

# Chapter 4

## Estimation and Inference

In this chapter we will use hypothesis tests, based on the distributions presented in Chapter 3, to test biological hypotheses about populations based on sample data. These questions include:

- Does the mean gene expression of the ALL patients differ from that of the AML patients?
- Is the mean gene expression different from zero?
- Are the gene expression values normally distributed?
- Are there outliers among a sample of gene expression values?
- How can an experimental effect be defined?
- How can genes be selected with respect to an experimental effect?

In the following chapters, many population parameters are used to define theoretical distributions that attempt to model biological phenomena. In any empirical research setting, the specific values of such population parameters are unknown and must be estimated. When any population estimate is postulated, this creates an hypothesis about the population that can be statistically tested using sample data. This chapter provides the framework for hypothesis testing and gives several basic real-world examples. In addition, robust types of testing are briefly introduced as well as outlier testing.

## 4.1 Statistical hypothesis testing

Let  $\mu_0$  be a number representing the hypothesized population mean proposed by a researcher on the basis of experience and knowledge from the field. The null hypothesis put forward by the researcher can be formulated as  $H_0 : \mu = \mu_0$  and the alternative hypothesis as  $H_a : \mu \neq \mu_0$ . These are two mutually exclusive statements of which the latter is the opposite of the first - either  $H_0$  or  $H_a$  is true, but not both. Note that the alternative hypothesis  $H_a : \mu \neq \mu_0$  is true if and only if  $H_a : \mu < \mu_0$  or  $H_a : \mu > \mu_0$  holds true. Because of this relationship, an  $H_a : \mu \neq \mu_0$  type of alternative hypothesis is called *two-sided*. In the case when the alternative hypothesis is just  $H_a : \mu > \mu_0$  (or  $H_a : \mu < \mu_0$ ), then the alternative is called *one-sided*.

In hypothesis testing, the null hypothesis is statistically tested against the alternative using a suitable distribution of a sample statistic (e.g. standardized mean). We test the validity of the null hypothesis about the population by assuming that it is true, and then calculating the probability of observing the alternative hypothesis just by chance in our sample data. If we consider the probability of observing the alternative in the sample data far to improbable to occur just by chance if the null hypothesis is true, then we reject the null hypothesis in favor of the alternative. Otherwise, we conclude that there isn't enough evidence to reject the null hypothesis in favor of the alternative. After conducting an experiment or obtaining sample data, the value of the sample statistic is computed from the data. Next, we assume that null hypothesis is true. Then by comparing the value of the sample statistic with the  $H_0$  hypothesized distribution, we draw a conclusion with respect to the null hypothesis:  $H_0$  is rejected or it is not. The probability of observing  $H_a$  at which we reject  $H_0$  is called the *significance level* and is generally denoted by  $\alpha$ . The significance level commonly used is  $\alpha = 0.05$ , but sometimes other significance levels are desired.

Note that it is not correct to say that we *accept* the null hypothesis. Instead, we can only say that we *fail to reject* or *cannot reject* the null hypothesis - given the observed evidence for or against the null hypothesis. Just because we cannot reject the null hypothesis given the sample data does not imply that the null hypothesis is necessarily true - so we should not *accept* it.

### 4.1.1 The Z-test

The Z-test applies to the situation where we want to test  $H_0 : \mu = \mu_0$  against a two-sided ( $H_a : \mu \neq \mu_0$ ) or one-sided ( $H_a : \mu > \mu_0$  or  $H_a : \mu < \mu_0$ ) hypothesis and the population standard deviation  $\sigma$  is known (which is usually not the case). Assuming that the gene expression values  $(x_1, \dots, x_n)$  are from a normal distribution, we compute the standardized *test statistic*  $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ . Next we define the so-called *p-value* as the probability of  $Z$  attaining values being equal to or more extreme than  $|z|$ , which for a two-sided test is the probability of occurring to the left of  $-|z|$  or to the right of  $|z|$ .<sup>1</sup> Accordingly, the two-sided *p-value* equals:

$$P(Z \leq -|z|) + P(Z \geq |z|) = 2 \cdot P(Z \leq -|z|)$$

The conclusion from the test is now as follows: if the *p-value* is larger than the significance level  $\alpha$ , then  $H_0$  is not rejected and if it is smaller than the significance level, then  $H_0$  is rejected in favor of  $H_a$ .

Note that these *p-values* constitute the areas under the tails of the normal probability density function (PDF). In general, when calculating *p-values* from any probability density function, we wish to find the probability of observing our sample in the extreme tail(s) of the null-hypothesis distribution. Table 4.2 summarizes the three different types of *p-value* calculations given the null hypothesis  $H_0 : \mu = \mu_0$ :

Table 4.1: Types of one-sample *p-value* calculations for  $H_0 : \mu = \mu_0$ .

Type	Alternative Hypothesis	Probability
(Left) lower-tailed test	$H_a = \mu < \mu_0$	$P(X \leq \text{test statistic})$
(Right) upper-tailed test	$H_a = \mu > \mu_0$	$P(X \geq \text{test statistic})$
Two-tailed test	$H_a = \mu \neq \mu_0$	$P(X \leq - \text{test statistic} ) + P(X \geq  \text{test statistic} )$

**Example 1: Hypothetical Gdf5 gene expression.** To illustrate the Z-test we will look at the Gdf5 gene from the Golub et al. (1999) data<sup>2</sup>. The Gdf5 expression values are contained in row 2058. A quick search through the NCBI site makes it likely that this gene is not directly related to leukemia. Hence, we may hypothesize that the population mean of the ALL expression

<sup>1</sup>Recall from a calculus course that  $|-2| = 2$  and  $|2| = 2$ .

<sup>2</sup>We will work with `golub` throughout this chapter, so it is essential to load these data and to define the factor `gol.fac`.

values equals zero. Accordingly, we test  $H_0 : \mu = 0$  against  $H_a : \mu \neq 0$ . For the sake of illustration we shall pretend that the population standard deviation  $\sigma$  is known to be equal to 0.25. The *test statistic*  $z$ -value can be computed as follows:

```
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl, levels=0:1, labels= c("ALL", "AML"))
> sigma <- 0.25; n <- 27; mu0 <- 0
> x <- golub[2058, gol.fac=="ALL"]
> z.value <- sqrt(n)*(mean(x) - mu0)/sigma
> z.value
[1] 0.001116211
```

Then, the  $p$ -value can now be computed as follows:

```
> 2*pnorm(-abs(z.value), 0, 1)
[1] 0.9991094
```

Since it is clearly larger than 0.05, we conclude that the null hypothesis of mean equal to zero is not rejected.

Note that the above procedure implies rejection of the null hypothesis when  $z$  is highly negative or highly positive. More precisely, if  $z$  falls in the region  $(-\infty, z_{0.025}]$  or  $[z_{0.975}, \infty)$ , then  $H_0$  is rejected. For this reason these intervals are called *rejection regions*. If  $z$  falls in the interval  $(z_{0.025}, z_{0.975})$ , then  $H_0$  is not rejected and consequently this region is called the *acceptance region*, although a far better moniker would be the *non-rejection region*. The situation is illustrated in Figure 4.1. We can create Figure 4.1 using the `dnorm()`, `polygon()`, `arrows()`, `seq()`, and `text()` functions:

```
> f<-function(x){dnorm(x,0,1)}
> x1 <- seq(-4,-1.96,0.01)
> y1 <- dnorm(x1,0,1)
> x2 <- seq(-1.96,1.96,0.01)
> y2 <- dnorm(x2,0,1)
> x3 <- seq(1.96,4,0.01)
> y3 <- dnorm(x3,0,1)
> plot(f, # function
+      -4, # begin x-value
+      4, # end x-value
+      cex.lab=1.5, # make axis labels big
+      xlab="x",
+      ylab="Normal probability density function f(x)")
> polygon(c(-4,x1,-1.96), c(0,y1,0), col="red")
> polygon(c(-1.96,x2,1.96), c(0,y2,0), col="lightblue")
> polygon(c(1.96,x3,4), c(0,y3,0), col="red")
> arrows(-3,0.15,-3,0.03)
> text(-3,0.23,"Rejection")
> text(-3,0.20,"Region")
> text(-3,0.17,expression(alpha/2))
> arrows(3,0.15,3,0.03)
```

## The Z-test Using the Normal PDF

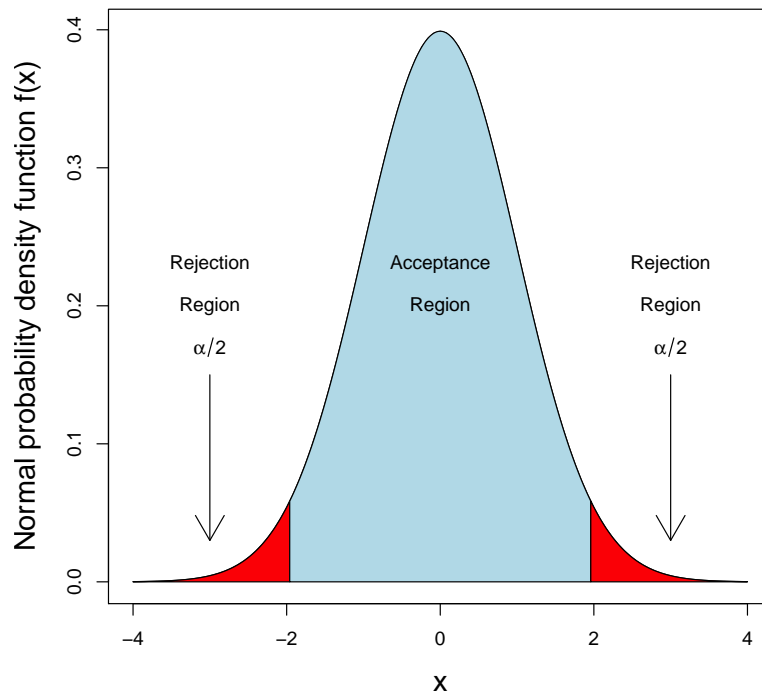


Figure 4.1: Acceptance and rejection regions of the two-tailed  $Z$ -test as determined by the significance level  $\alpha$ .

```
> text(3,0.23,"Rejection")
> text(3,0.20,"Region")
> text(3,0.17,expression(alpha/2))
> text(0,0.23,"Acceptance")
> text(0,0.20,"Region")
```

The interval  $(z_{0.025}, z_{0.975})$  is often named the *confidence interval* because if the null hypothesis is true, then we are 95% confident that the observed  $z$ -value falls in that range. It is custom to rework the confidence interval into an interval with respect to  $\mu$  (Samuels & Witmer, 2003, p. 186). In particular, the 95% confidence interval for the population mean  $\mu$  is:

$$\left( \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.975} \frac{\sigma}{\sqrt{n}} \right). \quad (4.1)$$

That is, we are 95% certain<sup>3</sup> that the true mean falls in the confidence interval. Such an interval is standard output of most statistical software.

