Applied Statistics and Data Analysis
One Sample Problems

Joaquín Ortega
KAUST

August, 2020

Contents

1 One sample problems

1.1 Example: The speed of light

In 1879 Albert Michelson carried out a series of experiments to measure the speed of light. Some of the experiments (but not the one we are going to consider) were made together with Edward Morley and they are known as the Michelson-Morley experiments. Michelson went on to win the Nobel prize in Physics in 1907. At the time the 'accepted' value for the speed of light was 299,990 km/s.

The file `michelson` in the `MASS` package has the data for this experiment. It is also stored as `morley` in the `base` package. There were five experiments, each consisting of 20 different runs. The response is the speed of light measurement (with 299,000 subtracted).

```
library(MASS)
data(michelson)
str(michelson)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Speed: int  850 740 900 1070 930 850 950 980 980 880 ...
##  $ Run  : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Expt : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The results of the commands above show that `michelson` is a data frame with three variables,

- `Speed`, an integer corresponding to the measured speed of light (minus 299.000)
- `Run`, a factor indicating the order of the runs in each experiment (20 possible values), and
- `Expt`, a factor identifying the experiment (5 possible values)

We are going to use the data from one of the five experiments, the first, to see if the value measured by Michelson differs from the accepted value at the time of 299,990 km/s. For this, we need to compare the average of the speed measurements with the accepted value of 990 (recall we have subtracted 299,000 from the results).

We extract the values corresponding to the first experiment. We show two ways of doing this. In the first, we use square brackets and the `$` notation to specify that we want the data that corresponds to `Expt` equal to 1.

```
(mich.exp1 <- michelson[michelson$Expt == 1,1])
```

```
##  [1]  850  740  900 1070  930  850  950  980  980  880 1000
## [12]  980  930  650  760  810 1000 1000  960  960
```

Since the column is set to 1 inside the square brackets, we only get the value corresponding to the first component of the data frame, which is `Speed`.

Another way to do this is using the command `subset`:

```
subset(michelson, Expt == 1, select = 1)
```

```
##    Speed
## 1    850
## 2    740
## 3    900
## 4   1070
## 5    930
## 6    850
## 7    950
## 8    980
## 9    980
## 10   880
## 11  1000
## 12   980
## 13   930
## 14   650
## 15   760
## 16   810
## 17  1000
## 18  1000
## 19   960
## 20   960
```

Observe that even though the numbers are the same, the format is different.

The summary values for the data set are

```
summary(mich.exp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     650     850     940     909     980    1070
```

The average value for the first experiment is

```
(meanspeed <- mean(mich.exp1))
```

```
## [1] 909
```

We want to compare this value with the accepted value at the time, which was 990. The question of interest is whether the difference between these two values is too large to be due to randomness.

1.2 A general setting for our problem

Let's look at this problem in a general context. Suppose we have a sample $x_1, x_2, \dots, x_n$. We assume that these values are independent (in the statistical sense) but, for the time being, make no assumptions about the distribution.

The mean of these values, denoted by $\bar{x}_n$ or $\hat{\mu}_n$ is given by

$$\bar{x}_n = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We can think of the values $x_1, x_2, \dots, x_n$ as the *realization* of a collection of random variables $X_1, X_2, \dots, X_n$ where X_i represents the result of performing the i-th experiment. We assume that these variables all have the same (unknown) distribution and are independent. We denote this by **iid** (independent and identically distributed). We will denote random variables with capital letters: X, Y, Z and their values, once the experiment has been performed and we have a specific results, by small case letters: x, y, z. Also, rv stands for random variable or variables.

We also assume that the unknown common distribution for the X_is has a finite mean and variance, which we denote by

$$E(X) = \mu \qquad \text{and} \qquad Var(X) = \sigma^2.$$

If X has a continuous distribution with density $f(x)$ (as the normal distribution, for example) then

$$E(X) = \int x\, f(x)\, dx = \mu$$

and

$$\begin{aligned} Var(X) &= E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 = \int x^2 f(x)\, dx - \mu^2. \end{aligned}$$

On the other hand, if X has a discrete distribution with values $y_1, y_2, \dots$ and probability function $p_1, p_2, \dots$ then

$$E(X) = \sum_{i \geq 1} y_i p_i$$

and

$$Var(X) = E((X-\mu)^2) = E(X^2) - \mu^2 = \sum_{i \geq 1} y_i^2 p_i - \mu^2.$$

These are location and shape parameters.

We now have two means, the (unknown) theoretical or population mean μ and the empirical or sample mean $\hat{\mu}_n$ that we obtain from the sample. An important result in Probability Theory, known as the Law of Large Numbers, says that as the sample size grows, the empirical mean will converge to the population mean:

$$\hat{\mu}_n \to \mu \quad \text{as } n \to \infty.$$

with probability one. Therefore it makes sense to use $\hat{\mu}_n$ to estimate μ.

In our example, the 'accepted' value for the mean of the distribution was 990 and the sample mean obtained from Michelson's experiment was 909. These are the quantities we want to compare.

1.3 The sample mean is a random variable

Observe that $\hat{\mu}_n$ is also a random variable: it depends on the sample. If the experiments were to be repeated, different values would be obtained, but intuitively we would expect to see more variability in the sample than in their averages.

In figure 1 we show the results for all runs of the five experiments in the `michelson` data frame. Experiments are identified by color and the average for the ten runs in each experiment are represented by a star. From the graph we can see that there is more variability in the samples than in their averages. The code to produce this graph is in the gray box below.

```
library(MASS)
attach(michelson)
plot(Speed, col=Expt, pch=16)
mich.mean <- tapply(Speed, Expt, mean)
points(c(10,30,50,70,90),mich.mean,pch=8, cex=2, col=1:5)
```

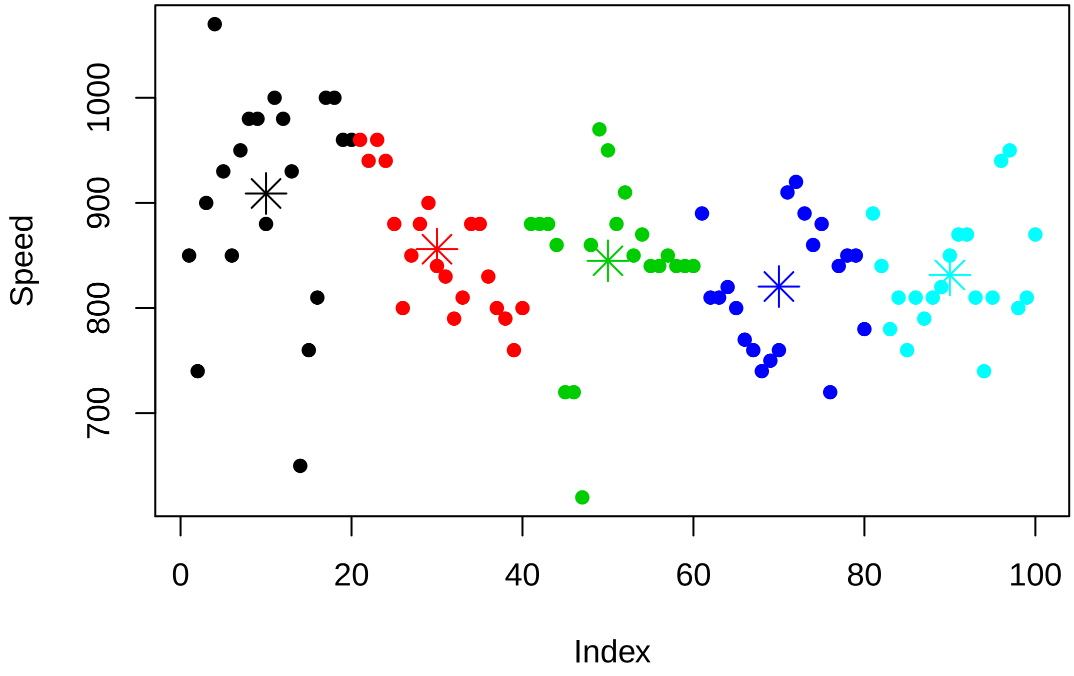

Figure 1: Results of Michelson's experiments

Since $\hat{\mu}_n$ is a random variable, we can consider its mean and variance:

$$E(\hat{\mu}_n) = E(\frac{1}{n}\sum_{1}^{n} X_i) = \frac{1}{n}\sum_{1}^{n} \mu = \mu.(\#eq:1) \tag{1}$$

So the mean value of $\hat{\mu}_n$ is μ. (Observe that here n is fixed). On the other hand, the variance is (using independence)

$$Var(\hat{\mu}_n) = Var\Big(\frac{1}{n}\sum_{1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{1}^{n} Var(X_i) = \frac{\sigma^2}{n}(\#eq:2) \tag{2}$$

Therefore the standard deviation of the sample mean decreases with the square root of the sample size. This is known as the *standard error of the mean*.

$$se = \frac{\sigma}{\sqrt{n}}$$

Since this is a dispersion parameter, we would expect to see dispersion decrease as the sample size increases.

How can we use these tools to compare Michelson's results with the previously accepted value? Could the difference be attributed to randomness?

1.4 First alternative: Gaussian distribution

Let us assume that the speed of light measurements come from a Normal or Gaussian distribution.

1.4.1 Gaussian distribution

The Gaussian distribution plays a central role in Statistics. The standard Gaussian density is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for $x \in \mathbb{R}$. It has mean 0 and variance 1.

