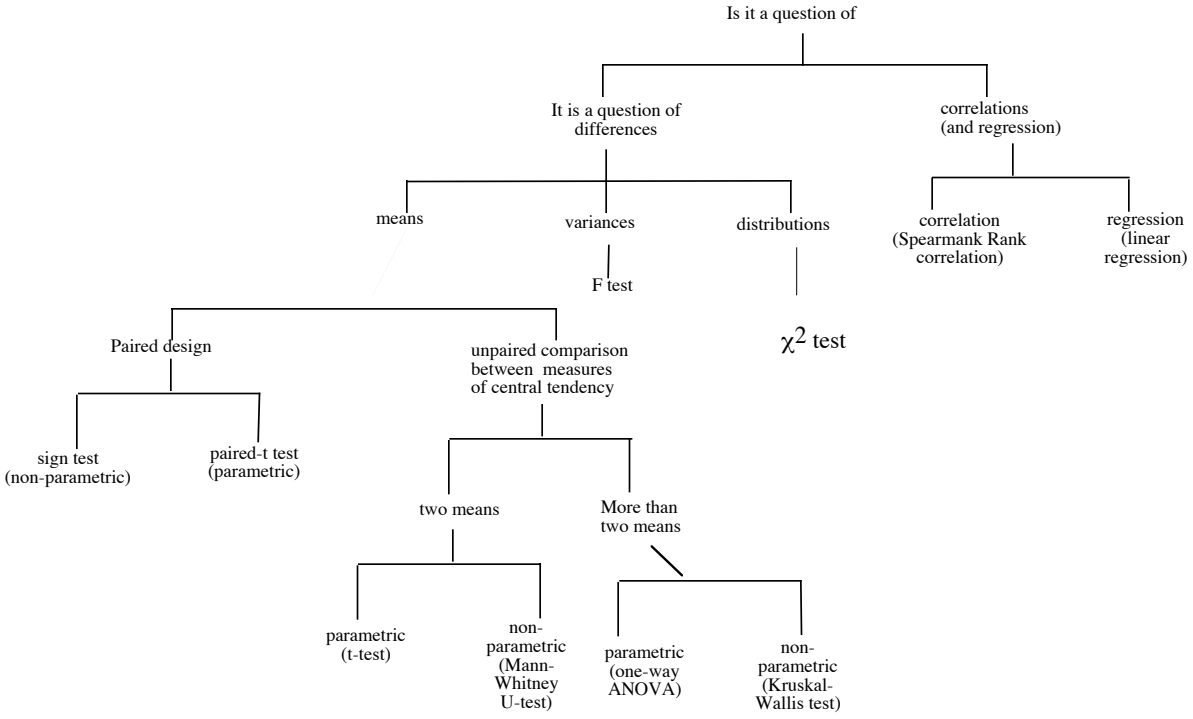
Statistics a very basic introduction

Carlos Martínez del Rio
Department of Zoology
University of Wyoming
Laramie, WY 82071-3166



A flow chart of statistical tests. Use it judiciously....I modified it from Ambrose, H. W. and K. P. Ambrose. 1995. Handbook of biological investigation.

STATISTICS A (VERY) BASIC INTRODUCTION

Carlos Martínez del Rio
Department of Zoology and Physiology
University of Wyoming, Laramie. WY 82071-3166
(cmdelrio@uwyo.edu)

Introduction

If you read the scientific literature on seed dispersal, you will find that it is full of statistics. The articles and books are full of acronyms like ANOVA, LSD (it is not a drug!), HSD, and letters (p, q, F and $\chi 2$) It is safe to say that you cannot truly understand this literature without some knowledge of what statistics are and what their purpose is. Worse, you cannot publish in the scientific literature without knowing statistics. The two lectures that I will give you about statistics should give you enough background to get started in thinking about the world in a statistical fashion. They should also help you to read –and understand, at least some of the statistical sections in the scientific literature. I must warn you. The two lectures are not a substitute for a detailed course (or courses) in statistics. You may be understandably annoyed when you learn that you need to learn a lot of statistics. After all, your goal is to understand the natural world. For good or bad, statistics are an essential part of your toolbox as a biologist. Without them, it is very unlikely that you will make a big contribution to ecology. Ecology is a complicated subject and ecological systems are variable. Variability makes statistics absolutely necessary.

Before I even begin the lectures I must make a disclaimer. The few lectures that I will give you are incomplete, even as an introduction. There are many important topics that I will not discuss. For example, I will never talk about what the normal, t, and F distributions are, and why we can use them to make statistical inferences. I will use them, but I will not justify their use. Because understanding these distributions is essential to fully comprehend statistical procedures, the lecture will have a bit of a magical flavor. You plug in data in the computer or in formulas and out come statistics and p-values. I do not thing that this is a good situation, but it is unavoidable. I have placed an annotated list of books and references at the end of these notes. The list includes the books and articles that I believe should be in any practicing ecologist's library. Reading these books and articles will give you a better understanding of statistics. I hope that these lectures will do 3 things:

1) Give you an idea of why you need to think statistically (i.e. think of scientific questions as questions that must be addressed by observations and experiments that yield data that can be analyzed with statistics).

2) Provide you with a few basic statistical tools, so that you can begin designing your own observations and experiments, and analyzing the data resulting from them.

And

3) Motivate you to take more statistics courses and to learn more about statistics on your own.

Statistics: what are they good for? It is sensible to begin this lecture by asking: why do we need statistics in ecology? There are two approaches to answer this question. I will begin by the cynical one. There are 4 bad reasons to learn statistics:

1) using statistics makes your papers more difficult to read, and hence, makes them scientifically respectable.
2) Everyone else uses statistics.
3) Computers make it easy to get a bunch of numbers, and it is easier to spend time sitting in front of a computer (in an air-conditioned room) than sweating in the field or thinking.
4) You cannot publish a scientific paper if you do not use statistics and you need to publish if you want to get a job that will allow you to sit in front of a computer.

Of course, these reasons are not really valid. A less disrespectful view recognizes that we need statistics for two interrelated reasons: The first one has to do with the tendency that humans have to find pattern in nature, sometimes even when it does not exist. We need a tool that can tell us if differences among classes of phenomena in nature are associated with understandable processes or are simply the result of random variation. If you are a half-decent scientists, you discover regular patterns in nature and pose hypotheses to explain them all the time. Statistics give you a tool to determine if the patterns that you discover are really regular and if your hypotheses to explain them are correct. The second reason is that science is about making measurements that are repeatable and about making predictions. Statistics allow us to determine the repeatability of our measurements and the accuracy of our predictions.

In summary, statistics are important and useful because they allow us to evaluate the confidence that we can place on the regularity a pattern, and because they allow us to determine if our hypotheses are supported or not by reality.

Know your data: describing pattern

Ecology is a quantitative science. It deals with numbers. Before you do an experiment to test a hypothesis you have to describe a phenomenon. Phenomena in ecology are described with numbers. These numbers describe, hopefully in a more accurate fashion, your qualitative observations. How do we describe pattern? First we need a collection of observations. An observation is often called a *datum*, and a collection of observations is a *sample*. This sample is a subset of all possible observations about which you want to draw a conclusion. The set of all possible observations is a statistical population or sampling universe.

Types of data.- Your data will be numbers and can be of several types. It is important to recognize the types of variables that you are using because these will determine the kind of statistical analysis that you will use. Data can be divided into two broad classes: categorical variables and measurement variables.

Categorical (or nominal) variables are characteristics of an object that can be broken down into classes or categories (fruit can be green, red, and purple, it can be juicy or fatty). *Binary variables* are categorical variables for which only two categories exists (alive or dead, parasitized or non-parasitized). *Measurement variables* are associated with measurements on an object and that are associated with a number. Measurement variables can be of many different types:

1) Continuous ratio scale data. This category includes lengths, volumes, weights, capacities, rates, and so on. All data for which you can unambiguously assign a 0 and that you can express as a real number are of this type.
2) Discrete data. Whenever you count objects, you obtain discrete data that has the same properties as the natural numbers (0, 1, 2, 3, 4,…, etc.).
3) Ordinal data. Some measurements are better expressed as ordered sets. Plant A is smaller than plant B, and plant C is bigger than plant B (A < B < C). We can assign the numbers 1, 2, and 3 to plants A, B, and C. Note, however, that all we can say is that C is bigger than B, but we cannot say by how much. Continuous ratio scale data and discrete data contain more information (we can say that a tree has 3 times more fruit than another, or a fruit is twice as long).

Getting a representative sample is not as easy as it may seem, but it is something that you should strive to do. The main problem that plagues samples is lack of statistical independence. This problem is best illustrated with and example. Suppose that you would like to know the sugar concentration in the fruit of a plant. You measure this continuous variable (mg of sugar/100 mg of fruit pulp) in 100 fruits from one tree and in 5 of another. The measurements that you have done on each tree are not statistically independent. The fruits in each tree are more likely to be more similar to each other than to fruits in other trees. Your sample is biased because one tree is better represented than the other. To obtain a representative sample you would have to measure 1 fruit per tree in a 100 trees. In subsequent sections we will ask the question, how big should your sample be (i.e. how many fruits and trees should you measure). In addition to being representative, some tests may require that you have a random sample. To achieve a random sample, you may want to number 1000 trees and draw a random sample of a 100. Although obtaining random samples can be very important, it is sometimes not possible. For some problems you may want to measure many fruits per tree in many trees, and then ask the question how much of the variation in sugar content is explained by variation within and among trees. This is a question that we will address later. For now, it is sufficient to emphasize that when you collect samples, you should attempt to keep them as statistically independent as possible. If you cannot do this (say, it is impossible to find 100 trees, you can only find 5 individuals), then you should always identify all the possible sources of variation. That is you need to identify which fruit came from which tree. As we will see determining the contribution of different sources of variations to the total variation in a sample is one of the goals (and one of the useful tools) is statistical analysis. Assuming that data are independent when they are not in statistical analysis is called pseudoreplication (Hulbert 1983). It is considered an ugly sin by most ecologist, who nevertheless continue committing it. We have our first commandment:

First Commandment: You shall not pseudoreplicate.

Here I must make a very important comment. In general, it is not a good idea to go out and collect a data set without knowing in advance how you will analyze it. Many of the following sections will describe how to conduct certain statistical tests. You should use them as guides to how to design a study. Few things irritate statisticians more than having a biologist come with a big table of numbers and asking the statistician how to analyze them. As a general rule, you must know the statistical test that you will use before you even collect a single observation.

Second commandment: you should think about how you will analyze your data before gathering it.

