The Quantitative Analysis of Natural Populations:
Some Common Statistics and What They Mean

Valentin Schaefer and Andrew Elves 2018

Population Sampling

A natural population consists of individuals of the same species occupying a continuous space. The size of this space is arbitrary and depends upon the interests of the person investigating the particular subject or species.

A population of organisms has many properties. The organisms themselves have height, weight, colour, appendages, metabolic rates, and so on. These are physical and physiological properties. Another type of category would be behavioural properties. These would include such things as diet and habitat preferences (temperature and humidity, for example).

If you are a researcher and are interested in knowing a particular property or parameter of a population, you would ideally go into the population and examine each individual for that property. However, natural populations generally have too many individuals, or the individuals are too spread out or elusive to do such an exhaustive survey.

As a result, the numerical characteristics distinguishing a population (parameters) are inferred by using a random sample taken from that population. This sample is assumed to be representative of the population as a whole. Figure 1 shows the localities from which samples were collected for a study of fish populations in the Fraser valley.

The kind of technique used in sampling a population depends on the type of organism, the habitat, and what it is you want to know. For example, if you wish to determine the density of rodents or grasshoppers, a likely sampling technique is the mark-and-recapture method. For many types of vegetation analysis, the quadrat method is appropriate.

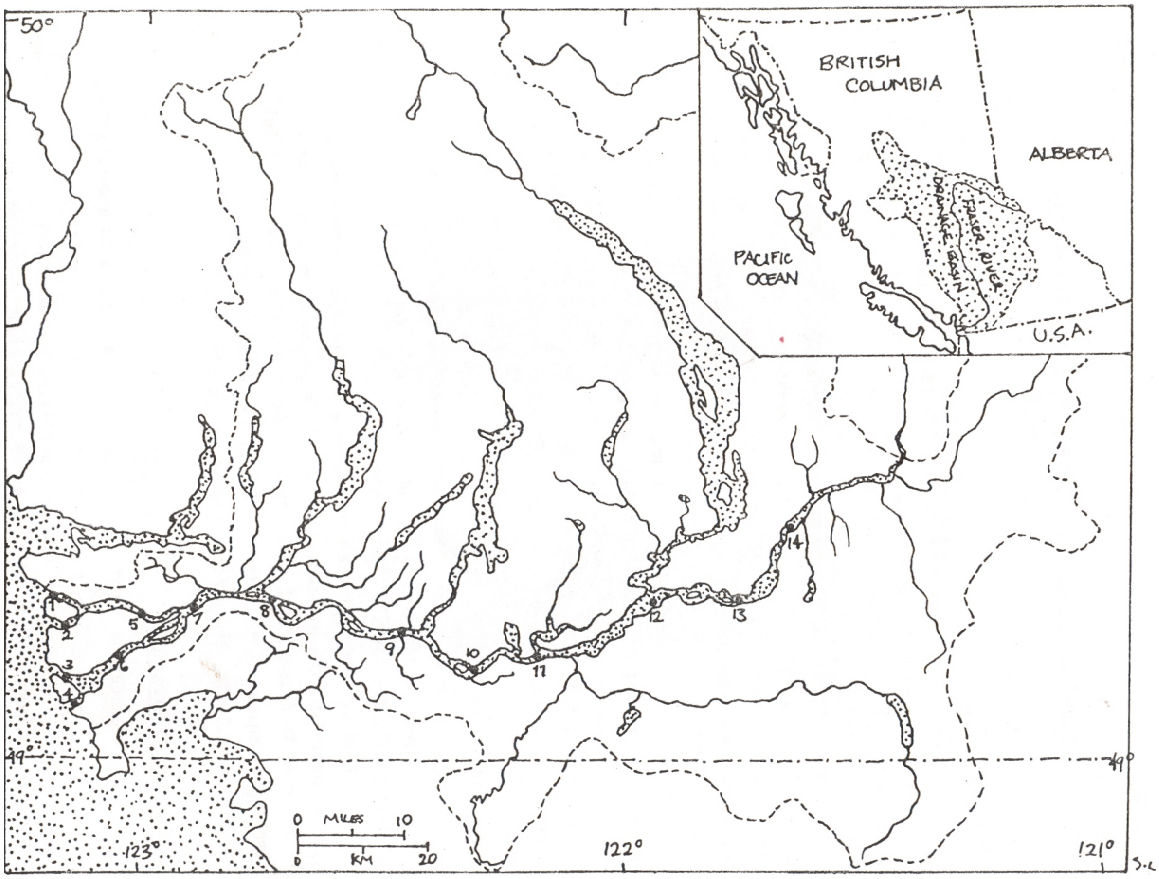


Figure 1. Locations of fourteen fish sampling stations on the lower Fraser River. The broken line shows the major part of the lower drainage basin. The total drainage basin within British Columbia is shown in the inset.

Descriptive Information

If you measure the lengths of a number of individuals in a population, and you feel that your sample is representative of the population or the species, you can say that the average or mean (calculated by summing all the observations of a sample and dividing by the number of observations: the mean is represented by the symbol X), is the average for the species. So, if you have measured the lengths of a sample of red foxes, you can say that the red fox (*Vulpes fulva*) has an average length of perhaps 89 cm. You may also mention the maximum and minimum lengths you measured, called the range, and say that the red fox varies in length from 81-102 cm. Such measurements (the mean and the range) are useful for describing the appearance of a population or species. They may also be illustrated graphically as shown in Figure 2. Here, the means and the ranges for the functional responses (numbers of cocoons opened for food) of three deer mice (*Peromyscus maniculatus*) are illustrated.

The **mean** or average is one measure of central tendency of a characteristic for a population. It is one way of describing what most of the individuals would be like. Other measures of central tendency are the median and the mode.

The **median** is the value for the individual in the exact middle of the population when arranged in order from least to most. The median is the value for which half the population has a lower value and the other half has a higher value.

The **mode** is the value for a population characteristic that is possessed by the greatest number of individuals compared to other values.

You may also be interested in determining if there is any seasonal variation in the abundance of a species throughout the year. Such quantification is illustrated in Table 1. In this case, the occurrence of two species, A and B, in pitfall traps, as well as the percentage of the traps which contain individuals, is recorded. You can examine such a table and conclude that species A is less abundant in winter and spring. You may even speculate that this species is migrating, or hibernating, or dying-off, depending on what you know of the life history of species A. You could also compare the information for the two species and try to draw some conclusions. You may say, for example, that the two species differ in the way their numbers vary seasonally.

Figure 2. Functional responses of three caged *Peromyscus* (means and ranges shown)

