

Introductory Biostatistics

Course notes

Frederick S. Scharf
Biology and Marine Biology
UNCW

These course notes represent a set of lectures that I wrote and organized for an introductory graduate level course in biometry. Although I organized the notes and contributed my own ideas throughout, I have drawn extensively from several texts. Many of the ideas contained in these notes build upon or are taken directly from ideas presented by the authors of those texts. When an example or an idea that improves explanation of a concept is based on material presented in a previous text and used with little or no modification on my part, I have tried to cite the text and the location of the material. Any omissions of such citations are my errors. The list below includes published texts that I have drawn from in the creation of these course notes. The first three texts listed were used most extensively.

1. **A Primer of Ecological Statistics** (1st edition; 2004) by Nicholas J. Gotelli and Aaron M. Ellison
2. **Biometry** (3rd edition; 1995) by Robert R. Sokal and F. James Rohlf
3. **Biostatistical Analysis** (4th edition; 1999) by Jerrold H. Zar
4. **Design and Analysis of Ecological Experiments** (2nd edition; 2001) by Samuel M. Scheiner and Jessica Gurevitch
5. **Ecological Methodology** (2nd edition; 1999) by Charles J. Krebs

6. **Experimental Design and Data Analysis for Biologists** (1st edition; 2002) by Gerry P. Quinn and Michael J. Keough
7. **Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance** (1st edition; 1997) by A. J. Underwood

In addition to the above texts, these course notes also benefited from ideas and examples contained in the course notes for Public Health 540 and 640; two graduate level biostatistics courses taught during the 1994-95 academic year at the University of Massachusetts, Amherst by Drs. David W. Hosmer and Stanley Lemeshow.

Introduction to Biostatistics

First, some definitions:

What is Biostatistics exactly?

- The application of statistical methods to the solution of biological problems

Statistics = the scientific study of data describing natural variation (Sokal and Rohlf 1995)

Scientific study = objectivity

Data = information about populations or groups of individuals (data is plural since statistical testing can't be performed on a single datum)

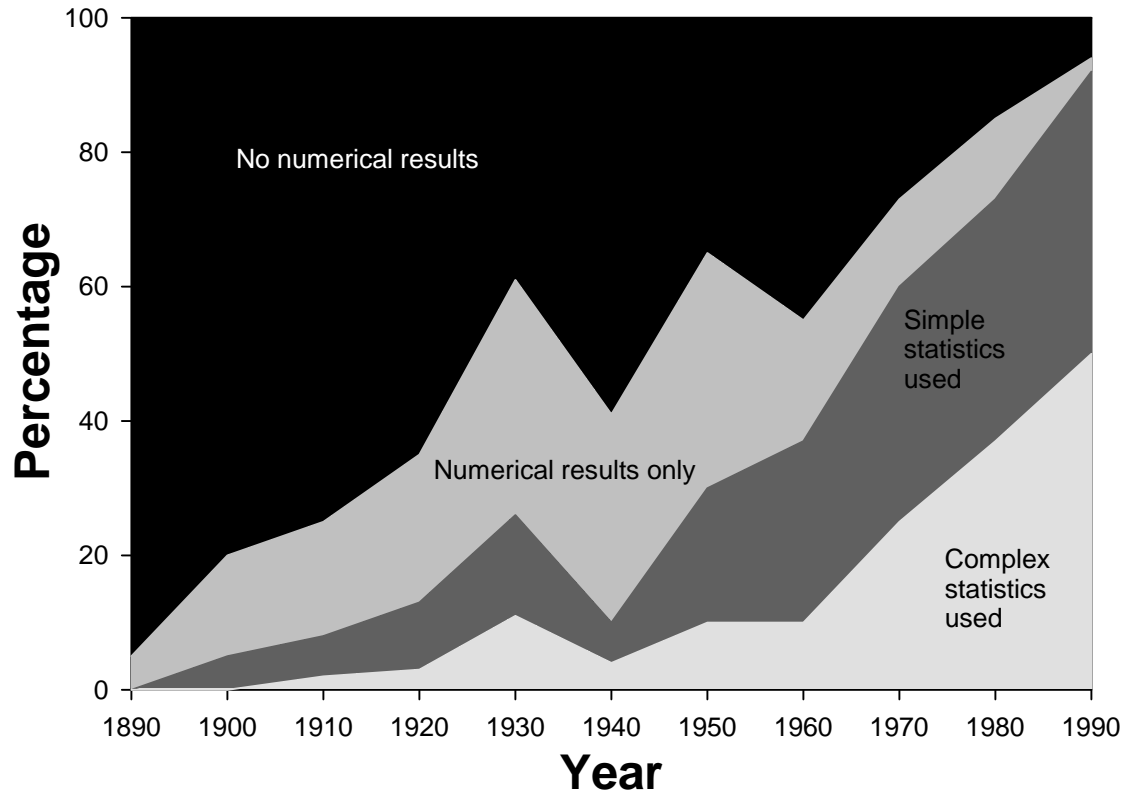
Natural variation = events that happen in nature not under the direct control of the investigator, plus those events that are evoked by and are, at least partly, under the control of the investigator (Sokal and Rohlf 1995)

Why do we care?

-Increased use of statistics in all disciplines within biology

-Realization that biological phenomena are affected by multiple causal factors that cannot always be identified or controlled

-These factors vary and their interactions generate large amounts of variation



Percentage of articles published in *The American Naturalist* using statistical analyses (modified from Sokal and Rohlf 1995)

-We need statistics to generate quantitative measures of observed phenomena and to assess the probability of measured differences

-Statistics, thus, places biological phenomena within a probabilistic framework (Sokal and Rohlf 1995)

-It represents a common language with which we can interpret the quantitative measures of our observations

A conceptual example:

Suppose you are walking through campus and are interested in quantifying the density of students. Your question might be “What is the best estimate of student density at UNCW?” Is it 1 student per 10m^2 ? or 5 students per 10m^2 ? How should you measure student density? Does it vary in different places on campus? at different times of the academic year?

Ultimately you should ask “What mechanisms or hypotheses might account for the variation observed?” and “What experiments or observations could be made to test these hypotheses?”

Statistics allows us to summarize and interpret the data (quantitative measurements) after we have made our observations. We can then test and differentiate among our hypotheses.

For many people, in the simplest sense,

Statistics \approx Patterns

Biological Data

Individual observations – measurements or data taken on the smallest sampling unit

Sample = a collection of individual observations

Population = totality of individual observations about which inferences are to be made (defined and justified by the investigator; *often not explicitly defined, but implied instead*)

When we make individual observations, the actual property measured is called a **variable** (length of a fish; number of plant leaves; etc.), and there are many types of variables

Types of variables

Ratio scale data

- Constant size interval between adjacent units on the measurement scale
- There exists a zero point on the measurement scale, which allows us to talk in terms of the ratios of measurements (e.g., x is twice as large as y)
- Most data on a ratio scale (examples include lengths, weights, numbers of items, volume, rates, lengths of time)

Interval scale data

- Constant interval, but no true zero, so can't express in terms of ratios
- Temperature scale is a good example (zero point is arbitrary; *can't say 40° is twice as hot as 20°*)
- Other biological examples could be time of day and lat/long

Ordinal scale data

- Data consist of an ordering or ranking of measurements only
- Exact measurement data unknown or not taken (e.g., we may only know larger/smaller, lighter/darker, etc.)
- Often ratio or interval data is converted to ordinal data to aid interpretation (i.e., exact measurements assigned ranks) and statistical analysis (e.g., grades)

Nominal scale data

- Data doesn't have a numerical measurement
- Eye color, sex, with or without some attribute

Continuous and Discrete data

- A continuous variable can take any value within the measured range

For example, if we measure fish length, the variable can be an infinite number of lengths between any two integers (*thus, we are only limited by the sensitivity of our measurement devices*)

- A discrete variable can generally only take on values that are consecutive integers (no fractional values are possible)

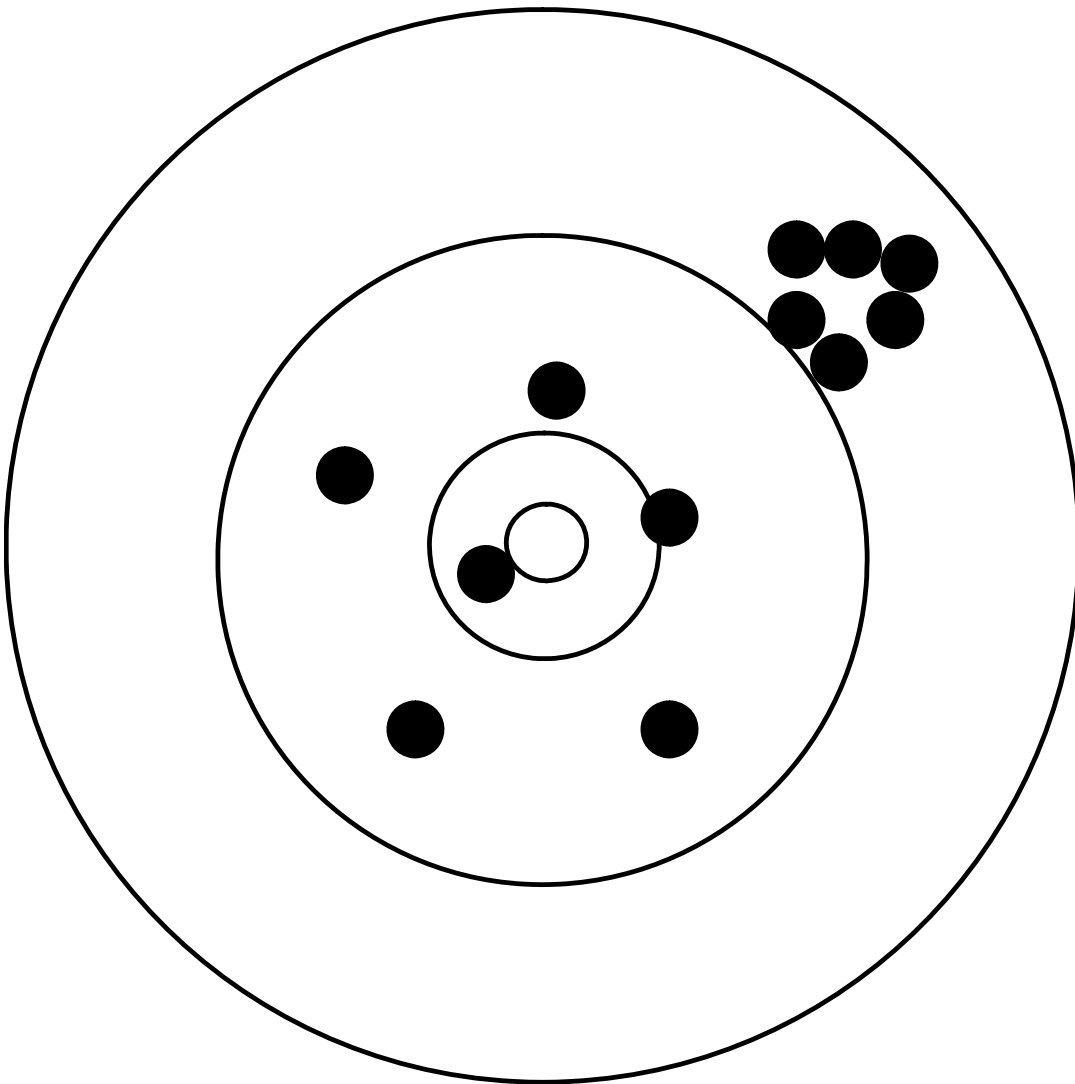
For example, if we count the number of ants in a colony there can be 221 ants or 222 ants, but not 221.5 ants

Nominal scale data are *always* discrete; other data types can be either continuous or discrete

Accuracy and Precision

Accuracy = closeness of a measured value to its true value
(**Bias** = inaccuracy)

Precision = closeness of repeated measurements of the same quantity
(**Variation** or variability = imprecision)



Many fields within biology differ in their ability to measure variables accurately and precisely

Most continuous variables are approximate, while discrete are exact

Significant Figures

The last digit of measurement implies precision = limits of measurement scale between which the true measurement lies

A length measurement of 14.8 mm implies that the true value lies between 14.75 and 14.85

***The limit always carries one figure past the last significant digit measured by the investigator

Rule of thumb for significant figures (Sokal and Rohlf, p. 14)

The number of unit steps from the smallest to the largest measurement in an array should usually be between 30 and 300

Example: If we were measuring the diameter of rocks to the nearest mm and the range is from 5-9mm, that is only four unit steps from smallest to largest and we should measure an additional significant figure (e.g., 5.3 – 9.2 mm, with 39 unit steps). In contrast if we were measuring the length of bobcat whiskers within the range of 10-150mm, there would be no need to measure to another significant figure (we already have 140 unit steps)

Reasoning: The greater the number of unit steps, the less relative error for each mistake of one measurement unit. Also, the proportional error reduction decreases quickly above high numbers of unit steps (300), making measurement to this level of precision not worthwhile

Examples of significant figures

22.34 (4)	25 (2)	0.065 (2)	0.1065 (4)
14,212 (5)	14,000 (2)		

Derived variables

A variable expressed as a relation of two or more independently measured variables (e.g., ratios, percentages, or rates)

These type of variables are very common in the field of biology; often times their construction is the only way to gain an understanding of some observed phenomena.

We will deal with the statistical issues with ratio data a bit more later, but for now we just need to mention that they present certain disadvantages when it comes to analysis. These are related to their inaccuracy (compounded when independent variables are combined) and their tendency to not be distributed normally

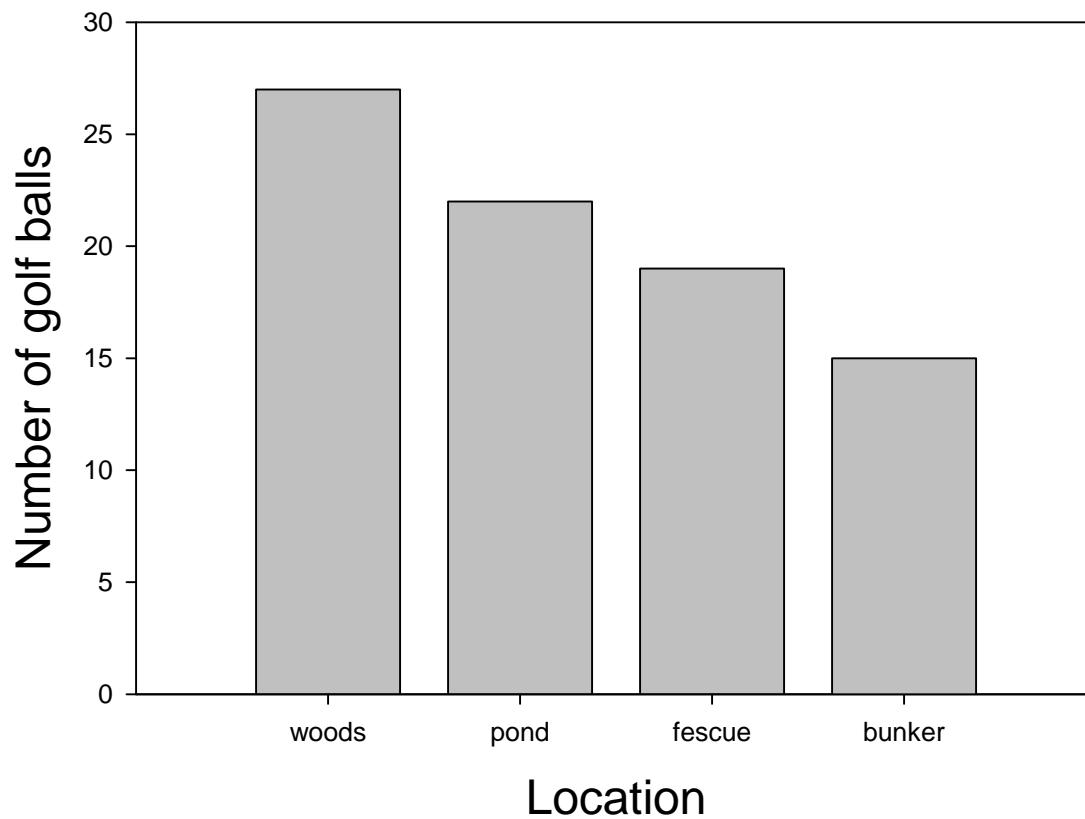
Frequency Distributions

A logical first step when collecting large amounts of data is to summarize it in a simple way on a routine basis. This is best done while collecting the data (i.e., continuously), rather than waiting until all of the data are collected to look at the patterns. Often, the patterns that begin to emerge from early data collections may enable adjustments to be made in your sampling approach that couldn't be done if you wait until data collection is completed before summarizing

Most investigators will start out by entering data into a common spreadsheet software package (e.g., EXCEL). This allows for easy computation of frequency tables and distributions. A frequency table is just a list of all of the values observed for a variable and how often each value was observed

Example:

<u>Location</u>	<u>Number of golf balls recovered</u>
Woods	27
Pond	22
Fescue	19
Bunker	15



*** Note that this example uses nominal data

The y-axis scale should begin at zero and the bars should be equal width, this ensures that the frequencies are expressed clearly

Bar graphs are straightforward to construct for nominal, ordinal, and discrete ratio-scale data (see examples 1.1-1.3 in Zar)

When ratio-scale data is distributed continuously, however, individual observations must be grouped before they can be tabulated (this is because continuous data can take on an infinite number of values)

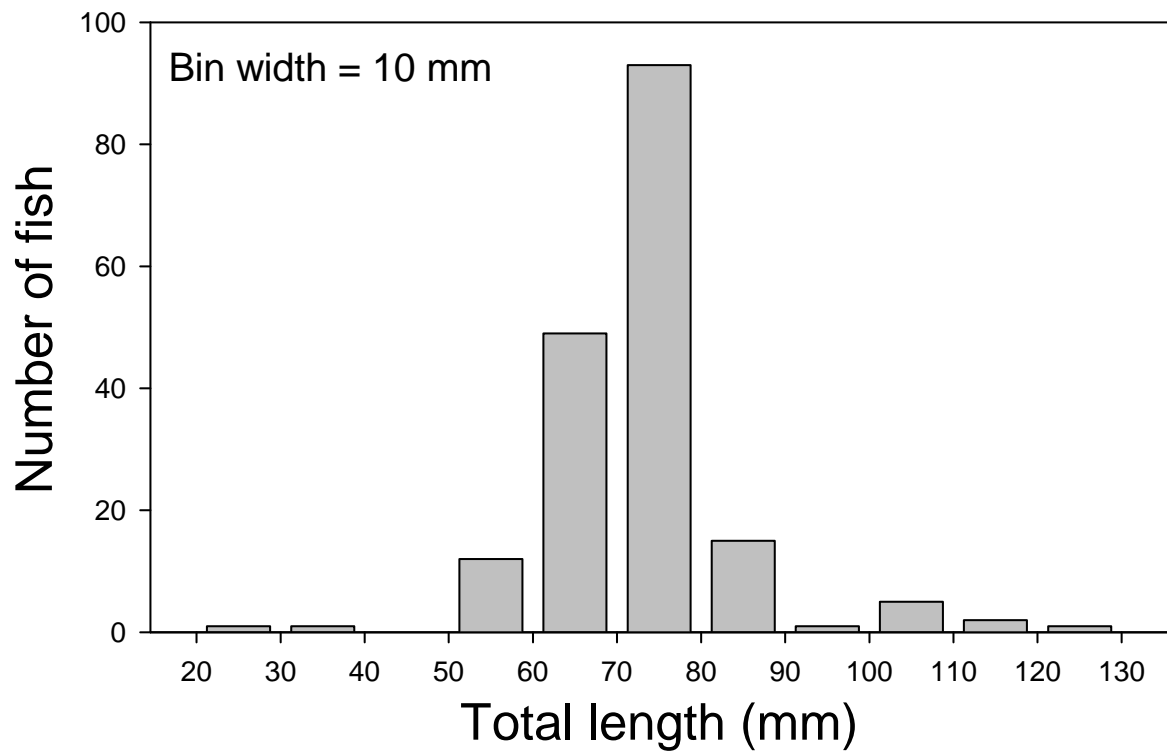
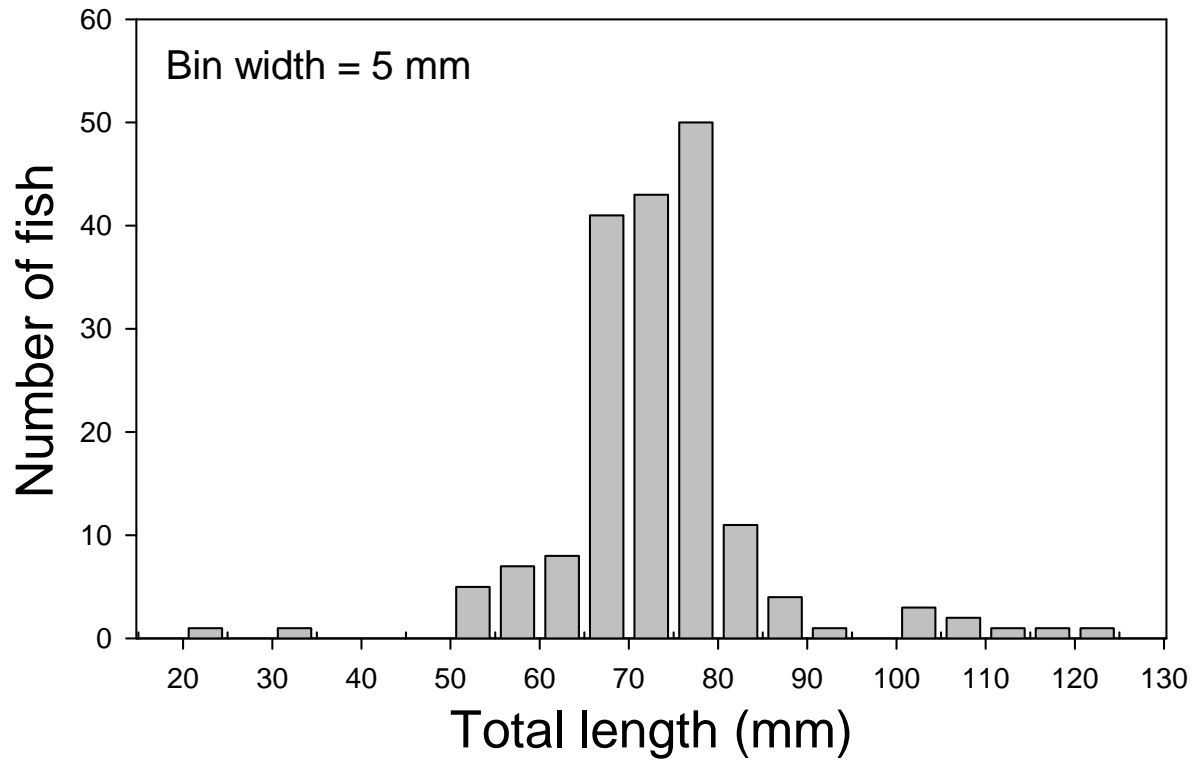
Sometimes discrete data is also grouped to ease the procedures of tabulation and graphing (see examples 1.4a and b in Zar).

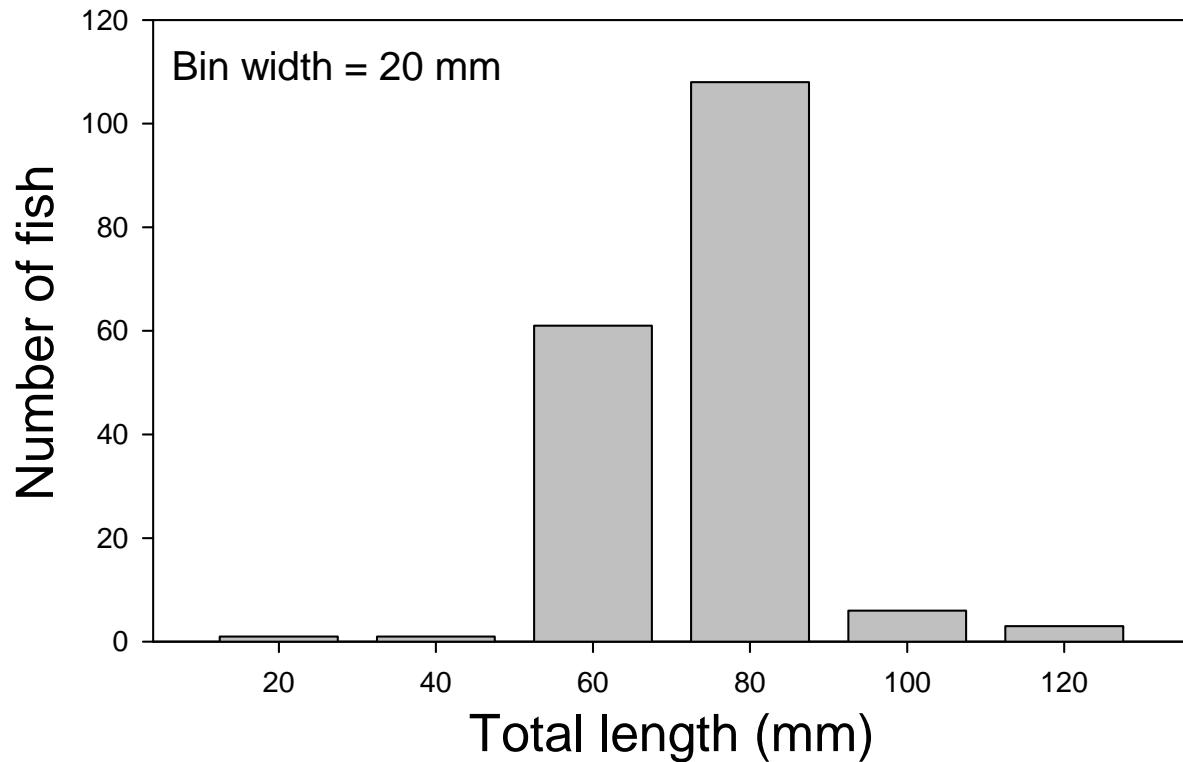
But keep in mind that grouping always results in a loss of information in the graph

Example:

Total lengths (mm) of Atlantic silversides collected in the Hudson River (n = 180;
range = 23 – 125mm)

23	65	69	70	73	75	76	79	82
32	65	69	70	74	75	76	79	82
51	66	69	70	74	75	77	79	82
55	66	69	71	74	75	77	79	83
55	66	69	71	74	75	77	79	84
55	66	69	71	74	75	77	79	85
55	66	69	72	74	76	77	79	85
57	66	69	72	74	76	77	80	86
58	67	69	72	74	76	77	80	89
60	67	70	72	74	76	77	80	90
60	67	70	72	74	76	77	80	90
60	67	70	72	74	76	77	80	92
60	67	70	72	75	76	78	80	101
60	67	70	73	75	76	78	80	105
62	67	70	73	75	76	78	80	105
63	68	70	73	75	76	78	80	107
65	68	70	73	75	76	78	81	109
65	68	70	73	75	76	78	81	115
65	68	70	73	75	76	79	82	118
65	68	70	73	75	76	79	82	125





Good rule of thumb: **Bin width** = $2 * \text{IQR} / n^{1/3}$ from Freedman and Diaconis (1981) on the histogram as a density estimator

*** IQR = Interquartile range = 75th quartile – 25th quartile

Quartile is a statistical function in EXCEL (Quartile (array, quartile))

For this example:

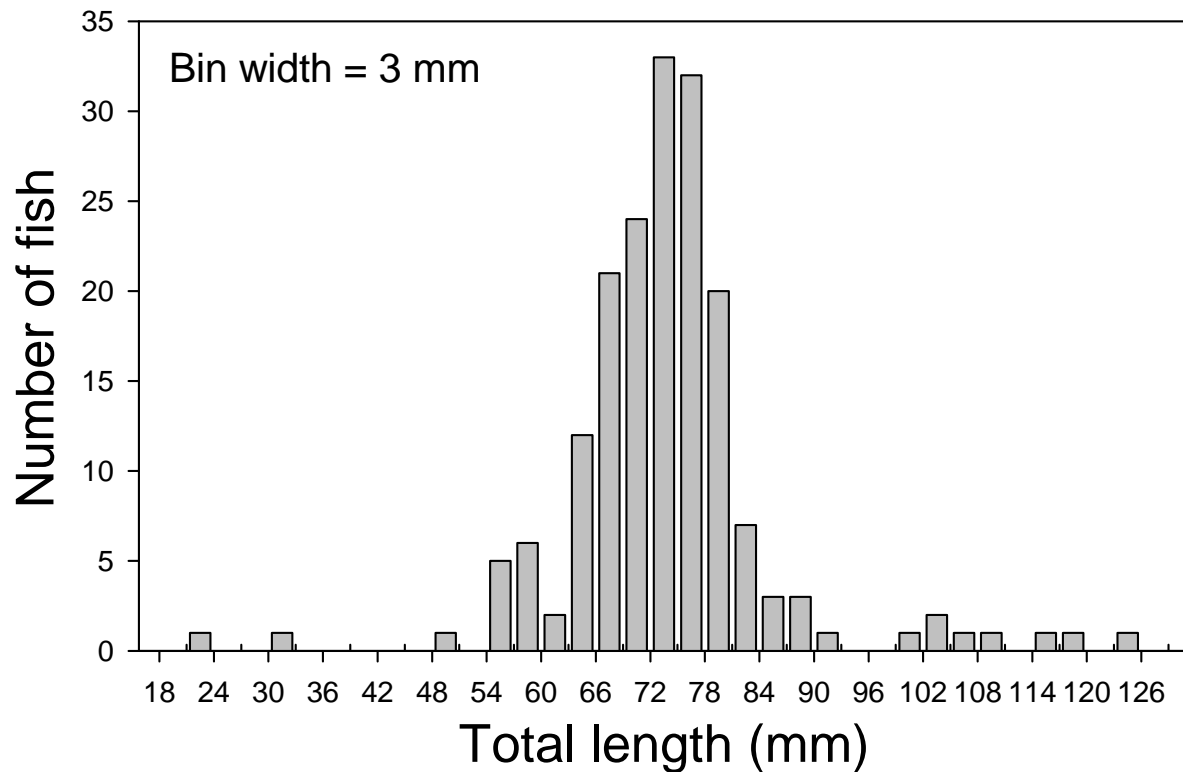
75th quartile = 78

25th quartile = 69

N = 180

So, $2 * \text{IQR} / n^{1/3} = 2 * 9 / 5.65 = 3.19$

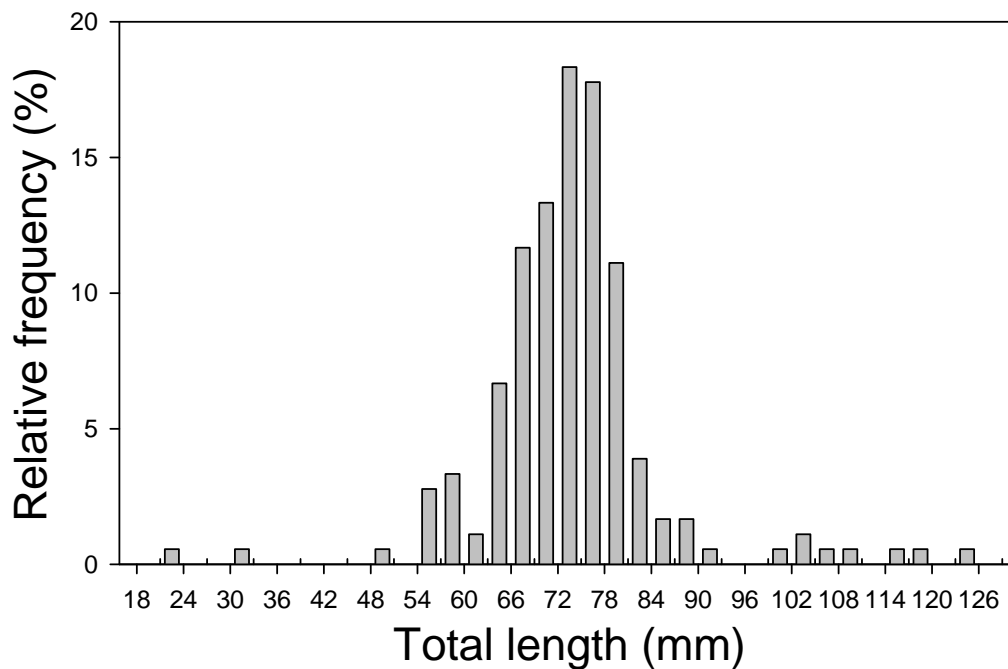
Let's see what a histogram with a bin width of 3mm looks like



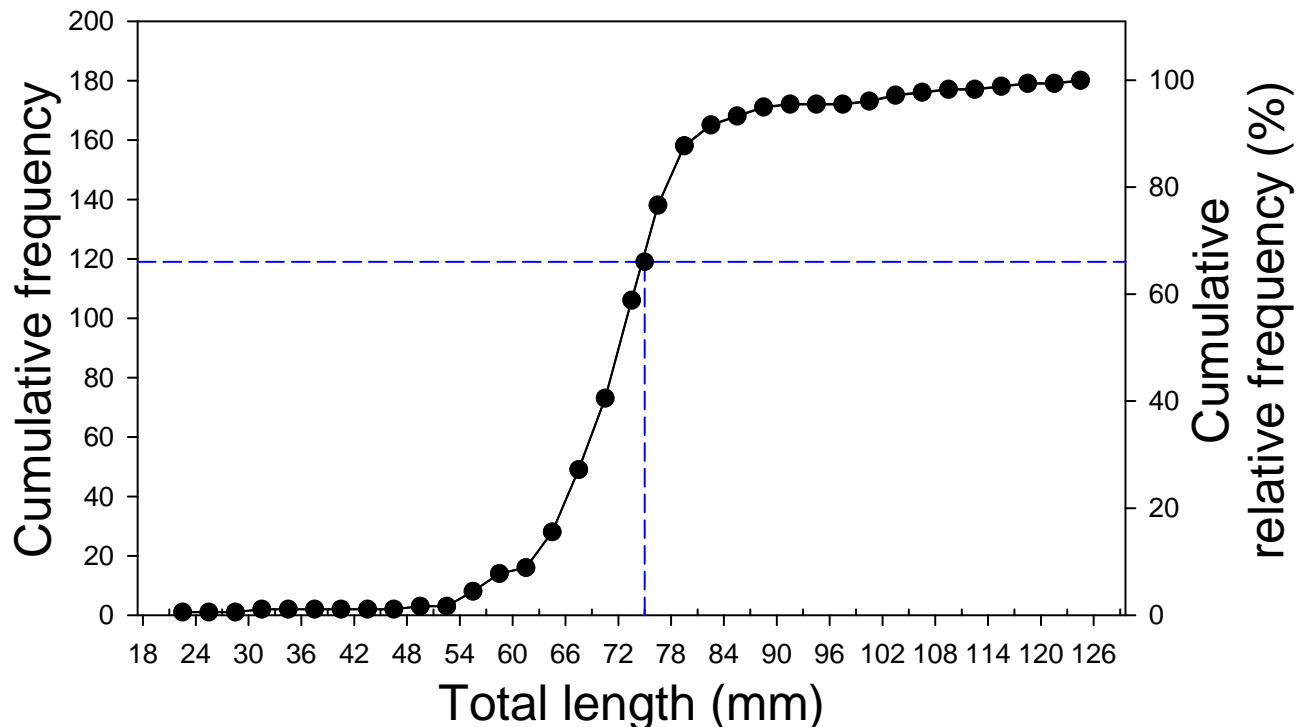
We can see more detail in the density distribution (number of fish per bin) with the smaller bin width. This histogram does a pretty good job of illustrating the underlying density distribution of silverside total lengths.

Obviously, if the bin width becomes as small as units of the last significant digit in our measurement scale, the histogram simply becomes the underlying density distribution

Often, you will see histograms plotted in terms of relative frequency (%) as opposed to frequency (n). This doesn't change the appearance of the histogram, but enables comparison with other data sets because the numbers of observations are scaled to 100%



Often, we are interested in knowing how many observations occur above or below some value (e.g., how many fish were larger or smaller than 75mm?). We can construct cumulative frequency (or relative frequency) distributions to evaluate these questions quickly






Measures of central tendency and variation

Now that you have begun to examine the general structure and distribution of your data by plotting it as a histogram or some other graphical display (stem and leaf plot, box plot, etc.), you need a way to describe the tendencies and variability present

We do this by estimating parameters for our population of interest by sampling

Some population parameters and their corresponding sample statistics:

Population mean = μ		Sample mean = \bar{x}
Population variance = σ^2		Sample variance = s^2
Population St. Dev. = σ		Sample St. Dev. = s

The most common measure used to make inferences about sample data is a measure of **central tendency** (location of the peak)

Different measures of central tendency

Mode = the most frequent observation

Median = the middle observation when the data is ranked (50% of observations above and below the median)

Mean = the sum of all observations divided by the sample size (n)

***The mean is the most commonly calculated measure of central tendency

Before defining the mean, some statistical symbols:

X = each observation is usually referred to as a variate X

\sum = Greek capital letter sigma denotes "the sum of"

$\sum_{i=1}^n X_i$ = "the sum of the X_i 's from $i = 1$ to n "

The arithmetic mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

In words, the mean is equal to the sum of the variates divided by the sample size

But, how does our sample estimate of \bar{x} relate to μ ?

\bar{X} will be an unbiased estimator of μ if:

1. observations (X_i 's) are random
2. observations (X_i 's) are independent
3. observations (X_i 's) are drawn from a larger population which can be described by a normal random variable

The **Law of Large Numbers** establishes that \bar{X} will approach μ as the sample size (n) gets large

Other measures of central location:

The Geometric Mean

The GM is calculated as:

$$e^{\frac{\sum_{i=1}^n \ln X_i}{n}}$$

The GM is used routinely for count data that fluctuate dramatically as it reduces the influence of large outliers on the mean

The Harmonic Mean

The HM is calculated as:

$$\frac{n}{\sum_{i=1}^n \frac{1}{X_i}}$$

The HM is very sensitive to small values and can be used to evaluate the potential effect on a group of low values that occur sporadically

Which measure of central location is best?

The arithmetic mean is widely used (and is assumed when someone uses the term 'mean' or 'average') because of the Central Limit Theorem

The **Central Limit Theorem** states that the averages of large (n), independent samples will follow a normal distribution regardless of the underlying population distribution

Stated differently: The distribution of sample means from a non-normal population will tend toward normality as n (the number of sample means drawn) increases

Importance: Enables us to use statistical tests that require our samples to be drawn from a normally distributed population, even when our data isn't normal, as long as n is large and our observations are independent (more on the significance of the CLM later.....)

The Geometric Mean and the Median (or other quantiles) are well suited to estimate central tendency when our data includes extreme observations that would have large leverage on the arithmetic mean

Weighted means can be used to calculate a 'grand mean' from several sample means of different n

$$\frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

Measures of spread

There are several measures that provide an indication of the spread of observations about the center of the distribution

The sample range

= the difference between the highest and lowest observations in a data set

Provides information on the boundaries of the sample data (but is a relatively crude measure of dispersion, and is a biased estimate of the population range)

Interquartile range IQR

= 75th percentile – 25th percentile

This measure indicates the boundaries of the majority of the sample data and is less sensitive to outliers

The IQR is the default box edge when constructing a box plot

Other percentiles (e.g., 90th-10th, 95th-5th) can also be used

The Mean deviation

$$\frac{\sum |X_i - \bar{X}|}{n}$$

aka, the Mean absolute deviation is a measure of the difference between each observation and the mean expressed in the same units as the data

The Variance

Introducing...the Sum of Squares

$$SS = \sum (X_i - \bar{X})^2$$

The SS is the preferred measure used to represent the differences between observations and the mean (like absolute values, squaring also removes the negative signs, but we really use it for reasons related to bias and additivity that we'll talk about later)

The mean SS is the **Variance** (or Mean Square) and is signified using σ^2 for a population and s^2 for a sample

$$\frac{\sum (X_i - \bar{X})^2}{n-1}$$

We also have what is referred to as a working or machine formula that simplifies the computation of the variance

$$\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$

* Note that we don't divide the Sum of squares by n, but rather by n-1

The quantity (n-1) represents the **Degrees of Freedom (df)**



So what exactly do we mean by Degrees of freedom?

The true definition actually stems from multi-dimensional geometry and sampling theory and is related to the restriction of random vectors to lie in linear subspaces.....

For our purposes, the definition used by Gotelli and Ellison (2004) will suffice: the number of independent pieces of information (i.e., n) in a data set that can be used to estimate statistical parameters

Essentially, we've already used up 1 degree of freedom to estimate the mean (\bar{X})

Box 2.1 on p. 20 of Quinn and Keough (2002) explains degrees of freedom as the number of observations in our data set that are "free to vary" when estimating the variance

Example from Quinn and Keough (2002): Suppose you have a data set with three observations (3, 4, and 5) and you know the sample mean = 4 and want to estimate the variance. Knowing the mean and one of the observations doesn't tell what the values of the other two observations must be, but if you know the mean and two of the observations, the third is fixed. So, once you know the mean, only two ($n-1$) of the observations are "free to vary".

Dividing our Sum of squares by $n-1$ generates an unbiased estimate of the variance

* Note that the variance is expressed in square units (relative to the mean)

We now introduce another statistic to express the spread in our data using the same units as the mean

The Standard deviation

$$\sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

The standard deviation (s) is often signified using an SD or sd, and is sometimes referred to as the *root mean square*

This is the most commonly reported measure of dispersion in the biological sciences

However, in reading the biological literature, you will often see another measure of dispersion reported, namely the standard error (SE or se). These are not the same quantities and we will deal with the standard error a bit later when we get to construction of confidence intervals and hypothesis testing

A way to compare measures of spread

Since your measure of spread is linked to the magnitude of the mean, how can you compare measures of spread when the means differ appreciably?

The **coefficient of variation (CV)** is calculated as $SD/\bar{X} \times 100$ (to convert it to a percentage)

This statistic enables comparison of variation on a relative scale

Skewness, Kurtosis, and Central Moments

The variance (and SD) are examples of central moments

A central moment in statistics is:

$$(1/n) \sum (X_i - \bar{X})^r$$

The first central moment equals zero and the second central moment is simply the variance

$$(1/n) \sum (X_i - \bar{X})^2$$

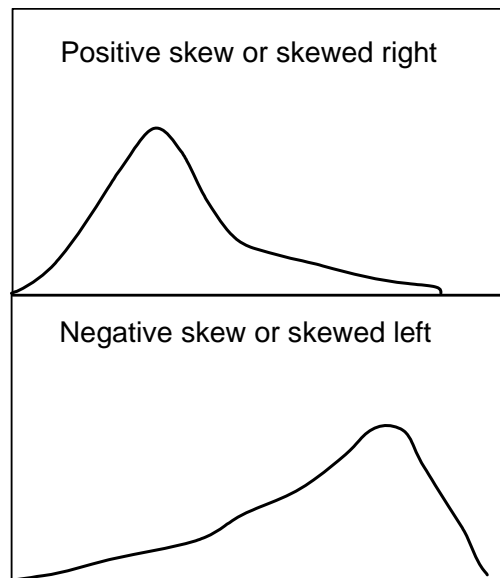
The third central moment divided by the cube of s is **g1** and is known as the **skewness**

$$(1/ns^3) \sum (X_i - \bar{X})^3$$

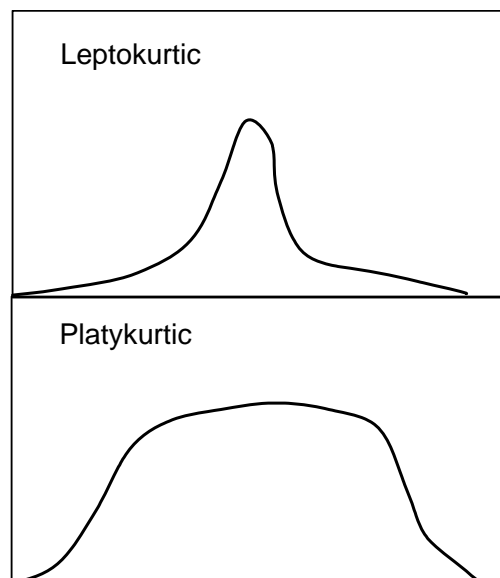
The fourth central moment divided by the 4th power of s and then minus 3 is **g2** and is known as the **kurtosis**

$$\left[(1/ns^4) \sum (X_i - \bar{X})^4 \right] - 3$$

Skewness measures asymmetry in the distribution; whether long tails exist on the right (positive skew; $g_1 > 0$) or left (negative skew; $g_1 < 0$) side



Kurtosis measures the proportion of the distribution in the center and tails relative to the shoulders. Leptokurtic ($g_2 > 0$) = more observations in the center and tails; Platykurtic ($g_2 < 0$) = more observations in the shoulders



Skewness and Kurtosis not used as much in modern literature as they are very sensitive to outliers and the magnitude of the mean

An Introduction to Probability

(Note: the following ideas are generously borrowed from Chapter 1 in Gotelli and Ellison (2004), who do a nice job of placing probabilistic ideas in a biological/ecological context)

Goal: To develop a conceptual understanding of basic probability calculations which are the backbone of the 'probabilistic framework' upon which all statistical analyses rest

We generally have an intuitive feel for what is meant by the term probability. If we make a statement that there is a 40% chance that a hurricane is going to make landfall at a specific location, we have a pretty good idea what that means because we understand that there is a level of uncertainty due to *natural variation*

How do we measure probability exactly?

Rather than use common examples of a coin flip or the toss of a die, we'll use a biological example (Gotelli and Ellison use pitcher plant ecology; I'll incorporate some ideas from fish predator-prey interactions instead)

Imagine a small cove in a large estuarine system. Young-of-the-year bluefish routinely enter this cove in search of prey, which are other fishes that inhabit the cove. As an individual bluefish swims through the cove, it either encounters a prey fish or it doesn't (it's a discrete outcome). Once a prey fish is encountered, an attack may or may not follow (also a discrete outcome); and if an attack occurs, it may or may not be successful (another discrete outcome)

Therefore, we are interested in estimating the probability that a single bluefish encounters, attacks, and captures a prey fish during a search of the cove.

The search is called an **event** (it has a beginning and an end), an encounter (or not), an attack (or not), and a capture (or not) are all considered **outcomes**, with the set of all possible outcomes = the **sample space**

***The sample space should be defined carefully because it limits our scope of inference

Events are often referred to as trials, with a set of trials making up an experiment (either controlled or natural, more on this later.....)

$$P = \text{number of outcomes/number of trials}$$

and

$$0 \leq P \leq 1$$

You can't have more outcomes than trials

Next, we would sample to generate data on the fraction of bluefish that encounter, attack, and capture prey. We might make observations from a platform above the water from which we could distinguish a bluefish and its behavior, or we might fix a camera or some other device to an individual bluefish and use video technology to generate our data (this technology is not quite there yet for a small fish). In any case, we use an effective sample design (more on this topic soon.....) and collect our data.

In our example, say we observed 100 bluefish search the cove and found that 72 of them encountered a prey fish (determined by some behavioral reaction by the bluefish to a nearby prey fish), and that 44 of those encounters elicited an attack, resulting in 11 captures

So, we now have

$$P(\text{encounter}) = 72/100 = 0.72$$

$$P(\text{attack}) = 44/100 = 0.44$$

$$P(\text{capture}) = 11/100 = 0.11$$

We'll return to this example a bit later when we discuss conditional probabilities

For now, let's just focus on the number of captures per search. Suppose that an individual bluefish searches our cove twice each day (once at dawn and once again at dusk), and that each search lasts for about $\frac{1}{2}$ hour. We know, because of the time it takes to manipulate and swallow a prey fish, that the maximum number of prey fish that a single bluefish can eat per search is 2. We now have 3 possible outcomes for each search (0,1, or 2) and two searches per day, so all total we have nine possible outcomes $\{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)\}$. It is important to note that these outcomes are said to be **mutually exclusive**.

This means that the sum of all of the probabilities of the outcomes will be equal to 1.0 (**The First Axiom of Probability**). If the outcomes are not mutually exclusive, this will not be true.

In our example, each outcome has a probability of $1/9$ and they sum to 1.0.

Complex and Shared Events

A **complex event** is one that can occur by multiple different pathways

A **shared event** is one that requires the simultaneous occurrence of two or more simple events

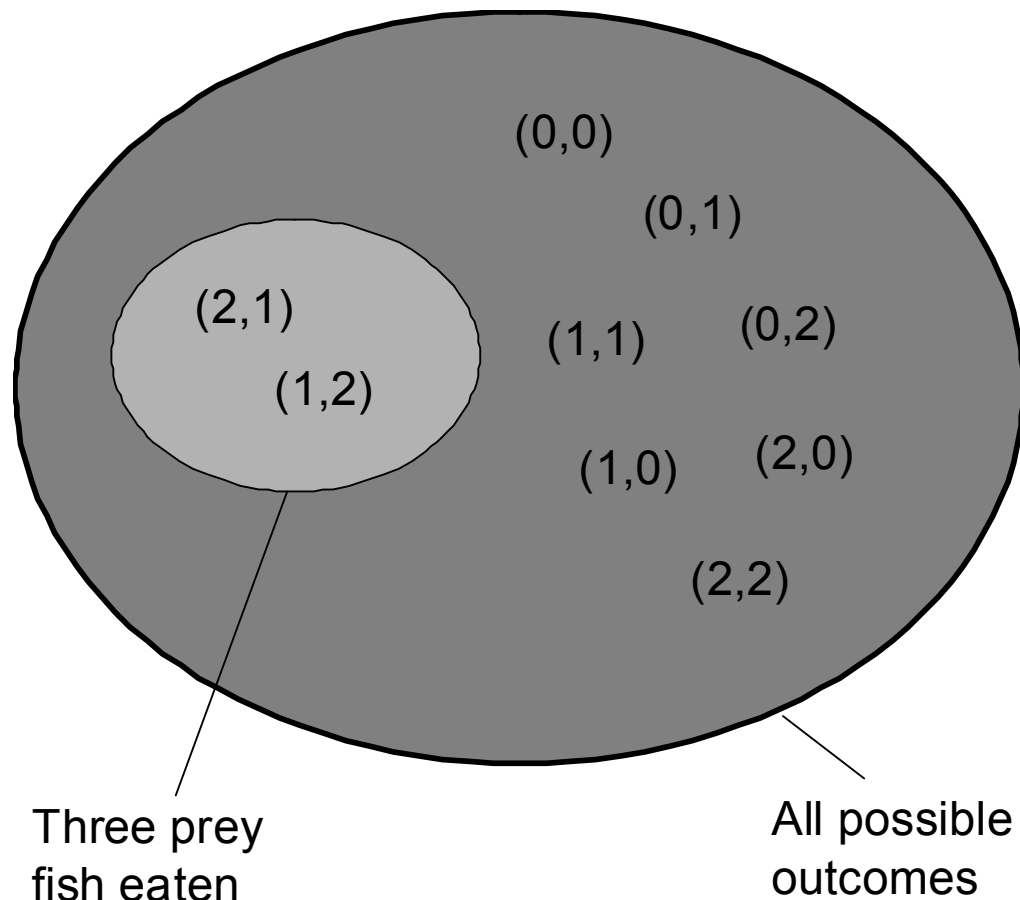
Complex events are represented using an 'or' statement (e.g., event A or event B or event C) and equal the **union** of simple events. We simply **sum** the probabilities of simple events

Shared events are represented using an 'and' statement (e.g., event A and event B and event C) and equal the **intersection** of simple events. In this case, we **multiply** the probabilities of simple events

Complex event example:

What is the probability that a bluefish captures three prey fish over the course of a single day? This event can occur in two ways:

$$\text{Two prey fish} = \{(1,2), (2,1)\}$$



Since two of the possible nine outcomes yield three prey fish eaten, we would estimate this probability as $1/9 + 1/9 = 2/9$

This is the **Second Axiom of Probability**: The probability of a complex event equals the sum of the probabilities of the outcomes that make up that event

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

Shared event example: Suppose instead that we were interested in estimating the probability of a bluefish catching 1 prey fish during each of its two search events. The probability of each of these events is $1/3$, so the probability of obtaining both is $1/3 \times 1/3 = 1/9$.

$$P(A \cap B) = P(A) \times P(B) \text{ (if A and B are **independent**)}$$

Now, an example from Gotelli and Ellison (2004) pp. 16-17 that illustrates probability calculations for both complex and shared events:

Suppose you are sampling a set of rock outcroppings in which there exist populations of the milkweed plant and populations of caterpillars that eat the milkweed. Some of the milkweed populations have developed chemical resistance (R) to predation by caterpillars and others have not. After sampling, you note that 20% of milkweed populations are resistant, $P(R) = 0.20$, which means that $P(\text{not } R) = 1 - P(R) = 0.80$. Sampling also reveals that 70% of the outcroppings contain caterpillars (C), so $P(C) = 0.70$, and $P(\text{not } C) = 1 - P(C) = 0.30$.

Now we define some ecological rules based on our knowledge of the movements and interactions of the two species. The first rule is that all milkweeds and caterpillars can disperse and reach all of the outcroppings. Second, all milkweeds can persist when caterpillars are

absent, but only resistant milkweeds can persist when caterpillars are present. Third, the initial colonization of outcroppings are independent events.

What are the different combinations of outcomes that can occur? We have four possible outcomes that are each the result of shared events (in this case, two events occurring simultaneously).

Shared event	Probability calculation	Milkweed present?	Caterpillar present?
NR milkweed and no caterpillar	$[1-P(R)] \times [1-P(C)] = (1.0-0.2) \times (1.0-0.7) = 0.24$	Yes	No
NR milkweed with caterpillar	$[1-P(R)] \times [P(C)] = (1.0-0.2) \times (0.7) = 0.56$	No	Yes
R milkweed and no caterpillar	$[P(R)] \times [1-P(C)] = (0.2) \times (1.0-0.7) = 0.06$	Yes	No
R milkweed with caterpillar	$[P(R)] \times [P(C)] = (0.2) \times (0.7) = 0.14$	Yes	Yes

***Note that the probabilities add to 1.0

We can now add some of these probabilities to obtain the probabilities of complex events. For instance, what is the probability that a milkweed population will be resistant? We simply add the two probabilities for the shared events that contain resistant milkweed (0.06 without caterpillars and 0.14 with caterpillars) to obtain 0.20, which matches the original probability of resistance at the outset.

We also know that non-resistant milkweed will not persist in the presence of caterpillars. This shared event is represented by a probability of 0.56. The compliment of this event $(1 - 0.56) = 0.44$ is

the probability that an outcropping will contain milkweed, a probability which can also be obtained by adding the probabilities from the other three shared events ($0.24 + 0.06 + 0.14$). Therefore, although the probability of resistance is only 0.20, we expect to find milkweed in 44% of outcroppings because not all non-resistant milkweed will be occupied by caterpillars.

Rules for combining complex and shared events

Returning to our bluefish foraging example, suppose we wish to know the combined probability of a bluefish consuming zero prey fish during its first search, and two prey fish during its second search. We now have two events (searches) each with sets of possible outcomes. We've already seen that we can have a total of 9 possible outcomes in the entire set. For this particular question, we'll call the first set of outcomes 1^{st} and the second set 2^{nd} :

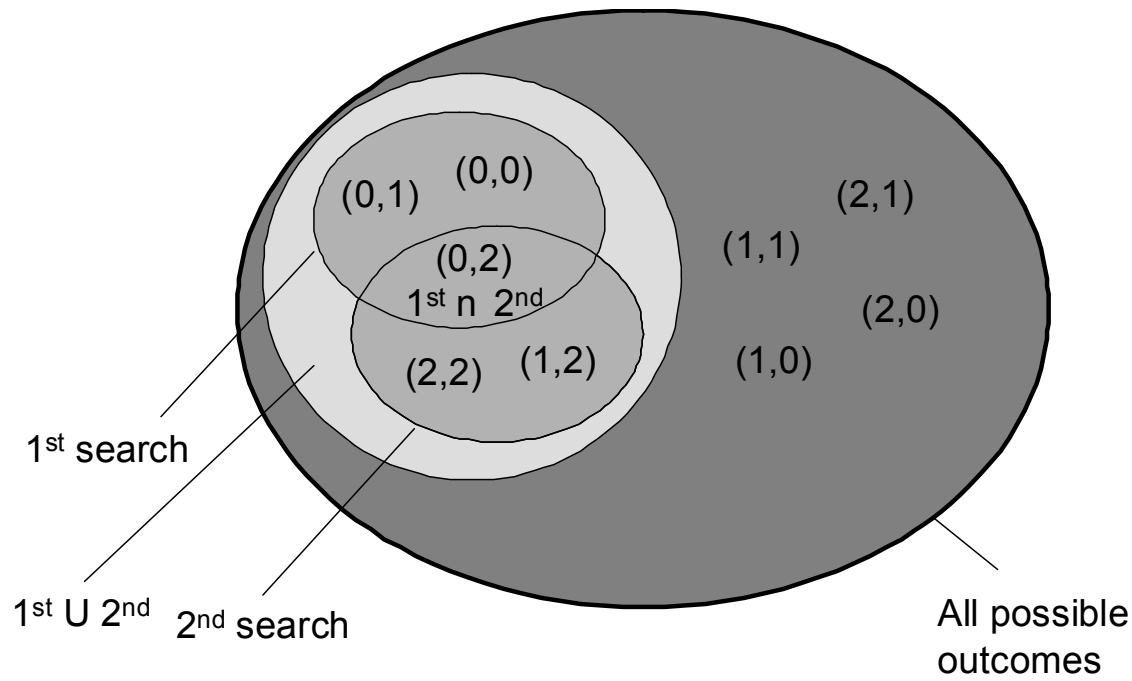
$$\begin{aligned} 1^{\text{st}} &= [(0,0), (0,1), (0,2)] \\ 2^{\text{nd}} &= [(0,2), (1,2), (2,2)] \end{aligned}$$

We can now construct two new sets of combined outcomes. The **union** of 1^{st} and 2^{nd} contains all of the outcomes that are in 1^{st} or 2^{nd} alone and is represented by $1^{\text{st}} \cup 2^{\text{nd}}$. It is basically the addition of 1^{st} and 2^{nd} sets

$$1^{\text{st}} \cup 2^{\text{nd}} = [(0,0), (0,1), (0,2), (1,2), (2,2)]$$

The second new set of combined outcomes is the **intersection** of 1^{st} and 2^{nd} sets and contains only those outcomes common to both 1^{st} and 2^{nd}

$$1^{\text{st}} \cap 2^{\text{nd}} = [(0,2)]$$



We can also create what are called complimentary sets. For instance, the complimentary set of 1^{st} would be denoted 1^{stc} and would include all outcomes not included in 1^{st}

$$1^{\text{stc}} = [(1,0), (1,1), (1,2), (2,0), (2,1), (2,2)]$$

Lastly, we need to have an empty set, which contains no outcomes and is written as $\{\emptyset\}$. The intersection of 1^{st} and 1^{stc} would be the empty set.

Returning to our question about bluefish foraging, the union of 1^{st} and 2^{nd} only contains 5 outcomes yielding a probability of $5/9$ that either the 1^{st} search would have 0 prey eaten or the 2^{nd} search would have 2 prey eaten. This seems to violate the 2^{nd} axiom of probability which states that the probability of a complex event equals the sum of the probabilities of the outcomes that make up the event. But, this axiom only holds true if the events 1^{st} and 2^{nd} are mutually exclusive. In this case, they are not, they have the outcome (0,2) in common.