Describing the data.- How can you make sense of the data and communicate it to others? We are primates, and hence we are a very visual species. One of the first recommendations that I can give you is to always draw a picture of your data. In this case a picture would be a histogram. A histogram or bar graph is a pictorial description of the frequency distribution of values in your sample. Making a histogram is a bit of an art, but once you have a bit of experience it will be second nature. To make a histogram, you must first divide the values into intervals. For discrete or categorical values, these intervals are really easy to construct. They are {0, 1, 2, 3,….,N} or {green, yellow, red, and black }. When you have continuous variables, it is a bit more difficult. Using too many or too few intervals, will obscure the shape of the distribution (which is something that sometimes can be of interest). There are some rules of thumb about how to construct these intervals, but the choice is generally left to good judgment, bearing in mind that 10 to 20 groups is about right for biological work (but please use your judgment!). In general, intervals of the same size should be used. Once you have determined the intervals, you must determine the absolute frequency in each interval. The absolute frequency is the number of observations or measurements in each interval. The best way to do this is to place the results in a table. The following two data sets are examples of what your data tables will look like:

TABLE 1

Category (fruit color)	Absolute frequency	Relative frequency Absolute frequency/N
A=green	56	0.27
B=yellow	60	0.28
C = red	46	0.22
D=black	49	0.23
Total = N	211	

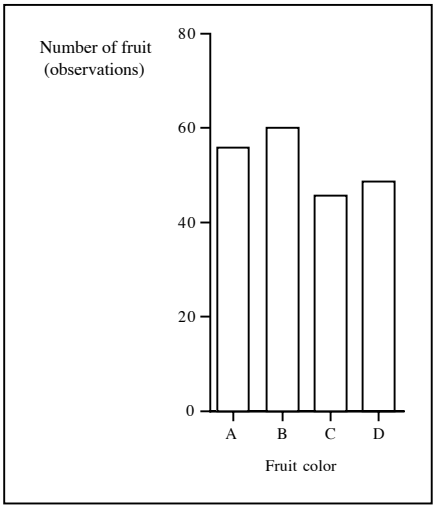
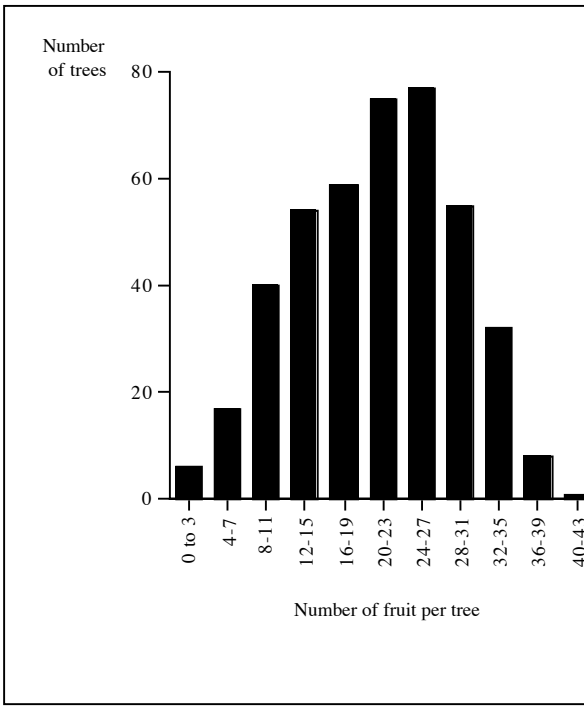
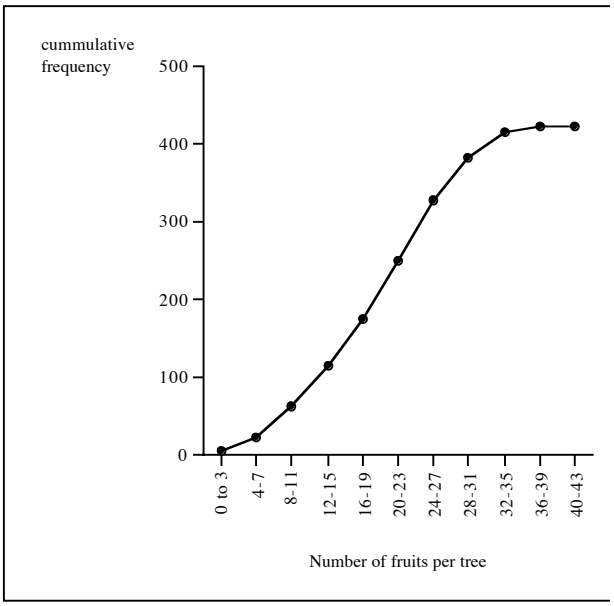


TABLE 2

Category (number of ripe fruit per tree)	f_i	F_i	cumulative
0	0.00707547	3	3
1	0.00235849	1	4
2	0.00235849	1	5
3	0.00235849	1	6
4	0.00471698	2	8
5	0.00707547	3	11
6	0.01179245	5	16
7	0.01650943	7	23
8	0.01886792	8	31
9	0.0259434	11	42
10	0.02358491	10	52
11	0.0259434	11	63
12	0.03066038	13	76
13	0.02830189	12	88
14	0.03773585	16	104
15	0.03066038	13	117
16	0.03301887	14	131
17	0.03773585	16	147
18	0.03537736	15	162
19	0.03301887	14	176
20	0.04009434	17	193
21	0.04245283	18	211
22	0.05424528	23	234
23	0.04009434	17	251
24	0.04481132	19	270
25	0.04245283	18	288
26	0.04481132	19	307
27	0.0495283	21	328
28	0.04245283	18	346
29	0.03066038	13	359
30	0.02358491	10	369
31	0.03301887	14	383
32	0.02122642	9	392
33	0.02358491	10	402
34	0.08018868	34	436
35	0.01179245	5	441
36	0.00943396	4	445
37	0.00235849	1	446
38	0.00471698	2	448





39	0.00235849	1	449
40	0	0	449
41	0.00235849	1	450

In a histogram you plot the absolute frequencies (often denoted by F_i) for each interval (or category i) as a bar in the y axis and the intervals in the x axis. The height of the bars is proportional to the value of the frequency. You will sometimes find that sometimes researchers use relative frequencies ($f_i = F_i/N$) or percentages ($\%i = [F_i/N]x100$) instead of absolute frequency. Another way that you will find is plots of the cumulative frequency against the interval:

$$\text{Cumulative frequency}_i = \sum_{j=0}^{i} F_j = F_1 + F_2 + F_3 + F_4 + \ldots + F_i.$$

The symbol $\sum$ means sum, and the expression means the sum of the frequencies F_j from j equal to 0 to i where i is the number of the interval (1,2,3,…, up to N, where N is the number of observations). These plots are called cumulative frequency distributions.

Third commandment: Always provide a visual representation of your data.

Parameters and statistics.- I cannot overemphasize the importance of making a pictorial representation of your data. However, you will also have to give numerical summaries of large data sets. These numerical summaries that, ideally, characterize your data can be divided into several types: the two most widely used are 1) *measures of central tendency* (arithmetic mean, median and mode), and 2) *measures of dispersion* (variance, standard deviation, coefficient of variation, and range). A quantity such as a measure of dispersion is called a *statistic*. Lets begin by describing measures of central tendency. The most commonly used measure of central tendency is the arithmetic mean or average:

$$\text{Arithmetic mean} = \overline{X} = 1/N\left(\sum_{i=0}^{N} x_i\right) = (1/N)(x_0 + x_1 + x_2 + \ldots x_N).$$

Where N is the total number of observations and xi equals the value of each one of the observations. If your measurements are discrete, a handy formula for the arithmetic mean is

$$\text{Arithmetic mean} = \overline{X} = 1/N\left(\sum_{i=0}^{K} F_i i\right) = \left(\sum_{i=0}^{K} f_i i\right) = f_0 0 + f_1 1 + f_2 2 + \ldots + f_K K$$

In this equation you have N observations (or data points) that can be divided into K catergories, each one of which has an absolute frequency F_i and a relative frequency f_i. For example you can calculate the arithmetic mean for the data set in table 2 as

$$\overline{X} = (0+0+0+1+2+3+4+4+5+5+5+\ldots..+41)/450 = 23.0$$

or you can calculate it as

$\overline{X}$ = (3/450)0 + (1/450)1 + (1/450)2+ (2/450)4 + (3/450)5 +…..= 23.0.

It is the same thing.

In addition to the mean, it is useful to consider the median and the mode as measures of central tendency. The *median* is defined as the value that divides your observations into two sets of equal size. 50% of your observations have a value that is lower than the median and 50% have a value that is higher than the median. For example if you take a look at table 2, you will find that 50% of the observations (i.e. 225 observations) fall below about 21.5. If N is odd, the median will be an integer. If it is even (as is the case that we have just calculated) it will be a half integer. The *mode* is defined as the most frequently occurring measurement in a data set, and can be found by taking a look at the frequency distribution. The data in table 3 has a mode when x = 22. The data that we have used are very well behaved in that all the measurements of central tendency (the mean, the median, and the mode) are about the same. When data have a mound-shaped, or approximately "normal" distribution, this is the case (the definition of normal in statistics is different than that in normal life). If the distribution is not normal, the values of the mean, median, and mode will not be the same. The next figure, which I shamelessly stole from Zar's (1996. Biostatistical analysis. Prentice Hal) excellent statistics textbook, shows some instances in which these measurements of central tendency differ. This figure illustrates several additional noteworthy comments. Often your data will have asymmetrical distributions, and often they will have more than one mode. A lot of interesting biological processes create asymmetrical and multimodal distributions.

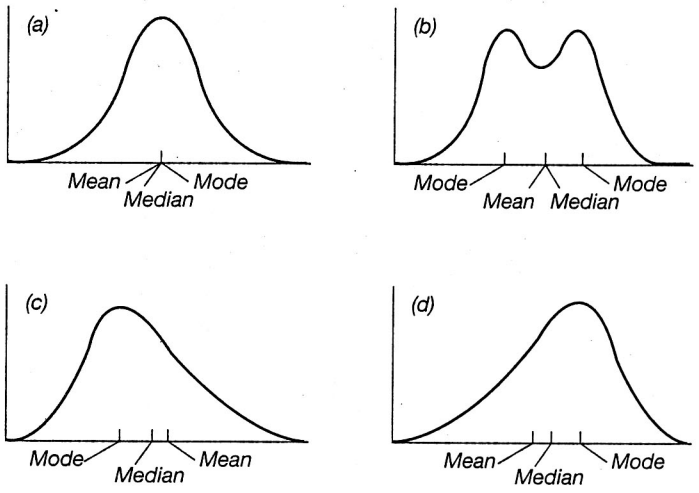


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.†

In addition to having a statistic that measures central tendency (or "average"), it is useful to have measurements of how far variable the data set is. A measurement of variability indicates how spread out measurements are around the average or center of the distribution. The four most commonly used statistics used to measure dispersion are the range, the variance, the standard deviation, and the coefficient of variation. The *range* is the difference between the highest and the lowest value of a data set. For example the range of the data in table 2 is 41. Rather –or in addition to, reporting the range as a number, I prefer to report the range of values by mentioning the lower and upper limits of the measurements in a data set. For the data set in table 3, I would report the range with a sentence:

"The number of fruits per tree ranged from 0 to 41"

The range, by itself, can be ambiguous. You can have the same range if trees produce from 1000 to 1041 fruits, or if they produce from 0 to 41 fruits. The *variance* (s^2) is defined as follows:

$$\text{Variance} = s^2 = SS/N = [1/(N-1)]\ \{\sum(x_i - X)^2\} = [1/(N-1)]\{\sum x_i^2 - (1/N)(\sum x_i)^2\},$$

or using relative frequencies

$$s^2 = SS/N = \sum f_i(i - X)^2$$

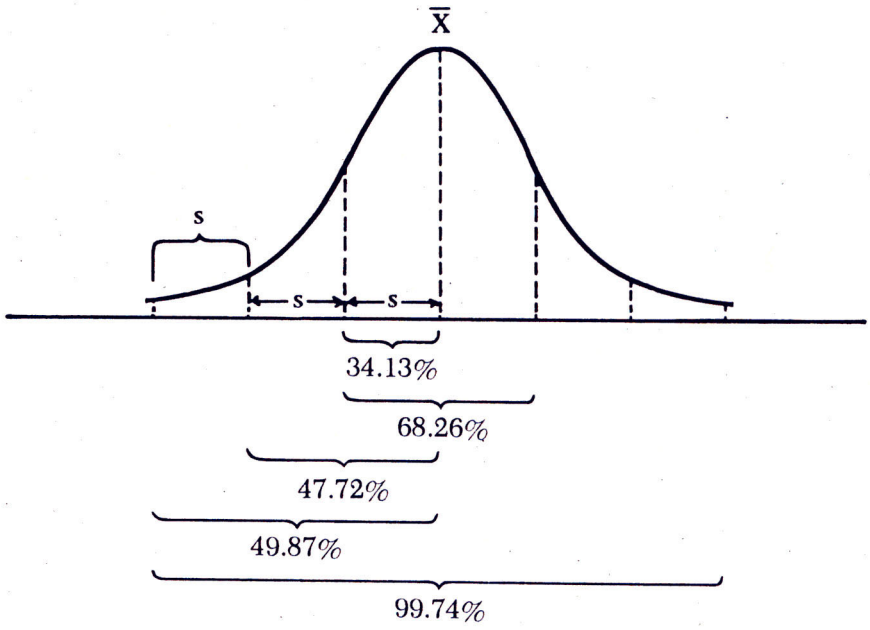
In words, the variance equals the sum of the squares of the deviations (called *the sum of squares* and denoted by SS) around the mean (i.e. the sum of $(xi - X)^2$ for all i) divided by N-1. We use N-1 rather than N, because using N yields a biased estimate. N-1 is often called the *degrees of freedom*. The variance, in essence, is the average of the squared deviations from the mean. Perhaps the *standard deviation* (denoted by s or SD) is used more frequently. The standard deviation is simply the square root of the variance, and therefore has the same units as the measurements in your data set:

$$\text{Standard deviation} = s = \sqrt{s^2}.$$

For normally (i.e. mound shaped) distributed data sets, about 70% (68.27%) of the data lie within one standard deviation from the mean, and about 95% (95.44%) of the data lie within 2 standard deviations from the mean.

