2

Empirical Statistical Methods

Statistics deals to a large extend with the process of transforming observed data from one or more experiments into a probabilistic model – that is a probability measure on the sample space E. In many cases we may want to obtain an understanding of the data in a more superficial way – if we just want to understand some aspects of the probability measure, say. The methods developed in this chapter are called empirical methods or descriptive statistics, because the methods developed mostly serve the purpose of summarising and describing the empirical data in a convenient way. We will also discuss some of the theoretical properties of the empirical methods that we develop.

The assumption throughout the entire chapter is that we observe n iid variables $X_1, \ldots, X_n$ taking values in a sample space E. A realisation of these variables (the outcome from the experiment) is denoted $x_1, \ldots, x_n$. Depending on the nature of the sample space (is it continuous or discrete, is it multivariate or univariate) we develop some of the most frequently used methods for summarising the data.

2.1 Continuous distributions and Quantiles

Definition 2.1.1. *A histogram with break points $q_1 < q_2 < \ldots < q_k$, chosen so that*

$$q_1 < \min_{i=1,\ldots,n} x_i \leq \max_{i=1,\ldots,n} x_i < q_k,$$

is the function h given by

$$h(x) = \frac{1}{q_{i+1} - q_i} \varepsilon_n\left((q_i, q_{i+1}]\right) \quad \text{for } q_i < x \leq q_{i+1}. \tag{2.1}$$

together with $h(x) = 0$ for $x \notin (q_1, q_n]$. Usually the plot of h with a box of height $h(q_{i+1})$ located over the interval $(q_i, q_{i+1}]$ is what most people associate with a histogram.

The function h is constructed so that

$$\begin{aligned}\int h(x) &= \sum_{i=1}^{k-1} \int_{q_i}^{q_{i+1}} \frac{1}{q_{i+1} - q_i} \varepsilon_n\left((q_i, q_{i+1}]\right) \mathrm{d}x \\ &= \sum_{i=1}^{k-1} \varepsilon_n\left((q_i, q_{i+1}]\right) \\ &= \varepsilon_n((q_1, q_n]) = 1\end{aligned}$$

where we use that all the data points are contained within the interval $(q_1, q_n]$. Since the function h integrates to 1 it is a probability density. The purpose of the histogram is to approximate the density of the true distribution of X – assuming that the distribution has a density.

Sometimes one encounters the *unnormalised* histogram, given by the function

$$\tilde{h}(x) = n\varepsilon_n(q_i, q_{i+1}] \quad \text{for } q_i < x \leq q_{i+1}.$$

Here $\tilde{h}(x)$ is constantly equal to the number of observations falling in the interval $(q_i, q_{i+1}]$. Since the function doesn't integrate to 1 it can not be compared directly with a density.

Example 2.1.2. We consider the histogram of 100 and 1000 simulated $N(0, 1)$ iid stochastic variables. We choose the breaks to be equidistant from -4 to 4 with a distance of 0.5, thus the break point are

$$-4 \quad -3.5 \quad -3 \quad -2.5 \quad \ldots \quad 2.5 \quad 3 \quad 3.5 \quad 4.$$

We find the histograms in Figure 2.1. Note how the histogram corresponding to the 1000 simulated stochastic variables approximates the density more closely.

Example 2.1.4. Throughout this section we will consider data from a microarray experiment. It is the so-called ALL dataset (Chiaretti et. al., Blood, vol. 103, No. 7, 2004). It consists of samples from patients suffering from Acute Lymphoblastic Leukemia. We will consider only those patients with B-cell ALL, and we will group the patients according to presence or absence of the BCR/ABL fusion gene.

On Figure 2.2 we see the histogram of the log (base 2) expression levels[1] for six (arbitrary) genes for the group of samples without BCR/ABL.

On Figure 2.3 we have singled out the gene with the poetic name 1635_ at, and we see the histograms for the log expression levels for the two groups with or without BCR/ABL. On the figur you also find the empirical distribution functions.