**Example 2: Confidence interval.** Using the data from Example 1, the 95% confidence interval given by Equation 4.1 can be computed as follows:<sup>4</sup>

```
> mean(x)+c(-1,1) * qnorm(c(0.975),0,1)*s/sqrt(n)
[1] -0.09424511  0.09435251
```

Hence, the rounded estimated 95% confidence interval is  $(-0.094, 0.094)$ . Since  $\mu_0 = 0$  falls within this interval,  $H_0$  is not rejected. It is instructive to run the  $Z$ -test from the `TeachingDemos` package, as follows:

```
> library(TeachingDemos)
> z.test(x,mu=0,sd=0.25)

      One Sample z-test

data:  x
z = 0.0011, n = 27.000, Std. Dev. = 0.250, Std. Dev. of the sample mean
= 0.048, p-value = 0.9991
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.09424511  0.09435251
sample estimates:
 mean of x
5.37037e-05
```

From the  $z$ -value, the  $p$ -value, and the confidence interval, the conclusion is not to reject the null-hypothesis that the mean is equal to zero. This illustrates that testing by either of these procedures yields equivalent conclusions.

**Example 3: Teaching demonstration.** To develop an intuition with respect to confidence intervals, load the package `TeachingDemos` and give the following command:

```
> ci.examp(mean.sim=0, sd = 1, n = 25, reps = 100,
+ method = "z", lower.conf=0.025, upper.conf=0.975)
```

<sup>3</sup>If we were to repeat the procedure thousands of times

<sup>4</sup>These computations only work together with those of Example 1, especially the definition of `x`.

In the above example, 100 samples of size 25 from the  $N(0, 1)$  distribution are drawn, and for each of these the confidence interval for the population mean is computed and represented as a line segment. Apart from sampling fluctuations, the confidence level corresponds to the percentage of intervals containing the true mean (colored in black), and the significance level corresponds to intervals not containing the true mean (colored in red or blue).

### 4.1.2 One-sample $t$ -test

In almost all research situations, the population standard deviation  $\sigma$  is unknown - and so the above  $Z$ -test is usually not applicable. In such cases,  $t$ -tests are very useful for testing  $H_0 : \mu = \mu_0$  against  $H_a : \mu \neq \mu_0$ , where the sample standard deviation  $s$  is used in lieu of the population standard deviation  $\sigma$ . The test is based on the *test statistic*  $t$ -value defined by  $t = \sqrt{n}(\bar{x} - \mu_0)/s$ . The corresponding  $p$ -value is defined by  $2 \cdot P(T_{n-1} \leq -|t|)$ . As with the  $Z$ -test,  $H_0$  is not rejected if the  $p$ -value is larger than the significance level and  $H_0$  is rejected if the  $p$ -value is smaller than the significance level. Equivalently, if  $t$  falls in the acceptance region  $(t_{0.025, n-1}, t_{0.975, n-1})$ , then  $H_0$  is not rejected and otherwise it is. For  $n = 6$ , the acceptance and rejection regions are illustrated in Figure 4.2. We can create Figure 4.2 using the `dt()`, `polygon()`, `arrows()`, `seq()`, `mtext()`, and `text()` functions:

```
> f<-function(x){dt(x,5)}
> x1 <- seq(-4,qt(0.025,5),0.01)
> y1 <- f(x1)
> x2 <- seq(qt(0.025,5),qt(0.975,5),0.01)
> y2 <- f(x2)
> x3 <- seq(qt(0.975,5),4,0.01)
> y3 <- f(x3)
> plot(f, # function
+      -4, # begin x-value
+      4, # end x-value
+      xlab="x",
+      ylab="t-Distribution probability density function f(x)")
> polygon(c(-4,x1,qt(0.025,5)), c(0,y1,0), col="red")
> polygon(c(qt(0.025,5),x2,qt(0.975,5)), c(0,y2,0), col="lightblue")
> polygon(c(qt(0.975,5),x3,4), c(0,y3,0), col="red")
> arrows(-3,0.15,-3,0.03)
> text(-3,0.23,"Rejection")
> text(-3,0.20,"Region")
> text(-3,0.17,expression(alpha/2))
> arrows(3,0.15,3,0.03)
> text(3,0.23,"Rejection")
> text(3,0.20,"Region")
> text(3,0.17,expression(alpha/2))
> text(0,0.23,"Acceptance")
> text(0,0.20,"Region")
```

```
> mtext(expression(t[0.025]),side=1,at=qt(0.025,5), col="red")
> mtext(expression(t[0.975]),side=1,at=qt(0.975,5), col="red")
```

The 95% confidence interval for the population mean is given by  $(\bar{x} + t_{0.025} \cdot s/\sqrt{n}, \bar{x} + t_{0.975} \cdot s/\sqrt{n})$ , where the expression  $s/\sqrt{n}$  gives the *standard error of the mean*.

### The $t$ -test Using the $t$ -Distribution PDF

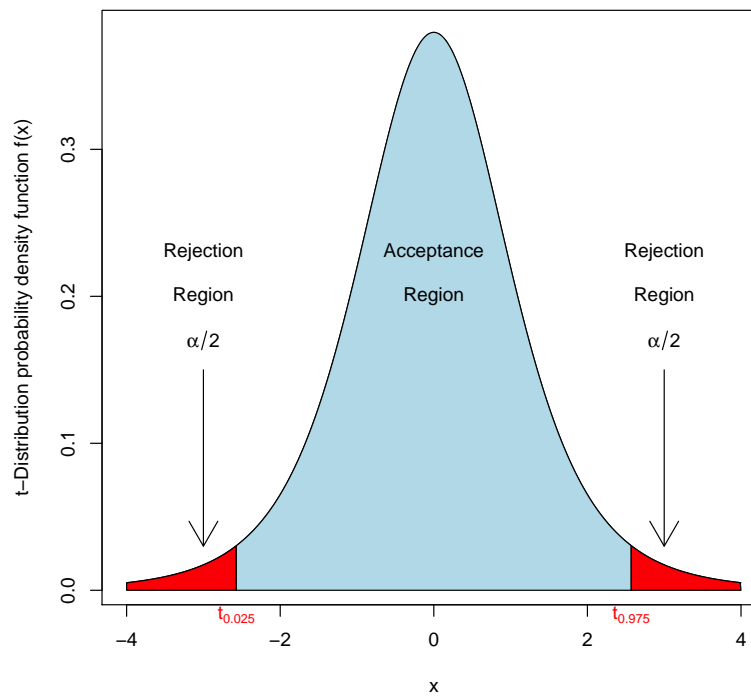


Figure 4.2: Acceptance and rejection regions of the two-tailed  $T_5$ -test as determined by the significance level  $\alpha$ .

**Example 1: Actual Gdf5 gene expression.** Let's test  $H_0 : \mu = 0$  against  $H_a : \mu \neq 0$  for the ALL population mean of the Gdf5 gene expressions. The latter are collected in row 2058 of the golub data. The  $t$ -value is computed as follows:



```
> x <- golub[2058,gol.fac=="ALL"]; mu0 <- 0; n <- 27
> t.value<-sqrt(n)*(mean(x) - mu0)/sd(x)
> t.value
[1] 0.001076867
```

The corresponding  $p$ -value is  $2 \cdot P(T_{26} \leq -0.0010) = 0.9992 > \alpha$  and can be computed by:

```
> 2 * pt(-0.0010,26)
[1] 0.9992097
```

so that the conclusion is not to reject the null hypothesis of the mean equal to zero.

To see whether the observed  $t$ -value belongs to the 95% confidence interval, we compute

$$(t_{0.025,26}, t_{0.975,26}) = (-2.055, 2.055)$$

as follows:

```
> c(qt(0.025,n-1),qt(0.975,n-1))
[1] -2.055529 2.055529
```

Since this interval does contain the  $t$ -value, we do not reject the hypothesis that  $\mu$  equals zero. The left boundary of the 95% confidence interval for the population mean can be computed as follows:

```
> mean(x)+qt(0.025,26)*sd(x)/sqrt(n)
[1] -0.1024562
```

The 95% confidence interval equals  $(-0.1025, 0.1025)$ . Since it contains zero, we do not reject the null-hypothesis.

In practice, it is much more convenient to use the built-in function `t.test()`. We illustrate its usage with the current problem:

```
> t.test(x,mu=0)

One Sample t-test

data:  x
t = 0.0011, df = 26, p-value = 0.9991
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1024562  0.1025636
sample estimates:
 mean of x
5.37037e-05
```

The `t.test()` function yields with one command line the observed  $t$ -value, the  $p$ -value, and the 95% confidence interval for  $\mu_0$ .

In the previous example, the test was two-sided because  $H_a$  holds true if  $\mu < \mu_0$  or  $\mu > \mu_0$ . However, if the researcher desires to test  $H_0 : \mu = \mu_0$  against  $H_a : \mu > \mu_0$ , then the alternative hypothesis is *one-sided* and this makes the procedure slightly different:  $H_0$  is not rejected if  $P(T_n \geq t) > \alpha$  and it is rejected if  $P(T_n \geq t) < \alpha$ . We shall illustrate this by a variant of the previous example.

**Example 2: CCND3 gene expression.** In Chapter 2, a box-and-whiskers plot revealed that the ALL gene expression values of CCND3 (Cyclin D3) are positive. We can test  $H_0 : \mu = 0$  against  $H_a : \mu > 0$  using the built-in `t.test()` function. Recall that the corresponding gene expression values for CCND3 are collected in row 1042 of the `golub` data matrix:

```
> ccnd3 <- grep("CCND3",golub.gnames[,2], ignore.case = TRUE)
> ccnd3
[1] 1042
> t.test(golub[ccnd3,gol.fac=="ALL"],mu=0, alternative = c("greater"))

      One Sample t-test

data:  golub[ccnd3, gol.fac == "ALL"]
t = 20.0599, df = 26, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.732853      Inf
sample estimates:
mean of x
 1.893883
```

The large  $t$ -value indicates that, relative to its standard error, the mean differs greatly from zero. Accordingly, the  $p$ -value is very close to zero, and the conclusion is to reject  $H_0$  in favor of the alternative  $H_a$ .

### 4.1.3 Two-sample $t$ -test with unequal variances (Welch's two-sample $t$ -test)

Suppose that gene expression data from two groups of patients (experimental conditions) are available and that the hypothesis is about the possible difference between the population means  $\mu_1$  and  $\mu_2$ . In particular,  $H_0 : \mu_1 = \mu_2$  is to be tested against  $H_a : \mu_1 \neq \mu_2$ . Note that these hypotheses can also be re-formulated as  $H_0 : \mu_1 - \mu_2 = 0$  and  $H_a : \mu_1 - \mu_2 \neq 0$ . Suppose that gene expression data from the first group are given by  $\{x_1, \dots, x_n\}$  and that of the second by  $\{y_1, \dots, y_m\}$ . Let  $\bar{x}$  be the mean of the first and  $\bar{y}$  that of the second, and  $s_1^2$  the variance of the first and  $s_2^2$  that of the second. Then

the  $t$ -statistic can be formulated as:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n + s_2^2/m}}. \quad (4.2)$$

The decision procedure with respect to the null-hypothesis is completely similar to the above tests. Note that the  $t$ -value is large if the difference between  $\bar{x}$  and  $\bar{y}$  is large<sup>5</sup>, the standard deviations  $s_1$  and  $s_2$  are small, and the sample sizes are large. This test is known as the Welch two-sample  $t$ -test (Lehmann, 1999). Table 4.2 summarizes the three different types of  $p$ -value calculations given the null hypothesis  $H_0 : \mu = \mu_0$ :

Table 4.2: Types of one-sample  $p$ -value calculations for  $H_0 : \mu = \mu_0$ .