Because this observation is true only for data that are normal, you must view with caution. But it is useful in back of the envelope calculations. The last measure of dispersion that we will consider is the coefficient of variation (or CV). In contrast with s which gives you absolute variation, CV tells you how variable a data set is (in %) relative to the mean:

$$\text{Coefficient of variation} = CV = (\text{standard deviation/mean})100 = (s/X)100.$$



Statistical inferences: getting started

One of the most difficult tasks when you are beginning to do science is to select a research topic. Because you are in this course, I assume that you are interested in seed-dispersal in particular and in ecology in general. After you have selected a research topic, you have another difficult chore: selecting a specific question that can be answered. Asking good, interesting, and relevant questions that can be answered is one of the traits that characterizes good scientists.

Variables.- The questions that you will ask most frequently involve associations between variables. Variables can be divided into *control* (or *independent*) variables, and *response* (or *dependent*) variables. This distinction is better illustrated with examples:

Example 1.- Do ripe and unripe fruits of Phoradendron sp differ in sugar content?

The control variable is degree of ripeness which we can measure as a nominal variable (ripe and unripe), or as a an ordinal one (unripe (1), a little ripe (2), ripe (3), very ripe (4), rotten(5)). The response variable is sugar content and is a continuous variable.

Example 2.- Does the rate at which birds visit fruit-bearing trees increase with the size of a fruit crop?

Control variable: fruit crop size (discrete [1000, 2225, 8000 fruits) or ordinal [no fruits, a few fruits, many fruits, gazillions of fruits]).

Response variable: visitation rate (continuous, visits/time).

Example 3.- Is the abundance of *Cecropia sp*. seedlings higher in tree-fall gaps than in the forest interior?

Control variable: Nominal (interior or gap)

Response variable: Abundance of seedlings (seedlings/m2, continuous).

The null hypothesis.- The logic of the traditional scientific framework is based on the rejection of hypotheses (I do not believe in this philosophy, but you must be aware of it and use it). Therefore your questions must be framed as "falsifiable" hypothesis. It is a good exercise to always ask yourself when you pose a hypothesis if there is a set of observations that can show that the hypothesis is wrong. If you cannot think of a data set that can potentially show that the hypothesis is not true, you may want to change the question. Untestable hypothesis that cannot be contrasted with reality (i.e. with a set of measurements on real objects) are not the material of science. This restrictive definition of science forces every hypothesis that you have to be accompanied by a *null hypothesis*. The null hypothesis (denoted by H_0) is a hypothesis of no difference (or no association between the control and the response variable). Normally as you plan your experiment or observation you may reduce your project to a form of shorthand. If we call the alternative hypothesis H_a then for Example 1 you have several possible outcomes that depend on your data set:
Example 1:
H_0: There is no difference in sugar content in fruits of different degrees of ripeness.
H_a: At least one of the degrees of ripeness has higher (or lower) sugar content.
If you are using a binary variable (ripe or unripe), then H_a may take the form
H_a: sugar content(ripe) $\neq$ sugar content(unripe)
For many reasons that will become clearer (I hope!) a bit later, it is better to frame alternative hypothesis in a directional fashion. So instead of asking whether there is a difference, you make a prediction of the direction that the difference will have:
H_a: sugar content(ripe) > sugar content(unripe)
(i.e. sugar content will be higher in ripe than in unripe fruit).
You can even pose a directional alternative hypothesis for ordinal or continuous variables. In the case of example 1 this hypothesis takes the form:
H_a: sugar content will increase with degree of ripeness.
You can restate this hypothesis as follows:
H_a: there will be a positive correlation (or association between sugar content and degree of ripeness).

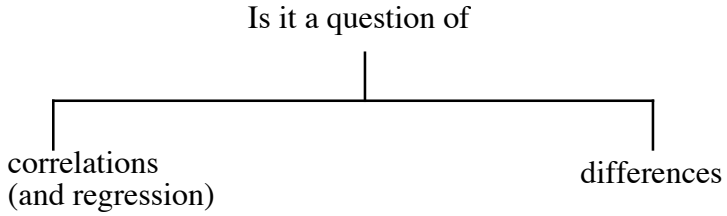
Making your questions "falsifiable" (i.e. having a null hypothesis) is necessary, but it does not make a scientific question good. Many studies pose dumb uninteresting questions, that nonetheless, can be framed as valid testable hypothesis. In addition to making your questions testable, you must make sure that they are interesting and worthy of your time and effort. Asking interesting questions is something that comes from intuition, experience, and good luck. It cannot be taught in a few hours. Some people never learn.

Population and sample statistics.- Statistical methods attempt to do something difficult. They attempt to make inferences about populations from samples. You use measurements in a sample to estimate a statistic that (hopefully…) describes a population. It is useful to have a notation that distinguishes population statistics from their estimates that you calculate from samples. Statisticians use Greek symbols for

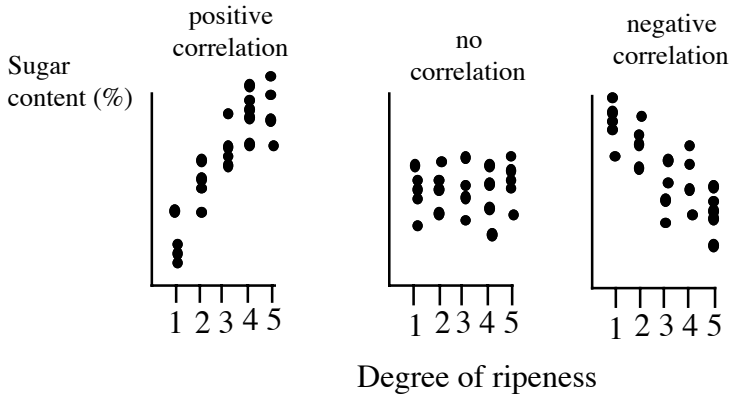
population statistics and Roman letters for their estimates. Sometimes, researchers put a hat on a letter (e.g. $\hat{y}$) to denote that this value is an estimate. Some commonly used symbols are shown in the following table:

Statistic	population	sample estimate
Mean	μ	$\overline{X}$
Variance	σ^2	s^2
Std. Deviation	σ	s
Slope	β_0	$\hat{\beta}_0$
Intercept	β_1	$\hat{\beta}_1$

Is it a question of differences or a question of correlation? One of the first decisions that you will have to make is what is the kind of question that you want to ask. As the examples described above illustrates, we can coarsely divide questions into questions of differences or questions of correlation (and regression).



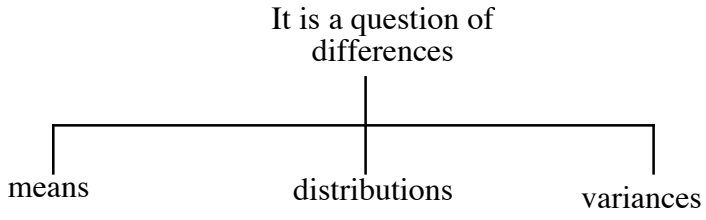
Once again lets use example 1 to illustrate the meaning of correlation and differences. Recall that one of our alternative hypothesis was a positive correlation between degree of ripeness (which was the independent variable) and sugar content. The following figure shows three possible outcomes for this hypothesis:



These diagrams illustrates that sugar content can increase, decrease, or remain the same with degree of ripeness. In following sections, we will describe how we can differentiate among these three possibilities. Before we discuss the question of differences, it is worth mentioning the difference between questions in correlations and questions in regression.

These questions are related, but they are not the same. Questions of correlation simply address the question of whether a dependent variable increases (or decreases) when another one does. Questions of regression attempt to establish the mathematical form of the relationship between two variables, and therefore allows you to make quantitative predictions (if the value of x is such, then the value of y will be….). It also allows you to estimate the degree of uncertainty in your predictions. As we will see, regression is a very powerful tool. Because one of the key ingredients of good science is quantitative prediction, understanding regression is tremendously important.

The questions of differences take three different forms. 1) You can ask whether two (or more) measurements of central tendency differ or if a mean differs from what you expect (for simplicity we will call this difference between, or among, means). In addition, you can ask if the distribution of values in the data set differs from an expected theoretical distribution. Finally, you may want to know if two samples differ in variation.



One sample and paired tests.- Once again lets use example 1 to illustrate how one goes about testing differences between a mean and an expected value. Suppose that you predict that sugar content will be different between ripe and unripe fruit. There are many ways to test this hypothesis. The two simplest ones are as follows: 1) you collect unripe fruit from 100 trees and ripe fruit from 100 **different** trees and then test if there is a difference in average sugar content between the two samples. You need to use two different set of trees because the data points must be independent. We will discuss how you would analyze this measurement design in the following section (comparing between two or more means). The second alternative is to collect 1 ripe and 1 one unripe fruit from each of a hundred trees. This alternative is preferable for two reasons. First, it requires less effort. Second (and much more importantly) if you use a paired design you control for the variation among trees. You end up with the following table:

Tree	ripe	unripe	ripe-unripe	score
1	x_1	y_1	x_1-y_1	+
2	x_2	y_2	x_2-y_2	-
3	x_3	y_3	x_3-y_3	+
4	x_4	y_4	x_4-y_4	0
.				
.				
100	x_{100}	y_{100}	$x_{100}-y_{100}$	+

in which x_i and y_i are the sugar contents in the ripe and unripe fruit of tree i, respectively. The table also includes the difference between x_i and y_i, and a score. We assign a + is

this difference is positive (x_i-$y_i > 0$) a – if the difference is negative (x_i-$y_i < 0$), and a 0 if there is no difference.

What are our hypothesis? If you have a directional prediction then:

H_0: There is no difference in the mean sugar content between ripe and unripe fruits of different degrees of ripeness (i.e $\mu_{x-y} = 0$). (remember that $\overline{x - y}$ is an estimate of μ_{x-y})
H_{a1}: The mean sugar content of ripe fruits is higher than that of unripe fruits ($\mu_{x-y} > 0$)
If you do not have a directional prediction H_a is transformed into
H_{a2}: The mean sugar content of ripe fruits and unripe fruits is different ($\mu_{x-y} \neq 0$).

These two alternatives are tested differently. We call a test that starts with a directional prediction a *one-tailed test*. If the test presumes difference, but has no prediction about the direction of this difference, we call it a *two-tailed test*. We will see why in a moment. Depending on the structure of our data and on our goal we can use a *parametric* or a *non-parametric* test to test the hypotheses that we have posed. Lets begin by describing the parametric test, then we will describe the non-parametric one, and we will finish the section with a discussion about when to use one or the other.

