Experimental design and statistics for Marine Ecology

Part 1

STATISTICAL ANALYSIS AND DESIGN OF ECOLOGICAL EXPERIMENTS - INTRODUCTION

The scientific method

One definition of science is that it aims to explain observable phenomena through studies of the patterns and processes generating them. One example of a phenomenon that requires an explanation is the observation that newly settled larvae of cockles (a bivalve) are negatively correlated to adult cockles. There may be several explanations, or models, for this observed pattern, e.g.:

1. Predation on settling larvae by adult cockles
2. Burrowing activity by adult cockles kills larvae
3. Other predators on settling larvae hide among adult cockles
4. An active choice by larvae to not settle among adult cockles

From these models it is possible to generate the following testable hypotheses:

- Hypothesis for model 1: larval remains should be found in stomachs of adult cockles
- Hypothesis for model 2: preventing adults from burrowing should result in more settling larvae
- Hypothesis for model 3: cages that exclude other predators should result in more settling larvae
- Hypothesis for model 4: preparations with extract of adult cockles should inhibit settling of larvae

After proposing these hypotheses we then proceed to plan different studies to collect observations or to do experiments in order to test if any of these hypotheses are supported or should be rejected. This process is at the heart of scientific investigations and it takes considerable understanding and practice to develop the skills necessary to build knowledge based on the scientific method. This module of experimental design and statistical analysis aims to an integrated understanding of the scientific method with emphasis on practical use. During the course you will carry out practical projects applying the theory and methods of hypothesis testing, experimental design and statistical analysis.
One version of the scientific method, as defined here, is summarized in Fig. 1 below.

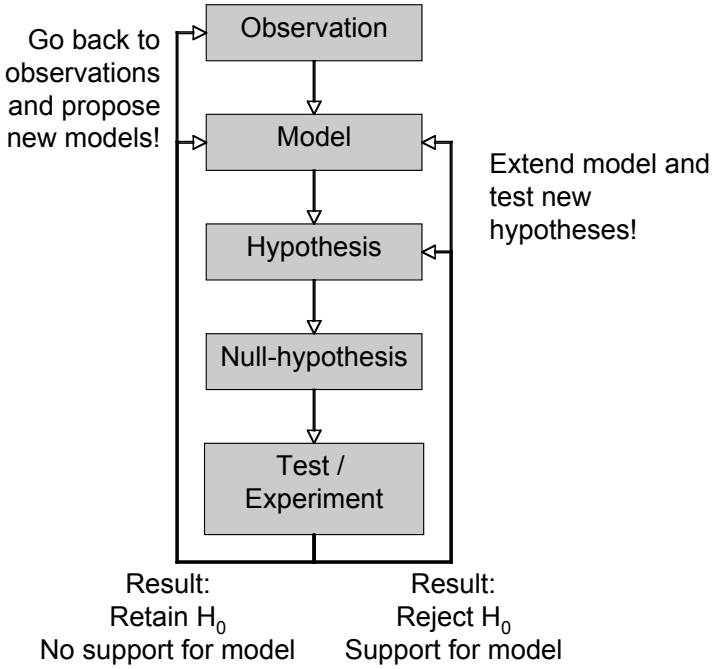


Fig. 1 Formal logical loop of the scientific method.

How can we reliably detect patterns of interest?

Often a scientific study begins with the observation of a phenomenon, or pattern, that we want to explain and understand. But how do we know that the pattern is real? It is well known that we do not observe the world objectively. Instead we are often biased in our observations by both conscious and unconscious aspects. Do we already have a lot of knowledge of a particular environment? Have we recently read an influential book about similar patterns? Are we tired or hungry? All such factors may contribute to how we observe the world.

In the example above about cockles we may one day observe the pattern "there are fewer settled larvae where I also find adult cockles". But is this observation generally true? In a complex world patterns often vary in space and time. Was our observation just a coincidence (a random pattern) or were we unconsciously longing to see this pattern? Clearly, it would be a step forward if we could document this observation in an objective way, find out if it is a general pattern, and maybe also something about the magnitude of the pattern, i.e. how few larvae that settled among how many adults.

From the way we observe the world it is clear that already the detection and description of patterns require a systematic method. In our cockle example we could design a structured observation as follows:

1. Identify the area that we will include in our observation
2. Randomly select plots with many and plots with few adult cockles

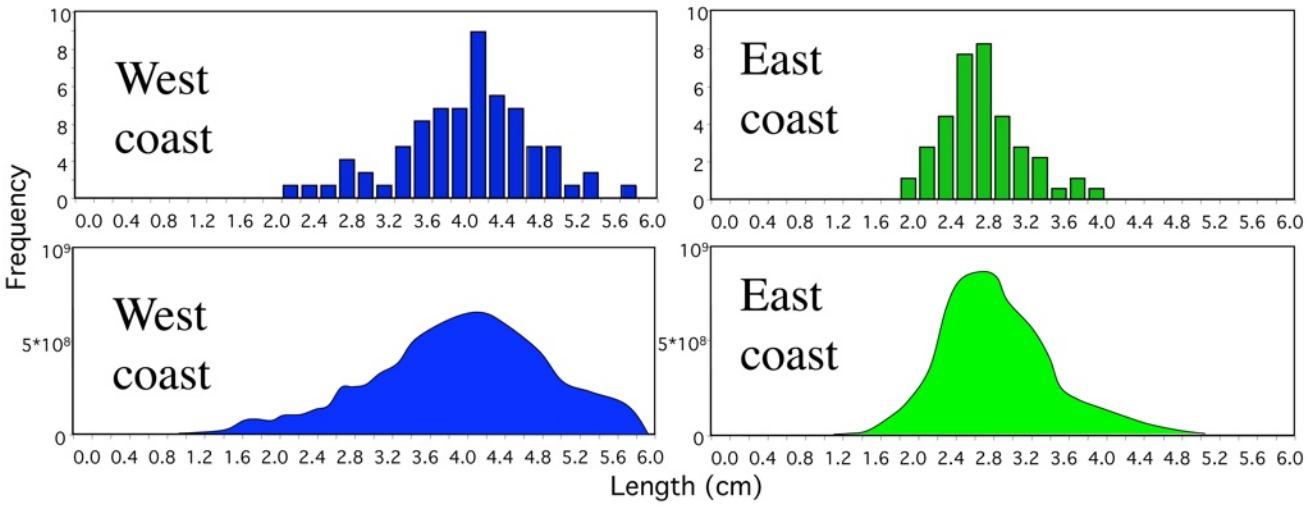
3. Take representative samples of sediment from each of these plots, and count all larvae found.
4. Compare the number of larvae found in plots with few to plots with many adult cockles, possibly with some statistical method to find out if the difference is greater than could be expected by chance only.

In ecological studies, and indeed in most scientific disciplines, quantitative measurements form a very important part of most methods. With quantitative knowledge it is possible to make clear descriptions of patterns, and hypotheses about what causes these patterns can be precisely formulated. In many cases it is necessary to include quantitative information to be able to separate different competing hypotheses.

When we quantify things, e.g. number of settling larvae, this is called a variable, because it can take on different values. In this case we counted the frequency of larvae for two different classes or categories (high and low abundance of adults). This type of variable is called nominal. Often things are measured on a scale, e.g. the length of mussels or the age of a sea-weed. Such variables are called ordinal and they can be continuous or discrete. Two ordinal data have a defined interval between them, e.g. the difference in length between two mussels. In contrast, for ranked variables, only the order of the different data is known. Ordinal variables can be transformed into ranked by sorting, e.g. three mussels with lengths of 4.3, 3.6 and 5.7 cm will have the ranks 2, 3 and 1.

Frequency distributions

Fig. 2a (upper panels) Samples from the two true populations. 2b (lower panels) Frequency



distributions of two imagined true populations of mussel shell lengths (see text for explanations).

Biological systems is characterized by their tremendous variability at most levels. This biodiversity makes biology so fascinating. But the variability also makes it more difficult to detect patterns and to find out their causes. In comparison, empirical data in chemistry and physics are often much less variable making it easier to study cause-effect relationships. Experimental design and statistics form an important part of ecological science just because the studied systems are so variable in space and time.

What does all this variation mean? Let us take an example. We want to know if, and how much, blue mussels differ in length between the Swedish west and east coasts. We could take one mussel from each coast and compare their lengths. But we soon discover that not all mussels on the same coast have identical lengths. So we begin to measure many individuals, record their lengths and plot them as shown in Fig. 2a. As we measure more and more mussels we approach the true frequency distributions of all mussels shown in Fig. 2b.