Type	Alternative Hypothesis	Probability
(Left) lower-tailed test	$H_a = \mu < \mu_0$	$P(X \leq \text{test statistic})$
(Right) upper-tailed test	$H_a = \mu > \mu_0$	$P(X \geq \text{test statistic})$
Two-tailed test	$H_a = \mu \neq \mu_0$	$P(X \leq - \text{test statistic} ) + P(X \geq  \text{test statistic} )$

**Example 1: CCND3 gene expression.** Golub et al. (1999) argue that gene CCND3 (Cyclin D3) plays an important role with respect to discriminating ALL from AML patients. The box plot in Figure 2.5 suggests that the ALL population mean differs from that of AML. The null hypothesis of equal means can be tested by the function `t.test()` and the appropriate factor and specification `var.equal=FALSE`:

```
> t.test(golub[ccnd3,] ~ gol.fac, var.equal=FALSE)

Welch Two Sample t-test

data: golub[ccnd3, ] by gol.fac
t = 6.3186, df = 16.118, p-value = 9.87e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8363826 1.6802008
sample estimates:
mean in group ALL mean in group AML
 1.8938826         0.6355909
```

The  $t$ -value above is quite large, indicating that the two means  $\bar{x}$  and  $\bar{y}$  differ greatly from zero relative to the corresponding standard error (denominator in Equation 4.2). Since the  $p$ -value is extremely small, the conclusion is to reject the null-hypothesis of equal means. The data provide strong evidence that the population means do differ.

<sup>5</sup> Assuming  $\mu_1 - \mu_2 = 0$ .

When the first group is an experimental group and the second a control group, then  $\mu_1 - \mu_2$  is the experimental effect in the population and  $\bar{x} - \bar{y}$  that in the sample. Likewise, the  $t$ -value is the experimental effect in the sample relative to the standard error. The size of the effect is measured by the  $p$ -value in the sense that it is smaller for larger effects.

#### 4.1.4 Two-sample $t$ -test with equal variances (pooled two-sample $t$ -test)

Suppose exactly the same setting as in the previous paragraph, but now the variances  $\sigma_1^2$  and  $\sigma_2^2$  for the two groups are known to be equal. Then the testing procedure simplifies considerably. To test  $H_0 : \mu_1 = \mu_2$  against  $H_a : \mu_1 \neq \mu_2$ , there is a  $t$ -test which is based on the so-called pooled sample variance  $s_p^2$ . The latter is defined by the following weighted sum of the sample variances  $s_1^2$  and  $s_2^2$ :

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

Then the  $t$ -value can be formulated as

$$t = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

**Example 1: CCND3 gene expression.** The null hypothesis  $H_0$  for gene CCND3 (Cyclin D3) that the mean of the ALL patients differs from that of AML patients can be tested by the two-sample  $t$ -test using the specification `var.equal=TRUE`.

```
> t.test(golub[ccnd3,] ~ gol.fac, var.equal = TRUE)

Two Sample t-test

data: golub[ccnd3, ] by gol.fac
t = 6.7983, df = 36, p-value = 6.046e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8829143 1.6336690
sample estimates:
mean in group ALL mean in group AML
 1.8938826          0.6355909
```

From the  $p$ -value  $6.046 \cdot 10^{-8}$ , the conclusion is to reject the null hypothesis of equal population means. Note that the  $p$ -value is slightly smaller than that of the previous test.

### 4.1.5 Paired, two-sample $t$ -test

The *paired* family of tests is adequate for paired difference tests between matched, labeled samples and repeated measurements. For example, if we wish to test the hypothesis that the mean gene expression is the same for the same patients before and after chemotherapy treatment, then we would use a paired, two-sample  $t$ -test. In the case of repeated measurements in time, as in the previous example, the paired  $t$ -test is also sometimes referred to as the *repeated measures t*-test.

**Example 1: pathway gene expression.** For a family of genes, our null hypothesis is that the mean gene expression for each gene in a certain pathway did not change after chemotherapy treatment. We can test this hypothesis with the `t.test()` function and including the `paired=TRUE` option:

```
> before = c(112.9, 133.5, 122.8, 145.6, 167.2, 169.2, 174.6, 175.3, 184.4, 121.3)
> after  = c(120.7, 123.6, 134.0, 141.2, 156.8, 170.0, 182.0, 185.9, 196.0, 115.1)
> t.test(before, after, paired=TRUE)

      Paired t-test

data:  before and after
t = -0.65483, df = 9, p-value = 0.529
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.240989  4.540989
sample estimates:
mean of the differences
          -1.85
```

With a  $p$ -value of 0.529, we cannot reject the null hypothesis that the mean gene expression for each gene in a certain pathway did not change after chemotherapy treatment.

### 4.1.6 F-test on equal variances

The assumption of the pooled  $t$ -test is that the two population variances are equal. Such an assumption can also serve as a null hypothesis for statistical testing. That is, we desire to test  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_0 : \sigma_1^2 \neq \sigma_2^2$ . This can be accomplished by the so-called  $F$ -test, as follows. From the sample variances  $s_1^2$  and  $s_2^2$ , the *test statistic*  $f$ -value  $f = s_1^2/s_2^2$  can be computed, which is  $F_{n_1-1, n_2-1}$  distributed with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom. If  $P(F_{n_1-1, n_2-1} < f) \geq \alpha/2$  for  $f < 1$  or  $P(F_{n_1-1, n_2-1} > f) \geq \alpha/2$  for  $f > 1$ , then  $H_0$  is not rejected. Otherwise,  $H_0$  is rejected.

**Example 1: CCND3 gene expression.** The null hypothesis for gene CCND3 (Cyclin D3) that the variance of the ALL patients equals that of the AML patients can be tested by the built-in function `var.test()`, as follows:

```
> var.test(golub[ccnd3,] ~ gol.fac)

      F test to compare two variances

data:  golub[ccnd3, ] by gol.fac
F = 0.7116, num df = 26, denom df = 10, p-value = 0.4652
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2127735 1.8428387
sample estimates:
ratio of variances
 0.7116441
```

From the  $p$ -value 0.4652, the null-hypothesis of equal variances is not rejected.

#### 4.1.7 Binomial test

Suppose that for a certain microRNA a researcher wants to test the hypothesis that the probability of a purine equals a certain value  $p_0$ . However, another researcher has reason to believe that this probability is larger. In such a setting, we want to test the null-hypothesis  $H_0 : p = p_0$  against the one-sided alternative hypothesis  $H_a : p > p_0$ . Suppose that sequencing reveals that the microRNA has  $k$  purines out of a total  $n$ . Assuming that the hypothesized binomial distribution holds, the null-hypothesis can be tested by computing the  $p$ -value  $P(X \geq k)$ . If the  $p$ -value is larger than the significance level  $\alpha = 0.05$ , then  $H_0$  is not rejected. If it is smaller, then it is rejected.

**Example 1: A microRNA of length 22 contains 18 purines.** The null hypothesis  $H_0 : p = 0.7$  is to be tested against the one-sided  $H_a : p > 0.7$ . We calculate  $P(X \geq 18) = 0.1645 \geq 0.05 = \alpha$  as follows:

```
> 1 - pbinom(17,22,0.7)
[1] 0.1645488
```

the conclusion follows not to reject the null-hypothesis. This test can also be conducted with the `binom.test()` function as follows:

```
> binom.test(18, 22, p = 0.7, alternative = c("greater"),
+ conf.level = 0.95)
      Exact binomial test
```

```

data: 18 and 22
number of successes = 18, number of trials = 22, p-value = 0.1645
alternative hypothesis: true probability of success is greater than 0.7
95 percent confidence interval:
 0.6309089 1.0000000
sample estimates:
probability of success
 0.8181818

```

The  $p$ -value 0.1645 is larger than the significance level 0.05, so the null hypothesis is not rejected.

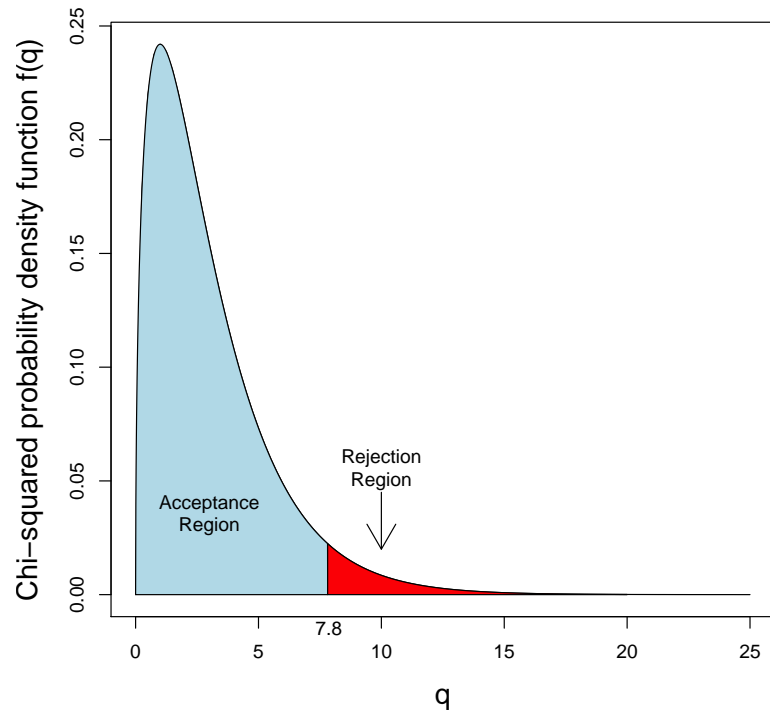
### 4.1.8 Chi-squared test

It often happens that we want to test a hypothesis with respect to more than one probability. That is, we want to test  $H_0 : (\pi_1, \dots, \pi_m) = (p_1, \dots, p_m)$  against  $H_a : (\pi_1, \dots, \pi_m) \neq (p_1, \dots, p_m)$ , where  $p_1$  to  $p_m$  are numbers corresponding to the hypothesis of a researcher. By multiplying the probabilities with the total number of observations, we obtain the expected number of observations ( $e_i = n \cdot p_i$ ). Now we can compute the test statistic  $q = \sum_{i=1}^m (o_i - e_i)^2 / e_i$ , where  $o_i$  is the  $i$ -th observed and  $e_i$  the  $i$ -th expected frequency. If the null hypothesis is true, then this statistic is chi-squared ( $\chi_{m-1}^2$ ) distributed with  $m - 1$  degrees of freedom. The  $p$ -value of the chi-squared test is defined as  $P(\chi_{m-1}^2 \geq q)$ . If it is larger than the significance level, then the null hypothesis is not rejected, and otherwise it is.