The appropriate parametric test is called a paired-t test. To conduct a paired t test you estimate the t-statistic. You will find this test in books as

$$t_{v,\alpha} = ((\overline{x - y}) - \mu)/s_{(x-y)}$$

where v equals the degrees of freedom (in this case n-1), μ is the expected mean with which we are comparing our variable of interest (in this case $\mu = 0$), and $s_{(x-y)}$ is called the standard error (SE) associated with the mean difference between x and y, and equals the ratio of the standard deviation of the differences (not of the values of x and y, but of **their difference**):

$$s_{(x-y)} = s/\sqrt{N}.$$

What the paired t-test is asking is what is the probability (the chance) that you would get such a t value if the sample was gathered from a population with mean (μ) equal to 0. If the probability is small, then you have a good reason to reject the null hypothesis. The symbol α needs a bit of an explanation. Scientists are very worried about rejecting a null hypothesis when it is true (believeing that there is a difference when there is none). *Rejecting a null hypothesis when it is true is called committing a type I error*. To that effect, they usually set a fixed and relatively low rejection level. α is the rejection level and it is customary to set it at $\alpha = 0.05$. What this means is that we are willing to reject the null hypothesis if the test tells us that there is a probability of 0.05 or less than the sample that we have comes from a population with mean equal to 0 (that is only 5 (or less) in a hundred samples of size N in a population with mean 0 show a mean that is equal that you found in your sample). The custom of setting an α level of 0.05 is, to a

certain extent, the consequence of lack of computers. In the old days, you had to look in a table that listed t values for different values of α and v. Now the computer gives you an exact probability. If p (the probability of getting such a t value) is lower than 0.05, you say that you found a *statistically significant* result.

The way you would report the results of this test would be:

"I found that ripe fruits had significantly higher sugar contents than unripe fruits (mean difference ± SD = 15%, paired two-tailed $t = 6.7$, $p < 0.02$, $N = 100$)"

Note that this sentence includes several elements:

1) the mean difference (or effect size).
2) the type of test and the value of the t statistic.
3) The probability of getting such a difference under the null hypothesis of no difference.
4) whether the test was 1 or 2-tailed
and
5) N, the sample size.

Make absolutely sure to always include these ingredients when reporting the results of a parametric statistical test. As we will see, non-parametric test may not allow you to report the effect size, but you must include all the other ingredients. If you have a directional prediction, the value of t that you need for statistical significance is smaller (remember, the higher the value of t, the lower the p value will be). Always mention if you used 1 or two tailed tests. If you do not mention if the test is one or two tailed, readers will assume that the test is two-tailed.

In addition to this sentence, you may want to include a histogram or a table showing the mean sugar composition of ripe and unripe fruit in your results. Many people get so happy to find a significant result, that they forget to describe what they found. Here I must make a really important point.

Fourth commandment: Do not confuse statistical significance with biological significance

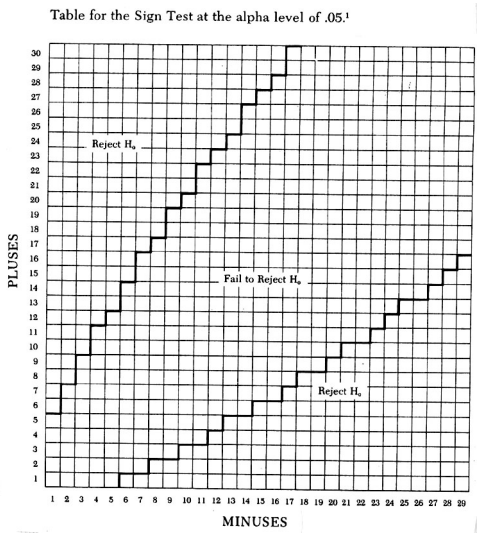
I used big bold letters to really emphasize this point. You must attempt to report effect sizes and mean values, because without them it is often (not always, but often) not possible to know what the biological significance of your result is. For example, the difference in sugar composition between ripe and unripe fruit may be statistically significant, but very small and hence make no difference to fruit eating birds.

The paired t-test is very powerful and very easy to use but it has some requirements and assumptions. The requirement is that the variables in question are continuous data. The assumption is that the sample comes from a normally distributed population. If your

sample is big enough (say N > 30), this assumption is not crucial and the test works just fine.

Fifth commandment: Know the assumption of the statistical test that you are using.

In many cases, you may be comparing paired samples and measuring an ordinal or even a binary variable. For example, you may be looking at the preferences of birds between ripe and unripe fruits (or between two fruit species). The birds will either accept (+) or reject (-) the fruit. Or you may have an unsensitive instrument that simply tells you that the content of sugar in fruit is none (0), a little (1), or a lot (3). You can still use a paired design and the appropriate test is a *sign test*. Sign tests are very easy to use, simply count the + and – (ignore the 0s), and look at the enclosed figure.



If the sample size is bigger than 60 you can use a "normal" approximation:

$$Z = \frac{X - N/2}{\sqrt{\frac{N}{4}}}$$

Where X is the number of +. If Z > 1.96 you can reject the null hypothesis with $p < 0.05$.

The sign test examines the following null hypothesis:

H_o: The sample comes from a population in which 50% of the observations (paired comparisons) show a + and 50% show a – (i.e. ripe fruit and an unripe fruit are as likely to be preferred by birds. The sugar content in ripe fruit and unripe fruit is about equal and hence when you compare them, one is as likely to have a high sugar content as the other).

H_a: The sample comes from a population in which more (or less) than 50% of the individuals. Ripe and unripe fruit have consistently different sugar contents, birds consistently prefer either ripe or unripe fruit over its alternative.

You should report a sign test differently than a paired t-test:

"I found that ripe fruit had more sugar than unripe fruit. In 27 out of 31 pairs of fruits the ripe fruit had a higher sugar content (sign test, $p < 0.05$)".

Note that in this case you are reporting the type of test, the probability of getting such a difference under the null hypothesis of no difference, and N the sample size (N = 31). Because you do not mention whether the test is 1- or 2-tailed, the reader assumes that you used a 2-tailed test. You are not reporting the effect size because the test does not estimate the mean difference. You should use a sign test whenever 1) you have small sample sizes and you suspect that the difference between pairs is not normally distributed, and/or 2) when the response variable is binary or ordinal. The price that you pay from adopting the non-parametric alternative is that you cannot estimate the effect size and that often the test is less powerful.

Confidence intervals.- The parametric alternative allows you to do something very useful, it allows you to construct a confidence interval for the population (not the sample mean). It allows you to say with some probability p, that the population mean estimated by $\bar{x}$ is within a certain interval. Often you will see data reported as

$$\bar{x} \pm 95\%CI$$

This expression gives you the sample mean ± a 95% confidence interval (CI). It allows you to say that the population mean is between the numbers $\bar{x} - 95\%CI$ and $\bar{x} + 95\%CI$ with a probability of 0.95. The confidence interval for a mean is really easy to calculate. If the population has a normal distribution, then

$$95\%\text{CI} = t_{0.05,v}\ \text{SE} = t_{0.05,v}\ (\frac{s}{\sqrt{N}}).$$

Note that the paired t-test asks whether 0 is included in the confidence interval for μ_{x-y}.

This confidence interval ranges from $(\overline{x-y} - t_{0.05,v}\ (\frac{s}{\sqrt{N}}))$ to $(\overline{x-y} + t_{0.05,v}\ (\frac{s}{\sqrt{N}}))$.

The question of power.- We have now used the word "powerful" twice when referring to statistical tests. What do statisticians mean by "statistical power"? To answer this question, we must introduce what statisticians refer to as a *type II error*. A type two error is failing to reject the null hypothesis when you should have rejected it. Most statistics textbooks pay little attention to type II errors. However ecologists and, especially, managers and conservation biologists must pay some attention to it. Why? Biologists may be fooled into believing that there is no pattern simply because the samples size that

they used is too small or because the test that they used is not powerful enough (I am using the word powerful again!).

Imagine that you conduct a test to determine if some human intervention (hunting or logging) has an effect on the abundance of a frugivore. You do this by comparing a large number of hunted and unhunted (or logged and unlogged) plots. You fail to reject the null hypothesis and hence you conclude that this human intervention has no effect. The effect is that the human intervention continues. You used the statistics that scientists always use, and that tend to minimize a or the probability of committing a type I error. Now suppose that the frugivore is endangered. If your conclusion is wrong (i.e. you committed a type II error) and is used to support continued logging, the frugivore goes extinct. It is useful to know the probability of committing a type one error. This probability is called β and there are many statistical methods to estimate it. The statistical power of a test is

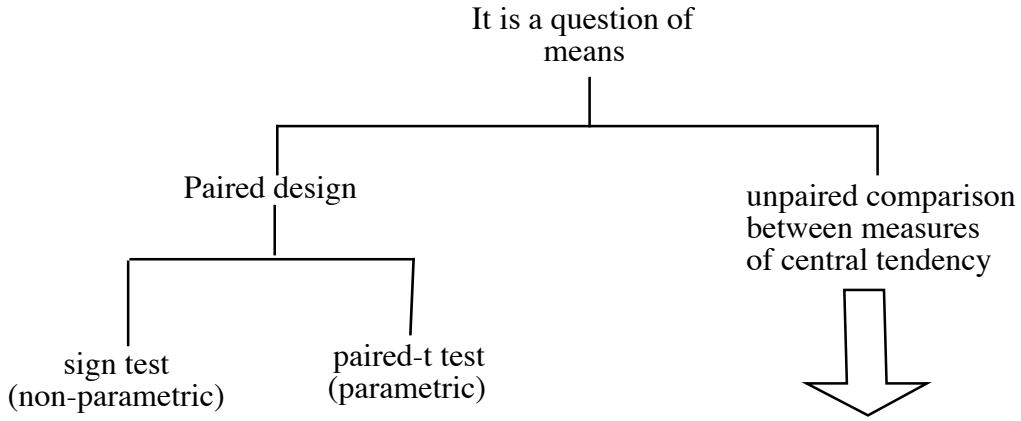
Power = 1-β.

Thus, a powerful test is one in which β is small and therefore, one in which the probability of committing a type II error is small. Calculating β and power can be complicated mathematically. Some of the good computer programs include these calculations. The following web site lists many interactive web sites that allow you to compute power for a given test and data set:

http://members.aol.com/johnp71/javastat.html#Power.

Sixth commandment: Use the most powerful test available, but take care not to violate its assumptions.

Testing for differences between 2 means.- We now have two tests that allow us to compare between measures of central tendency for paired designs. Paired designs are very powerful and you should try to use them whenever you can. It is not always possible to use a paired design, so we need to examine tests that compare among measures of central tendency and that are not based on a paired design.



The unpaired t-test (also called *the two sample t-test*) can be used to test for differences between two means. We have already outlined how to do a one sample test in which we compare a mean to an expected value, so describing a two sample t-test is easy. The test requires that 1) all observations are independent, 2) that you are dealing with continuous data, 3) that the data come from normally distributed populations (although if each sample is larger than about 30m this requirement is not crucial), and if the variances differ significantly, you must use a slightly different formula. After this section, we will explain how to find out if two variances are statistically different.