```r
curve(dnorm(x), -3,3,lwd=2)
```

The Gaussian distribution with parameters μ and σ^2, denoted $N(\mu, \sigma^2)$ has density

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

for $x \in \mathbb{R}$. The parameters are the mean μ and the variance σ^2.

If $X \sim N(\mu, \sigma^2)$ then it is easy to see that

$$\frac{X-\mu}{\sigma} \sim N(0,1).$$

We say that

$$\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$$

is a parametric family of distributions with parameters

$$\mu \quad \text{and} \quad \sigma^2.$$

μ and σ are location and scale parameters.

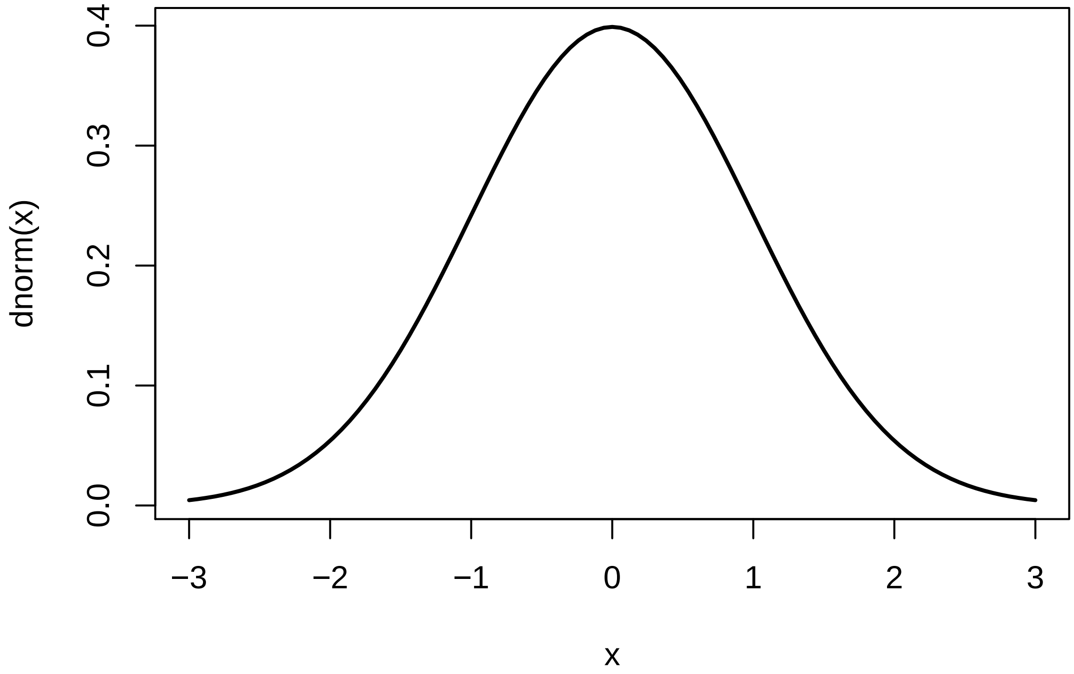

Figure 2: Standard Gaussian density

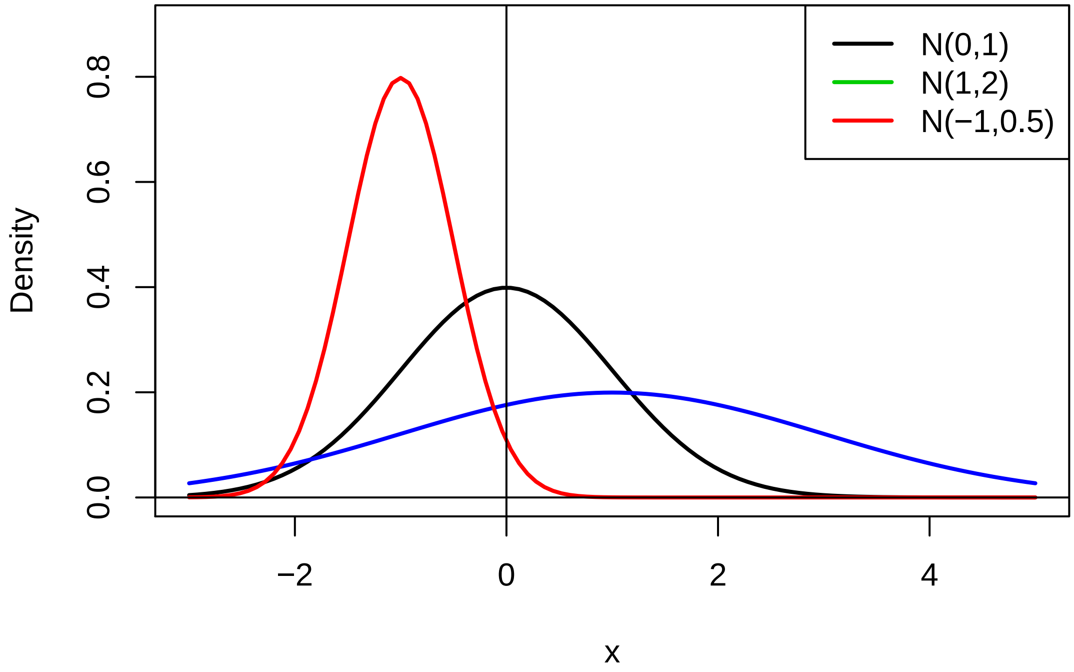

Figure 3: Three Gaussian densities with different parameters

A very important property of the normal family of distributions is that the sum of independent variables having normal distribution, also has a normal distribution. More precisely, if X has a normal distribution with mean μ_X and variance σ_X^2, denoted $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$, and they are independent, then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_X^2).$$

Also, if we multiply X times a, for any real number a, then $aX \sim N(a\mu_X, a^2\sigma_X^2)$. Therefore, linear combinations of normal variables produce normal variables.

This result has the following implication for the empirical mean $\hat{\mu}_n$: If the variables $X_i, i = 1, \dots, n$ are iid $N(\mu, \sigma^2)$ then

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

We just need to remember equations (1) and (2) to get the correct parameters.

1.4.2 Back to the speed of light experiment

As we have just seen, if we assume that the measurements come from a normal distribution then the average over a sample of size n has distribution

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

Now we can gauge how likely it is for the difference between what Michelson observed and what was thought at the time to be the true value to be due to chance.

Accepted value $\mu = 990$

Observed average $\mu_n = 909$

Difference $= 81$

We want

$$P(|\hat{\mu}_n - \mu| \geq 81).$$

However, we are not ready yet to answer our question because the distribution of $\hat{\mu}_n$ depends on another (unknown) parameter: σ^2.

Solution: Estimate the variance from the sample and use it in the formula.

This is not a bad solution if n is large, again due to the Law of Large Numbers. In our case $n = 20$, which is not too big.

The estimated variance is denoted s_n^2, and the standard deviation is s_n. The sample variance is given by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu_n)^2$$

We can use `R` to calculate variance and standard deviation for the speed of light sample:

```r
var(mich.exp1); sd(mich.exp1)
```

```
## [1] 11009.47
```

```
## [1] 104.926
```

Using this value,

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \approx N(0, 1).$$

where we have changed the symbol '$\sim$' for '$\approx$' to indicate that the relation is only approximately true.

The probability we are interested in is:

$$\begin{aligned}
P(|\hat{\mu}_n - \mu| \geq 81) &= P(\frac{|\hat{\mu}_n - \mu|}{s_n/\sqrt{n}} \geq \frac{81}{s_n/\sqrt{n}}) \\
&= P(|Z| \geq 3.452) \\
&= 2P(Z \leq -3.452)
\end{aligned}$$

where the last relation follows from the symmetry of the Gaussian distribution. We can now use `R` to calculate this probability:

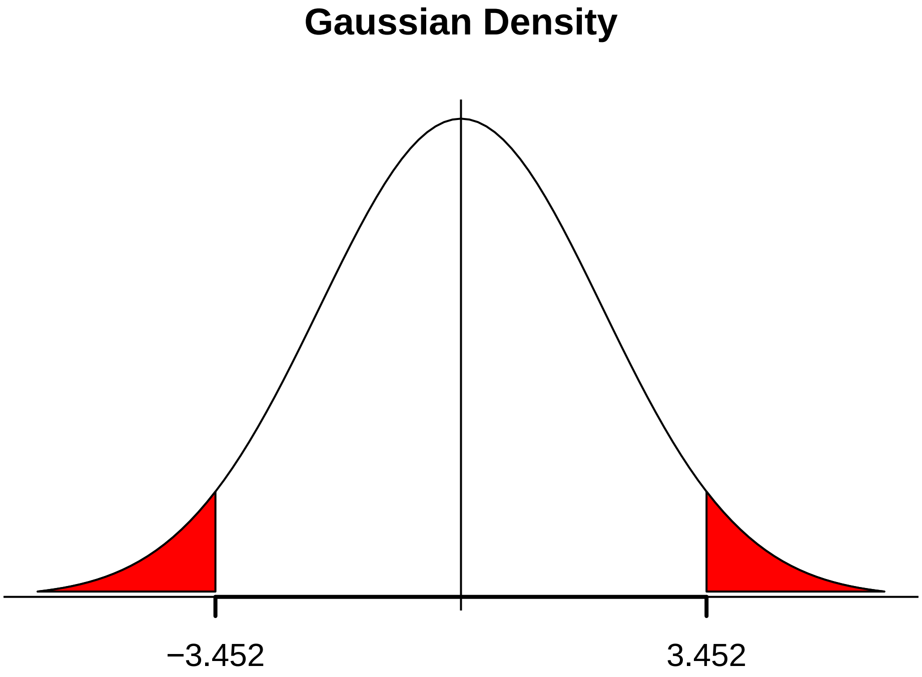

Figure 4: Standard Gaussian density. The area in red represents the probability that the absolute value of the difference between the empirical mean and the true population mean is greater than or equal to 81.

```
2*pnorm(-3.452)
```

```
## [1] 0.0005564477
```

Thus, the probability of observing a difference as large or larger than the one we observed is about 0.00056. The odds are about one in 1800. This is a small number so it would make us doubt the validity of the accepted value.

We have performed a hypothesis test. Assuming that the sample is Gaussian, we have tested the null hypothesis that the mean of the common distribution is 990 versus the alternative that it is not:

$$H_0 : \mu = 990 \qquad vs. \qquad H_A : \mu \neq 990$$

However, we made a fundamental assumption to carry out the previous calculations: We have assumed that our sample comes from a Gaussian distribution.

We need to ask ourselves whether this is a reasonable assumption. We can use graphical methods to explore this.

```
library(MASS)
truehist(mich.exp1, xlab = 'speed of light')
```

```
curve(dnorm(x,mean = mean(mich.exp1),sd = sd(mich.exp1)),
      600, 1100, add = TRUE, lwd=2, col='brown3')
lines(density(mich.exp1),col='blue3',lwd=2)
```

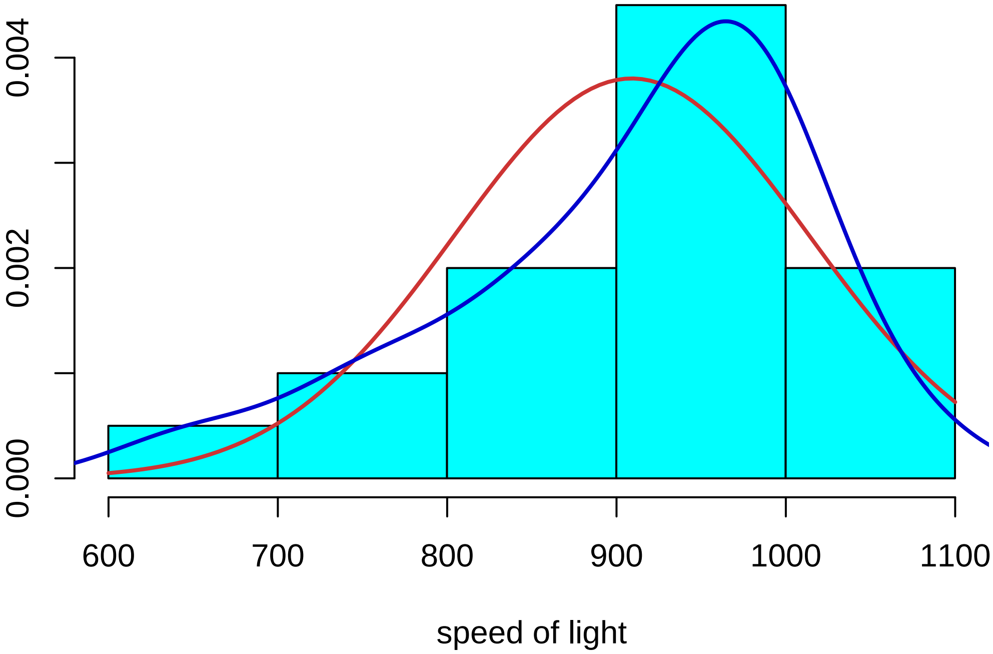

Figure 5: Histogram of the speed of light measurements in Michelson's first experiment. The blue curve represents the estimated density, the red curve is the Gaussian density with estimated parameters.

```
qqnorm(mich.exp1)
qqline(mich.exp1,lwd=2)
```

The qq plot does not look very good but a test of normality does not reject the null hypothesis.