The length of all mussels from the two coasts can be viewed as two frequency distributions. A frequency distribution (e.g. Fig. 2) is a plot of classes of observations (*x*-axis) and how often these classes occur (*y*-axis). There are several ways of characterizing frequency distributions. We can plot them as in Fig. 2 or we can calculate different parameters to describe important aspects of the distributions.

Parameters

1. The location parameter. There are several location parameters (e.g. mean, median and mode) that measures where on the x-axis the frequency distribution is located. However, the most common is the arithmetical mean which has the property that the sum of distances from each individual to the mean is zero. In shorthand the arithmetical mean (μ) of the true population is calculated as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

2. The dispersion parameter. The second most important parameter describing a frequency distribution is the dispersion parameter that indicates how much individual data are spread around the mean. The most common measure of dispersion is the variance. The variance (σ^2) of the true population is defined as the sum of squared deviations to the mean divided by the number of deviations, i.e. the number of observations. The variance is calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(X_i - \mu\right)^2}{N}$$

Often the variance is converted to the same units as the mean by taking the square root. This gives the standard deviation (*SD*):

$$\sigma = \sqrt{\sigma^2}$$

There are other parameters describing aspects of a frequency distribution. The most common are the skewness (σ^3) measuring the symmetry of the left and right tail of the distribution and kurtosis (σ^4) measuring the peakiness of the distribution. Note that information is lost by characterizing a frequency distribution with a few parameters. Thus, two different distributions may have the same parameters.

Samples

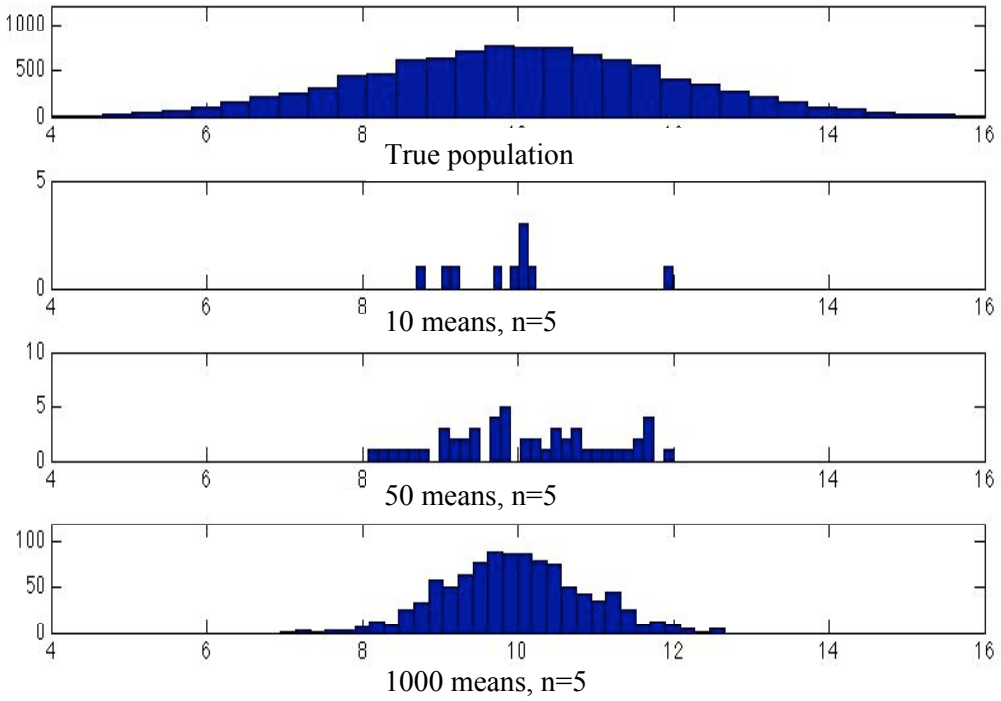
The objective with our study of mussels is of course to say something about the real world, in this case the length of mussels on different coasts. Out there on the coasts, there are two statistical populations, one on each coast, probably consisting of billions of mussels (Fig 2b). To learn something about these true populations we have drawn two samples (Fig 2a). To draw any valid conclusions about the true populations the samples need to be representative. There is actually only one representative sample from a true population; the sample that has the same relative frequency distribution as the true population. The best sampling strategy is to make it equally probable for each individual to be included in the sample. This is often a major challenge and requires much thought. How do we know if a sample is representative? In real life this is generally poorly known and requires a lot of knowledge about the population we want to sample. When little is known about the true population it is common that sampling is random. This prevents us from making subjective choices where to sample. Random sampling is not necessarily the best strategy if there is detailed information about, e.g. the spatial distribution of an organism. In this case different types of stratified sampling may lead to more representative samples.

The problem of obtaining representative samples is seriously overlooked. One example is that the efficiency of many marine sampling equipment, e.g. bottom grabs, differs among clay, silt and sand sediments. The differences in the sample frequency distributions may then just reflect the sampling efficiency and not differences among the true populations. Can you think of other cases where it is difficult to obtain representative samples?

The central limit theorem

When thinking closely, it seems almost impossible to conclude something reliable about millions of mussels in the field from a small sample of a few individuals. Fig. 3 shows an example where random samples with 5 mussels (n=5) are taken from a very large true population. Also shown are the frequency distributions of an increasing number of sample means. There are two things to observe:

Fig. 3. Frequency distributions showing the original true population and three sample mean distributions with increasing number of samples. The sample size is 5.



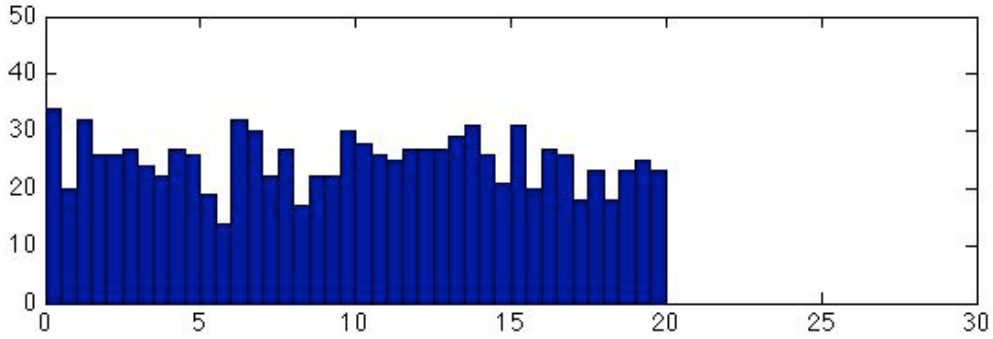
1. If the true population Fig. 3 forms a so called normal distribution, the distribution of sample means also approaches a normal distribution
2. The distribution of sample means is more narrow, i.e. has a smaller variance than the true population.

Now, it may not be so surprising that the sample means from a normal distribution also tend to a normal distribution. But the very interesting property of random samples is that their means will approximately form a normal distribution also when drawn from true populations that are *not* normally distributed. In Fig. 4 this is illustrated with a sample mean distribution drawn from a nearly uniform true population, and in Fig. 5 with sample means drawn from a skewed true population.

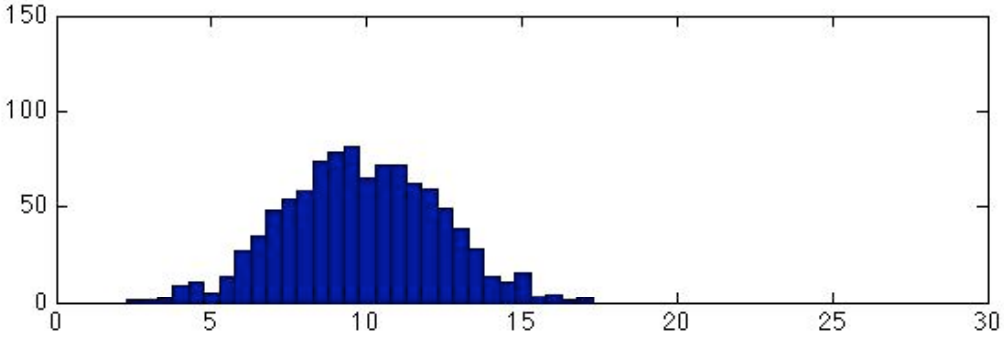
The tendency of sample means to be normally distributed regardless of the distribution of the true population is stated by the Central Limit Theorem (CLT). CLT further states that there is a specific relationship between the variance of the true population and the distribution of sample means. Figure 6 shows samples of size n drawn from a true population with mean μ and variance σ^2. For a very large number of samples the sample mean distribution also has mean μ but with variance σ^2/n.

Fig. 4. Distribution of sample means (lower panel) drawn from a near uniform true population.