The procedure is as follows. Go to the field and/or do an experiment in the lab, and when you come back you will have the following table:

Sample number	Treatment (independent variable)	Response (dependent variable)
1	ripe	x_{r1}
2	ripe	x_{r2}
.		
.		
.		
N_1	ripe	x_{rN1}
1	unripe	x_{u1}
2	unripe	x_{u2}
.		
.		
.		
N_2	unripe	x_{uN2}

You have N_1 and N_2 measurements o your two treatments. In this case, I have called the treatments ripe and unripe, but in general, you may want to associate a number with the treatments. It is a good idea to keep N_1 and N_2 equal or as similar as possible. The null hypothesis, of course, is that there is no difference between the means of samples (H_0: $\mu_1 - \mu_2$). Once again, you can use one- or two-tailed tests. If the variances are equal use the following formula:

$$t = \frac{\left(\left|\overline{X}_1 - \overline{X}_2\right|\right)}{\left(\sqrt{\frac{(N_1 - 1)s_1^1 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}\right)\left(\frac{1}{N_1} + \frac{1}{N2}\right)}$$

The test has $N_1+N_2 - 2$ degrees of freedom (d.f.). This formula looks really complicated, but do not worry, you will never have to calculate it. The computer will do it for you. If the variances are unequal, the formula to calculate t becomes easier, but calculating the degrees of freedom (d.f.) becomes difficult:

$$t = \frac{\left(\left|\overline{X}_1 - \overline{X}_2\right|\right)}{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N2}\right)}$$

$$d.f. = \left(\frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\left(\left(\frac{(\frac{s_1^2}{N_1})^2}{N_1 + 1}\right) + \left(\frac{(\frac{s_2^2}{N_2})^2}{N_2 + 1}\right)\right)}\right) - 2$$

Again, please do not worry about having to use these horrible formulas. The computer will calculate them for you. You must remember, however to make sure that the computer is using the correct formulas (you must "tell" the program to do so). Some programs will perform a test for equality of variances and conduct the appropriate test. If you are using a computer, the machine will calculate the p value. If you are not, you can consult the tables in a book (I recommend Zar's (1996)). I printed one of these tables so that you can practice. Note that if the sample size is small, the value of t needed to reach statistical significance is very large. For values of d.f. > 100, you can use a t value of 1.96

as the cut-off for significance at the $p < 0.05$ level. The way you report t tests is as follows:

"Ripe and unripe fruit had significantly different sugar contents (2-tailed t-test, $t = 3.7$, $p < 0.02$). Ripe fruit had higher (mean ± SD = 30 ± 7 %, N = 75) content than unripe fruit mean ± SD = 5 ± 3 %, N = 75)".

Note that, once again, the ingredients in these sentences: test, p-value, mean values, and sample sizes.

Seventh commandment: Always report the test that you used, the statistic for the test, the sample size, the p value, and the effect size of your treatments.

Table of critical values of t for the t Test at the .05 alpha Level.[1]

degrees of freedom	t critical	degrees of freedom	t critical
1	12.706	21	2.080
2	4.303	22	2.074
3	3.182	23	2.069
4	2.776	24	2.064
5	2.571	25	2.060
6	2.447	26	2.056
7	2.365	27	2.052
8	2.306	28	2.048
9	2.262	29	2.045
10	2.228	30	2.042
11	2.201	40	2.021
12	2.179	60	2.000
13	2.160		
14	2.145		
15	2.131		
16	2.120		
17	2.110		
18	2.101		
19	2.093		
20	2.086		

1. Adapted from J.P. Guilford. *Fundamental Statistics in Psychology and Education.* McGraw-Hill Book Company. New York. 1956.

Suppose that your data does not satisfy some of the assumptions of a t-test. For example your data may be ordinal, rather than discrete or continuous, or maybe the data set is small ($N_i < 30$) and does not have a normal distribution (you can check this with a formal test or simply by looking at the frequency distribution of your measurements and finding out if it looks "mound shaped"). Then you can use a non-parametric alternative called the Mann-Witney-U test. Like many statistical tests, you begin this test by ranking the data from highest to lowest. The largest value is given a rank of 1, the second largest the rank of 2,…, until the smallest value which is ranked as N (clearly, $N = N_1 + N_2$). You then calculate the Mann-Whitney U statistic U as

$$U = N_1 N_2 + \frac{N_1(N_1 - 1)}{2} - R_1$$

where R_1 equals the sum of the ranks in sample 1. For a two-tailed test, you must also compute a statistic called U'

$$U' = N_1 N_2 + \frac{N_2(N_2 - 1)}{2} - R2$$

where R_2 equals the sum of the ranks of sample 2. If either U or U' are as great or greater than $U_{\alpha/2, N1, N2}$ (which you must get in tables) than you have the two samples differ significantly. Once again, the computer will perform all these calculations for you. I have described the test in some detail, because you must know something about how the test is done in order to interpret the computer's output. Because you use ranks (rather than the actual measurements) in this test, the test does not allow you to estimate effect size. A statistically significant Mann-Whitney U test, tells you that the distribution of the two samples are shifted relative to each other. Therefore, it is perhaps appropriate to always present the results of this test accompanied by a figure showing both distributions. The results of this test can be written as:

"Ripe fruit and unripe fruit differed sin sugar content (Mann-Whitney U = 36, $p < 0.05$). Ripe fruit had higher sugar content than unripe fruit (Fig. 1)"

You must add the sample sizes to the figure.

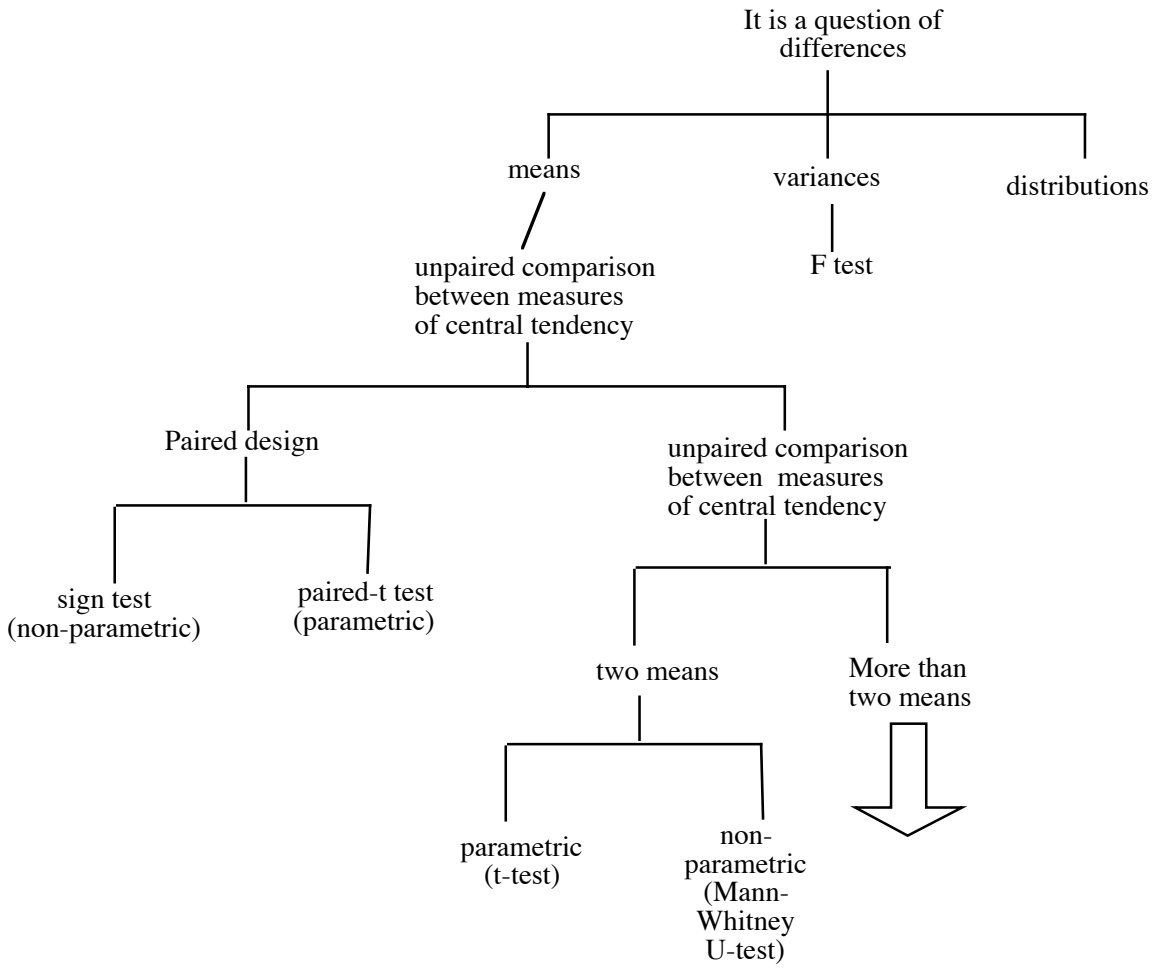
If the assumptions of a parametric t-test are satisfied, use it. The parametric procedure, in general, will be more powerful than the non-parametric one. However, if your data set does not meet the assumptions of the parametric test, the non-parametric test is preferable.

Comparing 2 variances.- Recall that sometimes to choose a statistical test you will need to assess if the variances of two samples are equal. Sometimes it may be really interesting to know if two samples differ in variation. There is a very simple test that answers this question. This test relies on a statistic that will play a central role in the next section (F). This test has equality of variances as a null hypothesis (H_0: $\sigma_1^2 = \sigma_2^2$). Choose the largest value between s_1^2 and s_2^2 and divide this number by the smaller one. As an example let's assume that $s_2^2 > s_1^2$:

$$F_{\frac{N_2 - 1}{N_1 - 1}} = \frac{s_2^2}{s_1^2} \quad F_{N-1, N-1} = \frac{s_2^2}{s_1^2}$$

This test has degrees of freedom for the numerator (N_2-1) and for the denominator (N_1-1). If $F_{calculated}$ is smaller than the $F_{critical}$ that you can find in tables, then you cannot reject the

null hypothesis. Because F values are associated with two sets of degrees of freedom, they can be tricky to use. We will illustrate how to use them in the next section.



Comparing among more than 2 means.- The parametric test that we will use to test for differences between means is called analysis of variance and is often simply referred to as ANOVA. There are many modes of ANOVA. Here I will describe the simplest ANOVA that we will use to determine if two or more means are different from each other. Suppose that you have k treatments (1, 2, 3,…,k), then the null hypothesis is:

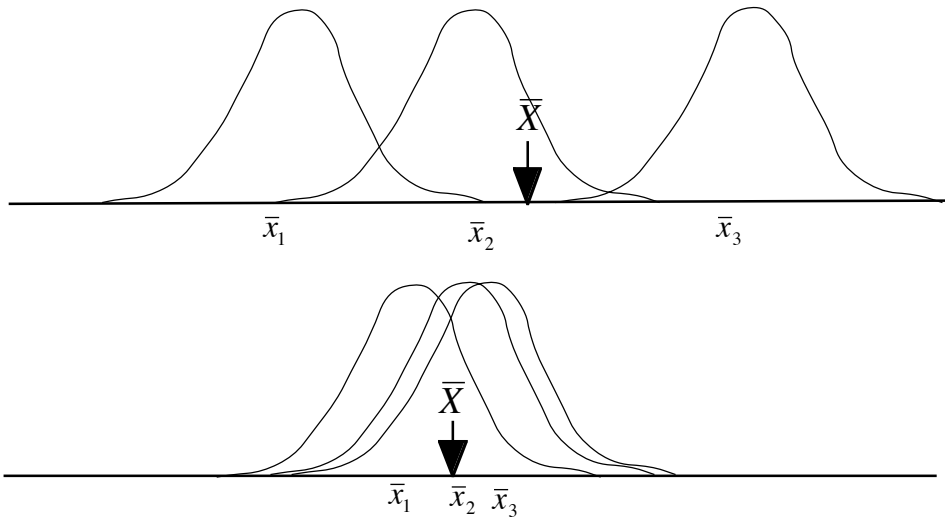
H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

The alternative hypothesis is not that all the means differ, but that at least one of them is different from another.

Although one way ANOVA is a fairly robust test, (meaning that you can violate assumptions a little bit), you must recognize its assumptions. 1) Individual observations must be statistically independent of one another; 2) The observations must be from a continuous scale of measurement (or discrete if the sample in each treatment is large, $n_i > 20$); 3) The observations must be normally distributed (again, this assumption can be

violated if n_i is large), and; 3) The variances of the samples are approximately equal. You can check this assumption with and F test (see Comparing 2 variances) by dividing the largest s^2 and dividing it by the smallest one.

Here you may ask yourselves, why do we call a test about means "analysis of variance"? The reason is subtle. The test asks whether the variation that you find among means is large relative to the variation within treatments. If this value is large, you conclude that there are significant differences among the means.



