Experimental design and statistics for Marine Ecology

Part 2

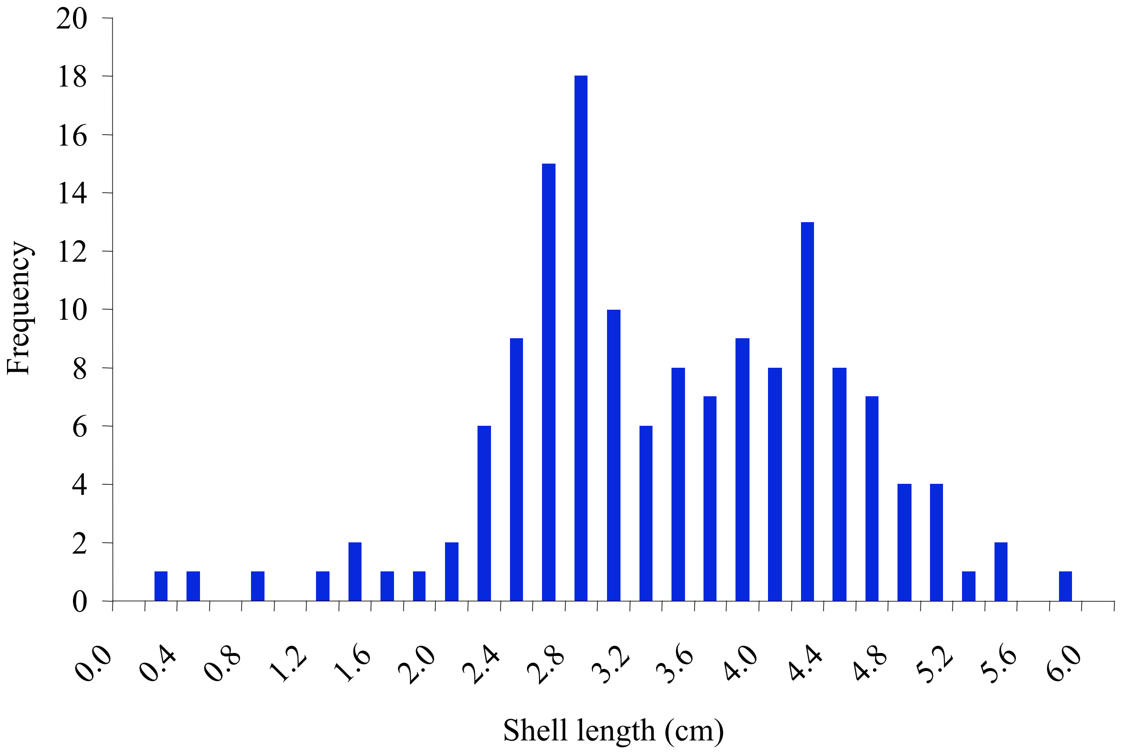
STATISTICAL ANALYSIS AND DESIGN OF ECOLOGICAL EXPERIMENTS - STATISTICAL MODELS AND REGRESSION

Linear statistical models

After the introduction in Part 1 about distributions, tests of hypotheses and statistical errors, this Part 2 will introduce a few more concepts that will be helpful in Part 3 about analysis of variance.

We start with an example recycling the mussels in Part 1. Mussels are sold at an outdoor market place and from the large variability of shell lengths we formulate a model that mussels from both the east- and west coast are sold. The frequency distribution of the shell lengths in the sample is shown in Fig. 1.

Fig. 1 Frequency distribution of mussel shell lengths sold at a market.



