves finding the regression relationship between the response variable Y and the explanatory variable x, on the basis of n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ on (x, Y). We refer to this is a regression of Y on x.
The simplest form of regression is linear regression, in which the mean of the response variable Y is linearly related to the single explanatory variable x, so that,

$$E(Y) = \alpha + \beta x,$$

where α and β are (unknown) parameters to be estimated.

4.1 Least Squares

We wish to estimate the values of the parameters α and β, i.e. to fit the model to data. One way to do this is by *least squares*. Consider the vertical distances

$$\epsilon_i = y_i - (\alpha + \beta x_i), \qquad i = 1, \ldots, n,$$

between the observed values y_i and the corresponding expected values $\mathbb{E}(Y_i) = \alpha + \beta x_i$ from the linear model.

Sum of squares

A mathematically convenient way of measuring the difference between the data and the model is by the "sum of squares":

$$\begin{aligned} S(\alpha, \beta) &= \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2 \\ &= \sum_{i=1}^{n} \epsilon_i^2. \end{aligned} \tag{4}$$

The *method of least squares* identifies the values $\hat{\alpha}$ and $\hat{\beta}$ of α and β that minimise the sum of squares, $S(\alpha, \beta)$. We can do this via calculus - differentiate $S(\alpha, \beta)$ with respect to α and β, set these differentials equal to zero and solve the simultaneous equations (this may look/sound familiar ...).

First we differentiate with respect to α:

$$\begin{aligned} \frac{\partial S(\alpha, \beta)}{\partial \alpha} &= -\sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) = 0 \\ \Rightarrow \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}. \end{aligned}$$

Now differentiating with respect to β:

$$\begin{aligned}
\frac{\partial S(\alpha,\beta)}{\partial\beta} &= -\sum_{i=1}^{n} 2x_i(y_i-\alpha-\beta x_i)=0 \\
\Rightarrow \sum_{i=1}^{n} x_iy_i &= \hat{\alpha}\sum_{i=1}^{n} x_i+\hat{\beta}\sum_{i=1}^{n} x_i^2 \\
\Rightarrow \sum_{i=1}^{n} x_iy_i &= (\bar{y}-\hat{\beta}\bar{x})\sum_{i=1}^{n} x_i+\hat{\beta}\sum_{i=1}^{n} x_i^2 \\
\Rightarrow \sum_{i=1}^{n} x_iy_i-\sum_{i=1}^{n} y_i\sum_{i=1}^{n} x_i/n &= \hat{\beta}\left(\sum_{i=1}^{n} x_i^2-(\sum_{i=1}^{n} x_i)^2/n\right) \\
\Rightarrow \hat{\beta} &= \frac{\sum_{i=1}^{n} x_iy_i-\sum_{i=1}^{n} y_i\sum_{i=1}^{n} x_i/n}{(\sum_{i=1}^{n} x_i^2-(\sum_{i=1}^{n} x_i)^2/n)}
\end{aligned}$$

The least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are often expressed as

$$\begin{aligned}
\hat{\beta} &= \frac{S_{xy}}{S_{xx}} && (5)\\
\hat{\alpha} &= \bar{y}-\hat{\beta}\bar{x}, && (6)
\end{aligned}$$

where

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \\
S_{xy} &= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right) \\
S_{yy} &= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2.
\end{aligned}$$

(Although S_{yy} is not needed for $\hat{\alpha}$ and $\hat{\beta}$ it will be used later.)

Notes:

- Estimation by least squares requires no assumptions about the distributions of $Y_1,\ldots,Y_n$.
- Since $Y_1,\ldots,Y_n$ are random variables, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ (which are now expressed as a function of the random variables $Y_1,\ldots,Y_n$ instead of the observed data $y_1,\ldots,y_n$) are also random variables.

Properties of $\hat{\alpha}$ and $\hat{\beta}$

1. $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β, respectively.
2. Var$(\hat{\beta})=\frac{\sigma^2}{S_{xx}}$ and Var$(\hat{\alpha})=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$.
3. $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators of α and β, respectively.

4.2 Normal Linear Regression

In general, we need to do more than just find (point) estimates of the parameters α and β; we often wish to test hypotheses about these parameters and/or construct associated confidence intervals. In order to do this, we need to make assumptions about the distributions of $Y_1, \ldots, Y_n$. We assume that $Y_1, \ldots, Y_n$ are independent, normally distributed random variables with the same variance, i.e.

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent.} \tag{7}$$

Equivalently we can write, for $i = 1, \ldots, n$,

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently.
The probability density function for observation $i = 1, \ldots, n$ is given by,

$$f(y_i; \alpha, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right).$$

The corresponding likelihood function of the parameters α and β is given by,

$$\begin{aligned} L(\alpha, \beta, \sigma^2; (x_1, y_1), \ldots, (x_n, y_n)) &= \prod_{i=1}^n f(y_i; \alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \end{aligned}$$

Hence the log-likelihood function is given by

$$\begin{aligned} l(\alpha, \beta, \sigma^2; (x_1, y_1), \ldots, (x_n, y_n)) &= \log L(\alpha, \beta, \sigma^2; (x_1, y_1), \ldots, (x_n, y_n)) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \end{aligned}$$

Notes:

1. The likelihood (and log-likelihood) can be maximised with respect to α and β when σ^2 is unknown.
2. Maximising the log-likelihood is equivalent to minimising the $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ - this is equal to the sum of squares, $S(\alpha, \beta)$, given in equation (4).

Distributional results

Property 2 above means that, for the normal linear regression model given in (7), the maximum likelihood estimates (MLEs) are equal to the least squares estimates $\hat{\alpha} = \bar{y} + \hat{\beta}\bar{x}$ and $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$. We note that this means that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of the normal random variables $Y_1, \ldots, Y_n$, and hence are also normally distributed. Further the mean and variances of these estimators are given in §4.1, so that putting these results together we have:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \tag{8}$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right). \tag{9}$$

The variance σ^2 in (7) is typically unknown, however, it can be estimated from the data. The variance is estimated by S^2 such that

$$S^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{y}_i)^2, \tag{10}$$

where

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

and corresponds to the *fitted value* corresponding to x_i.
It can be shown that S^2 is an unbiased estimator of σ^2, with the distributional result that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{11}$$