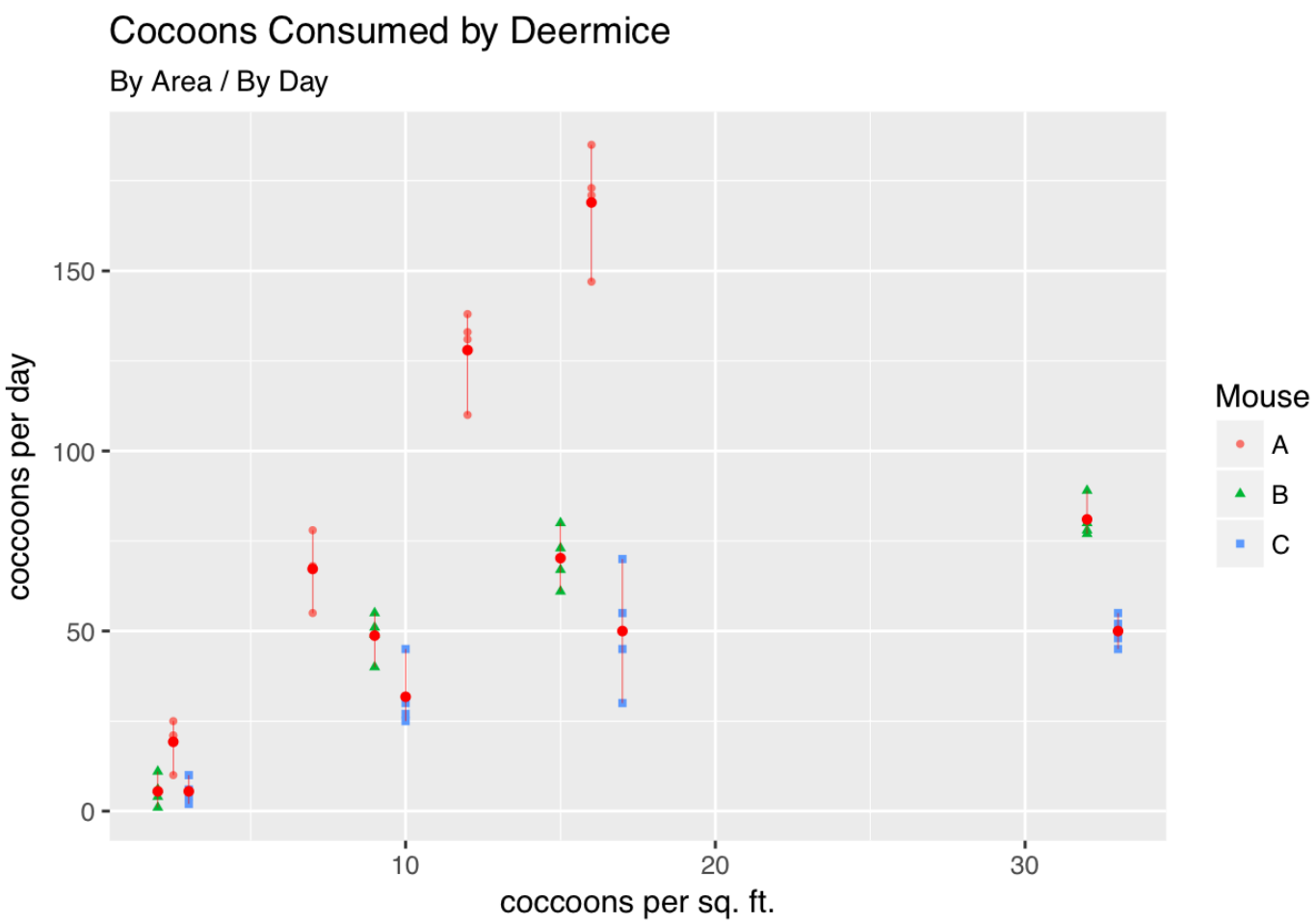


Table 1. Seasonal occurrence of species A and species B in 4 years' sampling by pitfall traps (6 traps per sample) operated 1 day and 1 night throughout the year.

		Spring	Summer	Autumn	Winter	Full Year
A	**No.**	48	222	140	3	413
	%	11	54	34	1	100
B	**No.**	585	302	165	66	1118
	%	52	27	15	6	100
Both spp.	**No.**	633	524	305	69	1531
	%	41	34	20	5	100

However, differences between species or population are not always striking or obvious. If the differences are not striking, and because we are dealing with samples, it is difficult to discern if the apparent differences are real for the population as a whole, or merely due to our particular samples. In other words, we suspect that if we took other samples the differences would no longer be present.

Correlations Between Factors

Look at the information in Table 2. Here we have a series of air temperatures and corresponding body temperatures for the larva of an insect. Looking at this information, you may ask: 1) Is larval temperature significantly influences by or related to air temperature, and 2) Is larval temperature significantly different from air temperature?

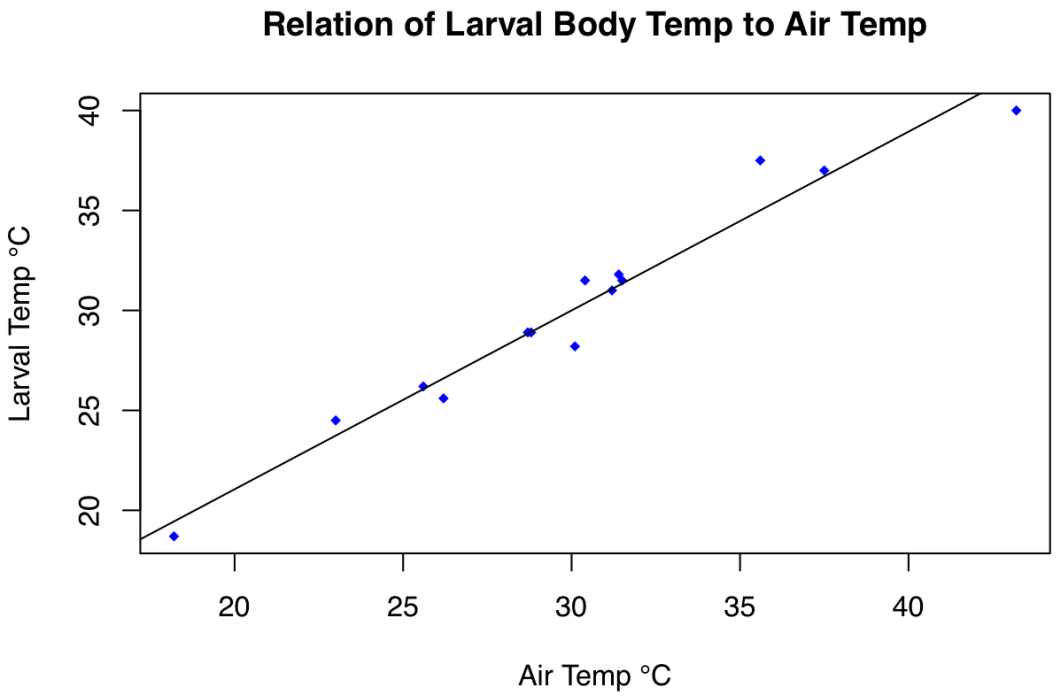
In this example, the values for air and larval temperatures are pretty close, and they do not always vary the same way (sometimes larval temperature goes down when air temperature goes up). But they may be related, and/or they may be significantly different from one another. You cannot tell by just looking at the numbers.

Table 2. Larval temperature taken by pressing the probe of a thermocouple in lateral folds of the cuticle. Each reading is the average from six larvae.

Air Temperature ^{o}C	Larval Temperature ^{o}C
25.6	26.2
43.2	40.0
28.7	28.9
31.2	31.0
31.5	31.5
26.2	25.6
30.1	28.2
31.4	31.8
30.4	31.5
28.8	28.9
37.5	37.0
35.6	37.5
23.0	24.5
18.2	18.7

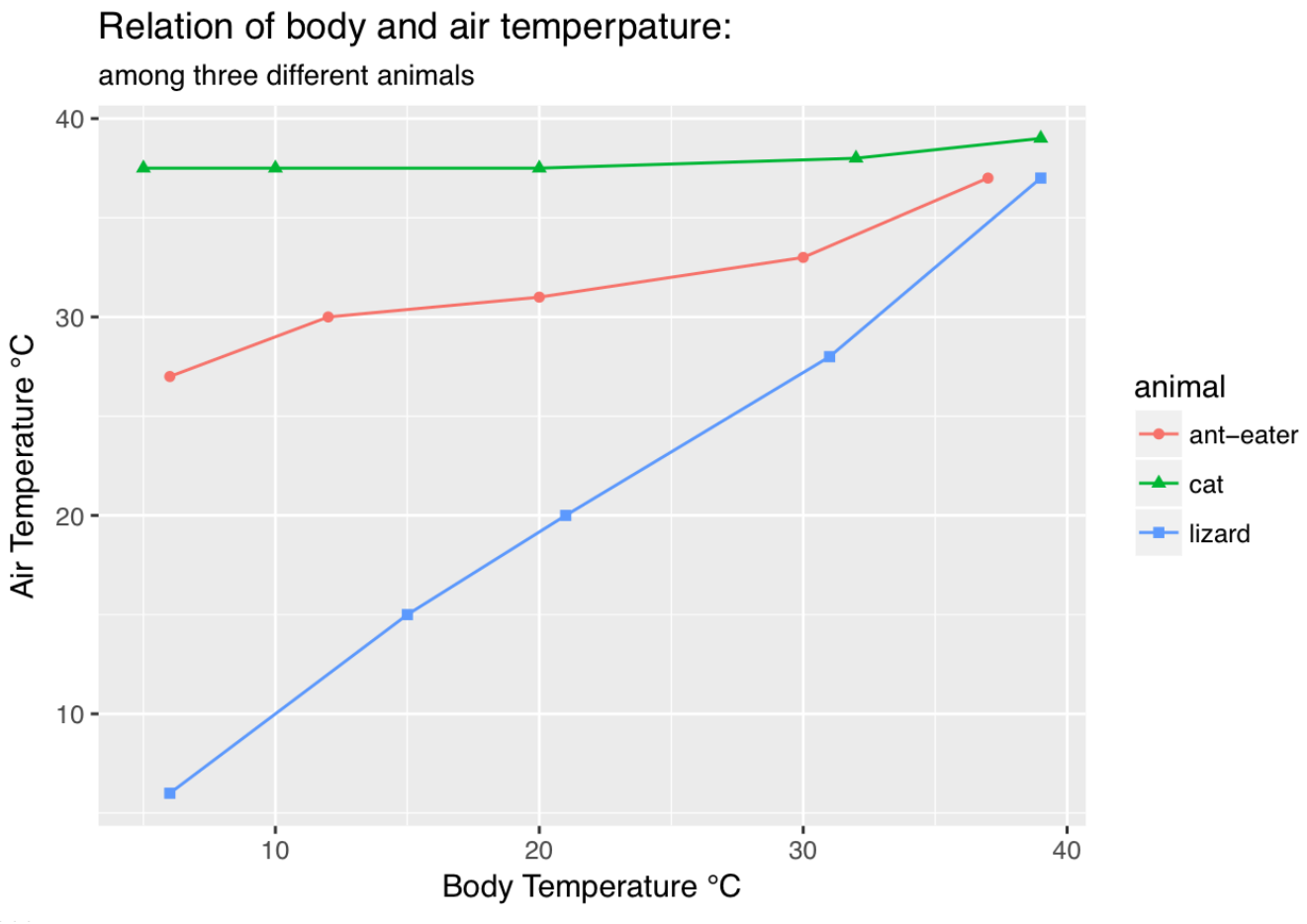
One way to determine if larval temperatures and air temperatures are related is to graph the two, a shown in Figure 3. In this case, it appears that the two are strongly, positively related. As one value goes up, so does the other, and in relatively proportional amounts.