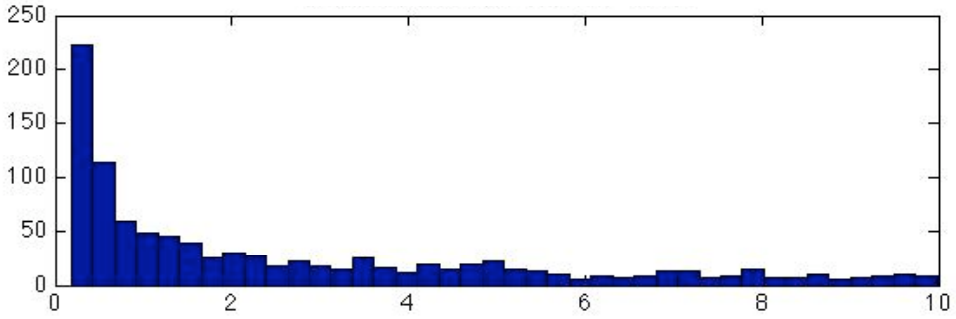
Fig. 5. Distribution of sample means (lower panel) drawn from a skewed true population.



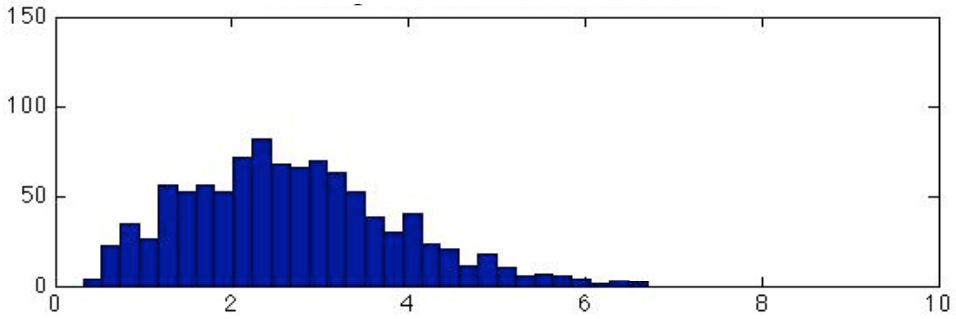
The true population



1000 means, n=5



The true population

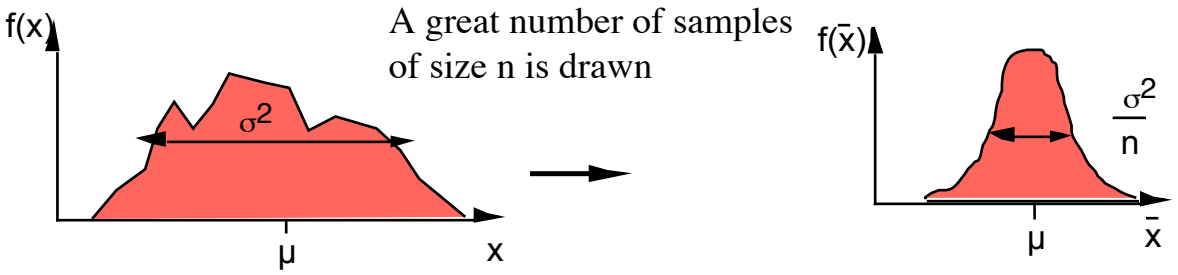


1000 means, n=5

The relationship between the variance of the true population and the distribution of sample means implies that as we increase the sample size *n* the variance of the sample mean

distribution (σ^2/n) becomes smaller and smaller. It makes intuitive sense that the mean of a large sample should vary less from sample to sample, than a small one. In a large sample

Fig. 6. Schematic drawing of the relationship between the distribution of the true population and the



sample means. Note that X is the value of a single individual while $\overline{X}$ is the mean of a sample.

it is less likely to obtain means that contain measurements from only one tail of the true population (Fig. 6). The CLT can also be mathematically derived in terms of probability theory.

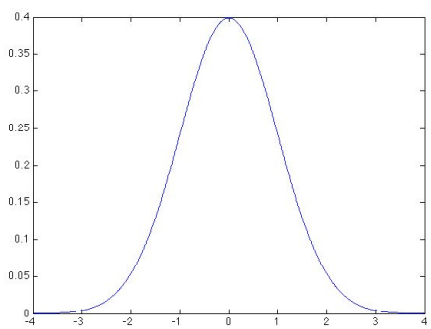
As we will see later, many statistical methods work surprisingly well for our ambitious objective to draw conclusions about natural phenomena from rather small samples. However, it should be pointed out that the CLT does not apply well for some distributions of true populations. The most important cases are highly skewed distributions (more than in Fig. 5) and distributions with many peaks (multimodal).

The normal distribution

The normal distribution (Fig. 7, "the bell curve") was introduced above, and the normal distribution is found in any textbook on statistics. Why is it so important? Without going into any mathematical details there are two major reasons:

1. Many natural processes, e.g. biological, that depends on the collective effect of many small contributions tend to be normally distributed. And many other non-normal

Fig. 7. The standard normal distribution with μ=0 and σ^2=1.



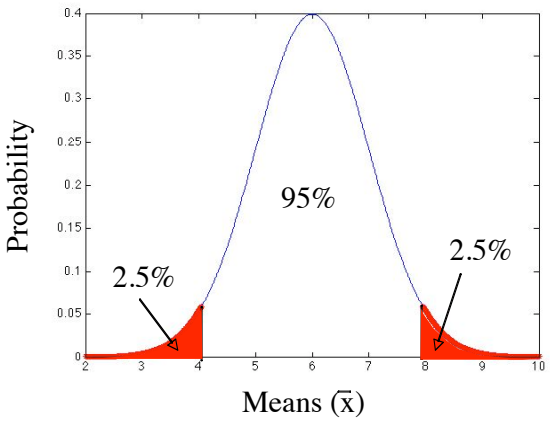
distributions occurring in nature can often be easily transformed to a normal distribution.

2. As explained above and stated by the Central Limit Theorem, sample means tend to be distributed as a normal distribution regardless of the original distribution of data.

Because the normal distribution has a defined shape and can be easily determined if we know the mean and the variance we here have an important link between our sample and the unknown true population.

One important use of frequency distributions is to convert them to probability distributions. The distribution of mussel lengths in the true population (e.g. the lower panel in Fig. 2) shows how many mussels there are in each length class. If we divide the frequency of each class by the total number of mussels, we get the proportion of mussels in each length class. Since all mussel lengths that can exist are represented in the distribution, the proportions of all classes should sum up to 1 (or 100%). We can also express this as a probability. If the proportion of mussels in the length class 5-6 cm is 0.3, the probability is also 0.3 that one mussel drawn from the true population comes from this class. With a probability distribution we can now ask questions like: what are the 5% most extreme means? Fig. 8 shows the boundaries of the 2.5% shortest and 2.5% longest mussel lengths, with the 95% most common lengths in between. As seen below this can be developed into a tool to say something about how close our sample is to the true population mean.

Fig. 8. Probability distribution of sample means showing probabilities of extreme values.



Standard deviation, standard error and the precision of means

The first important question we want to answer is:

- If we have taken a sample from an unknown true population is it then possible to say something about the true mean μ based on the data we have in our sample?

We will approach this essential question in steps.

1. We first make the unrealistic assumption that the true mean μ in Fig. 6 is unknown but that we actually know the true variance $\sigma^2 = 4$.
2. We draw a sample of 9 mussels with lengths 6, 4.5, 7, 3, 7.5, 9.5, 4, 7.5 and 9 cm. The sample mean ($\overline{X}$) of the 9 mussels is 6.4 cm.

