Applications of Computational Statistics

Application to biological problems

Programmation in 

Kim-Anh Le Cao & Peter Bailey
University of Queensland

Contents

Chapter 1

Data description

The field of statistics can be divided into two major branches: *descriptive* statistics and *inference* statistics. In both branches, we work with a set of measurements, and we need to organize, summarize and describe the data. Whether we are describing an observed population or using sampled data to draw an inference from the sample to the population, an insightful description of the data is an important step in drawing conclusions from it.
In this Chapter, we will review both graphical techniques and numerical descriptive techniques as two major methods for describing a set of measurements.

1.1 Visualising distributions

We will first start by describing the variability present in a single quantitative variable, which pattern we see is called the *distribution* of the variable. Three main aspects of the distribution can be described:

1. The location or centre of the variability
2. The spread of the variability
3. The shape of the variability

Once these patterns have been described, we can look for values that do not match this general description, for example *outliers* are values that do not match the rest of the pattern or *bimodal* distributions where there are two distributions of variability in our values.

1.1.1 Histogram

Histograms are appropriate for displaying frequency data for *quantitative* variables. A histogram displays the distribution of the data (its shape) as it is an estimate of the probability distribution of a continuous variable. The values need to be broken into a number of *bins* or *classes*. A histogram is obtained by drawing rectangles whose bases are the bins intervals and whose heights are in the counts in each bin.
We examine the overall shape in the histograms, often to compare different populations or samples. It actually helps to choose the appropriate measures to summarize the data (see following Sections). Common shapes (distributions) include:

- Unimodal / Bimodal
- Uniform
- Symmetric / left or right skewed

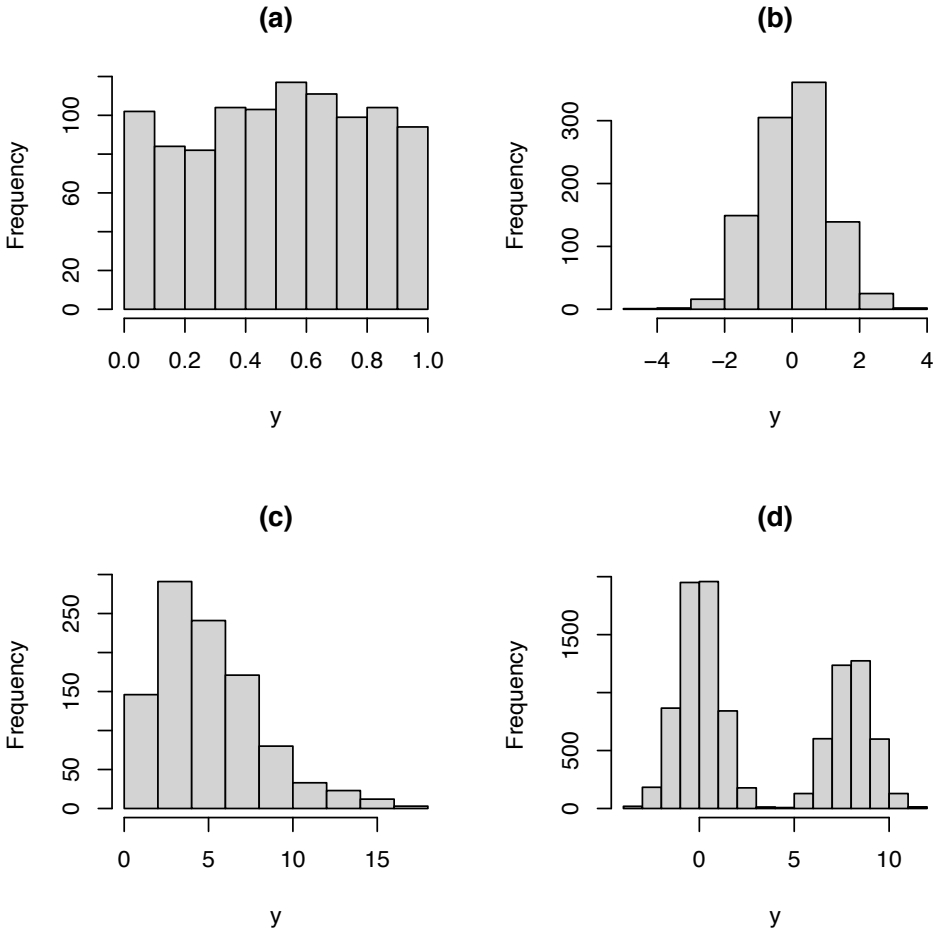

Figure 1.1: Some common shapes of distributions.

Exercise 1 *Characterise the different shapes of distribution from Figure 1.1.*

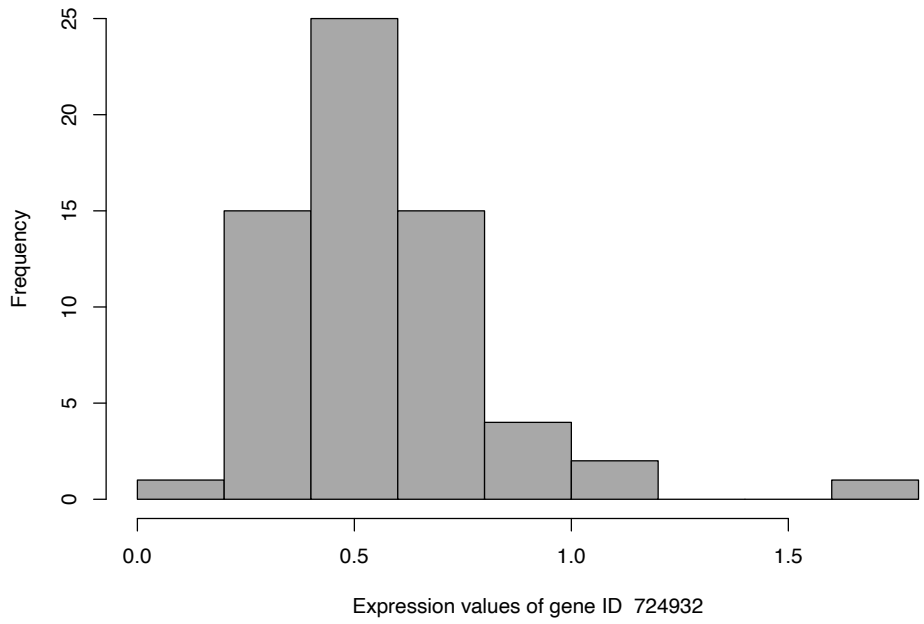

Figure 1.2: SRBCT gene expression data. Frequency histogram of the expression values of one gene.

Exercise 2 *The microarray study of Khan et al. 2001*[1] *measures the gene expression from 63 patients affected by Small Round Blue-Cell Tumors (SRBCT). The diagnostic for each of these patients is grouped into four different categories: EWS (Ewing family of tumours, $n_{EWS} = 23$), BL (Burkitt Lymphoma $n_{BL} = 8$), NB (Neuroblastoma, $n_{NB} = 12$) and RMS (Rhabdomyosarcoma, $n_{RMS} = 20$).*
We focus on the expression of one gene, which is a G protein-coupled receptor kinase 6 (gene ID: 724932), across the 63 samples. Discuss the distribution of the expression value of this gene displayed in Fig. 1.2.

Remark 1 *The shape of the histogram can be strongly affected by the number of bins used (argument* `breaks` *in the* `hist()` *function, set to a default algorithm). One case where the bins are predetermined is with discrete data (e.g. age of patients).*

Remark 2 *By default the* `R` *function* `hist()` *displays the frequency of the data (i.e. the counts in each bin), but the relative frequency is also available. The relative frequency histogram has a total area of one, and instead of using raw counts, represents the proportion of counts in each bin. For example Fig. 1.3 show the frequency and the relative frequency histogram of the same data: the same shape is observed but the y-axis is different.*

[1] Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine* **7** 6

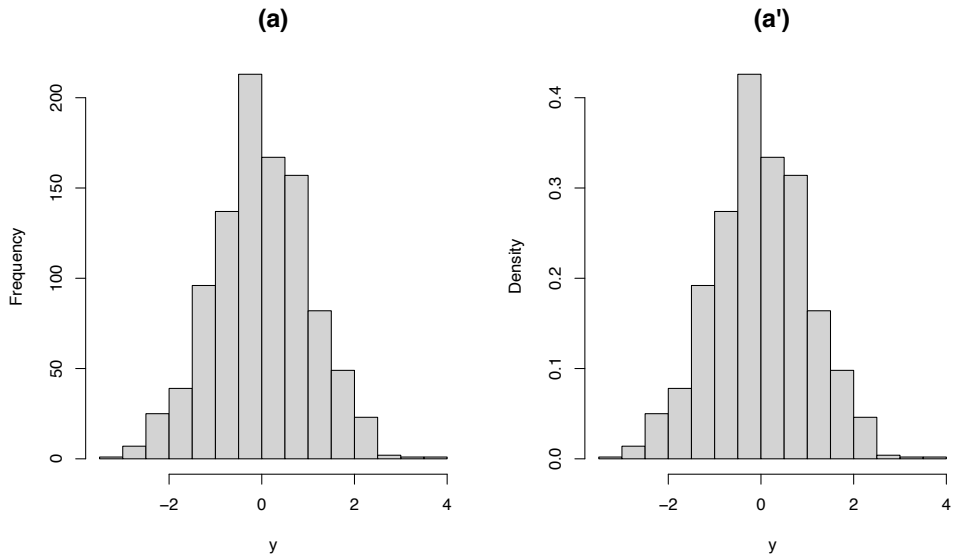

Figure 1.3: Frequency (a) and relative frequency (also called *density*) (a') histograms for the same data.

1.1.2 Density plots

The histogram is an example of a plot that estimates the *density* of a distribution (although not smoothed and highly dependent on the width of the bins). Kernel density estimators put a little lump (*kernel* of the density estimate) centered on each data value and then add all their densities together to get the final curve (Figure 1.4).

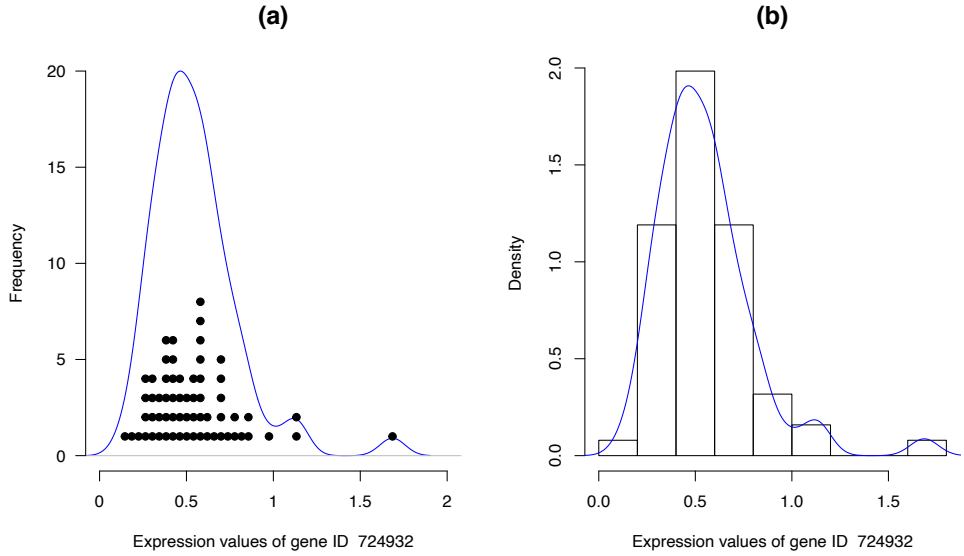

Figure 1.4: SRBCT gene expression data. Density curve with dotplot (a) and histogram of counts (b) for the expression values of one gene.

Density plots are very useful to compare multiple distributions in a single plot (Fig. 1.4).

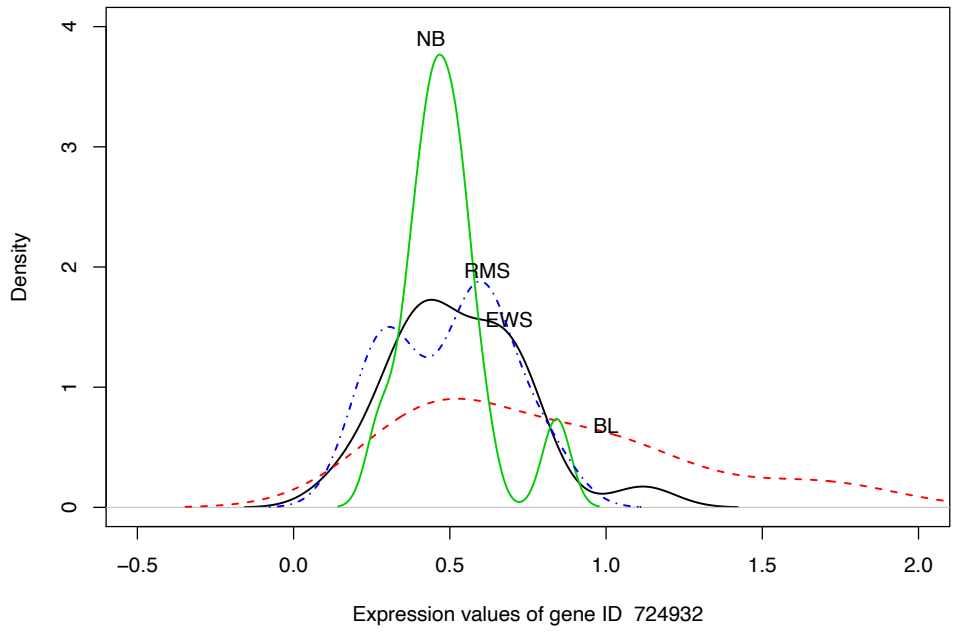

Figure 1.5: SRBCT gene expression data. Density curves with respect to each tumour category for the expression values of one gene.

Exercise 3 *Comment on the density plots obtained in Figure 1.5.*

1.2 Measures of central tendency and variability

Numerical descriptive measures enable statistical inference as graphical descriptive measures are inappropriate for that purpose. Common numerical descriptive measures are measures of central tendency (describe the center of the distribution of measurements) and measures of variabilities (how the measurements vary about the center of the distributions).

Remark 3 *There is a distinction between numerical descriptive measures for a population, called* ***parameters*** *and numerical descriptive measures for a sample, called* ***statistics****. In problems requiring statistical inference, we will not be able to calculate values for various parameters, but we will be able to compute corresponding statistics from the sample and use these quantities to estimate the corresponding population parameters.*

1.2.1 Measures of central tendency

Definition 1 *The **mode** of a set of measurements is defined to be the measurement that occurs most often (with the highest frequency).*

The mode can be applied to both qualitative and quantitative data. We can also encounter distributions with more than one measurement that occurs at the highest frequency (bimodal, trimodal ..).
Exercise 4 *What is the mode of histogram from Figure 1.2?*

Definition 2 *The **median** of a set of measurements is defined to be the middle value when the measurements are arranged from lowest to highest.*

The median reflects the central value of the data. The median for an even number of measurements is the average of the two middle values when the measurements are ordered from lowest to highest.

Definition 3 *The arithmetic mean, or **mean**, is defined to be the sum of the measurements divided by the total number of measurements.*

Population mean and **sample mean** are different. The population mean (denoted by $\boldsymbol{\mu}$) is unknown, while the sample mean (denoted by $\bar{\boldsymbol{y}}$) is used to make inferences about the corresponding μ.
If we let $y_1, y_2, \ldots, y_n$ denote the measurements observed in a sample size of size n (for example we record the weight of n individuals), then the sample mean $\bar{y}$ can be written as

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

The sample mean $\bar{y}$ is then used as an estimate of the mean value μ.

