Chapter 3

Statistics for one and two samples

Contents

3.1 Learning from one-sample quantitative data

Statistics is the art and science of learning from data, i.e. statistical inference. What we are usually interested in learning about is the population from which our sample was taken, as described in Section 1.3. More specifically, most of the time the aim is to learn about the mean of this population, as illustrated in Figure 1.1.

|||| Example 3.1 Student heights

In examples in Chapter 1 we did descriptive statistics on the following random sample of the heights of 10 students in a statistics class (in cm):

$$168 \quad 161 \quad 167 \quad 179 \quad 184 \quad 166 \quad 198 \quad 187 \quad 191 \quad 179$$

and we computed the sample mean and standard deviation to be

$$\begin{aligned} \bar{x} &= 178, \\ s &= 12.21. \end{aligned}$$

The population distribution of heights will have some unknown mean μ and some unknown standard deviation σ. We use the sample values as point estimates for these population parameters

$$\begin{aligned} \hat{\mu} &= 178, \\ \hat{\sigma} &= 12.21. \end{aligned}$$

Since we only have a sample of 10 persons, we know that the point estimate of 178 cannot with 100% certainty be exactly the true value μ (if we collected a new sample with 10 different persons height and computed the sample mean we would definitely expect this to be different from 178). The way we will handle this uncertainty is by computing an interval called the *confidence interval* for μ. The confidence interval is a way to handle the uncertainty by the use of probability theory. The most commonly used confidence interval would in this case be

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}},$$

which is

$$178 \pm 8.74.$$

The number 2.26 comes from a specific probability distribution called the t-distribution, presented in Section 2.86. The t-distributions are similar to the standard normal distribution presented in Section 2.5.2: they are symmetric and centred around 0.

The confidence interval interval

$$178 \pm 8.74 = [169.3,\ 186.7],$$

represents the plausible values of the unknown population mean μ in light of the data.

So in this section we will explain how to *estimate* the mean of a distribution and how to quantify the *precision*, or equivalently the *uncertainty*, of our estimate.

We will start by considering a population characterized by some distribution from which we take a sample $x_1, \ldots, x_n$ of size n. In the example above X_i would be the height of a randomly selected person and $x_1, \ldots, x_{10}$ our sample of student heights.

A crucial issue in the confidence interval is to use the correct probabilities, that is, we must use probability distributions that are properly representing the real life phenomena we are investigating. In the height example, the population distribution is the distribution of all heights in the entire population. So, this is what you would see if you sampled from a huge amount of heights, say $n = 1000000$, and then made a density histogram of these, see Example 1.25. Another way of saying the same is: the random variables X_i have a probability density function (*pdf* or $f(x)$) which describe exactly the distribution of all the values. Well, in our setting we have only a rather small sample, so in fact we may have to assume some specific *pdf* for X_i, since we don't know it and really can't see it well from the small sample. The most common type of assumption, or one could say *model*, for the population distribution is to assume it to be the normal distribution. This assumption makes the theoretical justification for the methods easier. In many cases real life phenomena actually indeed are nicely modelled by a normal distribution. In many other cases they are not. After taking you through the methodology based on a normal population distribution assumption, we will show and discuss what to do with the non-normal cases.

Hence, we will assume that the *random variable* X_i follows a *normal distribution* with mean μ and variance σ^2:

|||| Remark 3.2 How to write a statistical model

In all statistical analysis there must be an assumption of a *model*, which should be stated clearly in the presentation of the analysis. The model expressing that the sample was taken randomly from the population, which is normal distributed, can be written by

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, \ldots, n. \tag{3-1}$$

Hence we n random variables representing the sample and they are *independent and identically distributed* (i.i.d).

Our goal is to learn about the mean of the population μ, in particular, we want to:

1. *Estimate* μ, that is calculate a best guess of μ based on the sample
2. Quantify the precision, or equivalently the uncertainty, of the estimate

Intuitively, the best guess of the population mean μ is the sample mean

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

Actually, there is a formal theoretical framework to support that this sort of obvious choice also is the theoretically best choice, when we have assumed that the underlying distribution is normal. The next sections will be concerned with answering the second question: quantifying how precisely $\bar{x}$ estimates μ, that is, how close we can expect the sample mean $\bar{x}$ to be to the true, but unknown, population mean μ. To answer this, we first, in Section 3.1.1, discuss the distribution of the sample mean, and then, in Section 3.1.2, discuss the *confidence interval* for μ, which is universally used to quantify precision or uncertainty.

3.1.1 Distribution of the sample mean

As indicated in Example 3.1 the challenge we have in using the sample mean $\bar{x}$ as an estimate of μ is the unpleasant fact that the next sample we take would give us a different result, so there is a clear element of randomness in our estimate. More formally, if we take a new sample from the population, let us call it $x_{2,1}, \ldots, x_{2,n}$, then the sample mean of this, $\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2,i}$ will be different

from the sample mean of the first sample we took. In fact, we can repeat this process as many times as we would like, and we would obtain:

1. Sample $x_{1,1}, \ldots, x_{1,n}$ and calculate the average $\bar{x}_1$
2. Sample $x_{2,1}, \ldots, x_{2,n}$ and calculate the average $\bar{x}_2$
3. Sample $x_{3,1}, \ldots, x_{3,n}$ and calculate the average $\bar{x}_3$
4. etc.

Since the sample means $\bar{x}_j$ will all be different, it is apparent that *the sample mean is also the realization of a random variable*. In fact it can be shown that if X is a random variable with a normal distribution with mean μ and variance σ^2, then the random sample mean $\bar{X}$ from a sample of size n is also a normally distributed random variable with mean μ and variance σ^2/n. This result is formally expressed in the following theorem:

|||| Theorem 3.3 The distribution of the mean of normal random variables

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{3-2}$$

Note how the formula in the theorem regarding the mean and variance of $\bar{X}$ is a consequence of the *mean and variance of linear combinations* Theorem 2.56

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu, \tag{3-3}$$

and

$$\mathrm{V}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}, \tag{3-4}$$

and using Theorem 2.40 it is clear that the mean of normal distributions also is a normal distribution.

One important point to read from this theorem is that it tells us, at least theoretically, what the variance of the sample mean is, and hence also the standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \tag{3-5}$$

Let us elaborate a little on the importance of this. Due to the basic rules for mean and variance calculations, i.e. Theorem 2.56, we know that the difference between $\bar{X}$ and μ has the same standard deviation

$$\sigma_{(\bar{X}-\mu)} = \frac{\sigma}{\sqrt{n}}. \tag{3-6}$$

This is the mean absolute difference between the sample estimate $\bar{X}$ and the true μ, or in other words: this is the mean of the error we will make using the sample mean to estimate the population mean. This is exactly what we are interested in: to use a probability distribution to handle the possible error we make.

In our way of justifying and making explicit methods it is useful to consider the so-called *standardized sample mean*, where the $\bar{X} - \mu$ is seen relative to its standard deviation, and using the standardization of normal distributions in Theorem 2.43, which states that the standardized sample mean has a standard normal distribution:

|||| Theorem 3.4 The distribution of the σ-standardized mean of normal random variables

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N\left(\mu, \sigma^2\right)$ where $i = 1, \ldots, n$, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1^2\right). \tag{3-7}$$

That is, the standardized sample mean Z follows a standard normal distribution.

However, to somehow use the probabilities to say something clever about how close the estimate $\bar{x}$ is to μ, all these results have a flaw: the population standard deviation σ (true, but unknown) is part of the formula. And in most practical cases we don't know the true standard deviation σ. The natural thing to do is to use the sample standard deviation s as a substitute for (estimate of) σ. However, then the theory above breaks down: the sample mean standardized by the sample standard deviation instead of the true standard deviation no longer has

a normal distribution! But luckily the distribution can be found (as a probability theoretical result) and we call such a distribution a t-distribution with $(n-1)$ *degrees of freedom* (for more details see Section 2.10.2):

|||| Theorem 3.5 The distribution of the S-standardized mean of normal random variables

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, where $X_i \sim N\left(\mu, \sigma^2\right)$ and $i = 1, \ldots, n$, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1), \tag{3-8}$$

where $t(n-1)$ is the t-distribution with $n-1$ degrees of freedom.

A t-distribution, as any other distribution, has a probability density function, presented in Definition 2.86. It is similar in shape to the standard normal distribution: it is symmetric and centred around 0, but it has thicker tails as illustrated in the figure of Example 2.92. Also, the t-distributions are directly available in Python, via the SciPy package as seen also for the other probability distributions, see the overview of distributions in A.2.1. So we can easily work with t-distributions in practice. As indicated, there is a different t-distribution for each n: the larger the n, the closer the t-distribution is to the standard normal distribution.

|||| Example 3.6 Normal and t probabilities and quantiles

In this example we compare some probabilities from the standard normal distribution with the corresponding ones from the t-distribution with various numbers of degrees of freedom.

Let us compare $P(T > 1.96)$ for some different values of n with $P(Z > 1.96)$:

```
# The P(T>1.96) probability  for n=10
print(1-stats.t.cdf(1.96,df=9))

0.04082220273020831

# The P(Z>1.96) probability
print(1-stats.norm.cdf(1.96))

0.024997895148220484

# The P(T>1.96) probability  for n-values, 10, 20, ... ,50
print(1-stats.t.cdf(1.96,df=np.linspace(10, 50, 5)-1))

[0.041 0.032 0.030 0.029 0.028]

# The P(T>1.96) probability  for n-values, 100, 200, ... ,500
print(1-stats.t.cdf(1.96,df=np.linspace(100, 500, 5)-1))

[0.026 0.026 0.025 0.025 0.025]
```

Note how the t-probabilities approach the standard normal probabilities as n increases. Similarly for the quantiles:

```
# The standard normal 97.5% quantile
print(stats.norm.ppf(0.975,loc=0,scale=1))

1.959963984540054

# The t-quantiles for n-values: 10, 20, ... ,50
# (rounded to 3 decimal points)
print(stats.t.ppf(0.975,df=np.linspace(10, 50, 5)-1))

[2.262 2.093 2.045 2.023 2.010]

# The t-quantiles for n-values: 100, 200, ... ,500
# (rounded to 3 decimal points)
print(stats.t.ppf(0.975,df=np.linspace(100, 500, 5)-1))

[1.984 1.972 1.968 1.966 1.965]
```

The sample version of the standard deviation of the sample mean $s/\sqrt{n}$ is called the *Standard Error of the Mean* (and is often abbreviated *SEM*):

Definition 3.7 Standard Error of the mean

Given a sample $X_1, \ldots, X_n$, the *Standard Error of the Mean* is defined as

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}}. \tag{3-9}$$

It can also be read as the *Sampling Error* of the mean, and can be called the standard deviation of the *sampling distribution* of the mean.

Remark 3.8

Using the phrase *sampling distribution* as compared to just the *distribution* of the mean bears no mathematical/formal distinction: formally a probability distribution is a probability distribution and there exist only one definition of that. It is merely used to emphasize the role played by the distribution of the sample mean, namely to quantify how the sample mean changes from (potential) sample to sample, so more generally, the sample mean has a distribution (from sample to sample), so most textbooks and e.g. Wikipedia would call this distribution a sampling distribution.

3.1.2 Quantifying the precision of the sample mean - the confidence interval

As already discussed above, estimating the mean from a sample is usually not enough: we also want to know how close this estimate is to the true mean (i.e. the population mean). Using knowledge about probability distributions, we are able to quantify the uncertainty of our estimate even without knowing the true mean. Statistical practice is to quantify precision (or, equivalently, uncertainty) with a *confidence interval (CI)*.

In this section we will provide the explicit formula for and discuss confidence intervals for the population mean μ. The theoretical justification, and hence assumptions of the method, is a normal distribution of the population. However,

it will be clear in a subsequent section that the applicability goes beyond this if the sample size n is large enough. The standard so-called one-sample confidence interval method is:

|||| Method 3.9 The one sample confidence interval for μ

For a sample $x_1, \ldots, x_n$ the $100(1-\alpha)\%$ confidence interval is given by

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \tag{3-10}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile from the t-distribution with $n-1$ degrees of freedom.[a]
Most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}. \tag{3-11}$$

[a] Note how the dependence of n has been suppressed from the notation to leave room for using the quantile as index instead - since using two indices would appear less readable: $t_{n-1,1-\alpha/2}$

We will reserve the **Method** boxes for specific directly applicable statistical methods/formulas (as opposed to theorems and formulas used to explain, justify or prove various points).

|||| Example 3.10 Student heights

We can now use Method 3.9 to find the 95% confidence interval for the population mean height from the height sample from Example 3.1. We need the 0.975-quantile from the t-distribution with $n-1=9$ degrees of freedom:

```
# The t-quantiles for n=10:
print(stats.t.ppf(0.975,df=9))

2.2621571628540993
```

And we can recognize the already stated result

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}},$$

which is

$$178 \pm 8.74 = [169.3,\ 186.7].$$

Therefore with high confidence we conclude that the true mean height of the population of students to be between 169.3 and 186.7.