3. From the relationship in Fig. 6 we know that the variance of the distribution of many sample means of size *n* is σ^2/n, i.e. 4/9=0.44. Note this small trick; we only have one sample mean, but if we had repeated this sampling many times we would expect (from CLT) that the distribution of means has a variance of σ^2/n.
4. The variance is calculated from the sum of the squared deviations to the mean so all the data are squared compared to the mean and now have unit cm^2. We can easily convert the variance to the same dimension as the mean by taking the square root. The square root of the variance is called the standard deviation (*SD*):

   $SD = \sigma = \sqrt{\sigma^2}$

   and the square root of the variance of the mean distribution is called standard error (*SE*):

   $$SE = \frac{\sigma}{\sqrt{n}} = \frac{SD}{\sqrt{n}}$$
5. Note that *SD* is a biological characteristic of the true population. Because of, e.g. genetics, food availability, spatial variability etc. the mussels show a particular mean length with a particular *SD*. *SE* on the other hand is the standard deviation of our distribution of means. We can affect *SE* by choosing a particular sample size *n*. The larger the sample size, the more narrow *SE* and the less difference among sample means.
6. It turns out that for a normal distribution (Fig. 7) the width of one *SD* or *SE* on either side of the mean includes 68% of all values. So for a distribution of means the ±*SE* includes 68% of all expected means. However, 68% is not a very round figure. Since all normal distributions have the same basic shape and are only determined by the mean and the variance, it is possible to increase this area of means to 95% just by multiplying the *SE* with 1.96. The number 1.96 is just a mathematical feature of normal distributions. If we instead are interested in the area including 99% of all means we multiply *SE* with 2.6.
7. We can now construct what is called a confidence interval for our sample mean. In our mussel example we can calculate an interval that is expected to include 95% of all the means we can draw from the true population. The *SE* is $2/\sqrt{9}$ =0.67, so the lower boundary of the interval is 6.4-1.96*0.67 = 5.1 and the upper boundary is 6.4+1.96*0.67 = 7.7. We conclude that 95% of all means from a sample of size=9, that can be drawn from the true population of mussels, will lie within this interval. The sample mean estimates the true mean but with some error depending on not drawing a perfect representative sample. So, if this interval contains all possible means with 95% probability, this allows us to make the following very powerful statement about the unknown mean of the mussels in the true population: There is a 95% probability that the true mean lies in the interval 5.1 – 7.7 cm. And this is the answer to our question

about what we can say about the true mean μ based on the data we have in our sample.

8. It is important that whenever a mean of a measured variable is presented it should always include the *SE* and the sample size, or a confidence interval for some probability, e.g. 95%.

Sampled variance and the t-distribution

In the first step of the sequence leading to the confidence interval above, we assumed the unrealistic case that we knew the true variance,σ^2, of the mussel population. Obviously, this is normally not the case and the whole exercise of finding the confidence interval for the true mean may again seem futile. However, there is a solution. We begin by estimating the true variance from our sample. This is similar to the calculation of the true variance with two exceptions. We, of course, do not know the true mean so instead the sample mean is used to calculate the sum of squared deviations, and we divide this sum by the sample size minus one (n-1). The sample variance, called s^2, is thus found as:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

First, we see that all we need for the sample variance is found in our sample. Secondly, it seems strange why we should divide by n-1 and not n. This is not so simple to explain. Although there seems to be n independent squared deviations in the calculation of s^2, there are in fact only n-1. This is because we need to calculate the mean from the same sample that we use to calculate s^2. Here we lose one degree of freedom since if we know the sample mean, here 6.4 cm, and the lengths of the first 8 mussels, the 9th mussel must be 9 cm. The same applies to the squared deviations; if we know the mean and 8 of them the 9th is locked. So, to estimate the true variance from the sample the sum of the squared deviations are divided by the available, degrees of freedom, in this case by n-1.

The standard deviation, *SD*, is estimated from the sample variance $\sqrt{s^2}$ and the *SE* is $s/\sqrt{n}$. Armed with an *SE* calculated from the sample it should now be possible to calculate a confidence interval without the need to know the true variance. However, one serious obstacle still remains. It turns out that when we have to estimate the true variance from the sample we add a new uncertainty because this estimate usually differs from the true variance, i.e. there is some sampling error. The smaller the sample size, the larger this error can be. This has an important effect when we want to construct the confidence intervals around our means. If the sample variance adds an additional uncertainty to the uncertainty already present for the sample mean, this should act to expand the confidence interval. That is, we need some way to find new boundaries for the 95% confidence interval we constructed under point 7 above; a boundary that takes into account the added uncertainty from the sample variance. Obviously,

the number 1.96 we multiplied the *SE* to include 95% of all means should be larger, and should increase as sample size gets smaller. One way to solve this problem is as follows:

1. We call the new and yet unknown number that we must multiply with *SE* to get the true 95% confidence interval for $t_{95\%}$. A confidence interval is the same as saying that the difference between our sample mean and the true mean should be less than $SE*t_{95\%}$ with 95% probability, or:

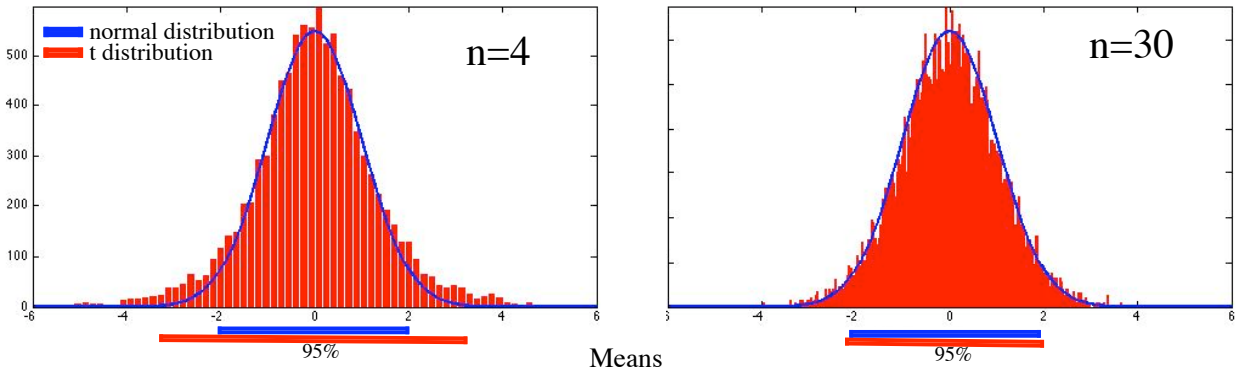
$$-\mathrm{SE}\cdot \mathrm{t}_{95\%} \le \overline{\mathrm{X}} - \mu \le \mathrm{SE}\cdot \mathrm{t}_{95\%}$$

   if we divide by *SE* we get:

$$-\mathrm{t}_{95\%} \le \frac{\overline{\mathrm{X}} - \mu}{\mathrm{SE}} \le \mathrm{t}_{95\%}$$

2. We then identify a large known population, e.g. 1000 dead mussel shells. All these are first measured to calculate the true mean shell length and the true variance and *SD*. From this true and known population we draw a sample of let's say 4 mussels. The sample mean and *SE* are calculated. We also calculate $\left(\overline{X} - \mu\right)/SE$, save this result and repeat this procedure with a new sample of 4 shells. This is done for hundreds of samples and we then plot the frequency distribution of our measure $\left(\overline{X} - \mu\right)/SE$.

Fig. 9. The *t*-distribution for *n*=4 and *n*=30 (shown in red). The normal distribution is shown in blue.



Also indicated are the intervals for the 95% probability area for the *t*- and normal distributions.

3. In Fig. 9 the results of our exercise is plotted for two sample sizes, *n*=4, and *n*=30. The red bars in Fig. 9 show the frequencies of the quantity $\left(\overline{X} - \mu\right)/SE$ called *t* and the distribution of *t* is not surprisingly called the *t*-distribution. Also shown in blue in Fig. 9 is the normal distribution with mean=0 and *SD*=1. It is clear that for small sample sizes the *t* values has "broader shoulders" than the normal distribution, and there is an increased probability for large deviations between the sample and the true mean. The red line below the graphs shows the 95% confidence interval indicating that 95% of all *t*-values are found within these boundaries. The critical *t*-values that delimit this

interval at n=4 are -3.18 and +3.18. This is at last the $t_{95\%}$ that we should multiply our *SE* to find the correct 95% confidence interval! Clearly, for a small sample size like n=4, the t-value (3.18) is substantially greater than the 1.96 we could use when the true variance was known. In fact the 95% confidence interval based on a sample of only 4 mussels became 62% broader.

4. So, we have now reached the goal of being able to estimate where the true mean should lie with some probability based on any sample size. Note that we need a new t-distribution for each sample size. These distributions or rather the critical t-values for a desired probability are found in statistical tables (or as a formula in Excel). Also note in the right graphs in Fig. 9 that as sample size increases the t-value approaches that for the normal distribution (e.g., 1.96 for 95% probability). Finally, note that tables list t-values after the degrees of freedom of the sample variance, i.e. n-1. In our example in the previous section with the sample of 9 mussels the t-value with 8 degrees of freedom for 95% probability is 2.31 and the true mean should lie within 6.4 ± 1.5 cm (check this yourself).