If $x_1, \ldots, x_n \in \mathbb{R}$ are n real observations from an experiment, we can order the observations

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

[1] Some further normalisation has also been done, that we will not go into here

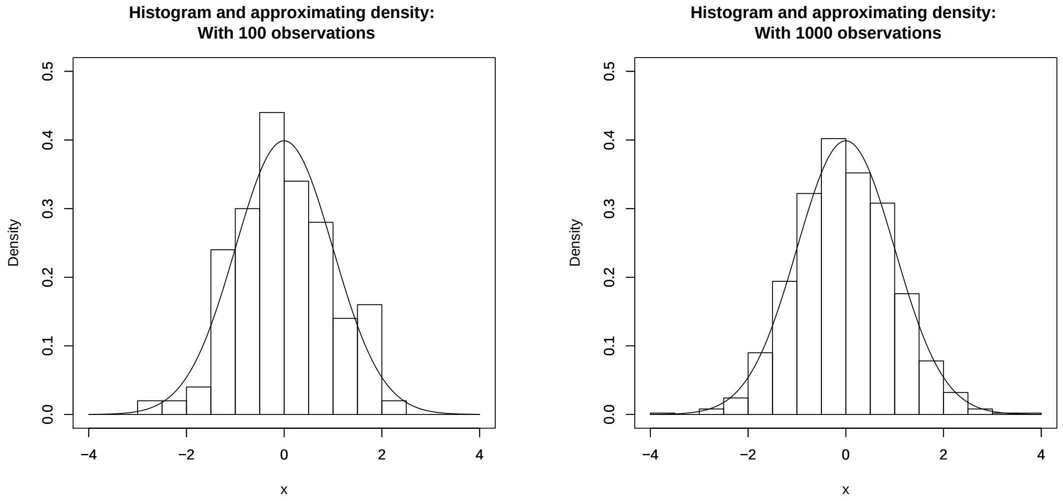


Figure 2.1: The histograms for the realisation of 100 (right) and 1000 (left) simulated iid $N(0,1)$ stochastic variables. For both histograms we compare the histogram with the corresponding density for the normal distribution.

R Box 2.1.3 (Histograms)**.** A histogram of the data in the numeric vector `x` is produced in R by the command

```
> hist(x)
```

This automatically opens a graphics window and plots a histogram using default settings. The break points are by default chosen by R in a suitable way. It is possible to explicitly set the break points by hand, for instance

```
> hist(x,breaks=c(0,1,2,3,4,5))
```

produces a histogram with break points $0,1,2,3,4,5$. Note that R will produce an error if the range of the break points does not contain all the data points in `x`. Note also that the default behaviour of `hist` is to plot the *unnormalised* histogram if the break points are equidistant. Otherwise it produces the normalised histogram. One can always make `hist` produce normalised histograms by

```
> hist(x,freq=FALSE)
```

thus $x_{(1)}$ denotes the smallest observation, $x_{(n)}$ the largest and $x_{(i)}$ the observation with $i-1$ smaller observations. If $q = i/n$ for $i = 1,\ldots,n$, then $x \in \mathbb{R}$ is called a *q-quantile*

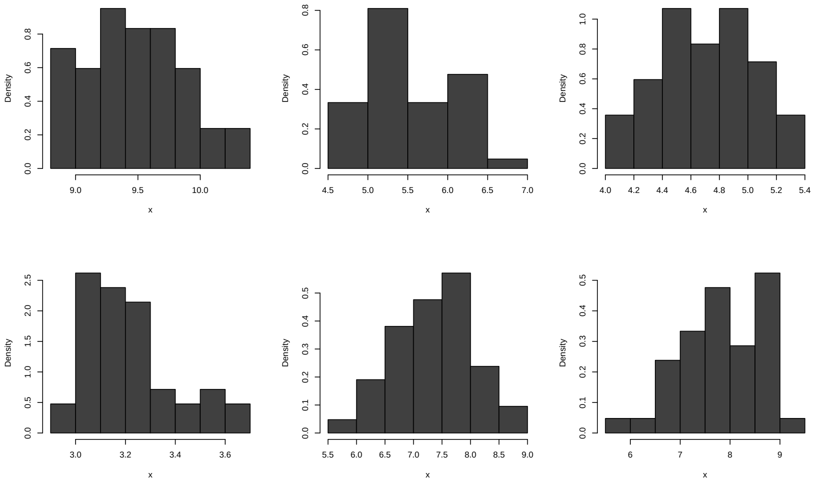


Figure 2.2: Histograms

(for the dataset) if the fraction q of the observations that are $\leq x$. This means that if $x_{(i)} \leq x \leq x_{(i+1)}$ then x is a i/n-quantile. Note that for a given $q = i/n$ there is a whole range of q-quantiles, namely the whole interval $[x_{(i)}, x_{(i+1)}]$.