```
shapiro.test(mich.exp1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mich.exp1
## W = 0.91992, p-value = 0.09876
```

1.5 Second alternative: Student's t distribution

In the previous section, we assumed that the sample had a Gaussian distribution with mean equal to 990. Since the variance was unknown, we used the sample variance in the probability calculation. Even though this may be a reasonable approach for large sample sizes, which was not our case, there is a better solution, using a different distribution for the sample mean.

The t distribution arises as the sampling distribution of the (empirical) mean $\hat{\mu}_n$ when the data come from a normal distribution with unknown variance.

We saw that if $\hat{\mu}_n = \frac{1}{n}\sum_1^n X_i$ and the X_i are iid with $N(\mu, \sigma^2)$ distribution then

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

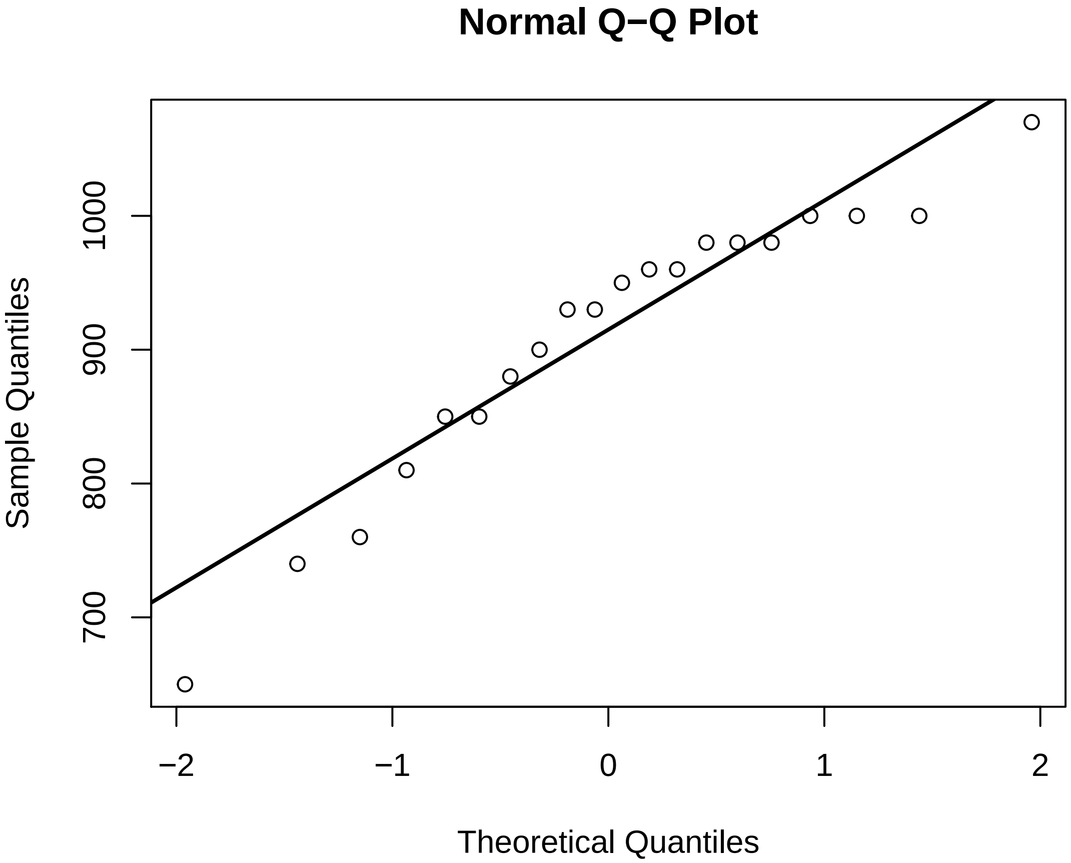

Figure 6: QQ-plot for the speed of light measurements.

In practice, we never know the true value for the variance σ^2 and we must estimate it by the empirical variance s_n^2:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

It was shown in 1908 by W.S. Gosset in a paper in Biometrika, published under the pseudonym *Student*, that

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom. For $n \geq 30$ the t distribution is very similar to the normal distribution but for n small there are important differences. The t distribution has 'heavier' tails than the normal, which means that large values are more probable. The next figure shows several examples of densities for small values of the degrees of freedom parameter.

```r
cols <- rainbow(100)
curve(dnorm(x),-3,3,lwd=2, ylab='density',col='grey25')
for (i in c(2,4,6)) {
  curve(dt(x,i),-3,3,lwd=2, add = TRUE, col=cols[55+3*i])}
curve(dt(x,30),-3,3,lwd=2, add = TRUE, col=cols[85])
legend('topright',c('normal','t2','t4','t6','t30'),
       col=c('grey25',cols[c(61,67,73,85)]),lwd=rep(2,5))
```

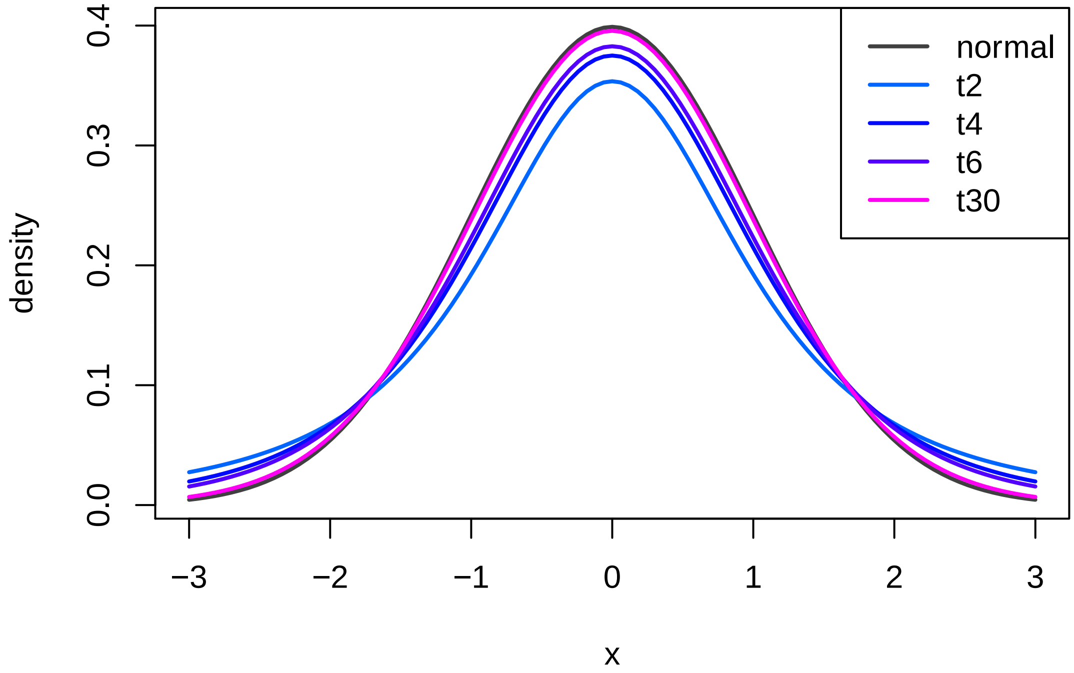

Figure 7: Density of the t distribution for different values of the frgrees of freedom parameter. The normal standard density (black) is included as reference.

Now that we know the true distribution for the sample mean with unknown variance (under the hypothesis that the data are normal), we can do a more precise test. We want to test

$$H_0 : \mu = 990 \qquad vs. \qquad H_A : \mu \neq 990$$

and our *test statistic* is

$$\bar{x}_n = \hat{\mu}_n = \frac{1}{n}\sum_{1}^{n} x_i$$

where the x_i are the observed values.

Our null hypothesis specifies a single value for the mean of the distribution: $\mu = 990$. Thus, *under the null hypothesis*, i.e. assuming the null hypothesis is true, we have that

$$\frac{\hat{\mu}_n - 990}{s_n/\sqrt{n}} \sim t_{n-1}$$

Recall that we have a sample of size $n = 20$ and the standard deviation is

```
sd(mich.exp1)
```

```
## [1] 104.926
```

```
sd(mich.exp1)/sqrt(20)
```

```
## [1] 23.46218
```

Thus

$$\frac{\hat{\mu}_n - 990}{23.46} \sim t_{19}.$$

We observed $\hat{\mu}_n = 909$. How likely is this under this distribution?

We want $P(|\hat{\mu}_n - 990| \geq 81)$ and this calculation is similar to what we did before,

$$\begin{aligned} P(|\hat{\mu}_n - 990| \geq 81) &= P(\frac{|\hat{\mu}_n - 990|}{23.46} \geq \frac{81}{23.46}) \\ &= P(|t_{19}| \geq 3.542) \\ &= 2P(t_{19} \leq -3.542) \end{aligned}$$

where the last equality follows from the symmetry of the t distribution.

To calculate this value use the distribution function for the t distribution in `R`, given by `pt`:

```
2*pt(-3.452,df=19)
```

```
## [1] 0.002670794
```

which is bigger than what we obtained before assuming a normal distribution and using the estimated standard deviation. The odds now are about one in 375.

The reason for this is that for small samples we do not expect the estimated variance to be accurate, and there will be more variability in the sample. Hence we need a distribution that makes having larger values more likely, which is the t distribution in this case.

It is not necessary to do all the calculation every time we want to do a t test. The function `t.test` in `R` will do this.

```
t.test(mich.exp1, mu=990)
```

```
##
##  One Sample t-test
##
## data:  mich.exp1
## t = -3.4524, df = 19, p-value = 0.002669
## alternative hypothesis: true mean is not equal to 990
## 95 percent confidence interval:
##  859.8931 958.1069
## sample estimates:
## mean of x
##       909
```

The output indicates that we are carrying out a one sample t-test, the data set is `mich.exp1`, gives the value for the t statistic, the number of degrees of freedom and the p-value. It reminds us what the alternative hypothesis is, gives a 95% confidence interval for the true mean, and the value for the sample mean.

1.6 Third alternative: Non-parametric test

So far we have used the hypothesis of normality as the basis to build our tests. What if we don't want to make this assumption?

There are tests that are **distribution-free**, i.e. they do not make distributional assumptions. They are known as **non-parametric tests**, because they are not based on the asumption of a parametric family of distributions. For the one-sample problem we are considering, the most popular choice is known as Wilcoxon's signed-ranks test.

Many non-parametric methods, Wilcoxon's test among them, are based on **order statistics** and **ranks**. Assume you have a sample $x_1, x_2, \dots, x_n$ and that all values are different. The **order statistics** for this sample are the ordered values:

$$x_{(1)} < x_{(2)} < \cdots < x_{(n-1)} < x_{(n)}$$

The **rank** is the position that a particular values has in the ordered sample.

1.6.1 Example

In this example we draw a sample of size 5 from the uniform distribution in (-1,1) using `R`, and find the order statistics and the ranks:

```
(unif.spl <- runif(5,-1,1))
```

```
## [1] -0.000809934 -0.494444656  0.545963046  0.899801027
## [5]  0.443136170
```

```
sort(unif.spl)
```

```
## [1] -0.494444656 -0.000809934  0.443136170  0.545963046
## [5]  0.899801027
```

```
rank(unif.spl)
```

```
## [1] 2 1 4 5 3
```

Assume that the sample comes from a continuous distribution which is symmetric with respect to its average value μ, such as the Gaussian distribution or the t distribution. We want to test the (null) hypothesis $H_0 : \mu = \mu_0$ versus the alternative that this is false

Under the symmetry condition stated above, it is equally likely that values will be above or below the mean, and also positive and negative differences of the same magnitude have the same probability of occurring. If the sample values are $x_1, \dots, x_n$ define

$$d_i = x_i - \mu_0.$$

To compute the test statistic follow these steps:

1. Take the absolute value of the d_i's.

2. Order these values and assign the ranks. If there are ties, use the midranks (average rank of the tied values).
3. Multiply the rank values obtained in 2 by the original signs of the d_i's.
4. Sum the positive values in 3 and denote the result by t^+.

t^+ is a sum of **ranks**, not of values, and is (the value) of the test statistic. We denote the corresponding r. v. by T^+. T^- is defined similarly using the negative values, but in fact we only need one of them to carry out the test.

If the sample has size n, the possible values of T^+ are between 0 (if all the values in the sample are negative) and

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

if all values are positive. Also,

$$T^+ + T^- = \frac{n(n+1)}{2}$$

and this is why we only need one of these random variables for the test.

If the hypothesis of symmetry is valid, we would not expect very small or very large values of T^+. Therefore, if we observe either of these situations we will reject the null hypothesis. The distribution of T^+ is difficult to calculate, particularly if there are ties in the sample. There is a normal approximation to the distribution that is frequently helpful. The test can be carried out in `R` with the command `wilcox.test`.