Figure 3. The relationship between the body temperature of larvae and air temperature.



Other examples showing the relationships between body and air temperatures are presented in Figure 4 below. Again, they are related in a straight-line (linear), and positive fashion.

Figure 4. The relationship between body and air temperatures for three different animals.



However, what if the points on the graph were more scattered? Drawing a neat straight line through the points would not be as easy. Nevertheless, there could still be a relationship between the two factors. In such a case a statistical test to determine if a relationship is significant would be necessary.

Testing for Significance

In our example of larval and air temperatures we are asking a question: Are the two related? We can formulate this question into a null hypothesis. The null hypothesis assumes that there is no real difference or relationship between the true value in the population with what you are testing. A statistical test determines if the null hypothesis is valid, or if your results are significantly different from the null hypothesis. The question we are asking in our example can be formulated into a null hypothesis, which is: Larval temperature is not related to air temperature. What we wish to do is verify this hypothesis, or prove it false. To verify the hypothesis we must determine if the relationship between larval temperature and air temperature is non-significant.

The word 'significant' here is not arbitrary. We are not merely going to see if something 'looks' significant. Rather, we wish to make some mathematical statement of the degree of relatedness between the two factors, and then compare this mathematical statement to some pre-determined level of significance which we feel is sufficient to establish relatedness. This pre-determined level is generally taken to be less than or equal to 5% (represented by $p,\leq 0.05$), probability in this case is represented by the letter 'p'. So, by accepting this level of significance we assume that we will be wrong in drawing our conclusion only 5% of the time in repeated samples.

Levels of significance are read from statistical tables which the researcher refers to when the results are obtained. The tables generally have various levels of significance listed across the top, and the "degrees of freedom' (represented by 'd.f') down the side. The degrees of freedom are related to sample size (represented by the letter 'N'), the nature of the relationship depending on the test used.

If a result is obtained which is significant, you would reject your null hypothesis (in our example, a significant result would cause us to reject our null hypothesis and conclude that larval body temperatures and air temperatures are related). If the test result is not significant at the 0.05 level, the null hypothesis is accepted (in our example, we would conclude that larval temperatures and air temperatures are not related).

Pearson Product-Moment Correlation Coefficient

A common statistical test to determine if the relationship between two factors is significant is the Pearson product-moment correlation which produces a coefficient 'r'. This is a value ranging from -1 to +1, with a value of -1 meaning complete negative relationship, a value of zero meaning absolutely no relationship and +1 meaning complete positive relationship.

Whether or not a particular value of 'r' is significant at the 0.05 level depends on the number of degrees of freedom. In a publication a researcher will say, for example, that he/she obtained a value of r=0.30, d.f=28, $p<0.05$. This indicates that a correlation coefficient of 0.30 was obtained with 28 degrees of freedom, and this was found to be significant at the 0.05 level. The researcher would then reject the null hypothesis (which said that the two factors being tested were not related), and conclude that there is a significant relationship between the two factors.

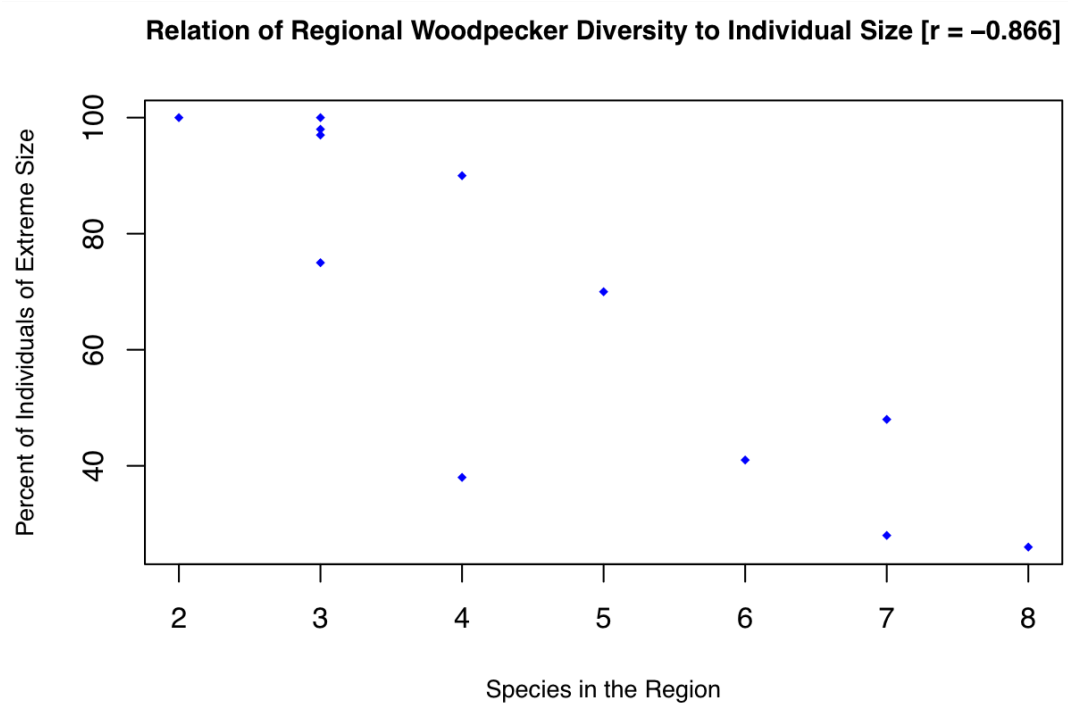
$$r = \frac{\sum (x- \bar{x})(y- \bar{y})}{\sqrt{[\sum (x- \bar{x})^2 \sum (y- \bar{y})^2]}}$$

If you have established that two factors are significantly related, you can speculate that the relationship is causal from your knowledge of the organism. In the example of larval temperature and air temperature, you may conclude that larval temperature is determined by air temperature. The larva perhaps has no homeothermic mechanisms for controlling its own body temperature, and it would be influenced by its surrounding temperatures.

Should you wish to more definitively demonstrate that a relationship is causal (that one factor actually controls the value of another), you can do so by means of an experiment.

Correlation coefficients are at times included along with a graph as shown in Figure 5, where an r = 0.859 is shown for the relationship between 'percent individuals of extreme size' and 'species in the region' for woodpeckers. Notice in this graph that it would be difficult to make a statement about the relationship between the two factors without the use of a statistical test – the points are too scattered to examine just visually.