Testing hypotheses – test a sample against a hypothetical value

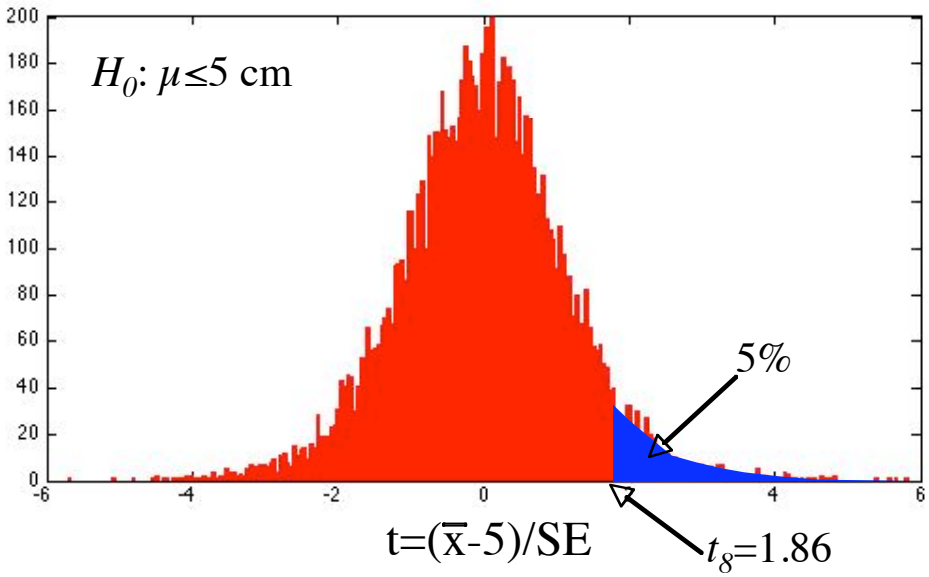
Apart from having solved the problem how to construct the confidence interval around a sample mean, the procedure we have walked through also demonstrates the principal for a statistical test. Statistical tests are often needed to test research hypotheses generated by some model we have proposed to explain something about the world. And, the testing of hypotheses is one of the main foci of this course.

Let us continue with the example with the blue mussels in Fig. 2. Our very simple model is that the west coast mussels have grown large enough for harvest. From previous experience we know that the most economical size at harvest is 5 cm. So we hypothesize that the mussels in the target area (west coast) have a mean length of 5 cm or longer. We can test this hypothesis either by collecting all mussels on the coast and determine if the true mean exceeds 5 cm, or we can take a smaller sample and work out the probability that the true mean is greater than 5 cm. Of course, under realistic conditions we almost always have to rely on small samples. The idea that we can make conclusions, e.g. test hypotheses, about the true world from small samples is called statistical inference. We approach the test of our hypothesis that the mussels now are ready for harvest along the following sequence:

1. We begin to formulate a so called statistical null-hypothesis, H_0. A statistical H_0 defines a frequency distribution of a parameter calculated from our sample (e.g. mean, *SE*, t) when the hypothesis in not true, in our case when mussels are shorter than 5 cm. The H_0 should include all the possibilities not covered by our hypothesis, here that mussels are longer than 5 cm. If our test rejects H_0 there is thus logical support for our hypothesis.

2. We then take a sample (the same 9 mussels as above). The mean and *SE* are 6.4 and 0.75 cm, respectively. If H_0 is true, the maximum true mean is at 5 cm. How should a sample mean of 9 mussels from such a population with a true mean of 5 cm be distributed? We can change this question a little bit by asking how the difference between the sample mean and the true mean ($\overline{X} - \mu$) should be distributed. The answer is that it should be distributed as *t*SE* with 8 degrees of freedom (as shown above). So, H_0 could be defined by: $t = \frac{\overline{X} - 5}{0.75}$.
3. In Fig. 10 the distribution of *t* is shown if our H_0 is true. As indicated by the blue area there is some low probability that we draw samples with a large difference between the sample mean and 5 cm even from a population with a true mean of 5 cm.

Fig. 10. The *t*-distribution for *n*=9 (8 degrees of freedom) if H_0 is true.



Now, here comes a very important way of thinking about statistical inference. In our sample of 9 mussels we calculated a *t*-value of (6.4-5)/0.75=1.87. What is the probability that we get a *t*-value of 1.87 or larger under the H_0 distribution in Fig. 10? The critical t-value delimiting the 5% area of the highest *t*-values is 1.86, so the probability of getting such a sample if H_0 is true is a little less than 5%. When we test if H_0 is true or not we have to state how low the probability should be for a sample statistic (in this case the *t*-value) to come from the distribution defining H_0. In this case we beforehand decided to use the 5% probability level. We can now conclude that the sample of the 9 mussels did not support the H_0 that the mussels were 5 cm or smaller, i.e. it is OK to harvest them. We have here performed the first statistical test!

One- and two-tailed tests

Note that a statistical test of H_0 can often be either one- or two-tailed. In the example above we formulated H_0 as a one-tailed test. The alternative to H_0 only considered a true mean greater than 5 cm. Accordingly, all our 5% most improbable *t*-values were found in the upper tail of the *t*-distribution in Fig. 10. However, in many cases the alternative may include both high and low *t*-values. Let's illustrate this with another example. A boat is seized by the coast guard because there are reasons to believe that mussels are stolen from a nearby mussel farm.

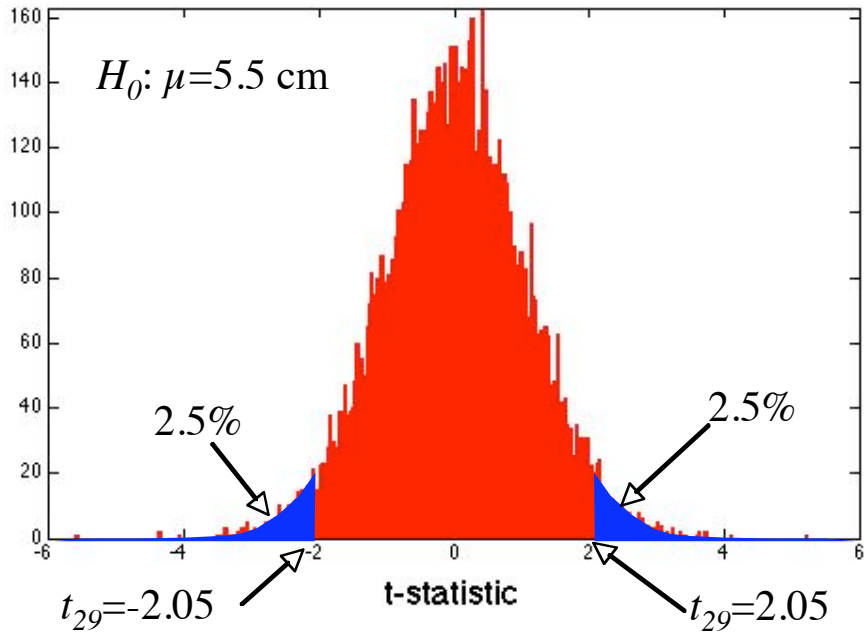


Fig. 11. The *t*-distribution for *n*=30 (29 degrees of freedom) if H_0 is true. Also shown in blue are the two tails with the lowest and highest *t*-values corresponding to 2.5% each of the total probability.

The boat owner claims that she has collected the mussels from a public beach. The mean size of mussels (mean=5.5 cm) in the farm is very well known, and a sample of 30 mussels is taken from the boat. The question is now if it is probable that the mussels from the boat come from the population of farmed mussels. H_0 is here that the mussels in the boat indeed come from the farm population, and in that case we expect the *t*-statistic (*t*-value) of the sample will belong to the most probable 95% area of the *t*-distribution (the red area in Fig. 11).

In contrast to the one-tailed previous example, this alternative hypothesis, i.e. that the mussels in the boat came from elsewhere include populations with both smaller and bigger mussels than a mean of 5.5 cm. Therefore both the lower and the upper tail of the *t*-distribution are used to reject H_0, i.e. we have a two-tailed test. Note that the critical *t*-statistic now should be selected to include the 2.5% lowest and the 2.5% upper tails of the distribution to add up to the 5% probability where we consider it unlikely that the mussels come from the farm. This is

achieved for a sample with 29 degrees of freedom if we (from a table) select a critical *t*-value of ±2.05.

Testing hypotheses – test of a difference between two samples

Most research hypotheses are tested with more complex information than a sample against a known true or theoretical mean. A simple example is to test the hypothesis that mussels are larger on the west coast than on the east coast. Here we have a hypothesis about two populations with unknown parameters. But fortunately we now know how to estimate the true mean and variance through samples. We can test our hypothesis by the following sequence.