Further,

$$S^2 \text{ is independent of } \hat{\alpha} \text{ and } \hat{\beta}$$

[**Warning**: $\hat{\alpha}$ and $\hat{\beta}$ are *not* independent!]
Putting (8)–(9) and (11) together, we obtain

$$\frac{\hat{\beta}-\beta}{\sqrt{\dfrac{S^2}{S_{xx}}}} \sim t_{n-2} \tag{12}$$

$$\frac{\hat{\alpha}-\alpha}{\sqrt{S^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}. \tag{13}$$

These results are the basis of inference on α and β. In particular, they enable us to (use R to) calculate confidence intervals for α and β. For example, 95% confidence intervals for β and α are

$$\hat{\beta} \pm t_{n-2;0.025}\sqrt{\frac{s^2}{S_{xx}}}$$

$$\hat{\alpha} \pm t_{n-2;0.025}\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$$

For convenience we let

$$\begin{aligned}\text{s.e.}\left(\hat{\beta}\right) &= \sqrt{\text{Var}(\hat{\beta})} = \sqrt{\frac{s^2}{S_{xx}}} \\ \text{s.e.}\left(\hat{\alpha}\right) &= \sqrt{\text{Var}(\hat{\alpha})} = \sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}.\end{aligned}$$

These quantities are the *standard errors* of $\hat{\beta}$ and $\hat{\alpha}$, respectively, and are usefully provided within R, when fitting a linear regression model. The confidence intervals can thus be specified in the form:

$$\hat{\beta} \pm t_{n-2;0.025} \times \text{s.e.}\left(\hat{\beta}\right)$$

$$\hat{\alpha} \pm t_{n-2;0.025} \times \text{s.e.}\left(\hat{\alpha}\right)$$

4.3 Confidence Intervals and Prediction Intervals

One of the main purposes in fitting a regression line (of Y on x, say) to data $(x_1, y_1), \ldots, (x_n, y_n)$ is to be able to say something about the values Y_0 of the response variable corresponding to any given value x_0 of the explanatory variable. Note that x_0 need not be one of $x_1, \ldots, x_n$. You may find it helpful to think of $y_1, \ldots, y_n$ as observations taken in the *past*, which we use to fit the sample regression line

$$\mathbb{E}(Y) = \hat{\alpha} + \hat{\beta}x.$$

Then, Y_0 is the random variable of a *future* observation of Y with $x = x_0$.
There are two types of interval which are of interest:

(i) confidence intervals for $\mathbb{E}(Y_0)$,

(ii) prediction intervals for Y_0.

Confidence intervals for $\mathbb{E}(Y_0)$

Here we consider the confidence interval for the fitted regression line, given the explanatory variable x_0. The regression equation for the response given the explanatory variable x_0 specifies,

$$\mathbb{E}(Y_0) = \alpha + \beta x_0.$$

Thus $\mathbb{E}(Y_0)$ is *fixed* and depends on the unknown parameters α and β. We estimate $\mathbb{E}(Y_0)$ by

$$\hat{\mathbb{E}}(Y_0) = \hat{\alpha} + \hat{\beta} x_0.$$

We know $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$, so that

$$\hat{\mathbb{E}}(Y_0) = \bar{Y} + \hat{\beta}(x_0 - \bar{x}).$$

It can be shown that $\bar{Y}$ and $\hat{\beta}$ are independent, and given the distribution of $\hat{\beta}$ from (8), it follows that

$$\begin{aligned}
\text{Var}\left[\hat{\mathbb{E}}(Y_0)\right] &= \text{Var}(\bar{Y}) + \text{Var}(\hat{\beta}(x_0 - \bar{x})) \qquad \text{since } \bar{Y} \text{ and } \hat{\beta} \text{ are independent} \\
&= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \text{Var}(\hat{\beta}) \\
&= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]
\end{aligned}$$

Finally, we note that $\hat{\mathbb{E}}(Y_0)$ is a linear function of normally distributed random variables, and so is also normally distributed. Estimating σ^2 by s^2 and standardizing, we therefore obtain

$$\frac{\hat{\mathbb{E}}(Y_0) - \mathbb{E}(Y_0)}{\sqrt{S^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

From which is follows that a 95% confidence interval for $\mathbb{E}(Y_0)$ is

$$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Note: The confidence interval for $\mathbb{E}(Y_0)$ become wider as x_0 moves away from $\bar{x}$. Thus estimates of $\mathbb{E}(Y_0)$ become less reliable, the further x_0 is from $\bar{x}$. This is one reason why it is unwise to extrapolate Y outside the range of $x_1, \ldots, x_n$.

Prediction intervals for Y_0

A prediction interval for Y_0 is a confidence interval for a (future) observation y_0 of Y_0, given the explanatory variable x_0 (rather than a confidence interval for the fitted regression line considered above). A prediction interval for Y_0 is centred on the fitted value $\hat{\alpha} + \hat{\beta} x_0$ corresponding to x_0. We can express a future estimate of Y_0 as

$$\hat{Y}_0 = \left(\hat{\alpha} + \hat{\beta} x_0\right) + \epsilon_0$$

for $\epsilon_0 \sim N(0, \sigma^2)$. The error ϵ_0 captures the additional randomness in predicting a single future observation, rather than the expected value, $\mathbb{E}(Y_0)$. Of course, we take $\hat{\epsilon}_0 = 0$, so that $\hat{Y}_0 = \hat{\mathbb{E}}(Y_0)$, but their variances differ. Since the ('future') random variable Y_0 is independent of the ('past') random variables $Y_1, \ldots, Y_n$ which are used to fit the model, ϵ_0 is independent of $\hat{\alpha}+\hat{\beta}x_0$, and so

$$\begin{aligned}
\text{Var}(\hat{Y}_0) &= \text{Var}\left(\left[\hat{\alpha}+\hat{\beta}x_0+\epsilon_0\right]\right) \\
&= \text{Var}\left(\hat{\alpha}+\hat{\beta}x_0\right)+\text{Var}(\epsilon_0) \\
&= \text{Var}(\hat{\mathbb{E}}(Y_0))+\text{Var}(\epsilon_0) \\
&= \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)+\sigma^2 \\
&= \sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right).
\end{aligned}$$

Finally we note that $\hat{Y}_0$ is a linear function of normally distributed random variables, and so is also normally distributed. Following the analogous steps as above we obtain a 95% prediction interval for Y_0 is

$$\hat{\alpha}+\hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}.$$

Notes:

- The prediction intervals for Y_0 become wider as x_0 moves away from $\bar{x}$.
- The prediction intervals for Y_0 are wider than the corresponding confidence intervals for $\mathbb{E}(Y_0)$. This is because the prediction interval takes into account the additional variability of the 'future' observation Y_0, whereas confidence intervals are concerned only with the mean value $\mathbb{E}(Y_0)$.