In the upper panel, the variation among means is large relative to the variation within each treatment. In the lower panel, the variation within means is larger than the variation among means. Suppose that you have k groups (or treatments), each one of which contains n_i observations ($n_1+n_2+n_3+\ldots+n_k= N$). ANOVA uses an F test to compare the variation among means with the variation within treatments:

$$F_{(k-1),(N-k)} = \frac{groups\ \text{MS}}{error\ \text{MS}}$$

Groups MS in this equation refers to an estimate of the variation among groups:

$$\text{group MS} = \frac{\sum_{i=1}^{N} n_i(\bar{x}_i - \bar{X})^2}{k-1} = \frac{SS_{group}}{\text{group degrees of freedom}}$$

The term SS_{group} ($\sum_{j=1}^{N} n_i(\bar{x}_i - \bar{X})^2$) is called the among-group sum of squares and has k-1 degrees of freedom (DF_{group}=k-1).

Errors MS refers to an estimate of the variation within treatments (or groups):

$$\text{error MS} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{N-K} = \frac{SS_{error}}{\text{error degrees of freedom}}$$

The term SS_{error} ($\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$) is called the error sum of squares.

There is another sum of squares that is worth mentioning. Total sum of squares (denoted by SS_{total}) is associated with total degrees of freedom (DF_{error}=N-1) and is calculated as:

$$SS_{total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{X})^2$$

This number is useful because it recognizes that the deviation from the grand mean ($\bar{X}$) of all data is attributable to a deviation of that datum from its group mean plus the deviation of that group from the grand mean ($(x_{ij}-\bar{X}) = (x_{ij}-\bar{x}_i) + (\bar{x}_i-\bar{X})$). The consequence of this is that the total sums of squares equals that sum of the group and error sums of squares:

$SS_{total} = SS_{groups} + SS_{error}$

with degrees of freedom

$DF_{total} = DF_{group} + DF_{error} = N\text{-}1$

You will rarely have to use the total degrees of freedom in a test. Then, why do I mention it? The reason is that in many cases, it is very interesting to find out what fraction of the total variation in a sample is accounted for by variation among groups.

The F test has k-1 and N-k degrees of freedom and is conducted as described in the Comparing 2 variances section. Once again, a high F value (given the appropriate degrees of freedom) leads you to reject the null hypothesis and to conclude that there are significant differences among the means. This does not mean that all the means are different from each other. All you know is that at least one mean differs from another. For example suppose that you have 3 treatments and you reject H_0: $\mu_1 = \mu_2 = \mu_3$. It can be that $\bar{x}_1 = \bar{x}_2 \neq \bar{x}_3$ or $\bar{x}_1 \neq \bar{x}_2 = \bar{x}_3$ or $\bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3$.. You may be tempted to compare between pairs of means using a t-test. AVOID THIS TEMPTATION. Comparing means in pairs will lead to a very high type I error. The following section will describe how to compare pairs of means after you the ANOVA revealed a significant difference.

Multiple comparisons among means.- Often after finding significance in an ANOVA you will want to conduct *multiple comparisons*. There are many, many, ways to do multiple comparisons. I will just describe the one that I use more often. It is called Tukey's honestly significant test (or Tukey's HSD). It considers the null hypothesis

H_0: $\mu_i = \mu_j$. The test is fairly simple. 1) rank the means from highest to lowest ($\bar{x}_a > \bar{x}_c > \bar{x}_d > \bar{x}_b$) and compare the two that are the most different using the formula:

$$q = \frac{\bar{x}_a - \bar{x}_b}{SE}$$

where

$$SE = \sqrt{\frac{MS_{error}}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

and DF = N-k. The critical value of q (that is the minimal value that will allow you to reject H0) depends on k (the number of treatments) and DF. The following table shows the critical values of q for $\alpha < 0.05$. More extensive tables can be found in Zar (1996). The conclusions reached by multiple comparisons testing depend on the order in which you do the comparisons. You should first compare the largest against the smallest, then the largest agains the second smallest, and so on.

TABLE B.5 (cont.) Critical Values of the *q* Distribution

$\alpha = 0.05$

v	*k*(or *p*) = 2	3	4	5	6	7	8	9	10
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474

Reporting ANOVA results correctly often requires a sentence and one or more tables. Suppose that you want to compare the seed content (as % of total weight) in 5 species of fruits:

spp1	spp2	spp3	spp4	spp 5
28.2	39.6	46.3	41.0	56.3
33.2	40.8	42.1	44.1	54.1
36.4	37.9	43.5	46.4	59.4
34.6	37.1	48.8	40.2	62.7
29.1	43.6	43.7	38.6	60.0
31.0	42.4	40.1	36.3	57.3
Means				
32.1	40.2	44.1	41.1	58.3

$n_1=n_2=n_3=n_4=n_5=6$

Source of variation	SS	DF	MS
Total	2437.57	29	
Groups	2193.44	4	548.36
Error	244.13	25	9.76

F=548.36/9.76=56.2

The critical value for $F_{4,25}$=2.76, and hence H_0 is rejected.

To conduct a q test, you rank the means

Species	1	2	4	3	5
Mean%	32.1	40.2	41.1	44.1	58.3

you calculate SE = $\sqrt{\frac{9.7652}{6}} = 1.28$

and for the comparison between species 1 and 5

$$q = \frac{\bar{x}_5 - \bar{x}_1}{SE} = \frac{58.3 - 32.1}{1.28} = 20.47$$

Because $q_{0.05,24,5} \approx 4.16$, we reject the null hypothesis and conclude that $\mu_5 \neq \mu_1$. You have to do that for all combinations of means. If you do it, you will find that species 1 had lower mean seed content than species 2, 4, and 3. You will also find that these 3 species did not differ from each other, but that species 5 had higher content than all the other species. Because the results are complicated, reporting ANOVA results can be cumbersome. This is the way that I would report these results:

"The percentage of seeds, relative to the total weight of the fruit, differed among species (One way ANOVA, $F_{4,25}$= 56.2, $p < 0.01$, table X). Table Y lists the mean percentage of seeds for each species. Briefly, spp. 5, contained the highest percentage of seeds whereas, spp.3 contained the lowest percentage of seeds".

Table X is the "sources of variation" shown above. Many researchers report this table to help readers reconstruct their analysis.

Table X.- The percentage of seeds relative to total fresh weight differed significantly among species. Lines connect means that did not differ from each other (Tukey's HSD test, $p < 0.05$). {OR means with the same letter did not differ from each other (Tukey's HSD test, $p < 0.05$)}.

Species	1	2	4	3	5
	32.1[a]	40.2[b]	41.1[b]	44.1[b]	58.3[c]

I encourage to reconstruct these results either by hand, or using a computer.

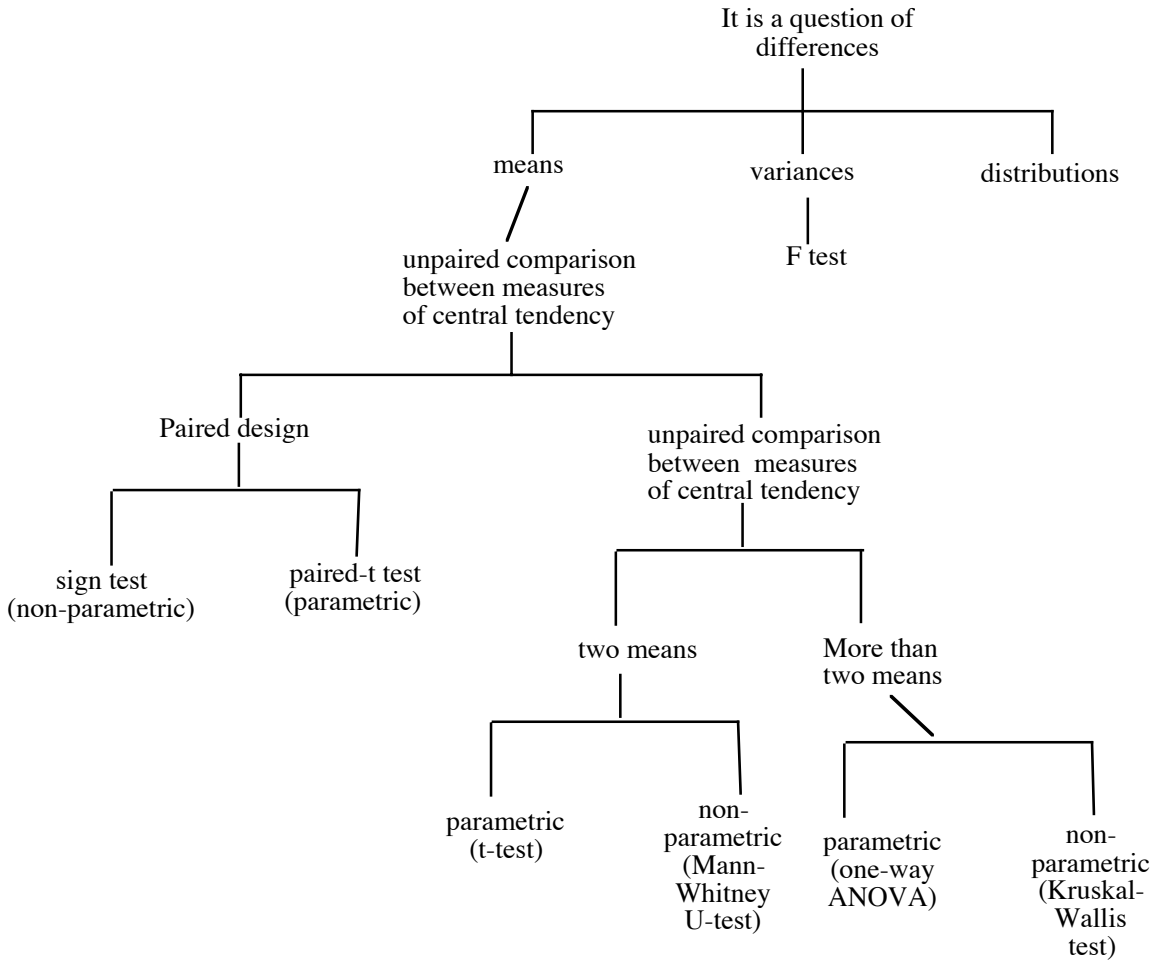
Non-parametric analysis of variance (the Kruskal –Wallis test).-The *Kruskal-Wallis test* is often called an "*analysis of variance by ranks"*. It can be used on ordinal data and with continuous or discrete data if the assumptions of one way ANOVA are not met. If the assumptions of ANOVA are met, use ANOVA because it is more powerful. However, if the samples are small ($n_1 < 20$) and come from populations that are clearly not normal, or if you reject the null hypothesis of equality of variances, then the Kruskal-Wallis test is an appropriate alternative. As in other non-parametric tests, we do not use population parameters in statements of hypothesis, and neither parameters nor sample statistics are used in the test calculations. The Kruskal-Wallis test statistic, H, is calculated as

$$H = \left\{ \left(\frac{12}{N(N+1)} \right) \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right\} - 3(N+1)$$

where n_i is the number of observations in group i, $N = \sum_{i=1}^{k} n_i$, and R_i is the sum of the ranks of the n_i observations in group i. Note that if k=2, the Kruskal-Wallis test becomes a Mann-Whitney U test. For intermediate to large samples (i.e. $n_i > 20$). H can be tested using a $\chi 2$ table with degrees of freedom equal to k-1. You I added this table to the end of these notes. You can conduct multiple comparisons after a Kruskal-Wallis test. However, these comparisons allow you to determine whether the sum of the ranks between two groups differ. I find these comparisons difficult to interpret. Most statistics packages have the Kruskal-Wallis test in their menus.

An opinionated note on parametric VS. non-parametric tests.- Although many researchers prefer non-parametric tests, I do not. I attempt, as frequently as possible to use parametric tests. The reason is that I am more interested in measuring effect sizes and in making predictions. As we have seen, non-parametric tests allow making statistical inferences,

but in general they do not allow to estimate size effects. The next section discusses some of the most basic statistical tools that you need to make predictions.