```
wilcox.test(mich.exp1, mu = 990)
```

```
## Warning in wilcox.test.default(mich.exp1, mu = 990): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mich.exp1
## V = 22.5, p-value = 0.00213
## alternative hypothesis: true location is not equal to 990
```

We get a warning saying that, since there are ties, the exact p value cannot be calculated. Then we get the value for the statistic `V=22.5`, the p value, which is 0.00213 and a reminder of what the alternative hypothesis is. Observe that the p-value is similar to what we got with the t-test.

Let's calculate the test statistic step by test to see how it works. We start by calculating the differences d_i and ordering them.

```
mich.dif <- mich.exp1-990
```

We see that there are several ties and more negative than positive values. In fact, there are

```
sort(mich.dif)
```

```
##  [1] -340 -250 -230 -180 -140 -140 -110  -90  -60  -60  -40
## [12]  -30  -30  -10  -10  -10   10   10   10   80
```

```
length((mich.dif)[mich.dif<0])
```

```
## [1] 16
```

16 negative values and only four positive ones. The function `rank()` in `R` calculates the ranks for the differences.

```
rank(abs(mich.dif))
```

```
##  [1] 15.5 19.0 13.0 12.0 10.5 15.5  9.0  3.5  3.5 14.0  3.5
## [12]  3.5 10.5 20.0 18.0 17.0  3.5  3.5  7.5  7.5
```

Now we multiply the rank by the original sign of the difference d_i and sum those that have positive sign. The result is the same as we had using `wilcox.test` in `R`.

```
mich.signrank <- rank(abs(mich.dif))*sign(mich.dif)
sum(mich.signrank[mich.signrank>0])
```

```
## [1] 22.5
```

2 Pointwise estimation

2.1 Introduction

One of the fundamental problems in Statistics is the study of a population using a sample drawn at random. This problem has many facets. Determining the characteristics of the population that we want to study, describing a probabilistic model for this characteristic, designing the procedure for drawing the sample, estimating parameters in the case of a parametric model and giving a quantification of the uncertainty in the estimation, are some of the problems of interest in this situation.

We assume that we have defined the characteristic we want to study and that the model consists of a family of distribution function $F(t;\theta)$ that depends on a parameter θ. The model gives a framework for statistical analysis and is usually a simple approximation to a complicated reality. A model that determines the distribution up to the value of one or several parameters is known as a **parametric model**.

2.2 Estimation

The parameter θ may be a number or a vector. The problem we want to consider here is the estimation of this parameter from a random sample drawn from the population. By **random sample**, we mean a subset $x_1, x_2, \ldots, x_n$ of elements in the population that are selected according to a random mechanism in which all elements of the population have the same probability of being selected.

We can think of the values $x_1, x_2, \ldots, x_n$ as the *realization* of a colection of random variables $X_1, X_2, \ldots, X_n$ where $X_i, i = 1, \ldots, n$ represent variables with distribution function $F(t;\theta)$ and these variables are independent. We use the notation **iid** for independent, identically distributed. As a rule, we will use capital letters X, Y, Z for denoting random variables and small case letter x, y, z for their values, once they have been observed. Also, rv stands for random variable.

To estimate the unknown parameter, we use a **statistic**, which is a function of the random sample. The sample mean is an example of a statistic, as are the sample variance, the standard deviation, the median, the quantiles, the correlation coefficient, or the slope of a regression line. Each of these functions may be used to estimate specific characteristics of the distribution that models our population of interest. A statistic is any function of the random sample. When we use a statistic to estimate a specific parameter of the population, we speak of an **estimator**. Thus, the sample mean is an estimator for the population mean. The sample mean is given by the formula

$$\bar{X} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i(\#eq : A1) \tag{3}$$

When we replace the variables $X_1, X_2, \ldots, X_n$ by their observed values $x_1, x_2, \ldots, x_n$, we have an **estimate** of the unknown population mean:

$$\bar{x} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.(\#eq : A2) \tag{4}$$

2.2.1 Example 1.

Suppose that we have a parametric model given by the family of Gaussian distributions $N(\mu, \sigma^2)$, with parameters μ and σ^2, where μ is the population mean and σ^2 is the variance. These parameters are unknown, and we draw a random sample from the population. Suppose the sample has size $n = 10$ and we observe the following values:

```r
round(smpl1,2)
```

```
##  [1] 4.56 5.46 5.01 4.64 3.73 1.17 7.08 5.73 4.79 0.94
```

We can use the sample mean $\bar{x}_{10}$ to estimate μ. In `R`, we use the function `mean`:

```r
(mean.smpl1 <- mean(smpl1))
```

```
## [1] 4.311172
```

Our estimate for the mean of the population, based on the sample contained in the vector `smpl1` is 4.31. We also use the notation $\hat{\mu}$ or $\hat{\mu}_n$ to denote this estimate

$$\hat{\mu} = \bar{x} = 4.31.$$

Similarly, we can use the sample variance s^2 to estimate σ^2. The formula for the sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

In `R` we use the function `var` to obtain the sample variance:

```r
(var.smpl1 <- var(smpl1))
```

```
## [1] 3.714826
```

We also use the notation $\hat{\sigma}^2$ to denote the estimate of the variance:

$$\hat{\sigma}^2 = 3.715$$

2.3 Sampling distribution

Since a statistic is a function of a random sample, and a random sample is a collection of iid random variables, a statistic is also a random variable. If we repeat the sampling procedure to obtain a new sample, we will get different values for the sample and therefore, a different value for the statistic.

The statistic has a probability distribution characterized by its distribution function. This distribution is known as the **sampling distribution** for the statistic. It is not always easy to determine the sampling distribution, but it is always possible to calculate the mean and variance for the sampling distribution (assuming that they exist).

Suppose we have a random sample $X_1, \ldots, X_n$ from a distribution function $F(t; \theta)$ that has finite mean and variance, denoted by μ and σ^2, respectively:

$$E(X_i) = \mu, \qquad Var(X_i) = \sigma^2, \qquad i = 1, 2, \ldots, n.$$

The sample mean is given by equation (3), and using the linearity of the expected value we get

$$E(\hat{\mu}_n) = E(\bar{X}) = E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X_i) = \mu(\#eq : A3) \tag{5}$$

When the expected value of the estimator is the parameter we want to estimate, we say that the estimator is **unbiased**. The sample mean is an unbiased estimator of the population mean.

To calculate the variance of $\hat{\mu}_n$ start with the definition:

$$Var(\hat{\mu}_n) = E\big[\big(\hat{\mu}_n - E(\hat{\mu}_n)\big)^2\big] = E\big[\big(\hat{\mu}_n - \mu\big)^2\big]$$

but

$$\hat{\mu}_n - \mu = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$$

In consequence

$$\begin{aligned}
Var(\hat{\mu}_n) = Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) &= E\big[\big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\big)^2\big] \qquad (6)\\
&= \frac{1}{n^2}E\big[\big(\sum_{i=1}^{n}(X_i - \mu)\big)^2\big] \qquad (7)\\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} E(X_i - \mu)^2 + \sum_{i\neq j} E(X_i - \mu)(X_j - \mu)\Big] \qquad (8)\\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\Big](\#eq:A4) \qquad (9)
\end{aligned}$$

where $Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big)$ is the covariance between X and Y with $\mu_X = E(X)$ and similarly for μ_Y. The formula (6) always holds. In the case we are considering, the variables in the sum are independent, and this implies that their covariance is zero. Let's prove this for two independent variables X and Y. Recall that the expected value of the product of *independent* variables is the product of the expected values. Using this

$$\begin{aligned}
Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) &= E(X-\mu_X)E(Y-\mu_Y) \qquad (10)\\
&= \big(E(X)-\mu_X\big)\big(E(Y)-\mu_Y\big) = 0(\#eq:A5) \qquad (11)
\end{aligned}$$

Using (10) in (6) we get

$$Var(\hat{\mu}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n}\sigma^2.(\#eq:A6) \tag{12}$$

Therefore, the variance of the sample mean decreases with n. Since the variance measures how much the distribution of a random variables is concentrated around its mean, this means that, as the sample size n grows, the distribution of the estimator $\hat{\mu}_n$ will concentrate around the true value μ.

2.3.1 Example 1 revisited

The data in the vector `smpl1` that we used in example 1 were simulated from a normal distribution with parameters $\mu = 4.5$ and $\sigma^2 = 4$. The sample average we obtained was $\hat{\mu} = 4.136$. If we draw another sample form the same distribution, we obtain a different estimate:

```r
smpl2 <- rnorm(10,4.5,2)
round(smpl2,2)
```

```
##  [1] 5.89 6.84 3.75 4.36 5.67 5.49 1.30 3.84 7.03 6.51
```

```r
mean(smpl2)
```

```
## [1] 5.066089
```

If we do this again, we will get yet another different value, since the sample average, as we have seen, is a random variable. What we proved in the previous section says that the expected value of these sample means we have obtained is the true value (4.5) and that their variance is equal to the population variance (4) divided by the sample size (10), i.e., 0.4. To show that this is the case we will carry out a simulation, to obtain an empirical approximation to the distribution of $\hat{\mu}$.

In the R code below, we will generate 1000 samples of size 10 from the population, which has $N(4.5, 4)$ distribution, and calculate the sample mean for each of these samples. This gives a sample of 1000 empirical averages of size 10.