We can start and look at the total sample (n=146) and calculate the mean, the sum of squared deviations, called Sum of Squares (SS), and the variance which is $SS/(n-1)$. Sum of squares is calculated as:

Written by Per Jonsson 1995, modified by Mats Lindegarth 2001 and by Per Jonsson 2008

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

We find that the sample mean ($\overline{X}$) is 3.3 cm, *SS* is 149 cm^2 and the sample variance (s^2) is 1.03 cm^2. We can write this as a linear statistical model:

$$L_i = \mu + e_i$$

Where L_i is the length of each mussel, μ is the overall true mean and e_i is the deviation (*e* stands for “error”) from the mean for each mussel. The subscript i=1,2,3...*n* are labels for each individual mussel. We could now extend the linear model in order to explain more of the total variation. The research model that the mussels sold at the market are a mix from both coasts can be included as a factor in the linear statistical model:

$$L_{ij} = \mu + Coast_i + e_{ij}$$

Where *Coast* is a factor with two levels, east coast (i=1) and west coast (i=2). The subscript *j* now indicates each individual mussel. The factor *Coast* potentially gives a specific contribution to the mussel length, i.e. the factor *Coast* potentially explains some of the variation in length. We apply a genetic analysis capable of assigning all the mussels to one of the two coasts. It turns out that the west- and east-coast mussels in the sample have means of 4.0 and 2.6 cm, respectively. In our linear model that determines the effect sizes of the factor *Coast* where $Coast_{i=1}$ = -0.7 cm and $Coast_{i=2}$ =+0.7 cm. With the means of the two levels of factor *Coast* we can now calculate a new *SS*. Instead of taking the sum of squared deviations from the total mean we now do it for all the east-coast mussels against the mean of east-coast mussels ($\overline{X}_{i=1)}$ and for the west-coast mussels against the west-coast mean ($\overline{X}_{i=2}$). The *SS* for the east coast is 33.4 and for the west coast 36.4. This sums up to a *SS* of 69.8. This means that by including the factor *Coast* we have reduced the unexplained variability from *SS*=149 to *SS*=69.8. With the factor *Coast* we now explained (149-69.8)/149=53% of the total *SS*. In principle we can now proceed to elaborate our research model with factors that may explain the remaining part of the unexplained variability, i.e. the remaining 69.8 *SS*. One example could be that age explains part of the variability in shell length. Determination of the age can be done by counting growth rings in the shell. The linear statistical model is now extended to:

$$L_{ijk} = \mu + Coast_i + Age_j + e_{ijk}$$

It turns out that when the *SS* is calculated for all mussels to each age mean within each coast the remaining *SS* is now only 23 so we have now explained 85% of the total variability with our research model about coastal origin and age. The remaining *SS* of 23 is called the residual *SS* and represents the unexplained variability caused by all the unknown factors we have not yet included in our research model. Examples of such unknown factors are genotype, food availability temperature etc. This method of partitioning the variance among hypothesized factors using a statistical linear model is a very powerful tool.

Factors

Let us look closer at the concept of statistical factors. A factor is here a variable that potentially explains (and sometimes causes) the variation we observe. A factor has two to an infinite number of levels (Table 1).

Table 1. Example of factors and their levels.

Factor	**Levels**
Coast	east coast, west coast
Age	1-year, 2-3-year, above 3 years
Temperature	8, 12, 17, 23°

There are two types of statistical factors, fixed factors and random factors. This is a difficult concept and needs careful explanation. A fixed factor includes all the relevant levels necessary to test a research hypothesis. A random factor only includes a sample of all possible levels. Here are some examples of fixed and random factors. We have identified *Season* as a potentially explaining factor for the growth rate of the knotted wrack (*Ascophyllum nodosum*). Our hypothesis is about seasonal effects and we include all the seasons: summer, autumn, winter and spring, and the factor is clearly fixed. In another study we have a hypothesis about temperature as an important factor to explain growth rate in a ciliate species. We know that the normal tolerance range for this ciliate is between 5-25°C, and we draw a random sample of 4 temperatures in this range to represent the factor *Temperature*. The levels happened to be 11, 13, 15 and 18°C. If we would repeat the experiment we would very likely have another set of levels. The factor *Temperature* is random. A good rule of thumb is to ask the question: Would I choose the same levels if I would repeat the study? If the answer is yes the factor is fixed and otherwise random. Note that, e.g. *Temperature* could be a fixed factor in another experiment. An example being a test if spawning of cockles is dependent on temperature. We include 5, 10, 15 and 20° and these levels represent low, medium, high and very high temperatures. If we repeat the experiment we would select the same levels and *Temperature* is a fixed factor. An obvious question is why it is important to distinguish between fixed and random factors. We will see later that this has consequences for how we statistically test if the proposed factors explain more of the variance than specified by H_0 of no effect. And equally important, the logical conclusions from fixed and random factors will be different. The logical conclusion from a random factor is more general since it applies to all possible levels, while conclusions from a fixed factor only apply to the levels we have selected.