The confidence interval is widely used to summarize uncertainty, not only for the sample mean, but also for many other types of estimates, as we shall see in later sections of this chapter and in following chapters. It is quite common to use 95% confidence intervals, but other levels, e.g. 99% are also used (it is presented later in this chapter what the precise meaning of "other levels" is).

|||| Example 3.11 Student heights

Let us try to find the 99% confidence interval for μ for the height sample from Example 3.1. Now $\alpha = 0.01$ and we get that $1 - \alpha/2 = 0.995$, so we need the 0.995-quantile from the t-distribution with $n - 1 = 9$ degrees of freedom:

```
# The t-quantile for n=10
print(stats.t.ppf(0.995,df=9))

3.2498355415921254
```

And we can find the result as

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}},$$

which is:

$$178 \pm 12.55 = [165.5,\ 190.5].$$

Or explicitly in Python:

```
# The 99% confidence interval for the mean
x = np.array([168, 161, 167, 179, 184, 166, 198, 187, 191, 179])
n = len(x)
print(x.mean() - stats.t.ppf(0.995,df=9) * x.std(ddof=1) / np.sqrt(n))
```

```
165.45078999139582
```

```
print(x.mean() + stats.t.ppf(0.995,df=9) * x.std(ddof=1) / np.sqrt(n))
```

```
190.54921000860418
```

Or using the function `stats.t.interval` from the SciPy package:

```
# The 99% confidence interval for the mean
stats.t.interval(0.99,df=n-1,loc=x.mean(),
scale=x.std(ddof=1)/np.sqrt(n))
```

```
(np.float64(165.45078999139582), np.float64(190.54921000860418))
```

Later we will introduce a function from the SciPy package that performs a "*t*-test", which can also be used to calculate confidence intervals.

In our motivation of the confidence interval we used the assumption that the population is normal distributed. Thankfully, as already pointed out above, the validity is not particularly sensitive to the normal distribution assumption. In later sections, we will discuss how to assess if the sample is sufficiently close to a normal distribution, and what we can do if the assumption is not satisfied.

3.1.3 The language of statistics and the process of learning from data

In this section we review what it means to make statistical inference using a confidence interval. We review the concepts, first presented in Section 1.3, of: a population, distribution, a parameter, an estimate, an estimator, and a statistic.

The basic idea in statistics is that there exists a *statistical* population (or just

population) which we want to know about or learn about, but we only have a *sample* from that population. The idea is to use the sample to say something about the population. To generalize from the sample to the population, we characterize the population by a distribution (see Definition 1.1 and Figure 1.1).

For example, if we are interested in the weight of eggs lain by a particular species of hen, the population consists of the weights of all currently existing eggs as well as weights of eggs that formerly existed and will (potentially) exist in the future. We may characterize these weights by a normal distribution with mean μ and variance σ^2. If we let X denote the weight of a randomly chosen egg, then we may write $X \sim N(\mu, \sigma^2)$. We say that μ and σ^2 are the parameters of this distribution - we call them *population parameters*.

Naturally, we do not know the values of these *true* parameters, and it is impossible for us to ever know, since it would require that we weighed all possible eggs that have existed or could have existed. In fact the true parameters of the distribution $N(\mu, \sigma^2)$ are unknown and will forever remain unknown.

If we take a random sample of eggs from the population of egg weights, say we make 10 *observations*, then we have $x_1, \ldots, x_{10}$. We call this the *observed sample* or just *sample*. From the sample, we can calculate the *sample mean*, $\bar{x}$. We say that $\bar{x}$ is an *estimate* of the true *population mean* μ (or just *mean*, see Remark 1.3). In general we distinguish estimates of the parameters from the parameters themselves, by adding a hat (circumflex). For instance, when we use the sample mean as an estimate of the mean, we may write $\hat{\mu} = \bar{x}$ for the estimate and μ for the parameter, see the illustration of this process in Figure 1.1.

We denote parameters such as μ and σ^2 by Greek letters. Therefore parameter estimates are Greek letters with hats on them. Random variables such as X are denoted by capital Roman letters. The observed values of the random variables are denoted by lower case instead – we call them *realizations* of the random variables. For example, the sample $x_1, \ldots, x_{10}$ represents actually observed numbers (e.g. the weights of 10 eggs), so they are not random and therefore in lower case. If we consider a *hypothetical* sample it is yet unobserved and therefore random and denoted by, say, $X_1, \ldots, X_n$ and therefore in capital letters, see also Section 2.1.

To emphasize the difference, we say that $X_1, \ldots, X_n$ is a *random sample*, while we say that $x_1, \ldots, x_n$ is a *sample taken at random*; the observed sample is not random when it is observed, but it was produced as a result of n random experiments.

A *statistic* is a function of the data, and it can represent both a fixed value from an observed sample or a random variable from a random (yet unobserved) sample. For example sample average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a statistic computed from an observed sample, while $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is also a statistic, but it is considered

a function of a random (yet unobserved) sample. Therefore $\bar{X}$ is itself a random variable with a distribution. Similarly the sample variance S^2 is a random variable, while s^2 is its realized value and just a number.

An *estimator* (not to be confused with an *estimate*) is a *function* that produces an estimate. For example, μ is a parameter, $\hat{\mu}$ is the estimate and we use $\bar{X}$ as an *estimator* of μ. Here $\bar{X}$ is the function that produces the estimate of μ from a sample.

Learning from data is learning about parameters of distributions that describe populations. For this process to be meaningful, the sample should in a meaningful way be representative of the relevant population. One way to ensure that this is the case is to make sure that the sample is taken completely at random from the population, as formally defined here:

Definition 3.12 Random sample

A random sample from an (infinite) population: A set of observations $X_1, ..., X_n$ constitutes a random sample of size n from the infinite population $f(x)$ if:

1. Each X_i is a random variable whose distribution is given by $f(x)$
2. The n random variables are independent

It is a bit difficult to fully comprehend what this definition really amounts to in practice, but in brief one can say that the observations should come from the same population distribution, and that they must each represent truly new information (the independence).

Remark 3.13

Throughout previous sections and the rest of this chapter we assume infinite populations. Finite populations of course exists, but only when the sample constitutes a large proportion of the entire population, is it necessary to adjust the methods we discuss here. This occurs relatively infrequently in practice and we will not discuss such conditions.

3.1.4 When we cannot assume a normal distribution: the Central Limit Theorem

The Central Limit Theorem (CLT) states that the sample mean of independent identically distributed (i.i.d.) random variables converges to a normal distribution:

Theorem 3.14 Central Limit Theorem (CLT)

Let $\bar{X}$ be the sample mean of a random sample of size n taken from a population with mean μ and variance σ^2, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \tag{3-12}$$

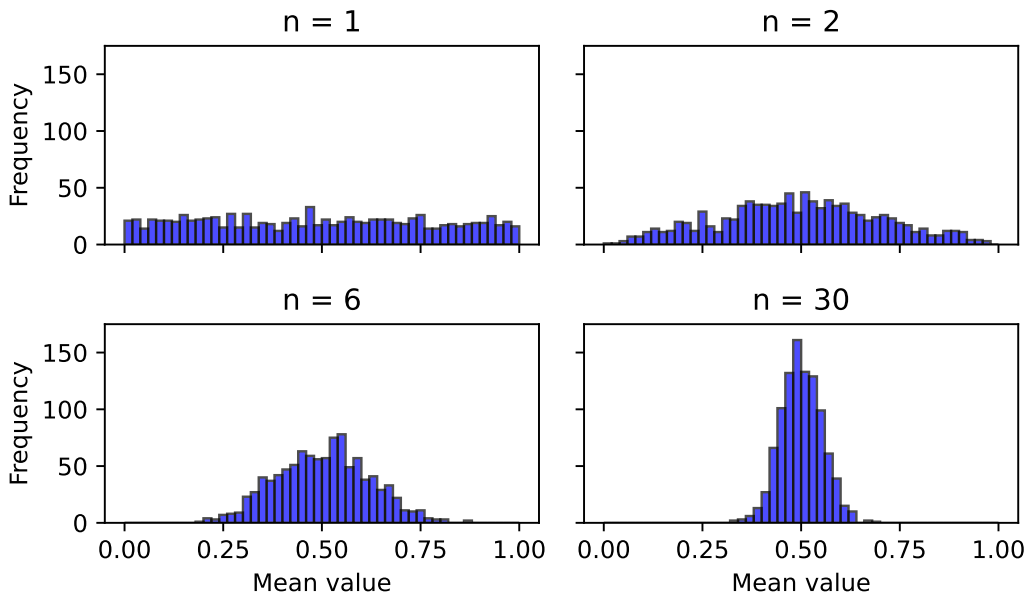
is a random variable which distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \to \infty$. In other words, for large enough n, it holds approximately that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2). \tag{3-13}$$

The powerful feature of the CLT is that, when the sample size n is large enough, the distribution of the sample mean $\bar{X}$ is (almost) independent of the distribution of the population X. This means that the underlying distribution of a sample can be disregarded when carrying out inference related to the mean. The variance of the sample mean can be estimated from the sample and it can be seen that as n increases the variance of the sample mean decreases, hence the "accuracy" with which we can infer increases.

Example 3.15 Central Limit Theorem in practice

```
# Number of simulated samples
k = 1000
# Number of observations in each sample
n = 1
# Simulate k samples with n observations
Xbar1 = stats.uniform.rvs(0,1, size=(k,n))
# Increase the number of observations in each sample
n = 2
Xbar2 = pd.DataFrame(stats.uniform.rvs(0,1, size=(k,n))).mean(axis=1)
# Increase the number of observations in each sample
n = 6
Xbar6 = pd.DataFrame(stats.uniform.rvs(0,1, size=(k,n))).mean(axis=1)
# Increase the number of observations in each sample
n = 30
Xbar30 = pd.DataFrame(stats.uniform.rvs(0,1, size=(k,n))).mean(axis=1)
# Plot the histograms
fig, axs = plt.subplots(2,2)
axs[0,0].hist(Xbar1, bins=50, range=[0,1], edgecolor='black', color='blue',
alpha=0.7)
axs[0,1].hist(Xbar2, bins=50, range=[0,1], edgecolor='black', color='blue',
alpha=0.7)
axs[1,0].hist(Xbar6, bins=50, range=[0,1], edgecolor='black', color='blue',
alpha=0.7)
axs[1,1].hist(Xbar30, bins=50, range=[0,1], edgecolor='black', color='blue',
alpha=0.7)
plt.tight_layout()
plt.show()
```



Notice how the plot resembles the front page.

Due to the amazing result of the Central Limit Theorem 3.14 many expositions of classical statistics provides a version of the confidence interval based on the standard normal quantiles rather than the t-quantiles

$$\bar{x} \pm z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}. \tag{3-14}$$

We present it here only as an interesting limit situation of the t-based interval in Method 3.9.

For large samples, the standard normal distribution and t-distribution are almost the same, so in practical situations, it doesn't matter whether the normal based or the t-based confidence interval (CI) is used. Since the t-based interval is also valid for small samples when a normal distribution is assumed, we recommend that the t-based interval in Method 3.9 is used in all situations. This recommendation also has the advantage that the SciPy-function `stats.t.interval`, which produces the t-based interval, can be used in all cases.

How large should the sample then be in a non-normal case to ensure the validity of the interval? No general answer can be given, but as a rule of thumb we recommend $n \geq 30$.

When we have a small sample for which we cannot or will not make a normality assumption, we have not yet presented a valid CI method. The classical solution is to use the so-called non-parametric methods. However, in the next chapter we will present the more widely applicable *simulation* or *re-sampling* based techniques.

3.1.5 Repeated sampling interpretation of confidence intervals

In this section we show that 95% of the 95% confidence intervals we make will cover the true value in the long run. Or, in general $100(1-\alpha)\%$ of the $100(1-\alpha)\%$ confidence intervals we make will cover the true value in the long run. For example, if we make 100 95% CI we cannot guarantee that exactly 95 of these will cover the true value, but if we repeatedly make 100 95% CIs then *on average* 95 of them will cover the true value.

|||| Example 3.16 Simulating many confidence intervals

To illustrate this with a simulation example, then we can generate 50 random $N(1, 1^2)$ distributed numbers and calculate the t-based CI given in Method 3.9, and then repeated this 1000 times to see how many times the true mean $\mu = 1$ is covered. The following code illustrates this:

```
# Simulate 1000 samples each with 50 observations
x = pd.DataFrame(stats.norm.rvs(loc=1,scale=1,size=(1000,50)))
# Calculate a 95% CI from each sample
CIs = stats.t.interval(0.95,df=50-1,loc=x.mean(axis=1),
scale=x.std(ddof=1,axis=1)/np.sqrt(50))
# Count how often 1 is covered
print(np.sum((CIs[0] <= 1) & (CIs[1] >= 1)))

954
```

Hence in 954 of the 1000 repetitions (i.e. 95.4%) the CI covered the true value. If we repeat the whole simulation over, we would obtain 1000 different samples and therefore 1000 different CIs. Again we expect that approximately 95% of the CIs will cover the true value $\mu = 1$.

The result that we arrived at by simulation in the previous example can also be derived mathematically. Since

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

where t is the t-distribution with $n-1$ degrees of freedom, it holds that

$$1 - \alpha = P\left(-t_{1-\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{1-\alpha/2}\right),$$

which we can rewrite as

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

Thus, the probability that the interval with limits

$$\bar{X} \pm t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \tag{3-15}$$

covers the true value μ is exactly $1 - \alpha$. One thing to note is that the only difference between the interval above and the interval in Method 3.9, is that the

interval above is written with capital letters (simply indicating that it calculated with random variables rather than with observations).

This shows exactly that $100(1-\alpha)\%$ of the $100(1-\alpha)\%$ confidence interval we make will contain the true value in the long run.

3.1.6 Confidence interval for the variance

In previous sections we discussed how to calculate a confidence interval for the mean. In this section we discuss how to calculate a confidence interval for the variance or the standard deviation.

We will assume that the observations come from a normal distribution throughout this section, and we will not present any methods that are valid beyond this assumption. While the methods for the sample mean in the previous sections are not sensitive to (minor) deviations from the normal distribution, the methods discussed in this section for the sample variance rely much more heavily on the correctness of the normal distribution assumption.

|||| Example 3.17 Tablet production

In the production of tablets, an active matter is mixed with a powder and then the mixture is formed to tablets. It is important that the mixture is homogeneous, such that each tablet has the same strength.