**Example 1: Nucleotide content.** Suppose we want to test the hypothesis that the nucleotides in the coding sequence of the Zyxin gene have equal probability. Let the probability of  $\{A, C, G, T\}$  occurring in the sequence be given by  $(\pi_1, \pi_2, \pi_3, \pi_4)$ . Then the null hypothesis to be tested is  $(\pi_1, \pi_2, \pi_3, \pi_4) = (1/4, 1/4, 1/4, 1/4)$ . In particular, for the sequence "X94991.1" from Table 1.1 the total number of nucleotides is  $n = 2166$ , so that the expected frequencies  $e_i$  are equal to  $2166/4 = 541.5$ . Then, the  $q$ -value equals:  $\sum_{i=1}^4 (o_i - e_i)^2 / e_i =$

$$\frac{(410 - 541.5)^2}{541.5} + \frac{(789 - 541.5)^2}{541.5} + \frac{(573 - 541.5)^2}{541.5} + \frac{(394 - 541.5)^2}{541.5} = 187.0674$$

Since  $P(\chi^2[3] \geq 187.0674)$  is close to zero, the null hypothesis is clearly rejected. The nucleotides within the Zyxin gene do not occur with equal probability.

The  $\chi^2$ -test Using the  $\chi^2$  PDFFigure 4.3: Rejection region of upper-tailed  $\chi^2_3$ -test.

A more direct manner to perform the test is by using the built-in function `chisq.test()` as follows:

```
> library(ape)
> zyxinfreq <- table(read.GenBank(c("X94991.1"),as.character=TRUE))
> chisq.test(zyxinfreq)

Chi-squared test for given probabilities

data:  zyxinfreq
X-squared = 187.0674, df = 3, p-value < 2.2e-16
```

Above, the observed frequencies are given as input to `chisq.test()` - which has equal probabilities as the default null hypothesis  $H_0$ . The  $q$ -value equals  $X$ -squared and the degrees of freedom  $df = 3$ . From the corresponding  $p$ -



value, the conclusion is to reject the null hypothesis of equal probabilities. The testing situation is illustrated in Figure 4.3, where the red colored area is the rejection region  $(7.81, \infty)$  and the blue area is the acceptance region. We can create Figure 4.3 using the `dchisq()`, `polygon()`, `arrows()`, `seq()`, `mtext()`, and `text()` functions:

```
> f<-function(x){dchisq(x,3)}
> plot(f, # function
+ 0, # begin x-value
+ 25, # end x-value
+ cex.lab=1.5, # make axis labels big
+ xlab="q",
+ ylab="Chi-squared probability density function f(q)")
> x1<- seq(0,qchisq(0.95,3),0.01)
> x2<- seq(qchisq(0.95,3),20,0.01)
> polygon(c(0,x1,qchisq(0.95,3)), c(0,f(x1),0), col="lightblue")
> polygon(c(qchisq(0.95,3),x2,20), c(0,f(x2),0), col="red")
> arrows(10,0.045,10,0.02)
> text(10,0.06,"Rejection")
> text(10,0.05,"Region")
> text(3,0.04,"Acceptance")
> text(3,0.03,"Region")
> mtext("7.8",side=1,at=7.86)
```

Remember from the previous chapter that the lower bound of this rejection interval can be found by `qchisq(0.95, 3)`. The observed  $q = 187.0674$  obviously falls far into the right hand side of the rejection region, so that the corresponding  $p$ -value is very close to zero.

**Example 2: Mendelian genetics.** In the year 1866, Mendel conducted a large number of experiments on the frequencies of characteristics of different kinds of seed and their offspring. One particular experiment yielded the frequencies 5474 and 1850 for two different seed shapes from ornamental sweet peas. A crossing of B and b yields off spring BB, Bb and bb with probability 0.25, 0.50, 0.25. Since Mendel could not distinguish Bb from BB, his observations theoretically occur with probability 0.75 (BB and Bb) and 0.25 (bb). To test the null hypothesis  $H_0 : (\pi_1, \pi_2) = (0.75, 0.25)$  against  $H_a : (\pi_1, \pi_2) \neq (0.75, 0.25)$ , we use the chi-squared test as follows:

```
> pi <- c(0.75,0.25)
> x <-c(5474, 1850)
> chisq.test(x, p=pi)

      Chi-squared test for given probabilities

data:  x
X-squared = 0.2629, df = 1, p-value = 0.6081
```

From the  $p$ -value 0.6081, we do not reject the null hypothesis  $H_0 : (\pi_1, \pi_2) = (0.75, 0.25)$ .

Below, we present another example to further illustrate the great flexibility of the chi-squared test.

**Example 3: Testing independence.** Given certain expression values for a healthy control group and an experimental group with a disease, we may define a certain cut off value and classify accordingly (e.g. smaller values to be healthy and larger ones to be infected). In such a manner, cut-off values can serve as a diagnostic instrument. The classification yields true positives (correctly predicted disease), false positives (incorrectly predicted disease), true negatives (correctly predicted healthy) and false negatives (incorrectly predicted healthy). For the sake of illustration, suppose that among twenty patients there are 5 true positives ( $tp$ ), 5 false positives ( $fp$ ), 5 true negatives ( $tn$ ), and 5 false negatives ( $fn$ ). These frequencies can be put in a two-by-two table like Table 4.3 giving the frequencies on two random variables: (1) the true disease state  $D$  of the patients, and (2) the predicted state of the patients by the test result  $T$  (as determined by the cut off value for the test).

Table 4.3: Overview of decisions for diagnosis.

true status	measurement result	
	positive $T_1$	negative $T_0$
disease $D_1$	true positive ( $tp$ )	false negative ( $fn$ )
no disease $D_0$	false positive ( $fp$ )	true negative ( $tn$ )

In the worst case, the prediction by the cut-off value is completely independent of the disease state of the patient. The null hypothesis of independence, can be tested by a chi-square test as follows:

```
> data <- matrix(c(5,5,5,5),2,byrow=TRUE)
> chisq.test(data)

Pearson's Chi-squared test with Yates' continuity correction

data: data
X-squared = 0.2, df = 1, p-value = 0.6547
```

Since the  $p$ -value is larger than the significance level, the null hypothesis of independence is not rejected.

Suppose that for another cutoff value we obtain 8 true positives ( $tp$ ), 2 false positives ( $fp$ ), 8 true negatives ( $tn$ ), and 2 false negatives ( $fn$ ). Then

testing independence yields the following:

```
> data <- matrix(c(8,2,2,8),2,byrow=TRUE)
> chisq.test(data)

Pearson's Chi-squared test with Yates' continuity correction

data: data
X-squared = 5, df = 1, p-value = 0.02535
```

Since the  $p$ -value is smaller than the significance level, the null hypothesis of independence is rejected.

### 4.1.9 Fisher's exact test

A frequently applied test in bioinformatics that is related to the Chi-squared is the Fisher's exact test. In a two-by-two contingency table with frequencies (actually counts)  $f_{11}$ ,  $f_{22}$ ,  $f_{12}$ , and  $f_{21}$ , this test is based on the so-called odds ratio  $f_{11}f_{22}/(f_{12}f_{21})$ .

**Example 1: Oncogenes on chromosome 1.** Suppose that the number of significant oncogenes in chromosome 1 is  $f_{11} = 100$  out of a total of  $f_{12} = 2000$ , and the number of significant oncogenes in the whole genome is  $f_{21} = 300$  out of a total of  $f_{22} = 6000$ . Then the odds ratio equals  $100 \cdot 6000 / (2000 \cdot 300) = 1$  and the number of significant oncogenes in chromosome 1 is exactly proportional to that in the genome.

	significant genes	non-significant genes
chromosome 1	100	1900
genome	300	5700

The null hypothesis of the Fisher's test is that the odds ratio equals 1 and the alternative hypothesis that it differs from 1.

Now suppose that the frequencies of significant oncogenes for chromosome 1 equals  $f_{11} = 300$  out of a total of  $f_{12} = 800$ , and for the genome  $f_{21} = 3,000$  out of  $f_{22} = 10,000$ :

	significant genes	non-significant genes
chromosome 1	300	3000
genome	500	7000

With the new values, it's not so clear whether or not oncogenes are significantly over- or underrepresented on chromosome 1 compared to the entire genome. However, we can test this hypothesis by testing if the odds ratio is significantly close to 1 using the `fisher.test()` function:

```
> data <- matrix(c(300,500,3000,7000),2,byrow=TRUE)
> fisher.test(data)

Fisher's Exact Test for Count Data

data: data
p-value = 1.336e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.201492 1.629240
sample estimates:
odds ratio
 1.399977
```

Since the  $p$ -value is smaller than the significance level, the null hypothesis of the odds ratio equal to one is rejected. There are more significant oncogenes in chromosome 1 than compared to that in the whole genome. Other examples of the Fisher's test will be given in Chapter 6.

Note that the `fisher.test()` function is very similar to the hypergeometric test `phyper()`. However, the `fisher.test()` function is more flexible than `phyper()` since it can perform a two-sided test and also give the confidence intervals of the odds ratio. For a one-sided test, the `fisher.test()` and `phyper()` functions will give the same  $p$ -value.

#### 4.1.10 Normality tests

Various procedures are available to test the hypothesis that a dataset is normally distributed. For example, the Shapiro-Wilk test is based on the degree of linearity in a Q-Q plot (Lehmann, 1999, p.347) and the Anderson-Darling test is based on the distribution of the data (Stephens, 1986, p.372).

**Example 1: CCND3 gene expression.** To test the hypothesis that the ALL gene expression values of CCND3 (Cyclin D3) from Golub et al. (1999) are normally distributed, the Shapiro-Wilk test can be used as follows:

```
> shapiro.test(golub[ccnd3, gol.fac=="ALL"])

Shapiro-Wilk normality test

data: golub[ccnd3, gol.fac == "ALL"]
W = 0.947, p-value = 0.1774
```

Since the  $p$ -value is greater than 0.05, the conclusion is not to reject the null hypothesis that CCND3 (Cyclin D3) expression values follow a normal distribution. The Anderson-Darling test function `ad.test()` is part of the `nortest` package which may need to be installed first. Running the Anderson-Darling test on our CCND3 (Cyclin D3) gene expression values produces the following:

```
> library(nortest)
> ad.test(golub[ccnd3, gol.fac=="ALL"])

Anderson-Darling normality test

data:  scale(golub[ccnd3, gol.fac == "ALL"])
A = 0.5215, p-value = 0.1683
```

Hence, the same conclusion is drawn as from the Shapiro-Wilk test. Note that the  $p$ -values from both tests are somewhat low. This confirms our observation in Section 2.2.6 based on the Q-Q plot that the distribution resembles the normal. From the normality tests the conclusion is that the differences in the *left, lower tail* are not large enough to reject the null-hypothesis that the CCND3 (Cyclin D3) expression values are normally distributed.