Therefore, the union of two non-mutually exclusive events is their sum minus their intersection:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

So, in our case, $P(1^{\text{st}} \cup 2^{\text{nd}}) = 3/9 + 3/9 - 1/9 = 5/9$. In addition, the probability that a bluefish eats 0 prey during its first search and 2 prey during its second search is the intersection of the two events = 1/9

Now, what if we know the outcome of the first search and wish to estimate the probability of the second search? The probability estimate for the second search is what we call a **conditional probability** and is written as:

$$P(B|A) \text{ or in our case, } P(2^{\text{nd}}|1^{\text{st}})$$

The **conditional probability** is calculated as:

$$P(B|A) = P(A \cap B)/P(A)$$

If outcome A has already occurred, then the outcomes for B need to be restricted to the set of outcomes in common with A, thus the intersection in the numerator. The denominator is the restricted sample space of possible outcomes for A, which has already occurred.

In our example the probability that a bluefish eats 2 prey fish during the second search after having already eaten 0 prey fish during the first search $P(2^{\text{nd}}|1^{\text{st}}) = 1/9 \div 1/3 = 1/3$. Note that the probability is higher than 1/9, which we calculated for the probability of both events occurring with no prior knowledge. *Having prior knowledge narrows the possibilities for subsequent events.*

As another example, consider the following questions based on our bluefish observations: What is the probability of a bluefish attack given that an encounter has already occurred? What is the probability of a successful capture for each attack? These are both questions that require us to estimate conditional probabilities.

Remember,

$$P(\text{encounter}) = 0.72$$

$$P(\text{attack}) = 0.44$$

$$P(\text{capture}) = 0.11$$

If an encounter has occurred, we can calculate the probability of an attack as:

$$\begin{aligned} P(\text{attack}|\text{encounter}) &= P(\text{encounter} \cap \text{attack})/P(\text{encounter}) \\ &= (0.44)/(0.72) = 0.611 \end{aligned}$$

$P(\text{encounter} \cap \text{attack}) = 0.44$ since all fish that were attacked had to be encountered first, but not all encountered fish are attacked.

$$\begin{aligned} P(\text{capture}|\text{attack}) &= P(\text{attack} \cap \text{capture})/P(\text{attack}) \\ &= (0.11)/(0.44) = 0.25 \end{aligned}$$

Again, $P(\text{attack} \cap \text{capture}) = 0.11$ since all fish that were captured had to be attacked first, but not all attacked fish are captured.

***Conditional probabilities are the foundation of Bayesian statistics, the framework for which we will describe a bit later relative to other statistical and model selection approaches

Probability distributions

All random variables will have an associated probability distribution with a range of values of the variable on the x-axis and the relative probabilities of each value on the y-axis

Most of the statistical procedures that you will use in the study of biology make some assumptions about the probability distribution of the variable you have measured (or about the distribution of the statistical errors). We also use probability distributions to generate models and make predictions, so they are very important to what we do.

Many (too many) probability distributions have been defined mathematically and there are several that work well in describing biological phenomena. We will focus on a few of the major ones.

Recall that a variable can be either discrete or continuous in its distribution, which creates some important differences in the probability distributions:

1. For discrete variables, the probability distribution will include measurable probabilities for each possible outcome
2. For continuous variables, there are an infinite number of possible outcomes. Thus, the probability distribution is what we call a **probability density function** (pdf), and it is used to estimate the probability associated with a range of values since the probability of any single value = 0.

Discrete probability distributions

Bernoulli random variables represent the simplest type of discrete variables because each event or trial can only have two outcomes

A collection of n independent Bernoulli trials results in a **Binomial random variable** (i.e., we perform many replicate Bernoulli trials)

A Binomial random variable, X , is defined by the number of successful results in n independent Bernoulli trials

$$X \sim \text{Bin}(n,p)$$

where n = number of independent Bernoulli trials and p = the probability of a successful outcome in any single trial

The Binomial probability distribution is calculated as:

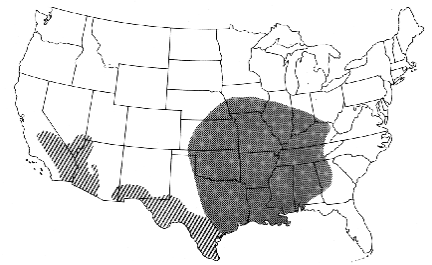
$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$$

Where n = number of trials and X = the number of successes, and $n!$ = n factorial = $n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$

p^X = probability of obtaining X independent successes each with probability p

$(1-p)^{n-X}$ = probability of obtaining $n-X$ failures each with probability $1-p$

$n!/X!(n-X)!$ = the binomial coefficient, which calculates the number of possible ways to get X successes, minus any double counting



Example: Suppose you are interested in estimating the probability that brown recluse spiders inhabit the UNCW campus. Based on previous research in the region, you know that the probability of a brown recluse being present ($X = 1$) in any single site in this region of the country is 0.04, so $P(X = 1) = p = 0.04$. So, if we search each of the buildings on campus there is only a 4% chance of a brown recluse spider being present in any one building. So, you set out and search each of the campus buildings and find that brown recluse spiders are present in 8 of the buildings on campus among a total of 64 buildings. You want to know the probability of this outcome given the probability of finding a spider in any one building is only 4%.

Based on the binomial distribution, our probability would be estimated as:

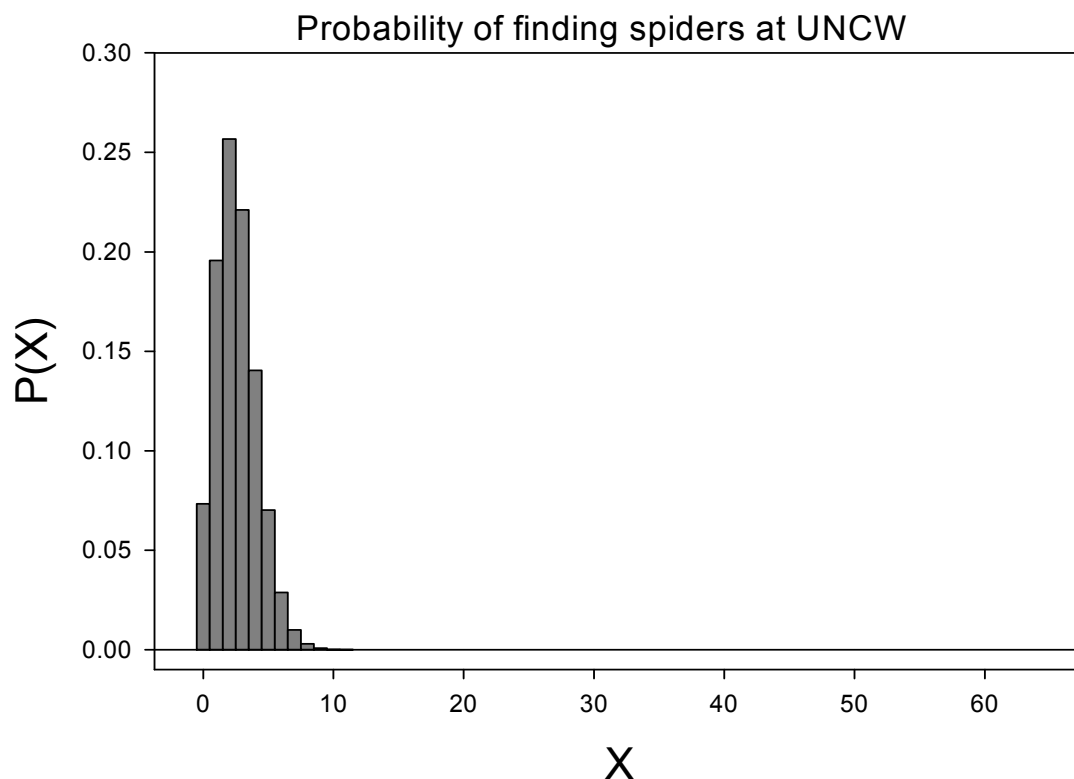
$$P(8) = \frac{64!}{8!(64-8)!} \cdot .04^8 (1-.04)^{64-8} = 0.003$$

Thus, there is only a 0.3% probability that we would find brown recluse spiders in 8 campus buildings given that the probability of finding a spider in any single building was 4%. I would be inclined to conclude that there are an unusually high number of brown recluse spiders on the UNCW campus.

Using the equation for the binomial distribution, we can estimate the probabilities for any value of X (up to $n = 64$ in this case) and plot them as a frequency distribution

Spider presence/absence probabilities

X	P(X)	X	P(X)	X	P(X)
0	0.07334304	22	2.54498E-15	44	2.68384E-46
1	0.19558144	23	1.93639E-16	45	4.97008E-48
2	0.25670064	24	1.37834E-17	46	8.55358E-50
3	0.22104778	25	9.18891E-19	47	1.36493E-51
4	0.14045744	26	5.74307E-20	48	2.01422E-53
5	0.07022872	27	3.36785E-21	49	2.74044E-55
6	0.02877427	28	1.85432E-22	50	3.42555E-57
7	0.00993397	29	9.59132E-24	51	3.91811E-59
8	0.00294915	30	4.66245E-25	52	4.08137E-61
9	0.00076459	31	2.13069E-26	53	3.85035E-63
10	0.00017522	32	9.1553E-28	54	3.26804E-65
11	3.584E-05	33	3.69911E-29	55	2.47579E-67
12	6.5956E-06	34	1.4053E-30	56	1.65789E-69
13	1.0993E-06	35	5.01893E-32	57	9.69529E-72
14	1.6685E-07	36	1.68459E-33	58	4.87551E-74
15	2.3174E-08	37	5.31178E-35	59	2.06589E-76
16	2.9571E-09	38	1.57257E-36	60	7.17324E-79
17	3.479E-10	39	4.36824E-38	61	1.9599E-81
18	3.785E-11	40	1.13756E-39	62	3.95141E-84
19	3.8182E-12	41	2.77454E-41	63	5.22674E-87
20	3.5795E-13	42	6.3308E-43	64	3.40282E-90
21	3.125E-14	43	1.34959E-44		



Poisson random variables represent another discrete random variable and are ideal for situations when p is very small and n is large. Therefore, we are talking about rare events in space or time. Often, biologists use the Poisson distribution to describe patterns resulting from counts or occurrences of plants or animals. This is because normally, within any single defined sample space or time interval, the most common count is 0.

A Poisson random variable is defined as the number of occurrences of an event in a fixed area or time interval. The probability of any value x is calculated as:

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

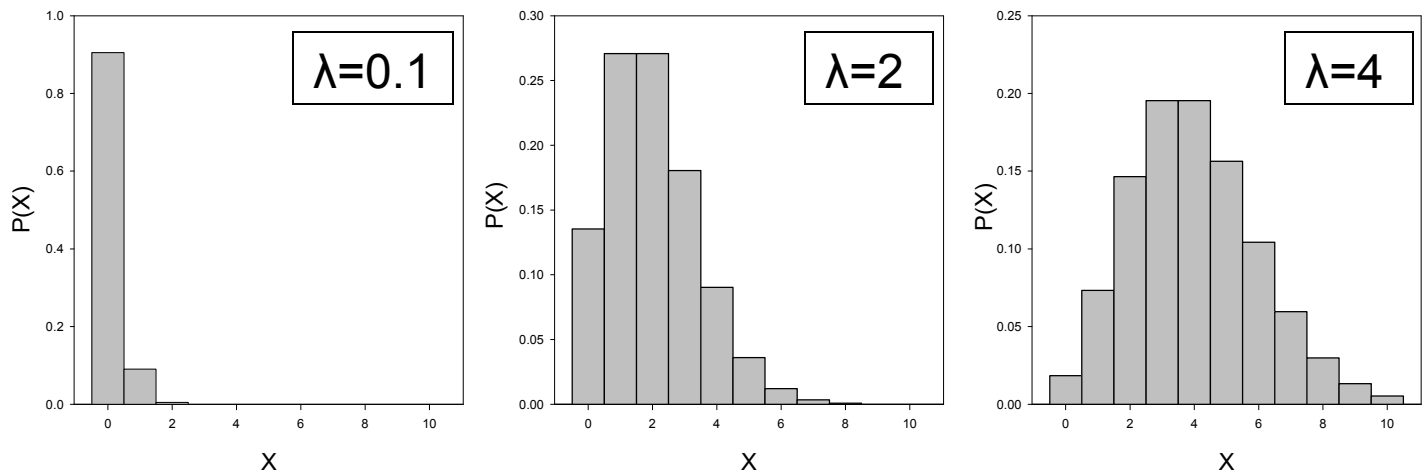
where λ = the average value of the number of occurrences of an event in each sample (space or time). The shape of the distribution depends only on λ , which differs from the binomial distribution that depended on both n and p .

Example: Suppose we surveyed multiple college campuses in the southeast US (*in this case the sample space is each campus, not a building*) and found that the average number of occurrences (λ) of brown recluse spiders was 2.56, then we can use the Poisson distribution to estimate the probability of having 8 occurrences on the UNCW campus as:

$$P(8) = \frac{2.56^8}{8!} e^{-2.56} = 0.0035$$

Recall, that our estimate using the binomial distribution = 0.00295

Changes in the shape of the Poisson distribution with changes in λ



Continuous probability distributions

As we mentioned earlier, continuous variables are not limited to take on integer values, but instead can take on an infinite number of values. Therefore, we can't estimate the probability of any single outcome and instead estimate the probability that an outcome will fall within a specific interval.

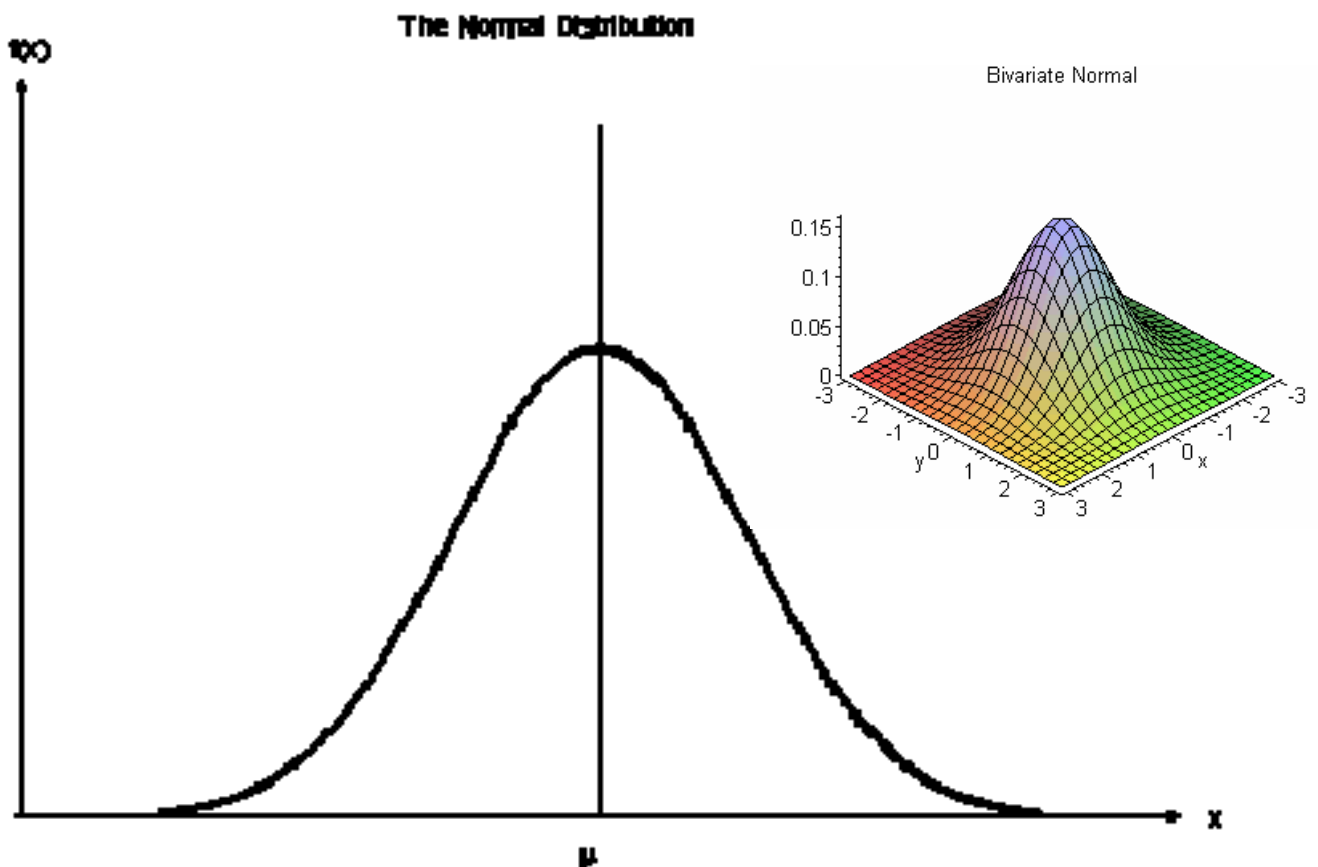
The probability distribution is now termed a **probability density function** (pdf), and we use it to estimate the probability of a variable falling within a certain range of values. Through integration, we can estimate the area under the curve (the curve is the pdf) for any range of values. Generally the pdf is normalized so that the area under the curve representing the total probability is approximately 1.

We can also generate **cumulative density functions** (cdf) to examine the probability of a variable being less than or greater than some value ($Y_i < Y$). These represent tail probabilities, which is where our familiar P-values come from.

The continuous distribution that fits the most patterns in nature is the **Normal (or Gaussian) distribution**, which has the familiar bell-shaped pattern. The normal distribution is symmetrical about the mean and is defined by the mean (μ) and the variance (σ^2). The probability density function for the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

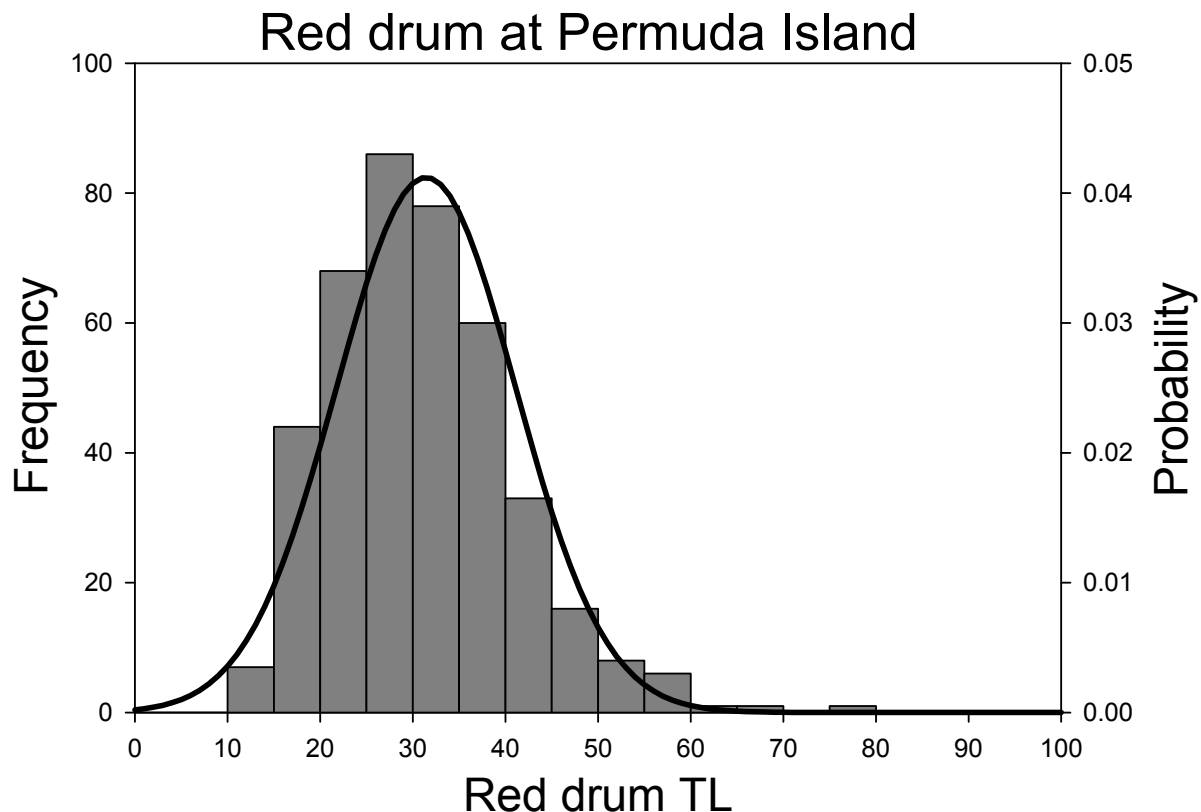
The general shape of a normally distributed variable



Many of the most common statistical models that are used in biology (e.g., linear models such as regression and ANOVA) have the assumption that the variables being analyzed (or their statistical errors = residuals about a fitted model) are normally distributed.

You can easily make a visual comparison between a normal distribution and your data just using your estimates of the mean and variance and the normal pdf

Example: Several years of sampling by the North Carolina Division of Marine Fisheries has produced a large amount of body size data for juvenile red drum collected near Permuda Island in Topsail Sound, NC. The sample size (n) = 409 with total lengths ranging from 13 – 76mm. The mean (\bar{x}) = 31.4 and the SD (s) = 9.68. Below is a histogram of the raw data along with a normal distribution (probabilities estimated using the mean, the SD, and equation for a normal pdf).



You can see that there are a few more fish between 15-30mm TL and a few less fish between 35-50mm TL than we would expect if TL were distributed exactly normally. Despite this, the data appear to follow an expected normal distribution fairly well. We will talk about methods to test for deviations from normality a bit later.

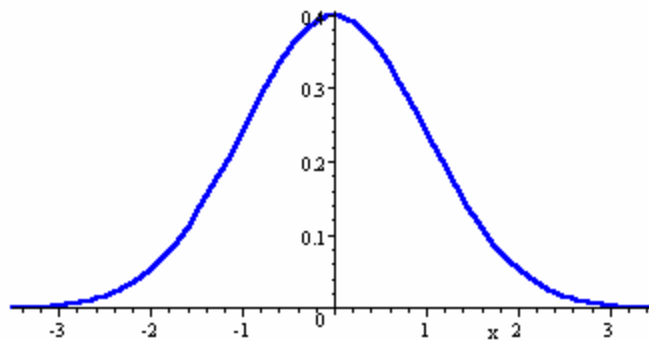
Properties of the Normal distribution

1. The normal distribution is symmetrical about the mean and its shape is determined only by the mean and variance
2. Normal distributions are additive (if A and B are normally distributed random variables, then their sum $(A + B = C)$ is also normally distributed)
3. Normal distributions can be easily transformed using **shift** (addition of a constant) and **scale** (multiplication by a constant) operations. Addition (shift) of a constant (a) to a normally distributed variable increases the mean by the value of the constant a , with no change to the variance. Multiplication (scale) of a normally distributed variable by a constant (a) multiplies the mean by a and the variance by a^2
4. An important combination of a scale ($X * 1/\sigma$) and shift ($X - \mu$) operation results in what we call a **standard normal random variable**

$$= \frac{X - \mu}{\sigma}$$

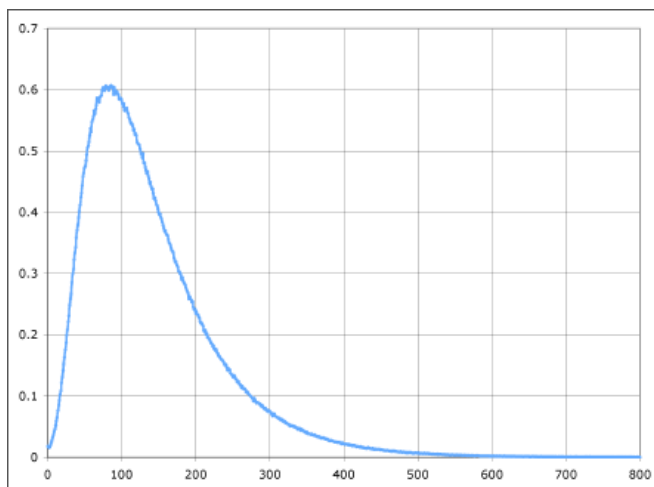
A standard normal distribution (called the **z distribution**) has a mean = 0 and a standard deviation = 1, and is expressed ($X \sim N(0,1)$). The conversion of any normal random variable to a standard normal random variable is what enables us to test hypotheses about the mean, which will be the first hypothesis tests that we perform

The standard normal distribution

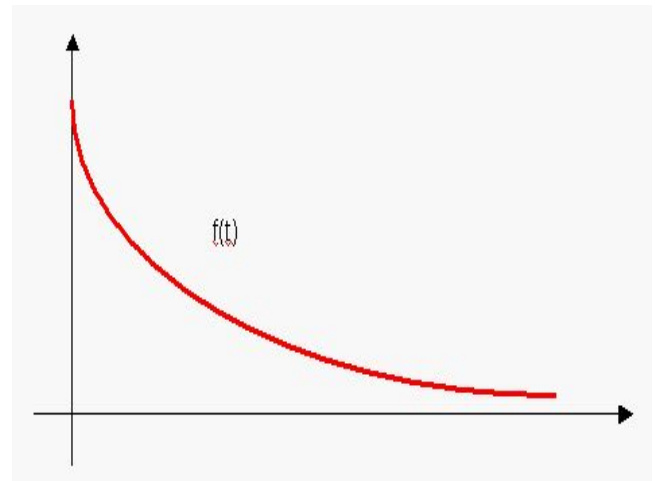


Continuous variables are not always distributed symmetrically. Many biological variables show right- or positive skewness, with long tails that include larger observations that occur with less frequency. The **lognormal distribution**, in which the log transformation of the variable is normally distributed, describes many biological data of this sort (i.e., measurement data that cannot be negative such as lengths and weights). Another asymmetric distribution observed, although less frequently in biology, for continuous random variables is the **Exponential distribution**

Examples of lognormal and exponential distributions



Lognormal distribution



Exponential distribution

We will encounter several other mathematical probability distributions throughout the course. There are several that are used to estimate the probabilities of sampling statistics and model parameters, as well as for hypothesis testing. We will briefly introduce a few of these here, and will devote more time to each later.

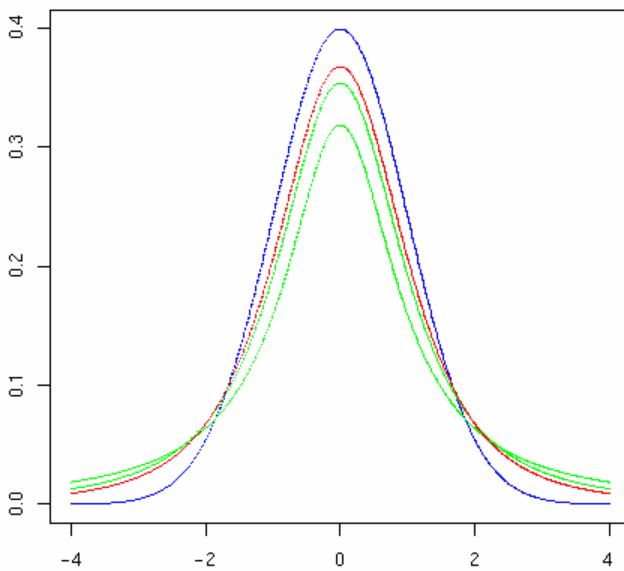
We have already mentioned the **z-distribution** that results when we standardize a normal random variable. It is used to test hypotheses concerning differences between sample statistics and population parameters when we know the standard deviation of the population parameter (which we never do).

The **t-distribution (or student's t-distribution)** is also used to test hypotheses concerning differences between sample statistics and population parameters. However, it accounts for the fact that we are estimating the standard deviation of the population parameter using our sample data (this is where the standard error becomes important as we will see soon).

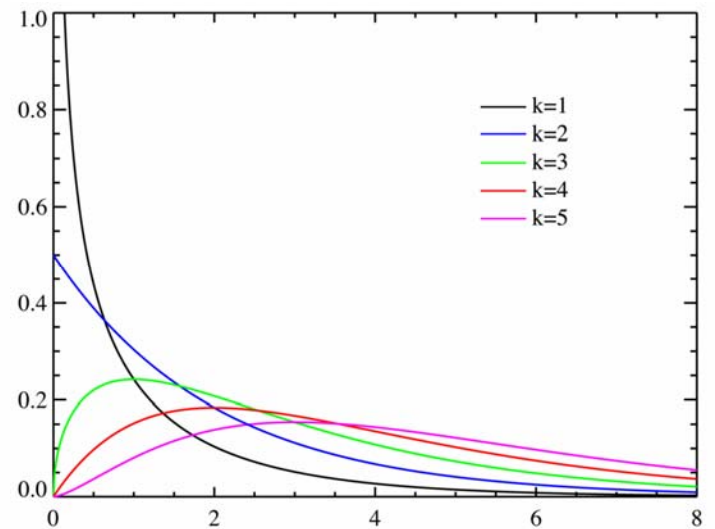
The **χ^2 (chi-square) distribution** is used for a variable that is distributed as the square of values from a standard normal distribution. Variances tend to follow a χ^2 distribution and we use the distribution to test for differences between observed and expected outcomes from a model (Categorical Data Analysis).

The **F-distribution** is a probability distribution for a variable that is distributed as the ratio of two χ^2 distributions and is used for testing hypotheses about the ratio of variances (this is a very important distribution for testing hypotheses using linear models, e.g., regression and ANOVA).