If we look back to our mussel example *Coast* was obviously fixed. The hypothesis was only about Sweden and there are only two coasts on this scale so all possible levels were included. We also identified the two coasts because we know that they represent known differences of environmental conditions that could explain the difference in shell size. With our model we wanted to test if *Coast* has a treatment effect on shell size. In our case we found that west coast mussels were on average 4.0 cm and east coast mussels 2.6 cm. The overall mean was 3.3 cm and we can say hat the treatment effect caused by *Coast* was 0.7 cm. If *Coast* has level *east* it subtracts 0.7 cm from the overall mean, and if *Coast* is *west* the factor adds 0.7 cm to the overall mean. This is how fixed factors work. How would this example look like if the factor *Coast* were to be random? In this case we can imagine that we have a general idea that mussels from different localities (stretches of coast) differ in length. We do not at this stage have any idea why they differ so these localities do not represent anything more than different places. Randomly, we draw 4 stretches of coast and within each stretch we sample a number of mussels. Had we repeated this we would have another set of coastal stretches. The factor *Coast* is here clearly random. The linear model looks the same:

$$L_{ij} = \mu + Coast_i + e_{ij}$$

We further assume that Coast explains the same amount of *SS* as in our fixed factor case. The difference between these two cases is how we logically interpret the results. As in the case with the fixed factor, the random factor *Coast* explained 53% of the total *SS*. But it would be meaningless to talk about a treatment effect since there are an almost infinite number of coastal stretches that do not represent anything else than that they are different places. Instead we say that the random factor *Coast* explained 53% of the variation in mussel length. We do not attempt to link any treatment effect to the different levels (coastal stretches). Why do we use random factors? There are two main reasons:

1. <u>Increase the generality of our research models</u>. We have a research model that states that snails grazing on the bladder wrack (*Fucus vesiculosus*) induce the wrack to produce a chemical compound defending the wrack from further grazing. This model is tested in an experiment where the fixed factor *Grazing* is manipulated by the levels presence and absence of snails. We indeed conclude from the experiment that the defence compound increases when snails are present. However, we only performed the experiment in one place. We now want to extend our model to include the whole geographic distribution where the wrack and snails co-occur. To do this we include the random factor *Place* and randomly sample 4 localities within the distribution area. The new H_0 is now

that the induction of the defence compound does not differ among localities, i.e. *Place* does not explain any significant *SS* of the variation in the induction of defence compound. We do not have any hypothesis about the different localities, they just represent a sample of different places to allow us to draw a more general conclusion, in this case about chemical defence.

2. A tool to explore patterns of variation. In an early phase of a scientific inquiry we usually have little information about factors potentially influencing the abundance of a species or a biological process. Before having any specific hypotheses about, e.g. factors like temperature, salinity and nutrient levels we may explore the more simple hypothesis if these factors explain any significant amount of the variation and also the relative magnitude of these. In this example we sample the levels for the factors *Temperature*, *Salinity* and *Nutrients* (but probably using some criteria, e.g. tolerance limits or commonly occurring ranges). We find that the *SS* explained by these factors are 300, 150, and 10, respectively. The residual unexplained *SS* is 30. This means that of the total *SS* temperature explains 61%, salinity 31% and nutrients only 2%. We can conclude that we should now focus on mechanisms that include temperature and salinity, probably leading to the identification of fixed factors.

Linear regression

We can further illustrate the use of statistical linear models by introducing the method called linear regression. Often we are interested how a factor with continuous levels may explain the variability in some studied system. As an example we propose a model that food abundance (e.g. microalgae) determines growth rate in a copepod. The data in Table 2 suggests that there is a relationship between food abundance and growth rate. In this case we have a fixed factor *Food* with 4 selected levels and we record the response in terms of growth rate. The factor *Food* is fixed because we let the levels represent more or less equal intervals within a food range normally encountered in the field.