Figure 5. Regions rich in woodpecker species tend to have more middle-sized individuals compared to species-poor areas; data from the 1974 Christmas count.



Correlation analysis works only for linear relationships. Figure 6 shows such a straight-line relationship which is perfectly correlated. (r=1).

Figure 6. The relationship is non-linear, correlation analysis would not apply. Figure 7 shows examples of non-linear relationships between organisms and their environment. Figures 8 and 9 show non-linear relationships between the densities of fruit-flies (*Drosophila* sp.) and the rate of reproduction and mean respectively.

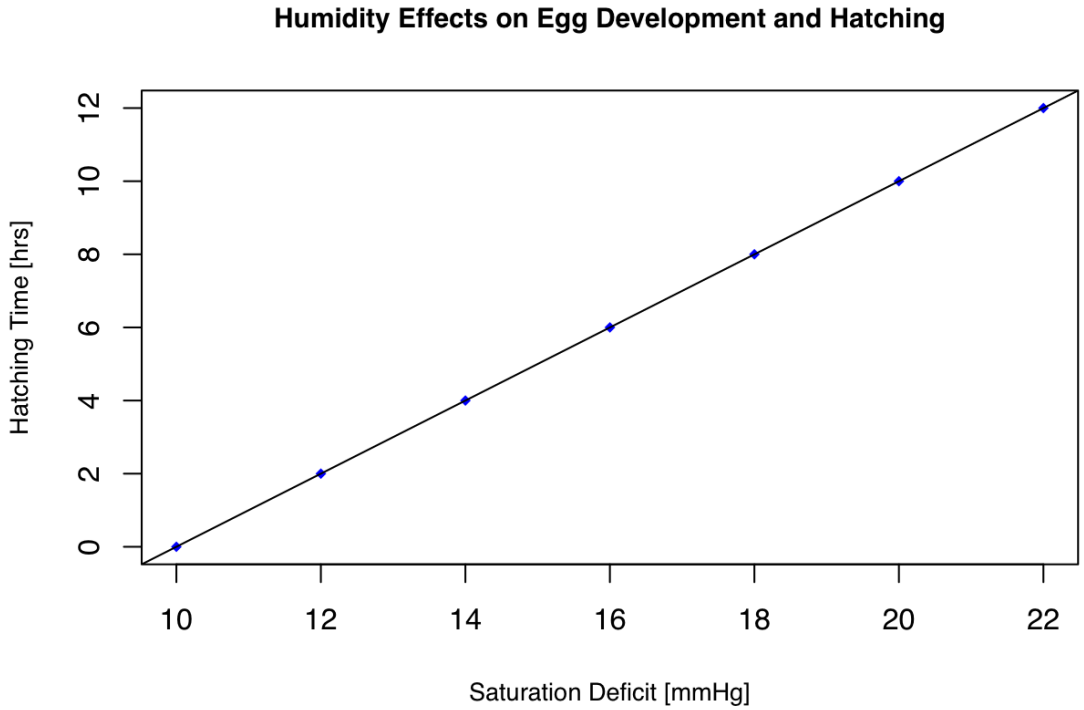


Figure 7. Activity of adult meal worms (*Tenebrio molitor*) and wood lice (*Porcellio scaber*) in uniform humidities.

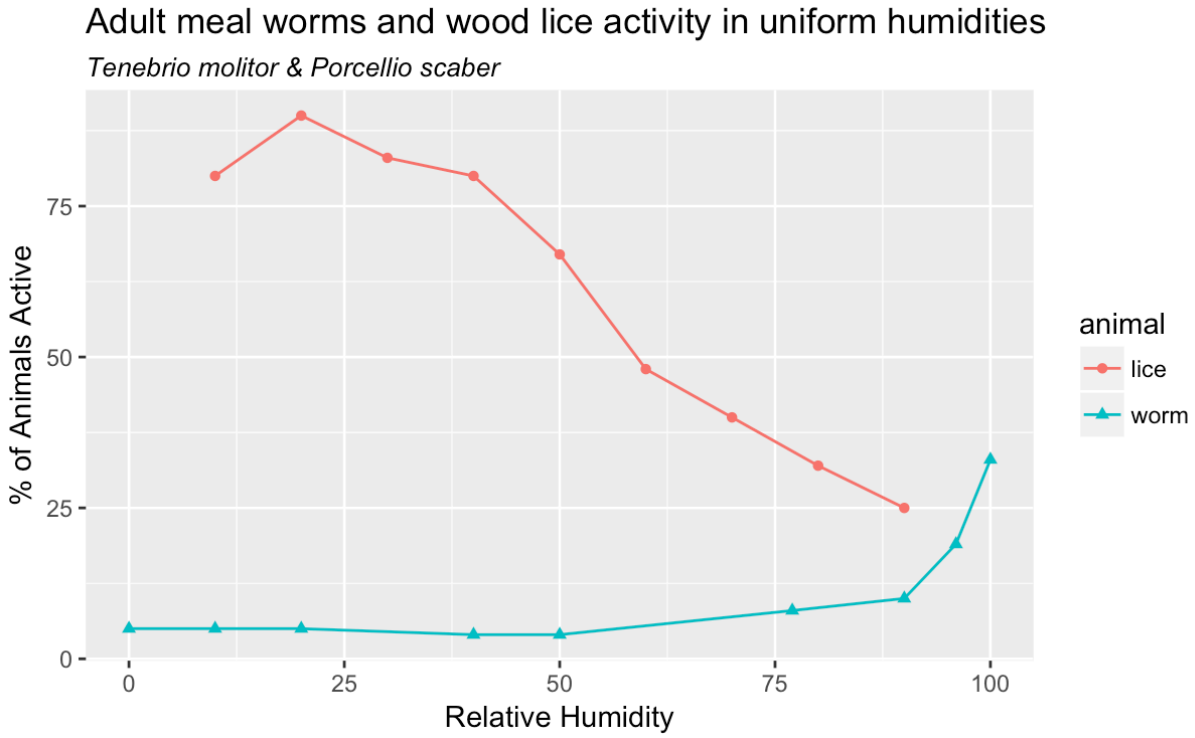


Figure 8. The relationship between rate of reproduction in fruit flies (Drosophila) and the density of the mated population.