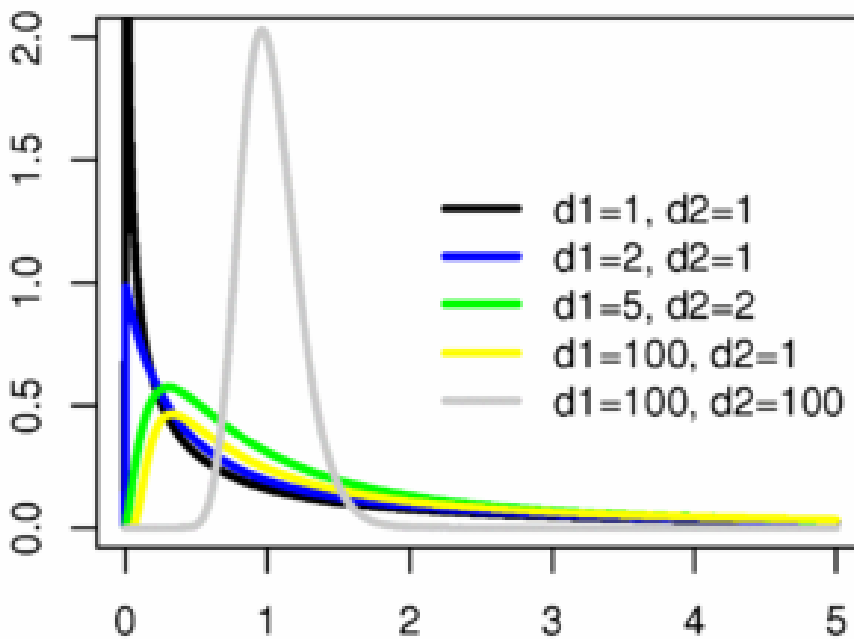
Examples of distributions used for statistical tests (for varying degrees of freedom):



t - distribution



χ^2 - distribution



F - distribution

Framing and Testing Hypotheses

(Based largely on Gotelli and Ellison (2004) chapters 4-7; as well as Underwood 1997 and Ecological Methodology by Krebs 1999)

Hypotheses can be simply defined as possible explanations for our observations. They often stem directly from our observations, the existing scientific literature, theoretical model predictions, intuition and reasoning, or all of the above

Good hypotheses:

- 1) Must be testable
- 2) Should generate unique predictions

In the biological sciences, we have two broad types of study designs that are fundamentally different. **Manipulative** studies involve the application of some treatment to a group of experimental units (e.g., one plot is burned and another is not, before measuring insect abundance). **Observational** (or mensurative) studies do not involve manipulative treatments, but only measurements (e.g., you might measure insect abundance in marsh vs. upland plots).

It is important to note that the statistical treatment of data is generally the same for each type of study. The difference is in the *confidence* we place in the inferences we make.

Comparing scientific methods

There are several methods of scientific reasoning that are used to make decisions about the hypotheses that we have formed.

Deduction is one such method. The logic of deductive reasoning proceeds from the general case to the specific case.

An example of deduction:

Statement 1 - All students at UNCW are from NC (major premise)

Statement 2 - I sampled this one student at UNCW (minor premise)

Statement 3 - This student is from NC (conclusion)

The three statements proceed logically so that the last must be true if the first two are true.

Inductive reasoning (**Induction**), in contrast, proceeds from the specific case to the general case.

Example of inductive reasoning:

Statement 1 – All 50 of these students are from NC

Statement 2 – All 50 of these students were sampled at UNCW

Statement 3 – All students at UNCW are from NC

Statement 3 represents a “probable inference” – it is likely to be true based on statements 1 and 2, but may be false

Statistics by its nature is an inductive process; we try to reach conclusions based on a sample of data (i.e., we very rarely have all of the data from the population of interest)

Inductive reasoning proceeds from (1) the development of models and hypotheses, to (2) predictions, and lastly to (3) observations which affirm the hypotheses

Advantages of the inductive method

1. Close link between data and theory
2. Modification of hypotheses based on data

Disadvantages of the inductive method

1. Only a single starting hypothesis (can be led down the wrong path)
2. Can encourage pet hypotheses
3. Derives theory only from empirical data, rather than theoretical models or intuition

Hypothetico-Deductive Method

- Begins with multiple working hypotheses to explain a set of observations
- Each makes unique predictions that can be tested
- Goal is not confirmation, but to falsify as many alternatives as possible
- Accepted explanation withstands repeated attempts to falsify it

Advantages of the H-D method

1. Simple explanations (parsimony) considered first
2. Forces consideration of multiple hypotheses from the start

Disadvantages of the H-D method

1. May not be multiple hypotheses available in all cases
2. The "correct" hypothesis must be among those at start

Platt (1964) attributed the success of molecular biology this century due to the widespread use of hypothetico-deductive logic trees

Statistical vs. Scientific Hypotheses

A statistical hypothesis tests for pattern in the data. The **statistical null hypothesis** would be one of "*no pattern*", meaning no difference between parameter estimates or no relationship between a variable and some measured factor. The **statistical alternative hypothesis** would be that "*some pattern exists*"

***But, how do these patterns relate to the scientific hypothesis?

If the statistical null hypothesis is rejected, it only tells us that there is pattern in the data. It doesn't automatically mean that we reject the scientific null hypothesis.

For example, the 'Bigger is better' hypothesis has the expectation that, at a given age, larger fish will survive at higher rates than smaller fish during early life. If we conducted an experiment and found no differences in our estimates of survival, we would fail to reject the statistical null hypothesis. This does not support the scientific hypothesis.

Alternatively, the 'Ideal Free Distribution' hypothesis predicts that densities of animals in given habitats will be adjusted to result in equal fitness. If we measured some component of fitness, say growth rate, among habitats and detected no differences, we would also fail to reject the statistical null hypothesis. However, in this example, failure to reject the statistical null actually supports the scientific hypothesis.

Statistical Significance and the P-value

A Conceptual example: Comparing two means

Suppose we measure fat content of squirrels in urban versus rural habitats (plots) and find the following:

Fat content of urban squirrels = 340 g of lipid per kg body weight

Fat content of rural squirrels = 810 g of lipid per kg body weight

How do we know if this difference is large enough to be attributable to the different environments?

First, we must define the statistical null hypothesis: that the difference represents random variation

H_0 = some specific mechanism does not operate to produce the observed differences

We can then define one or more statistical alternative hypotheses

H_A = observed difference is too large to be due to random variation alone

The set of H_A 's can be broadly defined as "not H_0 "

The statistical hypotheses are focused on pattern or no pattern in the data. Any inference of environmental effect is made later (mechanism inferred). This will depend separately on the quality of our design (e.g., if squirrels living in urban habitats were also smaller and we didn't account for it, then environmental and body size effects would be confounded). It would then be hard to infer that the difference in environment was the primary mechanism for the difference in fat content.

The P-value

The P-value that you see reported when data analyses are summarized in the papers you read has a clear definition.

P-value = The probability of observing a pattern as extreme or more extreme than the one observed if the null hypothesis is true.

The probability is stated $P(\text{data}/H_0)$ = the probability of observing the data given the null hypothesis

It is not the probability of the null hypothesis given the data

Thus, when P is low, we are saying that the probability of the observed pattern is very small if the null hypothesis were true, so we reject the null.

If we reject the null hypothesis of no pattern, we can “accept” H_A when our alternative hypothesis is stated broadly as:

$$H_A = \textit{pattern exists} = \textit{observed variation is not just random}$$

However, in most cases we are interested in a specific alternative hypothesis, such as:

$$H_A = \textit{difference in fat content due to environment}$$

Often, and depending on the quality of our experimental design, we can't simply accept this type of H_A , which is stated more narrowly.

For example, suppose our null hypothesis was that all students at UNCW are from North Carolina, and our alternative hypothesis was that at least 10 students at UNCW are not from North Carolina

$$\begin{aligned} H_O: & \textit{All students at UNCW from NC} \\ H_A: & \geq 10 \textit{ students at UNCW not from NC} \end{aligned}$$

After surveying several students around campus, we encounter 1 student from New Jersey. We can reject H_O , but we can't draw any conclusions about H_A . Had our H_A been less specific (e.g., all students at UNCW are not from North Carolina), then we could “accept” it. However, “acceptance” of a broadly stated H_A doesn't provide us with any information about the distribution of the home states of UNCW students.

How do we interpret a P-value that is very large (i.e., close to 1)?

Note that since P is a probability it is bound between 0 and 1

A large P-value signifies that there is a high probability that observed differences could have occurred simply due to random variation given that the null hypothesis is true. So, we cannot reject the null.

What determines the P-value?

1. The number of sample observations (n)
2. The differences between sample means ($Y_i - Y_j$)
3. The level of variation among individuals (s^2)

1. Higher n = lower P-value

The Law of Large Numbers states that as n increases, the more likely we are estimating the true population parameters (means, medians, regression coefficients, etc.) and can detect a real difference between them.

2. Greater difference between Y_i and Y_j = lower P-value

This is termed the **effect size**, and it is often what we are really interested in estimating. We want to know how different things are or how much of an effect a certain factor has on our response.

3. Smaller s^2 within each group (i & j) = lower P-value

If the variance within each group we are comparing, or the variance associated with an estimate of a parameter, is small, then we are more likely to detect differences or factor effects.

How small does P need to be?

Suppose in our squirrel example $P = 0.03$

This can be interpreted that, *if the null hypothesis were true (no pattern, instead only random variation)*, the probability of observing a difference in fat content as large or larger than 470g per kg body weight is 3 in 100

Stated another way, if we conducted this experiment 100 times, only three times would we expect to see a difference this large

This seems highly unlikely, so we reject the null hypothesis and conclude that there is a pattern related to environment.

Why $P < 0.05$?

It turns out that there is no threshold critical value for P , but the traditional operational value is 0.05

Keep in mind that P is distributed as a continuous variable that can take on an infinite number of values between 0 and 1. We have set up these dichotomous decision rules (reject, do not reject) that are so widely applied. There are different approaches, which we will discuss soon.

Restricting hypothesis rejection to $P < 0.05$ is actually very conservative (so, why do we choose to be so conservative?). If ocean conditions caused forecasters to predict that there was a 70% chance of drowning if you went surfing that day, you would probably stay home and work on your Biostats assignment.

For science, however, a 30% chance that you would have been incorrect in your rejection of the null hypothesis is too big a risk. Scientific progress, which builds on the previous work of others, depends on conservative decision making to keep false rejections low. In addition, psychology experiments have long noted that humans are predisposed to recognizing patterns, often seeing patterns where none exist.

A note of caution: *A low P-value is not a guarantee of good science, you can still get low P-values with poor experimental design.*

Remember:

- A *scientific hypothesis* poses a formal mechanism to account for patterns in the data
- A *statistical hypothesis* just establishes pattern

Refer back to manipulative vs. observational designs and our squirrel example (which is observational). If the differences in fat content between small and large squirrels are not accounted for, then our inferences related to environmental effects will not be strong

Errors in Hypothesis Testing

In all cases, the null hypothesis is either true or false (and we would know which if we had all possible information). Instead, our data is incomplete and represents a sample of the population. We are left to rely on methods of statistical inference to reject the null or not. This means that we are going to make mistakes sometimes and reject a null hypothesis that is actually true and fail to reject a null hypothesis that is in reality, false.

These possibilities generate a 2x2 table of potential outcomes:

	<u>Do not reject H_0</u>	<u>Reject H_0</u>
H_0 True	Correct	Type I error (α)
H_0 False	Type II error (β)	Correct

Type I Error

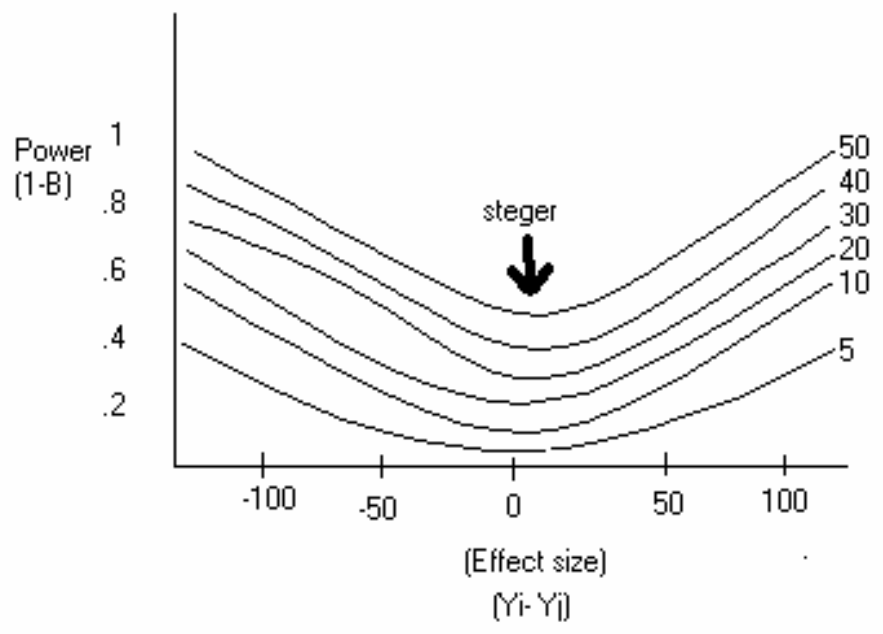
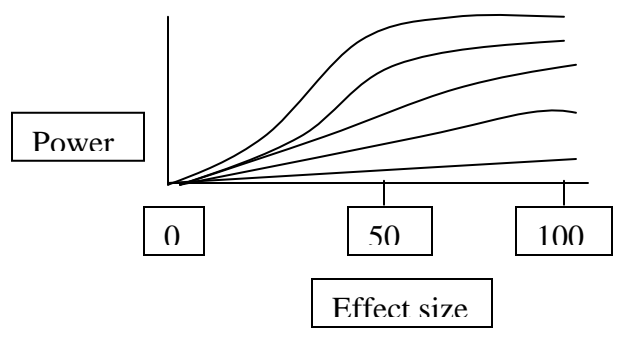
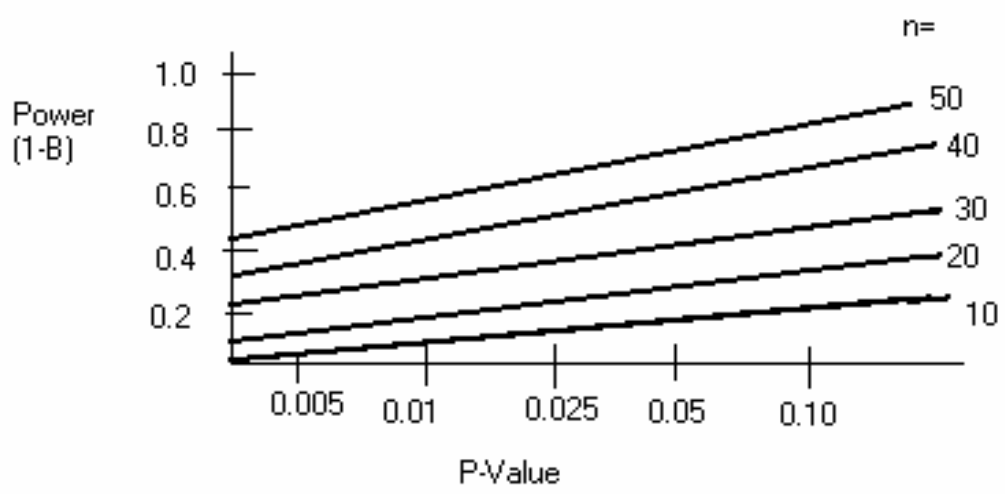
- Incorrectly reject a true null hypothesis
- Draw an incorrect inference that some factor beyond random variation is causing patterns in our data
- Denoted by alpha (α)
- Calculation of P is actually an estimation of alpha
- Smaller P = lower chance of Type I error
- We set rejection rule at $P < 0.05$ to minimize chances of committing a Type I error

Type II Error

- Fail to reject a false null hypothesis (*can remember by double-F*)
- Concluded incorrectly that only random variation is present
- Denoted by beta (β)

The **power** of a statistical test is calculated as $1-\beta$, which equals the probability of correctly rejecting the null when it is false

Generally, the probability of committing a Type I and II error is inversely related. But, there is no general formula, the relationship for any specific test depends on the effect size, the sample size, and the quality of the experimental design



The basic process of parametric statistical analyses

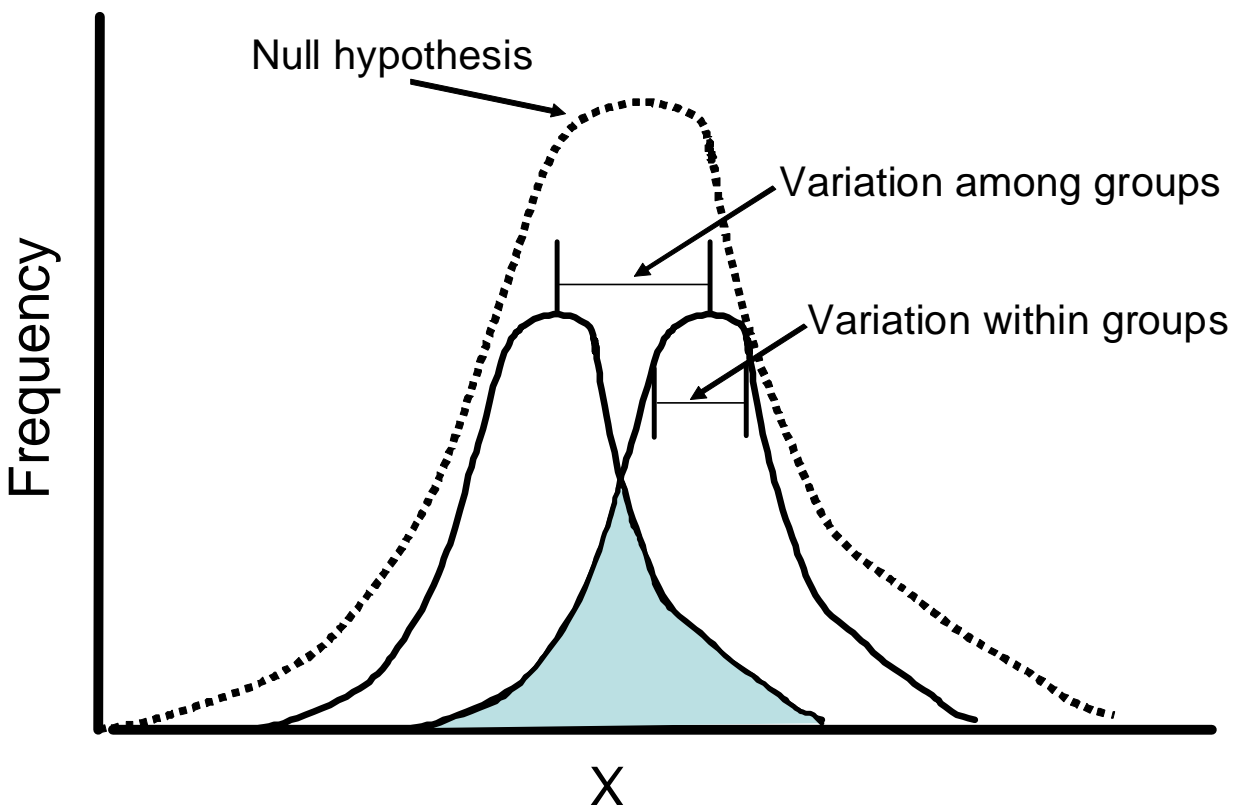
The most commonly applied statistical analyses (t-tests, ANOVA, regression) assume that the data were sampled from a specific distribution (*usually the normal*). Parameters (μ and σ^2) of this distribution are estimated and are then used to calculate tail probabilities of a true null hypothesis. Remember, we are estimating the probability of our observations given the null, $P(\text{data}|\text{null})$

There are 3 general steps in parametric analyses:

1. Specify the test statistic
2. Specify the null distribution
3. Calculate the tail probability

1. Test statistic

Example: Testing for a difference between two sample means



The null hypothesis (H_0) is that both groups of data are drawn from a single normal distribution. What we mean is that a single mean and variance (μ, σ^2) best represents both groups

The alternative hypothesis (H_A) is that the sample data for the groups is drawn from two different populations, each with its own mean and variance (although we assume variances are similar)

The closer the 2 curves are together, the more likely the data would have been collected under a true null

The more separate the 2 curves are, the less likely the data would be observed under a true null

The test statistic that is used is specific to different kinds of tests and we will cover them all in detail. When using Analysis of Variance (ANOVA), a test statistic was developed to quantify the overlap in the distributions as a ratio of the variances

F-statistic = the ratio of the variance among groups to the variance within groups. We have already introduced the F-distribution which is the null distribution used to estimate tail probabilities in ANOVA, and we will cover it in more detail soon.

2. Specify the null hypothesis

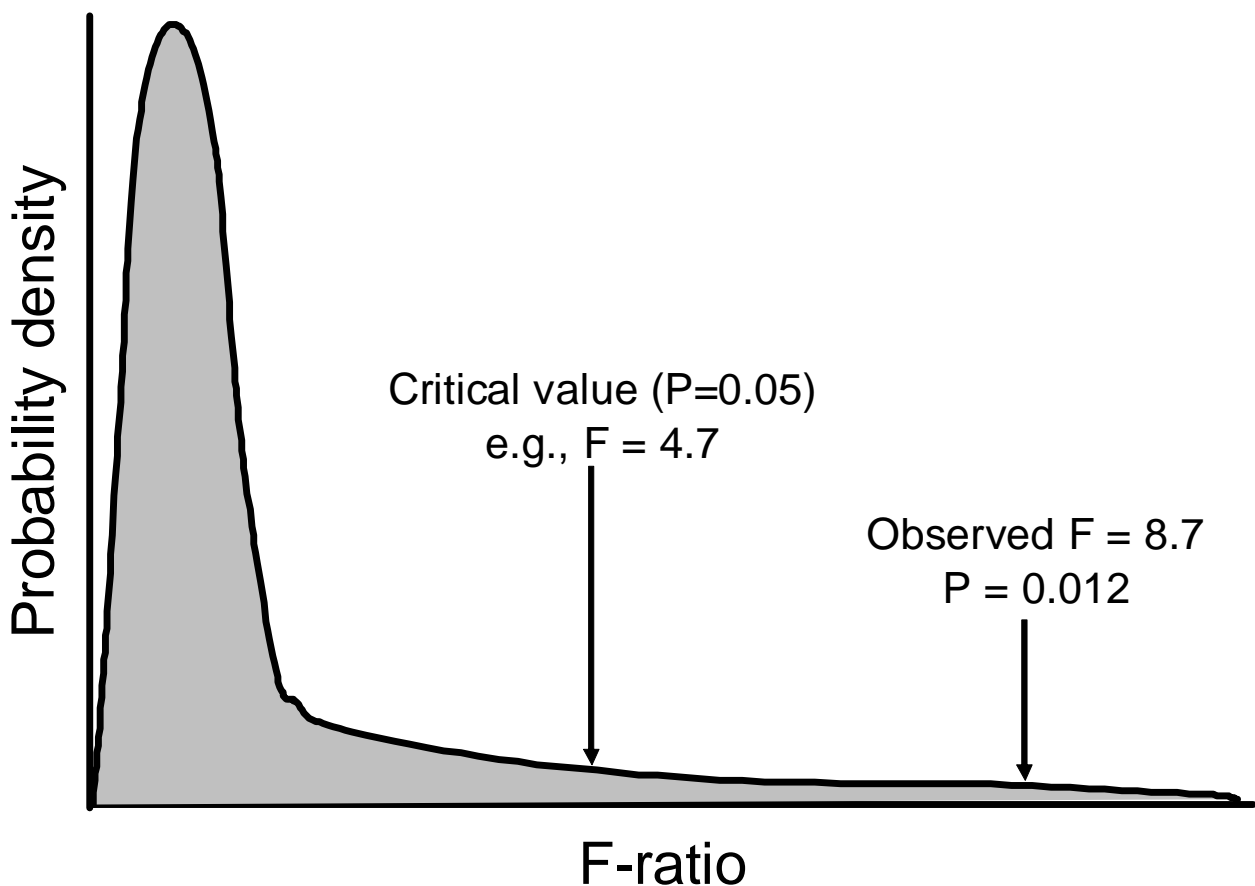
The null hypothesis is that all samples are drawn from one population. For this case specifically, the observed differences between group means are no larger than expected by chance (i.e., random variation)

If the null is true, then variation among groups should be small relative to variation within groups and the F-ratio should be close to 1.0

3. Calculating the tail probability

The P-value is an estimate of the probability of obtaining a specific F-ratio given a true null hypothesis [Remember, $P(\text{data}/\text{null})$]

For the F-distribution, the P-value is calculated as the proportion (%) of the area under the curve to the **right** of the observed F-ratio



To interpret this plot, we would expect that our observed F-ratio of 8.7 would only occur with probability 0.012 if the null were true. In this case, we would reject the null hypothesis and conclude that the two groups were sampled from two different populations

Assumptions of parametric statistical analyses

1. Data represent random, independent samples
2. Data sampled from a specific distribution

The first assumption is common to all statistical approaches (e.g., Bayesian, Monte Carlo, Information-theoretic). It is the second assumption that is unique to parametric approaches. Specific tests have additional assumptions (e. g., ANOVA requires homogeneity (*equality*) of variances)

Sampling and Experimental Design

What's the question?

Simple survey data can be used to address questions such as:

Q: Are there spatial or temporal differences in variable Y?

When conducting biological field research, a well designed survey can answer many questions of interest, but still often represents the initial step in a line of research that ultimately intends to address mechanistic questions (identifying important processes)

A more specific question might be:

Q: What is the effect of factor X on variable Y?

To address this question using a manipulative experiment, the investigator would establish different levels of factor X and then measure the response of variable Y (and then proceed to calculate a P-value and decide whether to reject the null hypothesis)

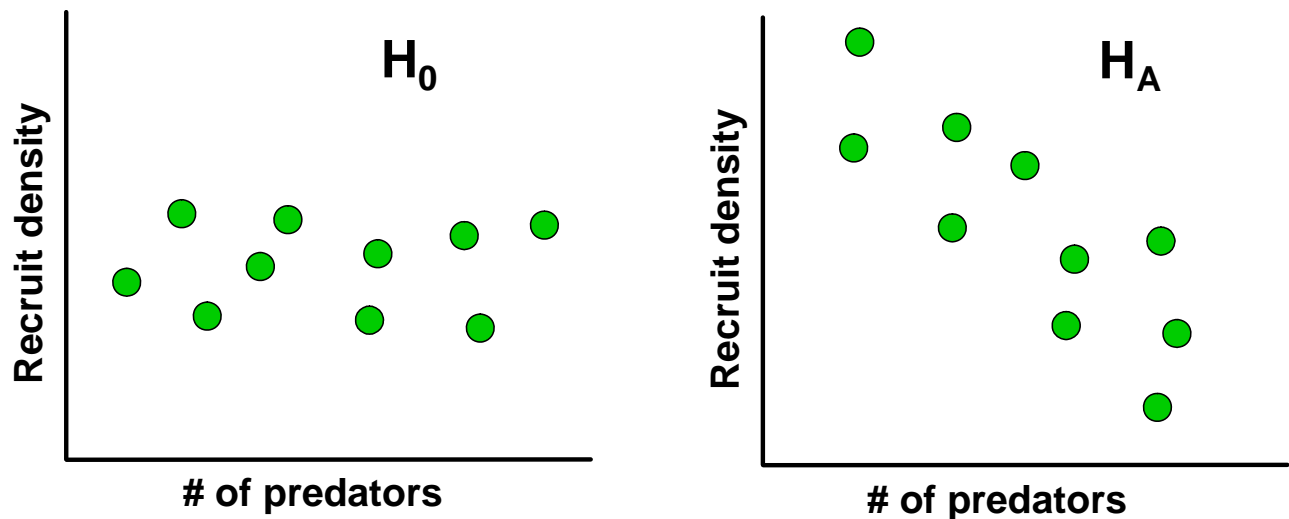
***Ultimately we want to ask if responses of variable Y are consistent with some *Hypothesis H* (our scientific hypothesis)

Design of manipulative experiments

Investigator alters levels of one or more factors (X) and measures the response of one or more variables (Y).

An example: Does predation by large predatory fish control the density of newly recruiting fishes on coral reefs?

One approach → alter predatory fish density and measure the density of new recruits (Plot X vs. Y)



Investigator could use a regression approach. If so, the slope would be an indication of the strength of relationship

Challenges of manipulative experiments (specifically field experiments)

1. The spatial scale often limited (80% of field experiments in biology conducted at scales less than 1 m²)
2. The results of even well-replicated small-scale experiments may not scale up well
3. If spatial scale is made large, replication is sacrificed
4. Often restricted to small-bodied, short-lived organisms
5. Difficult to manipulate one and only one factor (confounding problem)
6. Space, time, labor, cost, all limit # of replicates

Natural Experiments

Observational studies generally take advantage of the natural variation that is present in a variable of interest. They generate the same kind of data as manipulative experiments and are analyzed using the same statistical approaches. However, they lack the controls that we have when we conduct manipulative experiments and their interpretation is therefore, more difficult. It is harder to identify cause and effect relationships and identify the factors most responsible for an observed response.

There are two general types of natural experiments. Snapshot experiments are broad in spatial coverage, but limited to short time intervals (e.g., surveying multiple rivers in a single year). Trajectory experiments are broad in temporal coverage, but have limited spatial scope (e.g., surveying a single river for a decade).

Replication

How much replication do we need?

It depends..... on variation present in our sample data and the effect size we wish to be able to detect

***Remember, the P-value of any statistical test depends on n , s^2 , and the effect size (the differences you can detect)

But, how do we estimate variance (s^2) before we begin sampling?

Answer is that you can't, but investigators will often collect pilot data for this purpose or an estimate can be obtained from previously published studies that have measured the same response variable

Often the number of replicates we eventually measure comes down to affordability. Time, labor, and money often combine to dictate the number of replicates that is reasonable. Therefore, it is important that we estimate these *costs* up front when designing any experiment to ensure that we are realistic in what we can accomplish.