We consider a mixture (of the active matter and powder) from where a large amount of tablets is to be produced.

We seek to produce the mixtures (and the final tablets) such that the mean content of the active matter is 1 mg/g with the smallest variance possible. A random sample is collected where the amount of active matter is measured. It is assumed that all the measurements follow a normal distribution.

The variance estimator, that is, the formula for the variance seen as a random variable, is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{3-16}$$

where n is the number of observations, X_i is observation number i where $i = 1, \ldots, n$, and $\bar{X}$ is the estimator of the mean of X.

The (sampling) distribution of the variance estimator is the χ^2-distribution distribution: let S^2 be the variance of a sample of size n from a normal distribution with variance σ^2, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}, \tag{3-17}$$

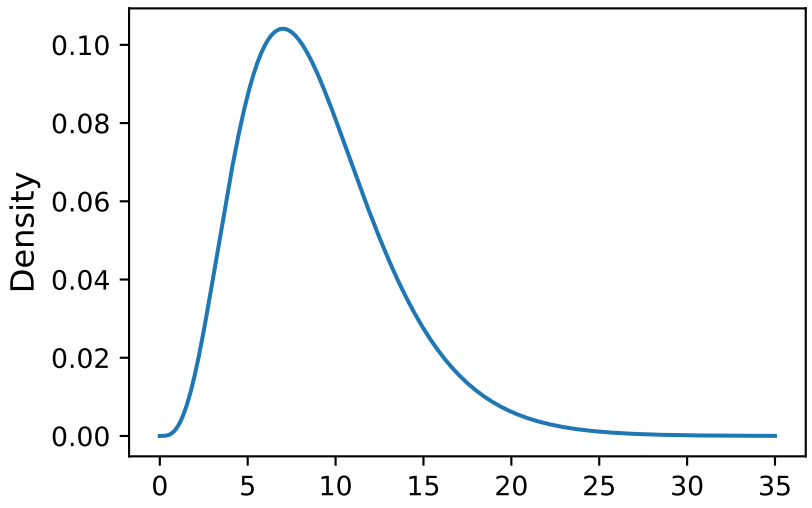
is a stochastic variable following the χ^2-distribution with $v = n - 1$ degrees of freedom.

The χ^2-distribution, as any other distribution, has a probability density function. It is a non-symmetric distribution on the positive axis. It is a distribution of squared normal random variables, for more details see Section 2.10.1. An example of a χ^2-distribution is given in the following:

|||| Example 3.18 The χ^2-distribution

The density of the χ^2-distribution with 9 degrees of freedom is:

```
# The chi-square-distribution with df=9 (the density)
x = np.linspace(0, 35, 1000)
plt.plot(x,stats.chi2.pdf(x,df=9))
plt.ylabel('Density',fontsize=12)
plt.tight_layout()
plt.show()
```



So, the χ^2-distributions are directly available in Python, via the SciPy package as seen for the other probability distributions presented in the distribution overview, see Appendix A.3.

Hence, we can easily work with χ^2-distributions in practice. As indicated there is a different χ^2-distribution for each n.

|||| Method 3.19 Confidence interval for the variance/standard deviation

A $100(1-\alpha)\%$ confidence interval for the variance σ^2 is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right], \tag{3-18}$$

where the quantiles come from a χ^2-distribution with $\nu = n-1$ degrees of freedom.

A $100(1-\alpha)\%$ confidence interval for the standard deviation σ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}\right]. \tag{3-19}$$

Note: The confidence intervals for the variance and standard deviations are generally non-symmetric as opposed to the t-based interval for the mean μ.

|||| Example 3.20 Tablet production

A random sample of $n = 20$ tablets is collected and from this the mean is estimated to $\bar{x} = 1.01$ and the variance to $s^2 = 0.07^2$. Let us find the 95%-confidence interval for the variance. To apply the method above we need the 0.025 and 0.975 quantiles of the χ^2-distribution with $\nu = 20 - 1 = 19$ degrees of freedom

$$\chi^2_{0.025} = 8.907,\ \chi^2_{0.975} = 32.85,$$

which we get from Python:

```
# Quantiles of the chi-square distribution:
print(stats.chi2.ppf([0.025,0.975],df=19))

[ 8.907 32.852]
```

Hence the confidence interval is

$$\left[\frac{19 \cdot 0.07^2}{32.85}, \frac{19 \cdot 0.07^2}{8.907}\right] \approx [0.00283,\ 0.0105],$$

and for the standard deviation the confidence interval is

$$\left[\sqrt{0.002834},\ \sqrt{0.01045}\right] \approx [0.053,\ 0.102]\,.$$

3.1.7 Hypothesis testing, evidence, significance and the p-value

|||| Example 3.21 Sleeping medicine

In a study the aim is to compare two kinds of sleeping medicine A and B. 10 test persons tried both kinds of medicine and the following 10 DIFFERENCES between the two medicine types were measured (in hours):

Person	$x =$ Beffect - Aeffect
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

For Person 1, Medicine B provided 1.2 sleep hours more than Medicine A, etc.

Our aim is to use these data to investigate if the two treatments are different in their effect on length of sleep. We therefore let μ represent the mean difference in sleep length. In particular we will consider the so-called null hypothesis

$$H_0:\ \mu = 0,$$

which states that there is *no difference* in sleep length between the A and B Medicines.

If the observed sample turns out to be not very likely under this null hypothesis, we conclude that the null hypothesis is unlikely to be true.

First we compute the sample mean

$$\hat{\mu} = \bar{x}_1 = 1.67.$$

As of now, we don't know if this number is particularly small or large. If the true mean difference is zero, would it be unlikely to observe a mean difference this large?

Could it be due to just random variation? To answer this question we compute the probability of observing a sample mean that is 1.67 or further from 0 – in the case that the true mean difference is in fact zero. This probability is called a p-value. If the p-value is small (say less than 0.05), we conclude that the null hypothesis isn't true. If the p-value is not small (say larger than 0.05), we conclude that we haven't obtained sufficient evidence to falsify the null hypothesis.

After some computations that you will learn to perform later in this section, we obtain a p-value

$$p\text{-value} \approx 0.00117,$$

which indicates quite strong evidence against the null hypothesis. As a matter of fact, the probability of observing a mean difference as far from zero as 1.67 or further is only ≈ 0.001 (one out of thousand) and therefore very small.

We conclude that the null hypothesis is unlikely to be true as it is highly incompatible with the observed data. We say that *the observed mean* $\hat{\mu} = 1.67$ *is statistically significantly different from zero* (or simply *significant* implying that it is different from zero). Or that *there is a significant difference in treatment effects of B and A*, and we may conclude that Medicine B makes patients sleep significantly longer than Medicine A.

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Table 3.1: A way to interpret the evidence for a given p-value.

The p-value

Definition 3.22 The p-value

The p-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

Interpretations of a p-value:

1. The p-value measures evidence
2. The p-value measures extremeness/unusualness of the data under the null hypothesis ("under the null hypothesis" means "assuming the null hypothesis is true")

The p-value is used as a general measure of evidence against a null hypothesis: the smaller the p-value, the stronger the evidence against the null hypothesis H_0. A typical strength of evidence scale is given in Table 3.1.

As indicated, the definition and interpretations above are generic in the sense that they can be used for any kind of hypothesis testing in any kind of setup. In later sections and chapters of this material, we will indeed encounter many different such setups. For the specific setup in focus here, we can now give the key method:

Method 3.23 The one-sample t-test statistic and the p-value

For a (quantitative) one sample situation, the p-value is given by

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|), \tag{3-20}$$

where T follows a t-distribution with $(n-1)$ degrees of freedom.
The observed value of the test statistics to be computed is

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \tag{3-21}$$

where μ_0 is the value of μ under the null hypothesis

$$H_0 : \ \mu = \mu_0. \tag{3-22}$$

The t-test and the p-value will in some cases be used to formalize actual decision making and the risks related to it:

Definition 3.24 The hypothesis test

We say that we carry out a hypothesis test when we decide against a null hypothesis or not, using the data.

A null hypothesis is *rejected* if the p-value, calculated after the data has been observed, is less than some α, that is if the p-value $< \alpha$, where α is some pre-specified (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Remark 3.25

Often chosen significance levels α are 0.05, 0.01 or 0.001 with the former being the globally chosen default value.

Remark 3.26

A note of caution in the use of the word *accepted* is in place: this should NOT be interpreted as having proved anything: *accepting* a null hypothesis in statistics simply means that we could not prove it wrong! And the reason for this could just potentially be that we did not collect sufficient amount of data, and *acceptance* hence proofs nothing at its own right.

Example 3.27 Sleeping medicine

Continuing from Example 3.21, we now illustrate how to compute the p-value using Method 3.23.

```
# Enter sleep difference observations
x = np.array([1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4])
n = len(x)
# Compute the tobs - the observed test statistic
tobs = (x.mean() - 0)/(x.std(ddof=1) / np.sqrt(n))
print(tobs)

4.671645978656775


# Compute the p-value as a tail-probability in the t-distribution
pvalue = 2 * (1-stats.t.cdf(abs(tobs),df=n-1))
print(pvalue)

0.0011658764685527068
```

Naturally, a function in Python can do this for us (the results differ slightly due to numerical inaccuracies). This function can also be used to calculate confidence intervals:

```
stats.ttest_1samp(x,popmean=0).pvalue

np.float64(0.0011658764685528319)


stats.ttest_1samp(x,popmean=0).confidence_interval()

ConfidenceInterval(low=np.float64(0.8613337442036719), high=np.float64(2.47866625579632
```

The confidence interval and the p-value supplements each other, and often both the confidence interval and the p-value are reported. The confidence interval covers those values of the parameter that we accept given the data, while the p-value measures the extremeness of the data if the null hypothesis is true.

Example 3.28 Sleeping medicine

In the sleep medicine example the 95% confidence interval is

$$[0.86,\ 2.48]\,,$$

so based on the data these are the values for the mean sleep difference of Medicine B versus Medicine A that we accept can be true. Only if the data is so extreme (i.e. rarely occurring) that we would only observe it 5% of the time the confidence interval does not cover the true mean difference in sleep.

The p-value for the null hypothesis $\mu = 0$ was ≈ 0.001 providing strong evidence against the correctness of the null hypothesis.

If the null hypothesis was true, we would only observe this large a difference in sleep medicine effect levels in around one out of a thousand times. Consequently we conclude that the null hypothesis is unlikely to be true and *reject* it.

Statistical significance

The word *significance* can mean *importance* or *the extent to which something matters* in our everyday language. In statistics, however, it has a very particular meaning: if we say that an effect is significant, it means that the p-value is so low that the null hypothesis stating *no effect* has been *rejected* at some *significance level* α.

Definition 3.29 Significant effect

An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α. [a]

[a] Often, $\alpha = 0.05$ is adopted.

At this point an *effect* would amount to a μ-value different from μ_0. In other contexts we will see later, *effects* can be various features of interest to us.

Example 3.30 Statistical significance

Consider the following two situations:

1. A researcher decides on a significance level of $\alpha = 0.05$ and obtains p-value $= 0.023$. She therefore concludes that the effect is *statistically significant*
2. Another researcher also adopts a significance level of $\alpha = 0.05$, but obtains p-value $= 0.067$. He concludes that the effect was not statistically significant

From a binary decision point of view the two researchers couldn't disagree more. However, from a scientific and more continuous evidence quantification point of view there is not a dramatic difference between the findings of the two researchers.

In daily statistical and/or scientific jargon the word "statistically" will often be omitted, and when results then are communicated as *significant* further through media or other places, it gives the risk that the distinction between the two meanings gets lost. At first sight it may appear unimportant, but the big difference is the following: sometimes a statistically significant finding can be so small in real size that it is of no real importance. If data collection involves very big data sizes one may find statistically significant effects that for no practical situations matter much or anything at all.

The null hypothesis

The null hypothesis most often expresses the *status quo* or that "nothing is happening". This is what we have to believe before we perform any experiments and observe any data. This is what we have to accept in the absence of any evidence that the situation is otherwise. For example the null hypothesis in the sleep medicine examples states that the difference in sleep medicine effect level is unchanged by the treatment: this is what we have to accept until we obtain evidence otherwise. In this particular example the observed data and the statistical theory provided such evidence and we could conclude a significant effect.

The null hypothesis has to be *falsifiable*. This means that it should be possible to collect evidence against it.

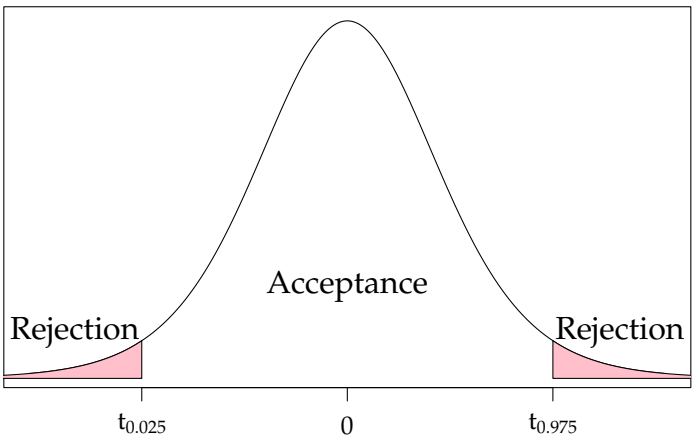


Figure 3.1: The 95% critical value. If t_{obs} falls in the pink area we would *reject*, otherwise we would *accept*

Confidence intervals, critical values and significance levels

A hypothesis test, that is, making the decision between *rejection* and *acceptance* of the null hypothesis, can also be carried out without actually finding the p-value. As an alternative one can use the so-called *critical values*, that is the values of the test-statistic which matches exactly the significance level, see Figure 3.1:

|||| Definition 3.31 The critical values

The $(1-\alpha)100\%$ critical values for the one-sample t-test are the $\alpha/2$- and $1-\alpha/2$-quantiles of the t-distribution with $n-1$ degrees of freedom

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}. \tag{3-23}$$

Method 3.32 The one-sample hypothesis test by the critical value

A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values

$$\text{If } |t_{\text{obs}}| > t_{1-\alpha/2} \text{ then } \textit{reject}, \tag{3-24}$$

otherwise *accept*.