Correlation and regression.- You ask questions of *correlation* and *regression* when you have measured two variables in each member of a sample (correlation and regression are methods on *paired data*). For example you may predict that the number (or biomass) of mistletoe parasites per tree host increases with the size of the tree (measured as diameter at breast height [DBH], or total height). In each tree you measure height (x) and parasite load (y). There are many, many questions that you can approach using correlation. When two variables are correlated, the magnitude of one changes with the magnitude of the other. However, we can establish no cause and effect relationships (to do so requires, often requires doing experiment). In questions of correlation, we are simply interested in asking whether the two variables increase or decrease together. Regression is a powerful statistical technique that allows you to estimate the mathematical form of the relationship between a dependent (response) variable (y= the number of mistletoes per tree) and an independent (control) variable (x = the size of the tree). Regression allows you to

estimate how accurate your predictions will be. It allows you to determine how well can you predict y from knowing x. Correlation and regression are closely related, but they are not the same.

Very often you will be confronted with situations in which you are interested in the relationship between 2 variables. Then you will have to ask yourself two questions: 1) am I interested simply in knowing if the variables are positively (or negatively correlated)? or 2) am I interested in the functional (mathematical) form of the relationship between these two variables? Sometimes you may be interested in answering 2) but your data set may not satisfy the assumptions of regression. If that is the case, you may have to simplify your question to one of correlation. I hope to convince you that if your data set satisfies the conditions for regression then you should use it. **Independently of whether the question that you have is one of correlation of regression, always plot your data.** The characteristics of your data and the form of this plot should tell you a lot about whether to ask questions of correlation or of regression.

There are several situations in which you have no alternative but to ask a question of correlation:

A) One (or both) of the variables is ordinal. For example you may ask if small, medium sized , or large trees have no mistletoes (1), a few mistletoes (2), or lots of mistletoes (3).

B) One (or both) of the variables is discrete or the independent variable is binary. In the past 15 years a large number of methods have been developed that allow using regression methods on discrete and binary response variables (they are called logistic and Poisson regression methods). These methods are relatively new, and therefore they are not described in most of the introductory statistics textbooks. However, they are tremendously important and their use is becoming widespread among ecologists. Because they are advanced, we cannot deal with them here. I recommend Ramsey and Schafer (1997. The statistical Sleuth. Duxbury Press.) as an introduction to these regression methods.

C) The relationship between x and y is clearly non-linear. In a subsequent section we will describe a few methods that will allow you to diagnose linearity. Sometimes if the relationship is non-linear, you can still fit a function. You can use very simple regression methods to fit polynomials (i.e. functions of the form $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n$) or you can use more complicated methods to fit other non-linear functions. Again, these notes will not deal with non-linear procedures. I recommend Motulsky and Ransanas (1987. FASEB Journal 1: 365-374) as a friendly and non-mathematical introduction to non-linear regression.

Lets begin by assuming that one or more of these caveats apply, and you must conduct a correlation rather than a regression analysis. The test that I recommend is the Spearman Rank Correlation. It is a simple non-parametric test. You can use this test for data that are ordinal, discrete , or continuous. The individual data points must be independent. The test

statistic is indicated by "r_s". And the null hypothesis is that there is no correlation between the two variables (H_0: $r_s = 0$). To do this test, you need to
1) Rank the variables for each data point within the two groups. Tied absolute values, each get the average rank of thos two values if they had not been tied. If this is not clear, see the example that follows.
2) Calculate the difference between the ranks (d_i) and calculate its square (d_i^2)
3) Sum the square of the differences($\sum_{i=1}^{N} d_i^2$)
4) apply the following formula:

$$r_s = 1 - \left(\frac{6 \sum_{i=1}^{N} d_i^2}{N^3 - N} \right)$$

5) Compare the calculated statistic r_s with the critical value given in the following table for the appropriate sample size.

In the following example, you are interested in finding out if visit rate (number of birds arriving to a tree per hour) is correlated with fruit abundance (measured as the number of ripe fruits per tree) in *Virola sebifera*. Because it is unclear if the relationship is linear or not (see figure), you decide to conduct a Spearman rank correlation test.

Tree	# fruit	Vi/h	Rank x	Rank y	d	d^2
1	30	10	7.5	21	-13.5	182.25
2	60	1.5	21	12	9	81
3	45	1.5	16	12	4	16
4	35	0	10.5	3.5	7	49
5	40	3	12	17.5	-5.5	30.25
6	45	1	16	9	7	49
7	0	0	1.5	3.5	-2	4
8	45	2	16	14	2	4
9	30	1.5	7.5	12	-4.5	20.25
10	0	0	1.5	3.5	-2	4
11	45	4	16	19	-3	9
12	45	7	16	20	-4	16
13	35	2.5	10.5	15.5	-5	25
14	50	2.5	20	15.5	4.5	20.25
15	45	3	16	17.5	-1.5	2.25
16	45	0	16	3.5	12.5	156.25
17	15	1	4	9	-5	25
18	10	0	3	3.5	-0.5	0.25
19	20	1	5	9	-4	16
20	30	0.5	7.5	7	0.5	0.25
21	30	0	7.5	3.5	4	16

Note how you you calculate the ranks: In the fruit # column, 2 trees had no fruit. Their ranks would have been 1 and 2, and hence their rank is $\frac{1+2}{2} = 1.5$. In the visits/h column (Vi/h), 6 trees received no visits and hence their rank is $\frac{1+2+3+4+5+6}{6} = 3.5$.

$$\sum_{i=1}^{N} d_i^2 = 726$$

and therefore

$$r_s = 1 - \left(\frac{6(726)}{21^3 - 21} \right) = 0.53$$

because $0.53 > 0.435$ (from the enclosed table), you reject H_0 and conclude that there is a positive correlation. How would you report this result?

" In *Viola sebifera,* visitation rate by frugivores increased significantly with the number of ripe fruits per tree (rs = 0.53, $p < 0.05$, N = 21, Fig. G)."

I hope that you have noted that I always report results using the past tense. Editors are reviewers of your manuscripts expect you to do it as well.

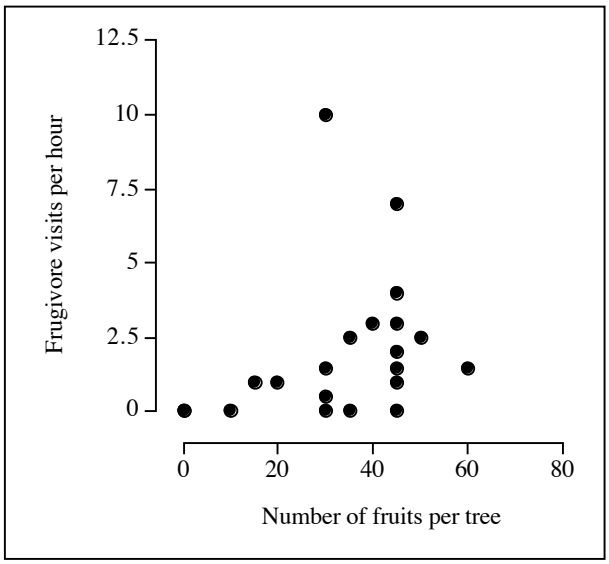


Table of critical values for different sample sizes at the .05 alpha level to be used with the Spearman's Rank Correlation test.[1] (n = sample size)

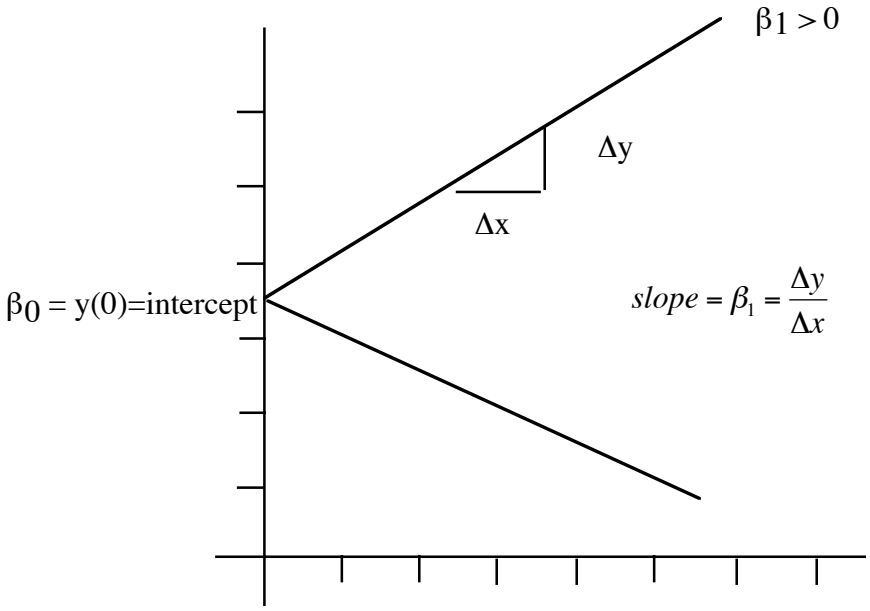
n	critical value	n	critical value	n	critical value	n	critical value
5	1.00	27	0.382	49	0.282	92	0.205
6	0.886	28	0.375	50	0.279	94	0.203
7	0.786	29	0.368	52	0.274	96	0.201
8	0.738	30	0.362	54	0.268	98	0.199
9	0.700	31	0.356	56	0.264	100	0.197
10	0.648	32	0.350	58	0.259		
11	0.618	33	0.345	60	0.255		
12	0.587	34	0.340	62	0.250		
13	0.560	35	0.335	64	0.246		
14	0.538	36	0.330	66	0.243		
15	0.521	37	0.325	68	0.239		
16	0.503	38	0.321	70	0.235		
17	0.485	39	0.317	72	0.232		
18	0.472	40	0.313	74	0.229		
19	0.460	41	0.309	76	0.226		
20	0.447	42	0.305	78	0.221		
21	0.435	43	0.301	80	0.220		
22	0.425	44	0.298	82	0.217		
23	0.415	45	0.294	84	0.215		
24	0.406	46	0.291	86	0.212		
25	0.398	47	0.288	88	0.210		
26	0.390	48	0.285	90	0.207		

1. Adapted from J.H. Zar. *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, N.J. 1974.

In *regression analysis* we are interested in two objectives: First, we are interested in finding out if there is a relationship between 2 variables (or between a dependent variable and several independent variables. Second, we also would like to find out, given a model to describe the data, what are the best possible estimates for the statistics in this model. In the case of linear regression, our model is a function of the form

$Y = \beta_1 X + \beta_0$

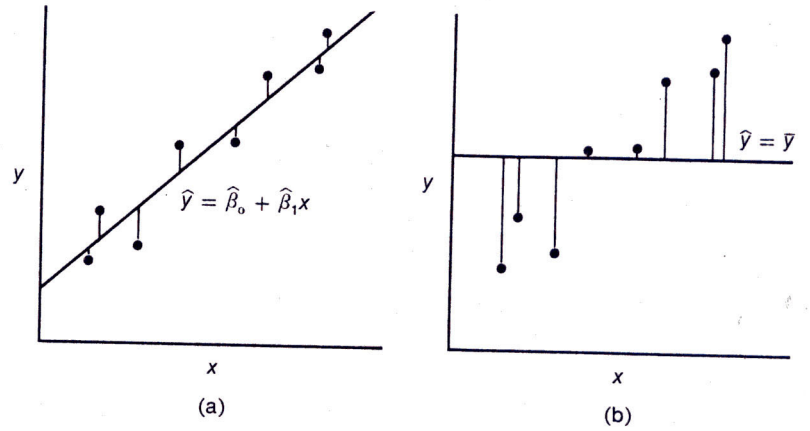
In which β_1 is the slope, and β_0 is the intercept. The meaning of the slope is the change in y when x increases by 1 unit. Its units are (units of y/units of x). If $\beta_1 > 0$, y increases with x. If $\beta_1 < 0$, y decreases as a a function of x. The meaning of the intercept is the estimated value of y when x = 0.