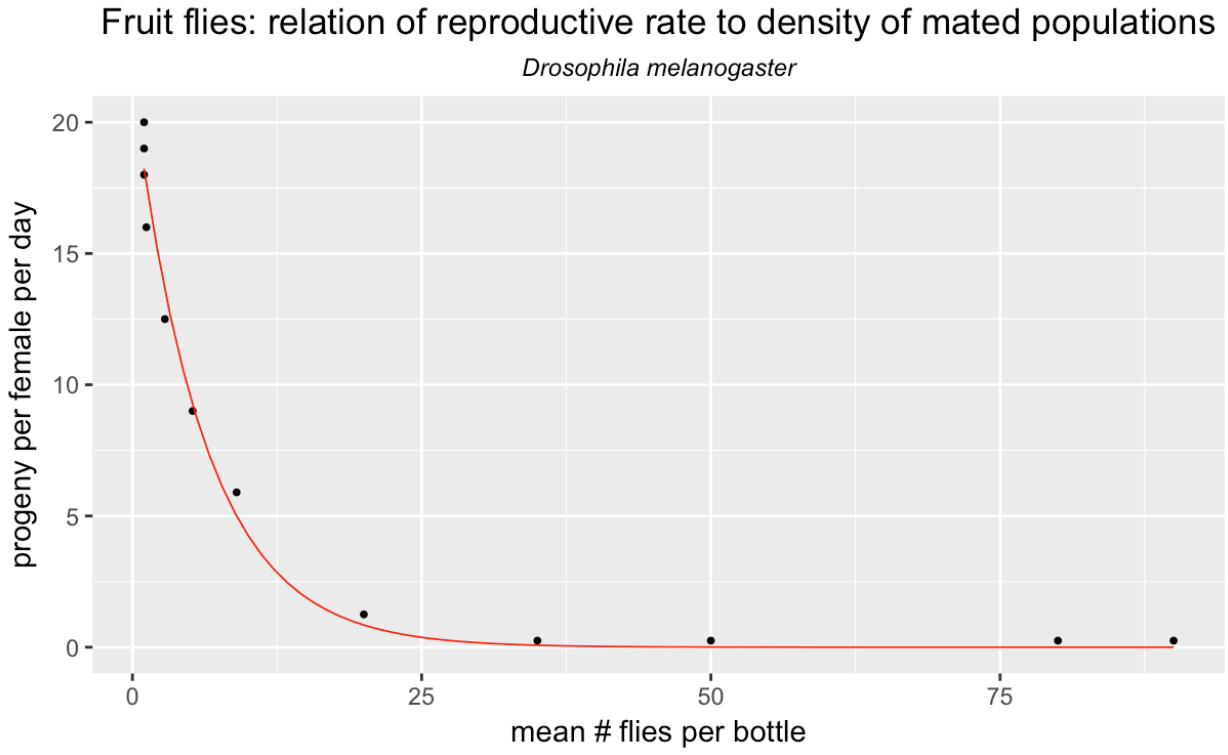
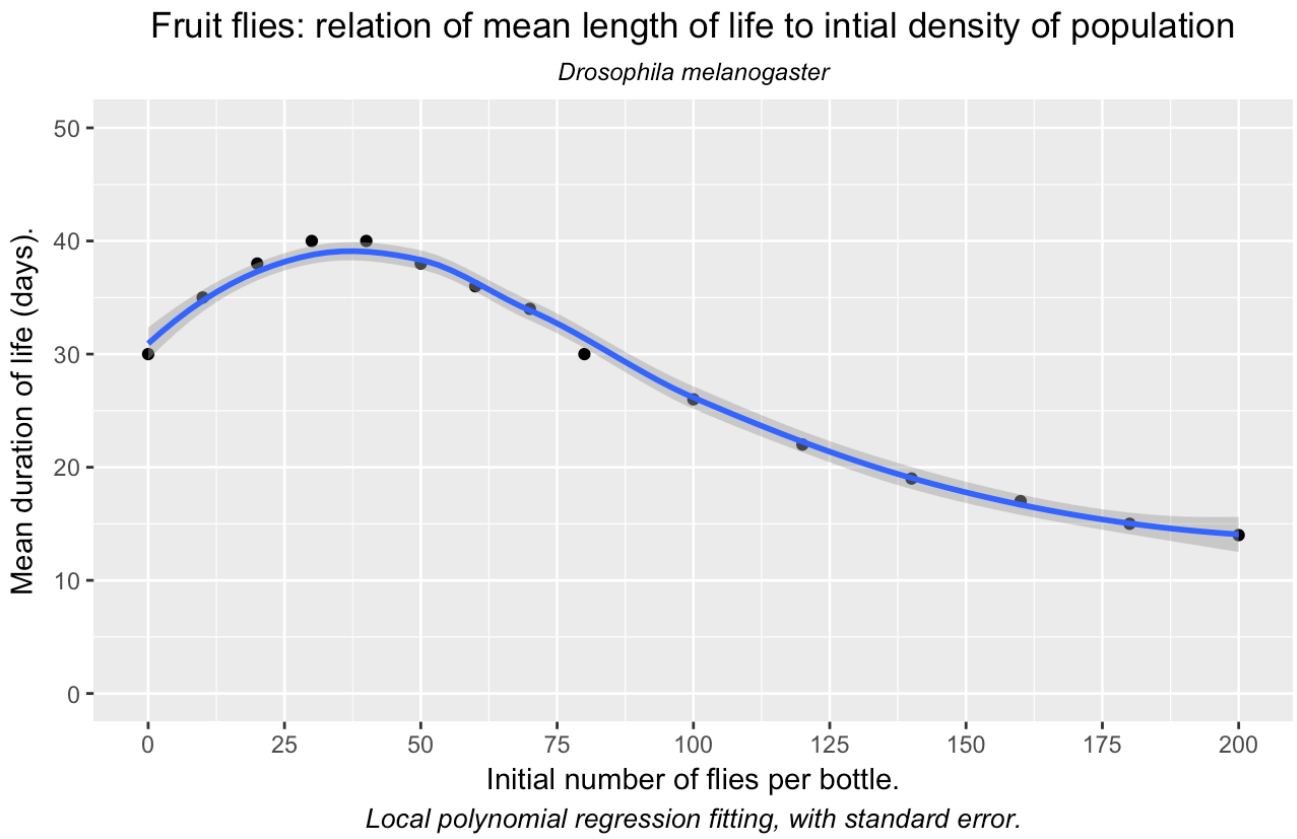


Figure 9. The relationship between the mean length of life and population density in fruit flies (*Drosophila*).



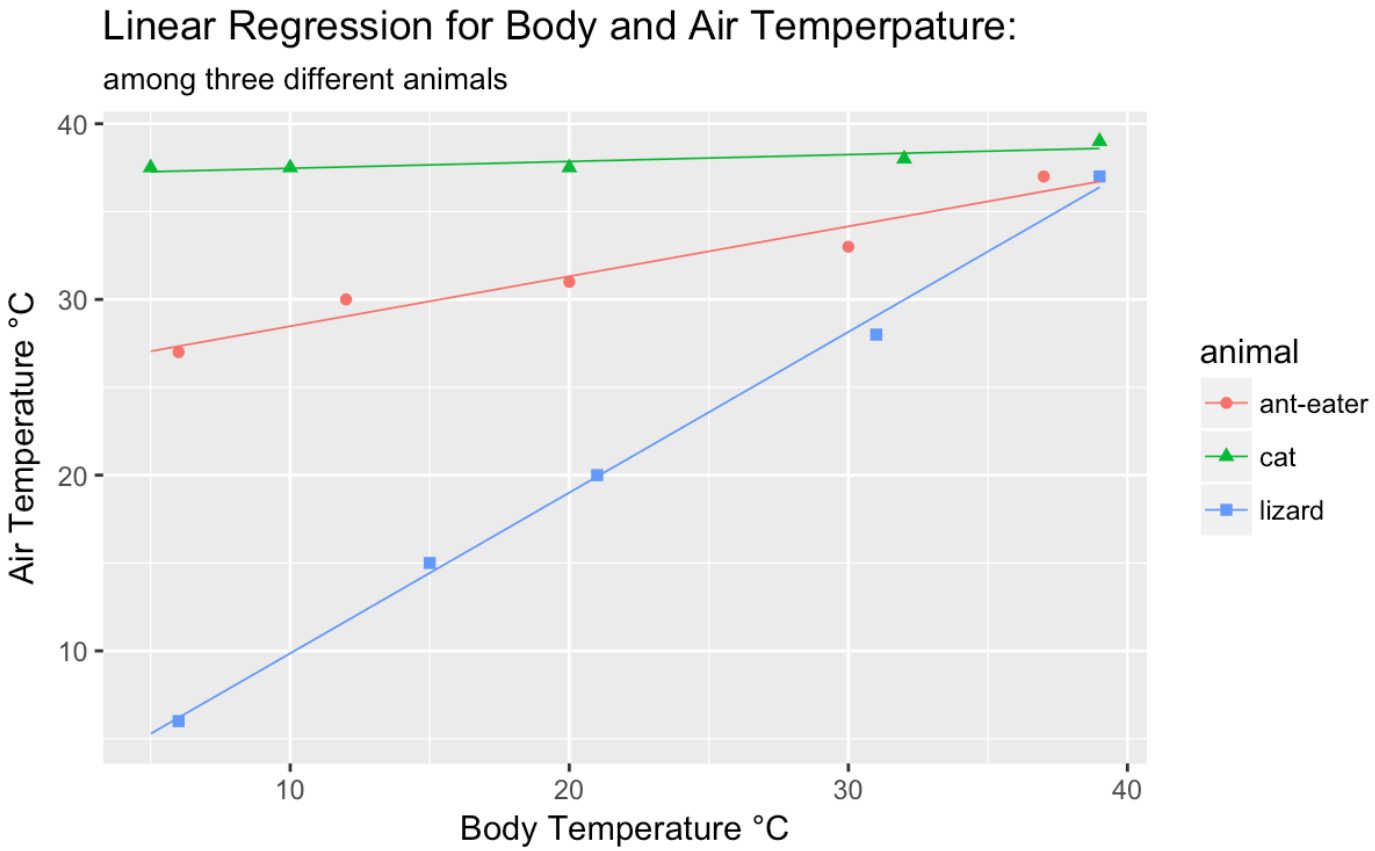
Regression Analysis

When analyzing the relationship between explanatory and response variables, the use of statistical methods for estimating the relationship is referred to as regression analysis. The most commonly used form is simple linear regression. Linear predictor functions, that rely on model parameter estimates derived from the dataset, are used to model the relationship between the response variable y and the explanatory variable X. In Figure 11, the relationship between body temperature and air temperature for three animal species was computed, with predicted y values a linear function of the conditional probability distribution of explanatory variable X.