Table 2. Growth rate of a copepod at different food levels.

Food abundance (µg Chl l^{-1})	Growth rate (µg day^{-1})
3	7.6
5	19.3
10	25.8
15	31.5

There are two things of interest here:

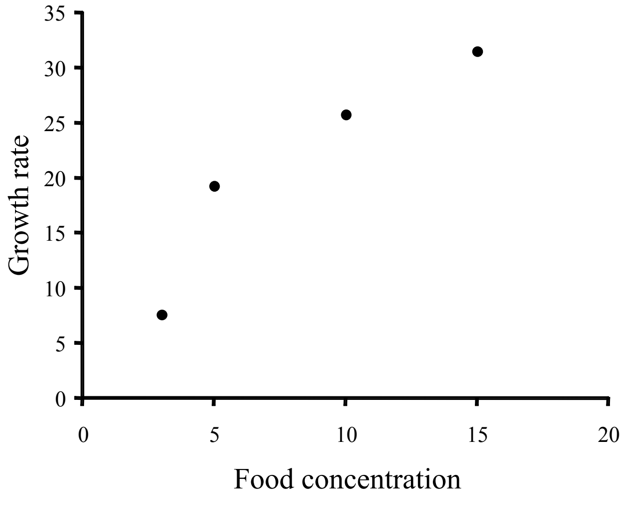
1. Test the H_0 that there is no relationship between the factor levels and the response (growth rate).
2. Describe the relationship numerically, i.e. to find an equation predicting growth rate if we know the food abundance.

A linear regression analysis addresses both these tasks. A linear regression analysis attempts to fit a straight line that best represents the relationship between the factor levels and the response variable. Often the factor levels are called the independent variable and the response is called the dependent variable, the logic being that we control the independent variable (a fixed factor) causing the dependent variable to respond.

A linear regression tests the hypothesis that there is a functional relationship between two variables *X* and *Y*. With a function we generally mean that a selected value of *X* determines the value of *Y*. Even if it is not always clear that *X* really causes *Y*, a statistical regression requires that we control the levels of the factor *X*. In the example above we control the levels of food, and we further assume that these levels do not have any variance (error), i.e. they have fixed values. In practice this is rarely true and this requirement can be relaxed by stating that the variance of *X* should be much smaller than the variance of *Y*. If we repeat the study in Table 2 we will find that for the same food level the dependent variable growth rate will differ each time.

In Fig. 2 the data in Table 2 are plotted in a graph. It is quite clear that as food increases so does growth rate. The growth rates in Fig. 2 can be viewed as a sample from a true population of growth rates represented by the frequency distributions shown in Fig. 3.

Fig. 2. Data on growth rates of a copepod offered different food concentrations.



We do not know the underlying frequency distributions, but the regression analysis assumes that these are normal distributions with the same variance (homogeneous variances).