The confidence interval covers the acceptable values of the parameter given the data:

Theorem 3.33 Confidence interval for μ

We consider a $(1-\alpha)\cdot 100\%$ confidence interval for μ

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}. \tag{3-25}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the hypothesis

$$H_0: \quad \mu = \mu_0. \tag{3-26}$$

Remark 3.34

The proof of this theorem is almost straightforward: a μ_0 inside the confidence interval will fulfil that

$$|\bar{x} - \mu_0| < t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \tag{3-27}$$

which is equivalent to

$$\frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} < t_{1-\alpha/2}, \tag{3-28}$$

and again to

$$|t_{\text{obs}}| < t_{1-\alpha/2}, \tag{3-29}$$

which then exactly states that μ_0 is accepted, since the t_{obs} is within the critical values.

The alternative hypothesis

Some times we may in addition to the null hypothesis, also explicitly state an *alternative hypothesis*. This completes the framework that allows us to control the rates at which we make correct and wrong conclusions in light of the alternative.

The alternative hypothesis is

$$H_1 : \mu \neq \mu_0. \tag{3-30}$$

This is sometimes called the two-sided (or non-directional) alternative hypothesis, because also one-sided (or directional) alternative hypothesis occur. However, the one-sided setup is not included in the book apart from a small discussion below.

Example 3.35 Sleeping medicine – Alternative hypothesis

Continuing from Example 3.21 we can now set up the null hypothesis and the alternative hypothesis together

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0. \end{aligned}$$

Which means that we have exactly the same setup just formalized by adding the alternative hypothesis. The conclusion is naturally exactly the same as in before.

A generic approach for tests of hypotheses is:

1. Formulate the hypotheses and choose the level of significance α (choose the "risk-level")
2. Calculate, using the data, the value of the test statistic
3. Calculate the p-value using the test statistic and the relevant sampling distribution, compare the p-value and the significance level α, and finally make a conclusion
   or
   Compare the value of the test statistic with the relevant critical value(s) and make a conclusion

Combining this generic hypothesis test approach with the specific method boxes of the previous section, we can now below give a method box for the one-sample t-test. This is hence a collection of what was presented in the previous section:

Method 3.36 The level α one-sample t-test

1. Compute t_{obs} using Equation (3-21)

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

2. Compute the evidence against the *null hypothesis*

$$H_0: \quad \mu = \mu_0, \tag{3-31}$$

vs. the *alternative hypothesis*

$$H_1: \quad \mu \neq \mu_0, \tag{3-32}$$

by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|), \tag{3-33}$$

where the t-distribution with $n-1$ degrees of freedom is used

3. If the p-value $< \alpha$, we reject H_0, otherwise we accept H_0,

or

The rejection/acceptance conclusion can equivalently be based on the critical value(s) $\pm t_{1-\alpha/2}$:
if $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

The so-called one-sided (or directional) hypothesis setup, where the alternative hypothesis is either "less than" or "greater than", is opposed to the previous presented two-sided (or non-directional) setup, with a "different from" alternative hypothesis. In most situations the two-sided should be applied, since when setting up a null hypothesis with no knowledge about in which direction the outcome will be, then the notion of "extreme" is naturally in both directions. However, in some situations the one-sided setup makes sense to use. As for example in pharmacology where concentrations of drugs are studied and in some situations it is known that the concentration can only decrease from one time point of measurement to another (after the peak concentration). In such case a "less than" is the only meaningful alternative hypothesis – one can say that nature really has made the decision for us in that: either the concentration has not changed (the null hypothesis) or it has dropped (the alternative hypothesis). In other cases, e.g. more from the business and/or judicial perspective, one-sided

hypothesis testing come up when for example a claim about the performance of some product is tested.

The one-sided "less than" hypothesis setup is: compute the evidence against the *null hypothesis* vs. the *one-sided alternative hypothesis*

$$H_0: \quad \mu \geq \mu_0 \tag{3-34}$$
$$H_1: \quad \mu < \mu_0, \tag{3-35}$$

by the

$$p\text{-value} = P(T < t_{\text{obs}}). \tag{3-36}$$

and equivalently for the "greater than" setup

$$H_0: \quad \mu \leq \mu_0 \tag{3-37}$$
$$H_1: \quad \mu > \mu_0, \tag{3-38}$$

by the

$$p\text{-value} = P(T > t_{\text{obs}}). \tag{3-39}$$

In both cases: if p-value $< \alpha$: We reject H_0, otherwise we accept H_0.

Note that there are no one-sided hypothesis testing involved in the exercises.

Errors in hypothesis testing

When testing statistical hypotheses, two kind of errors can occur:

Type I: Rejection of H_0 when H_0 is true

Type II: Non-rejection (acceptance) of H_0 when H_1 is true

|||| Example 3.37 Ambulance times

An ambulance company claims that on average it takes 20 minutes from a telephone call to their switchboard until an ambulance reaches the location.

We might have some measurements (in minutes): 21.1, 22.3, 19.6, 24.2, ...

If our goal is to show that on average it takes longer than 20 minutes, the null- and the alternative hypotheses are

$$H_0: \quad \mu = 20,$$
$$H_1: \quad \mu \neq 20.$$

What kind of errors can occur?

Type I: Reject H_0 when H_0 is true, that is we mistakenly conclude that it takes longer (or shorter) than 20 minutes for the ambulance to be on location

Type II: Not reject H_0 when H_1 is true, that is we mistakenly conclude that it takes 20 minutes for the ambulance to be on location

|||| Example 3.38 Court of law analogy

A man is standing in a court of law accused of criminal activity.

The null- and the alternative hypotheses are

$$\begin{aligned} H_0: &\quad \text{The man is not guilty,} \\ H_1: &\quad \text{The man is guilty.} \end{aligned}$$

We consider a man not guilty until evidence beyond any doubt proves him guilty. This would correspond to an α of basically zero.

Clearly, we would prefer not to do any kinds of errors, however it is a fact of life that we cannot avoid to do so: if we would want to never do a Type I error, then we would never reject the null hypothesis, which means that we would e.g. never conclude that one medical treatment is better than another, and thus, that we would (more) often do a Type II error, since we would never detect when there was a significance effect.

For the same investment (sample size n), we will increase the risk of a Type II error by enforcing a lower risk of a Type I error. Only by increasing n we can lower both of them, but to get both of them very low can be extremely expensive and thus such decisions often involve economical considerations.

The statistical hypothesis testing framework is a way to formalize the handling of the risk of the errors we may make and in this way make decisions in an enlightened way knowing what the risks are. To that end we define the two possible risks as

$$\begin{aligned} P(\text{"Type I error"}) &= \alpha, \\ P(\text{"Type II error"}) &= \beta. \end{aligned} \tag{3-40}$$

This notation is globally in statistical literature. The name choice for the Type I error is in line with the use of α for the *significance level*, as:

Theorem 3.39 Significance level and Type I error

The significance level α in hypothesis testing is the overall Type I risk

$$P(\text{"Type I error"}) = P(\text{"Rejection of } H_0 \text{ when } H_0 \text{ is true"}) = \alpha. \quad (3\text{-}41)$$

So controlling the Type I risk is what is most commonly apparent in the use of statistics. Most published results are results that became significant, that is, the p-value was smaller than α, and hence the relevant risk to consider is the Type I risk.

Controlling/dealing with the Type II risk, that is: how to conclude on an experiment/study in which the null hypothesis was not rejected (i.e. no significant effect was found) is not so easy, and may lead to heavy discussions if the non-findings even get to the public. To which extent is a non-finding an evidence of the null hypothesis being true? Well, in the outset the following very important saying makes the point:

Remark 3.40

Absence of evidence is NOT evidence of absence!

Or differently put:
Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!

The main thing to consider here is that non-findings (non-significant results) may be due to large variances and small sample sizes, so sometimes a non-finding is indeed just that we know nothing. In other cases, if the sample sizes were high, a non-finding may actually, if not proving an effect equal to zero, which is not really possible, then at least indicate with some confidence that the possible effect is small or even very small. The confidence interval is a more clever method to use here, since the confidence interval will show the precision of what we know, whether it includes the zero effect or not.

In Section 3.3 we will use a joint consideration of both error types to formalize the planning of suitably sized studies/experiments.

3.1.8 Assumptions and how to check them

The t-tests that have been presented above are based on some assumptions about the sampling and the population. In Theorem 3.3 the formulations are that the random variables $X_1, \ldots, X_n$ are independent and identically normally distributed: $X_i \sim N(\mu, \sigma^2)$. In this statement there are two assumptions:

- Independent observations
- Normal distribution

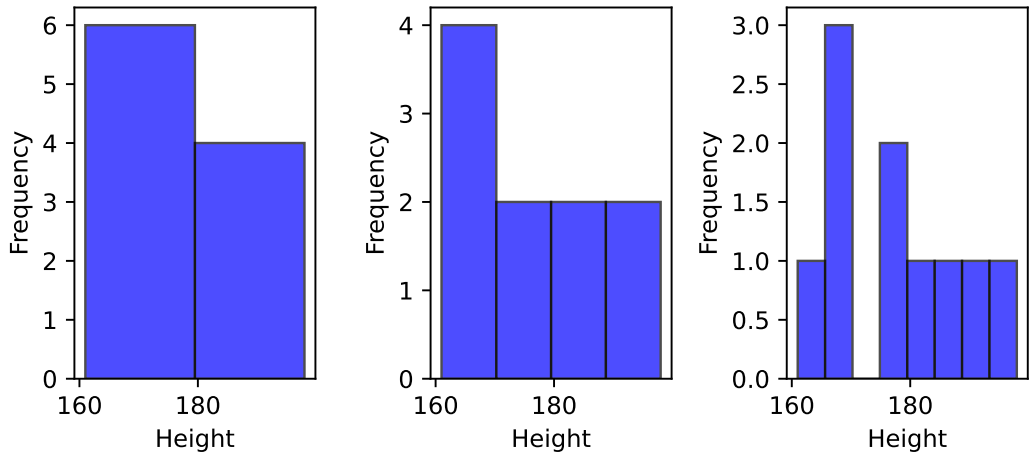
The assumption about independent observations can be difficult to check. It means that each observation must bring a unique new amount of information to the study. Independence will be violated if some measurements are not on randomly selected units and share some feature – returning to the student height example: we do not want to include twins or families in general. Having a sample of $n = 20$ heights, where 15 of them stem from a meeting with a large family group would not be 20 independent observations. The independence assumption is mainly checked by having information about the sampling procedure.

The assumption about normality can be checked graphically using the actual sample at hand.

|||| Example 3.41 Student heights

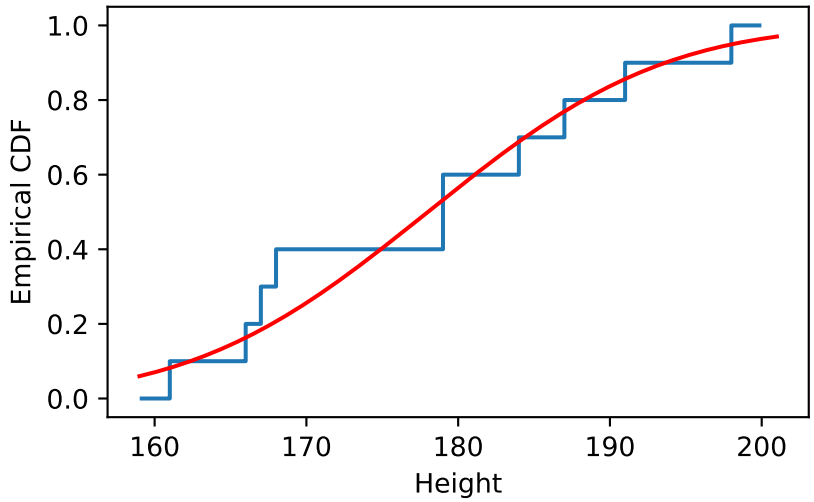
We will return to the height of the ten students from example 3.1. If we want to check whether the sample of heights could come from a normal distribution then we could plot a histogram and look for a symmetric bell-shape:

```
# The height sample
x = np.array([168,161,167,179,184,166,198,187,191,179])

# Using histograms
fig, (ax1, ax2, ax3) = plt.subplots(1, 3)
ax1.hist(x, bins=2, edgecolor='black', color='blue', alpha=0.7)
ax1.set(xlabel='Height', ylabel='Frequency')
ax2.hist(x, bins=4, edgecolor='black', color='blue', alpha=0.7)
ax2.set(xlabel='Height', ylabel='Frequency')
ax3.hist(x, bins=8, edgecolor='black', color='blue', alpha=0.7)
ax3.set(xlabel='Height', ylabel='Frequency')
plt.tight_layout()
plt.show()
```