Other forms of regression analysis used to estimate the relationship of explanatory and response variables include: multiple linear regression (more than one explanatory variable); multivariate linear regression (used with multiple correlated response variables); polynomial regression (used when the response of y is a nonlinear function of X: see Figure 9); logistic regression (used when the response variable is categorical); and other non-linear regression methods, which may include trigonometric functions and exponential functions (see Figure 8).

Recalling the plotting of fruit fly density in bottles as the explanatory variable X in Figures 8 and 9, the regression curves displayed show response variable y estimated from an exponential decay function and polynomial function based on the gamma probability distribution. The local polynomial regression curve in Figure 9 also displays the standard error estimate for this non-linear model as a shaded region bordering the line of best fit.

Figure 11. Linear regression of the same data used in Figure 4. Unlike the earlier example, the data points showing the relationship between body temperature and air temperature for three different animals are not connected by line segments. Instead, a line of best fit supplied by a linear function has been applied as a linear trend between the two variables. for each animal.



Sample Size (n)

Measurements for populations will become more accurate as more individuals are included in the sample. The more individuals that are sampled, the more confident we are that our results are representative of the entire population. It follows that the best approach to studying a population is to sample as many individuals as possible or is reasonable with the resources available for the study.

There is a mathematical reason for ensuring that a study has a certain minimum number of individuals in a sample. Quantitative statistics such as the correlation coefficient and t-test assume that the data have a normal distribution. This requires that the data cover a range of values to form this normal distribution. The minimum sample size for a study depends on a number of factors but a general rule of thumb is that there should be at least 30 in the sample.

Confidence Limits (Intervals)

Returning again to Table 2, suppose you wish to determine if air temperature is significantly different from larval temperature. One method of doing so is to calculate a mean for each group and to present confidence intervals. The word 'confidence' in this case refers to the 5% level of significance again, multiplied by a measure of the variability about the mean. An example would be, say for mean air temperature:
26 ± °C. This would indicate that 95% of the variation in temperature for that population occurs 3°C on either side of the mean 26°C. If the larval temperature was 27 ± 2°C, you can see that there is overlap with this confidence interval and that for air temperature. As a result, you would conclude that the two were not significantly different. If there were no overlap between the confidence intervals you could conclude that they were different. Table 3 illustrates a series of confidence intervals.

Two other statistics, the standard error (SE) and the standard deviation (SD), are also often used much like confidence limits. Again, the mean is presented, and then a '± ' value on either side of the mean, only now it is a standard error or standard deviation, and not a confidence limit. These values are also used to illustrate the lack of overlap between values for populations, but they are not necessarily significant at the $p \leq 0.05$ level.

As the size of a sample increases the confidence interval decreases because we are more certain that the sample mean is representative of the actual population mean.

Table 3. Effects of sea urchins on algal species diversity in the low intertidal zone in New England and the Bay of Fundy.

	New England		Bay of Fundy	
	Chamberlain, Maine (1)	**Canoe Beach Cove, Nahant, Mass. (2)**	**Cape Forchu. Yarmouth, Nova Scotia (3)**	**Quoddy Head, Maine (4)**
Herbivore densities *				
Strongylocentrotus	0	0	4.2±2.6	26.4±13.8
Acmaea	0	0.1±0.2	0.5±0.8	21.2±7.1
L. littorea	0	126.8±60.0	0	0
Algal diversity and percentage of cover				
No. species**	8	3	27	6
H'	1.20	0.23	2.03	1.14
$\bar{X}$% cover canopy	0	0	78.6±19.7	0.4±0.7
$\bar{X}$% cover understory	80.6 ±17.0	89.9±10.2	125.6±20.7***	2.6±2.1
$\bar{X}$% cover *Chondrus* (= in understory)	74.5±18.2	83.3±11.4	14.6±10.2	0

Note – Data are from June-July 1975-1976
*Densities are X±95% confidence intervals/0.25m^2 for 10 quadrats at each area
**Includes both canopy and understory species. In the low zone, *L. littorea* grazes and affects only epiphytic algae on *Chondrus*; sea urchins graze and affect *Chondrus* and most other understory and canopy species
***Percentage of cover >100% reflects the dense multilayer arrangement of the understory at Cape Forchu

Student's t-Test

A common statistical test used for determining significant differences between populations is the student's t-test, or simply, the t-test. The test is also used in analyzing experimental results – in this case, testing for differences between a control and an experimental group.

The t-test gives a value't'. Again, this is either significant at the 0.05 level (and the null hypothesis would be rejected), or it is non-significant, in which case there is no difference between the two populations or groups (the null hypothesis is accepted).

In an experimental situation a significant difference between a control and an experimental group is taken as sufficient evidence for causality. Thus, whatever was done differently to the experimental group of organisms is taken to be a cause of the significant difference obtained from the control group. Table 4 illustrates how the results of such an experiment may be presented.

Table 4. Number of frogs calling in the experimental (occupied by four loudspeakers) and control enclosures.

	No. Calling	
Trial	**Experimental Enclosure**	**Control Enclosure**
1	8	9
2	6	7
3	3	5
4	7	8
5	8	10
6	9	11
Total	41	50
Note: Significantly fewer frogs called in the experimental enclosure, p 0.005 (t=6.71; df=5)		

In comparing more than two populations at one time, an analysis of variance is used. An analysis of variance (ANOVA) produces a value 'F' which, like the value 't', can be significant or non-significant.

Chi-square Test (X^2)

Look at Table 5. Here you have information on the number of males and females of a species of moth, caught in traps set at two heights from the ground. The question you wish to answer from this information is: Is the proportion of males to females different in the upper and lower traps. The important word here is proportion. We are not interested in comparing absolute numbers of measurements. The proportion alone is of interest.

Table 5. Numbers of males and females caught in upper and lower traps.

	Females	**Males**	**Totals**
Upper Trap	(A) 17	(B) 246	(A + B) 263
Lower Trap	(C) 32	(D) 458	(C + D) 490
Totals	(A + C) 49	(B + D) 704	(N) 753

The test is used in this case is called the chi-square, represented by the letter: X^2. This value is again compared to values in prepared tables and checked for significance. If the value is not significant, then the proportions are said to be the same (the null hypothesis is accepted). If the value is significant, there is a significant difference between the proportions and they are said to differ.

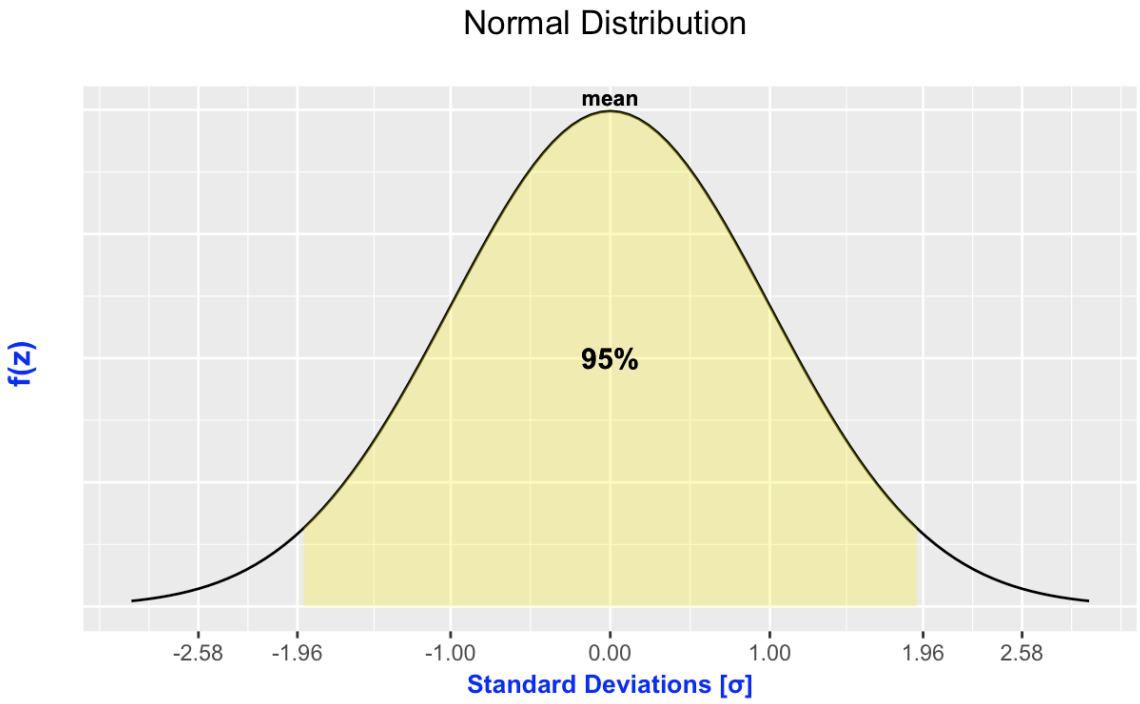
The chi-square is a common statistic in the field of genetics. Here, Mendelian ratios are compared to observed results to see if the characteristic is following a Mendelian form on inheritance.