4.4 Multiple Regression

So far, we have considered the relationship between the response variable and a *single* explanatory variable. In many cases, we wish to relate the response variable (Y, say) to several explanatory variables ($x_1,\dots,x_k$, say – note that the subscript now indicates which of the explanatory variables we are considering, not the value taken by a single explanatory variable).
The appropriate generalisation of the linear regression model is the multiple linear regression model

$$\mathbb{E}(Y)=\alpha+\beta_1x_1+\cdots+\beta_kx_k.$$

The least-squares estimates $\hat{\alpha},\hat{\beta}_1,\dots,\hat{\beta}_k$ of $\alpha,\beta_1,\dots,\beta_k$ are defined as the values which minimise

$$S(\alpha,\beta_1,\dots,\beta_k)=\sum_{i=1}^{n}\left\{y_i-(\alpha+\beta_1x_{1i}+\cdots+\beta_kx_{ki})\right\}^2,$$

where y_i is the response at the value $(x_{1i},\dots,x_{ki})$ of the explanatory variables $(x_1,\dots,x_k)$. The estimates $\hat{\alpha},\hat{\beta}_1,\dots,\hat{\beta}_k$ can be found by matrix algebra.
Further, we can write,

$$Y_i \sim N(\alpha+\beta_1x_{1i}+\cdots+\beta_kx_{ki},\sigma^2), \qquad i=1,\dots,n \qquad \text{with } Y_1,\dots,Y_n \text{ independent.}$$

Equivalently, we can write, for $i=1,\dots,n$,

$$Y_i=\alpha+\beta_1x_{1i}+\cdots+\beta_kx_{ki}+\epsilon_i,$$

where $\epsilon_i \sim N(0,\sigma^2)$, independently.

The corresponding maximum likelihood estimates of $\alpha,\beta_1,\dots,\beta_k$ are the least-squares estimates $\hat{\alpha},\hat{\beta}_1,\dots,\hat{\beta}_k$. The sum of squares statistic S^2, defined by

$$S^2=\frac{1}{n-(k+1)}\sum_{i=1}^{n}\left\{Y_i-(\hat{\alpha}+\hat{\beta}_1x_{1i}+\cdots+\hat{\beta}_kx_{ki})\right\}^2,$$

is an unbiased estimator of σ^2. In R, the observed value $s = \sqrt{s^2}$ is called the *residual standard error* (recall we have already met this terminology for the simple linear regression case).
Confidence intervals for α and β_j, $j = 1, \ldots, k$, are obtained using the same approach as for the simple linear regression case, although the degrees of freedom of the relevant t distribution are now $n - (k+1)$. For example, for 95% confidence intervals,

$$\hat{\alpha} \pm t_{n-(k+1);0.025} s.e.(\hat{\alpha}),$$
$$\hat{\beta}_j \pm t_{n-(k+1);0.025} s.e.(\hat{\beta}_j),$$

for $j = 1, \ldots, k$. The values of $s.e.(\hat{\alpha})$ and $s.e.(\hat{\beta})$ can again be obtained using matrix algebra.
The R command lm can be used for multiple linear regression. The various explanatory variables are joined together by +, within the command - see the following example.

F-test for model comparison

In multiple regression it is common to test whether specified regression coefficients are zero (representing the intuitive idea that the given explanatory variables 'have no effect'). It is tempting to think that, in general, when testing that several regression coefficients are zero, we should use a sequence of t-tests to test each of the coefficients in turn. However, this makes the calculation of p-values very complicated. It is better to use ANOVA (ANalysis Of VAriance) - this is a general way of testing hypotheses which are formulated as nested linear models. 'Nested' means that the model corresponding to the null hypothesis is a special case of the model corresponding to the alternative hypothesis, being obtained from it by placing (linear) restrictions on the parameters.

In the context of multiple regression, using slightly revised notation, we can write:

$$Y_i \sim N(\beta_1 + \beta_2 x_{1i} + \cdots + \beta_{p_1} x_{(p_1-1)i}, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent}$$

and we are interested in a submodel in which $p_1 - p_0$ of the regression coefficients $\beta_1, \ldots, \beta_{p_1}$ are zero. Thus we wish to test the null hypothesis

$$H_0 : \text{the specified regression coefficients are zero}$$

against the alternative hypothesis

$$H_1 : \text{there is no restriction on the specified regression coefficients.}$$

The idea is to see whether or not the *full model* (without any restrictions on the regression coefficients) gives a *significantly* better fit to the data than the submodel does (specified by H_0 with restrictions on the regression coefficients). Of course, the full model always fits the data a little more closely than the submodel, since it has more parameters. However, if H_0 is true then the difference in fit between the models should be quite small, while a big difference in fit would tend to suggest that H_0 is false.
The goodness-of-fit of either model to the data (i.e. how well the model fits the observed data) is measured by the *residual sum of squares* (rss), which is the sum of squares of differences between the model and data. For example for the full model, the residual sum of squares is

$$\mathrm{rss}_1 = \sum_{i=1}^{n} \left\{ Y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 x_{1i} + \cdots + \hat{\beta}_{p_1} x_{(p_1-1)i} \right) \right\}^2 .$$

The residual sum of squares rss_0 for the submodel is defined similarly, but using only the least squares estimates (under H_0) of the regression coefficients used in the submodel.
It can be shown that, under either model,

$$\frac{\mathrm{rss}_1}{\sigma^2} \sim \chi^2_{n-p_1}.$$

(This is comparable to the result stated in §2.4 for the S^2).
However, only if H$_0$ is true, do we have:

$$\frac{\mathrm{rss}_0}{\sigma^2} \sim \chi^2_{n-p_0}; \qquad \frac{\mathrm{rss}_0 - \mathrm{rss}_1}{\sigma^2} \sim \chi^2_{p_1-p_0}; \text{ and}$$

$$\mathrm{rss}_0 - \mathrm{rss}_1 \text{ is independent of } \mathrm{rss}_1.$$

Hence

$$\frac{(\mathrm{rss}_0 - \mathrm{rss}_1)/(p_1 - p_0)}{\mathrm{rss}_1/(n - p_1)} \sim F_{p_1 - p_0, n - p_1} \qquad \text{under } H_0.$$

Recall: the ratio of two independent chi-squared distributions divided by their respective degrees of freedom leads to an F-distribution). If H_0 is false this statistic will tend to be too large for consistency with $F_{p_1-p_0,n-p_1}$ (a small value of the test statistic does not provide evidence against H_0).