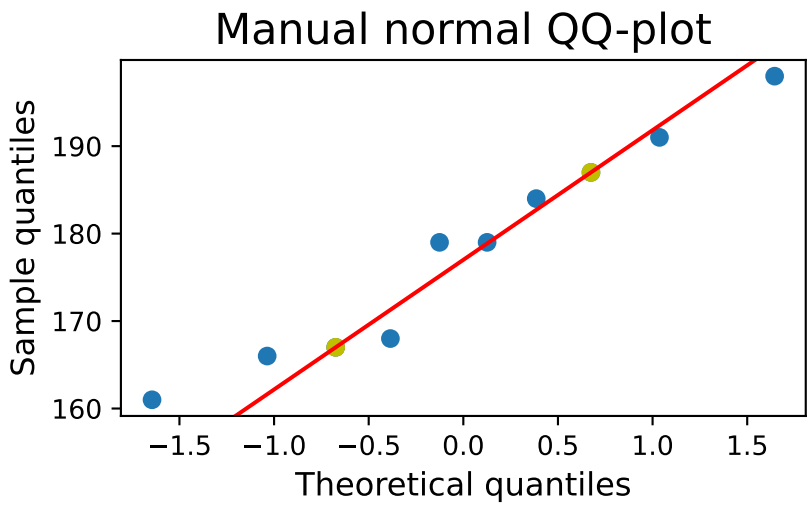
However, as we can see the histograms change shape depending on the number of breaks. Instead of using histograms, one can plot empirical cumulative distribution (see 1.6.2) and compare it with the best fitting normal distribution, in this case $N(\hat{\mu} = 178, \hat{\sigma}^2 = 12.21^2)$:

```
# Plot the empirical cdf
ecdf = stats.ecdf(x)
ax = plt.subplot()
ecdf.cdf.plot(ax)
ax.set(xlabel='Height', ylabel='Empirical CDF')
# Plot a normal cdf
y = np.linspace(159,201, 1000)
plt.plot(y,stats.norm.cdf(y,loc=x.mean(),
scale=x.std(ddof=1)),color="red")
plt.tight_layout()
plt.show()
```



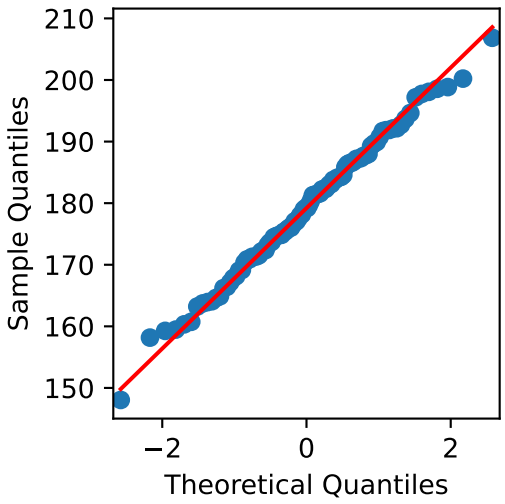
In the accumulated distribution plot it is easier to see how close the distributions are – compared to in the density histogram plot. However, we will go one step further and do the q-q plot: The observations (sorted from smallest to largest) are plotted against the expected quantiles – from the same normal distribution as above. If the observations are normally distributed then the observed are close to the expected and this plot is close to a straight line. In Python we can generate this plot by the following:

```
# A manual normal QQ-plot (normal Quantile-Quantile-plot)
# Calculate manual empirical CDF-values (p) for the observations in the
sample
n = len(x)
p = np.linspace(0.5/n,1-0.5/n,n)
# Plot the theoretical normal quantiles associated with p (x-axis)
against
# the observations. Note that the observations function as the sample
# quantiles. Thus, we compare the theoretical with the sample
quantiles.
plt.scatter(stats.norm.ppf(p),np.sort(x))
# Plot straight line thorugh (TQ1,SQ1) and (TQ3,SQ3).
# T: Theoretical - S: Sample
TQ1 = stats.norm.ppf(0.25)
TQ3 = stats.norm.ppf(0.75)
SQ1 = np.quantile(x,0.25,method='averaged_inverted_cdf')
SQ3 = np.quantile(x,0.75,method='averaged_inverted_cdf')
plt.plot((TQ1,TQ3),(SQ1,SQ3),'yo')
plt.axline((TQ1,SQ1),(TQ3,SQ3),color="red")
# Notice that this not generate the same plot as the standard
functions
plt.xlabel('Theoretical quantiles',fontsize=12)
plt.ylabel('Sample quantiles',fontsize=12)
plt.title('Manual normal QQ-plot',fontsize=16)
plt.tight_layout()
plt.show()
```



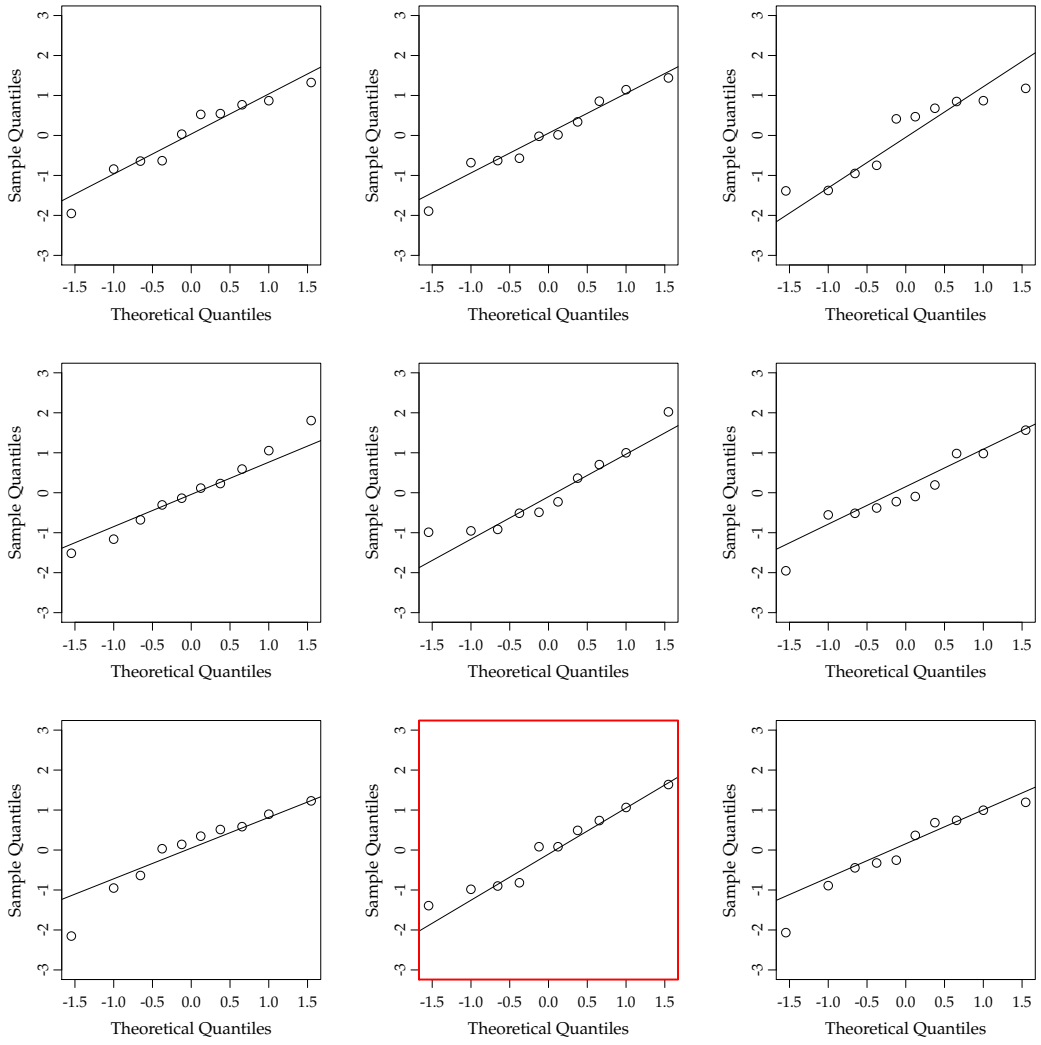
In the ideal normal case, the observations vs. the expected quantiles in the best possible normal distribution will be on a straight line, here plotted with the `line` argument of the `qqplot`-function from the `statsmodels` package:

```
# Simulate 100 observations
np.random.seed(31415)
simx = stats.norm.rvs(loc=x.mean(), scale=x.std(ddof=1), size=100)
# Do the normal QQ-plot and QQ-line with standard functions
sm.qqplot(simx,line="q",a=1/2)
plt.tight_layout()
plt.show()
```



Note that the inbuilt functions do exactly the same as the Python code generating the first q-q plot as described in Method 3.42.

In this example the points are close to a straight line and we can assume that the normal distribution holds. It can, however, be difficult to decide whether the plot is close enough to a straight line, so we write a function that generates one q-q plot of the observations and eight q-q plots with data simulated from a standard normal distribution. It is then possible to visually compare the plot based on the observed data to the simulated data and see whether the distribution of the observations is "worse" than they should be.



When we look at the nine plots then the original data are plotted in the frame with the red border. Comparing the observed data to the simulated data the straight line for the observed data is no worse than some of the simulated data, where the normality assumption is known to hold. So we conclude here that we apparently have no problem in assuming the normal distribution for these data.

Method 3.42 The Normal q-q plot

The ordered observations $x_{(1)}, \ldots, x_{(n)}$, called the sample quantiles, are plotted versus a set of expected normal quantiles $z_{p_1}, \ldots, z_{p_n}$. If the points are not systematically deviating from a line, we accept the normal distribution assumption. The evaluation of this can be based on some simulations of a sample of the same size.

The usual definition of $p_1, \ldots, p_n$ to be used for finding the expected normal quantiles is

$$p_i = \frac{i - 0.5}{n},\ i = 1, \ldots, n. \tag{3-42}$$

Hence, simply the equally distanced points between $0.5/n$ and $1 - 0.5/n$. This formula is suitable for samples with $n > 10$ and can be used in Python by specifying `qqplot(...,a=1/2)`. For samples with $n \leq 10$, the formula

$$p_i = \frac{i - 3/8}{n + 1/4},\ i = 1, \ldots, n, \tag{3-43}$$

which can be used in Python by specifying `qqplot(...,a=3/8)`, is preferred.

Example 3.43 Student heights

An example of how the expected normal quantile is calculated in Python can be seen if we take the second smallest height 166. There are 2 observations ≤ 166, so $166 = x_{(2)}$ can be said to be the observed $\frac{2-3/8}{10.25} = 0.1585$ quantile (where we use the formula for $n \leq 10$). The 0.1585 quantile in the normal distribution is `stats.norm.ppf(0.1585,loc=0,scale=1)` $= -1.00$ and the point $(-1.00,\ 166)$ can be seen on the q-q plot above.

3.1.9 Transformation towards normality

In the above we looked at methods to check for normality. When the data are not normally distributed it is often possible to choose a transformation of the sample, which improves the normality.

When the sample is positive with a long tail or a few large observations then the most common choice is to apply a logarithmic transformation, $\log(x)$. The log-

transformation will make the large values smaller and also spread the observations on both positive and negative values. Even though the log-transformation is the most common there are also other possibilities such as $\sqrt{x}$ or $\frac{1}{x}$ for making large values smaller, or x^2 and x^3 for making large values larger.

When we have transformed the sample we can use all the statistical analyse we want. It is important to remember that we are now working on the transformed scale (e.g. the mean and its confidence interval is calculated for $\log(x)$) and perhaps it will be necessary to back-transform to the original scale.

|||| Example 3.44 Radon in houses

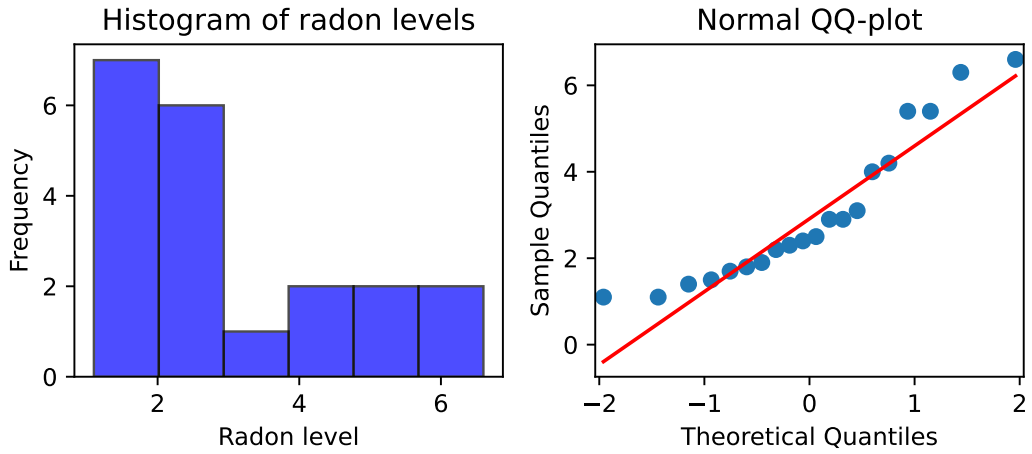
In an American study the radon level was measured in a number of houses. The Environmental Protection Agency's recommended action level is ≥ 4 pCi/L. Here we have the results for 20 of the houses (in pCi/L):

House	1	2	3	4	5	6	7	8	9	10
Radon level	2.4	4.2	1.8	2.5	5.4	2.2	4.0	1.1	1.5	5.4
House	11	12	13	14	15	16	17	18	19	20
Radon level	6.3	1.9	1.7	1.1	6.6	3.1	2.3	1.4	2.9	2.9

The sample mean, median and std. deviance is: $\bar{x} = 3.04$, $Q_2 = 2.45$ and $s_x = 1.72$.