#### 4.1.11 Outliers test

When gene expression values are not normally distributed, then outliers may appear with large probability. The appearance of outliers in gene expression data may influence the value of a (non-robust) statistic to a large extent. For this reason it is useful to be able to test whether a certain set of gene expression values is contaminated by an outlier or not. Accordingly, the null-hypothesis to be tested is that a set of gene expression values does not contain an outlier and the alternative is that it is contaminated with at least one outlier. Under the assumption that the data are realizations of one and the same distribution, such a hypothesis can be tested by the Grubbs' (1950) test. This test is based on the statistic  $g = |\text{suspect value} - \bar{x}|/s$ , where the suspect value is included for the computation of the mean  $\bar{x}$  and the standard deviation  $s$ .

**Example 1: CCND3 gene expression.** From Figure 2.5 we have observed that expression values of gene CCND3 (Cyclin D3) may contain outliers with respect to the left, lower tail. This can actually be tested by the function

`grubbs.test()` from the `outliers` package as follows:

```
> library(outliers)
> grubbs.test(golub[ccnd3, gol.fac=="ALL"])

      Grubbs test for one outlier

data:  golub[ccnd3, gol.fac == "ALL"]
G = 2.9264, U = 0.6580, p-value = 0.0183
alternative hypothesis: lowest value 0.45827 is an outlier
```

Since the  $p$ -value is smaller than 0.05, the conclusion is to reject the null-hypothesis of no outliers.

When the data are normally distributed, the probability of outliers is small. Hence, extreme outliers indicate that it's highly probable that the data are non-normally distributed. Outliers may lead to such an increase of the standard error that a true experimental effect remains uncovered (false negatives). In such cases, a robust test based on ranks may be preferred as a useful alternative.

#### 4.1.12 Nonparametric Tests

The assumption that the random variables are normally distributed may be difficult to defend in certain empirical situations. The  $t$ -test is fairly robust against departures from normality when the sample size is large (Ramsey, 1980). However, for smaller sample sizes such departures may lead to incorrect conclusions. For this reason tests are made available for which no specific distributional assumptions need to be made. In the literature these are known as *nonparametric* or *distribution free* tests. In general, *nonparametric* tests assign ordinal rankings to the data and then the rankings (not the original data) are used in a statistical test. Using the rankings of the data significantly reduces the impact of outliers, but also contributes to a loss of power compared to the corresponding *parametric* test.

##### One-sample sign test

The *one-sample sign test* is a nonparametric alternative to the one-sample  $t$ -test. The null hypothesis  $H_0$  is that the median  $m$  of the data set is equal to the hypothesized value  $m_0$ :  $H_0 : m = m_0$ . The two-sided alternative hypothesis  $H_a$  is that the median  $m$  of the data set is not equal to the hypothesized value  $m_0$ :  $H_0 : m \neq m_0$ . It is often less powerful than the signed-rank test,

but has the advantage of not requiring the data to be symmetric.

**Example 1: Small sample size.** Let's test the hypothesis  $H_0 : m = 6000$  with the following sample: 6003, 6304, 6478, 6245, 6134, 6204, 6150, 6345, 6298, 6194. To test this null hypothesis nonparametrically by the one-sample sign test we use the `SIGN.test()` function as follows:

```
> x<-c(6003, 6304, 6478, 6245, 6134, 6204, 6150, 6345, 6298, 6194)
> SIGN.test(x, md = 6000)
```

```

      One-sample Sign-Test

data:  x
s = 10, p-value = 0.001953
alternative hypothesis: true median is not equal to 6000
95 percent confidence interval:
 6139.191 6331.698
sample estimates:
median of x
 6224.5

Achieved and Interpolated Confidence Intervals:

              Conf.Level  L.E.pt  U.E.pt
Lower Achieved CI    0.8906 6150.000 6304.000
Interpolated CI      0.9500 6139.191 6331.698
Upper Achieved CI    0.9785 6134.000 6345.000
```

Since the  $p$ -value is smaller than  $\alpha = 0.05$ , the null-hypothesis  $H_0 : m = 6000$  is rejected.

### One-sample, Wilcoxon signed-rank test

The *one-sample, Wilcoxon signed-rank test* is another nonparametric alternative to the one-sample  $t$ -test. However, unlike the *one-sample sign test*, this test does assume a symmetric distribution around the median. Therefore, the null hypothesis  $H_0$  is that the distribution of the data is symmetric about the hypothesized median  $m_0$ . Likewise, the two-sided alternative hypothesis  $H_a$  is that the distribution of the data is not symmetric about the hypothesized median  $m_0$ . After subtracting  $\mu_0$  from the data, the values are ranked (ordered) and then the sum of the ranks is computed over all the positive and negative ranks. This produces a  $p$ -value on the basis from which the null hypothesis is rejected if it is smaller than the significance level  $\alpha$ .

**Example 1: Small sample size.** Let's test the hypothesis " $H_0 : \text{symmetry about } m = 6000$ " with the following sample: 6003, 6304, 6478, 6245, 6134,

6204, 6150. To test this null hypothesis nonparametrically by the Wilcoxon signed rank test we use the `wilcox.test()` function as follows:

```
> x <- c(6003, 6304, 6478, 6245, 6134, 6204, 6150)
> wilcox.test(x,mu=6000)

      Wilcoxon signed rank test

data:  x
V = 28, p-value = 0.01563
alternative hypothesis: true mu is not equal to 6000
```

Since the  $p$ -value is smaller than  $\alpha = 0.05$ , the null-hypothesis is rejected.

### Two independent samples: Wilcoxon rank-sum test

The *Wilcoxon rank-sum test* is a nonparametric alternative to the two-sample  $t$ -test. It's also sometimes called the *Mann-Whitney  $U$  test*. When the data are normally distributed with equal variance, the two-sample  $t$ -test is an optimal test for testing  $H_0 : \mu_1 = \mu_2$  against  $H_a : \mu_1 \neq \mu_2$  (Lehmann, 1999). However, if the data are not normally distributed due to skewness or otherwise heavy tails, then this optimality does not hold anymore and there is no guarantee that the significance level of the test equals the intended level  $\alpha$  (Lehmann, 1999). For this reason, rank-type tests have been developed that do not require any (or very minor) specific distributional assumptions.

For the *Wilcoxon rank-sum test* we switch from hypotheses about means to those about general distributions. For example, an alternative hypothesis may then be formulated that the distribution of a first group lays to the left of a second. Let the gene expression values of the first group ( $x_1$  to  $x_m$ ) have distribution  $F$  and those of the second group ( $y_1$  to  $y_n$ ) distribution  $G$ . The null hypothesis is that both distributions are equal:  $H_0 : F = G$  and the alternative is that they are not. A one-sided alternative is that the  $x$ 's are smaller (or larger) than the  $y$ 's. To calculate the two-sample Wilcoxon rank-sum test, the data  $x_1, \dots, x_m, y_1, \dots, y_n$  are ranked and the rank numbers of the  $x$ 's are summed to form the statistic  $W$  after a certain correction (Lehmann, 2006). The idea is that if the ranks of  $x$ 's are smaller than those of the  $y$ 's, then the sum is small. The distribution of the sum of ranks is known and a  $p$ -value can be computed - on the basis of which the null hypothesis is rejected if it is smaller than the significance level  $\alpha$ .

**Example 1: CCND3 gene expression.** The null hypothesis that the expression values for gene CCND3 (Cyclin D3) are equally distributed for the



ALL patients and the AML patients can be tested by the built-in `wilcox.test()` function as follows:

```
> wilcox.test(golub[ccnd3,] ~ gol.fac)

      Wilcoxon rank sum test

data:  golub[ccnd3, ] by gol.fac
W = 284, p-value = 6.15e-07
alternative hypothesis: true location shift is not equal to 0
```

Since the  $p$ -value is much smaller than  $\alpha = 0.05$ , the conclusion is to reject the null-hypothesis of equal distributions.

### Two-sample, paired sign test

The *two-sample, paired sign test* is a nonparametric alternative to the paired  $t$ -test. It's also sometimes called the *dependent-samples sign test*. The null hypothesis  $H_o$  is that the median of the differences between pairs is zero for pairwise values  $(x_1, x_2)$ :  $H_o : m(x_1 - x_2) = 0$ . Likewise, the two-sided alternative hypothesis  $H_a$  is that the median of the differences between pairs is not zero:  $H_a : m(x_1 - x_2) \neq 0$ . The *paired sign test* is often less powerful than the paired Wilcoxon rank-sum test, but has the advantage of not requiring the data to be symmetric.

**Example 1: pathway gene expression.** For a family of genes, our null hypothesis is that the mean gene expression for each gene in a certain pathway did not change after chemotherapy treatment. We can test this hypothesis with the `SIGN.test()` function and including the `alternative = "two.sided"` option:

```
> before = c(112.9, 133.5, 122.8, 145.6, 167.2, 169.2, 174.6, 175.3, 184.4, 121.3)
> after  = c(120.7, 123.6, 134.0, 141.2, 156.8, 170.0, 182.0, 185.9, 196.0, 115.1)
> SIGN.test(before, after, alternative = "two.sided")

      Dependent-samples Sign-Test

data:  before and after
S = 4, p-value = 0.7539
alternative hypothesis: true median difference is not equal to 0
95 percent confidence interval:
 -11.005333  8.699556
sample estimates:
median of x-y
      -4.1

Achieved and Interpolated Confidence Intervals:
```

	Conf.Level	L.E.pt	U.E.pt
Lower Achieved CI	0.8906	-10.6000	6.2000
Interpolated CI	0.9500	-11.0053	8.6996
Upper Achieved CI	0.9785	-11.2000	9.9000

With a  $p$ -value of 0.7539, we cannot reject the null hypothesis that the median gene expression for each gene in a certain pathway did not change after chemotherapy treatment.

### Two-sample, paired Wilcoxon rank-sum test

The *two-sample, paired Wilcoxon rank-sum test* is another nonparametric alternative to the paired  $t$ -test. However, unlike the *Two-sample, paired sign test*, this test does assume a symmetric distribution around the medians. Therefore, the distribution of differences between the paired samples should be symmetric in shape. The null hypothesis  $H_0$  is that the distribution of the differences between the paired values is symmetric around zero. Likewise, the two-sided alternative hypothesis  $H_a$  is that the distribution of the differences between the paired values is not symmetric around zero.

**Example 1: pathway gene expression.** For a family of genes, our null hypothesis is that the mean gene expression for each gene in a certain pathway did not change after chemotherapy treatment. We can test this hypothesis with the `wilcox.test()` function and including the `paired=TRUE` option:

```
> before = c(112.9, 133.5, 122.8, 145.6, 167.2, 169.2, 174.6, 175.3, 184.4, 121.3)
> after = c(120.7, 123.6, 134.0, 141.2, 156.8, 170.0, 182.0, 185.9, 196.0, 115.1)
> wilcox.test(before, after, paired=TRUE)

      Wilcoxon signed rank test

data:  before and after
V = 18, p-value = 0.375
alternative hypothesis: true location shift is not equal to 0
```

With a  $p$ -value of 0.375, we cannot reject the null hypothesis that the median gene expression for each gene in a certain pathway did not change after chemotherapy treatment.