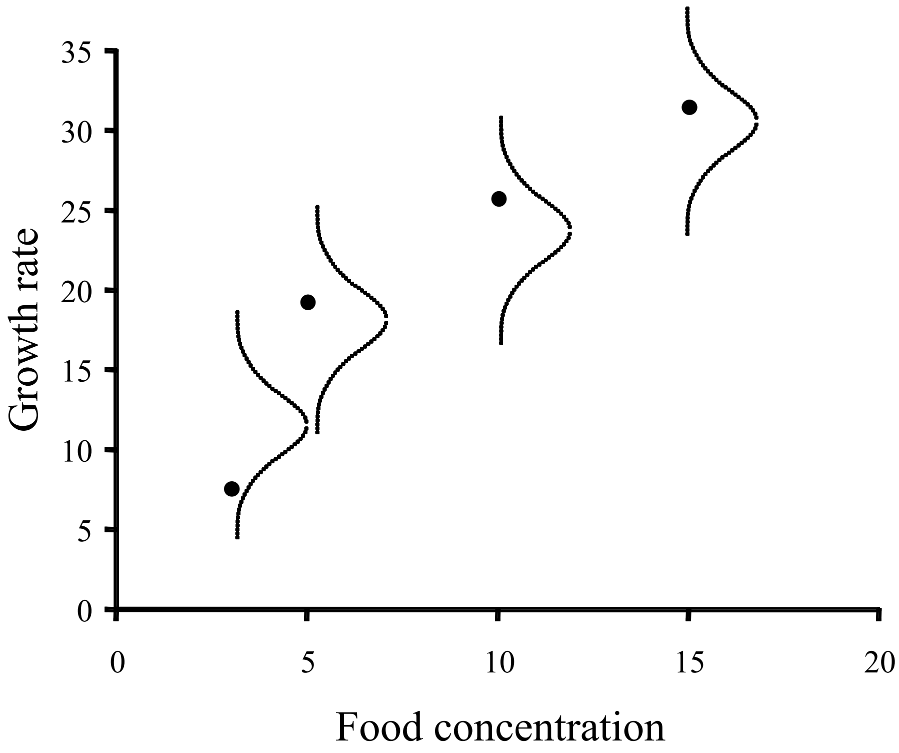


Fig. 3. Same data as in Fig. 2 but with assumed normal frequency distributions generating the sample in Table 2.

The straight line we will fit to the sample of growth rates in Fig. 3 may be viewed as a statistical model to explain some of the variance in growth rates. The linear model is:

$$Y_i = a + b*X_i + e_i$$

Where a is a constant determining the intercept of the regression line (growth rate when food concentration is 0), b is a constant determining the slope of the line (indicating how strongly growth rate depends on food), and e_i is the deviations between the regression line and each sampled growth rate, representing the unexplained or residual variability.

Now let us perform a partition of variation, in our case with copepod growth rates. As in the previous case with the mussels we can ask what is the total variability for the growth rates. This is measured by the *SS* to the mean of growth rates as shown in Fig. 4.

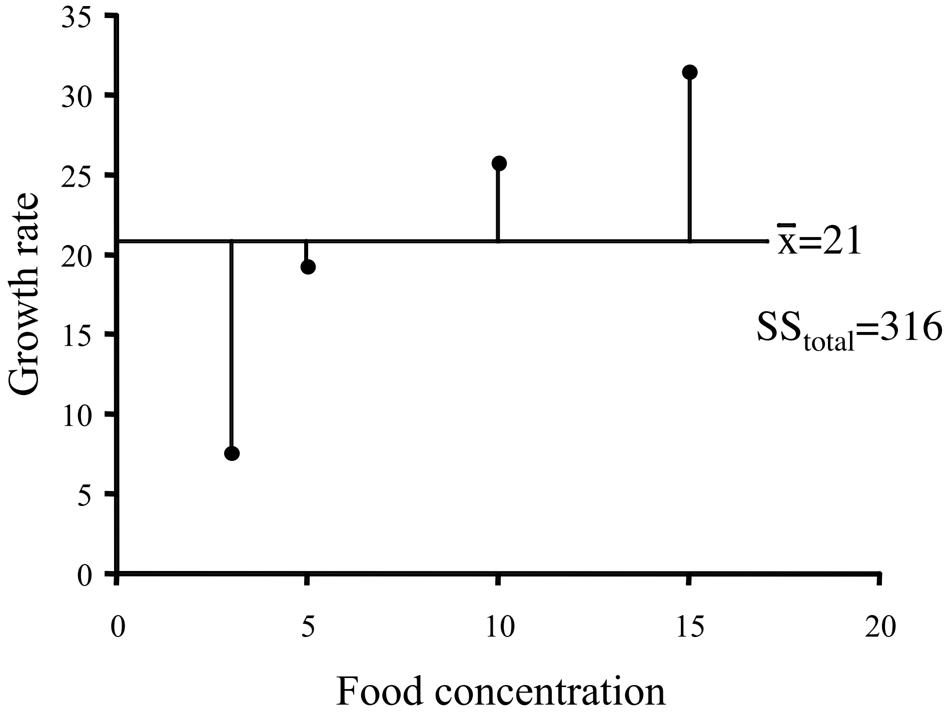


Fig. 4. The horizontal line represents the mean growth rate, and the vertical bars are the deviation from the mean to individual sample values. Also shown is the sum of squares (*SS*) of these deviations measuring the total variability.