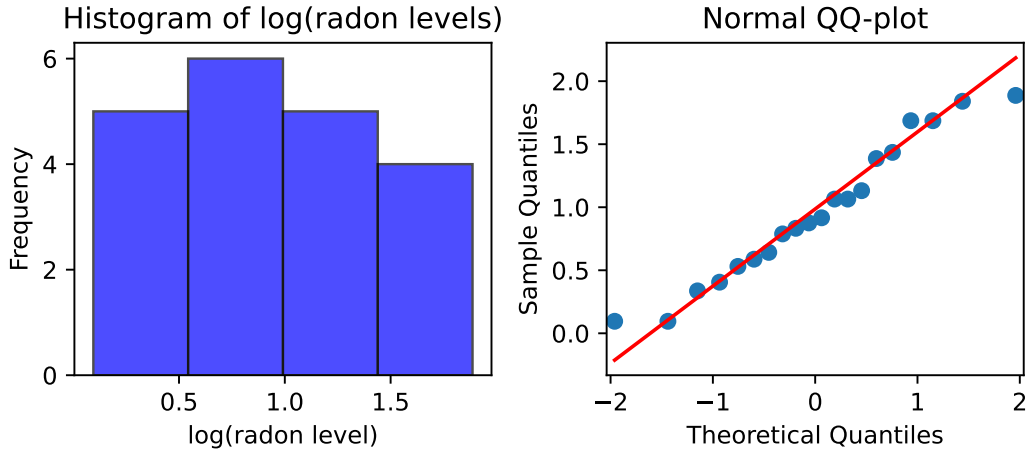
We would like to see whether these observed radon levels could be thought of as coming from a normal distribution. To do this we will plot the data:

```
# Reading in the sample
radon = np.array([2.4, 4.2, 1.8, 2.5, 5.4, 2.2, 4.0, 1.1, 1.5, 5.4,
                  6.3, 1.9, 1.7, 1.1, 6.6, 3.1, 2.3, 1.4, 2.9, 2.9])

# A histrogram and normal QQ-plot
fig, (ax1, ax2) = plt.subplots(1, 2)
ax1.hist(radon, bins=6, edgecolor='black', color='blue', alpha=0.7)
ax1.set(title="Histogram of radon levels",xlabel="Radon
level",ylabel="Frequency")
sm.qqplot(radon,line="q",a=1/2,ax=ax2)
ax2.set(title="Normal QQ-plot")
plt.tight_layout()
plt.show()
```



From both plots we see that the data are positive and right skewed with a few large observations. Therefore a log-transformation is applied:

```
# Transform using the natural logarithm
logRadon = np.log(radon)

# A histrogram and normal QQ-plot
fig, (ax1, ax2) = plt.subplots(1, 2)
ax1.hist(logRadon, bins=4, edgecolor='black', color='blue', alpha=0.7)
ax1.set(title="Histogram of log(radon levels)",xlabel="log(radon
level)",ylabel="Frequency")
sm.qqplot(logRadon,line="q",a=1/2,ax=ax2)
ax2.set(title="Normal QQ-plot")
plt.tight_layout()
plt.show()
```



As we had expected the log-transformed data seem to be closer to a normal distri-

bution.

We can now calculate the mean and 95% confidence interval for the log-transformed data. However, we are perhaps not interested in the mean of the log-radon levels, then we have to back-transform the estimated mean and confidence interval using $\exp(x)$. When we take the exponential of the estimated mean, then this is no longer a mean but a median on the original pCi/L scale. This gives a good interpretation, as medians are useful when the distributions are not symmetric.

```
# A confidence interval and t-test
n = len(logRadon)
test = stats.ttest_1samp(logRadon,popmean=0)
print(test.statistic,test.pvalue,test.df)

7.793651876947492 2.46529449526264e-07 19


CI = stats.ttest_1samp(logRadon,popmean=0).confidence_interval(0.95)
print(CI)

ConfidenceInterval(low=np.float64(0.7054264972507451), high=np.float64(1.22343071479501


# Alternatively, the CI can be obtained as
CI = stats.t.interval(0.95,df=n-1,loc=logRadon.mean(),
scale=logRadon.std(ddof=1)/np.sqrt(n))
print(CI)

(np.float64(0.7054264972507451), np.float64(1.2234307147950183))


# Back transform to original scale, now we get the median!
# This is a special case: In the lognormal distribution,
# the median coincides with the geometric mean value.
print(np.exp(logRadon.mean()))

2.623288297019726


# And the confidence interval on the original scale
print(np.exp(CI))

[2.025 3.399]
```

From the Python code we see that the mean log-radon level is 0.96 (95% CI: 0.71 to 1.22). On the original scale the estimated median radon level is 2.6 pCi/L (95% CI: 2.0 to 3.4).

Theorem 3.45 Transformations and quantiles

In general, the data transformations discussed in this section will preserve the quantiles of the data. Or more precisely, if f is a data transformation function (an increasing function), then

$$\text{The } p\text{th quantile of } f(Y) = f(\text{The } p\text{th quantile of } Y). \tag{3-44}$$

The consequence of this theorem is that confidence limits on one scale transform easily to confidence limits on another scale even though the transforming function is non-linear.

3.2 Learning from two-sample quantitative data

In this section the setup, where we can learn about the difference between the means from two populations, is presented. This is very often a setup encountered in most fields of science and engineering: compare the quality of two products, compare the performance of two groups, compare a new drug to a placebo and so on. One could say, that it should be called a two-population setup, since it is really two populations (or groups) which are compared by taking a sample from each, however it is called a two-sample setup (probably it sounds better to say).

First, the two-sample setup is introduced with an example and then methods for confidence intervals and tests are presented.

|||| Example 3.46 Nutrition study

In a nutrition study the aim is to investigate if there is a difference in the energy usage for two different types of (moderately physically demanding) work. In the study, the energy usage of 9 nurses from hospital A and 9 (other) nurses from hospital B have been measured. The measurements are given in the following table in mega Joule (MJ):

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

Our aim is to assess the difference in energy usage between the two groups of nurses. If μ_A and μ_B are the mean energy expenditures for nurses from hospital A and B, then the estimates are just the sample means

$$\hat{\mu}_A = \bar{x}_A = 8.293,$$
$$\hat{\mu}_B = \bar{x}_B = 10.298.$$

To assess the difference in means, $\delta = \mu_B - \mu_A$, we could consider the confidence interval for $\delta = \mu_B - \mu_A$. Clearly, the estimate for the difference is the difference of the sample means, $\hat{\delta} = \hat{\mu}_B - \hat{\mu}_A = 2.005$.

The 95% confidence interval is

$$2.005 \pm 1.412 = [0.59,\ 3.42],$$

which spans the mean differences in energy expenditure that we find acceptable based on the data. Thus we do not accept that the mean difference could be zero.

The interval width, given by 1.41, as we will learn below, comes from a simple computation using the two sample standard deviations, the two sample sizes and a t-quantile.

We can also compute a p-value to measure the evidence against the null hypothesis that the mean energy expenditures are the same. Thus we consider the following null hypothesis

$$H_0 :\ \delta = 0.$$

Since the 95% confidence interval does not cover zero, we already know that the p-value for this significance test will be less than 0.05. In fact it turns out that the p-value for this significance test is 0.0083 indicating strong evidence against the null hypothesis that the mean energy expenditures are the same for the two nurse groups. We therefore have strong evidence that the mean energy expenditure of nurses from hospital B is higher than that of nurses from hospital A.

This section describes how to compute the confidence intervals and p-values in such two-sample setups.

3.2.1 Comparing two independent means - confidence Interval

We assume now that we have a sample $x_1, \ldots, x_n$ taken at random from one population with mean μ_1 and variance σ_1^2 and another sample $y_1, \ldots, y_n$ taken at random from another population with mean μ_2 and variance σ_2^2.

Method 3.47 The two-sample confidence interval for $\mu_1 - \mu_2$

For two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{3-45}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile from the t-distribution with ν degrees of freedom given from Equation (3-50)

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}. \tag{3-46}$$

Note how the t-quantile used for the confidence interval is exactly what we called the critical value above.

Example 3.48 Nutrition study

Let us find the 95% confidence interval for $\mu_B - \mu_A$. Since the relevant t-quantile is, using $\nu = 15.99$,

$$t_{0.975} = 2.120,$$

the confidence interval becomes

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}},$$

which then gives the result as also seen above

$$[0.59,\ 3.42].$$

3.2.2 Comparing two independent means - hypothesis test

We describe the setup as having a random sample from each of two different populations, each described by a mean and a variance:

- Population 1: has mean μ_1, and variance σ_1^2

- Population 2: has mean μ_2, and variance σ_2^2

The interest lies in the comparisons of the means.

Method 3.49 The (Welch) two-sample t-test statistic

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\begin{aligned} \delta &= \mu_2 - \mu_1, \\ H_0 &: \ \delta = \delta_0, \end{aligned} \tag{3-47}$$

the (Welch) two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \tag{3-48}$$

Theorem 3.50 The distribution of the (Welch) two-sample statistic

The (Welch) two-sample statistic seen as a random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \tag{3-49}$$

approximately, under the null hypothesis, follows a t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}, \tag{3-50}$$

if the two population distributions are normal or if the two sample sizes are large enough.

We can now, based on this, express the full hypothesis testing procedures for the two-sample setting:

Method 3.51 The level α two-sample t-test

1. Compute the test statistic using Equation (3-48) and ν from Equation (3-50)

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \text{ and } \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

2. Compute the evidence against the *null hypothesis*[a]

$$H_0: \quad \mu_1 - \mu_2 = \delta_0,$$

vs. the *alternative hypothesis*

$$H_1: \quad \mu_1 - \mu_2 \neq \delta_0,$$

by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|),$$

where the t-distribution with ν degrees of freedom is used

3. If p-value $< \alpha$: we reject H_0, otherwise we accept H_0,

or

The rejection/acceptance conclusion can equivalently be based on the critical value(s) $\pm t_{1-\alpha/2}$:
if $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

[a] We are often interested in the test where $\delta_0 = 0$

An assumption that often is applied in statistical analyses of various kinds is that of the underlying variability being of the same size in different groups or at different conditions. The assumption is rarely crucial for actually carrying out some good statistics, but it may indeed make the theoretical justification for what is done more straightforward, and the actual computational procedures also may become more easily expressed. We will see in later chapters how this comes in play. Actually, the methods presented above do not make this assumption, which is nice. The fewer assumptions needed the better, obviously. Assumptions are problematic in the sense, that they may be questioned for particular applications of the methods.

However, below we will present a version of the two-sample t-test statistic, that actually is adapted to such an assumption, namely assuming that the two population variances are the same: $\sigma_1^2 = \sigma_2^2$. We present it here not because we really need it, we will use the above in all situations. But the version below will appear and be used many places and it also bears some nice relations to later multi-group analysis (Analysis of Variance (ANOVA)) that we will get to in Chapter 8.

If we believe in the equal variance assumption it is natural to compute a single joint – called the *pooled* – estimate of the variance based on the two individual variances:

|||| Method 3.52 The pooled two-sample estimate of variance

Under the assumption that $\sigma_1^2 = \sigma_2^2$ the *pooled* estimate of variance is the weighted average of the two sample variances

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{3-51}$$

Note that when there is the same number of observations in the two groups, $n_1 = n_2$, the pooled variance estimate is simply the average of the two sample variances. Based on this the so-called pooled two-sample t-test statistic can be given:

|||| Method 3.53 The pooled two-sample t-test statistic

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\begin{aligned} \delta &= \mu_1 - \mu_2, \\ H_0 &: \ \delta = \delta_0. \end{aligned} \tag{3-52}$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}. \tag{3-53}$$

And the following theorem would form the basis for hypothesis test procedures based on the pooled version:

|||| Theorem 3.54 The distribution of the pooled two-sample t-test statistic

The pooled two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}. \tag{3-54}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

A little consideration will show why choosing the Welch-version as the approach to always use makes good sense: First of all if $s_1^2 = s_2^2$ the Welch and the Pooled test statistics are the same. Only when the two variances become really different the two test-statistics may differ in any important way, and if this is the case, we would not tend to favour the pooled version, since the assumption of equal variances appears questionable then.

Only for cases with a small sample sizes in at least one of the two groups the pooled approach may provide slightly higher power if you believe in the equal variance assumption. And for these cases the Welch approach is then a somewhat cautious approach.

|||| Example 3.55 Nutrition study

Let us consider the nurses example again, and test the null hypothesis expressing that the two groups have equal means

$$H_0 : \delta = \mu_A - \mu_B = 0,$$

versus the alternative

$$H_0 : \delta = \mu_A - \mu_B \neq 0,$$

using the most commonly used significance level, $\alpha = 0.05$. We follow the steps of Method 3.51: we should first compute the test-statistic t_{obs} and the degrees of freedom ν. These both come from the basic computations on the data:

```
# Load the two samples
xA = np.array([7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9])
xB = np.array([9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68,
9.19])

# Summary statistics
print(xA.mean(),xB.mean())

8.293333333333335 10.297777777777776

print(xA.var(ddof=1),xB.var(ddof=1))

2.0394000000000005 1.954044444444444

print(len(xA),len(xB))

9 9
```

So

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01,$$

and

$$\nu = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99.$$

Or the same done in Python by ”manual” expression:

```
# Keep the summary statistics
ms = np.array([xA.mean(),xB.mean()])
vs = np.array([xA.var(ddof=1),xB.var(ddof=1)])
ns = np.array([len(xA),len(xB)])

# The observed statistic
t_obs = (ms[1]-ms[0])/np.sqrt(vs[0]/ns[0]+vs[1]/ns[1])

# The degrees of freedom
nu = ((vs[0]/ns[0]+vs[1]/ns[1])**2)/((vs[0]/ns[0])**2/(ns[0]-1)
+(vs[1]/ns[1])**2/(ns[1]-1))

# Print the result
print(t_obs)

3.009133495521211

print(nu)

15.992693827602634
```

Next step is then to find the p-value

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083,$$

where we use Python to find the probability $P(T > 3.01)$ based on a t-distribution with $\nu = 15.99$ degrees of freedom:

```
# The probability of observing a value greater that t_obs
print(1 - stats.t.cdf(t_obs,df=nu))

0.004161369978658014
```

To complete the hypothesis test, we compare the p-value with the given α-level, in this case $\alpha = 0.05$, and conclude:

> Since the p-value is less than α we *reject* the null hypothesis, and we have sufficient evidence for concluding: the two nurse groups have on average different energy usage work levels. We have shown this *effect* to be *statistically significant*.