The mean is a useful measure of the central values of a set of measurements, but is subject to distortion due to the presence of one or more extreme values called *outliers*. These outliers pull the mean in the direction of the outliers, distorting the mean as a measure of a central value. The median is often used in place of the mean when there are extreme values in the data set.

Exercise 5 *Indicate the mean, median and mode on each distribution shape in Figure 1.6.*
For each measure, indicate if: (a) there exists more than one of this measure for a set of measurements, (b) it is influenced by extreme values, (c) it is applicable to qualitative or quantitative data.

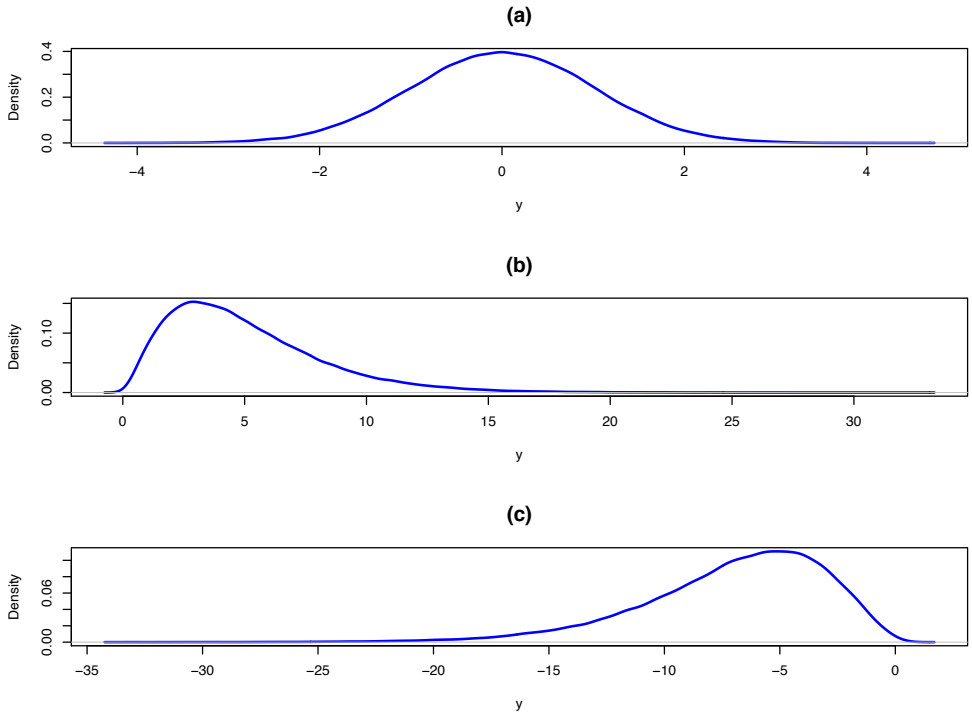

Figure 1.6: Relation between the mean, the median and the mode.

Remark 4 *We are not restricted to using only one measure of central tendency. For some data sets, it will be necessary to use more than one of these measures to provide an accurate descriptive summary of central tendency for the data.*

1.2.2 Measures of variability

Measures of variability are needed to determine how dispersed are the set of measurements around the mean. For example we can obtain relative frequency histograms with the same mean but different relative frequency histograms. For example figure 1.7 illustrates histograms with the same mean but a different spread (**variability**) about the mean.

Percentiles and interquartile range. A first simple measure is the range (difference between the largest and the smallest measurements of the set), and the use of percentiles.

Definition 4 *The p^{th} percentile of a set of n measurements arranged in order of magnitude is the value that has at most p % of the measurements below it and at most (100 - p) % above it.*

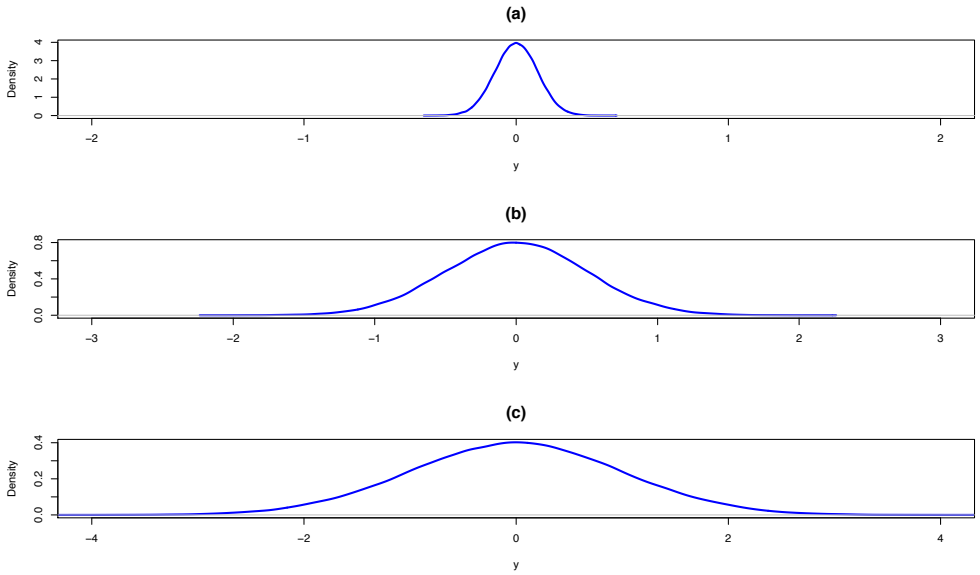

Figure 1.7: Variability around the mean with frequency histograms.

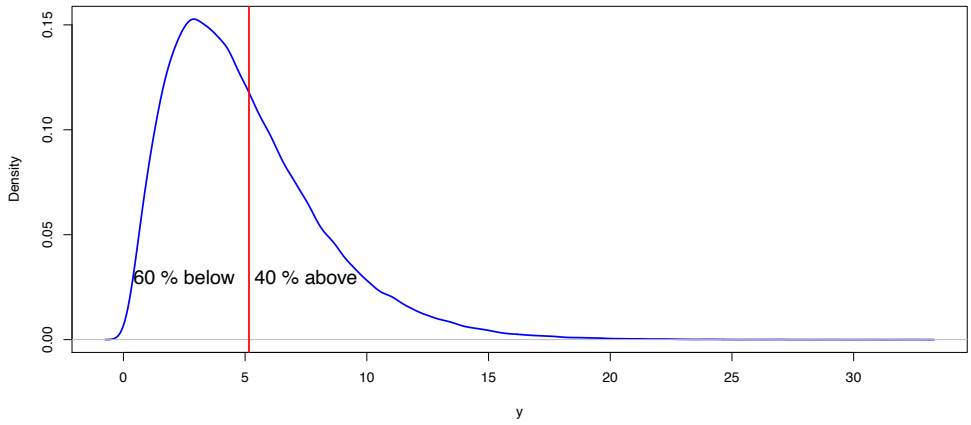

Figure 1.8: The 60th percentile of a set of measurements. The vertical red line indicates the 60th percentile.

Specific quartiles of interest are the 25th, 50th and 75th quartiles (called lower quartile, middle quartile and the upper quartile).

Definition 5 *The interquartile range (IQR) of a set of measurements is defined to be the difference between the upper and lower quartiles:*
IQR = 75th quartile - 25th quartile

The IQR completely ignores the extremes in the data. It can be quite useful to compare the variabilities of two or more sets of measurements, and is used to plot the boxplots (see following Section 1.4).

Exercise 6 *Similar to Figure 1.8, represent the lower, upper quartiles, median and IQR of a normal distribution.*

Exercise 7 *For the gene ID 724932, we order the expression values by increasing order. Give the lower, upper quartile, median and IQR.*

1	2	3	4	5	6	7	8	9	10	11
0.1458	0.2052	0.2464	0.2613	0.2619	0.2788	0.2903	0.3067	0.3094	0.3114	0.3278
12	13	14	15	16	17	18	19	20	21	22
0.3417	0.3447	0.3737	0.3795	0.3866	0.4024	0.4088	0.4092	0.4115	0.4366	0.4401
23	24	25	26	27	28	29	30	31	32	33
0.4403	0.4453	0.4469	0.4485	0.4615	0.4735	0.4792	0.4814	0.4964	0.5141	0.5212
34	35	36	37	38	39	40	41	42	43	44
0.5323	0.5345	0.5474	0.5592	0.5770	0.5846	0.5885	0.5922	0.6019	0.6022	0.6032
45	46	47	48	49	50	51	52	53	54	55
0.6053	0.6108	0.6134	0.6731	0.6848	0.6996	0.7045	0.7054	0.7212	0.7255	0.7803
56	57	58	59	60	61	62	63			
0.7974	0.8349	0.8415	0.8457	0.9782	1.1194	1.1367	1.6856			

Variance.

Exercise 8 *Suppose we have 5 measurements representing the percentage of registered voters in 5 cities:* $y_1 = 68$, $y_2 = 67$, $y_3 = 66$, $y_4 = 63$ *and* $y_5 = 61$. *Represent the data in a dot digram:*

The sample mean is

$$\bar{y} = \frac{68 + 67 + 66 + 63 + 61}{5} = 65$$

The deviations of the measurements are computed by $y - \bar{y}$.
What would 'little variability' mean in this example?

How can we measure the overall deviation?

The **variance** is useful not only to compare the variabilities of several sets of measurements, but also for interpreting the variability of a single set of measurements.

Definition 6 *The variance of a set of n measurements $y_1, y_2 \ldots, y_n$ with mean $\bar{y}$ is the sum of the squared deviations divided by $n - 1$:*

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Similarly to the sample and population means, the **sample variance** is denoted $\boldsymbol{s^2}$, and the corresponding **population variance** is denoted $\boldsymbol{\sigma^2}$.

Remark 5 *Some statisticians define the sample variance to be $\sum_i (y - \bar{y})^2/n$. However, the use of $(n-1)$ in the denominator makes this measure an unbiased estimator of the population variance σ^2, which means that for a very large number of samples, each of size n, and computed s^2 for each sample, the average sample variance would be equal to the population variance σ^2 (if we divided by n, then $s^2 \leq \sigma^2$).*

Definition 7 *The **standard deviation** of a set of measurements is defined to be the positive root of the variance.*

The standard deviation yields a measure of variability having the same unit of measurements as the original data, whereas the units for variance are the square of the measurements units.
The **sample standard deviation** is denoted $\boldsymbol{s}$, and the corresponding **population standard deviation** is denoted $\boldsymbol{\sigma}$.

1.3 Quantile-Quantile plot

The Q-Q plot enables to visualise if the data are normally distributed. In this plot, each point corresponds to each quantile[2] of the data against the corresponding quantiles of the normal distribution. The added straight line represents points which correspond exactly to the quantile of the normal distribution. The closer the the points appear to the line, the more likely the data are normally distributed.

[2]quantiles are points taken at regular intervals from the cumulative distribution of a random variable, e.g. the 4-quantiles are called the quartiles.

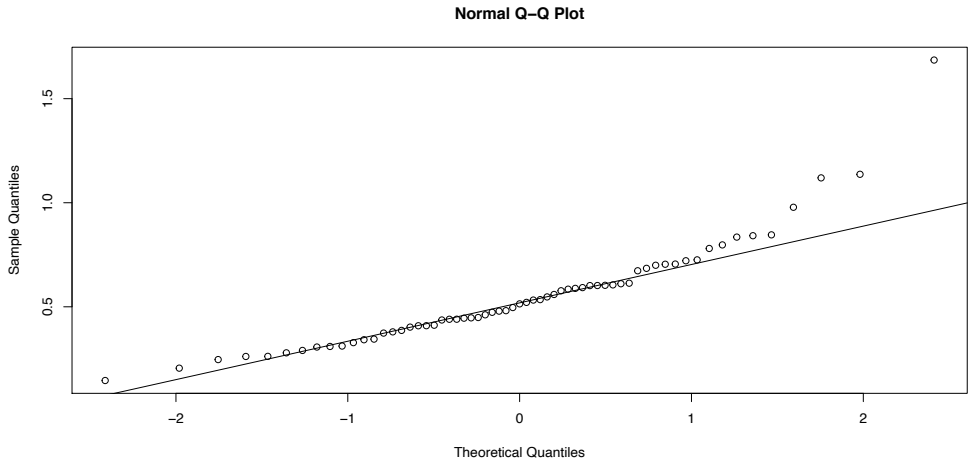

Figure 1.9: SRBCT gene expression data. Q-Q plot of the expression values of one gene.

Exercise 9 *On Figure 1.9, does the distribution of the expression values of one gene appear normally distributed?*

1.4 The boxplot

The boxplot (also called box-and-whiskers plot) displays the symmetry of the distribution, includes numerical measures of central tendency, and gives a very simple albeit thorough description of the data. The boxplot uses the median and quartiles of a distribution.

How to construct a boxplot.

1. Order the data from smallest to largest value
2. Divide the ordered data set into two data sets using the median $M = Q_2$ as the dividing values
3. The lower quartile Q_1 is the median of the set of values consisting of the smaller values. The upper quartile $Q3$ is the median of the set of values consisting of the largest values
4. Draw a box between Q_1 and Q_3 and draw a solid line to locate the median
5. The IQR is defined as the distance between Q_3 and Q_1
6. The end of the whiskers are usually defined as $Q_1 - 1.5IQR$ and the upper inner fence as $Q_3 + 1.5IQR$
7. Any data not included between the whiskers is plotted as an outlier with a dot.

This is the information that we can draw from a boxplot:

1. the median,
2. the variability given by the IQR,

3. the symmetry of the distribution,
4. the skewness is indicated by the length of the whiskers,
5. the outliers.

Exercise 10 *Using the results of Exercise* **??**, *draw the boxplot of the SRBCT gene expression data. Discuss or identify the 5 types of information from the boxplot that you represented.*

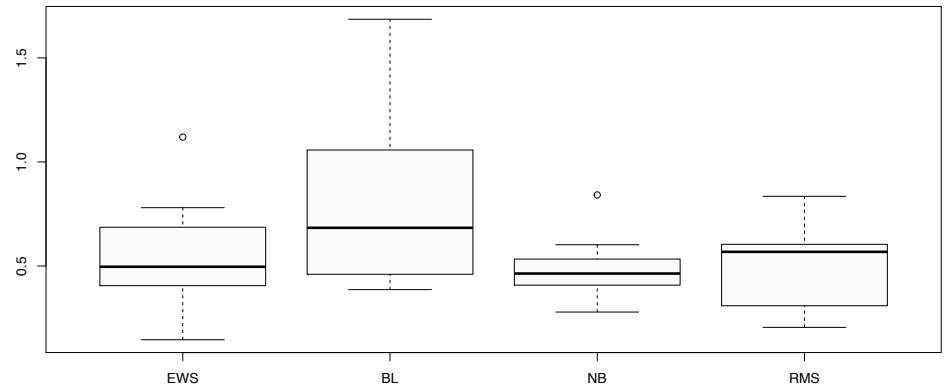

Figure 1.10: SRBCT gene expression data. Boxplot of the same gene with respect to the tumour classes.

Boxplots provide a powerful graphical technique for comparing samples from several different treatments or populations (Fig. 1.10).

Exercise 11 *Do we observe any difference in the gene expression values in the different tumour categories (Fig. 1.10? Compare to the information contained in the boxplot to the density plots from Fig. 1.5.*

1.5 Relationship between two variables

1.5.1 Scatterplot

A scatterplot displays the general shape and direction of the relationship between two quantitative variables. Each point on the plot represents the value of the two measured variables for each individual. The relationship can be summarized by fitting a straight line through the plotted points. There is a strong relationship between the two variables if the points are close to the line, and a weak relationship if the points are widely scattered about the line.