1. First, the target populations are defined, e.g. using maps. There are many difficult questions here. Should we include all of the coast or should we exclude areas (e.g. deep water) where mussels do not usually live? What about very exposed shores where wave action tend to exclude mussels? In the Baltic Sea we should probably exclude the coast above N 63° which is the approximate northern limit. Nevertheless, at last we have defined our target populations on the east and west coasts.
2. We then randomly collect mussels across the target areas on both coasts and we end up with the sample distributions shown in Fig. 2.
3. The samples suggest that the west-coast mussels indeed are larger. But, there is always some probability that we could have gotten two rather different samples although they are drawn from populations with the same true mean and variance. The question, as we saw above, is how small that probability is and if we regard this probability so small that we decide that the samples come from different populations.
4. We formulate a H_0 that the two samples come from a single population with the same mean and variance. If they come from the same population we expect that the difference of the sample means would be on average 0, i.e. $E\left(\overline{X}_w - \overline{X}_e\right) = 0$ (E indicates expected value). What about the *SE* of this difference between two samples? If we take the difference between two numbers that both have some uncertainty it seems intuitive that the difference has even more uncertainty. It can be shown that if two sample means have the uncertainty SE_1 and SE_2, the uncertainty of the difference (and also the sum) is $\sqrt{SE_1^2 + SE_2^2}$. So in analogy with the one-sample test above we should be able to calculate the *t*-statistic for this difference of sample means if H_0 is true. We get:

$$t = \frac{\overline{X}_w - \overline{X}_e}{\sqrt{SE_w^2 + SE_e^2}}$$

We also need to decide if the test is considered a one- or two-tailed test. If previous experience tells us that mussels on the west coast are always at least as large as on the east coast a one-tailed test is appropriate where H_0 includes no difference between

populations or that the east-coast mussels are larger. Another possible reason for a one-tailed test is that we want to test if it is more profitable to exploit mussels on the west coast because they could be larger. In this case we are only interested in the case where the comparison of the samples indicates that the west-coast mussels are larger, and a one-tailed test is appropriate. On the other hand, if we do not know anything specific about these mussel populations and we test the general hypothesis that the two mussel populations are different in size, then a two-tailed test applies. Finally, we decide that we will retain H_0 if the *t*-statistic falls within the 95% probability area.

5. In our example in Fig. 2 the means from the west- and east coasts are 4.03 and 2.56 cm, the *SD* are 1.0 and 1.3 cm, sample size was 73 mussels so the *SE* were 0.12 and 0.15 cm. If we calculate *t* using the equation above we should get t=7.6. One more thing needs to be known to be able to decide if we will retain or reject H_0. We need to know the degrees of freedom of the *t*-statistic. As before the degrees of freedom is a measure of how many independent data observations remain after the statistic is calculated (here *t*). In this case there where 2*73 independent mussel lengths and we needed two means to estimate the two *SE*, so there are 2*73-2=144 degrees of freedom (*df*). If we look up our *t*-statistic of 7.6 with 144 *df* in a statistical table we will see that it is extremely unlikely to obtain this large *t* under the H_0 distribution. So we reject H_0 and decide that mussels are bigger on the west coast. We have now completed a statistical test a hypothesis about the difference between two sampled populations.

Type I and type II errors in relation to a H_0

When we test a hypothesis by using statistical inference from samples we need to specify a probability where it is unlikely that H_0 is true. By convention we often use the 5% probability level. It is very important to carefully think about all logical possibilities in a statistical test. In one example above we rejected the H_0 that mussels were smaller than 5 cm and we gave the green light for mussel harvest (Fig. 10). When we say that we test H_0 at the 5% probability

	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct decision
Retain H_0	Correct decision	Type II error

Fig. 12. The four possible outcomes of a statistical test.

level we also accept that we will make an incorrect decision in 5% of all tests. After all the blue area in Fig. 10 includes all the high *t*-values that can come from samples when H_0 is true. The error to reject H_0 when it is actually true is called the type I error. There is one more kind of error that we need to worry about. We may also retain H_0 when H_0 in fact is false. This is called the type II error. An overview of the four possible outcomes of a statistical test is shown in Fig. 12. The correct decisions in Fig. 12 do not need any comments. As we have seen above the type I error is fairly easy to understand. Figure 13 summarizes the different features of the type I error.

The magnitude of the type I error is usually given as α, e.g. we have so far used an α=5% or more simply α=0.05. It is important to state tha α you use in hypothesis testing.

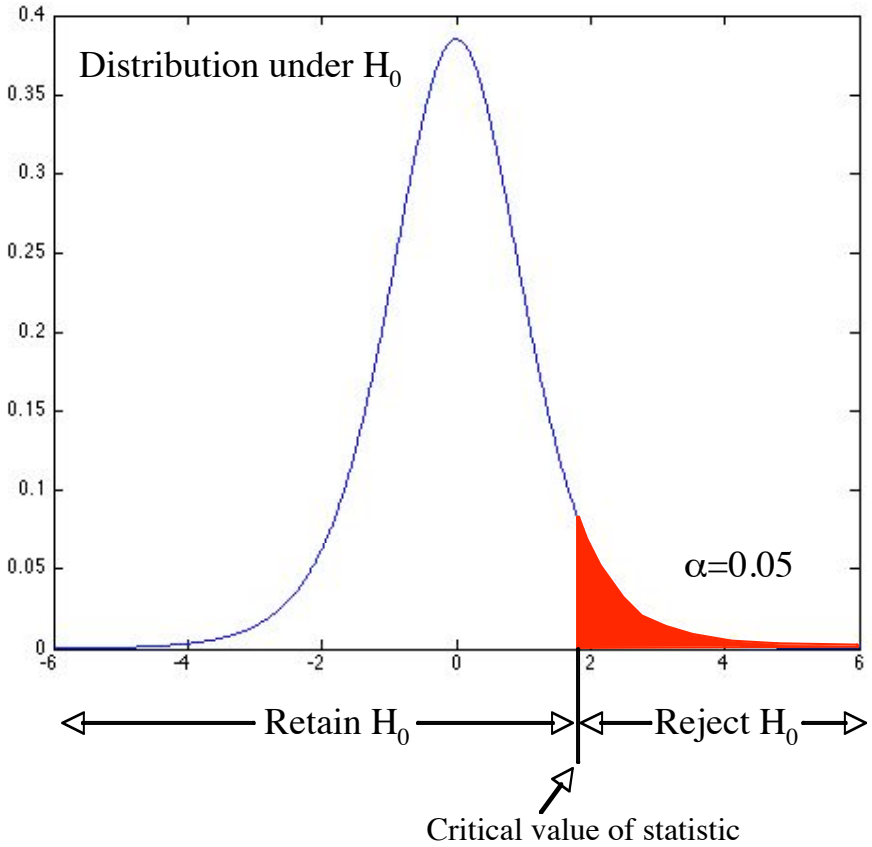


Fig. 13. The distribution of a statistic (here *t* for 7 df) if H_0 is true. The red upper tail indicates the 5% highest *t*-values where we decide to reject the H_0 (for a one-tailed hypothesis). This is the type I error we are prepared to make

The type I error is well known among scientists and normally reported in most scientific studies. Less well known and rarely seen is any information about the type II error. This is unfortunate because there are several arguments why type II error are more serious than are type I errors. Anyway, it is illogical to specify only the type I error and if nothing else is

known they should be similar in magnitude. The reason why the type II error is rarely considered is that it is more complicated to estimate. Again, let us walk through an example to find out more about the type II error.

1. We suspect that the fish in a polluted area may contain more PCB than the allowed concentration limit of 2 mg kg^{-1}.
2. Our model is that fish on average contain more than this limit, and our H_0 is that the PCB content is equal to or less than 2 mg kg^{-1}.
3. We take a sample (as representative as possible) of 16 fish. The sample mean is 2.16 and the *SE* is 0.13 mg kg^{-1}. We then proceed to calculate the *t*-statistic if H_0 is true: (2.16-2)/0.13=1.23. The critical *t*-value for a type I error of 0.05 (one-tailed) is 1.75 so we clearly retain H_0; we conclude that the PCB limit has not been exceeded.
4. All could end here. But wait, could it not be the case that we have falsely retained H_0 when the true population of fish actually have a PCB concentration greater than 2 mg kg^{-1}? This is of course the type II error (Fig. 12). If this error is very large the consequence would be that the study we employed would essentially never detect, in this case, an important and true increase in PCB. And with some knowledge about type II errors this may be evident even before we perform our study. Clearly, such studies are not worth the time nor the money.
5. To estimate the type II error we need something important. <u>We need to specify an alternative hypothesis</u>, i.e. at what concentration of PCB are we prepared to reject H_0. In this case we state that the alternative to H_0 is that the PCB concentration is $\geq$ 2.3 mg kg^{-1}. In other words, we want to know the type II error of retaining H_0 when the PCB content is equal to or above 2.3 mg kg^{-1}.