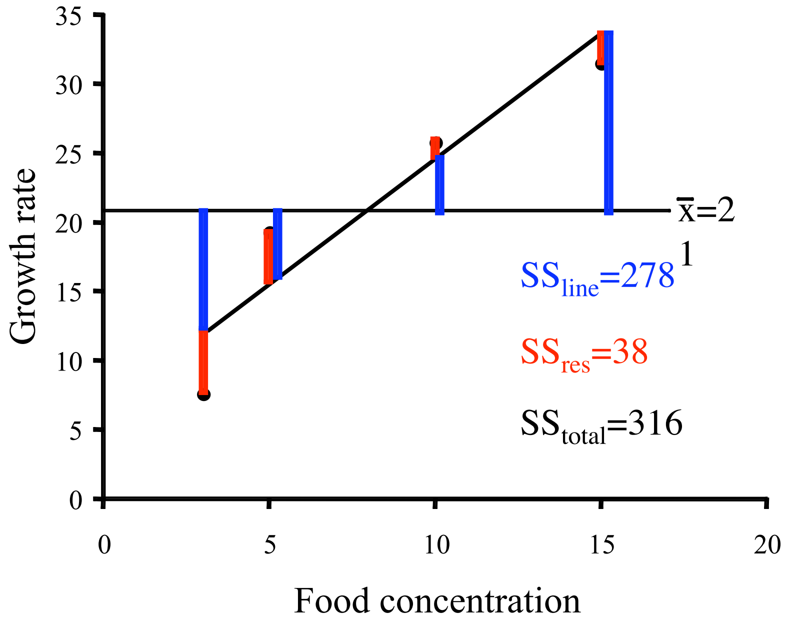


Fig. 5. The fitted regression line with the explained *SS* (278) and the residual *SS* (38).

Some of the SS_{total} of 316 can now be explained by the linear model above by fitting a straight line. The constants *a* and *b* are estimated to maximize the *SS* explained by the line (called the least square fit). In Fig. 5 the regression line is plotted together with the *SS* explained and the residual *SS* of the variability not explained by the line.

As pointed out in the beginning of this section we can determine the equation for the fitted line to allow prediction of growth rates for new levels of food concentration. The equation of the line is determined by the constants *a* and *b* and can be written as:

$$Y=1.79*X+6.3$$

We can also calculate how much of the total *SS* the line explained which in this case is 278 / 316=0.87 (or 87%). This quantity is called the coefficient of determination or simply r^2 ("r-square").

So far we have only described the relationship between food concentration and growth rate. Now we turn to the actual test of the hypothesis that food concentration affects growth rate. H_0 is in this case that there is no relationship. We will again use statistical inference and the statistical H_0 specifies the probability distribution of some statistic when there is no true relationship, i.e. when the slope *b* is zero. We have previously used the statistic *t* and we will now introduce a new statistic called *F*, a statistic that will be the main focus on the rest of the course.

Analysis of variance and the F-ratio

Previously we have calculated the *t*-statistic from the means of two samples and asked if the *t*-value is so large that it is unlikely that the two samples come from the same true population. If the *t*-value exceeds the critical value for the type I error that we have specified (e.g. α=0.05) we reject H_0. We can do a similar trick by using the sample variance. Assume a known true population that is normally distributed. If we take two samples, calculate the sample variance and then form a ratio, s_1^2/s_2^2; we expect that this ratio on average will be 1. Due to the fact that we will most of the time draw non-representative samples most ratios will deviate from 1, and the smaller the samples the more likely that we get large deviations. In Fig. 6 we can see the result of many such ratios of two samples (*n*=10) coming from the same true population. The ratio is called the *F*-ratio and the probability distribution it follows is called the *F* distribution (*F* after the famous statistician Ronald Fisher). As expected the peak of the distribution is close to 1. Also note that the distribution, of course, cannot be less than 0. One more thing, there was one *t*-distribution for each number of degrees of freedom (*df)*. The *F*-distribution is determined by two *df*: the *df* of the variance in the numerator of the ratio and the *df* of the denominator of the ratio. This is often presented as:

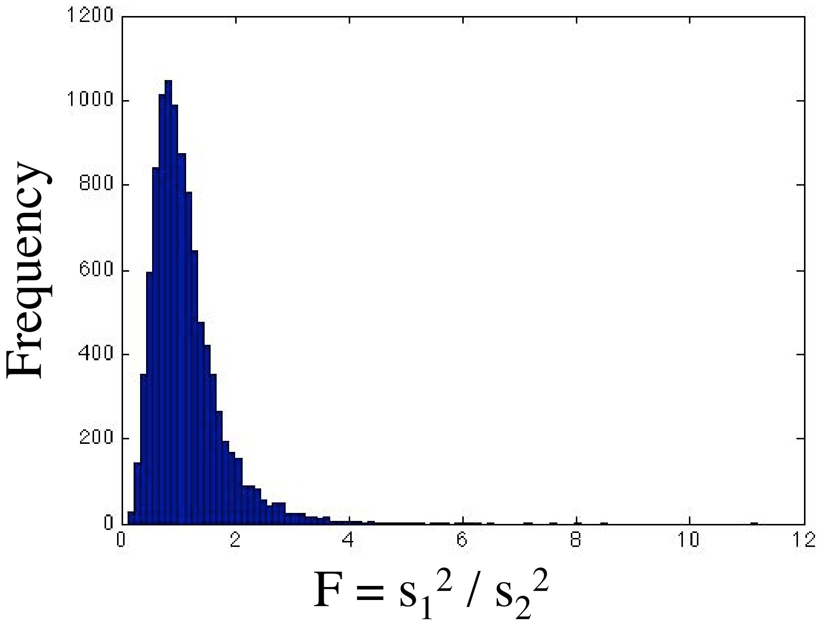


Fig. 6. Frequency distribution of the *F*-ratio between variances of 2 samples (*n*=10).

$$F_{(df1,df2)} = \frac{s_1^2}{s_2^2}$$

So how can we use this piece of knowledge? We are now ready to take a large step and carry out the first Analysis of variance, or in short: ANOVA. We resume our regression analysis from above. Table 3 below shows an ANOVA with the aim to test *H0* that there is no relationship between food and growth rate for the studied copepod. So what does Table 3 contain?

1. The first column shows the two sources of variation we have partitioned, and also their total sum of variation.
2. The second column contains the *SS*.
3. The third column contains the degrees of freedom (*df*) for the two sources of variation. The concept of *df* is still perhaps difficult but is as before: the number of independent observations (or numbers) minus the number of parameters needed to estimate that source of variation. For the residual *SS* we take the squared deviations between 4 independent growth rates and the line, where the line is determined by two parameters (slope and intercept), so the *df* is 4-2=2. The *SS* for the regression line depends on the two independent parameters (slope and intercept) and the *SS* is calculated from this line to the mean of growth rates, so *df* is 2-1=1. The *df* of the total *SS* is simply the number of growth rates minus their mean, i.e. 4-1=3. Note that the *df* for the different sources of variation must add up to the *df* of the total variation. It is good practice to always check that the *df* add up.

Table 3. Analysis of variance (ANOVA) for the test of regression of growth rate against food concentration.

Source of variation	***SS***	***df***	***MS***	***MS estimates***	***F***
Regression line	278	1	278	$\sigma_e^2 + \beta^2\Sigma x^2$	14.5
Residual	38	2	19	σ_e^2	
Total	316	3			

4. The fourth column labeled *MS* for Mean Square contains the standardized *SS* which is found by dividing the *SS* with the *df*. The reason we calculate the *MS* is that we do all this to arrive at a statistic that is based on the ratio between two variances (the *F*-ratio!). A variance is the *SS* divided by its *df*, essentially the *SS* is standardized by the number of independent deviations.
5. The fifth column is fundamentally important and is at the heart of any analysis of variance. If we start with the residual *MS*, what is it that we really calculate? The residual *MS* estimates the unexplained variance after the regression line has been fitted (see Fig. 5). We then walk up one notch to the source of variation of the regression line and ask what does this *MS* estimate if there is no true relationship between food and growth rate, i.e. the H_0 is true. If the true slope is zero most slopes will still deviate from zero due to non-representative samples of growth rates. The line will vary up and down from one experiment to the next. The variation in slope only depends on the residual variance, i.e. differences in growth rate from one food level to the next (when food level has no effect). Figure 6 shows how 100 sampled slopes look like when there is no true relationship.