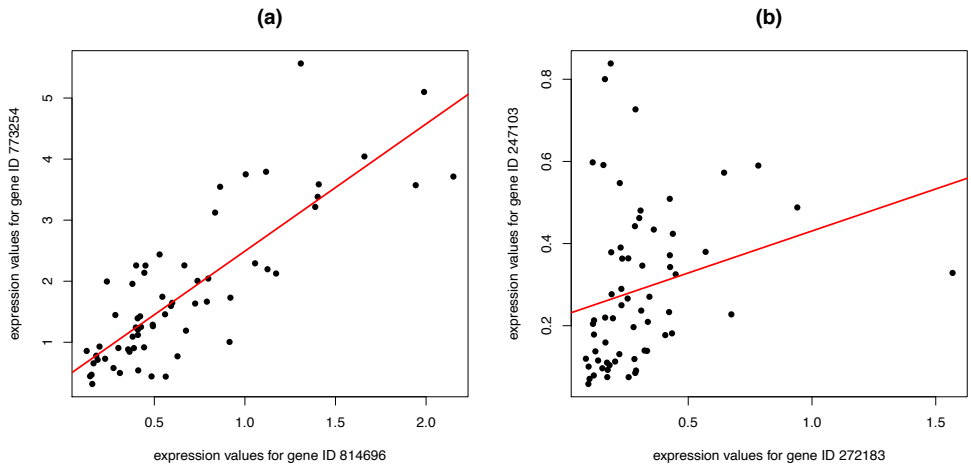

Figure 1.11: SRBCT gene expression data. Example of scatterplot for the expression values of 2 genes for all patients. A linear least squares regression model has been fitted through the points and the fitted line is represented in red.

Exercise 12 *Comment on the relationship between the expression values of the two genes in Figure 1.11 for both (a) and (b). Is a linear least squares regression model appropriate?*

1.5.2 Numerical measure: correlation coefficient

A numerical measure to evaluate the strength of a relationship between two quantitative variables is summarized with a statistic called the *Pearson's correlation coefficient*.

Definition 8 *The correlation coefficient r measures the strength of the linear relationship between two quantitative variables x and y. If the points in the scatter plot are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ then the correlation is defined by*

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The correlation coefficient r is a *unit-free measure* of the strength of linear relationship between the quantitative variables x and y.

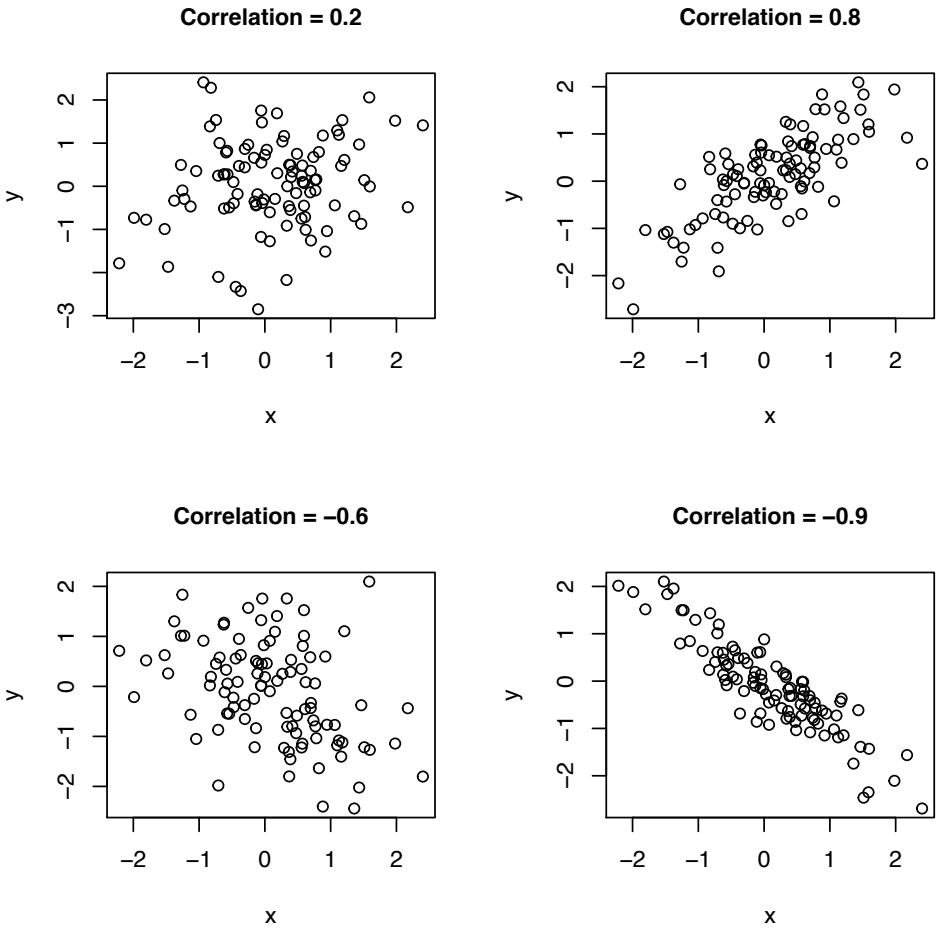

Figure 1.12: Examples of correlation coefficients for 100 observations (simulated data)

Remark 6 *The Pearson's correlation coefficient measures the strength and direction of a linear relationship. However, this measures is susceptible to outliers and does not apply when the relationship is not linear. In the data represented in Figure 1.11 (b), the Pearson correlation coefficient is -0.26 and seems to be affected by both several outliers at the tail of both variables and a possible non linear relationship. The Spearman correlation is less sensitive than the Pearson correlation to outliers (correlation coefficient is -0.25).*

Side-by-side boxplots

Side-by-side boxplots provide a visual assessment of the similarity in distributions in several variables at a time. Figure 1.13 displays such plot.

Exercise 13 *Comment on the similarity and differences of the distributions of the gene expression values.*

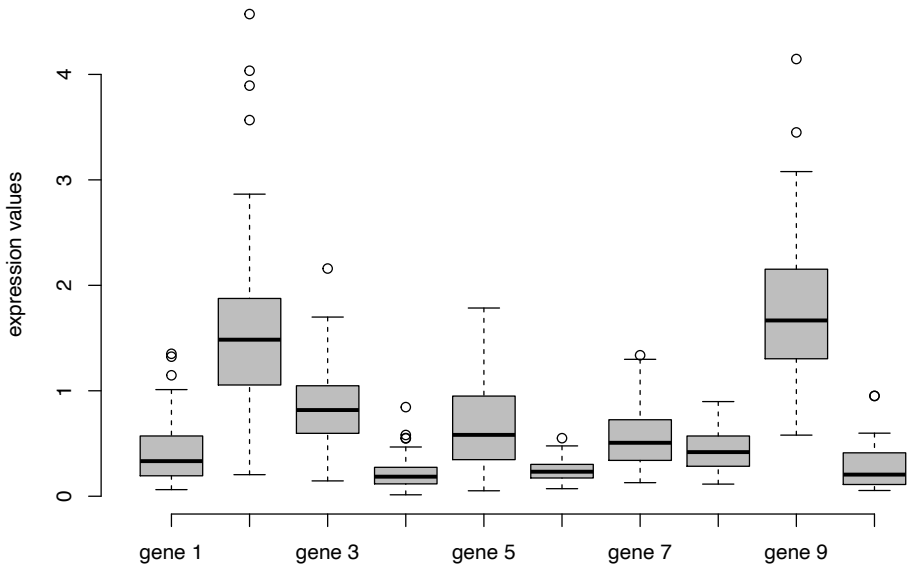

Figure 1.13: SRBCT data. An example of side-by-side boxplots for the expression values of 10 genes.

1.6 Summary

This chapter was concerned with graphical and numerical description of data. In particular, frequency, relative frequency histograms, boxplots, Q-Q plots are graphical techniques only applicable to quantitative data.
Numerical descriptive measures include measures of *central tendency* (mode, median, arithmetic mean), and measures of *variability* (range, interquartile range, variance and standard deviation of a set of measurements).
We examined plots for summarizing the relations between two quantitative variables. The material presented here will be expanded in later chapters.
Key formulas include sample mean, sample variance, sample standard deviation and correlation coefficient.

Chapter 2

Probability distributions

Last chapter presented graphical and numerical descriptive techniques to summarize and describe a sample. However, a sample is not identical to the population from which it was selected. This chapter is dedicated to *where* the data come from and some important statistical distributions that will enable us to perform statistical testing (see next Chapter).

2.1 Important terminology

Definition 9 *A **population** is a complete set of individuals or objects that we want information about. Ideally, we would collect the information about the whole population. However, this might be too expensive, in time or money or simply impossible.*
*Instead we have to take a **sample**, a subset of the population, and use the data from the sample to infer something about the population as a whole. The sample should be chosen so that it is representative of the population but also that is it not biaised in any way. One way of achieving this is to take a **random sample from the population**.*

Definition 10 *When we want to draw wider conclusions from an experiment, we need to be clear about what the population of interest will be and we need to obtain a representative sample from that population. **Selection bias** occurs when the sample itself is unrepresentative of the population we are trying to describe.*

Definition 11 ***Probabilities** are used to describe the process of sampling from a population. We need to know the probability of observing a particular sample outcome in order to make an inference about the population from which the sample was drawn. To do this we need to know the probability associated with each value of a given variable y (e.g. height). These probabilities generate a distribution of theoretical relative frequencies called the **probability distribution** of y. Probability distributions differ for discrete and continuous random variables. For discrete random variables, we will compute the probability of specific individual values occurring. For continuous random variables, the probability of an interval of values is the event of interest.*

Definition 12 *A **random variable** (also called stochastic variable) is a random process with a numerical outcome, i.e. a variable whose value is subject to variations due to chance. A random variable does not have a single fixed value, it can take a set of possible different values. Each of them are associated to a probability.*
*A **discrete random variable** is a random variable with discrete outcome, i.e. it assumes any of a specified list of exact values. A **continuous random variable** assumes any numerical value in an interval or collection of intervals (continuous outcome).*

The distinction between discrete and continuous variables is pertinent as we are seeking the probabilities associated with specific values of a random variable.

2.2 Probabilities distributions for discrete random variables

The probability distribution for a discrete random variable displays the probability $\mathbb{P}(y)$ associated with each value of y.

Properties of discrete random variables

- The probability associated with every value of y lies between 0 and 1,
- The sum of the probabilities for all values of y is equal to 1,
- The probabilities for a discrete random variable are additive: the probability that $y = 1$ **or** $y = 2$ is equal to $\mathbb{P}(1) + \mathbb{P}(2)$

2.2.1 Binomial distribution

Definition 13 *A binomial experiment has the following properties:*

1. *The experiment consists of n identical trials,*
2. *Each trial results in one of two outcomes (labelled 'success' and 'failure'),*
3. *The probability of success on a single trial is equal to π and π remains the same from trial to trial.*
4. *Each trial is independent (the outcome of one trial does not influence the outcome of any other trial).*
5. *The random variable y is the number of successes observed during the n trials.*

Exercise 14 *An experiment on 300 rats is performed to test if a drug is effective. 255 rats show a favourable response and 45 an unfavourable response. Under which conditions this study satisfies the properties of a binomial experiment? The binomial distribution for this example is plotted in Figure 2.1.*

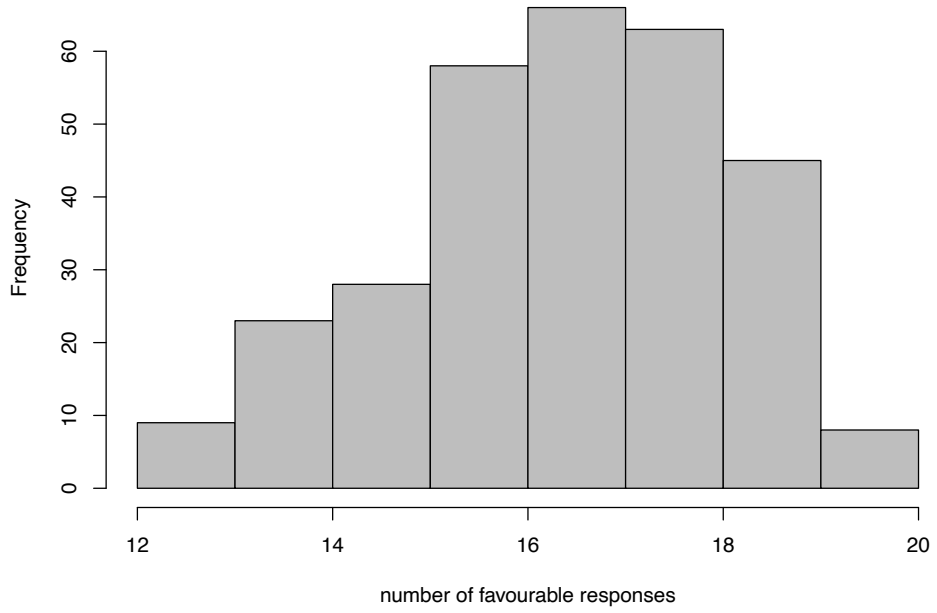

Figure 2.1: The binomial distribution for $n = 300$ and $\pi = 0.85$.

Let π be the probability of success on a single trial and X the random variable denoting the number of successes. The probability $\mathbb{P}$ of the event $(X = k)$ that k successes occur out of n trials is:

$$\mathbb{P}(X = k) = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{n-k} \qquad \text{for } k = 0, \ldots, n$$

where $n! = n(n-1)(n-2)\ldots(2)(1)$ (referred to as 'n factorial').

We denote by $X \sim Bin(n, \pi)$ a random variable having a Binomial distribution with n trials with success probability π.

Because the random value X is discrete, the probabilities can be easily computed or listed in a Binomial distribution table (see Appendix A.1 for an example of such table).
The built-in density function in `R`, `dbinom(k,n,p)`, where `p` $= \pi$ directly gives the values of the probability $\mathbb{P}(X = k)$. The cumulative probability distribution function `pbinom(k,n,p)` gives $\mathbb{P}(X \leq k)$.

Exercise 15 *Suppose that in the example above, the effectiveness of the drug is known to be* 85%. *A new experiment is performed on 10 rats. What is the probability that 8 rats or more will give a favourable response? If we increase the sample size in the experiment, what is the probability that 16 rats or more will give a favourable response for a sample of 20 rats?*

2.2.2 Poisson distribution

The Poisson distribution is a model for random counts events (i.e. π is very small and n is very large).

Definition 14 *Let y be the number of events occurring during a fixed time interval or fixed region of space. The probability disitrbuion of y is Poisson, provided the following conditions:*

1. *Events occur one at a time,*
2. *The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a non overlapping time period or region of space,*
3. *The expected number of events during one predion/region, λ is the same as the expected number of events in any other period/region.*

The probability $\mathbb{P}$ of the event $(X = k)$ that k successes occur out of n trials is:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where λ is the mean of the counts.
We denote by $X \sim \mathcal{P}(\lambda)$ a random variable having a Poisson distribution with mean λ.

Remark 7 *Unlike the binomial distribution, there is no upper limit for X for a Poisson distribution (even if it is unlikely that large values could occur, $\mathbb{P}(X = x)$ never equals 0.*

The `R` density function `dpois(k, lambda)` directly gives the values of the probability $\mathbb{P}(X = k)$. The cumulative probability distribution function `ppois(k,lambda)` gives $\mathbb{P}(X \leq k)$.