Gotelli and Ellison (2004, p. 150) suggest using a **Rule of 10** when deciding how many replicates to measure. As are all 'rules of thumb', this is subjective, but it's not a bad starting point. There are certainly many cases, as they point out, when less than 10 will be very much sufficient and others that will require many more than 10. What is most important is that you spend time thinking hard about your question, your effect size, and how much data you can reasonably expect to collect in a fixed time interval. *And always anticipate and plan for data loss.*

Independence of replicates

Definition of statistical independence: observations collected in one replicate do not have an influence on the observations collected in other replicates

Stated another way, replicates are statistically independent when **residual errors are independent** (meaning that a high residual for one observation doesn't necessarily cause a low/high residual for a separate observation)

Nearly all statistical analyses assume replicates are independent of one another. **Pseudoreplication** = statistical treatment of experimental units as independent when they are not (see Hurlbert 1984, *Ecol. Mono.* 54:187-211 for a detailed treatment of the subject).

An experimental example: Consider an experiment to examine salamander use of manipulated versus unmanipulated stream habitats. You set up several manipulated sections of stream bottom from which you've removed a fraction of the benthic invertebrate prey community. You also monitor several unmanipulated sections of stream bottom.

While snorkeling, you measure the number of salamanders using each section per hour and find:

- 20 salamander per hour to unmanipulated habitat
- 10 salamander per hour to manipulated habitat

However, you notice that salamanders that visited the manipulated habitat leave quickly and move to an unmanipulated section. Now, your observations are not independent. If the habitat treatments were farther apart, the pattern you observed would likely have been different.

A statistical test would likely produce a low P-value that could be spurious (Type I error). Non-independence could also result in a Type II error (failure to reject a false null). The exact effects on P-values and statistical power are unknown and will be specific to each experimental design.

In a case such as this, replicate treatments should be separated by enough space/time to ensure that they don't affect each other. *However, this generates several problems.* First, we often don't know the extent of spatial/temporal separation necessary to achieve independence. Second, it is usually expensive to separate samples in space/time. And third, a large degree of separation can ensure independence but may introduce new sources of variation.

Another example: Suppose you are interested in measuring the condition of bird fledglings in fragmented versus non-fragmented forests. You design a survey and sample 4 sites in each forest type, randomly select 20 trees per site, 5 nests per tree, and measure condition (weight per unit length) of 3 fledglings per nest.

What is the true replicate? Is it the fledgling, the nest, the tree, or the site? It will depend on how their residual errors are related. Clearly, fledglings within the same nest are not independent (their residuals may be positively or negatively correlated). The 5 nests have the same external factor (the tree) in common, but this can be accounted for in the statistical model so long as the residual errors of the 5 nests are independent.

Confounding factors

If we return to the salamander-stream habitat example, suppose that our habitat spatial separation was good, but one habitat treatment was located adjacent to deep pools and the other wasn't. Deep pools in streams likely contain more predators, which salamanders would do well to avoid. Now, we cannot separate the effects of our habitat treatment from the effects of deep pools.

The point is that although we think we understand the biology of the organism we are studying, there are likely *unmeasured* or *unknown* variables that can affect the response. In non-manipulated natural experiments, we often cannot avoid the presence of confounding variables that operate during our study.

Replication and Randomization

Through proper replication and randomization, we can offset problems introduced by confounding factors and non-independence

Replication = establishment of multiple plots or observations within the same treatment group

Randomization = random assignment of treatments or random selection of samples

Gotelli and Ellison (2004) point out that many samples or sites are really *haphazardly* chosen rather than being truly random (which implies using some mechanism to generate a random number). Haphazard means to follow some general criteria to achieve a homogeneous distribution of samples or sites.

If we return again to our salamander example, a properly replicated and randomized design would have a sufficient number of replicates (say 10, based on 'rule of 10') of each habitat treatment (manipulated and non-manipulated). The location of the habitat sites would be random and assignment of habitat treatment to the different sites would be random.

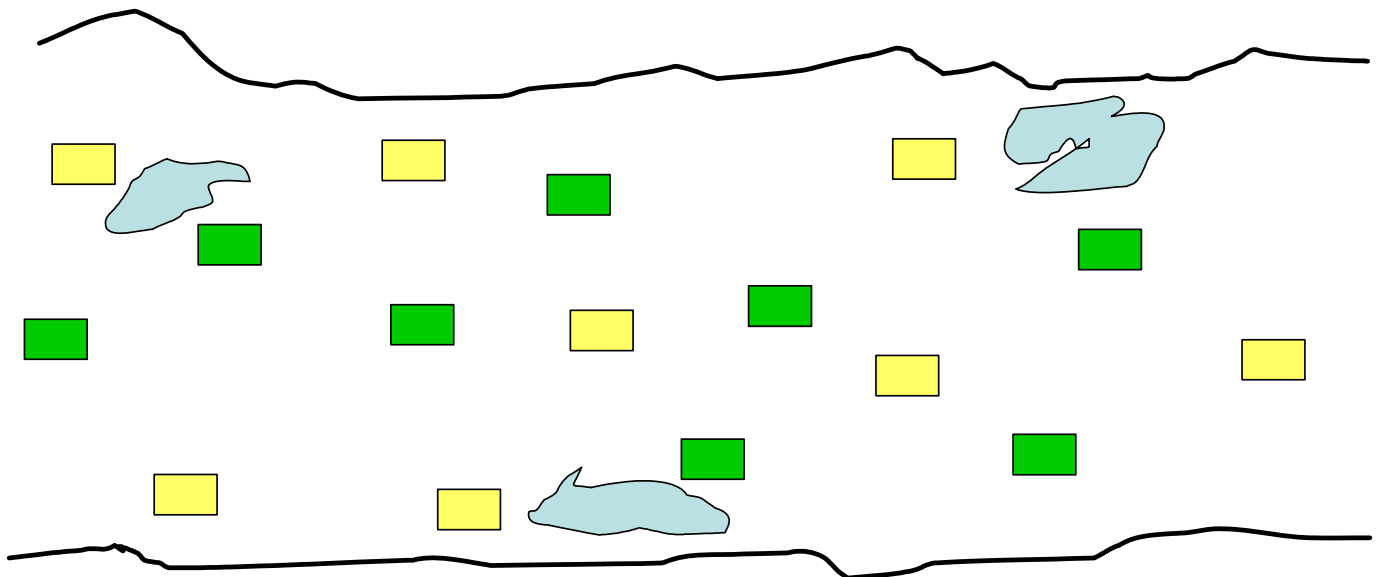
Proper replication and randomization (*must have both*) reduces the problems caused by confounding factors because all treatment levels occur within all levels of confounding factor (e.g., both manipulated and non-manipulated stream sections would be located adjacent to deep pools and shallow riffles)

We can now test for the effect of deep pools as a covariate (independent of our habitat treatment). The presence of deep pools adds more variation, but it is not biased because it is not systematic. We have randomized the placement of our habitat treatments to avoid any bias associated with deep pools.

Clearly, if we knew the proximity to deep pools was important ahead of time, we would have controlled for it

Proper randomization can also reduce the chance of bias due to non-independence. By locating our habitat treatments at random distances from each other (beyond some minimum distance) the effects of non-independence will vary with distance and may cancel each other out.

Again, we must both randomize and replicate to reduce the influence of confounding factors and non-independence



Questions to ask when designing a field experiment (from Gotelli and Ellison pp. 158-161)

1. Are plots or enclosures large enough to ensure realistic results?

Your design should ensure that the spatial scale of your experiment is appropriate relative to animal movement and behavior

2. What is the Grain and Extent of the study?

Grain = the size of the smallest unit of study (usually the size of a single replicate or plot)

Extent = the total spatial area encompassed by all sampling units

Often, it is hard to know which is best. An experiment with small grain and large extent is generally a good combination. Small grain allows manipulations and observations at the spatial scale of the organism you're interested in, and large extent expands your domain of interference.

3. Does the range of treatments span the range of possible environmental conditions?

For example, if you are testing the effects of temperature on an organism you should be sure to test temperatures near the extremes experienced by the organism, not just those close to the mean

4. Have appropriate controls been established?

In many cases, having simply unmanipulated plots to go along with your manipulated plots is not a sufficient control.

For example, caging experiments often include other effects due solely to the cages. A cage to exclude predators might also affect other processes, such as foraging by the treatment animal/plant.

If your response variable was the growth rate of salamanders, you would need to use:

1. Non-manipulated plots
2. Cage control (predators can enter, but simulates other cage effects)
3. No predator cage (full cage)

Then you can make all pairwise comparisons:

- 1 vs. 2 = cage effects
- 2 vs. 3 = predator effects
- 1 vs. 3 = combined effect of cages and predators

5. Have all replicates been manipulated in the same way except for the treatment application?

For example, transport and handling effects during a hooking mortality experiment for recreationally caught fishes. One needs to separate the effects on mortality rate of the hooking treatment and the transport and handling effects. Solution = a transport and handling treatment.

6. Have appropriate covariates been measured in each replicate?

Covariate = a continuous variable not under control of investigator that may affect response

****Measuring covariates is not a substitute for randomization and replication****

Types of Experimental Designs

(Gotelli and Ellison 2004, chap. 7)

In the broadest sense, our experimental design simply reflects our decisions about how our replicates will be physically arranged in space, and how they will be sampled through time. These decisions rest on the *definition of our replicate*, and our earlier decisions about the *number of replicates* we can realistically expect to collect, spatial and temporal *independence* of our replicates, and our strategy to *randomize*.

When we collect data, the variables that we measure are one of two general types:

Categorical vs. Continuous variables

Categorical variables = take on 2 or more discrete categories and are modeled as *discrete random variables*

Continuous variables = measured on a continuous numerical scale; can take on a range of real and integer values and are modeled as *continuous random variables*

We then designate our measured variables as **dependent or independent**:

The *dependent (response) variable* is the variable whose response we are interested in measuring and understanding (cause)

The *independent (predictor) variable* is the variable that is manipulated experimentally or varies naturally, and which we hypothesize may cause a response in our dependent variable

Four different design classes

Dependent variable (Y)	Independent variable (X)	
	Continuous	Categorical
Continuous	Regression	ANOVA
Categorical	Logistic regression	Tabular

*note that not all designs fit neatly into one of these categories (e.g., ANCOVA used when you have both categorical and continuous independent variables)

Regression designs

When the independent variable (X) is measured on a continuous scale, regression designs are appropriate. If the dependent variable (Y) is also continuous, then we use linear or non-linear regression models. If the dependent variable (Y) is categorical, then we can use logistic regression.

Single factor regression design example

Returning to our reef fish recruitment example. We wish to know if the density of predatory fish affects the number of reef fish recruits.

Dependent or response variable = number of reef fishes

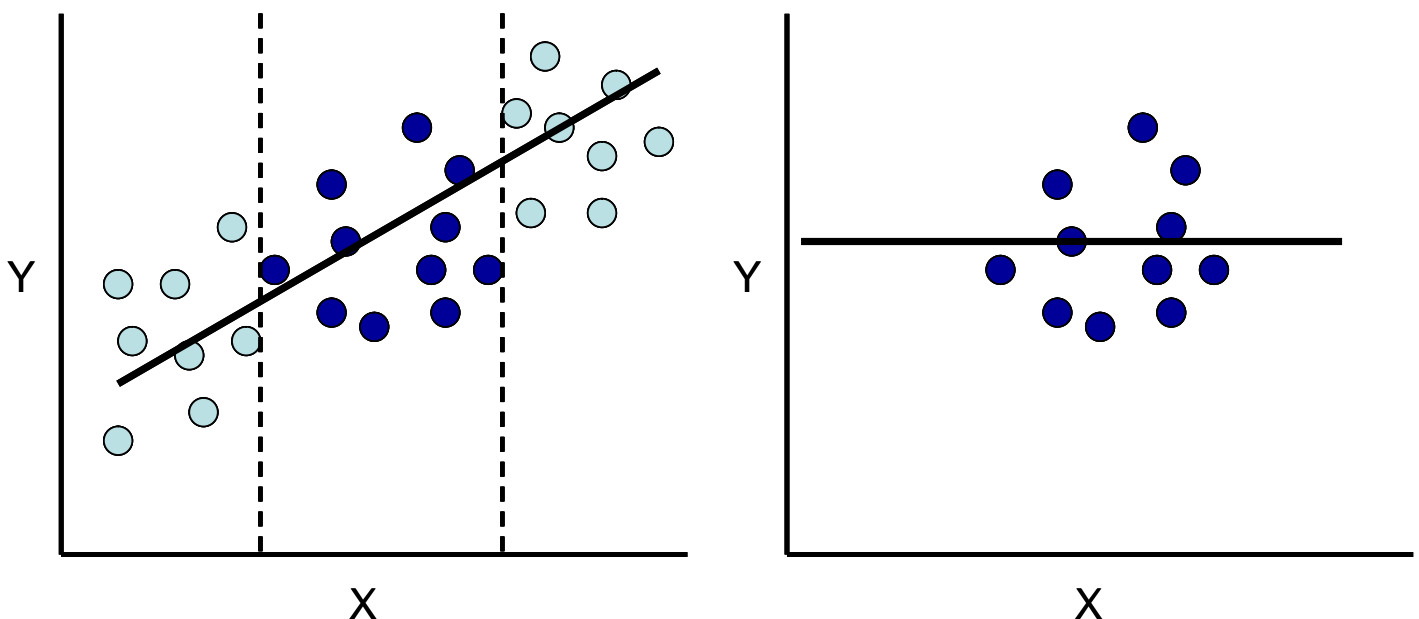
Independent or predictor variable = predator density

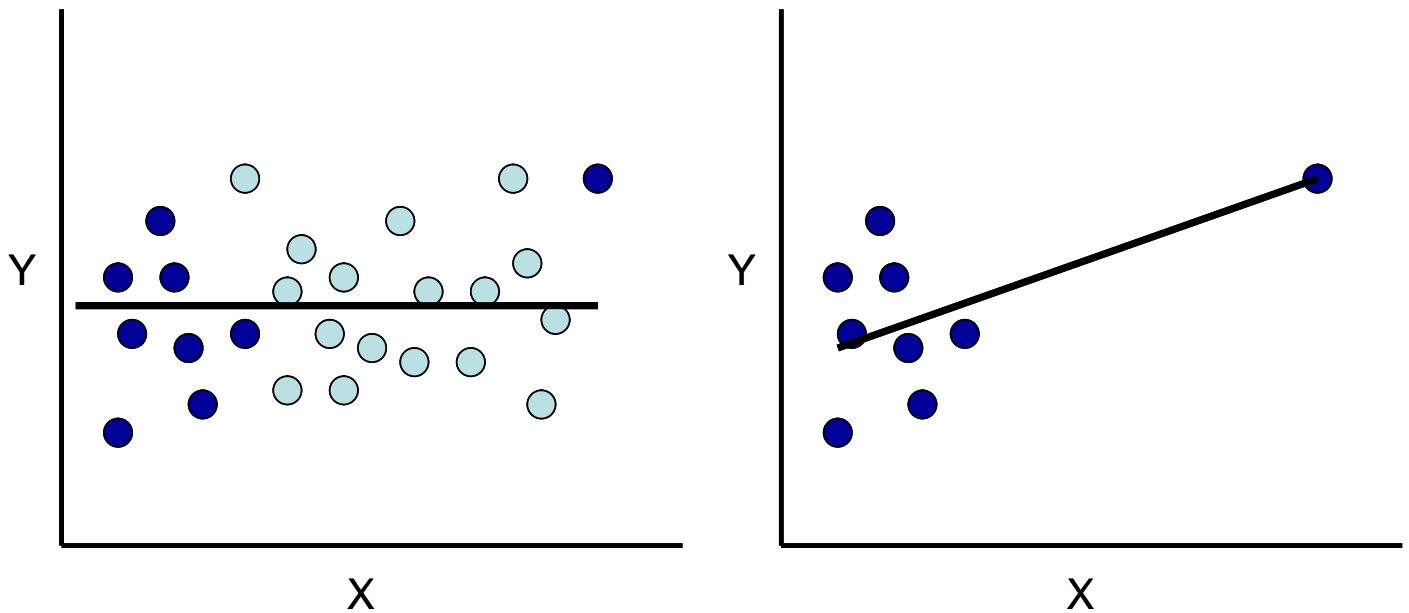
In an experimental study, we control or manipulate X and measure Y. In a natural experiment, we use the range in X that exists naturally, and measure Y along that range.

Regression assumes that variation in X causes variation in Y, i.e., that there exists a functional relationship between X and Y ($Y \sim f(X)$). This is different from *correlation* which does not specify cause and effect, but only tests for the strength of the association between two variables (X and Y).

2 Principles of regression designs:

1. We should ensure that the range of X is large enough to capture the full range of responses in Y (this helps to avoid Type II errors).
2. We should ensure that the distribution of X values is approximately uniform to minimize the influence of outliers or leverage points (this helps to avoid Type I errors).





Multiple regression designs

When we measure 2 or more predictor (X) variables, we need to use a multiple regression design. For example, suppose that our reef fishes were found in habitats that differed in their structural complexity. Now, in addition to the density of predatory fishes, differences in habitat may also affect the number of reef fishes that we measure.

An assumption is that the predictor (X) variables are independent of one another. However, in many natural experiments (observational), the predictors are often confounded (e.g., high predatory fish density with high structural complexity). This is referred to as *collinearity*, and it makes it hard to determine how much variation in our response is due to each predictor variable.

***We should be careful not to measure everything without adding more independent replicates. Instead, we should try to use only variables that we think may be biologically meaningful.

Stated differently, we shouldn't depend entirely on some statistical approach to select important variables for us after the fact, we should spend some time thinking hard about what we measure and why before we collect the data.

ANOVA designs

ANOVA = Analysis of Variance

These designs are widely applied throughout the biological sciences and are appropriate when the independent (predictor) variable is categorical and the dependent (response) variable is continuous.

In ANOVA, the categorical predictor variables are often referred to as **factors**. The different categories (or levels) of the predictor variable are referred to as **treatments** or treatment levels. Within each treatment, we will make multiple observations (**replicates**).

Single-factor and multi-factor ANOVA designs

You will often see researchers refer to a test as a one-way, two-way, or three-way ANOVA. These represent single- and multi-factor designs.

In a single treatment or single-factor design, we will have one predictor variable that acts as our factor of interest. We then test a response variable at several levels of our predictor variable. Each factor value = a treatment level (e.g., O₂ consumption by mice at different treadmill rates; the treatment = treadmill rate)

In a multi-factor design, we have 2 or more predictor variables (factors). Ideally then, each factor is applied with all levels of the other factors in a fully crossed design (n increases)

Example of 2-way ANOVA design:

We want to test the effects of treadmill rate and body size on the respiration rate of mice.

Single-factor design for treadmill rate

	Treadmill rate (cm per second)				Total
Level	1	2	3	4	
n	10	10	10	10	40

Single-factor design for body size

	Body size (g)				Total
Level	2	4	6	8	
n	10	10	10	10	40

2-Factor Design

Body size	Treadmill rate			
level	1	2	3	4
2	10	10	10	10
4	10	10	10	10
6	10	10	10	10
8	10	10	10	10

n = 160 (using "rule of ten")

Why not just run two separate single-factor designs with half the reps?

The advantage of the multi-factor design is that you can test for main effects and the **interaction** of factors

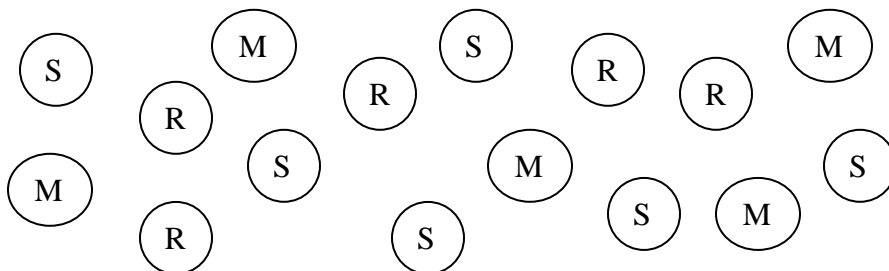
The main effect of any single factor in a multi-factor design is the response to each level of that factor *averaged over all levels of other factors*. In our example, the main effect of treadmill rate is examined by calculating the respiration rate at each speed, averaged over all body sizes.

The problem is that interaction terms cannot be predicted from simply adding up the main effects. Interactions represent unique responses to specific treatment combinations, and they may be greater (*synergistic*) or less (*antagonistic*) than expected from just the addition of main effects (e.g., all mice might respire at higher rates at faster treadmill speeds, but bigger mice may have higher respiratory costs than smaller mice only at high speeds).

Various single-factor ANOVA designs

Single-factor ANOVA is used to compare means among 2 or more levels of a factor or treatment. When we set up a single-factor ANOVA, there are various designs we can use. These include basic, randomized block, and nested designs. The randomized block and nested designs include a second factor. We are generally not interested in the response associated with that factor, it is included only to help control for sampling variation.

Example: Effect of substrate type on flatfish settlement; 3 substrate types: mud, sand, rock



To proceed, we identify “equal” sample sites for each substrate type. By “equal” we mean that other than differences in substrate type, the sites are similar in all other attributes that we think might affect flounder settlement. If we use the “rule of 10”, we would have 10 replicate sites for each substrate type, and we would measure flatfish settlers in some way.

We end up with a table of our data:

Treatment	Replicate	# of flatfish
Mud	1	12
Sand	1	6
Rock	1	0
Mud	2	9
Sand	2	4
Rock	2	1
Mud	3	11

It turns out that this simple design is very powerful for detecting the effects of our treatment. It can accommodate unequal sample sizes (n) among treatment levels and we can perform “post-hoc” or “*a posteriori*” tests to determine which treatment means are different from which others.

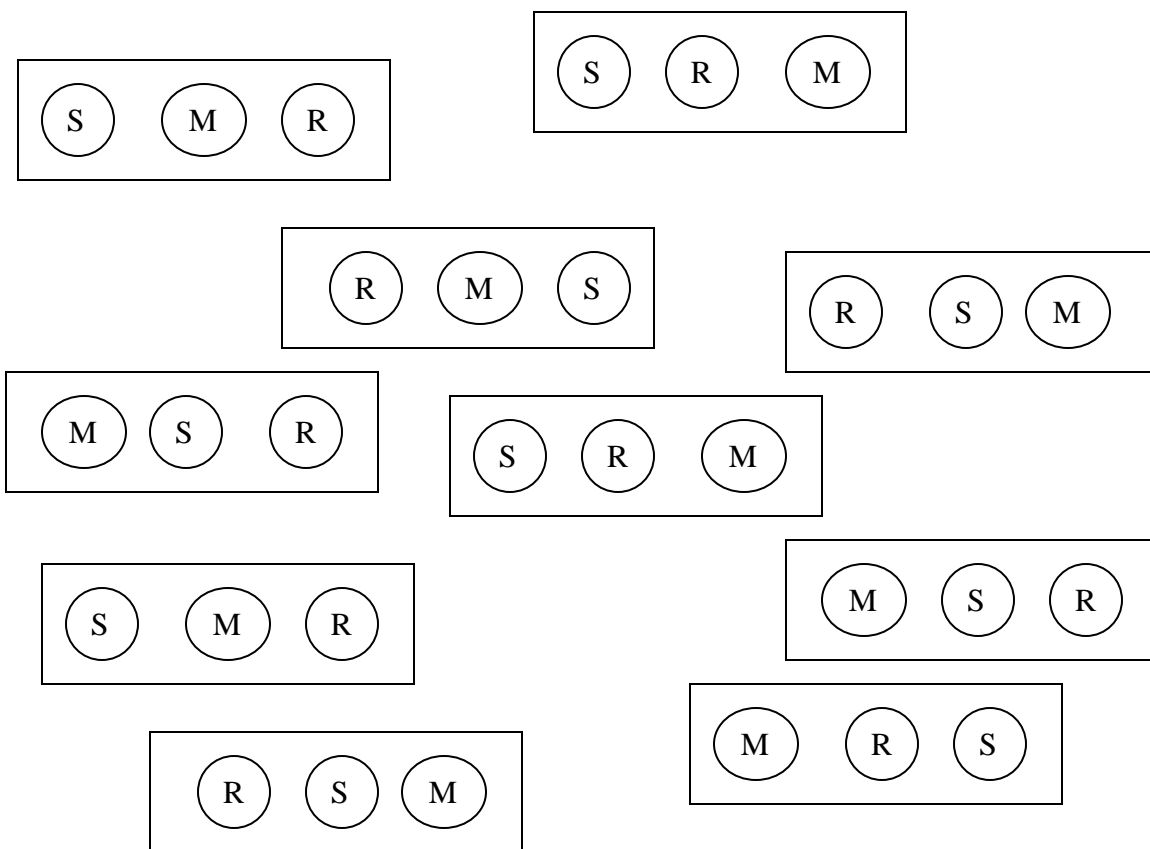
The primary disadvantage → the completely randomized design doesn’t explicitly account for differences in the environment (heterogeneity) among our sites. If our replicate sites are completely randomized, they should be distributed across a broad array of environmental conditions (**which is good!**), allowing our results to be generalized across many environments.

But....if the environmental “noise” is stronger than the treatment “signal”, our experiment will have low power (our ability to detect a pattern when one exists).

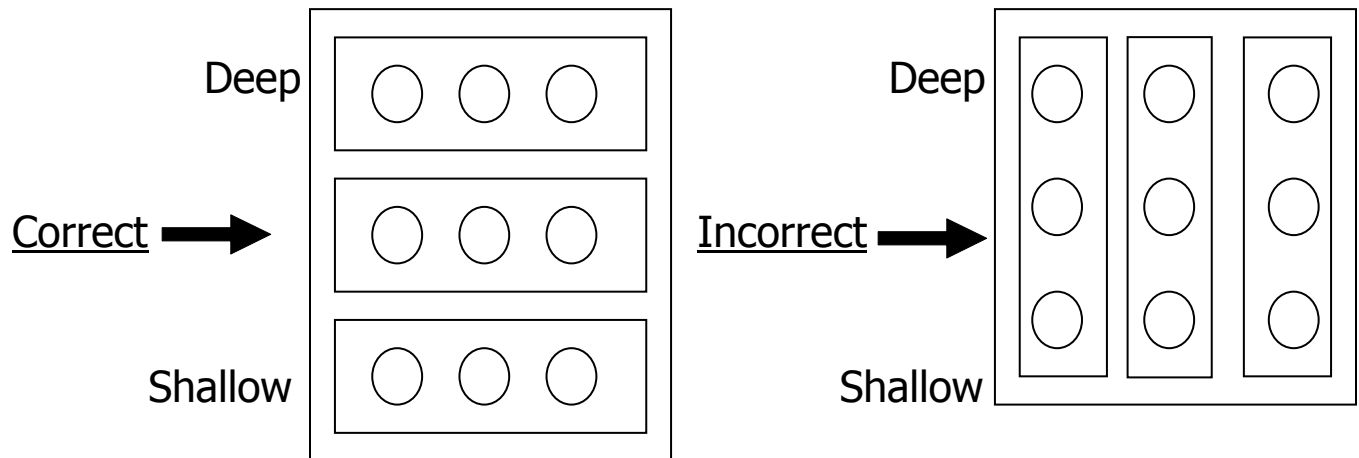
We can deal with this environmental “noise” in several ways.....

Randomized block design

In this type of design, we place our replicate sites within blocks, which are areas (space) or time periods within which the environmental conditions are relatively similar (*homogenous*). Our blocks should be arranged so that environments are more similar within a block than between blocks. Then, our replicate sites are assigned randomly within the blocks. In a simple randomized block design, each block contains exactly 1 replicate of each treatment level



This type of design helps us to deal with unknown environmental heterogeneity. If we know or suspect that an environmental gradient exists (e.g., depth) we would arrange our blocks along that gradient to ensure equal replication at each level of the suspected gradient.



We can also create blocks in *patchy* habitats, if we know the distribution of patches in the environment, or block *through time* (good for situations when you can't run all of your replicates simultaneously).

The randomized block design is more efficient than a completely randomized single-factor layout when there is sufficient environmental noise. The design reduces n while achieving the same power.

There are some disadvantages of the randomized block design:

1. Power is reduced if n is low and "noise" is weak
2. Potential for non-independence of replicates
3. If any replicates are lost, block is lost
4. **Assumes no interaction between block and treatment**

Underwood (1997) argues for replication within blocks to allow testing for the interaction (this would now be a 2-factor layout).

Nested ANOVA designs

Nested designs are those that include *subsampling* within the replicates. For instance, suppose that we measured the number of flatfish recruiting to each mud, sand, or rock site 3 times instead of just once. The total number of observations has increased from 30 to 90, but the number of *independent replicates* is still 30.

Subsampling within the replicates increases the *precision* with which we are able to estimate the response for each replicate. This is because the Law of Large Numbers tells us that higher n = more precise parameter estimates. In general, this will increase the power of our test.

Advantages of nested designs

1. Increased precision for each replicate (greater power)
2. We can test 2 hypotheses related to variation among treatments (using subsample averages) and variation among replicates

Nested designs lend themselves to a hierarchical sampling design. For example, in a single study of flatfish settlement you could look at subsamples nested within replicates, replicates nested within salinity zones, salinity zones nested within rivers, rivers nested within regions, etc. The variance in the response can be partitioned into components that represent each of the nested levels. We might find that most of the variation in flatfish settlement is found at the regional level. This will help us to identify important mechanisms as we move forward.

Some pitfalls of nested designs

1. Investigators may be tempted to treat subsamples as independent replicates (an example of *pseudoreplication*). This practice artificially boosts sample size and the probability of making a Type I error.

2. Complex nested designs can be hard to analyze if sample sizes are unequal
3. Subsampling often represents misplaced sampling effort
 - the power of ANOVA depends much more on true n than precision of replicates
 - subsampling is not a solution to inadequate replication
 - good if subsampling is cheap & easy or is necessary to avoid loss of replication (e.g., when using live animals)

Multi-factor ANOVA designs (2-factor layout)

In multi-factor designs, we examine the response to 2 or more factors simultaneously instead of just one. In terms of sampling and randomization, our approach is similar to a one-way layout.

Returning to the flatfish settlement example, suppose that in addition to substrate type, you want to test for the effects of predation by sea robins (an important predator of flatfish) on settlement rates.