In spite of a pre-defined α-level (whoever gave us that), it is always valuable to consider at what other α-levels the hypothesis would be rejected/accepted. Or in different words, interpret the size of the p-value using Table 3.1 and we thus sharpen the statement a little:

Since the p-value in this case is between 0.001 and 0.01 conclude: there is a *strong evidence* against equality of the two population energy usage means and it is found that *the mean is significantly higher* on Hospital B compared to Hospital A.

The last part, that the mean is higher on Hospital B, can be concluded because it is rejected that they are equal and $\bar{x}_B > \bar{x}_A$ and we can thus add this to the conclusion.

Finally, the t-test computations are actually directly provided by the `ttest_ind`-function from the SciPy package using the two data input vectors

```
# Use the automatic function for a t-test
test = stats.ttest_ind(xB,xA,equal_var=False)
tobs = test.statistic
pvalue = test.pvalue
df = test.df
print(tobs,pvalue,df)

3.009133495521211 0.00832273995731614 15.992693827602634


stats.ttest_ind(xB,xA,equal_var=False).confidence_interval(0.95)

ConfidenceInterval(low=np.float64(0.5922803841924627), high=np.float64(3.41660850469642
```

Note, how the default choices of the Python-function compare to our exposition:

- Default test version: the pooled test (assuming equal variances)
- Default α-level: 0.05
- Default "direction version": the two-sided (or non-directional) alternative hypothesis (see Section 3.1.7 about other alternative hypotheses)

Actually, the final rejection/acceptance conclusion based on the default (or chosen) α-level is not given by Python.

In the `ttest_ind` results the α-level is used for the given confidence interval for the mean difference of the two populations, to be interpreted as: we accept that the true difference in mean energy levels between the two nurse groups is somewhere between 0.6 and 3.4.

Remark 3.56

Often "degrees of freedom" are integer values, but in fact t-distributions with non-integer valued degrees of freedom are also well defined. The $\nu = 15.99$ t-distribution (think of the density function) is a distribution in between the $\nu = 15$ and the $\nu = 16$ t-distributions. Clearly it will indeed be very close to the $\nu = 16$ one.

We did not in the example above use Step 4. of Method 3.51, which can be called the critical value approach. In fact this approach is directly linked to the confidence interval in the sense that one could make a rapid conclusion regarding rejection or not by looking at the confidence interval and checking whether the hypothesized value is in the interval or not. This would correspond to using the critical value approach.

Example 3.57 Nutrition study

In the nutrition example above, we can see that 0 is not in the confidence interval so we would reject the null hypothesis. Let us formally use Step 4 of Method 3.51 to see how this is exactly the same: the idea is that one can even before the experiment is carried out find the critical value(s), in this case:

$$\text{The 5\% critical values} = \pm t_{0.975} = \pm 2.120,$$

where the quantile is found from the t-distribution with $\nu = 15.99$ degrees of freedom:

```
# The critical value for the test
print(stats.t.ppf(0.975,df=nu))

2.119984011855833
```

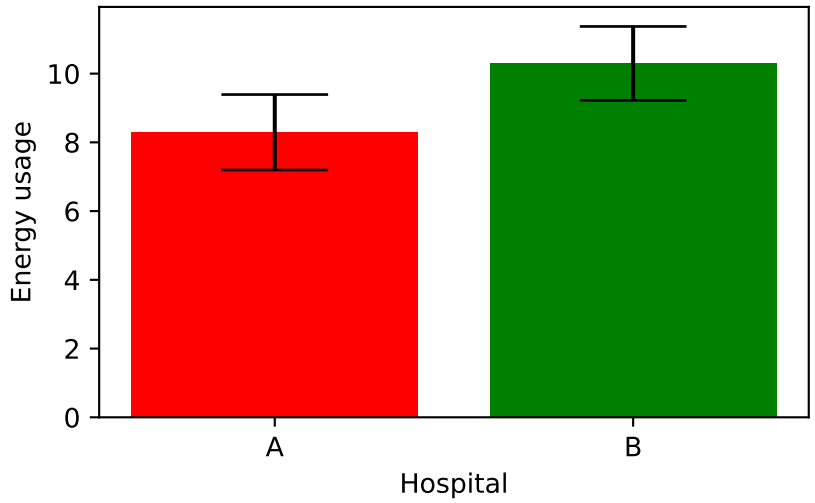
Now we conclude that since the observed t-statistic $t_{\text{obs}} = 3.01$ is beyond the critical values (either larger than 2.120 or smaller than -2.120) the null hypothesis is rejected, and further since it was higher, that $\mu_A - \mu_B > 0$ hence $\mu_B > \mu_A$.

Example 3.58 Overlapping confidence intervals?

A commonly encountered way to visualize the results of a two-sample comparison is to use a bar plot of the means together with some measure of uncertainty, either

simply the standard errors of the means or the 95% confidence intervals within each group:

```
# The confidence intervals
CIA = stats.ttest_1samp(xA,popmean=0).confidence_interval(0.95)
CIB = stats.ttest_1samp(xB,popmean=0).confidence_interval(0.95)

# Barplots with error bars
fig, ax = plt.subplots(1, 1)
ax.bar(x=[0,1],height=[xA.mean(),xB.mean()],
yerr=[(CIA[1]-CIA[0])/2,(CIB[1]-CIB[0])/2],capsize=20,color=("r","g"))
ax.set(xlabel="Hospital",ylabel="Energy usage")
ax.set_xticks([0,1],("A","B"))
plt.tight_layout()
plt.show()
```



Here care must taken in the interpretation of this plot: it is natural, if your main aim is a comparison of the two means, to immediately visually check whether the shown error bars, in this case the confidence intervals, overlap or not, to make a conclusion about group difference. Here they actually just overlap - could be checked by looking at the actual CIs:

```
# The confidence intervals
print(CIA)

ConfidenceInterval(low=np.float64(7.195617231957511), high=np.float64(9.391049434709158

print(CIB)

ConfidenceInterval(low=np.float64(9.223278703268573), high=np.float64(11.37227685228698
```

And the conclusion would (incorrectly) be that the groups are not statistically different. However, remind that we found above that the p-value $= 0.008323$, so we concluded that there was strong evidence of a mean difference between the two nurse groups.

The problem of the "overlapping CI interpretation" illustrated in the example comes technically from the fact that standard deviations are not additive but variances are

$$\begin{aligned} \sigma_{(\bar{X}_A - \bar{X}_B)} &\neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}, \\ \mathrm{V}(\bar{X}_A - \bar{X}_B) &= \mathrm{V}(\bar{X}_A) + \mathrm{V}(\bar{X}_B). \end{aligned} \tag{3-55}$$

The latter is what the confidence interval for the *mean difference* $\mu_A - \mu_B$ is using and what should be used for the proper statistical comparison of the means. The former is what you implicitly use in the "overlapping CI interpretation approach".

The proper standard deviation (sampling error) of the *sample mean difference* due to Pythagoras, is smaller than the sum of the two standard errors: assume that the two standard errors are 3 and 4. The sum is 7, but the square-root of the squares is $\sqrt{3^2 + 4^2} = 5$. Or more generally

$$\sigma_{(\bar{X}_A - \bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}. \tag{3-56}$$

So we can say the following:

Remark 3.59

When interpreting two (and multi-) independent samples mean bar plots with added confidence intervals:

When two CIs do NOT overlap: The two groups are significantly different

When two CIs DO overlap: We do not know from this what the conclusion is (but then we can use the presented two-sample test method)

One can consider other types of plots for visualizing (multi)group differences. We will return to this in Chapter 8 on the multi-group data analysis, the so-called Analysis of Variance (ANOVA).

3.2.3 The paired design and analysis

Example 3.60 Sleeping medicine

In a study the aim is to compare two kinds of sleeping medicine A and B. 10 test persons tried both kinds of medicine and the following results are obtained, given in prolonged sleep length (in hours) for each medicine type:

Person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

Note that this is the same experiment as already treated in Example 3.21. We now in addition see the original measurements for each sleeping medicine rather than just individual differences given earlier. And we saw that we could obtain the relevant analysis (p-value and confidence interval) by a simple call to the `ttest_1samp` function using the 10 differences:

```
# Read the samples
x1 = np.array([0.7, -1.6, -0.2, -1.2, -1.0, 3.4, 3.7, 0.8, 0.0, 2.0])
x2 = np.array([1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4])

# Take the differences
dif = x2 - x1

# t-test on the differences
test = stats.ttest_1samp(dif,popmean=0)
print(test.statistic,test.pvalue,test.df)

4.671645978656774 0.0011658764685528319 9


stats.ttest_1samp(dif,popmean=0).confidence_interval(0.95)

ConfidenceInterval(low=np.float64(0.8613337442036719), high=np.float64(2.47866625579632
```

The example shows that this section actually could be avoided, as the right way to handle this so-called paired situation is to apply the one-sample theory and methods from Section 3.1 on the differences

$$d_i = x_i - y_i \quad \text{for } i = 1, 2, ..., n. \tag{3-57}$$

Then we can do all relevant statistics based on the mean $\bar{d}$ and the variance s_d^2 for these differences.

The reason for having an entire section devoted to *the paired t-test* is that it is an important topic for experimental work and statistical analysis. The paired design for experiments represents an important generic principle for doing experiments as opposed to the un-paired/independent samples design, and these important basic experimental principles will be important also for multi-group experiments and data, that we will encounter later in the material.

Example 3.61 Sleeping medicine

And similarly in Python, they have prepared way to do the paired analysis directly on the two-sample data:

```
# Give both samples, but make paired t-test
test = stats.ttest_rel(x2,x1)
print(test.statistic,test.pvalue,test.df)

4.671645978656774 0.0011658764685528319 9


stats.ttest_rel(x2,x1).confidence_interval(0.95)

ConfidenceInterval(low=np.float64(0.8613337442036719), high=np.float64(2.47866625579632
```

Paired vs. completely randomized experiments

An experiment like the one exemplified here where two treatments are investigated can essentially be performed in two different ways:

Completely Randomized (independent samples) 20 patients are used and completely at random allocated to one of the two treatments (but usually making sure to have 10 patients in each group). So: different people in the different groups.

Paired (dependent samples) 10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments to make sure that it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random). So: the same people in the different groups.

Generally, one would expect that whatever the experiment is about and which observational units are involved (people, patients, animals) the outcome will be affected by the properties of each individual – the unit. In the example, some people will react positively to both treatments because they generally are more prone to react to sleeping medicines. Others will not respond as much to sleeping medicine. And these differences, the person-to-person variability, will give a high variance for the Welch independent samples t-test used for

the independent samples case. So generally, one would often prefer to carry out a paired experiment, where the generic individual variability will not blur the signal – one can say that in a paired experiment, each individual serves as his/her own control – the effect of the two treatments are estimated for each individual. We illustrate this by analysing the example data wrongly, as if they were the results of a completely randomized experiment on 20 patients:

|||| Example 3.62 Sleeping medicine - WRONG analysis

What happens when applying the wrong analysis:

```
# WRONG analysis
test = stats.ttest_ind(x2,x1,equal_var=False)
print(test.statistic,test.pvalue,test.df)

1.9334408348617207 0.06915652250932773 17.900065494971773
```

Note how the p-value here is around 0.07 as opposed to the 0.001 from the proper paired analysis. Also the confidence interval is much wider. Had we done the experiment with 20 patients and gotten the results here, then we would not be able to detect the difference between the two medicines. What happened is that the individual variabilities seen in each of the two groups now, incorrectly so, is being used for the statistical analysis and these are much larger than the variability of the differences:

```
# The sample variances of each sample and of the differences
print(x1.var(ddof=1))

3.4515555555555557


print(x2.var(ddof=1))

4.009


print((x2-x1).var(ddof=1))

1.2778888888888886
```

3.2.4 Validation of assumptions with normality investigations

For normality investigations in two-sample settings we use the tools given for one-sample data, presented in Section 3.1.8. For the paired setting, the investigation would be carried out for the differences. For the independent case the investigation is carried out within each of the two groups.

3.3 Planning a study: wanted precision and power

Experiments and observational studies are always better when they are carefully planned. Good planning covers many features of the study. The observations must be sampled appropriately from the population, reliable measurements must be made and the study must be "big enough" to be able to detect an effect of interest. And if the study becomes too big, effects of little practical interest may become statistically significant, and (some of) the money invested in the study will be wasted. Sample size is important for economic reasons: an oversized study uses more resources than necessary, this could be both financial but also ethical if subjecting objects to potentially harmful treatments, an undersized study can be wasted if it is not able to produce reliable results.

Sample size is very important to consider before a study is carried out.

3.3.1 Sample Size for wanted precision

One way of calculating the required sample size is to work back from the wanted precision. From (3-10) we see that the confidence interval is symmetric around $\bar{x}$ and the half width of the confidence interval (also called the margin of error (ME)) is given as

$$ME = t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{3-58}$$

Here $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile from the t-distribution with $n-1$ degrees of freedom. This quantile depends on both α and the sample size n, which is what we want to find.