Assumptions in Statistical Tests

Statistical tests are based on mathematical theory. In order for the equations to work properly (to be valid),it has been assumed that the values are normally distributed. Also, the variances (a measure of variability) are homogenous. If these two assumptions are not true for some data, then statistics (nonparametric) other than the ones mentioned in this handout must be used.

A normal distribution is illustrated in Figure 10. It has a characteristic bell-shape, so it is called a bell-shaped curve. Other names for it are the normal curve, and the Gausian curve. Notice that in a normal curve there are many values of intermediate range, and only a few very high or very low values.

Figure 10. Normal distribution curve. The most frequent value (greatest height of the curve) is the *mean,* with smaller or greater values decreasing in frequency the more they depart from the mean. The diagram shows the distribution or area of this curve measured in terms of departures from the mean called *standard deviations*.



Another assumption in statistical tests is that the individuals' measured were collected at random. It is necessary to sample at random (for example, using a table of random numbers to select quadrats for sampling) to avoid introducing a bias into your information. If you have a hypothesis which you are trying to prove or disprove, it is important that you not only collect those individuals which verify your hypothesis. They would not be representative of the population as a whole.

Notice in Figure 10 that part of the normal distribution containing 95% of the area of the curve. The confidence limits, discussed earlier, are usually 95% confidence limits, and they provide the values around the mean which contain 95% of the values in the population. These limits correspond to the area shown on the curve. The values outside the 95% area of the curve constitute the remaining 5% of the population. In discussing the levels of significance of statistical tests earlier, we referred to the 5 % level of significance. In a test such as the t-test for example, when we say the result was significant at the 5% level, we mean that the two populations only overlapped in this 5 % area of their curves, if that much. The conclusion would then be that we are dealing with two different populations.

Multivariate Statistics

Natural systems are complex, and multivariate by nature. Multivariate analysis is the application of multivariate statistics when dealing with many variables. There are many methods used to study multivariate data, all seeking a better understanding of systems' underlying processes, and the connection and relationships between variables. Multivariate analysis is used to reveal those variables that may determine most of the variability in a data set, and can aid in the identification of those variables that co-vary and might be highly correlated (see Figure 12). Multivariate analysis relies heavily on graphical visualization. Examples of multivariate analytical methods include: multivariate analysis of variance (MANOVA); multivariate analysis of covariance (MANCOVA); canonical correspondence analysis (CCA); artificial neural networks; principle response curves (PRC); and principal component analysis (PCA).

PCA is commonly used to simplify the analysis of data sets with multiple variables by combining them with the use of linear transformations, which is a form of multidimensional scaling. This preserves most of the original information regarding each variable, but allows for easier handling and interpretation of the data. In Figure 13 the principle components are labeled PC1-PC5. The figure shows biplots of each, and this allows for a comparison of multiple variables which would normally occupy more than three dimensions on a graphical matrix. The use of PCA can aid in the interpretation of multivariate data sets and help identify relationships between explanatory and response variables (see Figure 14).

Figure 12. Multivariate correlation of morphological traits between sexes and colour types for *Leptograpsus variegatus* (Purple Rock Crab). This pairwise scatterplot visualizes the correlation between different morphological measurements taken from 50 individual crabs: FL-frontal lobe; RW-rear width; CL-carapace length; CW-carapace width; BD-body depth.

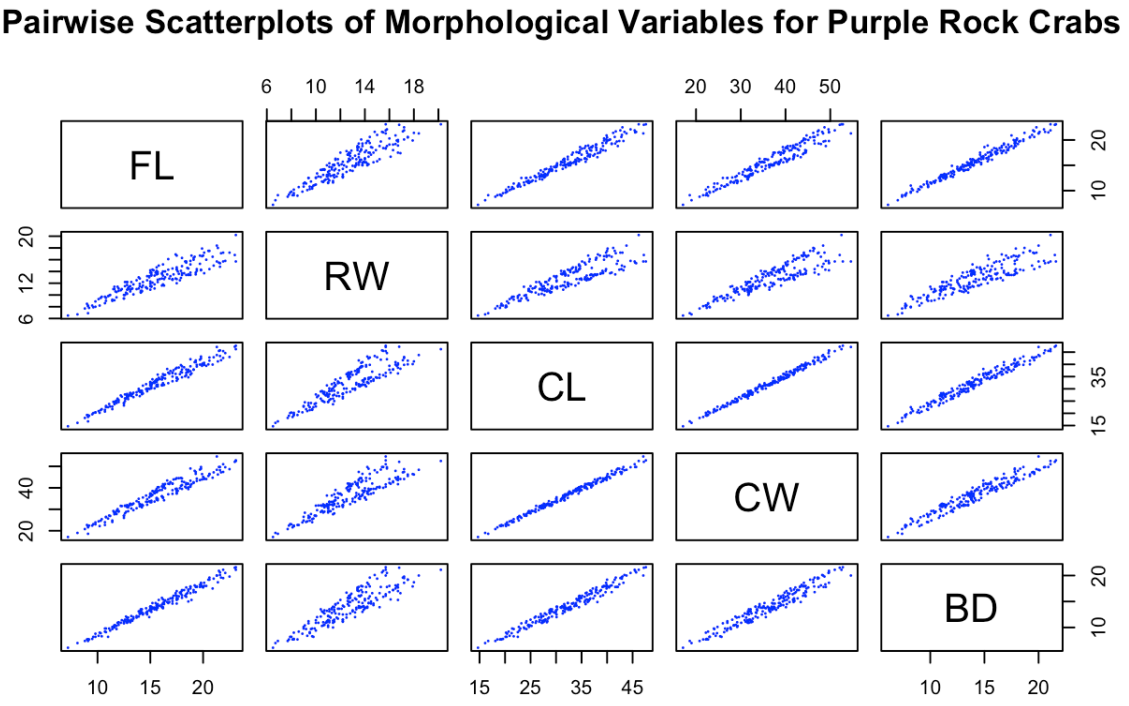


Figure 13. Principal Component Analysis (cont'd). This plot uses PCA to compare the relationship between the multiple morphological variables associated with crab development. In this figure the principle components are labeled PC1-PC5. These biplots show that PC2 and PC3 are not highly correlated based on the noticeable clustering, and could be examined further using the explanatory variables of crab sex and colour type. [The red arrows indicate the morphological response variables FL-BD, and arrows pointing in the same direction indicate positively correlated variables, while those in opposition indicate negatively correlated variability.]