We might come up with 3 levels of our predation treatment:

1. unmanipulated (this allows predation to occur naturally)
2. predator exclusion (substrate plots are surrounded with cages that allow flatfish to enter, but keep predators out)
3. cage control (a cage mimic, but predators move freely)

This is an example of a factorial design - testing 2 or more factors simultaneously.

*** A major key is that treatments are **fully crossed or orthogonal** = all treatment levels of each factor are represented with all treatment levels of each other factor

We now have 3 substrate treatments x 3 predator treatments = 9 treatment combinations

***If any of our 9 treatment combinations are missing, our design will be confounded (i.e., we won't be able to determine if our response is due to substrate or predator effects)

The main advantage of multi-factor designs, as mentioned earlier, is that we now have the ability to separate the main effects and estimate the non-additive interactions. Remember, treatment combinations may act additively, synergistically, or antagonistically. If we get a non-significant interaction term, it means that our main effects are simply additive (e.g., the effect of substrate type on flatfish settlement is the same whether sea robin predators are present or not).

The main disadvantage of multi-factor designs is that the treatment combination number can get large quickly, preventing adequate replication. Also, sometimes it is difficult to establish all orthogonal combinations.

Split-plot ANOVA designs

Split-plot designs represent an extension of the randomized block design for a two-factor layout, and were originally used mostly in agricultural experiments. A single block or plot is split into subplots and the second treatment factor is applied to the whole block or plot.

Returning again to our flatfish settlement example, we would block on substrate type just as in a simple randomized block design, but then we would apply one of our predator treatments to whole blocks

Predator treatment = whole plot factor
 Substrate type = subplot factor

Primary advantage is efficiency of the use of blocks:

For a 2-way layout, we would need:

30 cages for predator exclusion

30 simulated cages for cage controls

30 unmanipulated sites (no cages needed at these sites)

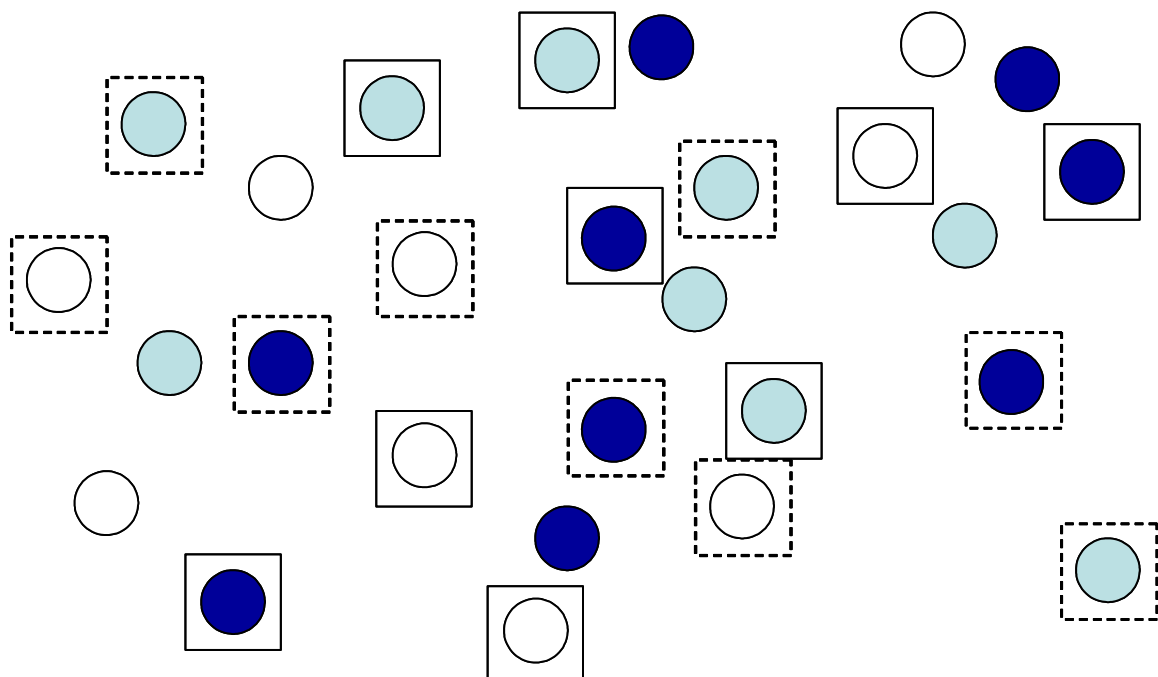
Alternatively, for a split-plot design, we only need:

10 cages for predator exclusion (each covers a block of 3 substrates)

10 simulated cages for cage controls (again, each covers a block)

10 unmanipulated blocks

Completely randomized 2-way layout



ANOVA designs for 3 or more factors

Not recommended because the treatment combinations get high quickly

For instance, 3 treatments each with 4 levels = $4^3 = 64$ treatment combinations x 10 replicates = 640 replicates

Field experiments with 3 factors are rare due to logistical problems

Repeated measures designs

Repeated measures designs incorporate the temporal variation that is introduced when multiple observations are made on the same replicate at different times. **Thus, the observations are not independent.**

Advantages:

1. Efficient – eliminates the need for unique replicates at each time interval
2. Each replicate serves as its own block or control - removes influence of individual variation
3. Allows one to test for time x treatment interaction (often most interesting term)

Both randomized block and repeated measures designs assume **Circularity** = variances of the differences between any 2 treatment levels within a block are the same across blocks (block design) or across time (repeated measures design)

Regression vs. ANOVA designs

Both regression and ANOVA represent powerful experimental designs that are based on the general linear model. But, how do we decide to use one or the other to address our question?

The use of ANOVA is widespread in the sciences and there are more books devoted to the design of experiments using ANOVA than one can count, let alone read. There are many investigators that feel that ANOVA is overused and that researchers are often constrained to think of their question in ANOVA terms. In many instances, a regression design may be more appropriate.

Considering that the independent variable in many ANOVA designs is, in reality, a continuous variable that has been grouped into categories. Rather than using 4-5 fixed levels of X , one can measure X across a range of values and measure the response in Y . This is the typical approach when we conduct an observational study, but is being more widely applied to controlled experimental situations (e.g., reef fish example).

Regression designs can identify thresholds and dynamic (non-linear) relationships between variables that could not be detected with ANOVA. Ideally, we would have some replication at each level of X , but often this isn't possible. We still obtain an unbiased estimate of the regression parameters (slope, intercept), we can construct confidence intervals around those estimates, and draw inferences about the relationship between X and Y .

In addition to the advantages in efficiency of regression designs, the resulting parameter estimates can be used for comparison with theoretical models (which are built mainly as differential equations).

One- and two-sample hypothesis tests

(Chapter 6 in Zar)

Recall our normal distribution with the property of symmetry and our ability to standardize any X_i value using:

$$Z = \frac{X_i - \mu}{\sigma}$$

We refer to Z as a normal deviate and it tells us how many standard deviations (σ) any value X_i is away from the mean. Therefore, if we know the mean and standard deviation of a normal distribution, we can calculate the proportions of that distribution.

Using Table B.2 in Zar, we obtain the following:

68.27% of observations lie within $\mu \pm 1\sigma$

95.44% of observations lie within $\mu \pm 2\sigma$

99.73% of observations lie within $\mu \pm 3\sigma$

50% of observations lie within $\mu \pm 0.67\sigma$

95% of observations lie within $\mu \pm 1.96\sigma$

97.5% of observations lie within $\mu \pm 2.24\sigma$

99% of observations lie within $\mu \pm 2.58\sigma$

99.5% of observations lie within $\mu \pm 2.81\sigma$

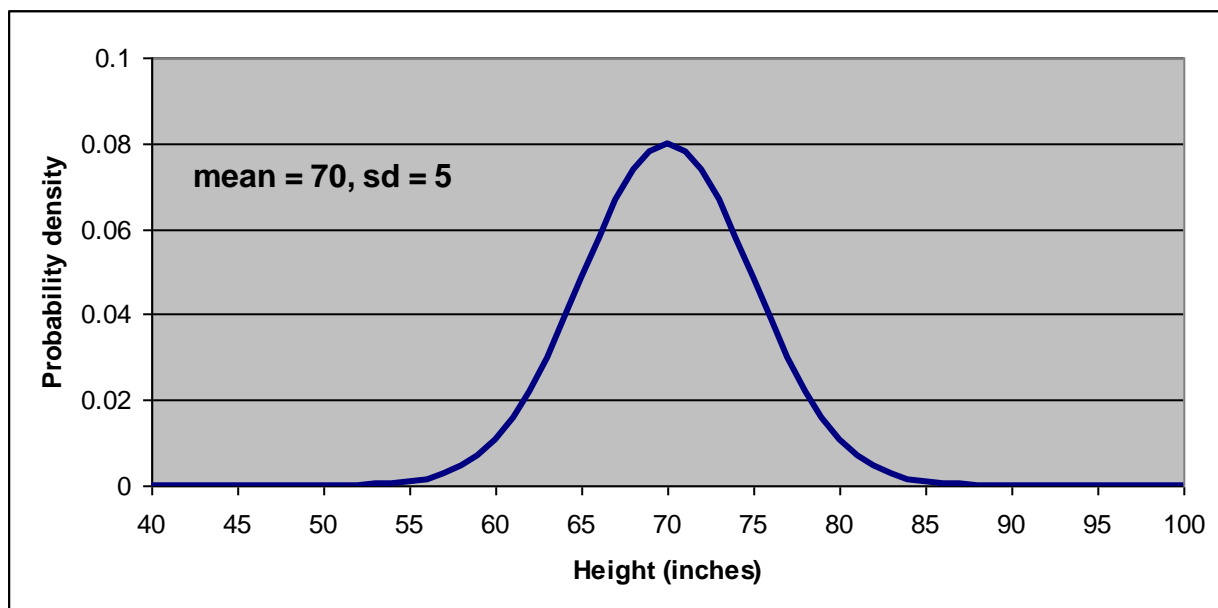
99.9% of observations lie within $\mu \pm 3.29\sigma$

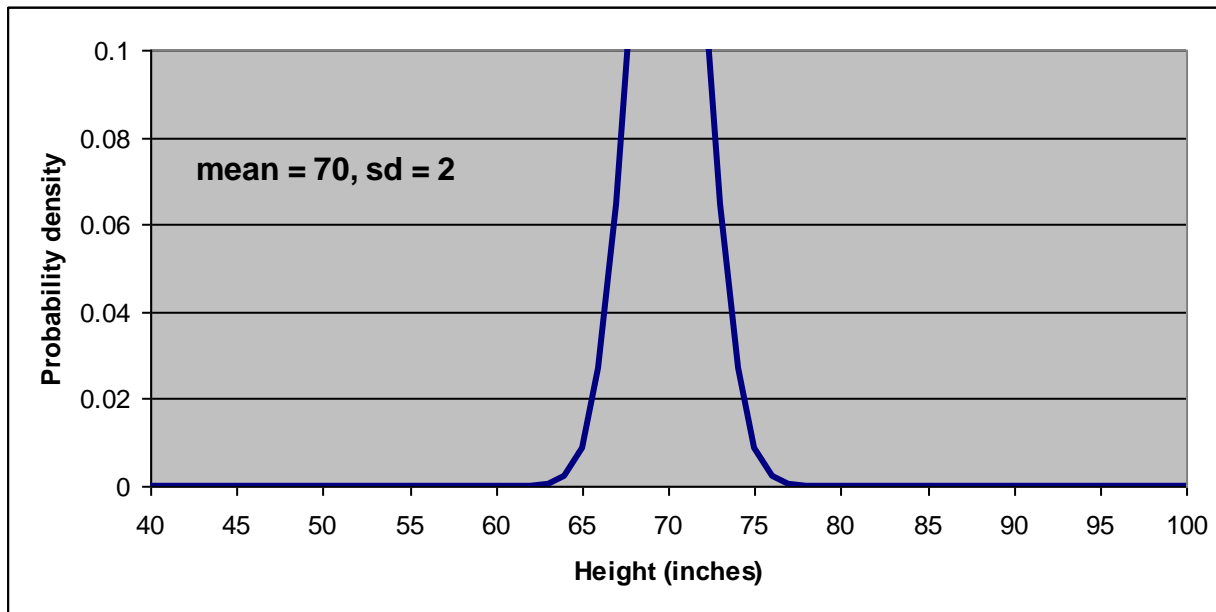
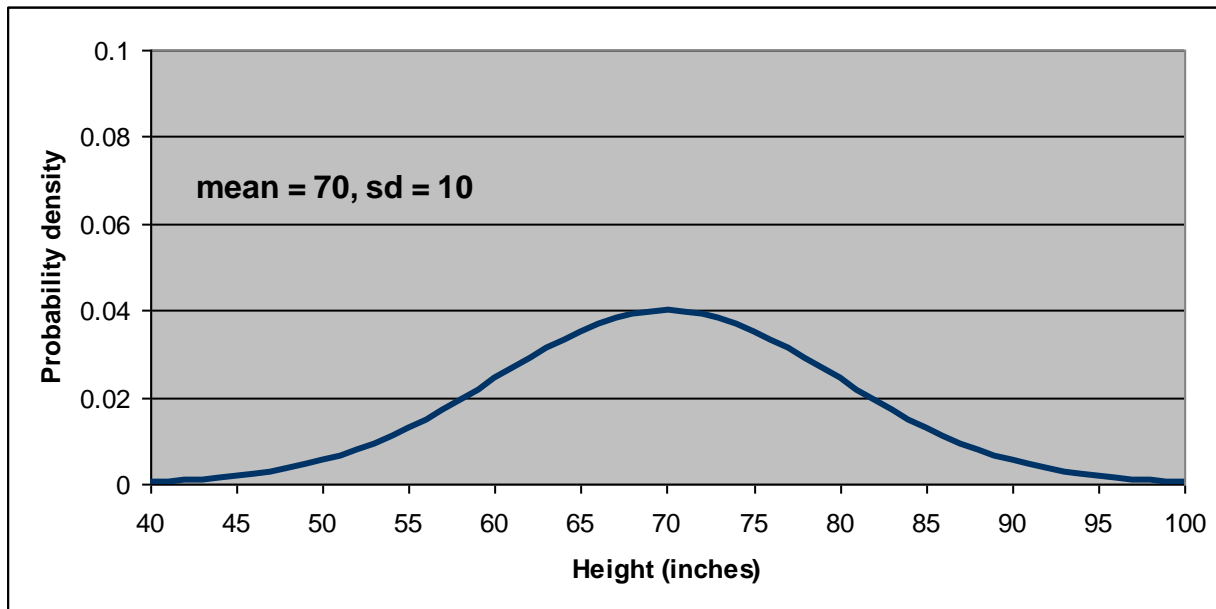
We can use Table B.2 to calculate the proportion of a normal distribution that lies beyond **any** value Z .

For example, we have measured the heights of students in this room and have calculated $\mu = 70$ inches and $\sigma = 5$ inches. The proportion of the distribution greater than or equal to $X_i = 70$ inches would be calculated as $Z = (70 - 70)/5 = 0$. From Table B.2, we see that $P(X_i \geq 70) = P(Z \geq 0) = 0.5000$ or 50%. To determine the proportion greater than or equal to 75 inches, $Z = (75-70)/5 = 1$. Therefore, from Table B.2, $P(Z \geq 1) = 0.1587$, so $P(X_i \geq 75) = 0.1587$ or 15.87%.

To estimate the probability of obtaining a height less than 75 inches, we simply subtract the probability of $X_i \geq 75$ from 1. $P(X_i < 75) = 1 - P(X_i \geq 75) = 1 - 0.1587 = 0.8413$.

***Remember, that the shape of a normal distribution is determined only by the mean and the standard deviation. For any mean (μ), there are an infinite number of normal distributions, each with a different standard deviation (σ), and vice versa.





****Review examples 6.3a and 6.3b in Zar here****

The standard error and hypothesis tests concerning the mean

Remember that the Central Limit Theorem states that the distribution of sample means taken from a non-normal population will tend toward normality as n increases. In addition, the variance of this distribution of means will decrease as n increases.

We end up with a parameter that we call the *variance of the mean*, denoted as:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

We can then obtain another parameter, the *standard deviation of the mean*, denoted as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of a statistic (the mean) is referred to as the **standard error**, so $\sigma_{\bar{x}}$ is often called the standard error of the mean, or just simply the standard error.

Now we can create a normal deviate for our \bar{x} values just like the one we had for our X_i values.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

And we can ask questions about the probability of obtaining a sample mean as large or larger than \bar{x} from a population with a known mean (μ) and standard deviation (σ).

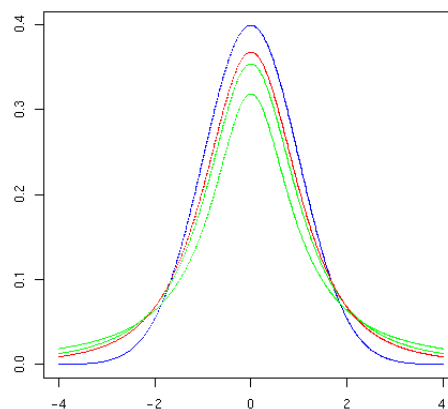
****Review examples 6.4 and 6.6 in Zar here****

The t-distribution (or Student's t-distribution) (chapter 7 in Zar)

In the previous section using normal deviates to test hypotheses about the mean, recall that our calculation of Z required us to know the value of $\sigma_{\bar{x}}$ (the standard error of the mean) which we won't unless we have data from all members of the population of interest. Instead, we calculate $s_{\bar{x}}$ as an estimate of the standard error. When n is very large, we can use $s_{\bar{x}}$ in our calculation of Z . However, in most cases, n isn't large enough to allow us to do this and we must turn instead to a distribution different from the standard normal.

The **t-distribution (or Student's t-distribution)** was developed by Gossett at the turn of the century. The distribution is leptokurtic relative to a normal distribution, but becomes normal as n approaches infinity. In general, the tails are slightly broader and flatter than a normal, and thus contain more of the probability density. This accounts for the extra variation that stems from our estimating the standard error of the mean as well as the mean using our sample data, so the tail probabilities are a little thicker. The test statistic is calculated as:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$



The shape of the t-distribution is affected by the degrees of freedom $\nu = n-1$.

Example:

Suppose I am interested in the mean January temperature in the lower Cape Fear River. I have data on the temperature every two days in January 2008.

Temperature (degrees C) = 3,2,6,2,1,4,4,5,1,5,1,3,4,2,5

My null hypothesis (H_0) is that the mean is five degrees: $H_0: \mu=5$

My alternative hypothesis (H_A) is that the mean is not five: $H_A: \mu \neq 5$

I set my alpha level (α) *a priori* to 0.05 and my $n = 15$.

Steps:

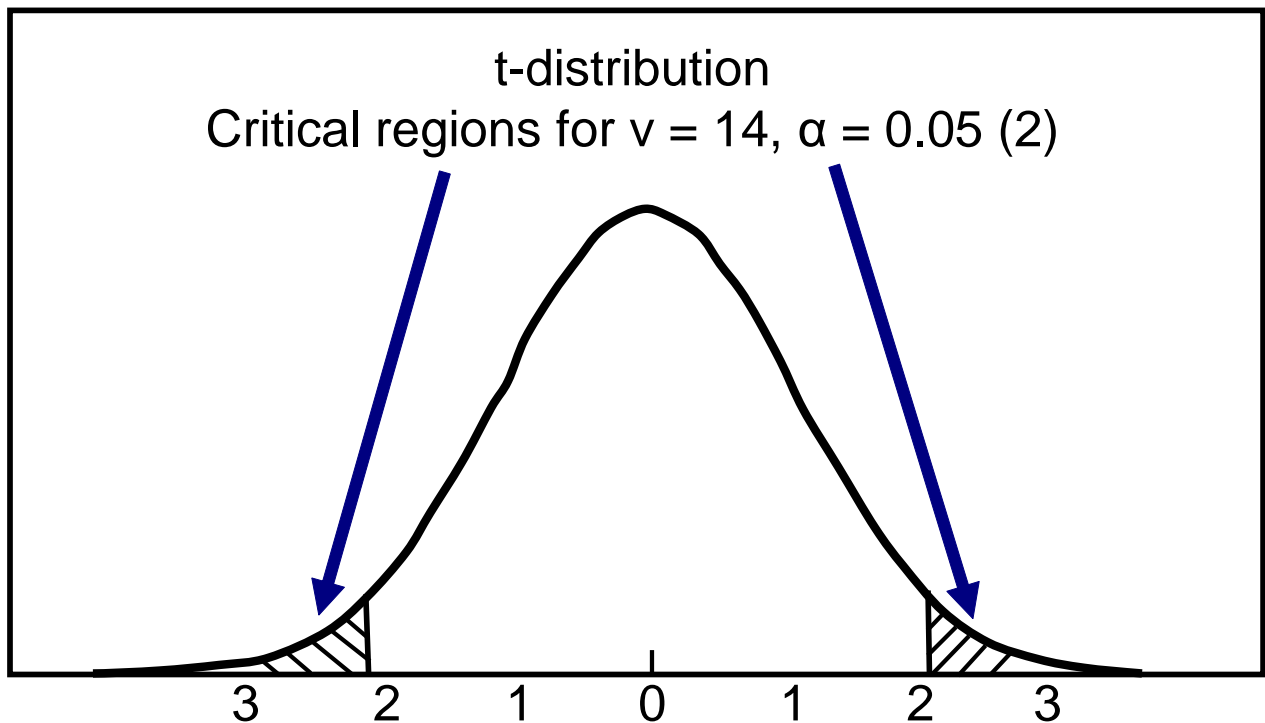
1. Calculate the sample mean: $\bar{x} = 3.20$
2. Calculate the sample variance: $s^2 = 2.74$
3. Calculate the standard error: $s_{\bar{x}} = 0.43$
4. Calculate the test-statistic:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{3.20 - 5}{0.43} = -4.17$$

5. Calculate the degrees of freedom (ν) = $n-1 = 14$.
6. Look up the critical value (Table B.3) = $t_{0.05(2), 14} = 2.145$.

Since $|t| = 4.17 > 2.145$, we would reject H_0 and conclude that the mean January temperature is not five.

Based on the values in Table B.3, our P-value is less than 0.001 ($P < 0.001$).



The critical regions account for 5% (2.5% in each tail) of the probability density of the t-distribution. Thus, any \bar{x} that generates a t-value that lies in either of the shaded areas would be expected to occur less than 5% of the time, if the null hypothesis of $\mu = 5$ were true.

In this case, we have completed a 'two-tailed' test. This means that an extreme value of \bar{x} in either direction will cause us to reject H_0 . We obtain the critical value for t ($t_{\alpha(2), v}$) from Table B.3 and compare it to the **absolute value** of our calculated t-value. For a two-tailed test:

$$\text{If } |t| \geq t_{\alpha(2), v}, \text{ then we reject } H_0$$

****Review examples 7.1 and 7.2 in Zar here****



One-tailed hypotheses

Often, we may only be interested in a mean that is different from zero or a hypothesized value in one direction. For such a 'one-tailed' test, our hypotheses are stated differently:

$$\begin{aligned} H_0: \mu \geq 0 \text{ or } \mu \geq \mu_0 \\ \text{and} \\ H_A: \mu < 0 \text{ or } \mu < \mu_0 \end{aligned}$$

Now, we only examine the critical region on one side of the t-distribution. In general:

$$\begin{aligned} \text{if } t \leq -t_{\alpha(1), v}, \text{ then reject } H_0 \\ \text{or} \\ \text{if } t \geq t_{\alpha(1), v}, \text{ then reject } H_0 \end{aligned}$$

Example:

Suppose I am interested in the mean number of absences by students in my Biostats class. I have data on the number of absences each day since the start of the semester.

Number of absences = 0,1,0,0,2,1,0,2,1,1,0,1,0,4

My null hypothesis (H_0) is that the mean is zero: $H_0: \mu = 0$

My alternative hypothesis (H_A) is that the mean is greater than zero: $H_A: \mu > 0$

I set my alpha level (α) *a priori* to 0.05 and my $n = 14$.

Steps:

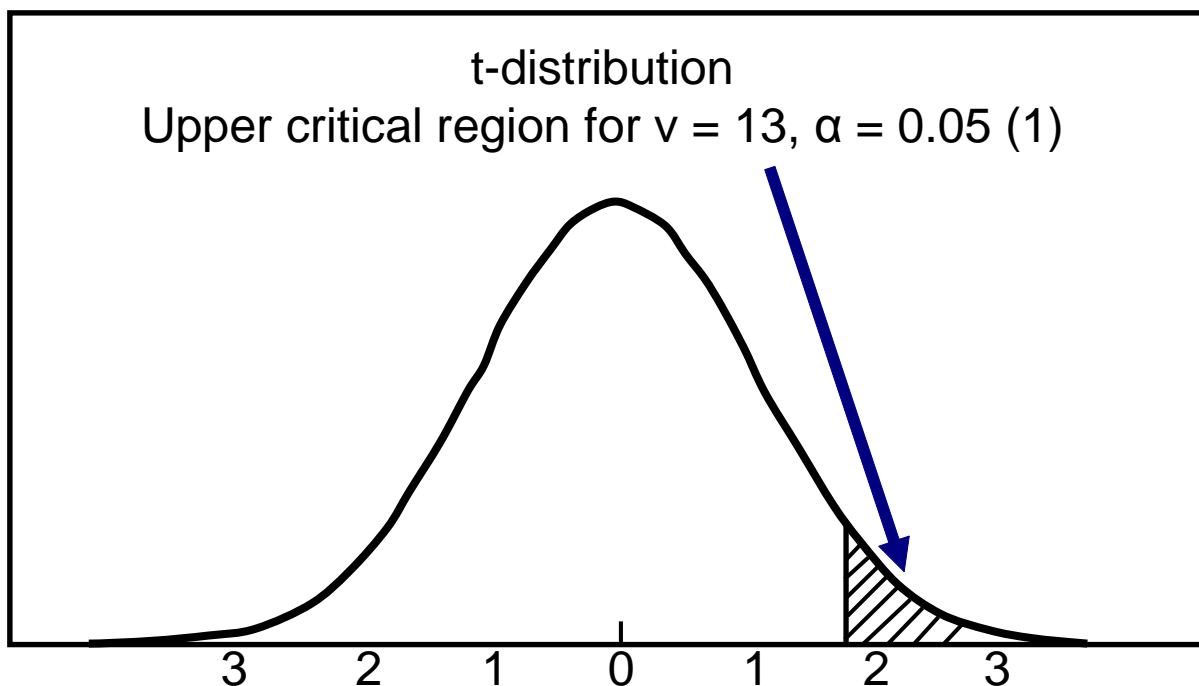
1. Calculate the sample mean: $\bar{x} = 0.93$
2. Calculate the sample variance: $s^2 = 1.30$
3. Calculate the standard error: $s_{\bar{x}} = 0.31$
4. Calculate the test-statistic:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{0.93 - 0}{0.31} = 3.0$$

5. Calculate the degrees of freedom (v) = $n-1 = 13$.
6. Look up the critical value (Table B.3) = $t_{0.05(1), 13} = 1.771$.

Since $t = 3.0 > 1.771$, we would reject H_0 and conclude that the mean number of absences is greater than zero.

Based on the values in Table B.3, our P-value is somewhere between 0.005 and 0.01 ($0.005 < P < 0.01$).



****Review examples 7.3 and 7.4 in Zar here****

Confidence limits for the mean

We can now use the critical values from our t-distribution to obtain the level of precision with which we are estimating the population mean. Our critical values told us that 5% of all possible sample means drawn from a population of mean μ will generate t-values that are either greater than $t_{0.05(2), v}$ or less than $-t_{0.05(2), v}$ (i.e., $|t| > t_{0.05(2), v}$). *This tells us that 95% of t-values will lie between these critical values.*

$$P\left[-t_{0.05(2), v} \leq \frac{\bar{x} - \mu}{s_x^-} \leq t_{0.05(2), v}\right] = 0.95$$

We can rearrange to obtain:

$$P\left[\bar{x} - t_{0.05(2), v} s_x^- \leq \mu \leq \bar{x} + t_{0.05(2), v} s_x^-\right] = 0.95$$

This is called the **confidence interval**. The general equation is:

$$P\left[\bar{x} - t_{\alpha(2), v} s_x^- \leq \mu \leq \bar{x} + t_{\alpha(2), v} s_x^-\right] = 1 - \alpha$$

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals *will include* the unknown population parameter. In other words, the parameter is either in or out of any calculated confidence interval (so we can't say we are 95% confident that this single interval contains μ). We usually calculate confidence intervals so that this percentage is 95%, but we can produce 90%, 99%, 99.9% (or whatever) confidence intervals we wish for the unknown parameter.

The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter (**precision**). A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter. Confidence intervals are more informative than the simple results of hypothesis tests (where we decide "reject H_0 " or "don't reject H_0 ") since they provide a range of plausible values for the unknown parameter.

Example:

Returning to our January water temperature data in the Cape Fear River:

Sample mean: $\bar{x} = 3.20$

Standard error: $s_{\bar{x}} = 0.43$

The degrees of freedom (ν) = $n-1 = 14$

The critical t-value (Table B.3) = $t_{0.05(2), 14} = 2.145$

95% CI = $\bar{x} \pm t_{0.05(2), 14} s_{\bar{x}}$

95% CI = $3.20 \pm (2.145)(0.43)$

95% CI = 3.20 ± 0.92

Lower CI limit = 2.28 Upper CI limit = 4.12

Power and sample size for one-sample t-tests

A common question voiced by researchers in biology is how many samples do I need to test a hypothesis related to the mean? This question can be answered before the samples are collected, but it requires several pieces of information to be specified first. We need to set acceptable levels of error probabilities (both Type I and Type II errors), we need to set our detection level (how small a difference between μ and μ_0 do we want to be able to detect), and we also need to have an idea about our sampling variance.