Remark The name *Analysis of Variance* comes from the fact that it is based on the decomposition

$$\frac{\mathrm{rss}_0}{\sigma^2} = \frac{\mathrm{rss}_1}{\sigma^2} + \frac{\mathrm{rss}_0 - \mathrm{rss}_1}{\sigma^2}$$

which splits up the (scaled) variability rss_0/σ^2 of the data about the submodel into the sum of the (scaled) variability rss_1/σ^2 of the data about the full model and the (scaled) difference $(\mathrm{rss}_0 - \mathrm{rss}_1)/\sigma^2$ between the two models.

4.5 Exercises

1. Consider the straight line model $\mathbb{E}(Y) = \alpha + \beta x$. Only one of the following assumptions is required when estimating α and β, using a simple linear regression of Y on x. Which one?

   (a) The observations on x are independent.

   (b) The observations on x are independent of those on Y.

   (c) The observations on Y are independent.

   (d) The observations on Y are normally distributed.

2. Consider the model $\mathbb{E}(Y) = \alpha + \beta x$, where observations $Y \sim N(\alpha + \beta x, \sigma^2)$, and $s^2 = \hat{\sigma}^2$ is estimated in the standard way. Which of the following expressions provides the 95% confidence interval for the mean response $\hat{\mathbb{E}}(Y_0) = \hat{\alpha} + \hat{\beta} x_0$?

   (a)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm z_{0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (b)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm z_{0.025} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (c)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (d)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   where $z_{0.025}$ is the upper 2.5% quantile of the $N(0,1)$ distribution, and $t_{n-2;0.025}$ is the upper 2.5% quantile of the t distribution with $n-2$ degrees of freedom.

5 Analysis of Variance

In multiple regression it is common to test whether specified regression coefficients are zero (representing the intuitive idea that the given explanatory variables ”have no effect”). It is tempting to think that, in general, when testing that several regression coefficients are zero, we should use a sequence of t-tests to test each of the coefficients in turn. However, this makes the calculation of p-values very complicated. It is better to use ANOVA (Analysis Of Variance) This is a general way of testing hypotheses which are formulated as nested linear models. ”Nested” means that the model corresponding to the null hypothesis is a special case of the model corresponding to the alternative hypothesis, being obtained from it by placing (linear) restrictions on the parameters.

5.1 One-way ANOVA

A common problem is that of comparing several populations. The usual technique for doing this is one-way analysis of variance (which is a special case of the ANOVA introduced in §4.4).
Consider k distributions (or populations) with means $\mu_1, \ldots, \mu_k$, and suppose that we wish to test

$$H_0 : \mu_1 = \cdots = \mu_k$$

against

$$H_1 : \mu_1, \ldots, \mu_k \text{ are not all equal.}$$

Notes:

- For the general case the alternative hypothesis is *not* ‘$H_1 : \mu_1, \ldots, \mu_k$ are all different’. Thus, if $k = 3$ and $\mu_1 = \mu_2 \neq \mu_3$, then H_1 would still hold.

Suppose that we have random samples of sizes $n_1, \ldots, n_k$, respectively, from the k distributions. Let y_{ij} denote the jth observation on the ith distribution, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. We think of y_{ij} as an observation on a random variable Y_{ij}.
We shall assume that

$$Y_{ij} \sim N(\mu_i, \sigma^2), \qquad i = 1, \ldots, k;\ j = 1, \ldots, n_i \quad \text{independently for all } Y_{ij}. \tag{14}$$

With its assumptions of independent normal distributions with the same variance, this model should be reminiscent of the simple linear regression model (7). It is actually a special case of the multiple regression model from §4.4. This can be seen as follows.
Define indicator variables $x_1, \ldots, x_k$ by

$$x_i = \begin{cases} 1 & \text{if the observation is from } i\text{th distribution,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the condition

$$\mathbb{E}(Y_{ij}) = \mu_i$$

and the mean of Y_{ij} can be written as

$$\mathbb{E}(Y_{ij}) = \mu_1 x_1 + \cdots + \mu_k x_k, \tag{15}$$

which is the multiple regression model with no constant term α. We can now apply the general results of §4.4. The full model (15) has k parameters, i.e. $p_1 = k$. The submodel given by H_0 is

$$\mathbb{E}(Y_{ij}) = \mu, \tag{16}$$

and so has one parameter, i.e. $p_0 = 1$. There are n observations, with

$$n = n_1 + \cdots + n_k.$$

We consider the test statistic:

$$F = \frac{(\mathrm{rss}_0 - \mathrm{rss}_1)/(k-1)}{\mathrm{rss}_1/(n-k)},$$

where rss_0 and rss_1 denote the residual sums of squares under the null model (16) and the alternative (full) model (15), respectively.

Then under H_0 (§4.4) gives

$$F = \frac{(\mathrm{rss}_0 - \mathrm{rss}_1)/(k-1)}{\mathrm{rss}_1/(n-k)} \sim F_{k-1,n-k}, \tag{17}$$

If H_0 is false, then the above statistic will tend to be too large for consistency with $F_{k-1,n-k}$. Small values are not inconsistent with H_0. Thus (only) large values of the statistic lead to rejection of H_0.

In one-way ANOVA, the residual sums of squares rss_0 and rss_1 have nice expressions, which we now derive. Let $\bar{Y}$ and $\bar{Y}_{i.}$ (for $i = 1, \dots, k$) denote the overall sample mean (of all n observations) and the sample mean of the ith sample, respectively. Mathematically,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad \bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

(i) The *total sum of squares* is

$$SS_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2$$

and represents the variability of all the observations about their overall mean.

(ii) The *between groups sum of squares* is

$$SS_B = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_{i.} - \bar{Y}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{Y}_{i.} - \bar{Y}\right)^2$$

and represents the variability between the k sample means.

(iii) The *within groups sum of squares* is

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

and represents the variability of the observations about their respective group means.