Figure 2.2 gives some example of Poisson distributions for different λ values.

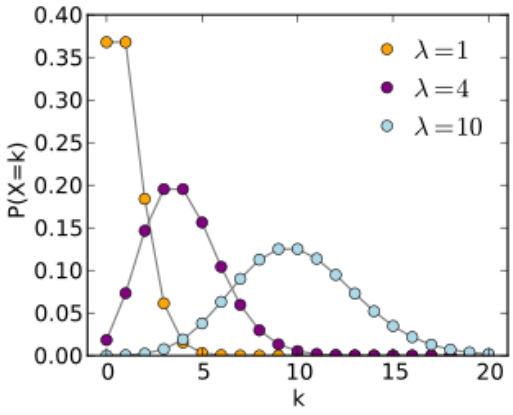

Figure 2.2: Poisson distribution for different λ values (source: Wikipedia). The horizontal axis indicates the number of occurrences k. The function is only defined at integer values of k.

Exercise 16 *A team of wildlife scientists is surveying the number of small mammals in the region. Let x denote the number of field mice captured in a trap over a 24-hour period. Suppose that x has a Poisson distribution with an average number of mice captured per trap equals 2.3.*

What is the probability of finding 4 mice in a randomly selected trap?
What is the probability of finding ***more*** *than 4 mice in a randomly selected trap?*

There exists a coefficient of dispersion (or index of dispersion) $CD = \frac{s^2}{\bar{x}}$ (ratio sample variance/mean) which indicates if the data comes from a Poisson distribution. The value should be around 1 for Poisson data. In biological studies, there sometimes will be a larger number of extreme values than the Poisson distribution predict. In that case (high counts), we would find $s^2 > \bar{x}$ (i.e. CD value >1, *overdispersion* occurs) compared to the expected $s^2 = \bar{x}$ for a Poisson distribution. In the case of overdispersion, other distributions can be used such as Negative Binomial or improvements of Poisson distribution for overdispersed data.

Remark 8 *When n is large and π is small in a binomial experiment, the Poisson distribution provides a good approximation to the binomial distribution. As a general rule, the Poisson distribution provides an adequate approximation to the binomial distribution when $n \geq 100$, $\pi \leq 0.01$ and $n\pi \leq 20$.*

2.3 Probability distributions for continuous random variables

Discrete random variables have possible values that are distinct and separate. Other random variables are most usefully considered to be continuous: their possible values for a whole interval (or range).
Theoretically, we can assume that continuous random variables can have values associated with infinitely many points in a line interval. Figure 2.3 illustrates the probability distribution for a continuous random variable (a) where the total area under the curve should be equal to 1. The probability that a continuous random variable falls in an interval (say between two points a and b) is equal to the area under the curve over the interval $[a, b]$, written $\mathbb{P}(a < y < b)$.

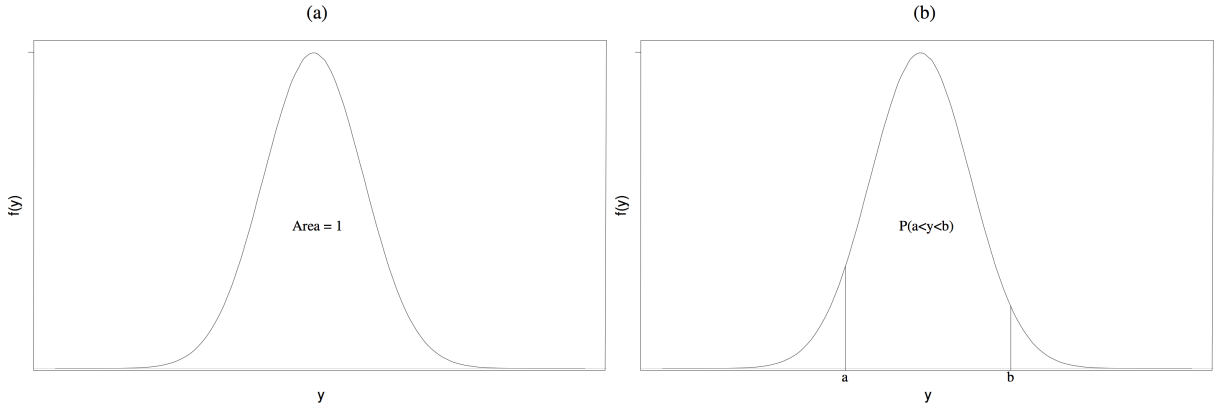

Figure 2.3: Probability distribution for a continuous random variables. (a) indicates the total area under curve, (b) the probability for a specific interval.

2.3.1 Normal Distribution

Many variables of interest have a normal distribution. For example pre processed gene expression values (for one given gene) are seen as a realisation of a random variable X having a normal distribution. The data values follow a normal distribution with mean μ and variance σ^2 (denoted '$\mathcal{N}(\mu, \sigma^2)$').

The value of the distribution function is given by $\mathbb{P}(X \leq x)$, which is the probability of the population to have values smaller than or equal to x.

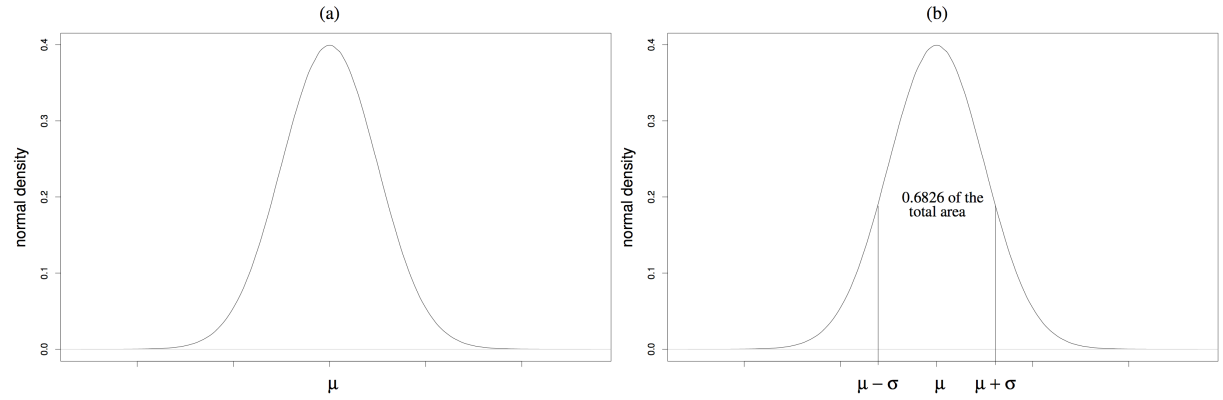

Figure 2.4: (a) Density of the standard normal distribution, (b) Area under normal curve within 1 standard deviation of mean.

The normal probability distribution is bell shaped and symmetrical about the mean μ (Fig. 2.4 (a)). When μ increases the distribution moves to the right, if σ is small (large) then the distribution is steep (flat).

We can calculate the probability that a measurement falls within any distance of the mean μ. For example, if we select a measurement at random from a population with a normal distribution, the probability is approximately 0.68 that the measurement will lie within 1 standard deviation of its mean (Fig. 2.4 (b)) (this value is given by the Empirical Rule). In fact, we can calculate the probability that a measurement falls within any distance of the mean μ for a normal curve, or the probability that a measurement is below or above a given value, using the `pnorm(x, μ, σ)` function.

Exercise 17 *An environmental protection agency has developed a procedure for measuring vehicle emission level of nitrogen oxide. Let P denote the amount of this pollutant in a randomly selected vehicle in Brisbane. Let suppose that the distribution of P can be adequately modelled by a normal distribution with a mean level 70 ppb (parts per billion) and standard deviation of 13 ppb.*
What is the probability that a randomly selected vehicle will have emission levels less than or equal 60 ppb? strictly greater than 90 ppb?

Evaluating whether or not a population distribution is normal. To assess wheter or not a random sample $y_1, y_2, \ldots, y_n$ was selected from a normal distribution, we use a **normal probability plot** of the data values. This plot is a variation of the quantile quantile plot introduced on Chapter 1 Section 1.3. In the normal probability plot, we compare the quantiles from the data observed from the population to the corresponding quantiles from the standard normal distribution.

Standard unit or Z score. There are numerous Normal distributions (with different means and different standard deviations) and hence it is more practical to focus on one Normal distribution with mean 0 and standard deviation 1 (the *standard Normal distribution*). We are going to subtract and divide the data following a Normal distribution as it doe not change the shape of the Normal distribution. The standardized values are now called *z-scores*, according to the following formula:

Definition 15 *If* $X \sim \mathcal{N}(\mu, \sigma)$, *then*

$$Z = \frac{X - \mu}{\sigma} \qquad \textit{so that} \qquad Z \sim \mathcal{N}(0, 1).$$

We can then use a Normal Standard distribution table for any kind of normal distribution (see Table in A.2).

Exercise 18 *Suppose that our data follow a normal distribution. The 1.5 IQR rule in a boxplot states that any observation above* $Q_3 + 1.5 \times IQR$ *or below* $Q_1 - 1.5 \times IQR$ *should be flagged as an outlier. What is the probablity that an observation is flagged as an outlier?*
Use the Normal cumulative distribution function in Table **??** *to obtain the values of* Q_1 *and* Q_3 *and then compute the IQR and then the probability that an observation is an outlier.*

2.3.2 T-distribution

The t-distribution has many useful applications for testing hypotheses about means of gene expression values, in particular when the sample size is lower than thirty. If the data are normally distributed with sample mean $\bar{y}$ and standard deviation s, then the transformed values of $\sqrt{n}(\bar{y} - \mu)/s$ (i.e. centered and standardized) follow a t-distribution with $n - 1$ degrees of freedom (denoted df)[1]. These transformed values are called *standardized* z values, as seen in previous Section 2.3.1.
The t-distribution is approximately equal to the normal distribution when the sample size is greater than or equal to thirty.
There exists many t distributions depending on the sample size (and therefore the number of degrees of freedom) as shown in Figure 2.5(a).

[1]The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df). In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself.

Definition 16 *Properties of the Student's t distribution*

1. *There are many different t distributions, specified by their degres of freedom,*
2. *The t-distribution is symmetrical about 0 (it has a mean equal to 0, similar to the z distribution),*
3. *The quantity*

$$T = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

is called the t-statistic and has a t distribution (also called Student's t distribution) with $n - 1$ degrees of freedom.

Remark 9 *When n increases, the distribution of t approaches the distribution of z.*

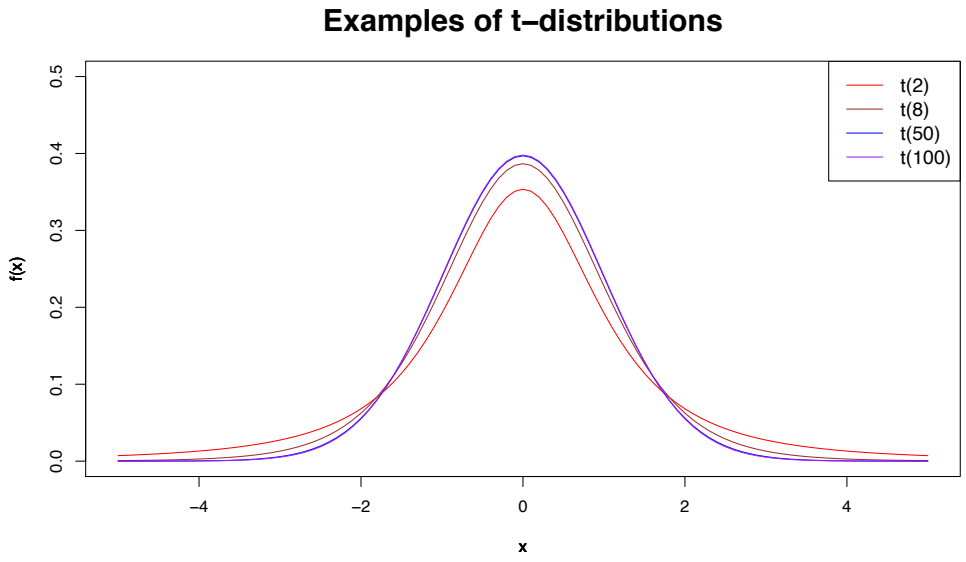

Figure 2.5: Density of t distribution with different degrees of freedom.

2.3.3 Chi-squared distribution

The χ^2 distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data (see following chapter and practical). Figure 2.6 shows examples of χ^2 distribution densities for different values of degrees of freedom (denoted *df*).

Definition 17 *Some of the properties of the χ^2 distribution are as follows:*

1. *The χ^2 distribution is positively skewed with values between 0 and ∞,*
2. *There are many χ^2 distributions labeled by their degrees of freedom,*
3. *The mean and variance are given by the degrees of freedom: $\mu = df$ and $\sigma^2 = 2 \times df$*

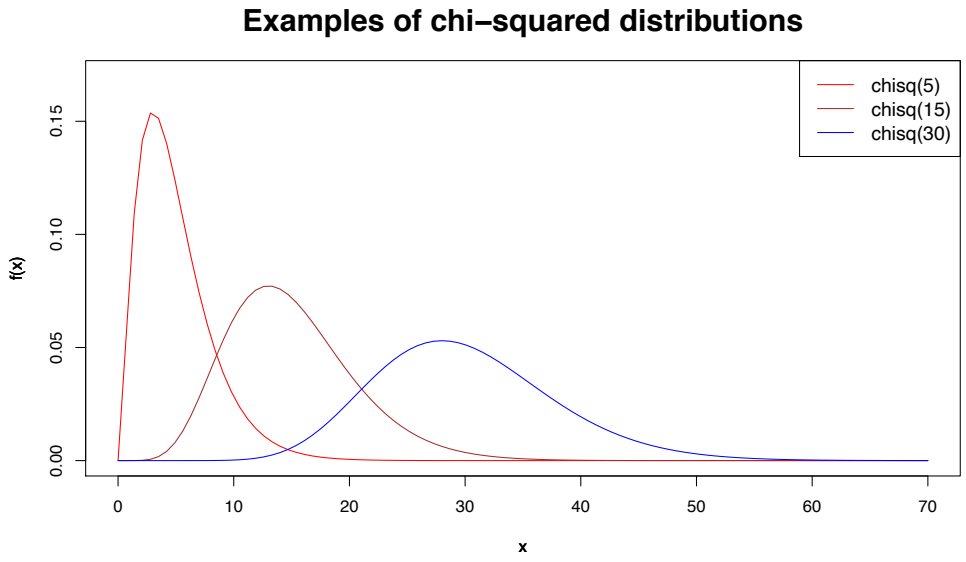

Figure 2.6: Density of χ^2 distribution with different degrees of freedom.

2.3.4 F-distribution

The F-distribution is important for testing the equality of two (or more) variances from two (or more) groups of samples. It can be shown that the ratio of variances from two independent sets of normally distributed random variables follows an F-distribution. There are two degrees of freedom involved with this distribution, one for the numerator, and one for the denominator. More specifically, if if two population have equal variances ($\sigma_1^2 = \sigma_2^2$), then s_1^2/s_2^2 follows an F-distribution with $df_1 = n_1 - 1, df_2 = n_2 - 1$ degrees of freedom ($\sim \mathcal{F}_{n_1-1,n_2-1}$), where s_1^2 (s_2^2) is the sample variance from the first (second) group and n_1 (n_2) is the number of measurements from the first (second) group.

Similar to the χ^2 distribution, the F-distribution has a skewed density curve. There exists many density curves for each combination of the numerator and denominator degrees of freedom (see Figure 2.7.