### Nonparametric bootstrapping

A general manner to test hypotheses with respect to parameters is by a computer intensive method called the bootstrap (Efron, 1979; Efron & Tibshirani, 1993). In particular, suppose that it is desired to test  $H_0 : \mu = \mu_0$

against  $H_0 : \mu \neq \mu_0$  and that  $\{X_1, \dots, X_n\}$  are independent and identically distributed with observations  $\{x_1, \dots, x_n\}$ . The idea is to take  $n^*$  random samples from the set  $\{x_1, \dots, x_n\}$  with replacement and to compute their mean - and to repeat this many, many times. For each iteration, this gives  $n^*$  values  $\{x_1^*, \dots, x_{n^*}^*\}$  by which the empirical distribution of  $x$  can be estimated. By computing quantiles of the empirical distributions, a bootstrap confidence interval for  $\mu$  can be estimated. It may be noted that when  $n$  increases and  $n^*$  is sufficiently large, then the bootstrap interval approaches the correct confidence interval. The mean of these  $n^*$  values is defined as  $\bar{x}^* = \sum_{i=1}^{n^*} x_i^*/n^*$  and is in fact another estimate of the unknown  $\mu$ . The bias may be estimated by the mean of  $\bar{x}^* - \bar{x}$ .

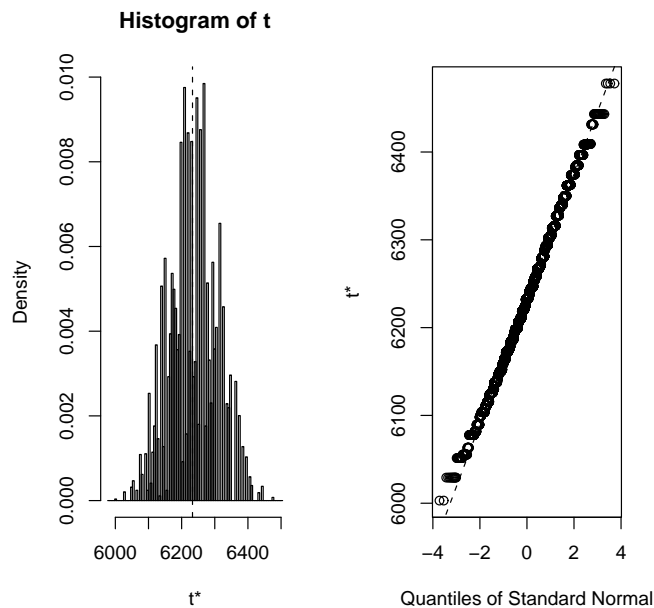


Figure 4.4: Bootstrap histogram and Q-Q plot of empirical distribution of light intensities.

**Example 1: Small sample size.** Lets use the bootstrap to test  $H_0 : \mu = 6000$  using the data from the previous examples. To install, load, and run the bootstrap on these data, we can execute the following commands:

```
> library(boot)
```

```

> x <- c(6003, 6304, 6478, 6245, 6134)
> boot.mean <- boot(x, function(x,i){mean(x[i])}, R=9999)
> boot.mean
ORDINARY NONPARAMETRIC BOOTSTRAP
Bootstrap Statistics :
  original    bias    std. error
t1*   6232.8 -0.4822082    71.179

```

The definition of `function` with the argument `i` is necessary to do the re-sampling properly. The re-sampled means are collected in the vector `boot.mean$t`, then the mean of this vector can be computed by `mean(boot.mean$t)`, and its standard deviation by `sd(boot.mean$t)`:

```

> mean(boot.mean$t)
[1] 6233.95
> sd(boot.mean$t)
[1] 71.60737

```

The latter is small relative to the difference between the bootstrap mean and mean of the null-hypothesis. The function `boot()` produces the object `boot.mean`, which is useful as input to the function `boot.ci()` to compute confidence intervals:

```

> boot.ci(boot.mean, conf = c(0.95), type = c("perc"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 9999 bootstrap replicates
Intervals :
Level      Percentile
95%    (6100, 6374 )
Calculations and Intervals on Original Scale

```

Since  $\mu = 6000$  does not fall within the confidence interval, the null hypothesis is rejected. The percentiles in `boot.ci()` are in fact exactly the same as from the `quantile()` method used in Chapter 1. That is,

```

> pvec <- c(0.025,0.975)
> quantile(boot.mean$t,pvec)
 2.5% 97.5%
6099.8 6374.4

```

Finally, the command `plot(boot.mean)` can be used to plot the empirical distribution. From the QQ plot at the right hand side in Figure 4.4, it can be observed that the distribution of the re-sample means is not far from being normal. Even for small sample sizes the bootstrap confidence interval is close to that of the one-sample *t*-test.

### 4.1.13 Robust estimation

In the case when outliers are expected in the data, it is useful to have an estimation method which is robust against them (Huber, 1964; Huber, 1981; Venables & Ripley, 2002). Roughly speaking, the idea is not to give every data point an equal weight in the estimation. While the mathematics behind robust estimators can get somewhat complicated, the idea behind robust estimation is pretty straight forward. For example, most robust estimation methods work in an iterative manner. In the first iteration, all the data points have equal weights. Then, for the next iteration outlier data points are downweighted according to their distance from the estimate. This process is typically iterated until the estimate stops changing within some epsilon. Through this procedure, the influence that the outliers have on the estimate is effectively reduced.

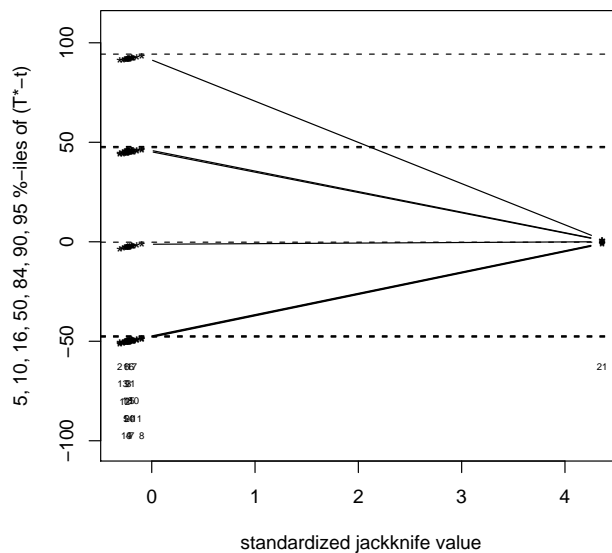


Figure 4.5: Jackknife after the bootstrap estimate the influence of data points.

**Example 1: Normal distribution with an outlier.** Lets construct a

sample from the normal distribution, add a clear outlier, and re-estimate the mean of the population.

```
> x <- rnorm(20,10,3)
> mean(x)
[1] 10.72456
> x[21] <- 1000
> mean(x)
[1] 57.83292
```

Hence, the mean of the sample with the 20 observation is fairly close to the population mean. However, by adding the clear outlier 1000 to the sample, the discrepancy between the sample mean and the population mean increases largely. The mean can be estimated robustly by using the function `huber()`, as follows:

```
> library(MASS)
> unlist(huber(x))
      mu      s
11.156620  3.022381
```

Obviously, the robust estimate is much closer to the population mean despite the presence of a clear outlier.

A technique to investigate the presence of an outlier is the leave-one-out method, also called *jackknifing*, after the bootstrap:

```
> library(boot)
> boot.mean <- boot(x, function(x,i){mean(x[i])}, R=9999)
> jack.after.boot(boot.mean)
```

This method makes it possible to estimate the influence of each data point on the results. The plot produced by the function `jack.after.boot()` is given by Figure 4.5. All first twenty data point are plotted to the left of the plot indicating their equality of influence. Data point twenty one, however, is plotted to the far right hand side of the plot indicating its large influence of the bootstrap estimate of the population.

#### 4.1.14 Type I and type II errors

When statistical hypotheses are tested, we assume that an empirical process of making many observations can be used to determine the true value of a population parameter. First, the researcher has a null hypothesis with respect to the true value of the population parameter and this hypothesis is either true or false. Next, the researcher makes observations of reality which yields data that forms the basis for the statistical test. Often, the statistical

test calculates the probability of observing the data when assuming that the null hypothesis is true. Highly improbable observations provide evidence to reject the null hypothesis. Lastly, given the outcome of the statistical test, the researcher makes a decision as to whether the data does or does not provide enough evidence to reject the null hypothesis in favor of the alternative hypothesis. Obviously, each of these decisions is either correct or incorrect. A schematic overview of the four possibilities of the decision situation is given in Table 4.4.

Under  $H_0$ , the statistic has probability  $1 - \alpha$  to be in the confidence interval (acceptance region). The remaining area is called the critical or rejection region, often illustrated in red in Figures such as 4.1 and 4.2. If  $H_0$  is true, then the statistic has probability  $\alpha$  to be in the rejection region. In this case, the incorrect conclusion follows that  $H_0$  is rejected. The error committed by rejecting  $H_0$  although it is true is traditionally called a Type I error or a false positive error. The probability to commit a Type I error is  $\alpha = P(\text{reject } H_0 | H_0 = \text{true})$ . And the probability to not reject  $H_0$  when it is true (a true negative) is  $1 - \alpha = P(\text{not reject } H_0 | H_0 = \text{true})$ .

Table 4.4: Overview of probabilities of decisions by researcher.

	decision $H_0 = \text{true}$	decision $H_0 = \text{false}$
reality $H_0 = \text{true}$	$1 - \alpha$	$\alpha$
reality $H_0 = \text{false}$	$\beta$	$1 - \beta$

It may happen that the researcher decides to not reject  $H_0$ , although in reality  $H_0$  is false. Such an error of incorrectly not rejecting  $H_0$  is called a type II error or a false negative error. The probability of this false negative event is  $\beta = P(\text{not reject } H_0 | H_0 = \text{false})$ . The last possibility is that the researcher decides to reject  $H_0$  and the alternative hypothesis  $H_a$  is indeed true. The probability of correctly rejecting  $H_0$  (a true positive) is called the *power* and is given by  $1 - \beta = P(\text{reject } H_0 | H_0 = \text{false})$ . The probabilities of the four events of the decision situation are summarized in the Table 4.4.