After some algebraic manipulation is can be shown that:

$$SS_{Tot} = SS_B + SS_W,$$

Further calculation shows that the maximum likelihood estimates (which are also the least squares estimates) $\hat{\mu}_1, \dots, \hat{\mu}_k$ and $\hat{\mu}$ of $\mu_1, \dots, \mu_k$ and μ in (15) and (16), respectively, are

$$\begin{aligned} \hat{\mu}_i &= \bar{y}_{i.} \qquad i = 1, \dots, k \\ \hat{\mu} &= \bar{y} \end{aligned}$$

and that

$$\begin{aligned} \mathrm{rss}_0 &= SS_{Tot} \\ \mathrm{rss}_1 &= SS_W \\ \mathrm{rss}_0 - \mathrm{rss}_1 &= SS_B. \end{aligned}$$

Thus the left hand side of (17) can be expressed in terms of SS_B and SS_W. It is useful to express the numerator and denominator of the left hand side of (17) as *mean squares*.

(i) The *between groups mean square* is

$$MS_B = \frac{SS_B}{k-1};$$

(ii) The *within groups mean square* is

$$MS_W = \frac{SS_W}{n-k}.$$

Then the left hand side of (17) is

$$F = \frac{MS_B}{MS_W} \tag{18}$$

and if H_0 is true,

$$F \sim F_{k-1,n-k}.$$

Many statistical packages (including R) set out the observed sums of squares, mean squares, etc. in an *ANOVA table*.

Source	d.f.	SS	MS	F	p
Between	$k-1$	SS_B	MS_B	F	p
Error	$n-k$	SS_W	MS_W		
Total	$n-1$	SS_{Tot}			

F is given by (18), d.f. denotes the number of degrees of freedom and p the associated p-value of F.
An important property of the within groups mean square MS_W is that it is an unbiased estimator of the variance σ^2 in (14). The observed *residual mean square* is

$$S^2 = MS_W, \tag{19}$$

with S often called the *residual standard error*.

Least Significant Differences

If the null hypothesis is rejected then the one-way ANOVA has told us only that $\mu_1, \dots, \mu_k$ are not all equal; it has not indicated which group means are significantly different. To do the latter, it is usual to compare the groups in pairs, using t-tests. If $\mu_i = \mu_j$ then we have

(i) The distributional result:

$$\bar{Y}_{i.} - \bar{Y}_{j.} \sim N\left(0, \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right) \quad \Rightarrow \quad \frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{\sqrt{\sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim N(0,1);$$

(ii) The distributional result:

$$S^2(n-k)/\sigma^2 \sim \chi^2_{n-k}$$

where $S^2 = MS_W$; and

(iii) The sample variance S^2 ($= rss_1/(n-k) = MS_W$, as in (19)) is independent of $\bar{Y}_i - \bar{Y}_j$.

Combining all of the above (as in §2.4) we have the result,

$$\frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{\sqrt{S^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{n-k}.$$

Note the differences between this and the 2-sample t-statistic. Here,

(i) S^2 is a function of all the groups (not just groups i and j),

(ii) the t-statistic has $n-k$ degrees of freedom (rather than $n_i + n_j - 2$).

If the above t-statistic is significant (i.e. has large absolute value), then we have evidence that $\mu_i \neq \mu_j$. The significant pairwise comparisons should be used *only as a guide* to which means differ from which, rather than as rigorous tests. By rearranging the above formulae we can work out the minimum observed difference $(\bar{y}_{i.} - \bar{y}_{j.})$ for which the difference in sample means would be significant (and hence would lead to reject the hypothesis of equal means). For a significance level of α this is given by:

$$t_{n-k;\alpha/2}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

This difference is often referred to as the *least significant difference (LSD)* between groups i and j.
If $n_1 = \cdots = n_k$ (i.e. the samples are of equal size), then it is particularly worth calculating the smallest difference in sample means that would lead to rejection of the null hypothesis that two groups have equal population means. This least significant difference (LSD) is then equal for all pairs of groups. Thus it is easy to look for the pairs of groups with sample means differing by more than the LSD. If there are k groups, each with m observations (so that $n = mk$), then the LSD for significance level α is

$$t_{n-k;\alpha/2}\sqrt{\frac{2s^2}{m}}.$$

5.2 Two-way ANOVA

In the previous one-way ANOVA, each observation belonged to just a single group. However, often observations may be cross-classified across two (or more) groups. We will consider the case where each observation is classified according to different two groups, leading to two-way ANOVA. When there is only one observation recorded for each combination of the two groups this is often referred to as a randomised block experiment (this term is often sometimes used when there are multiple replicates, but this is less common).

Two-way ANOVA with no replications

We initially consider the case where there is a single observation for each combination of groups. Notationally assume that the two groups are labelled “Block” and “Treatment”, such that there are a total of b blocks and k treatments. Let y_{ij} denote the observation corresponding to block $i = 1, \ldots, b$ and treatment $j = 1, \ldots, k$. We initially assume that there is only one observation per block and treatment combination - we remove this assumption later. Let $n = bk$ which corresponds to the total number of observations. We assume that each each observed value is an independent observation from the random variable $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ where $\mu_{ij} \equiv \mathbb{E}(Y_{ij})$. We let $= \{\mu_{ij} : i = 1, \ldots, b; j = 1, \ldots, k\}$ and $= \{y_{ij} : i = 1, \ldots, b; j = 1, \ldots, k\}$. Now there are several models that might be of interest, with regard to the mean. One of the most common models of interest is:

$$\mu_{ij} = \alpha_i + \beta_j.$$

This model assumes that the mean is the linear sum of an effect for each group - we will call this model 3.
However, we need to look at this model in a bit more detail. Consider the simplest example we can have where $b = 2$ and $k = 2$, so a 2×2 block design, with parameters α_1, α_2, β_1 and β_2. We have that,

$$\mu_{ij} = \alpha_i + \beta_j.$$

Now consider setting $\alpha_1^* = \alpha_1 + 10$, $\alpha_2^* = \alpha_2 + 10$, $\beta_1^* = \beta_1 - 10$, $\beta_2^* = \beta_2 - 10$. We then have that,

$$\mu_{ij}^* = \alpha_i^* + \beta_j^* = \mu_{ij}.$$

In other words there is not a unique solution of μ_{ij} in terms of α_i and β_j, as we can change all the values of the parameters of the model (to the starred parameter values) but would leave all the modelled μ_{ij} values unchanged. It is the μ_{ij} values that the likelihood is a function of so that this means that the model as written does not have a unique set of maximum likelihood estimates of the parameters. This problem is most easily eliminated by simply setting one of the parameters to zero (e.g. set $\alpha_1 = 0$ or $\beta_2 = 0$), and the problem disappears, while the resulting reduced model with the given constraint still fits the data just as well as the original model. The parameter which we set to 0 is arbitrary.