```
smpl.mat <- matrix(rnorm(25000,4.5,2), ncol = 10)
mean.vec <- apply(smpl.mat,1,mean)
hist(mean.vec, breaks = 20, freq = FALSE, xlab = 'Sample mean',
     main = 'Histogram of 25000 sample means', xlim = c(0,10))
lines(density(mean.vec, adjust = 1.5), col = 'red',lwd = 2)
curve(dnorm(x,4.5,2), 0, 10, add = TRUE, col = 'blue', lwd = 2)
```

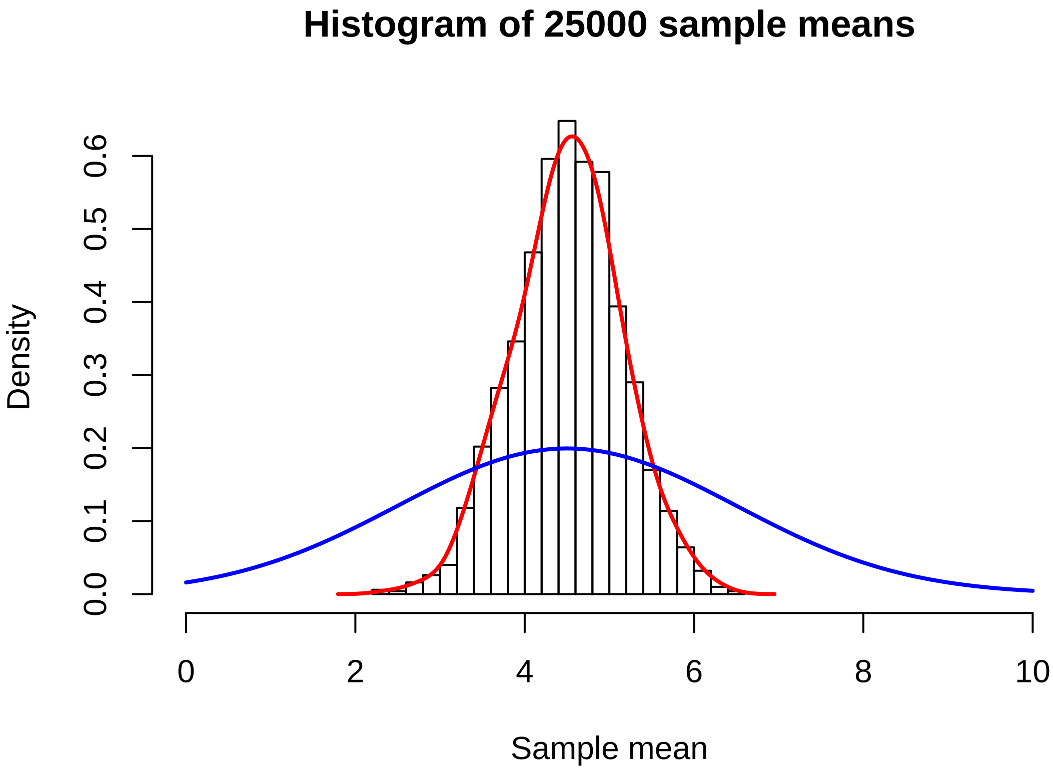

Figure 8: Histogram of 25000 empirical means for samples of size 10. In blue, population distribution, in red, estimated sampling density.

The figure shows that there is much less variability in the means that in the sample, as the variance has considerably reduced, even though the size of the sample is only 10. The red curve looks like a normal distribution, and we will see in the next section that this is indeed the case.

2.4 Sampling distribution for the mean of a normal sample

When data come from a normal distribution, it is possible to obtain the sampling distribution for the average of a random sample. To do this, we need the following property of independent normal random variables: Let X and Y be independent random variables with means μ_X, μ_Y, and variances σ_X^2, σ_Y^2, respectively. Then their sum $X + Y$ also has a normal distribution with mean $\mu_X + \mu_Y$, and variance $\sigma_X^2 + \sigma_Y^2$.

We can use this property to derive the sampling distribution for the mean. Equation (3) shows that $\hat{\mu}$ is a sum of independent normal random variables. Therefore, the distribution of the sum will also be normal, and in section 1.3 we calculated the mean and variance for $\hat{\mu}$. We conclude that $\hat{\mu} \sim N(\mu, \sigma^2/n)$.

2.5 Asymptotic sampling distribution for the mean

What happens if the original sample does not come from a normal distribution?

Let $X_1, X_2, \dots, X_n$ be independent random variables and suppose that all the variables have the same distribution with mean μ and variance σ^2, but we do not assume that they have normal distribution. As before, the sample average is given by

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

This variable has, approximately, a normal distribution. This is a consequence of the Central Limit Theorem, one of the fundamental results of Probability Theory.

2.5.1 Central Limit Theorem

Central Limit Theorem

Let $X_n, n \geq 1$ be a sequence of independent random variables having the same distribution with mean μ and variance σ^2. Then,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges to a standard normal distribution, as $n \to \infty$. We denote this by

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{w} N(0,1),$$

where the w stands for weak convergence, which convergence in distribution. What this means is that for any number x, and n large,

$$P\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\Big) \approx \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

We can also say that for a sample $X_1, \dots, X_n$ of size n, all variables identically distributed with mean μ and variance σ^2, the (sampling) distribution for the average $\bar{X}_n$ is approximately normal with mean μ and variance σ^2/n, if the size of the sample n is large. We denote this by

$$\bar{X}_n \approx N(\mu, \sigma^2/n)$$

The standard deviation of the sampling distribution is called the **standard error** and plays an important role in Statistics.

2.5.2 Simulation Example

We can use stochastic simulation to give an example of how the Central Limit Theorem works. For this, we will simulate 10000 samples of size 20 from an exponential distribution with parameter 1 and estimate the sample means. Then we will plot an histogram and the estimated density for the sample means and compare them with a normal distribution. Recall that the mean and variance for an exponential distribution of parameter 1 are both 1.

```r
exp.mean <- numeric(10000)
for (i in 1:10000) {
  exp.mean[i] <- mean(rexp(20))}
```

The graphs shows that the red (estimated density) and blue (normal density) curves are close, even though the sample size is small.

```
hist(exp.mean, breaks = 20, freq = FALSE, ylim=c(0,2), xlim = c(0.3,2.3),
     main='Sampling distribution for n=20', xlab='values')
lines(density(exp.mean), col='red', lwd=2)
curve(dnorm(x,1,1/sqrt(20)),-1,2.3,add = TRUE, col='blue2', lwd=2)
```

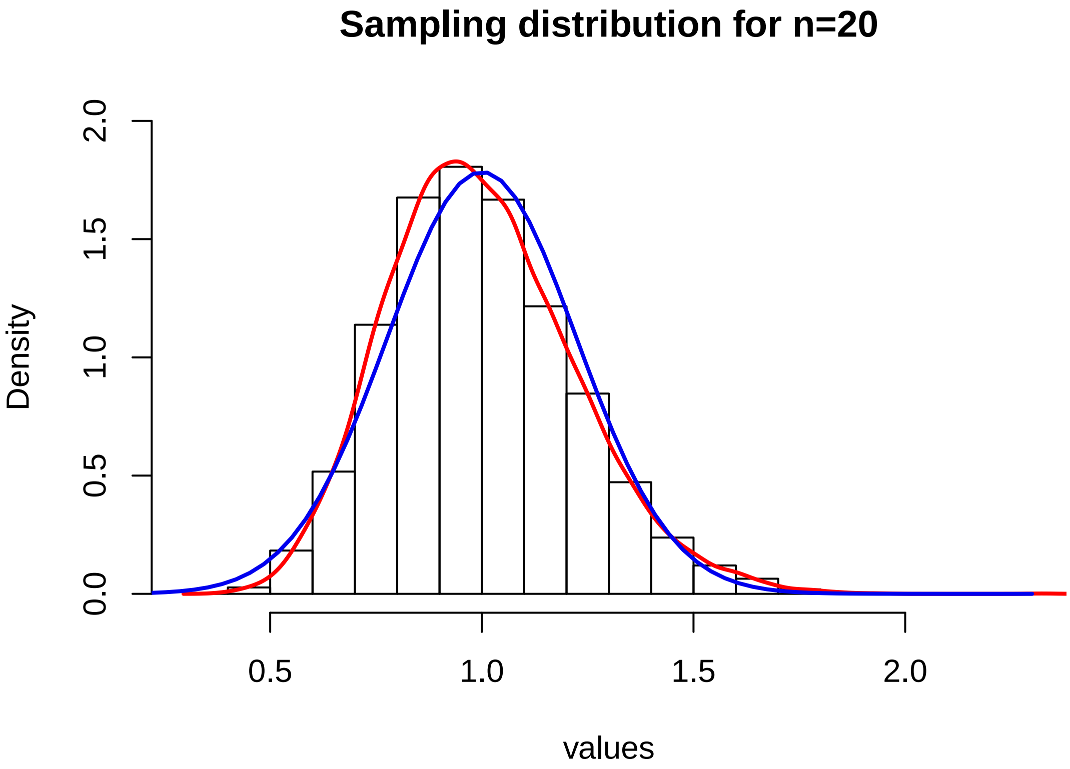

Figure 9: Sampling distribution for the empirical mean of exponential samples. The parameter for the exponential distribution is 1 and the sample size is 20. The blue curve represents the normal density.

We repeat this simulation with samples of size 50.

```
exp.mean <- numeric(10000)
for (i in 1:10000) {exp.mean[i] <- mean(rexp(50))}
hist(exp.mean, breaks = 20, freq = FALSE, ylim=c(0,3),xlim = c(0.3,2.3),
     main='Sampling distribution for n=50', xlab='values')
lines(density(exp.mean), col='red', lwd=2)
curve(dnorm(x,1,1/sqrt(50)),0,2.,add = TRUE, col='blue2', lwd=2)
```

Now the fit is better. Observe also how the standard deviation of the sampling distribution decreases as the sample size increases.

If we sample from a standard normal distribution, then, as we saw before, the average value for the sample also follows a normal distribution with mean zero and variance $1/n$. In this case the normal distribution is not an approximation to the sampling distribution: it is exactly the sampling distribution. Therefore we can plot the changes in the distribution as the sample size increases.

```
curve(dnorm(x,sd=1/sqrt(10)),-1,1,ylim=c(0,4),lwd=2,
      main = 'Sampling density', xlab = 'values', ylab = 'density')
curve(dnorm(x,sd=1/sqrt(25)),-1,1, add = TRUE, col=2,lwd=2)
```

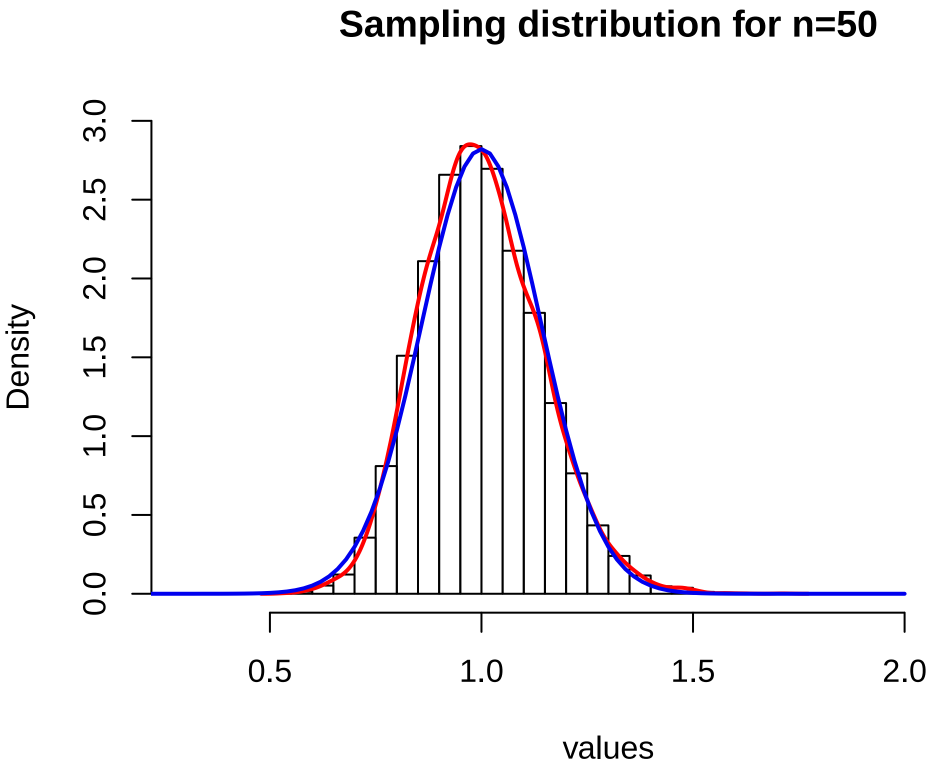

Figure 10: Sampling distribution for the empirical mean of exponential samples. The parameter for the exponential distribution is 1 and the sample size is 50. The blue curve represents the normal density.