If $i/n < q < (i+1)/n$ it is slightly more tricky how one should define a q-quantile, but the proper definition is that then $x_{(i+1)/n)}$ is the only q-quantile. This is the only definition that assures *monotonicity* of quantiles in the sense that if x is a q-quantile and y is a p-quantile with $q < p$ then $x < y$.

Some quantiles have special names, e.g. a 0.5-quantile is called a median, and the upper and lower quartiles are the 0.75- and 0.25-quantiles respectively. Note the ambiguity here. If n is even then *all* the x's in the interval $[x_{(n/2)}, x_{(n/2+1)}]$ are medians, whereas if n is odd the median is uniquely defined as $x_{((n+1)/2)}$. This ambiguity of e.g. the median and other quantiles can be a little annoying in practice, and sometimes one prefers to define a single (empirical) quantile function $Q : (0,1) \to \mathbb{R}$ such that for all $q \in (0,1)$ we have that $Q(q)$ is a q-quantile. Whether one prefers to say that $x_{(n/2)}$, $x_{(n/2+1)}$, or perhaps $(x_{(n/2)} + x_{(n/2+1)})/2$ is the median if n is even is largely a matter a taste.

Quantiles can also be defined for theoretical distributions. We prefer here to consider the definition of a quantile function only.

Definition 2.1.7. *If $F : \mathbb{R} \to [0,1]$ is a distribution function for a probability measure P on $\mathbb{R}$, then $Q : [0,1] \to \mathbb{R}$ is a quantile function for P if*

$$F(Q(y) - \varepsilon) \leq y \leq F(Q(y)) \tag{2.2}$$

for all $y \in [0,1]$ and all $\varepsilon > 0$.