There are two further sub-models (or reductions of this model) that may be of particular interest:

$$\mu_{ij} = \alpha_i$$

which corresponds to the mean being a function of only Factor 1. Call this model 2.

Alternatively we may also be interested in the sub-model:

$$\mu_{ij} = \beta_j,$$

corresponding to the mean being a function of only Factor 2. Call this model 1.
Thus there are two possible hypothesis that we may wish to consider.

Test 1:

$$H_0 : \alpha_1 = \cdots = \alpha_b \qquad \text{(Model 1)}$$

against

$$H_1 : \alpha_1, \ldots, \alpha_b \text{ are not all equal} \qquad \text{(Model 3)}.$$

This corresponds to testing whether or not there is a block effect on the mean.

Test 2:

$$H_0 : \beta_1 = \cdots = \beta_k \qquad \text{(Model 2)}$$

against

$$H_1 : \beta_1, \ldots, \beta_k \text{ are not all equal} \qquad \text{(Model 3)}.$$

This corresponds to testing whether or not there is a treatment effect on the mean.

The analysis of variance for conducting these hypothesis test (or equivalently for comparing these models) may be expressed in the following ANOVA table:

Source	d.f.	SS	MS	F	p
Blocks	$b-1$	SS_B	MS_B	F_B	p_B
Treatment	$k-1$	SS_T	MS_T	F_T	p_T
Error	$(b-1)(k-1)$	SS_W	MS_W		
Total	$bk-1$	SS_{Tot}			

The elements of the table are analogous to those from the one-way ANOVA. For example, SS terms corresponds to the sum of squares (between blocks, between treatments; within error; and total); MS terms to the mean squares; the F terms to the test statistic and the p terms the corresponding p-values for the above hypotheses.

Mathematically, let $\bar{Y}_{i\cdot}$ denote the sample mean for block $i = 1, \ldots, b$, so that $\bar{Y}_{i\cdot} = \frac{1}{k}\sum_{j=1}^{k} Y_{ij}$. Similarly let $\bar{Y}_{\cdot j}$ denote the sample mean for treatment $j = 1, \ldots, k$, so that $\bar{Y}_{\cdot j} = \frac{1}{b}\sum_{i=1}^{b} Y_{ij}$. Finally let $\bar{Y}$ denote the overall mean, so that $\bar{Y} = \frac{1}{bk}\sum_{i=1}^{b}\sum_{j=1}^{k} Y_{ij}$.
The sum of squares are then given by:

(i) The *total sum of squares* is

$$SS_{Tot} = \sum_{i=1}^{b}\sum_{j=1}^{k}\left(Y_{ij} - \bar{Y}\right)^2$$

and represents the variability of all the observations about their overall mean.

(ii) The *between blocks sum of squares* is

$$SS_B = k \sum_{i=1}^{b} \left(\bar{Y}_{i\cdot} - \bar{Y}\right)^2$$

and represents the variability between the sample means for the blocks, with $b - 1$ degrees of freedom.

(iii) The *between treatment sum of squares* is

$$SS_T = b \sum_{j=1}^{k} \left(\bar{Y}_{\cdot j} - \bar{Y}\right)^2$$

and represents the variability between sample means for the treatments, with $k - 1$ degrees of freedom.

(iv) The *within groups sum of squares* is

$$SS_W = \sum_{i=1}^{b} \sum_{j=1}^{k} \left(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}\right)^2$$

and represents the variability of the observations about their respective group means, with $(b - 1)(k - 1)$ degrees of freedom.

It can again be shown that:

$$SS_{Tot} = SS_B + SS_T + SS_W.$$

The associated mean squares are equal to the sum of squares divided by the associated degrees of freedom so that:

$$\begin{aligned} MS_B &= \frac{SS_B}{b-1} \\ MS_T &= \frac{SS_T}{k-1} \\ MS_W &= \frac{SS_W}{(b-1)(k-1)}. \end{aligned}$$

The test statistics are given by:

$$\begin{aligned} F_B &= \frac{MS_B}{MS_W} \\ F_T &= \frac{MS_T}{MS_W}, \end{aligned}$$

for Tests 1 and 2, respectively.

Consider Test 1 (corresponding to whether or not there is a block effect), then if H_0 is true we have that,

$$F_B \sim F_{(b-1);(b-1)(k-1)}.$$

The corresponding p-value is denoted by p_B.

Consider Test 2 (corresponding to whether or not there is a treatment effect), then if H_0 is true we have that,

$$F_T \sim F_{(k-1);(b-1)(k-1)}.$$

The corresponding p-value is denoted by p_T.

Recall:

- There is only evidence against H_0 for large values of the test statistic (small test statistics do not provide evidence against H_0 in favour of H_1), so that this is a one-tailed test;
- We have that $MS_W = S^2$ and is an unbiased estimator of σ^2.

Two-way ANOVA with multiple replications

The above approach can be extended such that there are multiple observations for each block and treatment combination. This leads to a similar ANOVA table. We omit the repetition of the mathematical details here - but note that the the ANOVA table that is obtained stays the same except that the degrees of freedom of the residuals increases to take into account the additional number of observations. Suppose that there are r replications for each combination of group and treatment. The associated ANOVA table is given by:

Source	d.f.	SS	MS	F	p
Blocks	$b-1$	SS_B	MS_B	F_B	p_B
Treatment	$k-1$	SS_T	MS_T	F_T	p_T
Error	$rbk-b-k+1$	SS_W	MS_W		
Total	$rbk-1$	SS_{Tot}			

LSD's for two-way ANOVA

If some effects turn out to be significant in a two-way ANOVA (as in the above example), then it is of some interest to know where the major differences lie. For example if school has an effect, it is a good idea to figure out which schools contribute most to this result. By similar reasoning to that employed in the one-way ANOVA case, the least significant differences can be obtained. We initially consider the case where there is only one observation for each combination of group and treatment, before the case with multiple replications.