```
curve(dnorm(x,sd=1/sqrt(50)),-1,1, add = TRUE, col=3,lwd=2)
curve(dnorm(x,sd=1/sqrt(100)),-1,1, add = TRUE, col=4,lwd=2)
legend('topright',legend = c(10,25,50,100), col=1:4, lwd=rep(2,4))
```

2.5.3 Summary

Let $X_1, \ldots, X_n$ be iid rv's with mean μ and variance σ^2.

1. The sampling density of $\bar{X}_n$ has mean μ .
2. The standard deviation of the sampling density, known as the standard error, is $\sigma/\sqrt{n}$.
3. When n is large, the sampling density of $\bar{X}_n$ approaches the normal distribution, regardless of the distribution of the population.
4. When the population distribution is normal, so is the sampling density for any value of n.
5. When n is small and the population distribution is not normal, we cannot assume that the sampling distribution is normal.
6. If the variance is not known and we need to estimate it from the sample, the sampling distribution for the normalized sample mean is t with $n-1$ degrees of freedom, where n is the size of the sample.

A Rule of Thumb

For a sample size $n > 30$, the sampling distribution for the mean for any population with mean μ and variance σ^2 can be approximated by a normal distribution with mean μ and variance σ^2/n.

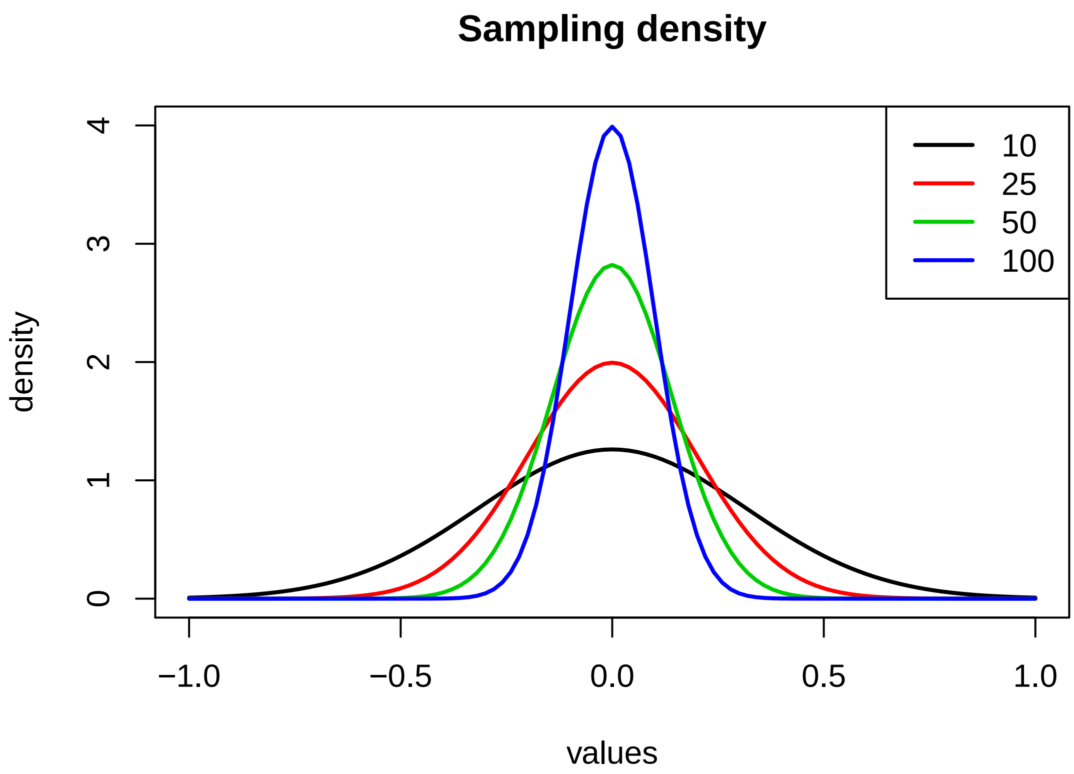

Figure 11: Sampling density for the empirical mean of standard normal samples for different sample sizes.

When n is small and the population is not normal, all we can say is that

1. The mean of the sampling density of the mean equals μ, the mean of the population.
2. The standard deviation of the sampling distribution of the mean is σ^2/n.

One way to proceed in this case is to use the bootstrap, a technique that will study later in this course.

3 Interval estimation

3.1 Introduction

In section 1 we studied how to get pointwise estimates for the mean of a distribution, using the sample mean. The result of this procedure is a number, and this is not very satisfactory, because we have no idea of how accurate this estimation is. If we have pointwise estimates coming from two different samples, one of size 1000 and another of size 10, we have no way of comparing the uncertainty associated to each one, if we only consider the single values we obtain in each case.

As we saw before, the variance associated to the first estimate will be 100 times smaller than the variance associated to the second, and this means that we should expect the first estimated to have less 'uncertainty'. One way to quantify this uncertainty is through an interval estimate, or confidence interval, that gives a range of values for the unknown parameter, and has an associated confidence level.

3.2 Confidence intervals

Let α be a number in the interval $(0, 1)$, and let $Z \sim N(0, 1)$. Let z_α be the real number defined by the relation

$$P(Z > z_\alpha) = \alpha.$$

```
points.x <- seq(-3,3,length.out = 101)
points.y <- dnorm(points.x)
xw <- points.x[points.x >= 1.74]; xw <-c(xw, 3.0,1.74)
yw <- points.y[points.x >= 1.74]; yw <- c(yw,0,0)
plot(points.x,points.y,type='l',xlab='',main='Gaussian Density',
     ylab='', axes=FALSE)
abline(h=0);  abline(v=0)
polygon(xw,yw, col='tan1')
axis(1,at=c(1.74),labels = expression(z[alpha]),line=0,pos=0,lwd=2)
text(2,.022,expression(alpha), cex =1.5)
```

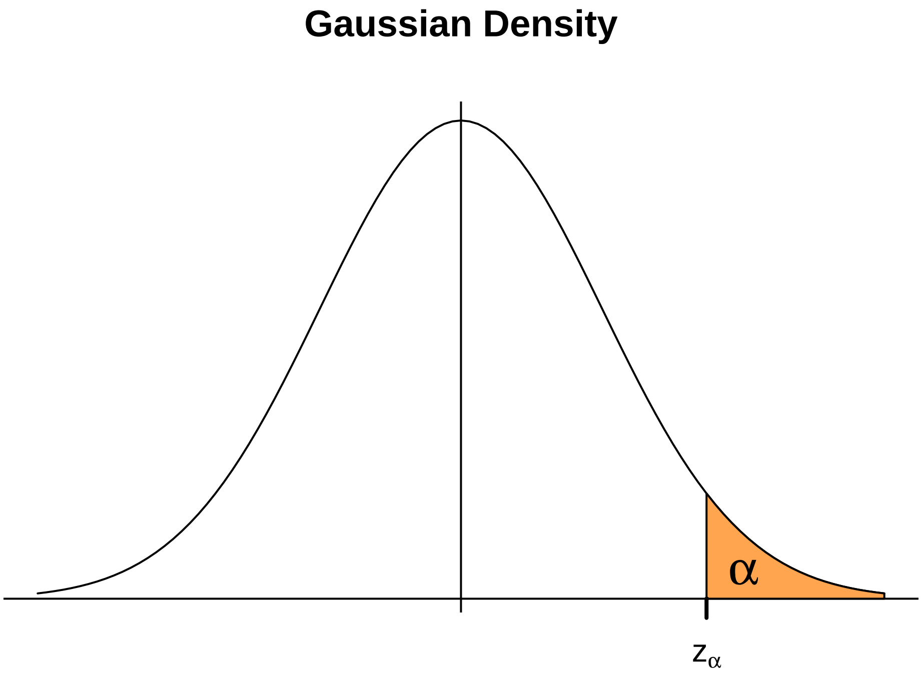

Figure 12: Definition of z_α for a Gaussian distribution

Figure 12 shows the relation between α and z_α. Observe that the value of z_α increases as α decreases: To make the area smaller, we need to move z_α further to the right. Since the normal distribution is continuous and strictly increasing, this number is unique. By the symmetry of the normal distribution,

$$P(Z < -z_\alpha) = \alpha$$

```
points.x <- seq(-3,3,length.out = 101)
points.y <- dnorm(points.x)
xw <- points.x[points.x >= 1.74]; xw <-c(xw, 3.0,1.74)
yw <- points.y[points.x >= 1.74]; yw <- c(yw,0,0)
xv <- points.x[points.x <= -1.74]; xv <-c(xv, -1.74, -3)
yv <- points.y[points.x <= -1.74]; yv <- c(yv,0,0)
plot(points.x,points.y,type='l',xlab='',main='Gaussian Density',
     ylab='', axes=FALSE)
abline(h=0.0);  abline(v=0)
polygon(xw,yw, col='tan1')
axis(1,at=c(1.74),labels = expression(z[alpha]),line=0,pos=0,lwd=2)
text(2,.022,expression(alpha), cex =1.5)
polygon(xv,yv, col='tan1')
```