For a t-test, the formula to calculate sample size needed is:

$$n = \frac{s^2}{\delta^2} (t_{\alpha, v} + t_{\beta(1), v})^2$$

Where s^2 = an estimate of the sampling variance, and δ = the difference between μ and μ_0 that you want to be able to detect. The t_α can be $t_{\alpha(1)}$ or $t_{\alpha(2)}$, depending on whether you are conducting a one-tailed or a two-tailed test.

****Review example 7.7 in Zar here****

For a given sample size (n), we can also determine our minimum detectable difference (δ) using:

$$\delta = \sqrt{\frac{s^2}{n}} (t_{\alpha, v} + t_{\beta(1), v})$$

****Review example 7.8 in Zar here****

Lastly, the power of the test can be estimated for a given sample size (n) and minimum detectable difference (δ) using:

$$t_{\beta(1),v} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,v}$$

****Review example 7.9 in Zar here****

The two main issues when conducting *a priori* power analyses are obtaining an estimate of the variance (s^2) and deciding what the minimum detectable difference (δ) should be. Reasonable variance estimates can generally be obtained through literature searches or the collection of pilot data. Selecting a minimum detectable difference requires some hard thinking!

Two-sample and paired-sample tests (chapters 8 and 9 in Zar)

In the previous section, we were focused on one-sample hypotheses such as whether the mean was or was not a specified value (i.e., $\mu = 0$ or $\mu = 5$). However, we are often more interested in comparing the parameters of two distributions and for this we need to use two-sample tests. We will again make use of the *t-distribution*.

Example:

We divide our Biostats class by gender and measure heights. We have 14 females and 10 males and their heights are listed below.

Female heights in inches (n=14)

62,65,66,68,65,66,65,67,65,64,64,62,65,66

Male heights in inches (n=10)

68,70,76,75,72,73,70,71,69,70

We wish to test the hypothesis that the mean height is different between the gender groups. The H_0 and H_A statements can be written in different ways.

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_A: \mu_1 - \mu_2 \neq 0$$

or

$$H_0: \mu_1 = \mu_2 \quad \text{and} \quad H_A: \mu_1 \neq \mu_2$$

We assume that the two samples were drawn from normal distributions with equal variances.

The t-statistic for a two-sample test is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

The denominator is the standard error of the difference between the sample means. We calculate this quantity using our sample data. It depends on the fact that *the variance of the difference between two variables is equal to the sum of the variances of the two variables.*

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

Since $\sigma^2_{\bar{x}} = \sigma^2/n$, then

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

And since we assume equal variances for a two-sample test,
 $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Now, we need an estimate of σ^2 . We have two estimates of σ^2 from our calculated sample variances (s_1^2 and s_2^2). Since they are both assumed to estimate σ^2 , we calculate a *pooled variance* s_p^2 using:

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2}$$

Then we can calculate the standard error of the difference between the sample means as:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Now, our calculation of the t-statistic becomes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Returning to our example,

Female heights

$$n = 14$$

$$v = 13$$

$$\bar{x} = 65 \text{ inches}$$

$$SS = 36 \text{ inches}^2$$

Male heights

$$n = 10$$

$$v = 9$$

$$\bar{x} = 71.4 \text{ inches}$$

$$SS = 60.4 \text{ inches}^2$$

$$s_p^2 = \frac{36 + 60.4}{13 + 9} = \frac{96.4}{22} = 4.38 \text{ inches}^2$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{4.38}{14} + \frac{4.38}{10}} = \sqrt{0.313 + 0.438} = 0.867$$

$$t = \frac{65 - 71.4}{0.867} = \frac{-6.4}{0.867} = -7.38$$

$$t_{0.05(2),v} = t_{0.05(2), 22} = 2.074$$

Since $|t| > 2.074$, we would reject H_0 ; $P < 0.001$

We would conclude that the mean heights differed between the gender groups, with males being taller than females.

***Review examples 8.1 and 8.2 in Zar here



Assumptions of the two-sample t-test

The two-sample t-test assumes that both samples were drawn from normal populations with equal variances. However, unless departures from these assumptions are severe, the t-test is quite robust. One-tailed tests are affected to a greater degree than two-tailed tests by departures from normality (i.e., skewed distributions). In general, the probability of committing a type I error will be inflated by non-normality and non-equality (heterogeneity) of variances, but the amount of increased error probability is lessened greatly by larger sample sizes. See Table 8.1 in Zar for type I error probabilities under assumption violations. There is also a modified t-test (*Welch's approximate t*) that is an improved test when the assumptions of normality and variance homogeneity have been violated. One can also use a non-parametric approach, which we will discuss shortly. In most cases, however, the researcher will be justified in simply employing the standard t-test.

Confidence intervals for two-sample tests

Since we assume that the variances are equal when we test for differences between two means, the confidence interval for either of the means is calculated using the pooled variance (s_p^2) to estimate the standard error. For either μ_1 or μ_2 , the confidence interval is calculated as:

$$\bar{X}_i \pm t_{\alpha(2),v} \sqrt{\frac{s_p^2}{n_i}}$$

In our height example:

$$\sqrt{\frac{s_p^2}{n_{females}}} = \sqrt{\frac{4.38}{14}} = 0.559$$

Since $t_{\alpha(2),v} = t_{0.05(2), 22} = 2.074$, the 95% CI would be $65 \pm (2.074) * (0.559 \text{ inches}) = 65 \pm 1.16 \text{ inches}$

For males, the 95% CI would be $71.4 \pm (2.074) * (0.662) = 71.4 \pm 1.37 \text{ inches}$

A confidence interval for the difference between two means ($\bar{x}_1 - \bar{x}_2$) can also be calculated. The equation is:

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha(2),v} s_{\bar{X}_1 - \bar{X}_2}$$

For our example this would be:

$$71.4 - 65 \pm (2.074) * (0.867) = 6.4 \pm 1.80 \text{ inches}$$

Power and sample size for two-sample tests

Similar to the procedures we used for one-sample tests, we can estimate the required sample size, minimum detectable difference, and/or power of our two-sample test. To estimate the sample size (n) required to detect a specific difference between two means we can use:

$$n \geq \frac{2s_p^2}{\delta^2} (t_{\alpha,v} + t_{\beta(1),v})^2$$

Here, S_p^2 is our pooled variance estimate and δ is the minimum difference we want to be able to detect. The $t_{\alpha,v}$ value can be for either a one-tailed (1) or a two-tailed (2) test. As we noted before, the required sample size (n) will be dependent on the difference we wish to be able to detect, our variance, the α -level, and the power ($1-\beta$) we wish to achieve.

*Note that if we are constrained to have *unequal sample sizes*, we will generally need a higher overall n to detect a specified difference with the same type I and type II error probabilities.

***Review example 8.4 in Zar here

For a given sample size (n), α -level, and power ($1-\beta$), we can estimate the minimum detectable difference (δ) using:

$$\delta = \sqrt{\frac{2s_p^2}{n} (t_{\alpha,v} + t_{\beta(1),v})^2}$$

***Review example 8.5 in Zar here

The power ($1-\beta$) of the test can be estimated for a given sample size (n), α -level, and minimum detectable difference (δ) using:

$$t_{\beta(1),v} = \frac{\delta}{\sqrt{\frac{2s_p^2}{n}}} - t_{\alpha,v}$$

***Review example 8.6 in Zar here

When estimating minimum detectable difference (δ) and power ($1-\beta$) for two-sample tests with unequal samples sizes ($n_1 \neq n_2$), the single n that is called for in the formulae above can be calculated as:

$$n = \frac{2n_1n_2}{n_1 + n_2}$$

Although Zar outlines the estimation of *a posteriori* (= after the test) power in example 8.7, this is not recommended. Obviously, if you failed to reject H_0 then you didn't have enough power to detect a difference. We will discuss this in more detail when we cover power analysis for ANOVA.

Testing for differences between two variances

In addition to comparing means from two samples, we can also ask questions about the variances. Hypotheses about differences

between variances can be addressed using a *variance ratio test*. The test-statistic is F and is calculated using:

$$F = \frac{s_1^2}{s_2^2}$$

The larger of the two variances is positioned in the numerator, so F can range from 1.0 to ∞ . If the calculated ratio for any two variances deviates greatly from 1.0, then we reject the null hypothesis $H_0: s_1^2 = s_2^2$ and conclude that they are different. We use table B.4 in Zar to obtain critical values for the F-distribution using the α -level and the degrees of freedom (n-1) for the numerator and denominator (**in that order**).

***Review example 8.8 in Zar here

Nonparametric tests for two samples

If a researcher believes that the assumptions of normality and equality of variance have been severely violated they may choose to use a *nonparametric* approach to test their hypothesis of interest. These approaches don't require us to estimate the mean and variance, and don't assume normality. Keep in mind, that whenever a parametric test can be used, it will be more powerful (lower probability of making a type II error) compared to a nonparametric test.

A common nonparametric test for comparing two means is the **Mann-Whitney test**. In this test, the actual measurements are

not used, but instead they are converted to ranks (either from highest to lowest or vice versa). We rank all the data together, not separately within the groups we wish to compare. The test statistic (U) is:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

where n_1 and n_2 are the sample sizes in each group and R_1 is the sum of the ranks for group 1. Alternatively, we can calculate U' :

$$U' = n_2 n_1 + \frac{n_2(n_2 + 1)}{2} - R_2$$

If we have already calculated U or U' , the other can be calculated more simply using:

$$U' = n_1 n_2 - U$$

or

$$U = n_1 n_2 - U'$$

For a two-tailed test, both U and U' are computed and we use whichever is larger to compare to the critical value $U_{\alpha(2)n_1, n_2}$ in table B.11

Example:

We are interested in comparing the ground speeds of lizards (*Anolis* spp.) and skinks (*Eumeces* spp.) as we chase them during capture attempts. We measure the speeds (cm/s) of 6 of each:

Lizard: 3.1, 4.2, 3.7, 3.6, 4.0, 3.3

Skink: 4.3, 4.8, 4.9, 4.1, 3.9, 5.0

If we rank these from fastest to slowest we have:

- | | |
|-----------------|------------------|
| 1. 5.0 (skink) | 7. 4.0 (lizard) |
| 2. 4.9 (skink) | 8. 3.9 (skink) |
| 3. 4.8 (skink) | 9. 3.7 (lizard) |
| 4. 4.3 (skink) | 10. 3.6 (lizard) |
| 5. 4.2 (lizard) | 11. 3.3 (lizard) |
| 6. 4.1 (skink) | 12. 3.1 (lizard) |

We can sum the ranks of the skinks = $(1+2+3+4+6+8) = 24$, and calculate our test statistic (U):

$$U = 6 * 6 + \frac{6(6+1)}{2} - 24 = 33$$

Then $U' = 6*6 - 33 = 3$. Since $U > U'$, we use U to test our hypothesis. From table B.11, we obtain $U_{0.05(2)6,6} = 31$. Since our calculated $U > U_{critical}$, we would reject the null hypothesis that lizards and skinks demonstrate the same ground speed when being chased, and conclude that skinks should be harder to catch than lizards.

*Note that when two observations would receive the same rank (i.e., they are tied), each observation is assigned the mean of the ranks that would have been assigned if they weren't tied (e.g., if the 3rd and 4th observation are tied, they each are assigned a rank of 3.5).

***Review example 8.13 in Zar here

Paired-sample tests

There are occasions when the observations from the two samples we wish to compare are *not independent*. This precludes us from using the two-sample testing procedures described above and requires us to use **paired-sample tests** instead.

Paired-sample tests evaluate hypotheses using only the differences between paired observations (*not the actual values themselves*). Therefore, instead of the null hypothesis being expressed as:

$$H_0: \mu_1 - \mu_2 = 0$$

it is expressed as:

$$H_0: \mu_d = 0$$

where $\mu_d = \mu_1 - \mu_2$. And the test statistic is calculated as:

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

where \bar{d} = the mean of the differences between paired observations and $s_{\bar{d}}$ = the standard error of the mean differences between paired observations. It is similar to a one-sample t-test, with n = number of observation pairs and $v = n-1$.

The paired t-test necessitates that each observation in one group is correlated with only one other observation from the second group. The observations for each sample do not need to be normally distributed, nor do the samples need to display homogeneous variances. However, it is assumed that the

differences (d 's) are drawn from a normally distributed population.

Example:

Suppose we were interested in the effect of sediment grain size on the burrowing depth of an intertidal shrimp. Previous data suggests that individual variation in burrowing depth is large, such that detecting small differences due to sediment grain size would be difficult (*low signal to noise ratio*) if we used a standard two-sample design. If, however, we set up experimental tanks that had coarse sediment on one side and fine sediment on the other, we could measure the burrowing depth of the *same individual* shrimp in both sediment types and analyze the **paired differences** using a paired-sample test. This will be more powerful than a standard two-sample t-test because we have eliminated the *noise* due to individual shrimp variation.

$$H_0: \mu_d = 0$$

$$H_A: \mu_d \neq 0$$

$$n = 10 \text{ paired observations, thus } v = n - 1 = 9$$

Observation (ind. shrimp)	coarse	fine	diff.
1	2.4cm	3.6cm	1.2cm
2	1.4cm	1.9cm	0.5cm
3	3.4cm	4.0cm	0.6cm
4	2.2cm	2.5cm	0.3cm
5	5.1cm	5.4cm	0.3cm
6	1.8cm	2.8cm	1.0cm
7	3.0cm	4.7cm	1.7cm
8	1.3cm	2.0cm	0.7cm
9	4.2cm	4.8cm	0.6cm
10	4.9cm	5.4cm	0.5cm

We first calculate the mean of the paired differences $\bar{d} = 7.4\text{cm}/10 = 0.74\text{cm}$. We then calculate the variance $s_d^2 = 0.194\text{cm}^2$, the standard deviation $s_d = 0.44\text{cm}$, and lastly, the standard error of the mean of the differences $s_{\bar{d}} = 0.139$. The test statistic is calculated as:

$$t = \frac{0.74}{0.139} = 5.32$$

We compare this t-value to our critical value $t_{0.05(2), 9} = 2.262$ and we would then reject H_0 and conclude that sediment grain size affects burrowing depth. Our P-value < 0.001

***Review examples 9.1 and 9.2 in Zar here

Confidence intervals and power for paired-sample tests

Since we have estimated the mean of a sample of differences, \bar{d} , the confidence intervals for a paired-sample test are estimated similarly to those for a one-sample t-test of the mean. The interval is calculated using:

$$\bar{d} \pm t_{\alpha(2), v} s_{\bar{d}}$$

For our example, the 95% CI would be $0.74\text{ cm} \pm (2.262) * (0.139) = 0.74\text{ cm} \pm 0.31\text{ cm}$

To calculate required sample size (n) and power ($1 - \beta$), we also can use the approaches for a one-sample t-test. We just need to replace \bar{x} with \bar{d} and s^2 with s_d^2 .

Paired-sample testing by ranks

We introduced the Mann-Whitney rank test earlier when assumptions of the two-sample t-test were violated to such a large degree to prevent its use. When performing paired-sample tests where the sample of paired differences severely violates the assumption of normality, we can use a similar test known as the **Wilcoxon paired-sample test** (*aka* Wilcoxon rank sum test or Wilcoxon signed rank test).

The procedure ranks the absolute values of the differences then assigns the sign (+ or -) of each difference to the ranks. All of the positive ranks are summed (T_+) and all of the negative ranks are summed (T_-). If we are conducting a two-tailed test, then if either T_+ or T_- is less than the critical value from Table B.12, we reject H_0 . The procedure for tied ranks is the same as that outlined for the Mann-Whitney test.

In our example of burrowing depths of shrimp in different sediment grain sizes, all of our differences were in the same direction (either all positive or negative). Our T_+ total would be 55 (the sum of all ten ranks) and our T_- total would be 0 (or vice versa). Since 0 is far lower than the critical value $T_{0.05(2), 10} = 8$, we would reject the null hypothesis of no difference in burrowing depth.

***Review example 9.3 in Zar here



Evaluating multi-sample hypotheses with ANOVA

Up to this point, we've concerned ourselves with only one- and two-sample hypothesis tests. However, as biologists, we are often interested in questions that involve more than two samples. It would be incorrect to apply a large number of separate two-sample t-tests to a multi-sample problem due to the inflation of the probability of making a Type I error (incorrectly rejecting a true null). As the number of tests increases, the α -level increases as $1 - (1-\alpha)^n$ (see Table 10.1 in Zar). This is because our multiple t-tests used on the same data set do not represent truly *independent* tests of our hypotheses.

For example, suppose we want to know if there are differences in bacteria present in five different types of frozen vegetables. We could go to the store and randomly select 5 boxes of each type of vegetable and measure the bacteria levels present.

Peas	Carrots	Broccoli	Beans	Spinach
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4	\bar{X}_5

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

H_A : At least one of the vegetables is different.

If we were forced to use our two-sample procedures, we would examine the samples two at a time and compare the means. This would require ten separate tests.

- | | | |
|-----------------------------|-----------------------------|-------------------------------|
| (1) $H_{01}: \mu_1 = \mu_2$ | (5) $H_{05}: \mu_2 = \mu_3$ | (8) $H_{08}: \mu_3 = \mu_4$ |
| (2) $H_{02}: \mu_1 = \mu_3$ | (6) $H_{06}: \mu_2 = \mu_4$ | (9) $H_{09}: \mu_3 = \mu_5$ |
| (3) $H_{03}: \mu_1 = \mu_4$ | (7) $H_{07}: \mu_2 = \mu_5$ | (10) $H_{010}: \mu_4 = \mu_5$ |
| (4) $H_{04}: \mu_1 = \mu_5$ | | |

If we didn't reject any of these hypotheses then we would not reject $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

Besides being inappropriate, this approach would be very inefficient. Instead, we should design our experiment or study to enable us to an appropriate multi-sample test.

The **Analysis of Variance (ANOVA)** is a multi-sample test that allows us to test the null hypothesis:

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

where k is the number of sample or experimental groups we wish to compare. For our vegetable example, we can now use a single test to evaluate the likelihood that the 5 vegetables were drawn from a population having the same mean bacteria level.

Note: Just like our t-tests, we don't really think the 5 means are precisely equal (they are certainly not). However, if an ANOVA fails to reject H_0 , then we may conclude that the differences are relatively small and don't warrant any further study (i.e., the effect size isn't big enough to be of interest to us). If we reject H_0 , then we will look further to see why the hypothesis was rejected. To study the means, and evaluate whether meaningful differences exist, it is necessary to "*analyze the variance*".

Keep in mind that despite the fact that our focus will be on examining the variance, the goal of ANOVA is still simply to compare means among randomly sampled groups. The ANOVA model is a special case of a more *generalized linear model*.

Let's start with a simple example: We have measured the diving depth of three groups of turtles exposed to different stress levels (our single treatment) before hand. The data are as follows:

Treatment level		
1	2	3
75	25	100
80	75	80
75	25	100
50	75	40

H_0 : There is no difference in diving depth among the three treatments ($H_0: \mu_1 = \mu_2 = \mu_3$)

H_A : The diving depths are not all equal

Notation

Each observation (X) is denoted with a number for the treatment level (i) and a number for the observation (j) within the treatment level. Different factors or treatments are denoted with capital letters (A, B, C,...). In our case, we have a single treatment, which would receive the symbol A. When we refer to the summation of items across levels of the treatment, we sum from i to k . So our summation totals look like:

$$\sum_{j=1}^n X_{ij} \quad \text{or} \quad \sum_{i=1}^k \sum_{j=1}^n X_{ij}$$

For our example of turtle diving depths we have:

Treatment level		
1	2	3
X_{11}	X_{21}	X_{31}
X_{12}	X_{22}	X_{32}
X_{13}	X_{23}	X_{33}
X_{14}	X_{24}	X_{34}

$$\sum_{j=1}^4 X_{ij} \quad \text{and} \quad \sum_{i=1}^3 \sum_{j=1}^4 X_{ij}$$

Partitioning the Sums of Squares

Recall that we have been computing sums of squares for our variance estimates all along. Now, we will separate the variance into different components. That is, we will partition the sums of squares into variation due to randomness (simply error) and variation due to our treatments.

Our ANOVA model depends on the same assumption of homogeneity of variance that our two-sample t-test did. We calculated a pooled variance by pooling the sums of squares and dividing by the pooled degrees of freedom. We will do the same for our ANOVA model. We start by pooling the sums of squares within each of our groups across all the groups and we obtain:

$$\text{within-groups SS} = \sum_{i=1}^k \sum_{j=1}^n \left(X_{ij} - \bar{X}_i \right)^2$$

Then we pool the degrees of freedom to obtain:

$$\text{within-groups DF} = \sum_{i=1}^k (n_i - 1) = N - k$$

In addition to within-groups SS and within-groups DF, these two quantities are also often referred to as the *error sums of squares* and the *error degrees of freedom*, as well as the *residual sums of squares* and the *residual degrees of freedom*. When we divide the within-groups SS by the within-groups DF, we obtain the best estimate of the variance that is common to all k groups.

Next, we calculate the amount of variation among each of our k groups. This is represented by the differences among the means of each of the groups. This source of variation is calculated as:

$$\text{among-groups SS} = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

where the \bar{X} with no subscript is equal to the **Grand Mean**, which is calculated as:

$$\frac{\sum_{i=1}^k \sum_{j=1}^n X_{ij}}{N}$$

We also need our among-groups DF, which is simply $k-1$. The among-groups SS and DF are often referred to simply as *groups sums of squares* and *groups degrees of freedom*.

We also calculate the total sums of squares of the data, which is the sum of the squared deviations of each observation (X_i) from the grand mean (\bar{X}). This is done using:

$$\text{total SS} = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2$$

And our total degrees of freedom (total DF) = $N - 1$.

We have now partitioned the total variance into two components. We can summarize this process by noting that each deviation of an observation from the grand mean represents the total of the deviation of that observation from its group mean plus the deviation of that group mean from the grand mean.

$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})$$

Both the sums of squares and the degrees of freedom are additive, so

$$\text{total SS} = \text{among-groups SS} + \text{within-groups SS}$$

and

$$\text{total DF} = \text{among-groups DF} + \text{within-groups DF}$$

You should practice calculating the different sums of squares by hand (*you'll have to for some upcoming assignments*). There are several machine formulas on p. 182 in Zar to speed things up.

Let's show that this principle of *additivity* holds for our turtle example:

Obs	X_{ij}	\bar{X} (grand)	$(X_{ij} - \bar{X})^2$	\bar{X}_i (group)	$(X_{ij} - \bar{X}_i)^2$	$(\bar{X}_i - \bar{X})^2$
X_{11}	75	66.67	69.44	70	25	11.11
X_{12}	80	66.67	177.78	70	100	11.11
X_{13}	75	66.67	69.44	70	25	11.11
X_{14}	50	66.67	277.78	70	400	11.11
X_{21}	25	66.67	1736.11	50	625	277.78
X_{22}	75	66.67	69.44	50	625	277.78
X_{23}	25	66.67	1736.11	50	625	277.78
X_{24}	75	66.67	69.44	50	625	277.78
X_{31}	100	66.67	1111.11	80	400	177.78
X_{32}	80	66.67	177.78	80	0	177.78
X_{33}	100	66.67	1111.11	80	400	177.78
X_{34}	40	66.67	711.11	80	1600	177.78
			7316.67		5450	1866.67

We have,

$$\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2 = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

total SS = within-groups SS + among-groups SS

7316.67 = 5450 + 1866.67

Hence, our principle of additivity holds.

The division of the sums of squared deviations (SS) by their respective degrees of freedom (DF) generates a variance estimate. In ANOVA, each of these variances is generally referred to as a **mean square (MS)**, which is short for mean squared deviation from the mean. Therefore we can calculate the within-groups MS and the among-groups MS.

$$\text{within - groups MS} = \frac{\text{within - groups SS}}{\text{within - groups DF}}$$

and

$$\text{among - groups MS} = \frac{\text{among - groups SS}}{\text{among - groups DF}}$$

Sometimes you will see the within-groups MS referred to as **MSE** (*mean square error*) or **residual MS**. All of these basic calculations are usually summarized in a particular order using a table (*the ANOVA table*).

Typical ANOVA table for a single factor design

Source of variation	Sum of squares	df	Mean square (MS)	F
among	$\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$	k-1	$\frac{\text{among SS}}{k-1}$	$\frac{\text{among MS}}{\text{within MS}}$
within	$\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$	N-k	$\frac{\text{within SS}}{N-k}$	
total	$\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2$	N-1		

***Note: we could calculate a MS for the total SS, but there is no need.

For our turtle example, the table would be generated as:

Source of variation	Sum of squares	df	Mean square (MS)	F
among	1866.67	2	933.33	1.54
within	5450.00	9	605.56	
total	7316.67	11		

To test the null hypothesis ($H_0: \mu_1 = \mu_2 = \mu_3$). We compare the variances (MS values) among and within our groups. If the groups were drawn from a population with a single mean, then the among-groups MS and the within-groups MS should each be an unbiased estimate of the population variance (σ^2). However, if the k group means are not the same, then the among-groups MS will be greater than the within-groups MS. The test for equality of means is then a one-tailed variance ratio test. The among-groups MS is placed in the numerator and we calculate F.

$$F = \frac{\text{among - groups MS}}{\text{within - groups MS}}$$

The critical value is $F_{\alpha(1), (k-1), (N-k)}$. If the calculated F is $\geq F_{\text{critical}}$, then we reject H_0 . Keep in mind that the order of the degrees of freedom matters (numerator df 1st, denominator df 2nd).

In our turtle diving depth example, the calculated $F = 1.54$ and the critical value, $F_{0.05(1), 2, 9} = 4.26$. We would fail to reject H_0 and our P-value > 0.25 . It appears that our stress treatment levels did not significantly affect turtle diving depth.

***Review example 10.1 in Zar here

The assumptions of ANOVA

1. The samples are random and independent
-as always, this assumption forms the basis for all statistical tests
2. The variances are homogeneous among groups
-similar to our two-sample t-test, we assume the groups were drawn from a population with the same variance, and each treatment group contributes equally to our within-groups sums of squares
3. The residuals are normally distributed
-the deviations follow a normal distribution with mean = 0. As long as n is fairly large, the Central Limit Theorem minimizes the problems caused by violation of this assumption
4. The samples are classified correctly
-each of the samples that are assigned to a particular treatment or treatment level are treated identically

Just like our t-test, the ANOVA model is robust with respect to violations of the assumptions of homogeneity of variances and normal error distributions. So long as our n is relatively large and we have close to equal n 's among our groups, we need not worry except for the most severe departures. Similar to normality tests, there are several tests to ensure homogeneity of variances (i.e., F-max test, Bartlett's test, Cochran's test, Levene's test). None are very powerful, and we can often homogenize our variances through data transformations (we will cover those soon). If we fear we have severe departures from our assumptions, we can employ a non-parametric ANOVA.



Power analysis and ANOVA

We introduced the concept of power previously in our discussions of t-tests, and it is important to revisit again now that we have started examining ANOVA designs. Zar includes several formulae for calculating required sample size, effect size, and power presuming we have an idea about the size of our expected within sample variance (s^2). For ANOVA, the formulae each involve calculation of the quantity phi (ϕ), which we also briefly introduced earlier. When our H_0 is false, our variance ratio follows what we call a *noncentral F-distribution*, and phi (ϕ) simply represents the amount of noncentrality in the distribution (which depends mostly on the effect size).

The power of an ANOVA is estimated using:

$$\phi = \sqrt{\frac{n\delta^2}{2ks^2}}$$

where k = number of groups, δ = minimum detectable difference, s^2 is our estimate of the variance, and n is the sample size per group. After we calculate the value of phi (ϕ), we use Appendix Fig. B.1 to determine the power.

***Review example 10.4 in Zar here

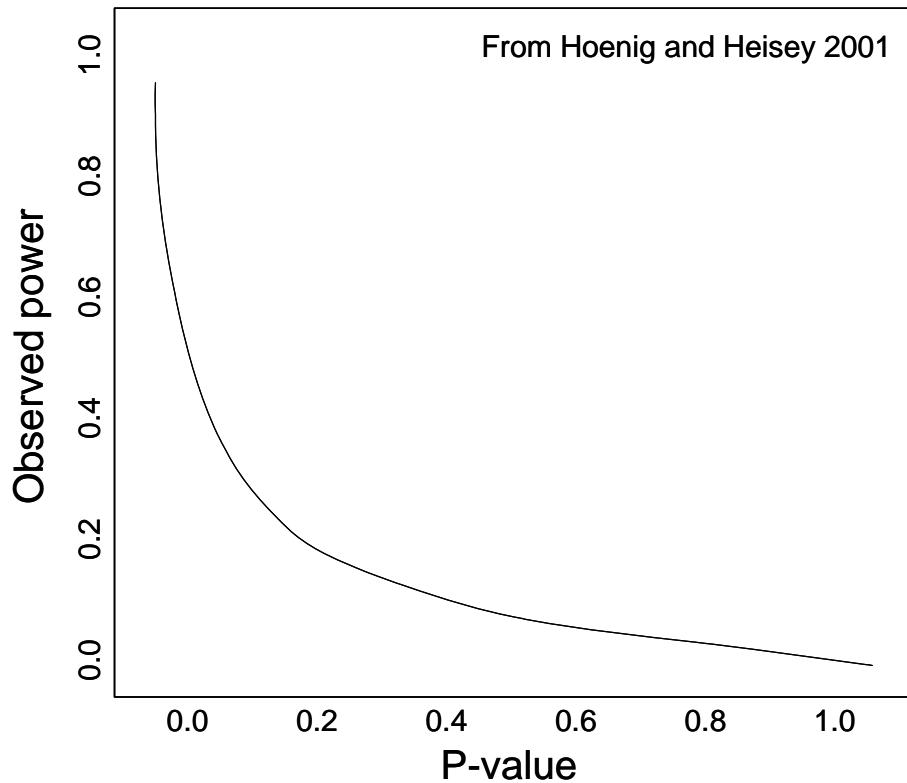
We can use the same formula above to estimate the required sample size for a given level of power and effect size through an iterative process similar to the one we used for the t-test. We can also rearrange the equation to estimate the minimum

detectable difference of an experiment for a given level of power and sample size:

$$\delta = \sqrt{\frac{2ks^2\phi^2}{n}}$$

***Review examples 10.6 and 10.7 in Zar here

Zar provides an example (10.5) to estimate the power of an ANOVA after it has been performed. *This represents flawed logic and should not be followed.* As we have spoken about, many journals began a push for researchers to provide information about the power of their tests after failing to reject null hypotheses. This was, in essence, an inappropriate way for authors to defend their experimental design or to cling to some notion that a pattern existed where it did not. Phrases such as *"we did not reject the null hypothesis, however a posteriori power analysis revealed that the test had low power to detect the observed difference. Therefore, a different design that lowered the variance or an increase in sample size may have succeeded in detecting the difference."* This represents incorrect thinking. If you didn't reject the null, clearly your test wasn't powerful enough. The a posteriori power analysis doesn't add anything new to our interpretation of the results. All power analyses should be performed *a priori* as a guide to setting up the experimental design.