The sample size now affects both n and $t_{1-\alpha/2}$, but if we have a large sample (e.g. $n \geq 30$) then we can use the normal approximation and replace $t_{1-\alpha/2}$ by the quantile from the normal distribution $z_{1-\alpha/2}$.

In the expression for ME in Equation (3-58) we also need σ, the standard deviation. An estimate of the standard deviation would usually only be available after the sample has been taken. Instead we use a guess for σ possibly based on a pilot study or from the literature, or we could use a scenario based choice (i.e. set σ to some value which we think is reasonable).

For a given choice of ME it is now possible to isolate n in Equation (3-58) (with the normal quantile inserted instead of the t-quantile):

Method 3.63 The one-sample CI sample size formula

When σ is known or guessed at some value, we can calculate the sample size n needed to achieve a given margin of error, ME, with probability $1-\alpha$ as

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2. \tag{3-59}$$

Example 3.64 Student heights

In Example 3.1 we inferred using a sample of heights of 10 students and found the sample mean height to be $\bar{x} = 178$ and standard deviation $s = 12.21$. We can now calculate how many students we should include in a new study, if we want a margin of error of 3 cm with confidence 95%. Using the standard deviation from the pilot study with 10 students as our guess we can plug into Method 3.63

$$n = \left(\frac{1.96 \cdot 12.21}{3}\right)^2 = 63.64.$$

These calculations show that we should include 64 students, the nearest integer to 63.64.

The formula and approach here has the weakness that it only gives an "expected" behaviour of a coming experiment - at first reading this may seem good enough, but if you think about it, it means that approximately half of the times the actual width will be smaller and the other half, it will be larger than expected. If the uncertainty variability is not too large it might not be a big problem, but nothing in the approach helps us to know whether it is good enough – we cannot guarantee a minimum accuracy with a certain probability. A more advanced approach, that will help us control more precisely that a future experiment/study will meet our needs, is presented now.

3.3.2 Sample size and statistical power

Another way of calculating the necessary sample size is to use the power of the study. The *statistical power of a study is the probability of correctly rejecting H_0 if H_0 is false*. The relations between Type I error, Type II error and the power are seen in the table below.

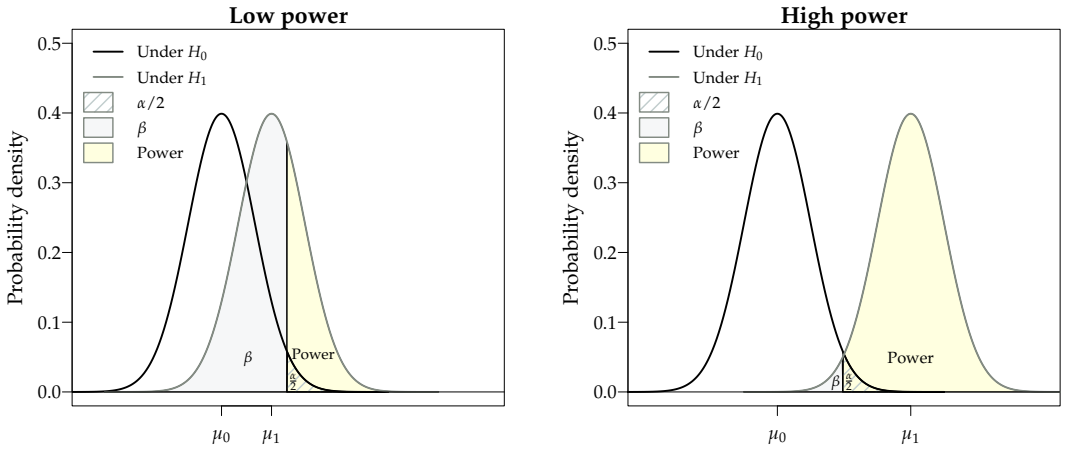


Figure 3.2: The mean μ_0 is the mean under H_0 and μ_1 the mean under H_1. When μ_1 increases (i.e. moving away from μ_0) so does the power (the yellow area on the graph).

	Reject H_0	Fail to reject H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0 (Power)	Type II error (β)

The power has to do with the Type II error β, the probability of wrongly accepting H_0, when H_0 actually is false. We would like to have high power (low β), but it is clear that this will be impossible for all possible situations: it will depend on the scenario for the potential mean – small potential effects will be difficult to detect (low power), whereas large potential effects will be easier to detect (higher power), as illustrated in Figure 3.2. In the left plot we have the mean under H_0 (μ_0) close to the mean under the alternative hypothesis (μ_1) making it difficult to distinguish between the two and the power becomes low. In the right plot μ_0 and μ_1 are further apart and the statistical power is much higher.

The power approach to calculating the sample size first of all involves specifying the null hypothesis H_0. Then the following four elements must be specified/chosen:

- The significance level α of the test (in Python: `alpha`)
- A difference in the mean that you would want to detect, delta
- The standard deviation σ (`sd` in the code)
- The wanted power $(1 - \beta)$ (in Python: `power`)

When these values have been decided, it is possible to calculate the necessary sample size, n. In the one-sided, one-sample t-test there is an approximate closed form for n and this is also the case in some other simple situations.

Python offers easy to use functions for this not based on the approximate normal distribution assumption, but using the more proper t-distributions. In more complicated settings even it is possible to do some simulations to find the required sample size.

Method 3.65 The one-sample sample size formula

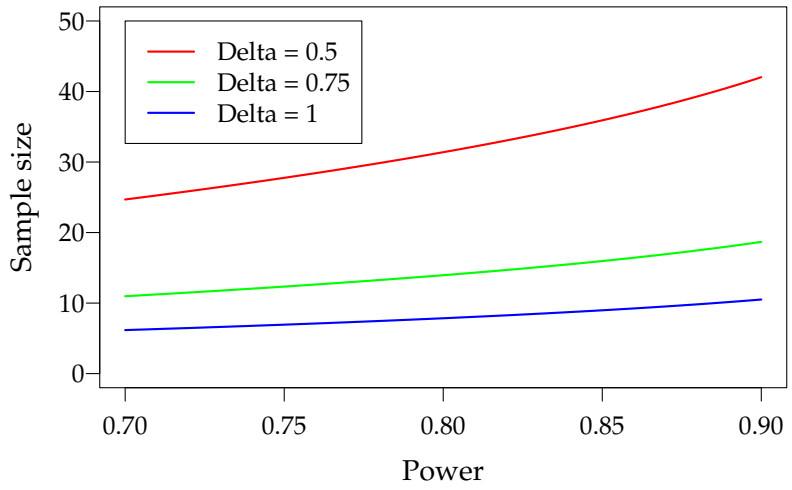
For the one-sample t-test for given α, β and σ

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)}\right)^2,$$

where $\mu_0 - \mu_1$ is the difference in means that we would want to detect and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Example 3.66 Sample size as function of power

The following figure shows how the sample size increases with increasing power using the formula in 3.65. Here we have chosen $\sigma = 1$ and $\alpha = 0.05$. Delta is $\mu_0 - \mu_1$.



Example 3.67 Student heights

If we return to the example with student heights 3.1, we might want to collect data for a new study to test the hypothesis about the mean height

$$H_0 : \mu = 180$$

Against the alternative

$$H_1 : \mu \neq 180$$

This is the first step in the power approach. The following four elements then are:

- Set the significance level α equal to 5%
- Specify that we want to be able to detect a difference of 4 cm
- We will use the standard deviation 12.21 from the study with 10 subjects as our guess for σ
- We want a power of 80%

Using the formula in 3.65 we get

$$n = \left(12.21 \cdot \frac{0.84 + 1.96}{4}\right)^2 = 73.05.$$

So we would need to include 74 students.
We could also use a Python-function for power and sample size based on the t-distributions:

```
# The sample size for power=0.80
delta = 4
sd = 12.21
alpha = 0.05
power = 0.8
smp.TTestPower().solve_power(effect_size=delta/sd, alpha=alpha,
power=power)

75.07715049712685
```

From the calculations in Python avoiding the normal approximation the required sample size is 76 students, very close to the number calculated by hand using the approximation above.

In fact the Python-function is really nice in the way that it could also be used to find the power for a given sample size, e.g. $n = 50$ (given all the other aspects):

```
delta = 4
sd = 12.21
nobs = 50
alpha = 0.05
smp.TTestPower().solve_power(effect_size=delta/sd, nobs=nobs,
alpha=alpha)

np.float64(0.6220915188555853)
```

This would only give the power 0.62 usually considered too low for a relevant effect size.

And finally the Python-function can tell us what effect size and delta that could be detected by, say, $n = 50$, and a power of 0.80:

```
nobs = 50
alpha = 0.05
power = 0.80
sd = 12.21
effect = smp.TTestPower().solve_power(nobs=nobs, alpha=alpha,
power=power)
delta = effect*sd
print(delta)

4.935074496518317
```

So with $n = 50$ only a delta as big as 4.9 would be detectable with probability 0.80.

To summarize: if we know/define 4 out the 5 values: significance level, power $(1 - \beta)$, n, *delta*, and σ, we can find the 5'th. In the Python-function, the arguments are called `alpha`, `power`, `nobs`, and `effect_size`, where `effect_size` is delta/σ.

In the practical planning of a study, often a number of scenario-based values of *delta* and σ are used to find a reasonable size of the study.

3.3.3 Power/Sample size in two-sample setup

For power and sample size one can generalize the tools presented for the one-sample setup in the previous section. We illustrate it here by an example of how

to work with the inbuilt Python-function:

|||| Example 3.68 Two-sample power and sample size computations in Python

We consider the two-sample hypothesis test

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2$$

```
# Finding the power of detecting a group difference of 2
# with sigma=1 for n=10
delta = 2
sd = 1
nobs = 10
alpha = 0.05
smp.TTestIndPower().solve_power(effect_size=delta/sd, nobs1=nobs,
alpha=alpha, ratio=1.0)

np.float64(0.9881789691948746)
```

```
# Finding the sample size for detecting a group difference of 2
# with sigma=1 and power=0.9
delta = 2
sd = 1
alpha = 0.05
power = 0.90
smp.TTestIndPower().solve_power(effect_size=delta/sd, alpha=alpha,
power=power, ratio=1.0)

6.386755384175011
```

```
# Finding the detectable effect size (delta)
# with sigma=1, n=10 and power=0.9
nobs = 10
alpha = 0.05
power = 0.90
sd = 1
effect = smp.TTestIndPower().solve_power(nobs1=nobs, alpha=alpha,
power=power, ratio=1.0)
delta = effect*sd
print(delta)

1.5336931237722076
```

Note that we now use the function `TTestIndPower`, which uses the arguments `nobs1` and `ratio` to specify the number of observations in the two samples.

Glossaries

Alternative hypothesis [Alternativ hypotese] The alternative hypothesis (H_1) is ofter the negation of the null hypothesis 30, 32, 33, 50, 66

Analysis of Variance [Variansanalyse]

χ^2-distribution [χ^2-fordeling (udtales: chi-i-anden fordeling)] 19, 20

confidence interval [Konfidensinterval] The confidence interval is a way to handle the uncertainty by the use of probability theory. The confidence interval represents those values of the unknown population mean μ that we believe is based on the data. Thus we believe the true mean in the statistics class is in this interval 9,

Central Limit Theorem [Centrale grænseværdisætning] The Central Limit Theorem (CLT) states that the sample mean of independent identically distributed outcomes converges to a normal distribution 14,

Critical value *Kritisk værdi* As an alternative to the p-value one can use the so-called critical values, that is the values of the test-statistic which matches exactly the significance level 28–30, 32, 50, 56

Degrees of freedom [Frihedsgrader] The number of "observations" in the data that are free to vary when estimating statistical parameters often defined as $n - 1$ 6, 9, 17, 19, 20, 24, 28, 32, 48–50, 52, 54, 56, 64

Empirical cumulative distribution [Empirisk fordeling] The empirical cumulative distribution function F_n is a step function with jumps i/n at observation values, where i is the number of identical observations at that value 37

Histogram [Histogram] The default histogram uses the same width for all classes and depicts the raw frequencies/counts in each class. By dividing the raw counts by n times the class width the density histogram is found where the area of all bars sum to 1 2, 36, 37

Independence [Uafhængighed] 13, 36

Independent samples [Uafhængige stikprøver] 59–61

(Statistical) Inference [Statistisk inferens (følgeslutninger baseret på data)] 1, 11, 14

Interval [Interval] Data in a specified range 1

Median [Median, stikprøvemedian] The median of population or sample (note, in text no distinguishment between *population median* and *sample median*) 44

Normal distribution [Normal fordeling] 1, 2, 4–6, 8, 11, 12, 14, 16, 18, 36–43, 64, 67

Null hypothesis [Nulhypotese (H_0)] 21–24, **25**, 26–30, 32, 33, 35, 47, 49–52, 54, 56

One-sample t-test Missing description 28, 31, 32, 66

One-sided (test) [Énsidet test] Is also called directional (test) 66

***P*-value** [*p*-værdi (for faktisk udfald af en teststørrelse)] 22–26, 28, 31, 35, 47, 54, 58, 62

Sample mean [Stikprøvegennemsnit] The average of a sample 1, 3–5, 8, 10, 14, 18, 21, 46, 65

Two-sided (test) [Tosidet test (test med tosidet alternativ)] Is also called non-directional (test) 55

Acronyms

ANOVA Analysis of Variance 51, 59, *Glossary:* Analysis of Variance

cdf cumulated distribution function *Glossary:* cumulated distribution function

CI confidence interval 1–3, 8–11, 16, 18, 20, 21, 26, 29, 30, 35, 42, 44, 46–48, 55–59, 62, 64, 65, *Glossary:* confidence interval

CLT Central Limit Theorem 14, 16, *Glossary:* Central Limit Theorem

IQR Inter Quartile Range *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function