Definition 18 *Some of the properties of the F distribution are as follows:*

1. *Like the χ^2 distribution, the F distribution only assumes positive values,*
2. *The F distribution is non symmetrical,*
3. *There are many different shapes of F distributions, specified by the degrees of freedom associated to s_1^2 and s_2^2.*

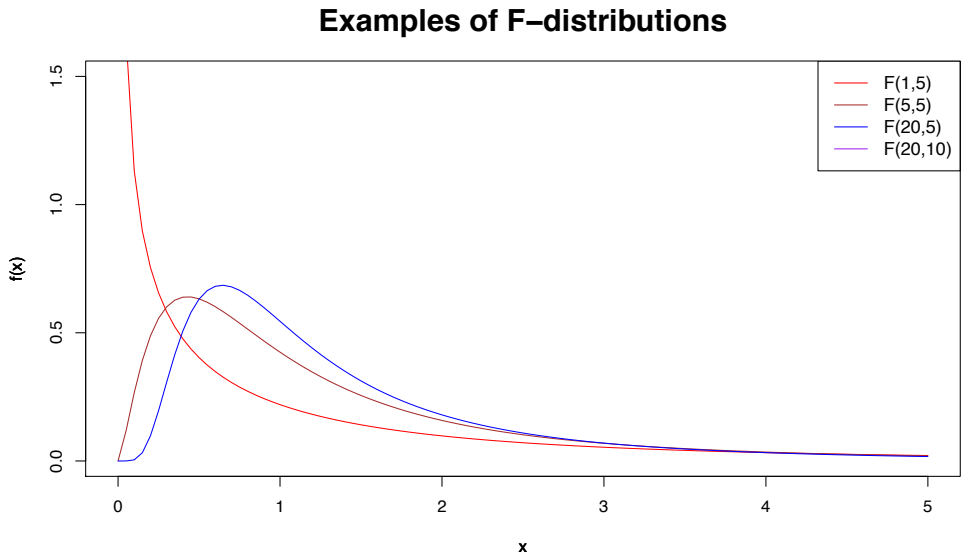

Figure 2.7: Density of F distribution with different degrees of freedom.

2.4 Summary

Distribution	Parameters	Density	Cumulative probability distribution	Quantiles	Random sampling of N observations
Binomial	n, p	`dbinom`(k, n, p)	`pbinom`(q, n, p)	`qbinom`(α, n, p)	`rbinom`(N, n, p)
Poisson	λ	`dpois`(k, λ)	`ppois`(q, λ)	`qpois`(α, λ)	`rpois`(N, λ)
Normal	μ, σ	`dnorm`(x, μ, σ)	`pnorm`(q, μ, σ)	`qnorm`(α, μ, σ)	`rnorm`(N, μ, σ)
t	df	`dt`(x, df)	`pt`(q, df)	`qt`(α, df)	`rt`(N, df)
χ^2	df	`dchisq`(x, df)	`pchisq`(q, n)	`qchisq`(α, n)	`rchisq`(N, df)
F	df_1, df_2	`dt`(x, df_1, df_2)	`pt`(q, n_1, n_2)	`qt`(α, n_1, n_2)	`rt`(N, df_1, df_2)

Table 2.1: `R` functions for random variables following distributions presented in this Chapter.

In this Chapter we have seen that there are many `R` functions that can be used for the distributions we have covered, where `d` stands for density, `p` for cumulative probability distribution, `q` for quantiles α and `r` for drawing random samples (Table 2.1).

The Binomial distribution is an important example of a discrete random variable and is used to model sampling from finite populations. The Poisson distribution is a model for random counts of rare events. The coefficient of dispersion gives a rough measure for determining whether data could come from a Poisson distribution. The Normal distribution is a model for continuous random variables. The normality of a variable distribution can be assessed via normal probability plot. The t distribution is used for a small number of samples. Both Normal and t distribution are similar when the number of samples is large.

In this Chapter we have covered some important distributions and shown how to calculate probabilities of events given the distribution of some random variables. Next Chapter will cover statistical testing and statistical inference.

Chapter 3

Statistical tests

In Chapter 2, we defined families of hypothetical distributions. *The objective of statistics is to make inferences about a population based on the information contained in a sample.* Populations are characterized by numerical descriptive measures called parameters (for example mean, median, standard error...). In any research setting, the specific values of such parameters are unknown and inferences must be made about these parameters.

3.1 Estimation and Hypothesis testing

Methods for making inferences about parameters fall into one of two categories: either we will *estimate* the value of the population parameter of interest, or we will *test a hypothesis* about the value of the parameter. This involves different procedures, answering different types of questions. In estimating a population parameter we are answering the question: 'What is the value of the population parameter?'. In testing a hypothesis, we are for example answering the question: 'Does the population parameter satisfy $\mu > 20$?'

3.1.1 Hypothesis testing

Assume we have μ_0 a number representing a hypothesized population mean in an experiment.
With respect to the real population mean μ the null hypothesis can be formulated as:

$$H_0 : \mu = \mu_0$$

This hypothesis is also called the **negation** of the alternative (or **research**) hypothesis:

$$H_1 : \mu \neq \mu_0$$

Either H_0 or H_1 is true. The alternative hypothesis is true if $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$ holds true. This type of alternative hypothesis is called *two-sided*.
A *one-sided* hypothesis would be for example: $H_1 : \mu > \mu_0$.

The null hypothesis is statistically tested against the alternative using a suitable distribution of a statistic, where the statistic is computed from the experimental data. By comparing the statistic with its distribution, we can draw a conclusion with respect to the null hypothesis and reject or not H_0. The probability to reject H_0, given the truth of H_0 is called the *significance level*, generally denoted α and usually set to $\alpha = 5\%$ (but this is not compulsory).

3.1.2 Statistical test

Definition 19 *A statistical test is based on the concept of proof by contradiction and is composed of the five parts listed here:*

1. *Null hypothesis H_0,*
2. *Research hypothesis (alternative hypothesis): H_1,*
3. *Test statistics (T.S),*
4. *Rejection region,*
5. *Check assumptions and draw conclusions.*

3.1.3 Example with a Z-test

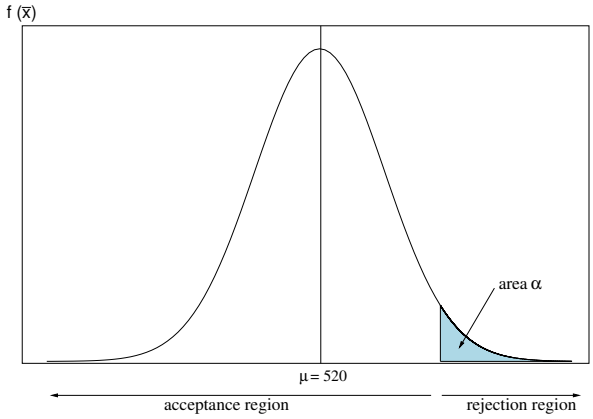

Figure 3.1: Rejection region for the soybean example for $H_1 : \mu > 520$

Definition 20 *If a set of measurements $(x_1, x_2, \ldots, x_n) \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \textit{so that} \qquad Z \sim \mathcal{N}(0, 1).$$

When applying a Z-test, we assume that

- *the **standard deviation σ is known***
- *the data follow a normal distribution.*

- *Hypotheses that can be tested with a Z-test:*
  - *Case 1: $H_0 : \mu \leq \mu_0$ vs. $H_a : \mu > \mu_0$ (right-tailed test)*
  - *Case 2: $H_0 : \mu \geq \mu_0$ vs. $H_a : \mu < \mu_0$ (left-tailed test)*
  - *Case 3: $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ (two-tailed test)*

- *The test statistics is*

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

- *The rejection region is:*
  - *Case 1. Reject* H_0 *if* $z \geq z_\alpha$
  - *Case 2. Reject* H_0 *if* $z \leq z_\alpha$
  - *Case 3. Reject* H_0 *if* $|z| \geq z_\alpha$

Example 1 *An agricultural service wants to determine whether the mean yield per acre (in bushels) for a particular variety of soybeans has increased since last year. The mean yield was 520 bushels per acre. We have a sample of* $n = 36$ *one-acre plot. From these data we compute the sample mean* $\bar{x} = 573$ *and the sample standard deviation* $s = 124$*. Can we conclude that the mean yield for all farms is above 520?*

1. $H_0 : \mu \leq 520$
2. $H_1 : \mu > 520$
3. *T.S:* $z = \frac{\bar{y}-520}{124/\sqrt{36}}$
4. *For* $\alpha = 0.025$*, we reject* H_0 *if T.S.* $> T.S_\alpha$*, or, equivalently, if the p-value is* $< \alpha$*.*

The shaded area in Figure 3.1 illustrates the rejection region with an area α *in the right tail of the distribution of* $\bar{x}$*. Determining the location of this rejection area is equivalent to determining the* z *value that has an area* α *to its right (here* $\alpha = 0.025$*). A statistics Table for the* $\mathcal{N}(0, 1)$ *indicates that this value is* $T.S_\alpha = 1.96$ *(also given by* `qnorm(0.975, 0, 1)`*)**.
We have

$$T.S = z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{573 - 520}{124/\sqrt{36}} = 2.56$$

5. *Since we have T.S.* $> T.S_\alpha$*, we reject* H_0 *in favor of the alternative hypothesis and conclude that the average soybean yield per acre is greater than 520.*

3.1.4 One and two sided tests

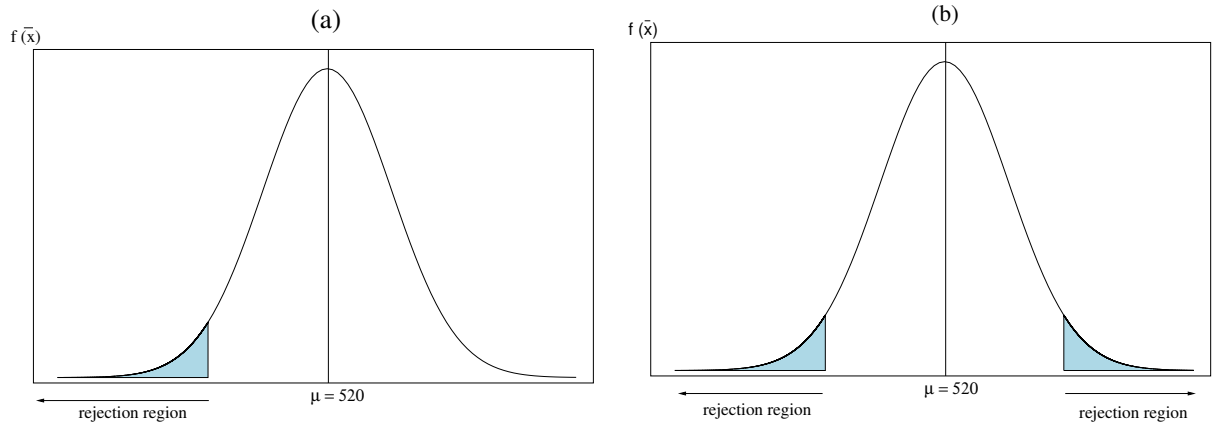

Figure 3.2: (a) One and (b) two sided tests rejection regions based on the the soybean example.

In the example above we conducted a one-sided test where $H_1 : \mu > 520$. If our alternative hypothesis was instead $H_1 : \mu < 520$, small values of $\bar{x}$ would indicate the rejection of null hypothesis. The rejection

region would be located in the lower tail of the distribution of $\bar{x}$ (Fig. 3.2 (a)).
A two-sided test could be formulated for $H_1 : \mu \neq 520$ where both large and small values of $\bar{x}$ would contradict the null hypothesis and the rejection region would be located in both tails of the distribution of $\bar{x}$ (Fig. 3.2 (b)).

3.2 Statistical tests with discrete variables

3.2.1 Binomial test

Suppose a coin is tossed 10 times and 9 times out of 10 it comes up heads. How suspicious is this? Let's calculate the probability that this could happen for a fair coin. Let X the number of heads, then for a fair coin, $X \sim$. We have

$$\mathbb{P}(X = 9) =$$

Remark. Samples independence: here we used the assumption that each toss is *independent* from each other.

In this example our assumption is that the coin is fair, so $H_0 : p = 0.5$, against the alternative hypothesis $H_1 : p > 0.5$ (i.e the coin is biased towards head). What is the probability that $\mathbb{P}(X \geq 9)$?

$$\mathbb{P}(X \geq 9) = 1 - \mathbb{P}(X \leq 8) = \texttt{1 - pbinom(8, 10, 0.5)} = 0.0107$$

Since the `p-value` is less than our significance level $\alpha = 0.05$, we reject the null hypothesis. Alternatively, we can use the `R` function `binom.test`:

```
> binom.test(9, 10, 0.5, alternative = 'greater')

        Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6058367 1.0000000
sample estimates:
probability of success
                   0.9
```

3.2.2 Chi-squared test

The chi-square test assesses whether data which appear to be dependent is the result of random variability, rather than real dependence.

Definition 21 *If two variables are dependent, it means that one variable has some value to predict the other.*

We arrange the data in a contingency table with r rows and c columns. The null hypothesis for the χ^2 test is independence: H_0 : the row and the column variables are independent, vs. H_1 : the row and the column variables are dependent (*associated*). Without going into too many details, the χ^2 statistic is the sum of all cells in the contingency table of

$$(\text{observed values} - \text{expected values})^2/\text{expected values}$$

Rejection of the null hypothesis indicates that the apparent association is not reasonably attributable to chance (it does not indicate anything about the strength of the type of association). This test statistic is

Severity	Age category				
	I	II	III	IV	All Ages
Moderate	15	32	18	5	70
Mildly severe	8	29	23	18	78
Severe	1	20	25	22	68
All severities	24	81	66	45	216

Table 3.1: Contingency table of example 19.

also called 'Pearson's χ^2 test' and follows a $|chi^2$ distribution with (nr-1)(nc-1) degrees of freedom (where nr = number of rows and nc = number of columns from the contingency table).

Exercise 19 *A random sample of 216 patients having a skin disease are classified into 4 age categories as represented in the contingency Table 3.1. We conduct a test to determine if the severity of the disease is independent of the age of the patient. State the null and alternative hypothesis and draw conclusions on the outputs obtained using the* `chisq.test` *function in* `R`*.*

```
> table.chisq = t(matrix(c(15, 32, 18, 5, 8, 29, 23, 18, 1, 20, 25, 22), nrow = 4, ncol = 3,
+          dimnames =list(c('I', 'II', 'III', 'IV'), c('Mod', 'Mild', 'Severe'))))
> chisq.test(table.chisq)

        Pearson's Chi-squared test

data:  table.chisq
X-squared = 27.135, df = 6, p-value = 0.0001366
```

Remark 10 *The accuracy of the approximation of the sampling distribution of χ^2 by a chi-square distribution depends on both the sample size n and the number of cells k. The approximation should be adequate if n/k exceeds 1, if no cell has a count less than 1 and no more than 20% of the cells have counts less than 5 counts. When the approximation is not valid, we can either combine levels of categorical variables to increase the observed cell counts (caution! on how to redefine the levels of the categorical variables), or use exact methods such as the Fisher's exact test (*`fisher.test()`*).*

3.3 Statistical tests with continuous variables

3.3.1 One sample tests

We have already seen an example of one sample test in Subsection 3.1.3. This type of test tests for the value of the mean of the data, for example: $H_0 = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ or $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$.