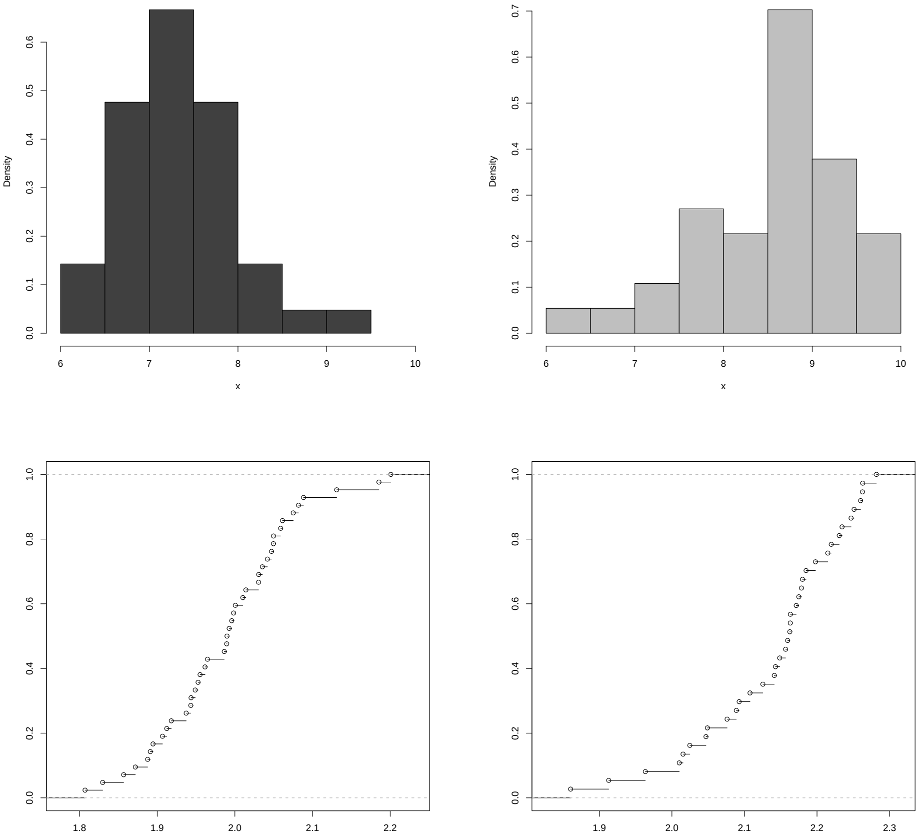


Figure 2.3: Histograms and empirical distribution functions of log (base 2) expression levels for the gene 1635_ at from the ALL microarray experiment with (right) or without (left) precense of the BCR/ABL fusion gene.

Theorem 2.1.8. *The generalised inverse distribution function $F^{\leftarrow}$, cf. Section 1.8, is a quantile function.*

Proof: To see this, first observe that with $x = F^{\leftarrow}(y)$ then

$$F^{\leftarrow}(y) \leq x \Rightarrow y \leq F(x) = F(F^{\leftarrow}(y))$$

by the definition of $F^{\leftarrow}$. On the other hand, suppose that there exists a $y \in [0, 1]$ and an $\varepsilon > 0$ such that $F(F^{\leftarrow}(y) - \varepsilon) \geq y$ then again by the definition of $F^{\leftarrow}$ it follows that

$$F^{\leftarrow}(y) - \varepsilon \geq F^{\leftarrow}(y),$$

which can not be the case. Hence there exists no such $y \in [0, 1]$ and $\varepsilon > 0$ and

$$F(F^{\leftarrow}(y) - \varepsilon) < y$$

R Box 2.1.5 (Empirical distribution functions)**.** If `x` is a numeric vector in R containing our data we can construct a `ecdf`-object (empirical cumulative distribution function). This requires the `stats` library:

```
> library(stats)
```

Then

```
> edf <- ecdf(x)
```

gives the empirical distribution function for the data in `x`. One can evaluate this function like any other function:

```
> edf(1.95)
```

gives the value of the empirical distribution function evaluated at 1.95. It is also easy to plot the distribution function:

```
> plot(edf)
```

produces a nice plot.

R Box 2.1.6 (Quantiles)**.** If `x` is a numeric vector then

```
> quantile(x)
```

computes the 0%, 25%, 50%, 75%, and 100% quantiles. That is, the minimum, the lower quartile, the medium, the upper quartile, and the maximum.

```
> quantile(x,probs=c(0.1,0.9))
```

computes the 0.1 and 0.9 quantile instead, and by setting the `type` parameter to an integer between 1 and 9, one can select how the function handles the non-uniqueness of the quantiles. If `type=1` the quantiles are given as the generalised inverse of the empirical distribution function. Note that some choices of `type` actually produce a result that violates our definition of quantiles.

for all $y \in [0,1]$ and $\varepsilon > 0$. This shows that $F^{\leftarrow}$ is a quantile function. $\square$

There may exist other quantile functions besides the generalised inverse of the distribution function, which are preferred from time to time. However, if F has an inverse function then the inverse is the only quantile function.

Definition 2.1.9. *If F is a distribution function and Q a quantile function for F the median or second quartile of F is defined as*

$$q_2 = \text{median}(F) = Q(0.5).$$

In addition we call $q_1 = Q(0.25)$ and $q_3 = Q(0.75)$ the first end third quartiles of X. The difference

$$\text{IQR} = q_3 - q_1$$

is called the interquartile range.

Note that the definition of the median and the quartiles depend on the choice of quantile function. If the quantile function is not unique these numbers are not necessarily uniquely defined. The median summarises in a single number the location of the probability measure given by F. The interquartile range gives a single value telling how spread out around the median the distribution is.

An important observation that binds the definition of a quantile function for any distribution function F and the quantiles defined for a given dataset is, that if

$$F_n(x) = \varepsilon_n((-\infty, x]) \tag{2.3}$$

denotes the empirical distribution function for the observations then any quantile function for the distribution function F_n also gives empirical quantiles as defined for the dataset.