No replications

Consider the block effects. We can once again derive the following distributional results:

$$\bar{Y}_{i.} - \bar{Y}_{j.} \sim N\left(\alpha_i - \alpha_j, 2\sigma^2\left(\frac{1}{k}\right)\right) \quad \Rightarrow \quad \frac{\bar{Y}_{i.} - \bar{Y}_{j.} - (\alpha_i - \alpha_j)}{\sqrt{2\frac{\sigma^2}{k}}} \sim N(0,1);$$

and,

$$(b-1)(k-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(b-1)(k-1)}$$

independently of each other.
Now, if H_0 is true and there is no block effect, $\alpha_i = \alpha_j$, and following the usual steps, we have that,

$$\frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{\sqrt{2\frac{S^2}{k}}} \sim t_{(b-1)(k-1)}$$

Thus for the block effect the least significant difference for the test at significance level α is,

$$t_{(b-1)(k-1);\alpha/2}\sqrt{2\frac{s^2}{k}}.$$

Similarly for the treatment effects we obtain the associated least significant difference for the test at significance level α to be:

$$t_{(b-1)(k-1);\alpha/2}\sqrt{2\frac{s^2}{b}}.$$

Multiple replications

When there are multiple replicates we need to take these into account within the least significant differences. Recall that we assume that there are r replicates per group and treatment combination. Following the same steps as above we can derive the following least significant differences at significance level α for the blocks:

$$t_{rbk-b-k+1;\alpha/2}\sqrt{2\frac{s^2}{rk}}.$$

Similarly for the treatments:

$$t_{rbk-b-k+1;\alpha/2}\sqrt{2\frac{s^2}{rb}}.$$

5.3 Exercises

1. The following analysis was conducted in R.

```
> data <- c(21,25,19,17,20,18,22,16,18,29,26,24)
> treatment <- c(rep(1,3),rep(2,3),rep(3,3),rep(4,3))
> analysis <- lm(data ˜ as.factor(treatment))
> anova(analysis)

Analysis of Variance Table

Response: data
                     Df  Sum Sq Mean Sq F value  Pr(>F)
as.factor(treatment)  3 123.583  41.194  6.0285 0.01890 *
Residuals             8  54.667   6.833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   Which of the following best describes the analysis?

   (a) A regression of `data` on `treatment`.

   (b) A regression of `treatment` on `data`.

   (c) A one-way analysis of variance to test whether the effect of the three treatments all have the same mean.

   (d) A one-way analysis of variance to test whether the effect of the four treatments all have the same mean.

2. Which of the following statements is the most reasonable interpretation of the analysis of the previous question?

   (a) We can accept the null hypothesis that the treatments do not differ.

   (b) There is no evidence to suggest that the treatments differ.

   (c) There is evidence that the treatments differ.

   (d) The treatments are not all the same.

3. The effect of vitamin C on tooth growth in guinea pigs has been studied. The response variable is the length of odontoblasts (cells responsible for tooth growth) of 36 guinea pigs. Each animal received one of three dose levels of vitamin C (dose level 1 = 0.5, dose level 2 = 1, and dose level 3 = 2 mg/day) by one of two delivery methods (method 1 = orange juice; method 2 = ascorbic acid). Each possible combination of dose and treatment were given to six different guinea pigs. The following analyses were conducted in `R`:

```
length <- c(4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 16.5, 16.5, 15.2, 17.3, 22.5,
17.3, 23.6, 18.5, 33.9, 25.5, 26.4, 32.5, 15.2, 21.5, 17.6,
9.7, 14.5, 10.0, 19.7, 23.3, 23.6, 26.4, 20.0, 25.2, 25.5,
26.4, 22.4, 24.5, 24.8, 30.9)
dose <- c(rep(1,6), rep(2,6), rep(3,6),rep(1,6), rep(2,6), rep(3,6))
delivery <- c(rep(1,18),rep(2,18))

> model <- lm(length ˜ as.factor(delivery)+as.factor(dose))
> anova(model)
```

```
Analysis of Variance Table
---
Response: length
Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(delivery)  1  137.28  137.28  8.3678 0.006817 **
as.factor(dose)      2 1388.54  694.27 42.3184 1.03e-09 ***
Residuals           32  524.99   16.41
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) What is the name of the analysis being conducted? State the corresponding model assumptions.

(b) State the null and alternative hypotheses for the two tests that have been conducted in R.

(c) State your conclusions regarding whether different doses and/or different delivery methods affect the tooth growth at the 5% significance level? Explain your answer.

(d) Calculate the least significant differences for the delivery methods and dose levels at significance level $\alpha = 0.05$, as appropriate. The following R output may be of use:

```
> qt(0.975, 1)
[1] 12.7062
> qt(0.975, 2)
[1] 4.302653
> qt(0.975, 31)
[1] 2.039513
> qt(0.975, 32)
[1] 2.036933
```

(e) Conclude which delivery methods and/or dose levels differ from each other in terms of mean length of odontoblasts. The following R output may be of use:

```
> c(mean(length[delivery==1]), mean(length[delivery==2]))
[1] 17.27222 21.17778
> c(mean(length[dose==1]), mean(length[dose==2]),
mean(length[dose==3]))
[1] 11.14167 20.29167 26.24167
```

(f) If there is evidence that there is a difference between the delivery method and/or dose with regard to tooth growth state which delivery method and/or dose should be recommended for a quicker tooth growth?

6 Exercise Solutions

Solutions for Exercises 2.5

1. (c) is true. In general we cannot say whether consistent estimators are biased or unbiased estimators so that (a) and (b) are not always true - consistent estimators are unbiased in the limiting case as $n \to \infty$. (d) is not true as Var$(T) \to 0$ (not θ) as $n \to \infty$.

2. (d) is false. Consider,

$$\begin{aligned}\mathbb{E}\left(\frac{X}{n}\right)^2 &= \frac{1}{n^2}\mathbb{E}(X^2) \\ &= \frac{1}{n^2}(\text{Var}(X) + (\mathbb{E}(X))^2) \\ &= \frac{1}{n^2}(np(1-p) + (np)^2) \\ &\neq p^2.\end{aligned}$$

Note that (a) is true (see Example 2.3 - result follows similarly for general n); (b) is true (see example 2.2); (c) is true due to the invariance property of MLEs (see §2.3) and as (b) is true).