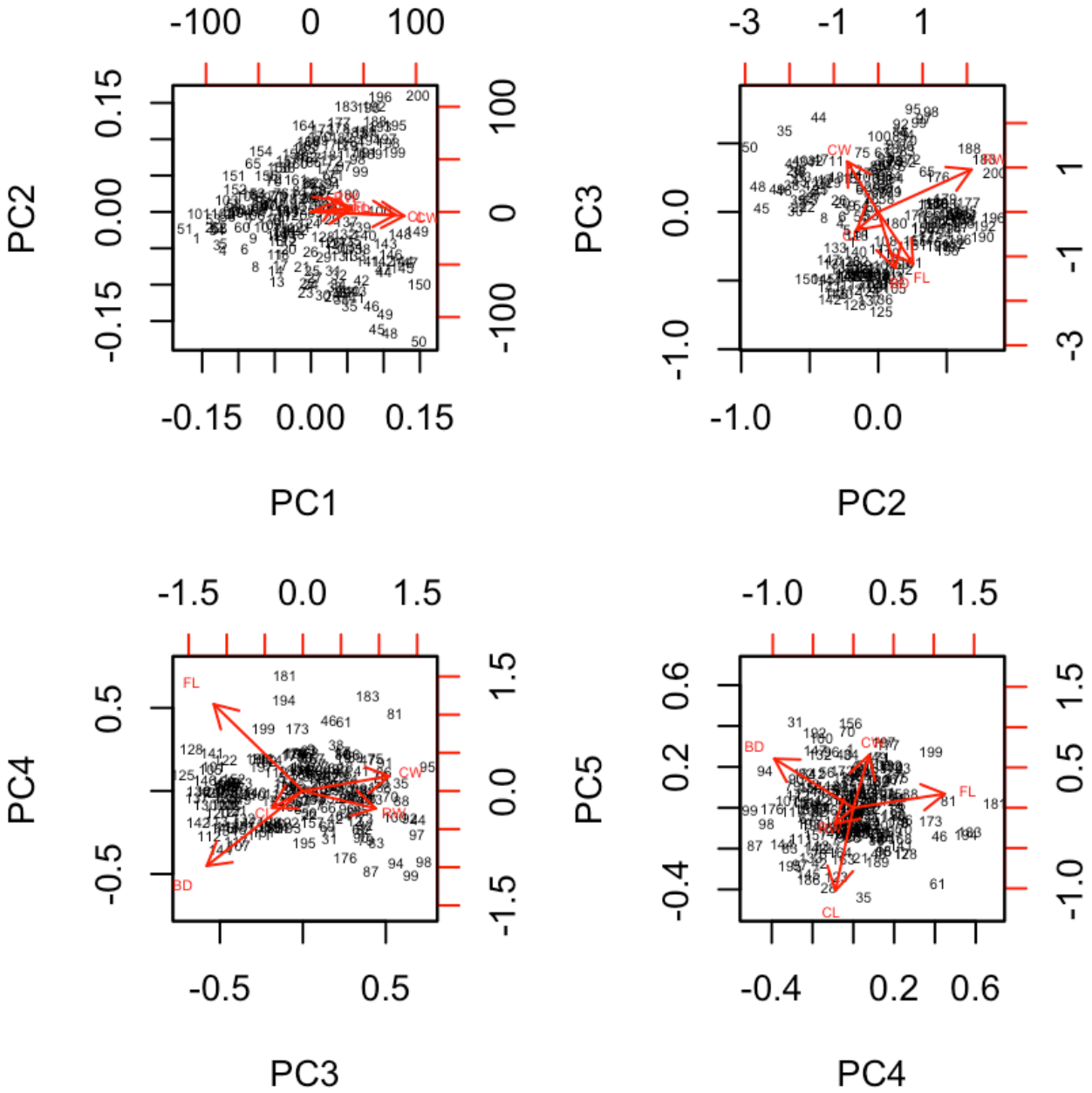
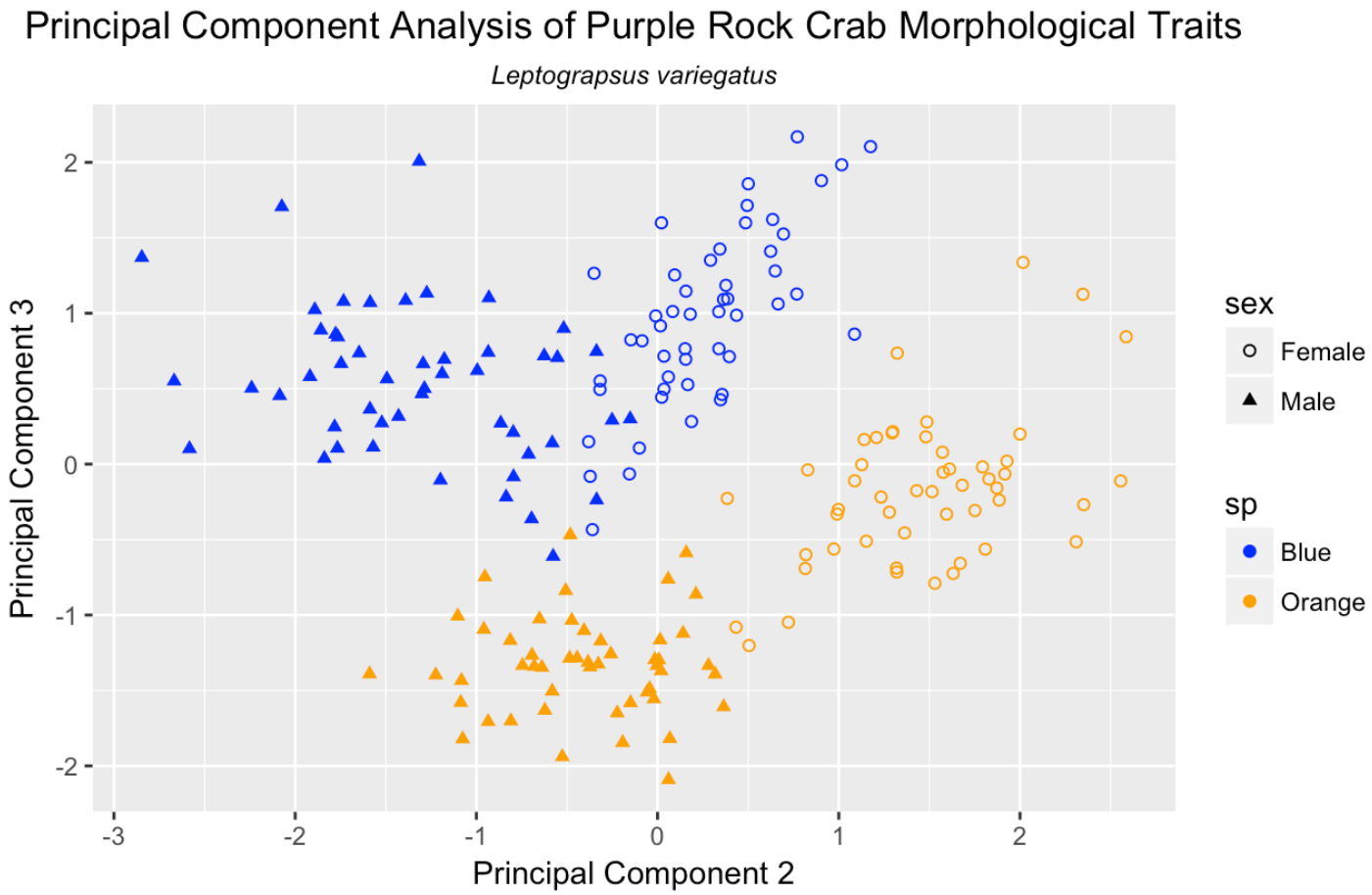


Figure 14. Principal Component Analysis (cont'd). This biplot overlays the explanatory variables "sex" and "colour type" with principal components 2 and 3. The clusters identified in Figure 13 are explained by these variables, and PCA has helped with the interpretation of this multivariate dataset.



R Statistical Software

R is a a free and open source statistical programming language and software environment based on the S programming language. R allows for statistical computing and data visualization with enhanced graphics. Data is readily explored, manipulated, computed and displayed in the R environment. Because R was designed as a “true computer language”, new functions and packages are easily integrated into the system, allowing for robust, effective and efficient data analysis. R supports most statistical techniques, with data analysts and researchers continuously adding to the environmental repertoire.

RStudio is an integrated development environment. RStudio provides the data analyst or researcher with a coherent user interface, simplifying the importing, exploration and manipulation of data. It makes carrying out statistical calculations, visualization and graphical representation of data easier, and it enhances the researcher’s ability to share their work and collaborate on projects.

All of the plots and graphical representations of statistical relationships found in this primer were produced using the R programming language in RStudio.