One of the applications of quantiles and the empirical quantile function is to compare two distributions by comparing their quantiles.

Definition 2.1.10. *If F_1 and F_2 are two distribution functions with Q_1 and Q_2 their corresponding quantile functions a QQ-plot is a plot of Q_1 against Q_2.*

R Box 2.1.11 (QQ-plots)**.** If `x` and `y` are numeric vectors then

```
> qqplot(x,y)
```

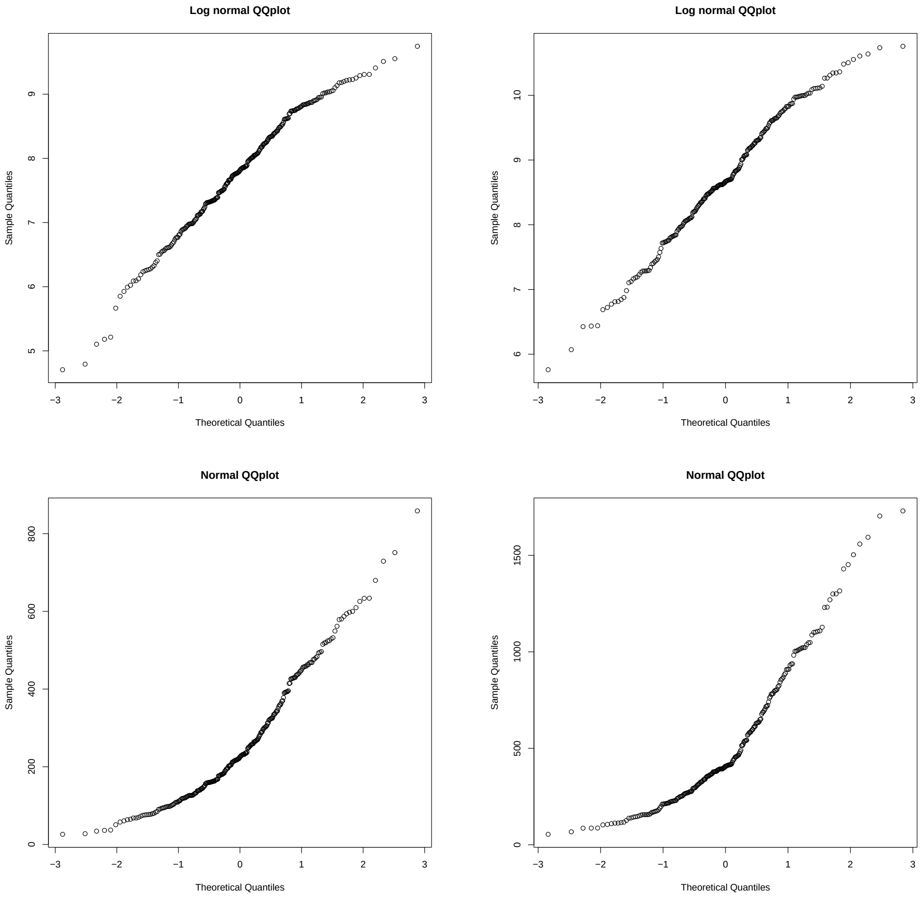
produces a QQ-plot of the empirical quantiles for `y` against those for `x`.

```
> qqnorm(x)
```

results in a QQ-plot of the empirical quantiles for `x` against the quantiles for the normal distribution.

Usually when making a QQ-plot one of the distributions, F_1, say, is empirical. It is then common only to plot

$$(Q_2(i/n), x_{(i)}), \quad i = 1, \ldots, n-1,$$


Log normal QQplot

Log normal QQplot

Normal QQplot

Normal QQplot

Figure 2.4: QQplots for gene 1635_ at from the ALL dataset. Here we see the expression levels and log (base 2) expression levels against the normal distribution with (right) or without (left) precense of the BCR/ABL fusion gene.

choosing the generalised inverse of F_1 as quantile function. If the empirical quantile function Q_1 is created from a realisation of n iid stochastic variables having distribution function F with quantile function Q_2 then the points in the QQ-plot should lie close to a straight line with slope 1 and intercept 0. It can be beneficial to plot the straight line to be able to visualise any discrepancies from the straight line.