```
axis(1,at=c(-1.74),labels = expression(-z[alpha]),line=0,pos=0,lwd=2)
text(-2,.022,expression(alpha), cex =1.5)
```

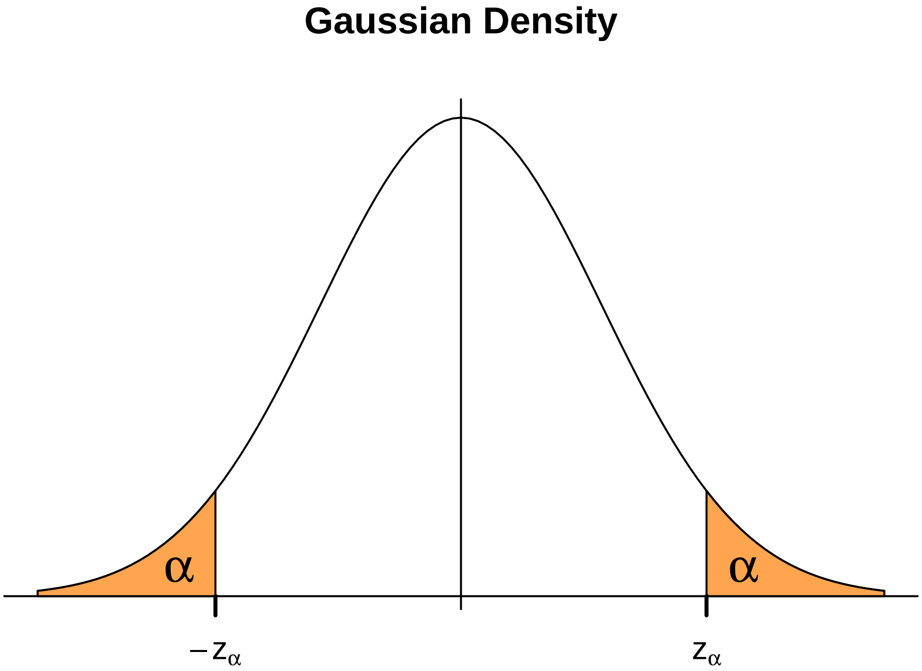

Figure 13: Definition of z_α and $-z_\alpha$ for a Gaussian distribution

This says that the probability that Z belongs to the interval $[-z_\alpha, z_\alpha]$ is $1-2\alpha$, or equivalently, using $\alpha/2$ instead of α,

$$P\big(|Z| \le z_{\alpha/2}\big) = 1-\alpha.(\#eq:A7) \tag{13}$$

Recall now that $\hat{\mu} \sim N(\mu, \sigma^2/n)$. Therefore

$$\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \sim N(0,1)$$

and this implies that $\sqrt{n}(\hat{\mu}-\mu)/\sigma$ has the same distribution as Z. Therefore, replacing Z by $\sqrt{n}(\hat{\mu}-\mu)/\sigma$ in (13),

$$P\Big(\Big|\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma}\Big| \le z_{\alpha/2}\Big) = 1-\alpha(\#eq:A8) \tag{14}$$

After some manipulation of the inequality in the expression above we get that

$$P\Big(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) = 1-\alpha.(\#eq:A9) \tag{15}$$

This equation says that, with probability $1-\alpha$, the interval

$$I(\alpha;\sigma,n) = \Big[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$$

contains the true value of the parameter μ. We say that $(1-\alpha)$ is the confidence level of the interval, frequently expressed as a percentage $100(1-\alpha)\%$. The smaller α, the higher the confidence level.

Observe that the extremes of the interval are random, and therefore the probability statement we just made applies to them, and not to the parameter, which has a fixed (but unknown) value.

The interval $I(\alpha;\sigma,n)$ is centered at $\hat{\mu}$ and depends on α, σ and n. The standard deviation σ is a parameter of the population, and is usually unknown, but fixed. We will see later on how we can deal with the situation where σ is unknown and we have to estimate it, adding uncertainty to the confidence interval. For the moment we consider it known. The width of the interval is

$$\big|I(\alpha;\sigma,n)\big| = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}(\#eq:A10) \tag{16}$$

and the smaller the width the sharper our estimate is. We saw before that making α smaller makes z_α bigger. Therefore, if we want to increase the confidence level of the interval, we need to reduce α, and this will increase z_α and make the confidence interval wider. *Higher confidence levels imply wider confidence intervals.* In other words, for fixed sample size the precision (width) of the interval and the confidence level go in opposite directions. If we want to reduce one of them, we have to increase the other.

On the other hand, if we increase sample size we reduce the width of the interval, and therefore increase its precision. For instance, if we want to reduce the width by 1/2, we need to increase sample size by 4, since we have a square root in the denominator. Increasing sample size is usually linked to costs, and therefore is not always feasible.

It is possible to use equation (16) to determine the sample size required to have a given width and confidence level, as long as we know the standard deviation σ or express the desired width in terms of σ.

3.2.1 Example 2

Suppose we want the confidence interval for the mean of a normally distributed variable to be less than or equal to one half of the (unknown) standard deviation, with a confidence level of 98%. This means that $\alpha = 0.2$. The interval width is given by expression (16) and we want this to be equal to $\sigma/2$. This gives

$$\sqrt{n} = 4z_{0.01}$$

We need to calculate $z_{0.01}$. Observe from figure 13 that $-z_\alpha$ is the α quantile of the standard normal distribution and therefore we can use the function `qnorm` to calculate its value.

```
(alp <- qnorm(0.01))
```

```
## [1] -2.326348
```

Thus, $z_{0.01} = 2.326348$ and the sample size is

```
(4*abs(alp))^2
```

```
## [1] 86.59031
```

So, a sample of size 87 would be large enough to satisfy the conditions of the example.

3.3 One-sided confidence intervals

In some situations we are only interested in one-sided confidence intervals. For example, in a chemical reaction it may be important that the temperature in the reaction does not exceed certain critical value or that a sanitizing agent is at least as effective as the standard required by a certain control agency.

From relations (13) and (14), the one-sided confidence intervals are

$$\Big(\hat{\mu} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\Big) \quad \text{and} \quad \Big(-\infty, \hat{\mu} + z_\alpha\frac{\sigma}{\sqrt{n}}\Big).$$

3.4 Confidence intervals for the mean when the variance is unknown

In the previous section we assumed that the variance was known, but this is not likely. Almost always both parameters of the distribution are unknown, so we also have to estimate the variance and this will add uncertainty to the confidence interval. In fact, the sampling distribution for the mean changes if we do not know the variance and have to estimate it from the sample.

We saw that if $\hat{\mu}_n = \frac{1}{n}\sum_1^n X_i$ and the X_i are iid with $N(\mu, \sigma^2)$ distribution then

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

In practice, we never know the true value for the variance σ^2 and we must estimate it by the empirical variance s_n^2:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

The t distribution arises as the sampling distribution of the (empirical) mean $\hat{\mu}_n$ when the data come from a normal distribution with unknown variance. It was shown in 1908 by W.S. Gosset in a paper in Biometrika, published under the pseudonym *Student*, that

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}(\#eq : A11) \tag{17}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom. For $n \geq 30$, the t distribution is very similar to the normal distribution, but for n small there are important differences. The t distribution has 'heavier' tails than the normal, which means that large values are more probable. The next figure shows several examples of densities for small values of the degrees of freedom parameter.

```
cols <- rainbow(100)
curve(dnorm(x),-3,3,lwd=2, ylab='density',col='grey25')
for (i in c(2,4,6)) {
  curve(dt(x,i),-3,3,lwd=2, add = TRUE, col=cols[55+3*i])}
curve(dt(x,30),-3,3,lwd=2, add = TRUE, col=cols[85])
legend('topright',c('normal','t2','t4','t6','t30'),
       col=c('grey25',cols[c(61,67,73,85)]),lwd=rep(2,5))
```

Now that we know the true distribution for the sample mean with unknown variance (under the hypothesis that the data are normal), we can obtain confidence intervals for the mean without assuming that the variance is known.

Let T_n be a random variable with t distribution with n degrees of freedom. We define $t_{\alpha,n}$ to be the real number that satisfies

$$P(T_n > t_{\alpha,n}) = \alpha.$$

By symmetry we have that $P(T_n < -t_{\alpha,n}) = \alpha$.

Following a similar argument as before, from equation (17) we get that

$$I^*(\alpha; n) = \left[\hat{\mu} - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right]$$

is a confidence interval of level $1-\alpha$ for the mean. One-sided confidence intervals can be obtained in a similar fashion.

3.4.1 Example 1 revisited

Recall that we have a sample of size 10 and the estimated value for the mean is

```
mean.smpl1
```

```
## [1] 4.311172
```

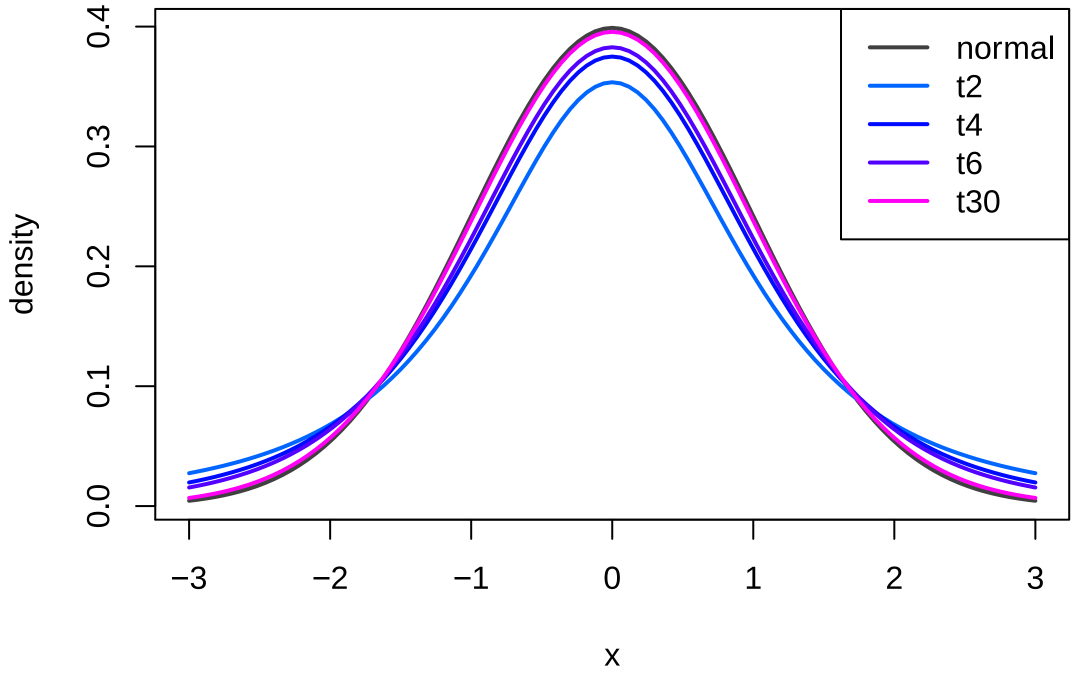

Figure 14: Density of the t distribution for different values of the frgrees of freedom parameter. The normal standard density (black) is included as reference.

We know that the sample was simulated from a normal distribution with mean 4.5 and variance 4. Since, in this case, we know the exact value for the parameters, we can build different confidence intervals and compare them. We will consider

- intervals built with the true value of the variance
- intervals built with the estimated value of the variance, but using the normal as sampling distribution.
- intervals built with the estimated value of the variance, but using the t_n as sampling distribution.

```
I10 <- matrix(numeric(6), ncol=3)
(zz <- abs(qnorm(0.025)))
```

```
## [1] 1.959964
```

Interval using known variance

```
(I10[,1] <- c(mean.smpl1 - (zz*2/sqrt(10)),mean.smpl1 + (zz*2/sqrt(10))))
```

```
## [1] 3.071582 5.550762
```

Interval using estimated variance and normal sampling distribution

```
(I10[,2] <- c(mean.smpl1 - (zz*sqrt(var.smpl1)/sqrt(10)),
  mean.smpl1 + (zz*sqrt(var.smpl1)/sqrt(10))))
```

```
## [1] 3.116586 5.505758
```

Intervals using estimated variance and t sampling distribution

```
tt9 <- abs(qt(0.025,9))
(I10[,3] <- c(mean.smpl1 - (tt9*sqrt(var.smpl1)/sqrt(10)),
  mean.smpl1 + (tt9*sqrt(var.smpl1)/sqrt(10))))
```