#### 4.1.15 Power of a statistical test

Obviously, a test procedure summarized in Table 4.4 is good if both  $\alpha$  and  $\beta$  are small or, equivalently, if both  $1 - \alpha$  and  $1 - \beta$  are large - thereby

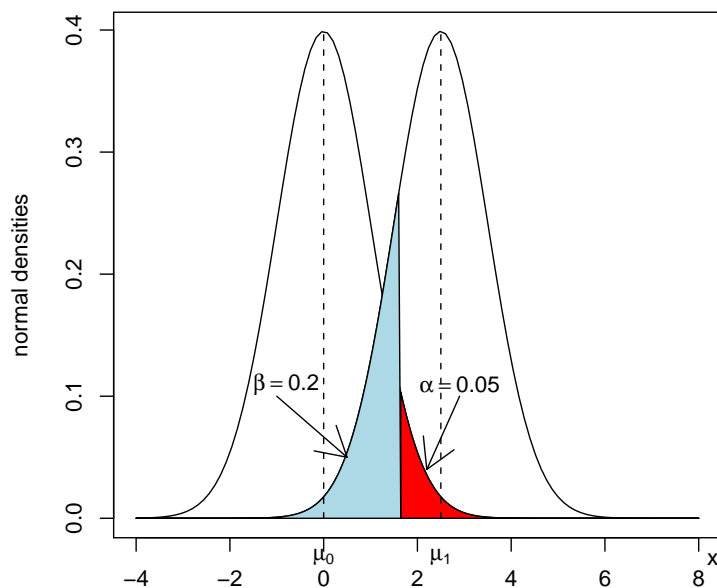


Figure 4.6: Type I (red) and Type II (blue) errors for two overlapping normal densities. The left normal distribution contains the negative samples while the right normal distribution contains the positive samples.

guaranteeing a large probability of drawing a correct conclusion. However, a large  $1 - \beta$  (true positive rate) is in general considered more important than a large  $1 - \alpha$  (true negative rate). For example, when screening for cancer it is far better to correctly detect all the cancer patients for treatment (true positives) than to just correctly predict those patients who for sure don't have cancer (true negatives). For this reason, the *power* of a statistical test is defined as the probability of correctly rejecting  $H_0$  (a true positive) - which is given by  $1 - \beta = P(\text{reject } H_0 | H_0 = \text{false})$ . If  $\mu_1$  from the alternative hypothesis is assumed to have a fixed value, then, given the sample size and the distributional assumptions, the power of the test becomes fixed and can be computed. To make the idea of power explicit, the decision situation with



respect to the hypotheses must be simplified to two possible values for the mean. That is,  $\mu = \mu_0$  under  $H_0$ , and  $\mu = \mu_1$  under  $H_a$ .

**Example 1.** With respect to the normal distribution it is desired to test  $H_0 : \mu = \mu_0$  against  $H_a : \mu = \mu_1$ . Furthermore, let  $\mu_0 = 0$ ,  $\mu_1 = 3.5$ , and  $\sigma^2 = 1$ ,  $\alpha = 0.05$ . Then we seek the quantile  $x_{0.95}$  such that

$$0.05 = \alpha = P(\text{reject } H_0 | H_0 = \text{true}) = P(X \leq x_{0.95}),$$

where  $X \sim N(0, 1)$ . Recall that this quantile can be found by `qnorm(0.95, 0, 1) = 1.644854`. The probability of the statistic to fall in the rejection region  $[1.644854, \infty)$  equals  $\alpha = 0.05$ . This corresponds to the red colored area in Figure 4.6. Furthermore, since  $\mu_1$  is known to be equal to 3.5, it follows that

$$\beta = P(\text{not reject } H_0 | H_0 = \text{false}) = P(Y \leq 1.644854) = 0.032,$$

```
> qnorm(0.95, 0, 1)
[1] 1.644854
> pnorm(1.644854, 3.5, 1)
[1] 0.03178769
```

where  $Y \sim N(3.5, 1)$ . The corresponding area of the type II error is colored in blue in Figure 4.6.

**Example 2: Teaching demonstration.** To further visualize the idea of the power of a test, load the `TeachingDemos` package and use the command `run.power.examp()`. Two densities are plotted: the normal  $N(0, 1)$  under  $H_0 : \mu = 0$ , and  $N(1, 1)$  under  $H_a : \mu = 1$ . In the first distribution the rejection region under  $\alpha = 0.05$  is colored in red, and in the second distribution the area corresponding to the power is colored in blue. Click on the "Sample Size" and move the slider to the right to increase it. Observe that by increasing the sample size, the variance of the statistic decreases, and the power of the test increases. Furthermore, by increasing the difference between  $\mu_0$  and  $\mu_1 = 1$ , the power is also increased.

**Example 3.** In the case of a one-sample  $t$ -test, if an estimate of  $\mu_1$  is made then the power can be computed. Let  $\alpha = 0.05$ ,  $n = 5$ ,  $\sigma = 175$ ,  $\mu_0 = 6000$ , and  $\mu_1 = 6300$  similar to Example 1 of Section 4.1.2. Then the function `power.t.test()` from the `stats` package can be used to estimate the power as follows:

```
> power.t.test(n=5, delta=300, sd=175, type = c("one.sample"))
One-sample t test power calculation
  n = 5
 delta = 300
  sd = 175
sig.level = 0.05
 power = 0.8138148
alternative = two.sided
```

The estimated power  $1 - \beta = 0.8138148$ , so that  $\beta = 0.1861852$ . Thus, the probability to not reject the null-hypothesis although the alternative is true (type II error) is quite large.

The larger the difference between  $\mu_0$  and  $\mu_1$ , the larger the power of the test. Furthermore, if  $\alpha$  increases then  $\beta$  decreases, and vice versa. However, commonly  $\alpha$  is fixed beforehand by the researcher and  $\mu_1$  is fixed by nature. Then the only manner to increase the power is by increasing the sample size. Consequently, the variance of the statistic decreases and the information in the data with respect to the population parameter increases.

**Power of parametric and nonparametric tests.** In general, for paired data the paired  $t$ -test will have more statistical power than an unpaired  $t$ -test. In addition, if the parametric assumptions hold, then a parametric test (e.g.  $t$ -test,  $F$ -test) will have more statistical power than the corresponding nonparametric test. Also, if the symmetry assumption holds, then the nonparametric rank-sum test will have more statistical power than the corresponding sign test. In conclusion, a statistical test that has more assumptions about the data will in general have more statistical power as long as those assumptions are true.

## 4.2 Maximum likelihood estimation (MLE)

Many models in bioinformatics are estimated by maximum likelihood because the estimation method has several desirable properties. First, estimation by maximum likelihood assumes that the density function is known. In addition, a parameter of the density function is generally denoted by  $\theta$ . For a particular parameter of the density function, it is assumed that distinct densities imply distinct parameter values, so that a density in fact identifies a parameter value. The true or population value of the parameter  $\theta$  is typically from a set of different parameter values. This set  $\{\theta_i\}$  of all the parameter values

will be denoted by  $\Theta$ . The idea behind maximum likelihood estimation is to use the parameter values which maximizes the joint density function as the estimator of the model to observe the data. In other words, we wish to find the model parameters  $\Theta$  that maximizes the likelihood of observing the data. In particular, suppose a sample from independent and identically distributed variables  $X_1, \dots, X_n$  is available with observations  $x_1, \dots, x_n$ . Recall that the density  $f(x_i|\theta)$  of  $X_i$  is assumed to be known. Also, the vertical bar “|” is typically read as “given”. The likelihood is defined as

$$L(\theta) = f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta),$$

where the last equality follows from the independence assumption (Billingsley, 1995, p. 262). The *maximum likelihood estimator*  $\hat{\theta}_n$  is defined as the values over the parameter set  $\Theta$  for which the likelihood is maximal for the data  $x_1, \dots, x_n$ . That is,

$$L(\hat{\theta}_n) = \arg \max_{\theta \in \Theta} (L(\theta))$$

There are many different *maximum likelihood estimator* implementations. However, many of them use an expectation-maximization (EM) algorithm in an iterative method to find the maximum likelihood estimates of the parameters in a statistical model. The EM iteration alternates between performing: (1) an expectation (E) step, and (2) a maximization (M) step of the expectation. The E-step creates a function for the expectation of the log-likelihood while using the current parameter estimates. Then the M-step computes new parameters that maximize the expected log-likelihood found in the E-step. The procedure is repeated with these new parameter estimates to determine the distribution of the variables in the next E-step. Note that the EM method is generally a greedy algorithm that is very dependent on the chosen initial parameters (conditions).

**Example 1.** Various parameters can be estimated by maximum likelihood using the built-in function `fitdistr()` from the `MASS` package. Importantly, for the normal, log-normal, geometric, exponential, and Poisson distributions the closed-form MLEs (and exact standard errors) are used by `fitdistr()`, and thus no initial parameters should be supplied. We shall illustrate this by a number of examples. First, let’s use `fitdistr()` to fit a Poisson distribution to a random sample (from a Poisson distribution):

```

> x <- rpois(100,4)
> fitdistr(x,"Poisson")
  lambda
 3.970000
(0.1992486)

```

The estimate of the Poisson parameter is 3.97 and its standard error equals  $\sqrt{\widehat{\lambda}_n/n} = \sqrt{3.97/100} = 0.1992486$ . Hence, the 95% confidence interval for  $\lambda$  becomes

$$[3.970 + z_{0.025} \cdot 0.199, 3.97 + z_{0.975} \cdot 0.199] = [3.579, 4.360]$$

Using the data from Example 1 of Section 4.1.2, the maximum likelihood estimates of the mean and the standard deviation of the normal distribution can be computed, as follows.

```

> x <- c(6003, 6304, 6478, 6245, 6134)
> fitdistr(x,"Normal")
  mean      sd
6232.80000 159.94424
( 71.52924) ( 50.57881)

```

From this the 95% confidence interval for  $\mu$  becomes

$$[6232.8 + z_{0.025} \cdot 71.52924, 6232.8 + z_{0.975} \cdot 71.52924] = [6092.605, 6372.995]$$

Recall that  $H_0 : \mu_0 = 6000$ . Since 6000 does not fall in the 95% confidence interval, it is concluded that the null hypothesis is rejected.

An example from the exponential distribution is the following.

```

> x <- rexp(20,5)
> fitdistr(x,"exponential")
  rate
 6.553413
(1.465388)

```

From this the 95% confidence interval becomes

$$[6.553 + z_{0.025} \cdot 1.465, 6.553 + z_{0.975} \cdot 1.465] = [3.681, 9.425]$$

Note that the confidence interval is large due to the small sample size.

**Example 2.** Frequently it is assumed that the data are identically distributed. However, in bioinformatics it may happen that the data come from a mixture of two separate distributions. For example, in a microarray study, some of the genes are "off" while others are "on". Let us assume that the gene expression values can be modeled by the following mixture model:

$$X_i \sim P \cdot N(3000, 300^2) + (1 - P) \cdot N(6000, 300^2).$$

where  $P$  is a Bernoulli random variable with probability 0.6 that the gene is "on" (equal to 1). Suppose that twenty data points are available giving the density plot in Figure 4.7.