We are often interested in comparing the empirical distribution with a distribution where we know the *form* of the distribution but not the location and scale. If X has distribution with quantile function Q_2 and our dataset is a realisation of n iid stochastic variables each having the same distribution as

$$\sigma X + \mu$$

for some unknown scale $\sigma > 0$ and position $\mu \in \mathbb{R}$, then if we make a QQ-plot of the empirical quantile function against Q_2 it will still result in points that lie close to a straight line, but with different slope and intercept.

One could also compare distribution functions directly instead of comparing quantile functions. It is, however, often more difficult to see the differences between two distribution functions. Especially if the differences are mostly occurring in the tails of the distribution functions. Then the differences will show up nicely on a QQ-plot but may be undetectable by comparing distribution functions directly.

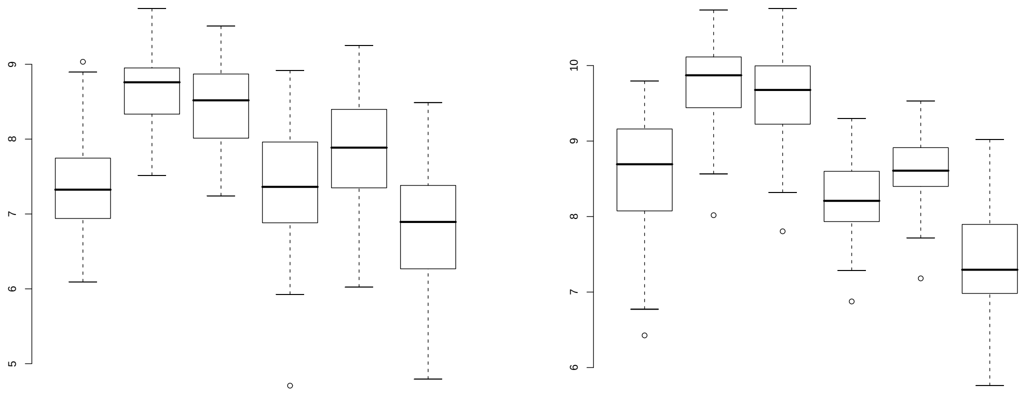


Figure 2.5: Comparing the empirical distributions of the six genes 1635_ at,1636_ g_ at,39730_ at,40480_ s_ at, 2039_ s_ at, 36643_ at for those with BCR/ABL (right) and those without (left) using boxplots

Histograms are useful for representing a single empirical distribution and QQ-plots are valuable for comparing an empirical distribution with another empirical distribution or a theoretical distribution. The *box plot* is a useful tool for visualising and comparing three or more empirical distributions. It may also be useful for visualising just a single empirical distribution if all you want is a rough picture of location and scale.

Definition 2.1.13. *One defines a box plot using quantile function Q and whisker coefficient $c > 0$ in terms of a five-dimensional vector*

$$(w_1, q_1, q_2, q_2, w_2)$$

with $w_1 \leq q_1 \leq q_2 \leq q_3 \leq w_2$. Here

$$q_1 = Q(0.25), \quad q_2 = Q(0.5), \quad q_3 = Q(0.75)$$

are the three quartiles and

$$\begin{aligned} w_1 &= \min\{x_i \mid x_i \geq q_1 - c(q_3 - q_1)\} \\ w_2 &= \max\{x_i \mid x_i \leq q_3 + c(q_3 - q_1)\} \end{aligned}$$

R Box 2.1.12 (Box plots)**.** For a numeric vector `x` we get a single box plot by

```
> boxplot(x)
```

If `x` is a dataframe the command will instead produce (in one figure) a box plot of each column. By specifying the `range` parameter (= whisker coefficient), which by default equals 1.5, we can change the length of the whiskers.

```
> boxplot(x,range=1)
```

produces a box plot with whisker coefficient 1.

are called the whiskers. The box plot is drawn as a vertical box from q_1 *to* q_3 *with "whiskers" going out to* w_1 *and* w_2*. If datapoints lie outside the whiskers they are often plotted as points.*