Z-test for known variance. We have seen in Subsection 3.1.3 an example of a Z-test, which tests for the value of the mean of the data, and which assumes that the **standard deviation σ is known** and that the data are assumed to follow a normal distribution.

t-test for unknown variance. In most research situation, the standard deviation σ **is unknown** and the Z-test cannot be applied. In such cases, a t-test is appropriate. The test statistic T.S is defined by

$$T.S = T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim \tau_{n-1}$$

where s is the sample standard deviation estimated from the data. The statistic table for the Student's t-test is given in Appendix A.3.

Exercise 20 *Let's go back to the SRBCT example from Chapter 1, Figure 1.2. We would like to test if the mean expression of this single gene across all 63 samples is equal to 0.5. State the null and alternative hypothesis, the T.S, the rejection region and conclude.*
We have $\bar{x} = 0.5488$, $s = 0.2525146$.

Alternatively, we can use the `t.test` function:

```
> t.test(c(data.gene), mu =0.5, alternative = 'two.sided')

        One Sample t-test

data:  c(data.gene)
t = 1.5339, df = 62, p-value = 0.1301
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.485205 0.612395
sample estimates:
mean of x
   0.5488
```

3.3.2 Two sample t-tests

T-tests are most often used when comparing two groups of patients (e.g. sick and normal) and we want to test the difference between the population means μ_1 and μ_2. For example in microarray data analysis, we would like to identify some genes for which the expression level means differ between the two groups as these genes might be crucial in explaining the development of a disease. The null hypothesis $H_0 : \mu_1 = \mu_2$ is to be tested against $H_1 : \mu_1 \neq \mu_2$ (two-sided test) or (less commonly) $H_1 : \mu_1 < \mu_2$ or $H_1 : \mu_1 > \mu_2$ (one-sided tests). The two-sided test can also be written as $H_1 : \mu_1 - \mu_2 = 0$.
In the following we denote by $\bar{x}_1$ ($\bar{x}_2$) the sample mean of the first (second) group, and by s_1 (s_2) the sample standard deviation of the first (second) group, and by n_1 (n_2) the number of patients in the first (second) group.

With unequal variance. When we make the assumption that the variances between the two groups s_1^2 and s_2^2 are unequal for a single variable, i.e. single gene expression, then, the t-test statistics is defined as

$$T.S = T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_1^2/n_2}}.$$

Under H_0, the t-test statistics is defined as

$$T.S_{H_0} = T_{H_0} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_1^2/n_2}} \sim t_{n_1+n_2-2}.$$

The test statistics is large if, under H_0, the difference between the means is large **and** the sample standard deviations are small.

Exercise 21 *Back to the the SRBCT experiment, the densities of a chosen gene were represented for each sample group in Figure 1.5 in Chapter 1. We would like to test if there is a difference between the expression means between the groups 'BL' and 'NB'. Is the use of a t-test with unequal variances valid? State the null and alternative hypothesis, the T.S, the rejection region and conclude.*
We have: $\mu_{BL} = 0.8093$, $s^2_{BL} = 0.201531$, $n_{BL} = 8$, $\mu_{NB} = 0.48885$, $s^2_{NB} = 0.01977323$, $n_{NB} = 12$.

Alternatively, we can use the built-in function `t.test` by specifying `var.equal = FALSE`:

```
> t.test(c(data.gene[class=='BL' | class == 'NB']) ~ as.factor(class[class=='BL' | class == 'NB']),
+   alternative = 'two.sided', var.equal = FALSE)

        Welch Two Sample t-test

data:  c(data.gene[class == "BL" | class == "NB"]) by as.factor(class[class == "BL" | class == "NB"])
t = 1.956, df = 7.924, p-value = 0.08653
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05796539  0.69886539
sample estimates:
mean in group BL mean in group NB
         0.80930          0.48885
```

The boxplot of the expression of this chosen genes for the two groups is displayed in Fig. 3.3. Does the boxplot reflect the conclusion of the statistical test?
This test is also called the Welch two sample t-test.

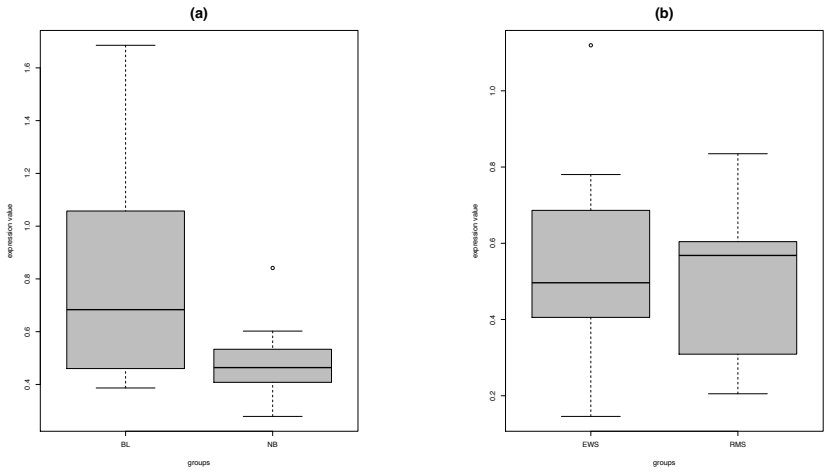

Figure 3.3: Boxplot of the expression levels of a chosen gene for each of the two groups (a): BL and NB (example 21) and (b): EWS and RMS (example 22).

With equal variance. We now make the assumption that the variances in each group are equal (i.e. $s_1^2 = s_2^2$). By making this assumption we can 'pool' the sample variance from the two groups and weight it with respect to the number of patients in each group:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The test statistic follows:

$$T.S = t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

Exercise 22 *For the same gene as illustrated in the above example 21, we would like to test if there is a difference between the expression means between the groups 'EWS' and 'RMS'. Is the use of a t-test with equal variances valid? Conclude on the t-test*
We have $\mu_{EWS} = 0.5311217$, $s^2_{EWS} = 0.04547513$, $n_{EWS} = 23$ *and* $\mu_{RMS} = 0.5009$, $s^2_{RMS} = 0.03608333$, $n_{RMS} = 20$.

Alternatively, we can use the built-in function `t.test` by specifying `var.equal = TRUE`:

```
> t.test(c(data.gene[class=='EWS' | class == 'RMS']) ~ as.factor(class[class=='EWS' | class == 'RMS']),
+   alternative = 'two.sided', var.equal = TRUE)

        Two Sample t-test

data:  c(data.gene[class == "EWS" | class == "RMS"]) by as.factor(class[class == "EWS" | class == "RMS"]
t = 0.4874, df = 41, p-value = 0.6285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09499128  0.15543476
sample estimates:
mean in group EWS mean in group RMS
        0.5311217         0.5009000
```

3.3.3 F-test

To compare the variance between two groups. To validate the assumption of the t-test on the equality of the variances, we can test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ (i.e. $H_0 : \sigma_1^2/\sigma_2^2 = 1$) against $H_0 : \sigma_1^2 \neq \sigma_2^2$. As mentionned in Chapter 2, Section 2.3.4, The F statistics

$$T.S. = F = \frac{s_1^2}{s_2^2} \sim \mathcal{F}_{(n_1-1),(n_2-1)}$$

Exercise 23 *We test the equality of variance for both cases presented in examples 21 and 22. State the null and alternative hypothesis in both cases and conclude on the F-test outputs using the* ***`var.test`*** *R function.*

```
> var.test(data.gene[class=='EWS'], data.gene[class=='RMS'], alternative = 'two.sided')
```

```
        F test to compare two variances

data:  data.gene[class == "EWS"] and data.gene[class == "RMS"]
F = 1.2603, num df = 22, denom df = 19, p-value = 0.6149
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5085289 3.0331510
sample estimates:
ratio of variances
          1.260281

> var.test(data.gene[class=='BL'], data.gene[class=='NB'], alternative = 'two.sided')

        F test to compare two variances

data:  data.gene[class == "BL"] and data.gene[class == "NB"]
F = 10.1921, num df = 7, denom df = 11, p-value = 0.0009716
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  2.711651 47.999456
sample estimates:
ratio of variances
          10.19211
```

To compare the means between more than two groups. More generally, the F-test, through an ANOVA (ANalysis Of VAriance), enables to test the equality of means in K groups

$$H_0 : \mu_1^2 = \cdots = \mu_K^2 \qquad \text{vs.} \qquad H_1 : \exists(j,k)/\mu_j^2 \neq \mu_k^2$$

for $(j \neq k) \in K$, i.e. the alternative hypothesis is: *'There is at least one group which mean is different from another group mean', or 'no mean is the same'.*

Assumption of the Analysis of Variance.

1. The samples are *independent* random samples, i.e. the results from one sample do not affect the measurements observed in another sample.
2. Each sample is selected from a *normal* population.
3. The mean and variance for population or group k are, respectively, μ_k and σ_k^2, $k = 1, \ldots, K$. The K *variances are equal*: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$.

Analysis of Variance.

The one-way ANOVA breaks the total variability of the data into the error variability within groups and the variability between groups. Each variability component is summarized by a sum of squares deviations and a degree of freedom. The formula for the one-way ANOVA F-test statistic is

$$F = \frac{\text{between group variability}}{\text{within group variability}} \sim \mathcal{F}_{K-1,n-K}$$

Traditionally, the information to test the nullity of the parameters in a linear regression is presented in an analysis of variance table (ANOVA), as shown in Table 3.2.

SST is the total sum of squared variation and since the ANOVA is breaking up the variance into different sources, we have $SST = SSW + SSB$.

Source	deg. Freedom	Sum of Squares	Mean Squares	F
Between group variability	$K-1$	SSB	SSB$/(K-1)$	F
Within group variability (Residuals)	$n-K$	SSW	SSW$/(n-K)$	
Total	$n-1$	SST		

Table 3.2: Analysis of Variance Table.

Exercise 24 *Combining the data from exercises 22 and 22, we perform an ANOVA on the expression of a given gene with respect to the four groups of tumor (see the* `R` *code below). State the null and alternative hypothesis, identify each cell from Table 3.2 and draw a conclusion.*
Based on the previous examples in this Chapter, do you think the assumptions of the ANOVA are verified in this example?

```
> summary(aov(data.gene ~ class))

             Df Sum Sq Mean Sq F value Pr(>F)
class         3  0.639 0.21303   3.792 0.0148 *
Residuals    59  3.314 0.05617
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Remark 11 *In the latter exercise we used the function* `aov`*. Another built-in function that can be used is* `oneway.test` *which gives the same results at the one-way ANOVA (note the argument* `var.equal = TRUE`*).*

```
> oneway.test(data.gene ~ class, var.equal = TRUE)

        One-way analysis of means

data:  data.gene and class
F = 3.7923, num df = 3, denom df = 59, p-value = 0.01483
```

Remark 12 *A one-way ANOVA testing the mean for two groups will give the same results as a t-test for equal variances.*

3.4 Multiple testing

Rationale behind multiple testing. In the ANOVA test in exercise 24, if the null hypothesis is rejected, the test tells us if at least one mean group is different from the other mean groups, but its does not tell us which groups have means that are significantly different. One straightforward procedure is to perform pairwise t-tests between each of the two groups. However, when K is large there are $K(K-1)/2$ comparisons to perform. If we set up a significance level at $\alpha = 5\%$ (i.e. there is 5% chance of making a mistake by wrongly rejecting the null hypothesis), it means that there is a 95% chance of not making a mistake. If we perform 3 pairwise t-tests, the probability of making no mistake is $0.95 \times 0.95 \times 0.95 = 0.8574$ (if we assume the tests independents). So the chance of making at least one mistake is 14.26%.

The probability of falsely rejecting at least one of the hypotheses increases as the number of tests increases. Thus, even if we have the probability of type I error at $\alpha = 5\%$ for each individual test, the probability of falsely rejecting *at least one* of those tests is larger that 0.05.

Some procedures were proposed to adjust the level $\alpha' < \alpha$ or to define the critical value $t_{1-\alpha/2,n-K}$. For example the Bonferroni test sets $\alpha' = \alpha/(K(K-1)/2)$. Tukey's 'Honest Significant Difference' method give studentized confidence intervals with *adjusted* p-values.

Exercise 25 *Interpret the Tukey multiple comparisons of means test following last exercise 24:*

```
> TukeyHSD(aov(data.gene ~ class), conf.level = 0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = data.gene ~ class)