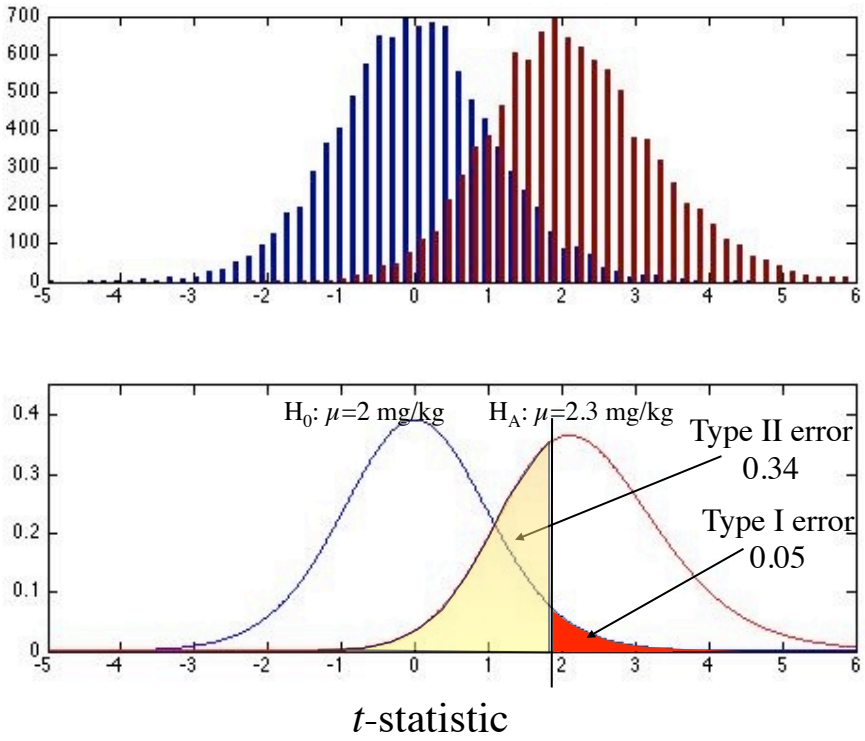


Fig. 14. *t*-distributions for H_0: 2 mg PCB kg^{-1} and the alternative hypothesis 2.3 mg PCB kg^{-1}. Also shown is the type 1 error (red area) and the type II error (shaded area) divided by the critical *t*-value.

6. With this alternative hypothesis we have all that is needed to estimate the type II error. The trick is now to compare the *t*-distributions for H_0 and the alternative hypothesis. In Fig. 14 this is done for 10000 samples in the upper panel. For clarity, in the lower panel we use the asymptotic (when number of samples approach infinity) distributions based on mathematical expressions. Figure 14 first shows the *t*-distribution for H_0 as we are used to (blue curve), with the type-1 error (here α=0.05). On top (in red) the *t*-distribution of the alternative hypothesis is plotted showing the *t*-values expected in a sample of 16 fish if the true population has a mean of 2.3 mg PCB kg^{-1}. Remember that we, of course, do not know the true mean, but if we get a sample where the *t*-value is less than the critical (here 1.75) we will retain H_0. This is a correct decision if the sample comes from a population with 2 mg kg^{-1}. However, if the sample comes from a population with 2.3 mg kg^{-1} all the *t*-values in the shaded region will also lead us to retain H_0 and we commit a type II error. The shaded area in Fig. 14 is the total type II error in our case. The type II error is denoted with β which here is 0.34, meaning that there is a 34% probability that we commit a type II error, i.e. deciding that the PCB limit is not exceeded when in fact the fish contain too much PCB.
7. If the type II error is large we say that the test was not powerful; it had low statistical power. Statistical power is simply 1-β which is the area in the *t*-distribution of the

alternative hypothesis to the right of the critical value (the unshaded area). How can we reduce the type II error, i.e. to increase statistical power of our tests. Statistical power depends on four things:

i) the level of type I error we choose. This is shown in Fig. 15 where we have increased α from 0.05 to 0.1 (compare Figs. 14 and 15) and we move the critical *t*-value to the left with the result that type II error decreases from 0.34 to 0.2 and consequently the statistical power increases to 0.8.

ii) the variance in the true population. Not surprisingly, it will be more difficult to detect an increase in PCB concentration if the true population is very variable. This is often not directly under our control, but sometimes the sample variance is inflated

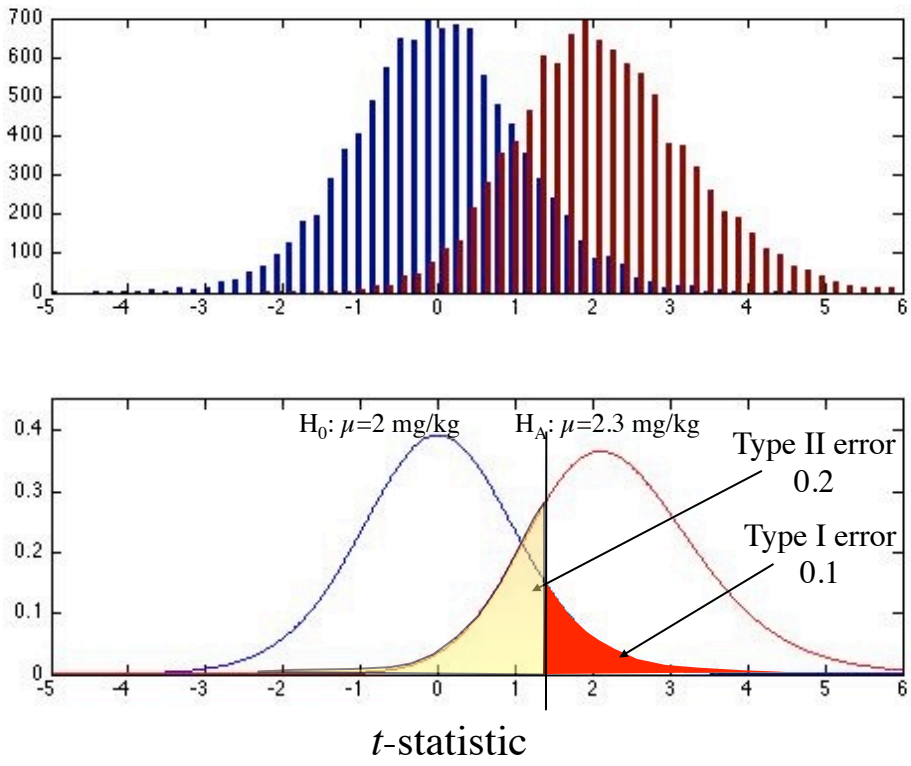


Fig. 15. Effect on type II error and power by changing the type I error.

due to poor sampling or inadequate analytical methods. Figure 16 shows the effect when the true variance is reduced by 33%. The type II error now decreased to only β=0.05.

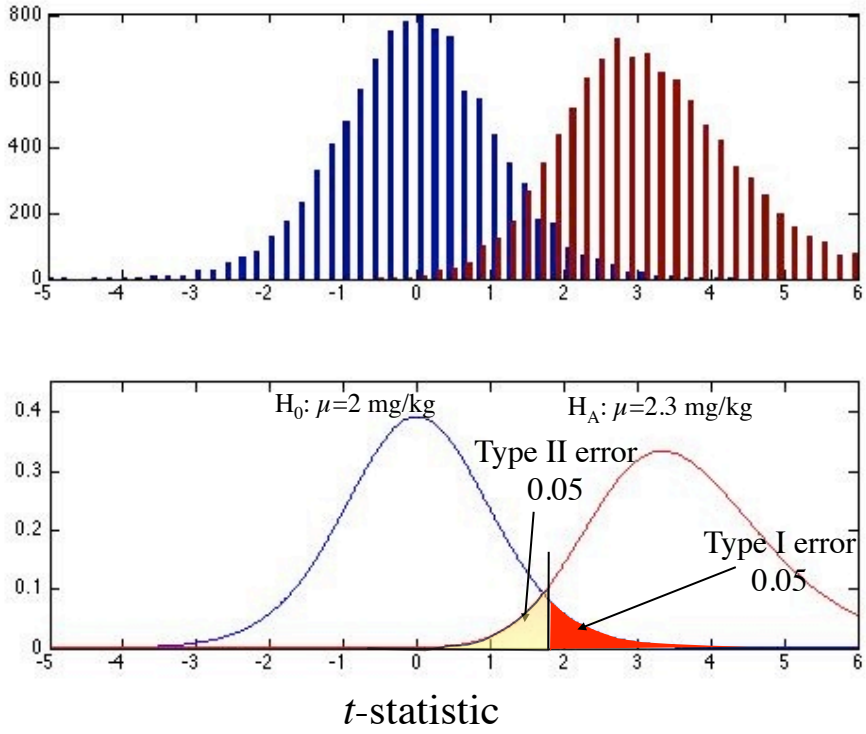


Fig. 16. Effect on type II error and power when the population variance is 33% lower than in Fig. 14.

iii) the so called effect size that we want to detect. Again, not surprisingly the probability will be higher to detect an increased PCB content if the true mean is even higher than in our alternative hypothesis of 2.3 mg kg^{-1}. If we change our alternative hypothesis and are content to detect the PCB concentration when it exceeds 2.5 mg kg^{-1} (Fig. 17) this will substantially reduce the type II error and we have a very powerful analysis to detect this greater change (β is only 0.012). The effect size that we want to detect is our choice. There may here be a compromise between available resources (time & money) and how small effect sizes we can expect to detect.