```
## [1] 2.932401 5.689942
```

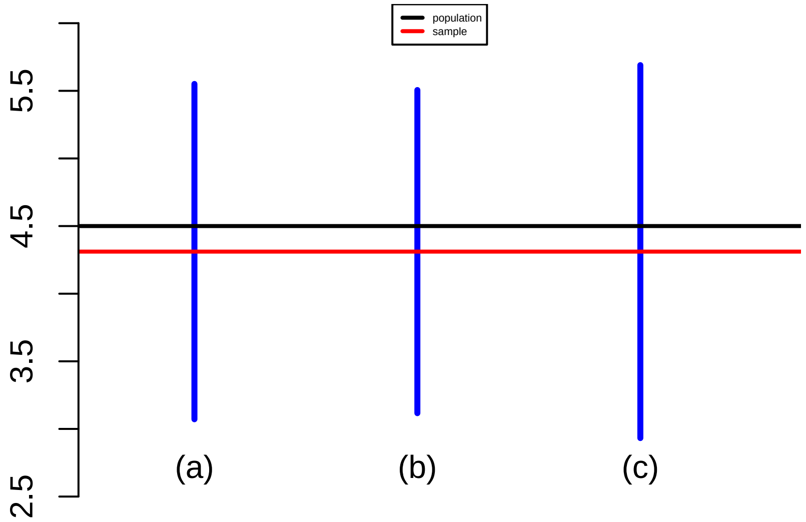

Figure 15: Confidence intervals for the mean in a sample of size 10 from the normal distribution N(4.5,4), (a) Interval using known variance, (b) Interval using estimated variance and normal sampling distribution, (c) Interval using estimated variance and t sampling distribution

Figure 15 shows the three intervals, from left to right, interval using the true value for the standard deviation, interval using estimated value for the standard deviation and a normal sampling distribution, and interval using estimated value for the standard deviation and a t_9 sampling distribution. The horizontal lines represent the true (black) and estimated (red) means. Observe that in this case, the second interval is shorter than the first. The reson for this is that for the calculation of these intervals, the only difference is the value of the standard deviation, which in the first case is the true value, 2, while in the second is the estimated value, 1.927388.

4 Hypothesis Tests

4.1 Hypothesis Tests

A *hypothesis* is an assumption about the value or values of a parameter or parameters of the population. Hypothesis are formulated in mutually exclusive pairs. If possible, hypothesis are chosen so that they are exhaustive (i.e. their union covers all possible outcomes of an experiment). This forces us to make a choice.

If θ is the parameter in question and Θ is the parameter space then the null hypothesis defines the region $\{\theta \in \Theta_0\}$ while the alternative hypothesis defines $\{\theta \in \Theta_1\}$. These regions are mutually exclusive.

When a hypothesis completely specifies the distribution of the population we say that the hypothesis is **simple**. Any hypothesis that is not simple is a **composite** hypothesis. We will only consider simple null hypothesis while the alternative will usually be composite.

Null hypothesis	Alternative hypothesis	Type of alternative
$H_0 : \theta = \theta_0$	$H_1 : \theta < \theta_0$	lower one-sided
	$H_1 : \theta > \theta_0$	upper one-sided
	$H_1 : \theta \neq \theta_0$	two-sided

Table 1: Simple null hypothesis and types of composite alternative hypotheses

In the Neyman-Pearson approach one must choose use one of the two alternatives using information from a sample, from which a test statistic is computed. The sample space is split into two regions $\mathcal{R}$, known as the

rejection region and $\mathcal{R}^c$, known as the **acceptance** region. The value of θ that separates theses two regions is known as the **critical** value. The test statitic is computed from the sample and if it falls in the acceptance region, the null hypothesis is accepted, otherwise the alternative hypothesis is accepted.

A different approach was proposed by R. Fisher. In this approach one determines how much evidence there is in the sample against the null hypothesis. The null hypothesis is not accepted but 'not rejected'. The test will determine if the sample collected can be due to chance alone under the null hypothesis. If this is not likely the researcher has evidence to reject the null hypothesis. A test with this characteristics is called a *significance test.*

4.2 p-value

The p-value for a hypothesis test is defined as the probability of observing a difference (or value) as extreme or more extreme than the observed difference (or value) under the assumption that the null hypothesis is true. If the p-value is less that the level of the test, the null hypothesis is rejected. Otherwise, there is not evidence in the sample to reject the null hypothesis. Table 2 gives formulas for calculating the p-value with different alternative hypothesis.

	p-value
$H_1 : \theta < \theta_0$	$P(t \leq t_{obs} \| H_0)$
$H_1 : \theta > \theta_0$	$P(t \geq t_{obs} \| H_0)$
$H_1 : \theta \neq \theta_0$	$2\min\{P(t \leq t_{obs} \| H_0), P(t \geq t_{obs} \| H_0)\}$

Table 2: Calculation of p-values for continuous distributions.

4.3 Types of error

Our decision is always subject to error. There are two types of error:

Type I: Reject H_0 when it is true

Type II: Fail to reject H_0 when it is false.

State of Nature	H_0 not rejected	H_0 rejected
H_0 is true	No error	**Type I error**
H_0 is false	**Type II error**	No error

Table 3: Types of error

Each error has a probability associated with it. We denote by α the probability of a Type I error and by β the probability of a Type II error. α is known as the **level** or **significance level** of the test. For a fixed sample size, α and β go in opposite ways: the smaller α, the larger β, so we cannot reduce both at the same time unless we change the sample size.

Shiny app for hypothesis tests: https://casertamarco.shinyapps.io/power/

4.4 Power

Consider

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta = \theta_1$$

where both hypotheses are simple. The **power** of this test is the probability of rejecting H_0 when H_1 is true.

$$P(\text{Reject } H_0 | \theta = \theta_1) = 1 - P(\text{Accept } H_0 | \theta = \theta_1) = 1 - \beta(\theta_1)$$

The power reflects the capacity of the test to detect the alternative hypothesis when it is true.Observe that we need to know the sampling distribution for the test statistic *under the alternative hypothesis.*

If instead of a simple alternative we have a composite hypothesis $H_1 : \theta \in \Theta_1$, we consider the power as a function of θ for values $\theta \in \Theta_1$.

4.4.1 Example (from Ugarte et al.)

Test the null hypothesis that for a certain age group the mean score on an achievement test is equal to 40 against the alternative that it is not equal to 40. Scores follow a normal distribution with $\sigma = 6$.

(a) Find the probability of type I error for $n = 9$ if the null hypothesis is rejected when the sample mean is less than 36 or greater than 44.

(b) Find the probability of type I error for $n = 36$ if the null hypothesis is rejected when the sample mean is less than 38 or greater than 42.

(c) Plot the power functions for $n = 9$ and $n = 36$ for values of μ between 30 and 50.

Let $\bar{X}_n$ denote the sample mean when the sample size is n. We know that

$$\bar{X}_n \sim N(\mu, \sigma/\sqrt{n}) = N(\mu, 6/\sqrt{n})$$

For n given, this distribution depends only on μ.

1. In the first question we reject H_0 when

$$\bar{X}_9 < 36 \quad \text{or} \quad \bar{X}_9 > 44.$$

Therefore, the probability of a type I error, α is

$$\begin{aligned}
\alpha &= P(\{\bar{X}_9 < 36\} \cup \{\bar{X}_9 > 44\} | \mu = 40) \\
&= P(\{\bar{X}_9 < 36\} | \mu = 40) + P(\{\bar{X}_9 > 44\} | \mu = 40) \\
&= P\Big(\frac{\bar{X}_9 - 40}{6/\sqrt{9}} < \frac{36 - 40}{6/\sqrt{9}}\Big) + P\Big(\frac{\bar{X}_9 - 40}{6/\sqrt{9}} > \frac{44 - 40}{6/\sqrt{9}}\Big) \\
&= P(N(0,1) < -2) + P(N(0,1) > 2) \\
&= 2P(N(0,1) < -2)
\end{aligned}$$

```
2*pnorm(-2)
```

```
## [1] 0.04550026
```

2. In the second question we reject H_0 when

$$\bar{X}_{36} < 38 \quad \text{or} \quad \bar{X}_{36} > 42.$$

A similar calculation as before shows that in this case we get exactly the same value for α.

3. The power function for $n = 9$ is

$$\begin{aligned}
Power(\mu) &= P(\bar{X}_9 < 36 | N(\mu, \frac{6}{\sqrt{9}})) + P(\bar{X}_9 > 44 | N(\mu, \frac{6}{\sqrt{9}})) \\
&= P(\frac{\bar{X}_9 - \mu}{2} < \frac{36 - \mu}{2}) + P(\frac{\bar{X}_9 - \mu}{2} > \frac{44 - \mu}{2}) \\
&= P(N < \frac{36 - \mu}{2}) + P(N > \frac{44 - \mu}{2})
\end{aligned}$$

while the power function for $n = 36$ is

$$\begin{aligned}
Power(\mu) &= P(\{\bar{X}_{36} < 38\} | N(\mu, \frac{6}{\sqrt{36}})) + P(\{\bar{X}_{36} > 42\} | N(\mu, \frac{6}{\sqrt{36}}) \\
&= P(\{\frac{\bar{X}_{36} - \mu}{1} < \frac{38 - \mu}{1}\}) + P(\{\frac{\bar{X}_{36} - \mu}{1} > \frac{42 - \mu}{1}\}) \\
&= P(N < 38 - \mu) + P(N > 42 - \mu)
\end{aligned}$$

```r
power.ex <- function(x,n,a){
  1-pnorm(40+a, x,6/sqrt(n)) + pnorm(40-a, x,6/sqrt(n))}
curve(power.ex(x,9,4),30,50, ylab=expression(Power(mu)),
      xlab=expression(mu), ylim=c(0,1), lwd=2, col='steelblue3')
curve(power.ex(x,36,2),30,50,add = TRUE, lwd=2, col='steelblue4')
arrows(32, 0.6 , 34.2, .78, lwd=2, length=0.05, col='steelblue3')
arrows(32, 0.35 , 37, .78, lwd=2, length=0.05, col='steelblue4')
arrows(40, 0.4 , 40, 0.06, lwd=2, length=0.05)
text(32,0.58, expression(n==9))
text(32.3,0.33, expression(n==36))
text(40,0.45, expression(alpha==0.045))
```

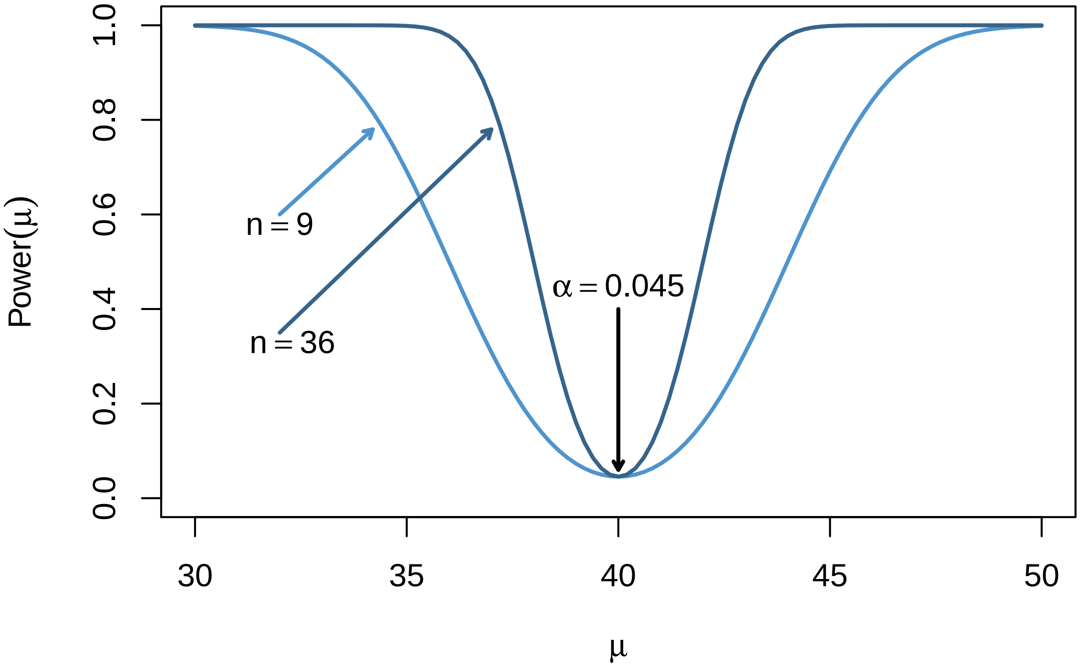

Figure 16: Power function.

‘