$class
               diff         lwr         upr     p adj
BL-EWS    0.27817826  0.02098086  0.53537566 0.0291015
NB-EWS   -0.04227174 -0.26541004  0.18086656 0.9585542
RMS-EWS  -0.03022174 -0.22180165  0.16135817 0.9753484
NB-BL    -0.32045000 -0.60645526 -0.03444474 0.0222160
RMS-BL   -0.30840000 -0.57052815 -0.04627185 0.0148513
RMS-NB    0.01205000 -0.21675421  0.24085421 0.9990257
```

Rationale about confidence intervals. Suppose that each random variables $(x_1, x_2, \ldots, x_n)$ follow a Normal distribution $\mathcal{N}(\mu, \sigma)$.

- The sample mean $\bar{x}$ roughly has a Normal distribution $\mathcal{N}(\mu, \sigma/\sqrt{n})$.
- In a Normal distribution, about 95% of the observations occur within 1.96 standard deviations of the mean μ. So in 95% of the samples, the sample mean $\bar{x}$ will be within $1.96\sigma/\sqrt{n}$ of μ.
- Reversing this, 95% of the samples of μ will be within $1.96\sigma/\sqrt{n}$ of $\bar{x}$

This means that when we use the sample mean to estimate the population mean, we can also give an idea of how far away the population mean could be from our estimate. We say that we are *95% confident* that the poulation mean is

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}},$$

or alternatively that the population mean is in the interval

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

which is the confidence interval for the population mean.

This allows us to say something about a population based on our sample, if σ was known.
The confidence interval is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest, if the experiment is repeated. Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter.
Caution: a confidence interval does *not predict* that the true value of the parameter has a particular probability of being in the confidence interval given the data actually obtained.

Multiple testing for highly dimensional data sets In high-throughput experiments such as microarrays or next-generation sequencing data, where the expression or the counts of hundreds of thousands of genes, transcripts are measured, we perform one test on each of these variables. Multiple correction needs to be applied in order to control the number of false positives genes or transcripts (i.e that are declared differentially expressed between different conditions although they are not). Commonly used multiple testing adjustments procedure include 'Benjamini and Hochberg' False Discovery Rate (FDR) or the most conservative Bonferroni correction. The practicals in `Bioconductor` will emphasise on this very important issue.

3.5 Summary

The traditional approach of hypothesis testing consists of 5 parts: research hypothesis, null hypothesis, test statistic, rejection region, checking asusmptions and drawing conclusions. A statistical test employs the technique of proof by contradiction (experiments are conducted to veryfy the hypothesis through the contradiction of the null hypothesis).
We considered statistical tests with categorical and continuous variables, one and two sample tests, F-test and ANOVA, as well as the problem of multiple testing.

Chapter 4

Clustering of large scale data

Many techniques have become available recently that produce vast amounts of quantitative biological data. These techniques include for example RNA sequencing, various gene expression techniques and protein expression analysis. In this Chapter, we will introduce common clustering techniques applied to high-throughput data.
High throughput, whole genome DNA microarrays enable to monitor the simultaneous expression of multiple genes. These techniques can reveal which genes are expressed together, or *co-expressed*, which might then lead to the identification of genes that might be functionally related. This information can then be used to help assign possible functions to unindentified genes with the same expression patterns.

In the first part of this Chapter, we will consider hierarchical clustering, one of the most widely used approach for analyzing patterns of gene expression microarray data. We will then introduce Principal Component Analysis (PCA), another popular tool used not only for dimension reduction (to highlight the most 'important' information from the data) but also for clustering. Finally, we will present the $k-$means algorithm as another clustering tool.

4.1 Distances

We are interested in genes that have similar expression profiles: each gene is measured across n samples and is represented by a vector of length n.

Definition 22 *Theoretical properties of a measure of similarity (or dissimilarity) between two sets of measurements (for example the expression levels of two genes, A and B):*

- *d(A, B)* ≥ 0. *The distance between two gene profiles must be strictly greater than 0.*
- *d(A, A) = 0. The distance between a profile and itself must be 0.*
  d(A, B) = 0 $\Longleftrightarrow$ *A = B. Conversely, if the distance between two profiles is zero, then the profiles must be identical.*
- *d(A, B) = d(B, A). The distance between profile A and profile B must be the same as the distance between profile B and profile A.*
- *d(A, C)* $\leq$ *d(A, B) + d(B, C) (also known as the triangle inequality rule).*

4.1.1 Euclidian distance

The Euclidian distance is an extension of distance we use in everyday life. It is the stright-line distance betwen points in a 2 or 3 dimensional space.

In two dimensions, the distance between two points is calculated using the Pythagerean theorem.
In three dimensions, say we have two points $\boldsymbol{A}$ and $\boldsymbol{B}$ with coordinates (x_A, y_A, z_A) and (x_B, y_B, z_B), then the Euclidian distance between them is defined as:

$$d(\boldsymbol{A}, \boldsymbol{B}) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$$

It is the same idea for the gene expression profiles measured in a n dimensional space. We can then extend the definition to higher dimensions:

$$d(\boldsymbol{A}, \boldsymbol{B}) = \sqrt{\sum_{i=1}^{n} (X_{i,A} - X_{i,B})^2}$$

The Euclidian distance is commonly used and is easy to evaluate. However, the problem with this distance is that it is *not scale invariant*. Two genes with similar shapes but with different magnitude will appear to be very distant. This may be observed for genes whose transcription is coordinated but do not necessarily produce equivalent response. This problem can be resolved by centering the profiles. Or by using an other distance, such as the correlation distance.

4.1.2 Pearson's correlation distance

Consider a distance between two points, $\boldsymbol{A}$ and $\boldsymbol{B}$. The definition of the Pearson's correlation measure for two sets of expression levels $X_A = \{X_{1,A}, X_{2,A}, \ldots, X_{n,A}\}$ and $X_B = \{X_{1,B}, X_{2,B}, \ldots, X_{n,B}\}$ is given by:

$$r(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_{i,A} - \bar{X}_A}{s_A}\right)\left(\frac{(X_{i,B} - \bar{X}_B}{s_B}\right)$$

where $\bar{X}_A$ is the sample mean of the values in X_A, and s_A is the sample standard deviation of the values in X_A (same for $\bar{X}_B$ and s_B).
As already seen in Chapter 1 subsection 1.5.2, the correlation value ranges from -1 (complete negative correlation) through 0 (no correlation) to +1 (perfect correlation).
In order to obtain a measure of similarity, we need to convert this measure into a distance measure with the properties listed above. We can use either

$$d((\boldsymbol{A}, \boldsymbol{B}) = 1 - |r(\boldsymbol{A}, \boldsymbol{B})|$$

or

$$d((\boldsymbol{A}, \boldsymbol{B}) = 1 - r(\boldsymbol{A}, \boldsymbol{B})^2$$

The Pearson's correlation coeffcient measures distance in terms of the shape of the patterns, not its absolute value (as opposed to the Euclidian distance). A limitation of this distance is that we might want to assign greater significance when both genes are highly expressed than when they are both poorly expressed.
Another limitation is that this standard correlation coefficient is susceptible to being skewed by outliers (as underlined in Chapter 1). Spearman's correlation is a non-parametric measure of correlation that is robust to outliers.

4.2 Hierarchical clustering

Definition 23 *A **clustering** or a **cluster analysis** aims at assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms.*

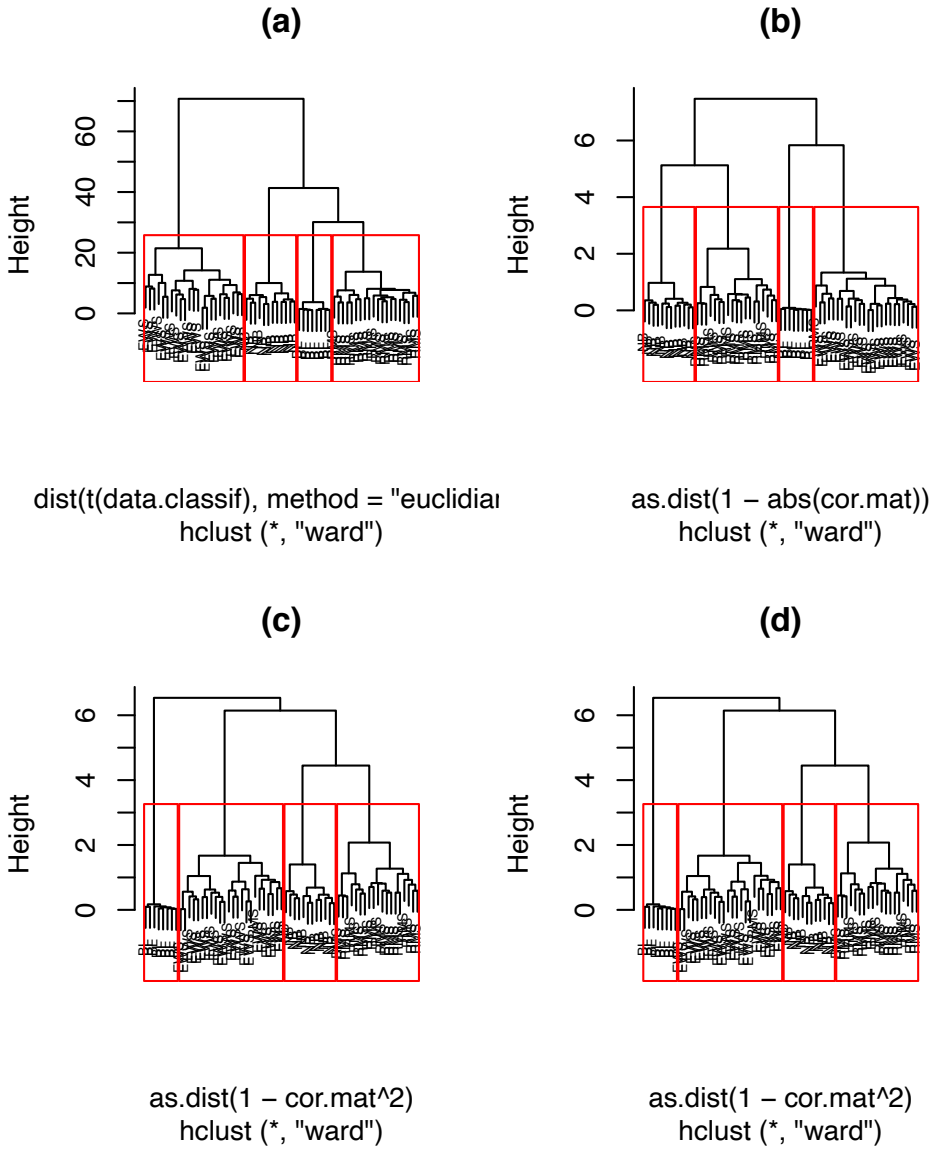

Figure 4.1: Examples of different cluster dendograms using different distances: (a): Euclidian, (b): Pearson's correlation coefficient distance $1-|\text{cor}|$, (c): Pearson's correlation coefficient distance $1-\text{cor}^2$, (d) Spearman's correlation coefficient distance $1-\text{cor}^2$. The red rectangles drawn around the branches indicate 4 clusters.

4.2.1 Dendrogram

A dendrogram is a tree diagram, visually representing clusters amongst variables or samples. The dendrogram does not only represent a single clustering, but rather a multilevel hierarchy, and the height of each node is proportional to the value of the intergroup similarity between the two lower nodes. A hierarchy of clusters is built, with usually a bottom up approach: the closest pair of observation are first joined together to form a cluster, and the process is repeated until there is no more observation to merge: at the lowest level, each cluster contains a single observation, and at the highest level there is only one cluster containing all the data. Therefore, it is essential to specify a metric - or distance between pairs of observations, as well as a linkage criterion to specify the dissimilarity of sets. The metric will influence the shape of the clusters

(Euclidian or 1-correlation distances are often used), whereas the linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations (often the Ward's method is used with the Euclidian distance).

There are some common ways to define the proximity of clusters. For example,

- *Single linkage clustering* uses the minimum distance between two objects in the clusters.
- *Complete linkage clustering* defines the proximity as the maximum of the distances between all possible pairs of objects in each cluster.
- The **Ward criterion** minimizes the total within-cluster variance. At each step the pair of clusters with minimum cluster distance are merged. The initial cluster distances in Ward's minimum variance method are therefore defined to be the squared Euclidean distance between points.

Further analysis is then needed after the clustering to decide which partitions are useful. By cutting off the dendrogram at various heights, different numbers of clusters emerge, and the sets of clusters are nested within one another.The choice of partitions to use in experimental interpretation is almost always subjective, but can be governed by extra information not available during the clustering process (for example set of genes having the same biological function, or patient outcomes, suggesting that the cluster has some biological meaning).

Exercise 26 *Comment on the dendrograms on the samples obtained in Fig. 4.1.*

4.2.2 Heatmap

Hierarchical clustering methods enable hierarchical representations of a measure of dissimilarity between groups of observations (i.e. groups of genes and groups of patients in our SRBCT example). The measure of dissimilarity is based on pairwise dissimilarities among the observations in the two groups.

In the context of microarray data, a **heatmap** representation arranges both the rows and the columns of the expression matrix in orderings derived from hierarchical clustering. By cutting the dendrograms at various heights, different number of clusters emerge and the set of clusters are nested within one another. This kind of representation is useful to interpret the gene clusters in terms of biological processes for example.

Figures 4.2 displays such dendrograms, using the Euclidian distance and the Ward method. The top of the figures shows that the highest break separates the class EWS from the other classes. If we cut the right hand side of the plot further, we are able to separate the NB patients from BL and RMS. The dendrogram on the left hand side clusters the genes with similar profiles across the samples.

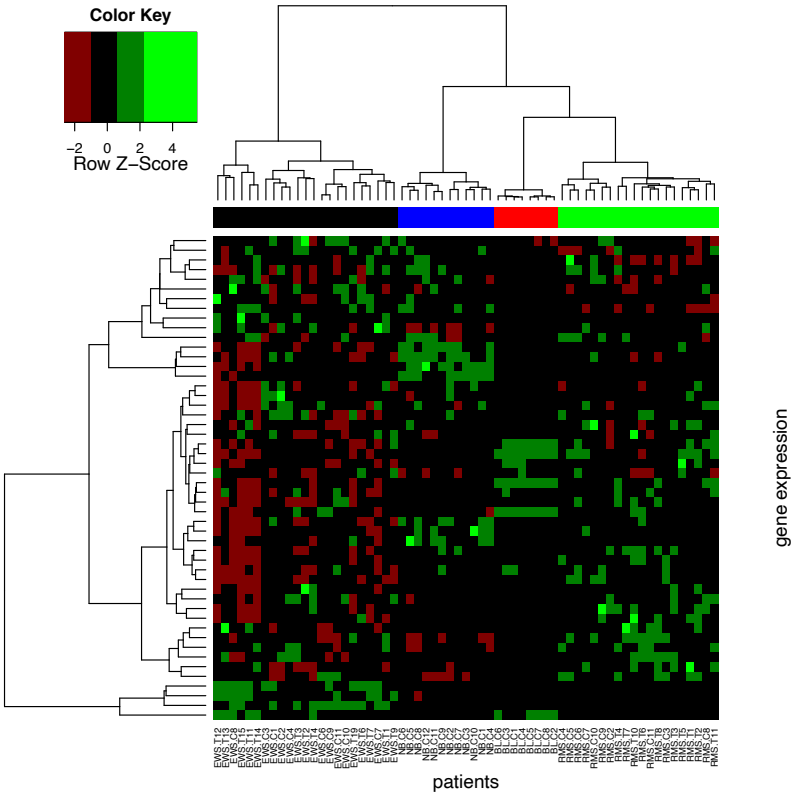

Figure 4.2: SRBCT data. Heatmap of hierarchical clustering applied independently to the rows (50 genes) and columns (63 patients) determining the ordering of the rows and columns. The colors range from brigh green (negative or under expression) to bright red (positive or overexpression). Euclidian distance and Ward linkage criterion were used.

4.3 Illustrative data set: metabolome analysis of yeast

The illustrative example that will be used in the remaining of this chapter is a data set from a yeast study (Villas-Boas et al (2005)[1]). In this data set, two Saccharomyces cerevisiae strains were used: a reference strain (wild-type: WT) and a mutant (MT) were carried out in batch cultures under two different environmental conditions in standard mineral media with glucose as the sole carbon source. The authors assayed metabolite levels in the two yeast strains (WT and MT) and two different environmental conditions, aerobic and anaerobic perturbations (AER and ANA). After normalization and pre processing, the metabolome data results in 37 metabolites and 55 samples which include 13 MT-AER, 14 MT-ANA, 15 WT-AER and 13 WT-ANA samples.