One of the purposes of linear regression analysis is to estimate the "best" value for m and b. The line that you derive using regression analysis is called the "line of best fit". The line of best fit is obtained by finding the numbers a and b that minimize the following sum of squares:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - [\beta_1 x_i + \beta_0])^2$$

Where $\hat{y}_i$ is the predicted value of y for $x = x_i$ ($\hat{y}_i = mx_i + b$). In words this means that the line of best fit is that for which the sum of the squares of the distance between the points and the line is as small as possible. Panel (a) describes a good fit in which the distance between the points and the regression line is small. In contrast, panel (b) describes a poor fit between the points and the regression line.



The assumptions of linear regression analysis are:

1) Pairs of measurements (x, y) are independent from each other.
2) The value of X is measured without error (or with a small relative error).
3) The y scale must be continuous (x can be discrete or continuous).
4)The test assumes that the variance around the regression line is the same (i.e that the scatter of points around the regression line is more or less the same for all values of x).

You calculate the following statistics:

$$\hat{\beta}_1 = \frac{SS_x}{SS_{xy}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

You can then test the following null hypotheses:
1) H_0: $\beta_1 = 0$ (this means that the slope = 0) using

$$t = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{SSE}{N-2}}} \right) \sqrt{SS_x}$$

with DF=N-2.

2) H_0: $\beta_0=0$ (is the intercept 0?)

$$t = \frac{\hat{\beta}_0}{\left(\sqrt{\frac{SSE}{N-2}}\right)\left(\frac{1}{N} + \frac{\bar{x}}{SS_x}\right)}$$

with N-2 degrees of freedom.
I would be very surprised if you ever have to calculate these statistics by hand (that is what computers are for!). But it is useful to know that they exist.

A very useful statistic in linear regression is the *coefficient of determination* r^2. You may decide to ignore many of the formulas that I have placed in this handout. DO NOT IGNORE THIS ONE.

$$r^2 = \frac{SS_y - SSE}{SS_y} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

It is useful to write this equation in words:

$$r^2 = \frac{\text{variation in y - variation in y explained by the regression line}}{\text{variation in y}}$$

The coefficient of determination r^2 varies from 0 to 1 and it tells you what fraction of the total variation in the dependent variable y is accounted for by the relationship between y and x. The coefficient of determination r^2 is a very important number because it tells you how well your linear model fits your data. If r^2=0.71, for example, this means that 71% of the variation in y is accounted for by the relationship between x and y.

Another statistic that you will encounter is r, the *Pearson product moment coefficient of correlation* (or simply *correlation coefficient*):

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

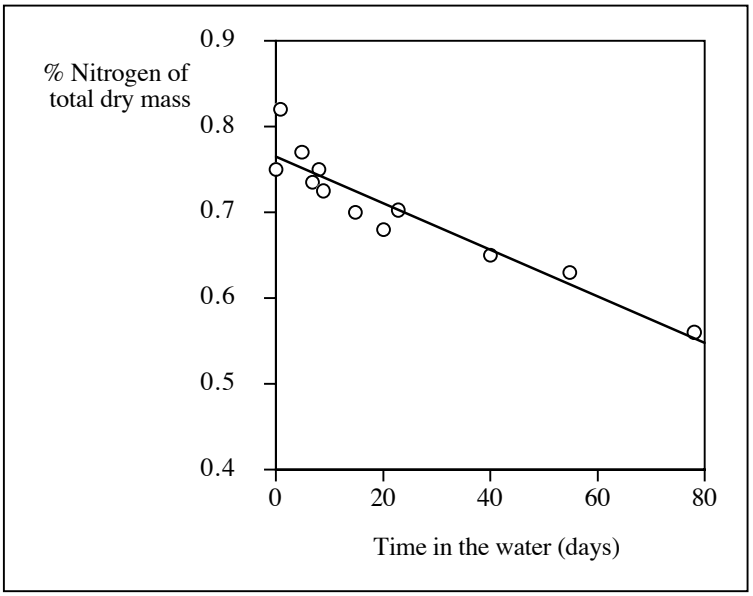
The coefficient of correlation ranges from –1 to 1. If its value is negative, x and y are negatively correlated. If it is positive, x and y are positively correlated. The coefficient of correlation is the parametric equivalent of the Speraman rank coefficient of correlation. You can test the null hypothesis of no correlation (H_0: r = 0) with the following test:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

which has N-2 degrees of freedom. I do not use r very frequently, but some researchers do.

Lets illustrate what we have learned about linear regression with an example. Susan Moegenburg hypothesized that the fruits of the palm *Euterpe oleracea* leached nutrients to water in flooded forests. She placed fruit in water and measured the percent nitrogen in fruit after different time intervals. The following table shows her results (frm Moegenburg 2002. Pp. 479-494 In Levey, Silva, and Galetti (eds.) Seed dispersal and frugivory. CABI Publishing).

Time	% Nitrogen
0	0.75
1	0.82
5	0.77
7	0.736
9	0.725
8	0.75
15	0.7
20	0.68
23	0.704
40	0.65
55	0.63
78	0.56



The results of a regression analysis are:

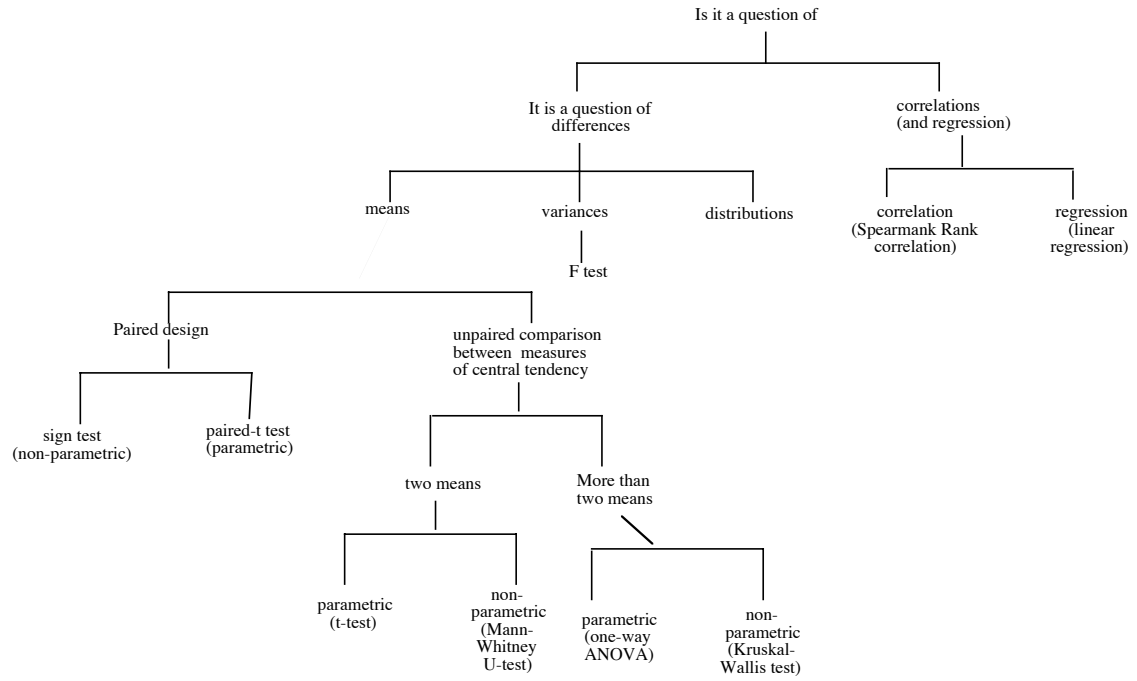
Parameter estimates	estimate	t	probability
β_0	0.77	78.1	<0.001
β_1	-0.003	8.8	<0.001

$r^2 = 0.89$

You conclude that fruit's nitrogen content decreases linearly with the time that it spends submerged in water. The relationship describing % nitrogen as a function of time is:

% Nitrogen = 0.77-0.003(time).

Note that variation in time "explains" about 90% of the variation in nitrogen content. The slope tells you that 0.003 % Nitrogen is lost per day.



The figure in this page is almost complete. We have discussed questions of correlation between variables, and of differences between means and variances. We will finish these notes by describing how to compare between frequency distributions.

Comparing an observed distribution with an expected one.- Lets use an example to motivate our description of these methods. Imagine that you are studying whether previous parasitism by a mistletoe influences the frequency with which birds deposit mistletoe seeds into tree hosts. You conduct a census and obtain the following data set:

	Parasitized	Non-parasitized	Total
Seeds present	19	5	24
No seeds	25	76	101
Total	44	81	125

This table is called a 2X2 contingency table (it has 2 factors and two levels in each factor). A 2X2 contingency table contains 4 cells representing all possible outcomes. You can construct contingency tables of any size ($N_1XN_2XN_n$). They are easy to analyze and almost impossible to interpret. I would stick to small contingency tables.

You are interested in finding out if seeds fall disproportionately into already parasitized trees. One possible way of answering this question is to compare the *observed frequency* (or distribution) with the distribution of frequencies that you would find if the trees received trees in proportion to their abundance. How can you construct this *expected*

distribution? To obtain the expected distribution we use the following elementary rule of probability: if two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. For example, you know that 0.192 = 24/125 of all trees received seeds. You also know that 44/125 =0.152 of all trees was parasitized. Therefore the probability of being parasitized and receiving seeds equals:

$$\left(\frac{24}{125}\right)\left(\frac{44}{125}\right) = 0.0292$$

and the expected number of parasitized trees receiving seeds should equal:

125(0.0292) = 3.65.

Lets call the levels of parasitism. Lets call the total number of observations N, and assume that each cell can be characterized by its two levels: i (i can be parasitized or non parasitized) and j (j can be received seeds or no seeds). For example $n_{\text{parasitized, no seeds}} = 5$. In general, you can easily estimate the expected value for each cell (E_{ij}) as:

$$E_{ij} = N\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right) = \frac{n_i n_j}{N}$$

To compare between observed and expected values, we can place the expected values in parenthesis in the contingency table:

	Parasitized	Non-parasitized	Total
Seeds present	19(3.65)	5(15.52)	24=n_{seeds}
No seeds	25(35.55)	76(65.45)	101=$n_{\text{no seeds}}$
Total	44	81	125
	$n_{\text{parasitized}}$	$n_{\text{non parasitized}}$	

This new table indicates that your conjecture may be correct. Parasitized host trees seem to have received more seeds that you would expect given their frequency and non-parasitized trees fewer seeds. You can test this conjecture using the following formula:

$$\chi^2 = \sum_j^r \sum_i^c \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum \frac{(observed - \exp ected)^2}{\exp ected}$$

Where O_{ij} and E_{ij} are the observed and expected absolute frequencies in a cell in column i and row j (I assumed a table with c columns and r rows). This test has (r-1)(c-1) degrees

of freedom. In the case of a 2X2 table, DF=1. Then you compare the value of $\chi 2$ with a critical value from the enclosed table. In the example above

$$\chi^2 = \frac{(19-3.65)^2}{3.65} + \frac{(25-35.5)^2}{35.5} + \frac{(5-15.52)^2}{15.52} + \frac{(76-65.45)^2}{65.45} = 76.48$$

Because the critical value for $\chi 2$ at the α=0.05 level is 3.84, we reject the null hypothesis. In this case the null hypothesis is that parasitized and unparasitized host trees receive mistletoe seeds in proportion to their abundance (H_0: $O_{ij} = E_{ij}$ for all i and j). The $\chi 2$ test ie enormously useful. It can be used in all cases in which you would like to compare and observed distribution with an expected one. Always report both the observed and the expected values when you conduct a contingency table (or distribution comparison) analysis.

X^2 Table at the .05 alpha level.[1]

d.f.	critical value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.1
6	12.6
7	14.1
8	15.5
9	16.9
10	18.3
11	19.7
12	21.0
13	22.4
14	23.7
15	25.0
16	26.3
17	27.6
18	28.9
19	30.1
20	31.4