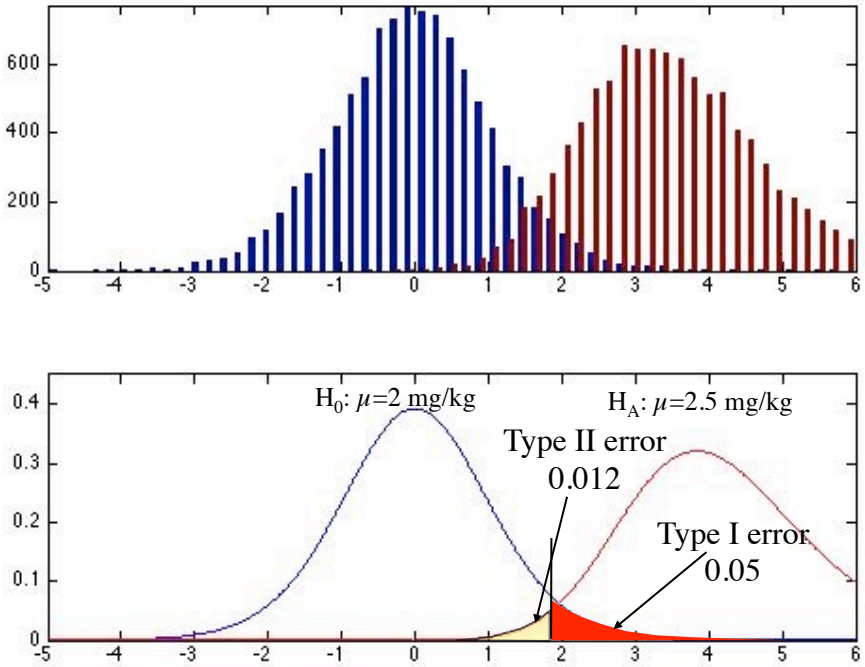


Fig. 17. The effect on type II error and power when the alternative hypothesis is changed to allow for a larger effect size (here from 2.3 to 2.5 mg PCB kg^{-1}.

iv) the final thing that affects statistical power is the sample size. Sample size is also under our control and as expected the bigger the sample size (or actually degrees of freedom) the less risk to commit a type II error. Adjusting the sample size is the most common way of controlling the type II error. In Fig. 18 the effect on type II error and power of
increasing sample size is clearly shown.

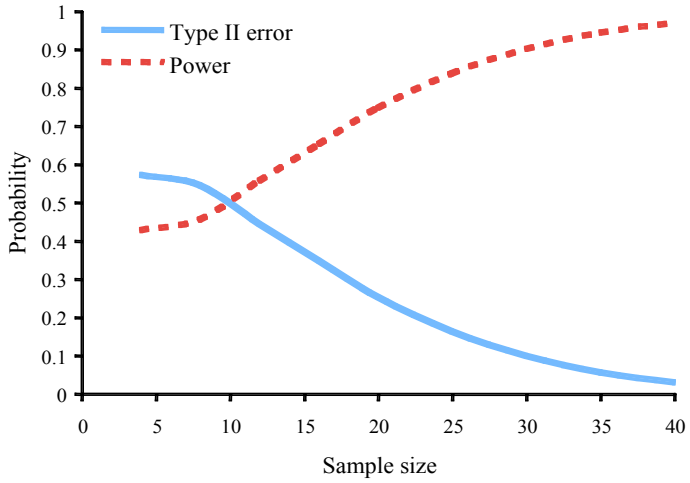


Fig. 18. The effect on type II error and power when sample size increases.

What level of statistical power is acceptable?

Because it is still rare to report the type II error no convention has developed about what level of statistical power is acceptable. Many argues that statistical power should be at least 0.8 ($\beta \leq 0.2$). However, logically, there are no reasons why we should by default select different levels of type I and II errors. In applied research, e.g. test of a new medical drug, it is common that the expected risk of committing type I and II errors are set similar or even that the type II error is lower. Risk is here the cost of making a wrong decision times its probability. When testing for harmful side-effects in a medical drug the costs of a type II may be much higher than a type I error. If we commit a type I error this means that we reject the H_0 of no side-effect when there in fact was no side-effect. The consequence may be that the drug developing program is terminated which will certainly involve a loss of money, but likely very much less than if we commit a type II error. In the case of a type II error we decide that there is no harmful side-effect when the drug in fact is harmful. No action is taken and the drug will maybe harm many people before this side-effect is discovered. As many examples show, this will be extremely costly and may even cause the close-down of a company. Here we should obviously set the type II error a lot lower than the type I error. Even in basic research there is logical arguments to care more about the type II error. In the logical loop in Fig. 1 we can see that when we reject H_0 this is taken as support for our model which we continue to develop and elaborate with tests of more hypotheses. If the model in fact is wrong we will probably discover that later. All that is lost is time. On the other hand if we retain H_0 when it is wrong we will abandon our model and it may take a very long time before we or others return to further testing of the model.

Important assumptions for parametric statistics

We have now learnt about how to test hypotheses by using the *t*-statistic. Before we continue our journey through the landscape of experimental design & statistics it is useful to stop a moment to consider any limitations of the approach we have used so far. The *t*-test belongs to a family of statistical techniques known as parametric statistics because they are based on estimates of the population mean and variance from samples. The other two methods we will encounter, linear regression and analysis of variance also belong to his family. For parametric statistics there are 3 major assumptions:

1. True populations are normally distributed. The parametric methods are exactly valid only when samples come from normal distributions. However, as we saw previously, the Central Limit Theorem ensures that sample means are approximately normally distributed regardless of how the true population is distributed. This fact makes most parametric methods surprisingly robust even for large deviations from the normal distribution.
2. The true populations included in a test of a hypothesis should have the same variance. This requirement is called homogeneous variances or in statistical jargon,

homoscedasticity. The requirement of homogeneous variances is more important than that of normal distribution, and when more than one sample is included we should test if the variances are sufficiently similar to allow a parametric test. We will come back to this later.

3. Independent samples. This is a very important assumption and is essential for all statistical tests. To ensure independence of samples is at the heart of sampling and experimental design and we will discuss this a lot during the course. A simple (and common) example of non-independent sampling is when we want to sample mussels to test if they increase in size in a selected locality. One sample of mussels are collected in May, their lengths are measured, they are individually marked and then returned to the sea. In August the same mussels are collected again, measured, and a *t*-test is performed to test the hypothesis of growth. Here the measurements in May certainly influence the measurements in August and the two samples are clearly not independent, they are after all from the same individuals. In this case the expected effect is that we underestimate the true variance in growth rate; we will get a too large *t*-value, and have a greater type I error than we think with the result that we are more likely to reject the H_0 when it is true.

From samples to experiments

So far we have only considered model and hypotheses that required one or two samples from field populations, e.g. mussels and fish. The rest of these days will be focused on a different method called manipulative experiments. In such experiments we change some aspects of a biological/ecological system in a controlled way. A correctly performed experiment is the most powerful scientific method to distinguish between different models and their predictions. The same basic principles for sampling and estimating the parameters of the true population apply to experiments. Let us take an example. We have a model (from observations) that the sea star *Asterias rubens* can sense some chemical compounds released by a predatory sea star, *Marthasterias glacialis*, making it possible for *A. rubens* to escape. To test this model we perform a manipulative experiment. We prepare a chemical extract from seawater where several *M. glacialis* have been living. We add this extract from pipettes in five aquaria with *A. rubens*. We also add an extract from seawater without *M. glacialis* to another 5 aquaria with *A. rubens*. Then the crawling velocities of all the *A. rubens* are measured and we test the H_0 that the means with extracts with and without *M. glacialis* are not different. Although slightly more abstract we can imagine a true population of crawling velocities when *A. rubens* is exposed to a *M. glacialis* extract. Our 5 aquaria is a sample of the much larger true population of aquaria containing *A rubens* and extracts. We again expect the means to be normally distributed, and under the conditions that the variances are homogeneous and all the aquaria are independent this allows us to use parametric statistics to say something about the real world from the outcome of experiments.