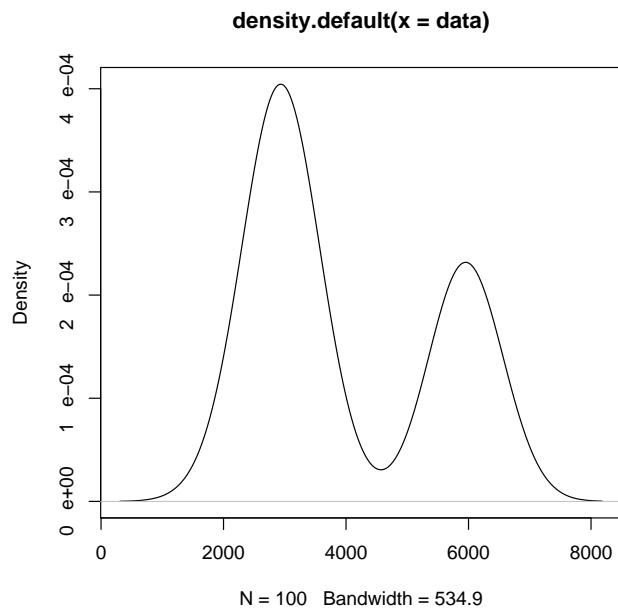


Figure 4.7: density of mixture of two normal distributions.

The parameters  $p, \mu_1, \sigma_1, \mu_2, \sigma_2$  of the defined mixture model can be estimated by maximum likelihood using the function `fitdistr()`.

```
> dmixnor <- function(x,p,m1,s1,m2,s2) {p * dnorm(x,m1,s1) + (1-p) * dnorm(x,m2,s2)}
> fitdistr(data,dmixnor,list(p=0.6,m1=3000,s1=300,m2=6000,s2=300))
      p          m1          s1          m2          s2
6.499808e-01  2.934890e+03  3.546717e+02  5.970190e+03  2.681543e+02
(4.769787e-02) (4.401108e+01) (3.044799e+01) (4.535870e+01) (3.132862e+01)
```

The object `data` contains the twenty data points. The 95% confidence interval for  $\mu_1$  becomes

$$[3073.21 + z_{0.025} \cdot 59.45, 6.55 + 3073.21 + z_{0.975} \cdot 59.45] = [2956.68, 3189.74]$$

### 4.3 Applications to gene expression data

So far in this chapter, we've applied various tests to a single vector of gene expressions for a single gene. In practice however, we commonly want to analyze a set of thousands of vectors of gene expression values which are collected in a gene expression matrix. This can be accomplished by taking advantage of the fact that R stores the output of a test as an object in where we can extract information such multiple  $p$ -values. Recall that the smaller the  $p$ -value the larger the experimental effect. Hence, by collecting  $p$ -values in a vector we can select genes with large differences between patient groups. We illustrate the technique by the two following examples.

**Example 1: Which gene expressions are normal?** With a data matrix of many gene expression values, a question one might ask is: What is the percentage of genes that passes a normality test? We can answer this question with the `apply()` and `shapiro.test()` functions as follows:

```
> data(golub, package="multtest")
> gol.fac <- factor(golub.cl, levels=0:1, labels= c("ALL", "AML"))
> sh <- apply(golub[, gol.fac=="ALL"], 1, function(x) shapiro.test(x)$p.value)
> sum(sh > 0.05)/nrow(golub) * 100
[1] 58.27598
```

Hence, according to the Shapiro-Wilk test, 58.27% of the ALL gene expression values are normally distributed (in the sense of non-rejection). For the AML expression values, 60.73% of the expression values are normally distributed. Therefore, we can conclude that about forty percent of the genes do not pass the normality test.

**Example 2: Differences between the  $t$ -test and Wilcoxon test.** In the case when the gene expression data are non-normally distributed, the  $t$ -test may indicate conclusions different from those of the Wilcoxon test. Differences between these can be investigated by collecting the  $p$ -values from both tests and looking for the largest differences.

```
> data(golub, package = "multtest");
> gol.fac <- factor(golub.cl, levels=0:1, labels= c("ALL", "AML"))
> pt <- apply(golub, 1, function(x) t.test(x ~ gol.fac)$p.value)
> pw <- apply(golub, 1, function(x) wilcox.test(x ~ gol.fac)$p.value)
> result <- data.frame(cbind(pw, pt))
> result[pw<0.05 & abs(pt-pw)>0.2,]
      pw      pt
456 0.04480288 0.2636088
1509 0.03215830 0.4427477
```

The  $p$ -values are extracted from the output of the `t.test()` function calls and stored in the vector `pt`. Likewise, the  $p$ -values are extracted from the output of the `wilcox.test()` function calls and stored in the vector `pw`. The logical operator `&` is used to select genes for which the Wilcoxon  $p$ -value is smaller than 0.05 and the absolute difference with the  $p$ -value from the  $t$ -test is larger than 0.2. Since there are only two such genes, we can draw the reassuring conclusion that the tests give similar results.

## 4.4 Overview and concluding remarks

Statistical hypothesis testing consists of hypotheses, distributional assumptions, and decisions (conclusions). The hypotheses pertain to the outcome of a biological experiment and are always formulated in terms of population values of parameters. Statistically, the outcomes of experiments are seen as realizations of random variables. The latter are assumed to have a certain suitable distribution which is seen as a statistical model for outcomes of an experiment. Then a statistic is formulated (e.g. a  $t$ -value) which is treated both as a function of the random variables and as a function of the data values. By comparing the distribution of the statistic with the value of the statistic, the  $p$ -value is computed and compared to the level of significance. A large  $p$ -value indicates that the model fits the data well and that the assumptions as well as the null-hypothesis are correct with large probability. However, a low  $p$ -value indicates, under the validity of the distributional assumptions, that the outcome of the experiment is so unlikely that this causes a sufficient amount of doubt for the researcher to reject the null hypothesis.

The quality of a test is often expressed in terms of efficiency, which is usually directly related to the (asymptotic) variance of an estimator. The relative efficiency is the ratio of the asymptotic variances. For Wilcoxon's rank-sum test versus the  $t$ -test this equals .955, which means that in the optimal situation where the (gene expression) data are normally distributed, Wilcoxon's test is only a little worse than the  $t$ -test. In the case, however, of a few outliers or a slightly heavier tail, the Wilcoxon test can be far more efficient than the  $t$ -test (Lehmann, 1999, p.176). Efficiency is directly related to power; the probability to reject a false hypothesis. Lastly, the probability of drawing correct conclusions can always be improved by increasing the sample size.

These considerations set the scene for making some recommendations,

which obviously should not be followed blindly. If gene expression data pass a normality test, then the Welch type of  $t$ -test provides a general test with good power properties (Ramsey, 1980; Wang, 1971). In the case where normality does not hold and the sample size per group is at least four, the Wilcoxon test is recommended.

Since the Wilcoxon  $p$ -values are based on ranks, and many of these ranks can be equal for different genes, the Wilcoxon rank-sum test is less suitable for ordering in the case when the sample size is small. On the other hand, it is obviously questionable whether extremely small differences in  $p$ -values produced by the  $t$ -test contribute to biologically relevant gene discrimination. That is, extremely small differences should not be over-interpreted.

## 4.5 Exercises

1. **CD33 gene.** Use `grep()` to find the index of the important gene CD33 among the list of genes in the `golub.gnames` table. For each test below formulate the null hypothesis, the alternative hypothesis, the  $p$ -value, and your conclusion.
  - (a) Test the normality of both the ALL and AML expression values separate from each other.
  - (b) Test for the equality of the variances between the ALL and AML patients.
  - (c) Test for the equality of the means by an appropriate  $t$ -test.
  - (d) Is the experimental effect strong?
2. **MYBL2 gene.** Use `grep()` to find the index of the gene “MYBL2 (V-myb avian myeloblastosis viral oncogene homolog-like 2)”.
  - (a) Use a box plot to construct a hypothesis about the experimental effect of ALL vs. AML for the MYBL2 gene.
  - (b) Test for the equality of the MYBL2 means by an appropriate  $t$ -test.
3. **HOXA9 gene.** The gene “HOXA9 (Homeo box A9)” can cause leukemia (Golub et al., 1999). Use `grep()` to find the index of this gene in the Golub data. (If your search returns more than one gene, then choose the first one in the list.)



- (a) Test the normality of the expression values of the ALL patients.
  - (b) Test for the equality of the means by the appropriate test.
4. **Zyxin gene.** On NCBI there are various cDNA clones of zyxin.
  - (a) Find the accession number of cDNA clone with IMAGE:3504464.
  - (b) Test whether the frequencies of the nucleotides are equal for each nucleic acid.
  - (c) Test whether the frequencies of “X94991.1” can be predicted by the probabilities of the cDNA sequence “BC002323.2”.
5. **Gene selection.** Select the genes from the `golub` dataset with the smallest two-sample  $t$ -test values for which the ALL mean is greater than the AML mean. Report the names of the best ten. Scan the Golub (1999) article for genes among the ten you found and briefly discuss their biological function.
6. **Antigens.** Antigens play an important role in the development of cancer. Order the antigens according to their  $p$ -values from Welch’s two-sample  $t$ -test with respect to gene expression values from the ALL and AML patients from the Golub et al. (1999) data.
7. **Mendelian genetic model.** A dihybrid cross in Mendelian genetics predicts that the four phenotypes associated with two independent traits show a 9:3:3:1 ratio in the F2 generation. In a certain experiment the offspring is observed with frequencies 930, 330, 290, 90. Do the data confirm the model?
8. **Comparing two genes.** Consider the gene expression values in rows 790 and 66 from the Golub et al. (1999) dataset.
  - (a) Produce a box plot for the ALL expression values and comment on the differences. Are there outliers?
  - (b) Compute the mean and the median for the ALL gene expression values for both genes. Do you observe a difference between the genes?
  - (c) Compute three measures of spread for the ALL expression values for both genes. Do you observe a difference between the genes?

- (d) Test by Shapiro-Wilk and Anderson-Darling the normality for the ALL gene expression values for both genes.
9. **Normality tests for gene expression values.** Perform the Shapiro-Wilk normality test separately for the ALL and AML gene expression values. What percentage passed the normality test separately for the ALL and the AML gene expression values? What percentage passes both tests?
10. **Two-sample tests on gene expression values.**
- (a) Perform the two-sample Welch  $t$ -test and report the names of the ten genes with the smallest  $p$ -values.
- (b) Perform the Wilcoxon rank-sum test and report the names of the ten genes with the smallest  $p$ -values.
11. **Biological hypotheses.** Suppose that the probability to reject a biological hypothesis by the results of a certain experiment is 0.05. Suppose that the experiment is repeated 1000 times.
- (a) How many rejections do you expect?
- (b) What is the probability of less than 10 rejections?
- (c) What is the probability of more than 5 rejections?
- (d) What is the probability that the number of rejections is between two and eight?
12. **Programming some tests.**
- (a) Program the two-sample  $t$ -test with equal variances and illustrate it with the expression values of CCND3 from the Golub et al. (1999) data.
- (b) Let the value  $W$  in the two-sample Wilcoxon rank-sum test equal the sum of the ranks of Group 1 minus  $n(n+1)/2$ , where  $n$  is the number of gene expression values in Group 1. Program this and illustrate it with the expression values of CCND3 from the Golub et al. (1999) data.

- (c) Let the value  $X$  in the two-sample Wilcoxon rank-sum test equal the number of values  $x_i > y_j$ , where  $x_i, y_j$  are values from Group 1 and 2, respectively. Program this and illustrate it with the expression values of CCND3 from the Golub et al. (1999) data.