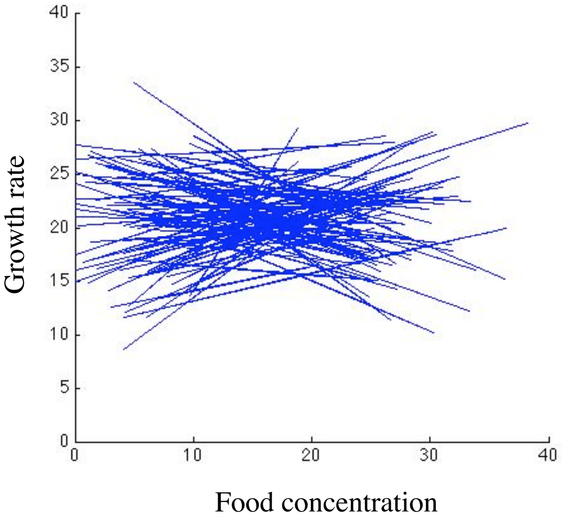


Fig. 6. Samples of 100 slopes when there is no true relationship between food and growth rate.

6. In fact, The *MS* for the regression line also estimates the residual variance (σ_e^2) when there is no true relationship. So, under H_0 the *MS* for the regression line and the residual *MS* estimate the same thing, and the ratio between them is expected to be close to 1. This is of course the *F*-ratio. Now note, that when there is a true relationship and the slope $b > 0$ there is an added component of variance ($\beta^2\Sigma x^2$) that will increase the *MS* for the regression line. Clearly, as the relationship between food and growth rate increases, the *F*-ratio is also expected to increase. The main question is now how great the *F*-ratio should be to make us reject H_0. As in the case with the *t*-statistic, there are tables of the *F*-statistic where we can look up the probability to obtain a particular *F* if H_0 is true. In our case the *F*-ratio is 14.5 with 1 *df* in the numerator and 2 *df* in the denominator. The probability of getting this *F* when H_0 is true is 0.06 and we consequently retain H_0 that there is no effect of food on growth rate. From the inspection of Fig. 2 we had probably expected a relationship, but obviously our test of the hypothesis had poor statistical power because we had so few data points, or in other words, there were few degrees of freedom for the statistical test. The risk that we have committed a type II error seems rather great. Had we performed some estimate of the statistical power before the experiment this may have made us include some more measurements of growth rate.

Also note that the ANOVA above was inherently one-sided. The alternative to H_0, i.e. that there was a relationship, could only increase the *F*-ratio. This is a very useful property of *F*-ratios that we will exploit in the next part.

Assumptions for a regression analysis

We conclude this part with a summary of assumptions underlying the regression analysis.

1. As before all measurements (the *Y*-values) must be independent. In regression analyses this is often negelected, e.g. growth rate is estimated from several measurements on the same individual at different times.
2. The variance of the measurements should be homogeneous. Often, in regression analyses only one data point for each value of *X* is collected and it may be difficult to evaluate if the variance is similar in the whole range of *X*. Many biological data are expected to show heterogeneous variances and this is a serious problem in regression analyses and may require statistical expertise.
3. As pointed out before, the regression analysis assumes that *X* is controlled by the researcher and that the variance of *X* is much less than variance in *Y*. If this

is not the case and both *X* and *Y* are sampled with some error you should contact statistical expertise.

4. The regression analysis is exactly valid only if the data come from normal distributions (see Fig. 3). Again, if only one data point for each *X* is measured this is difficult to evaluate. If the data are known or expected to deviate strongly from a normal distribution it is possible to exploit the Central Limit Theorem by measuring several *Y* for each *X* and then use the mean of the *Y:s* which will tend to be normally distributed.