Biological question. One of the main questions when analyzing high throughput data is whether the information provided by the metabolites spectra relate to the experimental conditions, or rather, to some interfering signals. In this chapter, we are focusing on different techniques to vizualize datasets in a 'blind' (unsupervised) way, i.e. when the biological background information, such as group affiliation or class label is not used in the statistical approaches. The aim is to represent the major or global information from the data sets without experimental knowledge.

[1]Villas-Boas et al., 2005 'High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts', *Biochem. J.* **388**, 669-677

4.4 Principal Component Analysis

A well-established technique for *vizualization* and *extraction of relevant information* is the popular Principal Component Analysis (PCA).

4.4.1 Principal Component Analysis

Principle. The aim of PCA[2] is to reduce the dimensionality of the data while retaining as much information as possible. 'Information' is referred here as *variance*. The idea is to create uncorrelated artificial variables called principal components (PCs) that combine in a linear manner the existing and possibly correlated variables (here the genes, or the metabolites). The dimension is reduced by projecting the data into the smaller subspace spanned by the PCs, while capturing the largest sources of variation between the samples. The principal components are obtained by maximising the variance-covariance matrix of the data, finding eigenvalue of the variance-covariance matrix or using singular value decomposition when the number of variables is very large. The data are usually centered, and sometimes scaled. Missing values are not allowed, unless using the NIPALS (nonlinear iterative partial least squares [3]) algorithm which also enables an estimation of the missing values.
The first PC is defined as the linear combination of the original variables that explains the greatest amount of variation. The second PC is then defined as the linear combination of the original variables that accounts for the greatest amount of the remaining variation subject of being orthogonal (uncorrelated) to the first component. Subsequent components are defined likewise for the other PCA dimensions. The user must therefore keep in mind how much information is explained by the first PCs as these are used to graphically represent the PCA outputs.

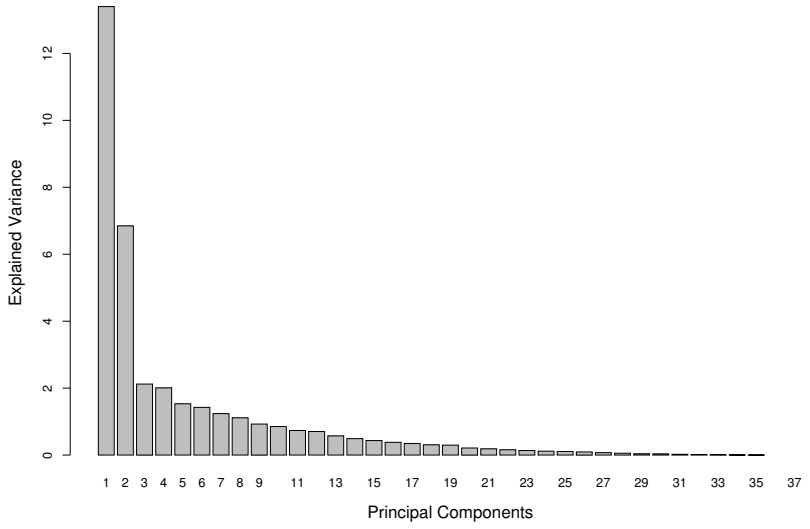

Figure 4.3: Principal Component Analysis of the yeast data set: barplot of the explained variance on each PC. This output is useful to choose the number of PCs to retain in the PCA analysis

Choosing the PCA dimension. We can obtain as many dimensions (i.e. number of PCs) as the number of variables. However, the goal is to reduce the complexity of the data and therefore summarize the data in fewer underlying dimension.
Fig. 4.3 displays the barplot of the eigenvalues associated with each PC . One criterion to select the number of PCs to retain in the analysis is to find the spot where the smooth decrease of the eigenvalues appears to level off to the right of the plot (i.e. when the 'elbow' appears). These eigenvalues correspond to the amount of variance explained by the components. Another criterion is the clarity of the final configuration (see next

[2] Joliffe (2002), 'Principal Component Analysis', *Springer-Verlag.*
[3] Wold (1987), 'Principal Component Analysis', *Chemometrics and Intelligent Laboratory Systems* **2**: 37-52

section). All of this is highly subjective and the reader must keep in mind that visualization becomes difficult above 3 dimensions.
Fig. 4.3 suggests that two PCs might be satisfactory to visualise most information from the yeast data. A usual output that can be obtained using statistical softwares is the cumulative percentage of explained variance (or information): in the yeast data, two PCs explained 54.72% of the total variance, and three PCs explained 60.45% of the total variance.

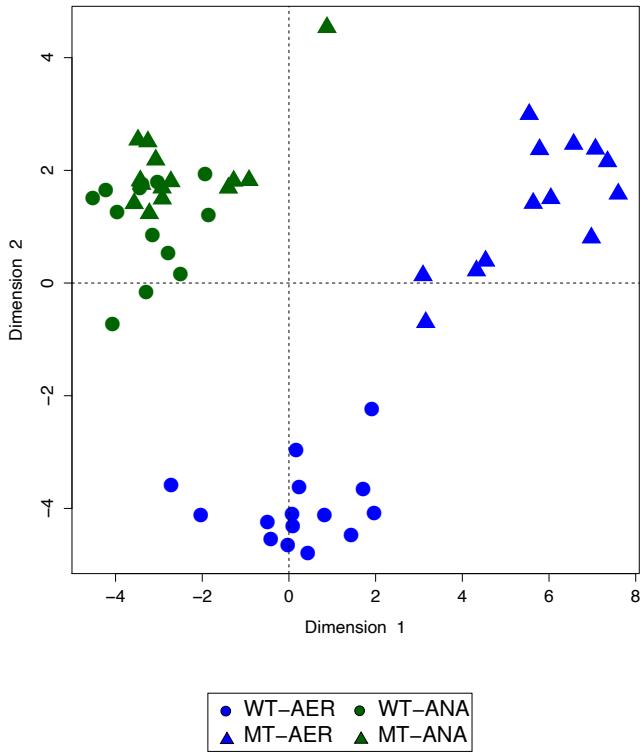

Figure 4.4: Principal Component Analysis of the yeast study and representation of the samples on the first two principal components (denoted 'Dimension 1' and 'Dimension 2'). Each dot represents a sample. Anaerobic and aerobic conditions are separated on the first PC.

Graphical outputs. PCA is an extremely valuable visualization tool to explore a dataset. It can reveal the discriminatory structure, as well as experimental bias in the data. Two types of graphical outputs can be obtained in PCA:

- *Sample representation* can be obtained by plotting the principal components to observe the similarities between the samples which account for most variation, but also to give a 'meaning' to the PCs. For example, in Figure 4.4, the first PC tends to discriminate anaerobic vs. aerobic conditions, whereas the second PC tends to discriminate the wild type aerobic vs. the other conditions. Remark that as noted above, only 2 PCs might be satisfactory enough to summarize most of the information from these data.
- A biplot allows to graphically display *both samples and variables*. Samples are displayed as dots while variables are usually displayed as vectors. If the data are centered and scaled, the cosine angle between the variable vector and the PC indicates the correlation coefficient between the variable and the PC. This is therefore a useful way to give a meaning to each PC. For example, Fig. 4.5 allows to identify

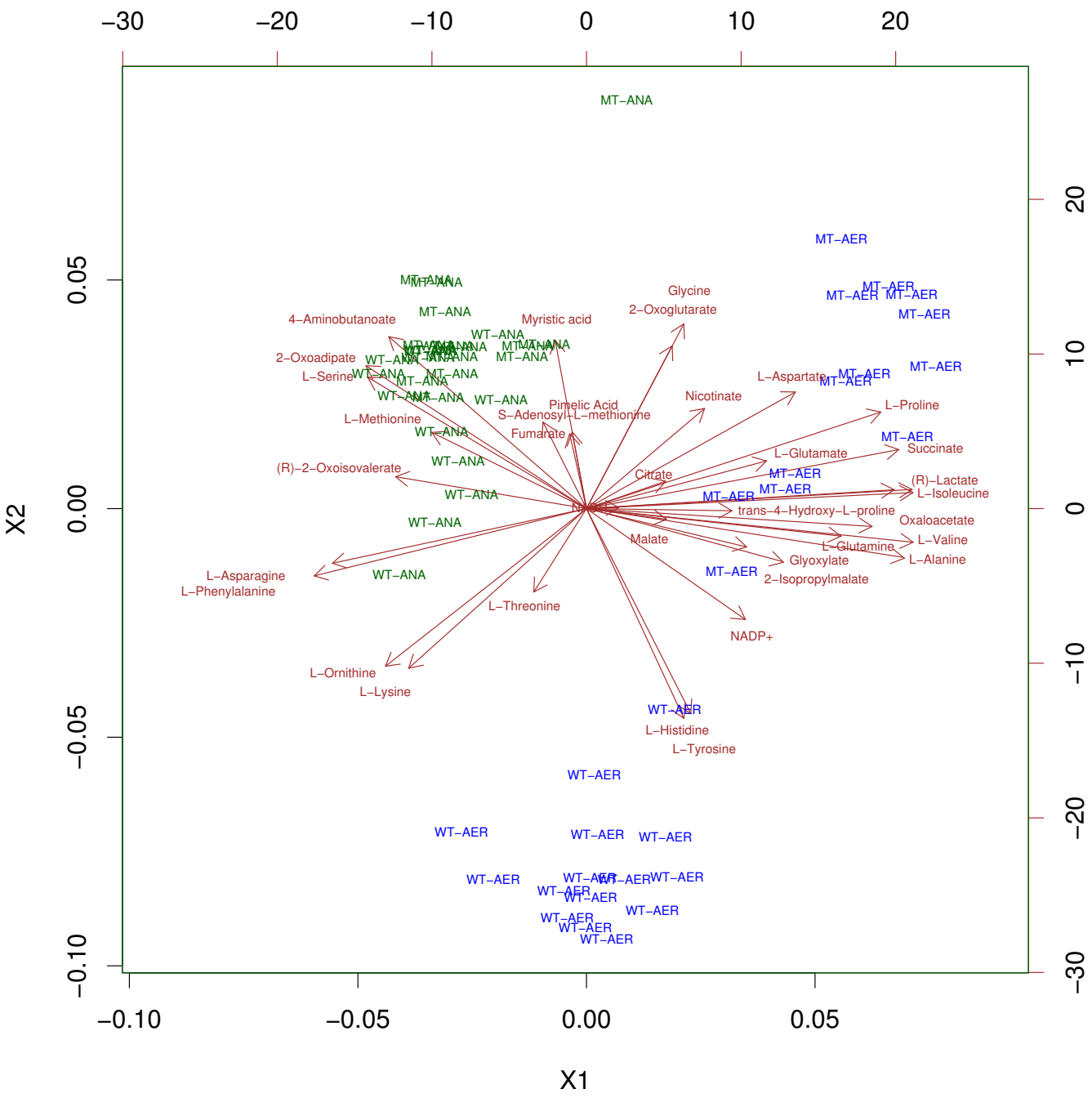

Figure 4.5: Biplot from the PCA analysis on the yeast data, simultaneous representation of the samples (dots) and variables (vectors) on the first 2 PCs. Clusters of metabolites correlated with the biological conditions can be identified.

which metabolites are highly correlated (negatively or positively) to the first and second PCs. The metabolites represented with long arrows and highly correlated with the principal components are the ones that explain most of the variation between the different conditions. For example the group of metabolites pointing towards MT-AER are highly expressed in this group, but under expressed in the anaerobic group.

Remark 13 *Note that using a smaller number of preselected metabolites (i.e. removing noise) may improve*

the quality of such graphical outputs.

4.5 K-means

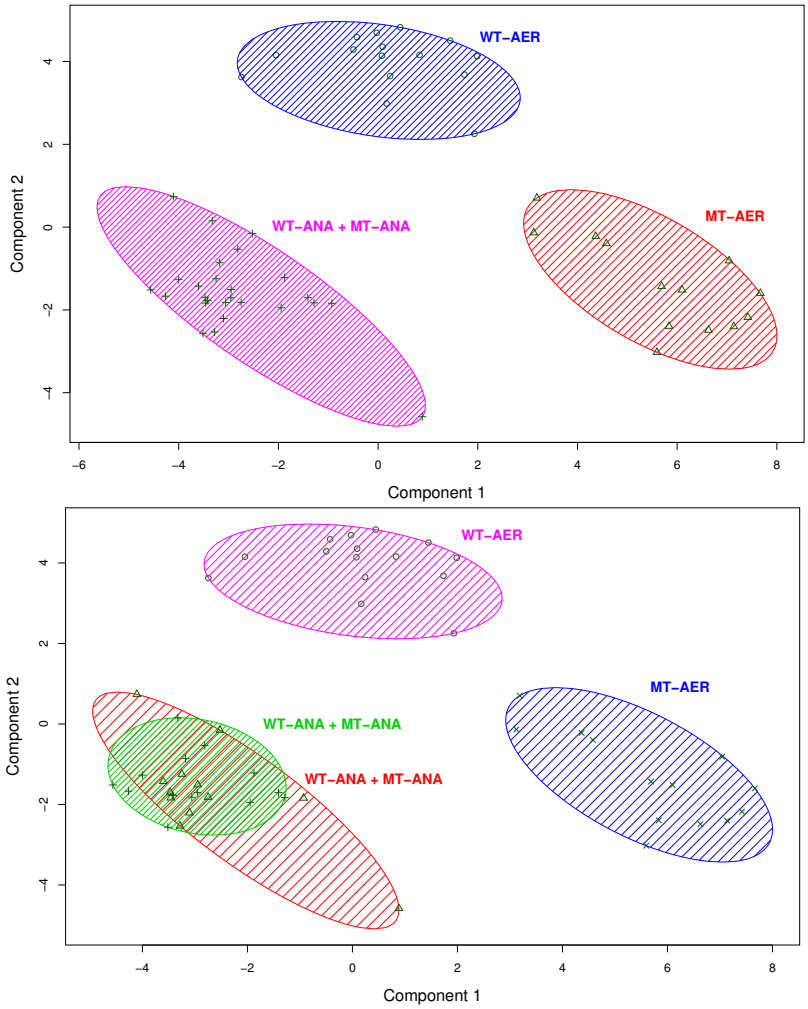

Figure 4.6: K-means clustering of the yeast data set and representation of the samples on the first two PCs components, for 3 clusters (top) and 4 clusters (bottom). Each dot represents a sample.

While hierarchical clustering and PCA do not require to specify a fixed number of clusters, another partitioning clustering method that does so is the $k-$ means approach[4] which partitions the data into k clusters. The choice of k is completely subjective.

The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The k-means algorithm will not necessarily find the global optimum solution and is sensitive to the initial randomly selected cluster centres. In step 1 in the algorithm described below, a different initial location of the cluster centroids can result in a different final partition. It is advisable to use several different starting points, generating several partitions. The $k-$ means algorithm also makes the assumption that clusters are spherical and of similar size. Despite these limitations, the algorithm is used

[4]MacQueen (1967). 'Some Methods for classification and Analysis of Multivariate Observations'. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.*

fairly frequently as a result of its ease of implementation and fast convergence. Depending on the data set, it might work well or fail on other data sets.

Description of the K-means algorithm.

1. Place k points into the space represented by the objects (the samples) that are being clustered. These points represent the initial set of centers.
2. Assign each object to the group that has the closest center.
3. When all objects have been assigned, calculate the means of each feature for the objects in each cluster. This mean vector becomes the new center for that cluster.
4. Repeat Steps 2 and 3 until the centers no longer move.

There is no general theoretical solution to find the optimal number of clusters for any given data set. A simple approach is to compare the results of multiple runs with different k classes and choose the best one according to a given criterion, such as Bayesian information criterion (BIC) or Akaike information criterion (AIC), or a visual assessment. Figures 4.6 show that $k = 3$ seemed to be more appropriate than $k = 4$, even though there are 4 real classes. The k-means algorithm was successful at grouping together WT-AER, ANA and MT-AER conditions (note that in order to visualise the data, they were first summarized in 2 dimensions using PCA). A shortcoming of k-means is that the clusters need not be nested within the three clusters. Therefore, hierarchical clustering described below might sometimes be preferable.

4.6 Summary

In this Chapter, we have seen that there are different ways of measuring the similarity between gene expression profiles. The distance measure that we use, for example, can affect the results.
Hierarchical clustering can be used to identify related genes or samples and portray them using dendrograms. Different distances and linkage methods will produce different results.
PCA is a dimension reduction technique and provides a good way to visualise the data, using sample and variable plots together (biplot).
K-means is also a well known clustering methodology that can be used on large scale data and requires the specification of a number of clusters.

Appendix A

Statistical tables

A.1 Binomial Table

Example of table of the binomial cumulative probability distribution function.

This table gives the cumulative distribution function (c.d.f) of the binomial distribution, i.e. the distribution of the number of successes in n independent trials of an experiment which leads to a success with probability p. The c.d.f. is

$$\mathbb{P}(X \leq k) = \sum_{i=0}^{k} \mathbb{P}(X = i)$$

Since this table only cover $p \leq 0.5$, the roles of successes and failures need to be reversed for $p > 0.5$ (see **Exercise** 15).

n	k	p	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	0		0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1		0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2		0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3		0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4		0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5		0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6		1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7		1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8		1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0		0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0004
	1		0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
	2		0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3		0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4		0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5		0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
	6		1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
	7		1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8		1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9		1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
	10		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
	11		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0		0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1		0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
	2		0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
	3		0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
	4		0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
	5		0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
	6		0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
	7		1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
	8		1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
	9		1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
	10		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
	11		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	12		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

A.2 Statistical table for a Normal Standard Distribution

This table gives $\mathbb{P}(Z \leq z)$ where $Z \sim \mathcal{N}(0,1)$.

Normal cumulative distribution function

	second decimal place of z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

A.3 Statistical table for Student t-test

Percentage points of Student's t distribution (right-tail probability α)

df	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Remark 14 *For two-tailed tests, use value in column headed by* $\alpha/2$