3. (a) Let X_i denote the random variable associated with the ith trial. We know that the first $y-1$ trials are all failures, and the final trial is successful, so that $X_1 = X_2 = \cdots = X_{y-1} = 0$ and $X_y = 1$. Since the trials are independent of each other, we have that for $y = 1, \dots,$

$$\begin{aligned}f(y) &= \mathbb{P}(Y = y) = \mathbb{P}(X_1 = 0, X_2 = 0, \dots, X_{y-1} = 0, X_y = 1) \\ &= \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0)\dots,\mathbb{P}(X_{y-1} = 0)\mathbb{P}(X_y = 1) \\ &\qquad\qquad\qquad\qquad \text{(since the } X_i\text{s are independent)} \\ &= (1-p)^{y-1}p = pq^{y-1}.\end{aligned}$$

(b) The likelihood for a single observation y is given by $L(p; y) = p(1-p)^{y-1}$. The corresponding log-likelihood is given by

$$l(p; y) = \log L(p; y) = \log p + (y-1)\log(1-p).$$

To find the MLE:

$$\begin{aligned}\frac{\partial l}{\partial p} &= 0 \\ \Rightarrow \quad \frac{1}{p} - \frac{y-1}{1-p} &= 0 \\ \Rightarrow \quad \frac{1}{\hat{p}} &= \frac{y-1}{1-\hat{p}} \\ \Rightarrow \quad \hat{p} &= 1/y.\end{aligned}$$

The second derivative is negative, confirming that this is a maxima.

(c) By the invariance property of MLE's (see §2.3), the MLE of the log odds is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log\left(\frac{1}{y-1}\right).$$

(Note that this is infinite if $y = 1$).

Solutions for Exercises 3.6

1. We have that,

$$\mathbb{P}\left(5 \leq \frac{3T}{\mu} \leq 15\right) = 0.90 \Rightarrow \mathbb{P}\left(\frac{3T}{15} \leq \mu \leq \frac{3T}{5}\right) = 0.90$$

   So, a 90% CI is, $\left(\frac{3T}{15}, \frac{3T}{5}\right)$. For $t = 5$, a 90% CI is $(1, 3)$.

   Note: probabilistic statements of the type $\mathbb{P}\left(\frac{3T}{15} \leq \mu \leq \frac{3T}{5}\right) = 0.90$ are valid, since T is part of the probability statement and T is a random quantity. Once T is replaced by $t = 5$, it is *wrong* to write something such as, $\mathbb{P}\left(\frac{3\times5}{15} \leq \mu \leq \frac{3\times5}{5}\right) = 0.90$, as there is no random quantity in the probability statement - μ is a fixed value not a random variable.

2. (c) is correct. (b) and (d) are direct probabilistic statements for θ, something we are not allowed to do since θ is not a random quantity. (a) is not always true because different experiments would, in general, produce different data sets and, consequently, different confidence intervals.

3. (a) We have that $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$ since:

   (i) a linear combination of Normally distributed random variables is also normally distributed;

   (ii) $\mathbb{E}\left(\sum_{i=1}^n X_i\right) = n\mu$; and

   (iii) Var$(\sum_{i=1}^n X_i) = n\sigma^2$ since the X_i's are independent.

   (b) We have that $\mathbb{P}\left(\sum_{i=1}^n X_i > n\mu\right) = 0.5$ since the distribution is symmetrical about its mean $n\mu$ (or use the standard method of converting to standard normal distribution and using R).

   (c) We have that $\frac{X_i - \mu}{\sigma} \sim N(0, 1)$, independently for each i, so that,

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

   (d) We have that,

$$\begin{aligned} \mathbb{P}\left(\chi^2_{n:1-\alpha/2} \leq \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \leq \chi^2_{n:\alpha/2}\right) &= 1 - \alpha \\ \Rightarrow \quad \mathbb{P}\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:\alpha/2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:1-\alpha/2}}\right) &= 1 - \alpha. \end{aligned}$$

   Thus,

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:\alpha/2}}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n:1-\alpha/2}}\right)$$

   is a $100(1 - \alpha)$% confidence interval for σ^2.

   (e) We have $\sum_{i=1}^{10} x_i^2 = 169.62$ and $\sum_{i=1}^{10} x_i = 40$, so that $\sum_{i=1}^{10}(x_i - \mu)^2 = \sum_{i=1}^{10}(x_i^2) - 2\sum_{i=1}^{10}(x_i)\mu + 10\mu^2 = 9.62$. In addition, to find the lower 2.5% quantile, in R use the command:

```
> qchisq(0.025,10)
[1] 3.246973
```

   Similarly for the upper 2.5% quantile, in R use:

```
> qchisq(0.975,10)
[1] 20.48318
```

   Substituting into the expression for the confidence interval in (a) above, we obtain a 95% CI of (0.4697, 2.9627).

Solutions for Exercises 4.5

1. (c) is the only necessary assumption. We do not need to make any assumptions on the explanatory variable x. We do not need to make a distributional assumption on the response, Y to fit the simple linear model (we can use least squares).

2. (c) is the correct answer.

Solutions for Exercises 5.3

1. (d)

2. (c)

3. (a) Two-way analysis of variance (with multiple replications). Assumptions are that the observations:
   - are from a normal distribution;
   - are independent;
   - have a common variance;
   - have a mean that is an additive function of the treatment and the block.
   The assumptions can be expressed as: $Y_i \overset{iid}{\sim} N(\mu_{ij}, \sigma^2)$, where $\mu_{ij} = \alpha_i + \beta_j$.

   (b) Hypothesis 1: $H_0 : \alpha_i = \alpha$ for all i, vs H_1: not all α_i's are equal.
   Hypothesis 2: $H_0 : \beta_j = \beta$ for all j, vs H_1: not all β_j's are equal.

   (c) For both tests, p-values are smaller than 0.05. Thus we would reject each individual null hypothesis in favour of the alternative hypothesis (assuming that the other factor is present in the model). So there is a delivery method and dose level effect, meaning that the mean of response variable is different at different levels of delivery and dose.

   (d) LSD at the 5% significance level for delivery method is given by,

$$t_{32;0.025}\sqrt{2 \times \frac{16.41}{18}} = 2.750489.$$

   LSD at the 5% significance level for dose factor is given by,

$$t_{32;0.025}\sqrt{2 \times \frac{16.41}{12}} = 3.368647.$$

   (e) Mean of delivery method 2 - mean of delivery method 1 = 3.90556,
   mean of dose level 2 - mean of dose level 1 = 9.15,
   mean of dose level 3 - mean of dose level 1 = 15.1,
   mean of dose level 3 - mean of dose level 2 = 5.95.
   Therefore, $\alpha_2 > \alpha_1$ and $\beta_2 > \beta_1$, $\beta_3 > \beta_1$, $\beta_3 > \beta_2$.

   Note that there are only 2 delivery methods and as we have concluded that there is a difference in delivery method we do not need to work out which delivery methods differ as methods 1 and 2 must differ!

   (f) Mean of tooth length is bigger for delivery method 2 and dose level 3.