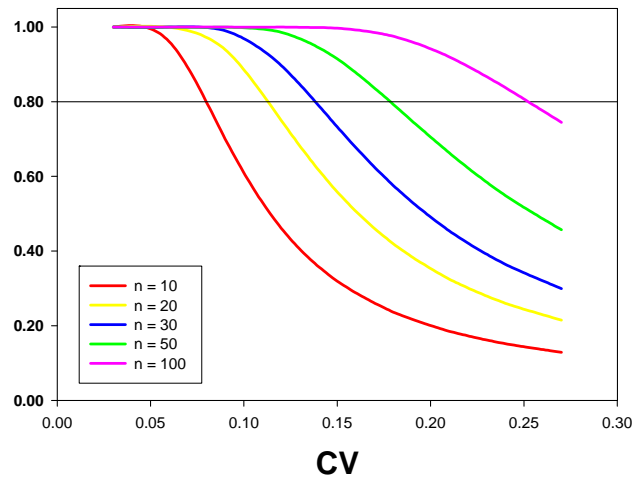
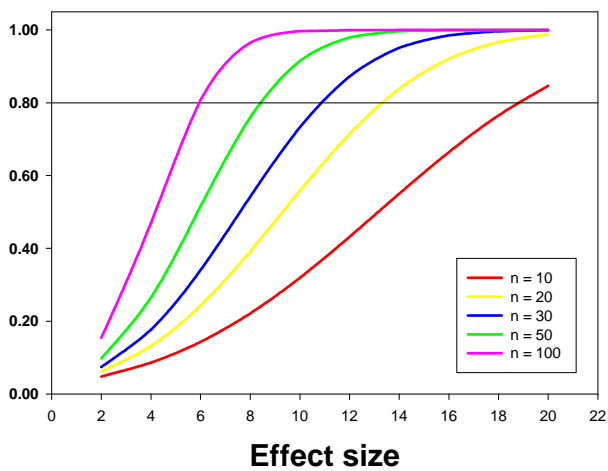
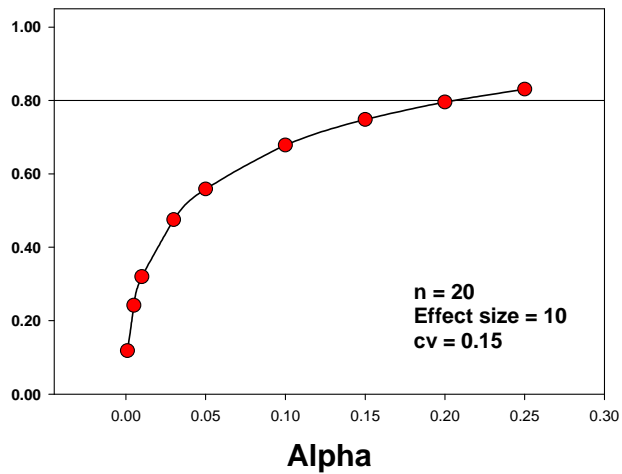
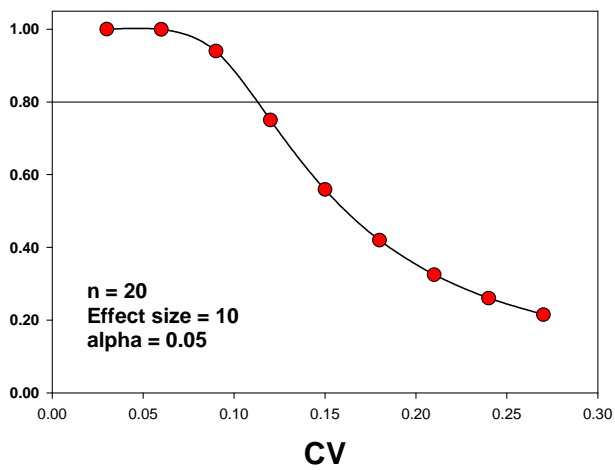
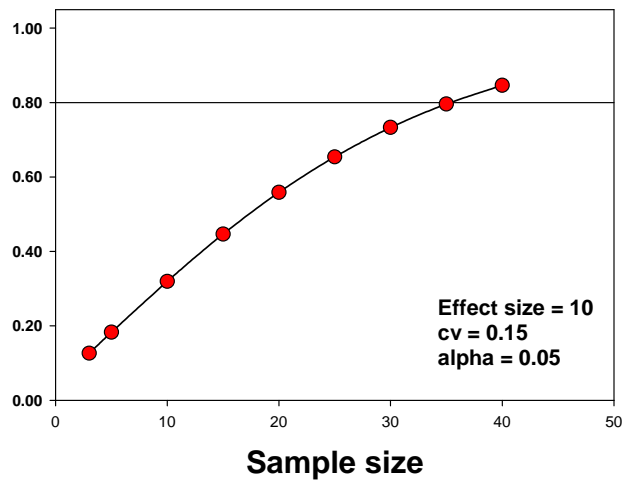
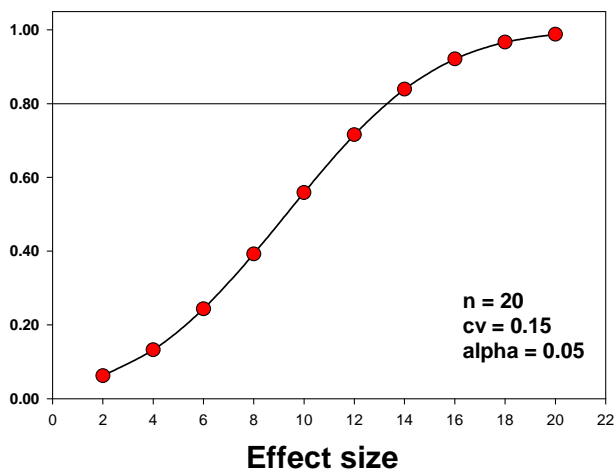


This figure shows the relationship between observed power and the P-value of a performed test. Notice that non-significant P-values always correspond to low power. This confirms that computing the power after the test tells us nothing new about the results of our test that we didn't already know from our P-value.

It is important to remember the major contributors to power when you are making decisions about your experiment:

1. sample size (n)
2. effect size (δ)
3. variance (s^2)
4. alpha (α) level
5. details of the design (one or two tailed hypothesis, equal variances)

Each of these will affect the power of your test



Recent literature on experimental power and sample size:

Gerow, K.G. 2007. Power and sample size estimation techniques for fisheries management: assessment and new computational tool. *North American Journal of Fisheries Management* **27**: 397-404.

Hayes, J.P., and Steidl, R.J. 1997. Statistical power analysis and amphibian population trends. *Conservation Biology* **11**(1): 273-275.

*Hoenig, J.M., and Heisey, D.M. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistical Association Journal* **55**: 19-23.

*Lenth, R. 2001. Some practical guidelines for effective sample size determination. *American Statistical Association Journal* **55**(3): 187-193.

Nakagawa, S., and Foster, T.M. 2004. The case against retrospective statistical power analyses with an introduction to power analysis. *Acta ethol* **7**: 103-108.

Reed, J.M., and Blaustein, A.R. 1997. Biologically significant population declines and statistical power. *Conservation Biology* **11**(1): 281-282.

Thomas, L. 1997. Retrospective power analysis. *Conservation Biology* **11**(1): 276-280.

Underwood, A.J., and Chapman, M.G. 2003. Power, precaution, Type II error and sampling design in assessment of environmental impacts. *Journal of Experimental Marine Biology and Ecology* **296**: 49-70.

Nonparametric ANOVA

Just like for our t-tests, we can perform analysis of variance without having to worry about distributional assumptions if we feel that our data are severely departed from normality. The nonparametric ANOVA is known as the **Kruskal-Wallis test**, and it is essentially an *analysis of variance by ranks*. Just like our other nonparametric tests, we don't compute parameter estimates (means and variances) to use in our test. Instead, we calculate the Kruskal-Wallis test statistic:

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

Where n_i = the number of observations in group i , N = the total number of observations in all k groups, and R_i is the sum of the ranks of the n_i observations in group i . The test doesn't explicitly assume homogeneity of variances among groups, but does assume the groups have generally the same shape and dispersion. However, the test is not very sensitive to violations of this assumption. As with the Mann-Whitney test, we can rank either from low to high or high to low, and we use the same procedures for tied ranks.

Example: A chemical analysis of three kinds of candy yields values of sugar content as shown below. We wish to determine whether the candy types differ in this trait, but the data are highly nonnormal and the group variances are heterogeneous. We have 5 observations for each candy type and we rank the data from low to high.

Runts	Jawbreakers	Gobstoppers
4.5 (5)	3.2 (1)	7.3 (13)
3.9 (2)	4.6 (6)	8.4 (15)
5.0 (9)	5.1 (10)	6.9 (12)
4.8 (7)	4.9 (8)	8.2 (14)
4.1 (3)	4.3 (4)	6.2 (11)
$R_1 = 26$	$R_2 = 29$	$R_3 = 65$

H_0 : the three candy types are identical in sugar content

H_A : at least one of the candy types has a different sugar content

We compute H:

$$H = \frac{12}{15(16)} \left[\sum \frac{26^2}{5} + \frac{29^2}{5} + \frac{65^2}{5} \right] - 3(16) = 9.42$$

We can look up the critical value for H ($H_{\alpha, n_1, n_2, n_3}$) using Table B.13 in Zar. For our example, $H_{.05, 5, 5, 5} = 5.780$. Since our calculated $H > H_{critical}$, we reject H_0 and conclude that at least one of the candy types has a different sugar content.

***Review examples 10.10 and 10.11 in Zar here

*Note that if the number of observations is greater than 8, you can compare H to a chi-square distribution or you can use the rank data to generate an approximate F-value and compare to an F-distribution. Also, if there are tied ranks, Zar presents a correction factor that is used in example 10.11. However, unless there are a very large number of ties, it doesn't amount to much.

Multiple comparisons

If we perform an ANOVA and reject our H_0 , we now want to know why! Several multiple comparison procedures have been developed to enable us to find out which of the means are different from others. We will examine a few of these procedures. One thing that many have in common is that the overall alpha (α) level is controlled (i.e., our probability of making a Type I error isn't inflated by performing multiple tests). In these cases, we refer to α as an *experimentwise error rate*.

There are two general approaches: (1) *a posteriori multiple comparison tests* where we make unplanned comparisons after the main test and (2) *a priori contrasts* in which we have specified beforehand (before performing the ANOVA) which hypotheses we want to test. There are advantages to each approach. We will outline some of the procedures for a posteriori multiple comparison tests first.

A posteriori multiple comparisons

There are many procedures for testing all pair-wise combinations of means after an ANOVA has rejected the overall null hypothesis. You will often see these procedures referred to as '**post-hoc**' tests, which just refers to the fact that they are unplanned comparisons that are performed after the main test. We will use an example to illustrate three of the more commonly used multiple comparison procedures.

Example: We have data on the calling frequency (#/hour) of three different species of frogs (*male frogs make calls to attract females, in case you didn't know). The data are below.

Species 1	Species 2	Species 3
10	4	2
8	6	0
6	2	2
8	5	4
8	3	2
$\bar{x} = 8$	$\bar{x} = 4$	$\bar{x} = 2$
$s^2 = 2.0$	$s^2 = 2.5$	$s^2 = 2.0$

Our estimate of $s_p^2 = 2.167 =$ within groups MS

Our grand mean $\bar{X} = 4.67$

Our among groups SS = 93.335

Our among groups MS = $93.335/2 = 46.665$

$F = 46.665/2.167 = 21.53$ and since $F_{0.05, 2, 12} = 3.89$, we reject the overall H_0 of our ANOVA.

In ANOVA table form, we have:

Source	SS	df	MS	F
Among	93.335	2	46.665	21.53
Within	26.004	12	2.167	
Total	119.339	14		

We can say that at least one of the frog species is different in its calling frequency, but we can't say which are different from which others.

The **Tukey HSD** (*honestly significant difference*) test considers the null hypothesis $H_0: \mu_a = \mu_b$, where the subscripts denote any possible pair of means. For k groups, $k(k-1)/2$ comparisons are possible. First, all group means are arranged in order of magnitude, then the test proceeds by comparing the largest vs. the smallest, followed by the largest vs. the next smallest and so

on, until all comparisons have been made. The test statistic is q (known as a *Studentized range*), which is calculated by dividing the mean differences by

$$SE = \sqrt{\frac{s^2}{n}}$$

where s^2 = the within groups (error) MS and n = the number of observations in each group (if all n_i are the same, then $n = n_i$; if the n_i 's are different, there are several approximations that are used to calculate SE; Zar uses the Tukey-Kramer approximation, which many times is the most powerful). The calculated q is then

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

If q is equal to or greater than the critical value, $q_{\alpha, v, k}$ (found in Table B.5 in Zar) then we reject the H_0 . Alpha (α) is the significance level of the test, v = the within groups (error) df from the ANOVA, and k = the total number of means being tested. The alpha (α) level in the case of the Tukey HSD test is now referred to as the *experimentwise error rate* and is the probability of encountering at least one Type I error among all mean comparisons.

For our frog calling example, we would have:

comparison	difference	SE	q	q_{critical}	conclusion
1 vs. 3	$8 - 2 = 6$	0.658	9.11	3.77	Reject H_0
1 vs. 2	$8 - 4 = 4$	0.658	6.08	3.77	Reject H_0
2 vs. 3	$4 - 2 = 2$	0.658	3.04	3.77	Fail to reject

*SE = $(2.167/5)^{0.5}$; $q_{0.05, 12, 3} = 3.773$

We conclude that frog species 1 calls with greater frequency than species 2 or 3, which aren't different from each other.

***Review examples 11.1 and 11.2 in Zar here

There will be times when your multiple comparison tests will generate results that seem confusing. For instance, we could have found that the means for frog species 1 and 2 were similar, the means for species 2 and 3 were similar, but the means for species 1 and 3 were different. This simply tells us that the test wasn't powerful enough to determine which population species 2 belonged to (either grouped with 1 or 3, or by itself). You can also have the situation where your overall ANOVA is significant, but the multiple comparison tests don't detect any pairwise differences. This can arise because the ANOVA is a more powerful test than any of the post-hoc tests.

The **Newman-Keuls test** (aka Student-Newman-Keuls test or SNK test) is another a posteriori multiple comparison approach. The test is performed exactly the same as the Tukey HSD test until the end. We first rank the means, determine the differences between the means, and calculate a SE and a q-value. However, the critical value of q is based on alpha (α), v , and p (instead of k), where p = the number of means in the range of means being compared. If we have three means and are comparing means 1 and 3, then $p = 3$; but if we are comparing means 1 and 2, then $p = 2$.

For our frog calling example, we would have:

comparison	difference	SE	q	q_{critical}	conclusion
1 vs. 3	$8 - 2 = 6$	0.658	9.11	3.77	Reject H_0
1 vs. 2	$8 - 4 = 4$	0.658	6.08	3.08	Reject H_0
2 vs. 3	$4 - 2 = 2$	0.658	3.04	3.08	Fail to reject



Now our difference between the means of species 2 and 3 is borderline (we are very close to the critical value). In general, the Newman-Keuls test will be more powerful compared to the Tukey HSD test (which is more conservative). Again, these are just two of many a posteriori multiple comparison procedures and there is no clear answer as to which is best. I tend to use the Tukey HSD test routinely, but it's comforting when other procedures detect the same differences.

***Review example 11.3 in Zar here

Another multiple comparison procedure known as **Scheffe's test** can also perform many pairwise tests, but it is generally less powerful than the tests we described above. The Scheffe's test is, however, more appropriate for testing what are referred to as 'multiple contrasts' a posteriori. Such contrasts are when we test groups of means against either a single mean or another set of means among our sample. For instance, in our frog calling example, we might wish to test whether species 1 is different from the average for species 2 and 3. In this case, H_0 would be stated: $(\mu_2 + \mu_3)/2 - \mu_1 = 0$. The Scheffe's test then expresses $(\mu_2 + \mu_3)/2$ as $\mu_2/2 + \mu_3/2$ and tests $H_0: \mu_2/2 + \mu_3/2 - \mu_1 = 0$. Now the μ_i 's are preceded by coefficients, c_i , of $c_2 = 1/2$, $c_3 = 1/2$, and $c_1 = -1$. Note that the sum of the coefficients = 0. The test statistic, S , is calculated as:

$$S = \frac{\left| \sum c_i \bar{X}_i \right|}{SE}$$

where

$$SE = \sqrt{s^2 \left(\sum \frac{c_i^2}{n_i} \right)}$$

The critical value for hypothesis tests is:

$$S_{\alpha} = \sqrt{(k-1)F_{\alpha(1),k-1,N-k}}$$

For the SE calculations and the critical value, s^2 = within groups (error) MS, and $k-1$ and $N-k$ are the among groups and within groups df, respectively. Keep in mind that our sets of means should have some logic behind their formation. For instance, in our frog calling example, maybe species 2 and 3 are smaller than species 1 and we expect that body size might affect calling frequency.

Example:

The critical value for our frog example would use $\alpha = 0.05$, $k-1 = 2$, $N-k = 12$.

$$S_{0.05} = \sqrt{(2)F_{0.05(1),2,12}} = 2.79$$

Contrast	SE	S	conclusion
$\frac{\bar{X}_2 + \bar{X}_3}{2} - \bar{X}_1$	$SE = \sqrt{2.167 \left[\frac{\left(\frac{1}{2}\right)^2}{5} + \frac{\left(\frac{1}{2}\right)^2}{5} + \frac{(-1)^2}{5} \right]} = 0.806$	$\frac{11}{0.806} = 13.65$	Reject H_0

***Review example 11.7

For cases when we have a control group that we wish to compare with all other sets of means, we can use a similar procedure called **Dunnett's test**. There are also multiple comparison procedures that follow nonparametric ANOVA (sections 11.6-7).

A priori contrasts

Unlike the procedures for unplanned 'post-hoc' comparisons outlined above, we use different methods when we have specified the contrasts we are interested in ahead of time (before running the ANOVA). These tests are more powerful because they don't require us to use the special tests of significance that were built into the above a posteriori procedures to protect against committing Type I errors.

First, we must decide how many and which planned comparisons to make. Technically, we can make as many as we would like, but many statisticians recommend that our planned contrasts be orthogonal to one another to ensure independence of results (i.e., that each contrast tests an independent relationship among the means). This way our P-values for each contrast are not correlated with one another. If there are k groups, then, at most, there can be $k-1$ orthogonal contrasts (although we can create the $k-1$ contrasts in multiple ways). We use an approach similar to the one outlined above for the Scheffe's test, in that we generate coefficients for each of the means in the contrast. The rules for building contrasts and assigning coefficients are presented by Gotelli and Ellison 2004 (pp. 339-341):

1. The sum of the coefficients for any contrast must equal 0
2. Sets of means averaged together have the same coefficient
3. Means not included in a contrast have a coefficient of 0
4. A maximum of $k-1$ orthogonal contrasts are possible
5. All of the pair-wise cross products must sum to 0

Rules 4 and 5 apply only when we want to limit our comparisons to orthogonal contrasts. If we chose to test non-orthogonal contrasts, we must adjust our alpha (α) level since the non-

independence of our tests will inflate our probability of making a Type I error. These types of adjustments to our alpha level are collectively referred to as **Bonferroni adjustments** and there are several types. The simplest is the Bonferroni method which sets $\alpha = \alpha/k$, where k = the number of tests performed.

To test planned contrasts, we construct a new mean square (MS):

$$MS_{contrast} = \frac{n \left(\sum_{i=1}^k c_i \bar{X}_i \right)^2}{\sum_{i=1}^k c_i^2}$$

This mean square has 1 df and we test it directly against the within-groups (error) MS from our ANOVA to determine if the contrast is significant.

Planned contrast example:

Returning to our frog calling example, suppose that we set up two contrasts beforehand that we are interested in testing. One involves comparing species 1 to the average of species 2 and 3, and the second involves testing whether species 2 and 3 are different from each other.

Contrast	coefficients	$MS_{contrast}$	F
$\frac{\bar{X}_2 + \bar{X}_3}{2} - \bar{X}_1$	$\bar{X}_1 (2), \bar{X}_2 (-1), \bar{X}_3 (-1)$	$\frac{5((2*8) + (-1*4) + (-1*2))^2}{2^2 + (-1)^2 + (-1)^2} = 83.33$	38.45
$\bar{X}_2 - \bar{X}_3$	$\bar{X}_1 (0), \bar{X}_2 (1), \bar{X}_3 (-1)$	$\frac{5((0*8) + (1*4) + (-1*2))^2}{0^2 + 1^2 + (-1)^2} = 10$	4.61



To test the significance of the contrasts, we compare our calculated F-ratios to $F_{\text{critical}} = F_{0.05,1,12} = 4.75$. We would conclude that frog species 1 is different than species 2 and 3, but that species 2 and 3 could not be distinguished. These conclusions are similar to those for our unplanned a posteriori comparisons, but this will not always be the case. The planned contrasts represent more powerful tests of differences between means or sets of means, but they must be identified before the ANOVA is run to be valid. Otherwise, we would be greatly inflating our probability of making a Type I error.

Two-factor ANOVA models

If we wish to examine the simultaneous effect of more than one factor on our response variable, then we must conduct what is referred to as a *factorial analysis of variance*. We introduced factorial designs earlier in our discussion of various experimental designs and noted that they were more efficient (time, labor, money) than conducting several single-factor ANOVAs. More importantly, they allow us to examine the **interactive effects** of multiple factors. Interactive effects are those that cannot be predicted from the additive effect of individual factors. We will begin with the layout for the two-factor model.

If we return to our example looking at the diving depths of turtles, recall that we had a single factor (exposure to stress) that we were interested in. Now suppose that in addition to exposure to stress, we have data on the body sizes of the turtles. We now want to run a 2-factor ANOVA examining the effect of stress, body size, and their interaction on diving depth. Therefore, we will have 3 null hypotheses related to the effects of each of our two factors and their interaction (see example 12.1 in Zar).

Also recall our ANOVA notation. We will make one change here, and refer to the number of levels of each treatment using subscripts a and b , instead of k . Each observation (X) is still denoted with a number for the treatment level (i or j) and a number for the observation within the treatment level (l up to n). We use l so we don't reuse k . Different factors or treatments are denoted with capital letters (A, B, C, \dots). In our case, we have two factors, which would receive the symbols A and B . When we refer to the summation of items across levels of the treatment, we sum from i to a or j to b . So our summation totals will now look like:

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n X_{ijl}$$

For our example of turtle diving depths we would have:

Small			Large		
1	2	3	1	2	3
X_{111}	X_{121}	X_{131}	X_{211}	X_{221}	X_{231}
X_{112}	X_{122}	X_{132}	X_{212}	X_{222}	X_{232}
X_{113}	X_{123}	X_{133}	X_{213}	X_{223}	X_{233}
X_{114}	X_{124}	X_{134}	X_{214}	X_{224}	X_{234}

We have two levels ($a = 1$ for small and 2 for large) of factor A (body size) and three levels ($b = 1, 2,$ or 3) of factor B (stress level). So we have $a \times b = 2 \times 3 = 6$ unique treatment combinations and $a \times b \times n = 2 \times 3 \times 4 = 24$ total replicates.

$$\sum_{i=1}^2 \sum_{j=1}^3 \sum_{l=1}^4 X_{ijl}$$

For our example of turtle diving depths the raw data is:

Small			Large		
1	2	3	1	2	3
75	25	100	120	75	130
80	75	80	100	100	100
75	25	100	100	75	90
50	75	40	80	100	110

Instead of just one mean square (MS) to represent the treatment effect like we had in a single-factor ANOVA, we will now partition the treatment effect into three mean squares that represent the effects of Factor A, Factor B, and their interaction. The factors are known as the **main effects** and the sum of squares (and mean square) for each main effect is calculated by averaging across all of the levels of the other factor. In our one-way design, we either controlled for the second factor (e.g., only examining small turtles) or its effect will simply contribute to the residual variation (our within-groups MS).

For our two-factor model:

$$\text{within-groups SS} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X}_{ij})^2$$

Then we pool the degrees of freedom to obtain:

$$\text{within-groups DF} = \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) = N - ab = ab(n-1)$$

Next, we calculate the amount of variation due to each of our factors. This is represented by the differences between each of the factor means (*averaged over all values of the other factor*) and the grand mean. This source of variation is calculated as:

$$\text{Factor A SS} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_i - \bar{X})^2$$

$$\text{Factor B SS} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{ij} - \bar{X})^2$$

The degrees of freedom for factor A and B sums of squares are $a-1$ and $b-1$, respectively.

Lastly, in addition to the variation due to the main effects, we need to calculate the variation due to their interaction. Remember, the interaction term represents the differences in the response that can't be predicted by simply adding up the two main factor effects.

The sum of squares for the interaction term is calculated using:

$$\text{Interaction (A} \times \text{B) SS} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2$$

The interaction term has $(a-1)(b-1)$ degrees of freedom.

As in the one-way layout, the division of the sums of squared deviations (SS) by their respective degrees of freedom (DF) generates a **mean square (MS)**. We can calculate our within-groups (error) MS, Factor A MS, Factor B MS, and Interaction MS in this way.

Then we can test each of our null hypotheses by comparing the variances (MS values) for factor A, B, and their interaction to our within-groups MS and calculating F-ratios for each test.

$$F_{\text{Factor A}} = \frac{\text{FactorA}_{MS}}{\text{Within - groups}_{MS}}$$

$$F_{\text{Factor B}} = \frac{\text{FactorB}_{MS}}{\text{Within - groups}_{MS}}$$

$$F_{\text{AxB}} = \frac{\text{Interaction(AxB)}_{MS}}{\text{Within - groups}_{MS}}$$

The ANOVA table for a two-factor design will be:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_i - \bar{X})^2$	a-1	$\frac{\text{FactorA}_{SS}}{a-1}$	$\frac{\text{FactorA}_{MS}}{\text{within}_{MS}}$
Factor B	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_j - \bar{X})^2$	b-1	$\frac{\text{FactorB}_{SS}}{b-1}$	$\frac{\text{FactorB}_{MS}}{\text{within}_{MS}}$
Interaction (A × B)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2$	(a-1) × (b-1)	$\frac{\text{Interaction}_{SS}}{(a-1)(b-1)}$	$\frac{\text{Interaction}_{MS}}{\text{within}_{MS}}$
Within-groups	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X}_{ij})^2$	ab(n-1)	$\frac{\text{within}_{SS}}{ab(n-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X})^2$	N-1		

Recall our turtle diving depth data:



Factor A	Small			Large		
Factor B	1	2	3	1	2	3
X_{ij}	75	25	100	120	75	130
X_{ij}	80	75	80	100	100	100
X_{ij}	75	25	100	100	75	90
X_{ij}	50	75	40	80	100	110
\bar{x}_{ij}	70	50	80	100	87.5	107.5

\bar{x} (grand mean) = 82.5

$\bar{x}_{\text{small}} = 66.67$ }
 $\bar{x}_{\text{large}} = 98.33$ } each of these 5 factor means is calculated
 $\bar{x}_{\text{stress 1}} = 85$ } by averaging over all the values of the
 $\bar{x}_{\text{stress 2}} = 68.75$ } other factor
 $\bar{x}_{\text{stress 3}} = 93.75$ }

To calculate our within-groups SS, we subtract each X_{ij} value from its corresponding \bar{x}_{ij} , square the differences and add them. For the 1st column: $\{(75-70)^2 + (80-70)^2 + (75-70)^2 + (50-70)^2\}$, and repeat for all 6 columns. Our within-groups SS = **7750.00**.

For each of our main factors, the SS is calculated by subtracting each of our 5 factor means listed above from the grand mean, squaring the differences, multiplying by the n used to estimate each mean, and adding them up. For the turtle data we have:

$$\text{Factor A SS} = \{(66.67-82.5)^2 + (98.33-82.5)^2\} * 12 = \mathbf{6016.67}$$

$$\text{Factor B SS} = \{(85-82.5)^2 + (68.75-82.5)^2 + (93.75-82.5)^2\} * 8$$

2575.00

For the Interaction SS, we start with each \bar{x}_{ij} and subtract the appropriate factor A mean and factor B mean, and then add the grand mean, square the result and multiply by the n that contributed to the \bar{x}_{ij} . We do this for each \bar{x}_{ij} and add them up. For the turtle data we have:

$$\text{Int SS}_{11} = (70.00 - 66.67 - 85.00 + 82.5)^2 * 4 = 2.78$$

$$\text{Int SS}_{12} = (50.00 - 66.67 - 68.75 + 82.5)^2 * 4 = 34.03$$

$$\text{Int SS}_{13} = (80.00 - 66.67 - 93.75 + 82.5)^2 * 4 = 17.36$$

$$\text{Int SS}_{21} = (100.0 - 98.33 - 85.00 + 82.5)^2 * 4 = 2.78$$

$$\text{Int SS}_{22} = (87.50 - 98.33 - 68.75 + 82.5)^2 * 4 = 34.03$$

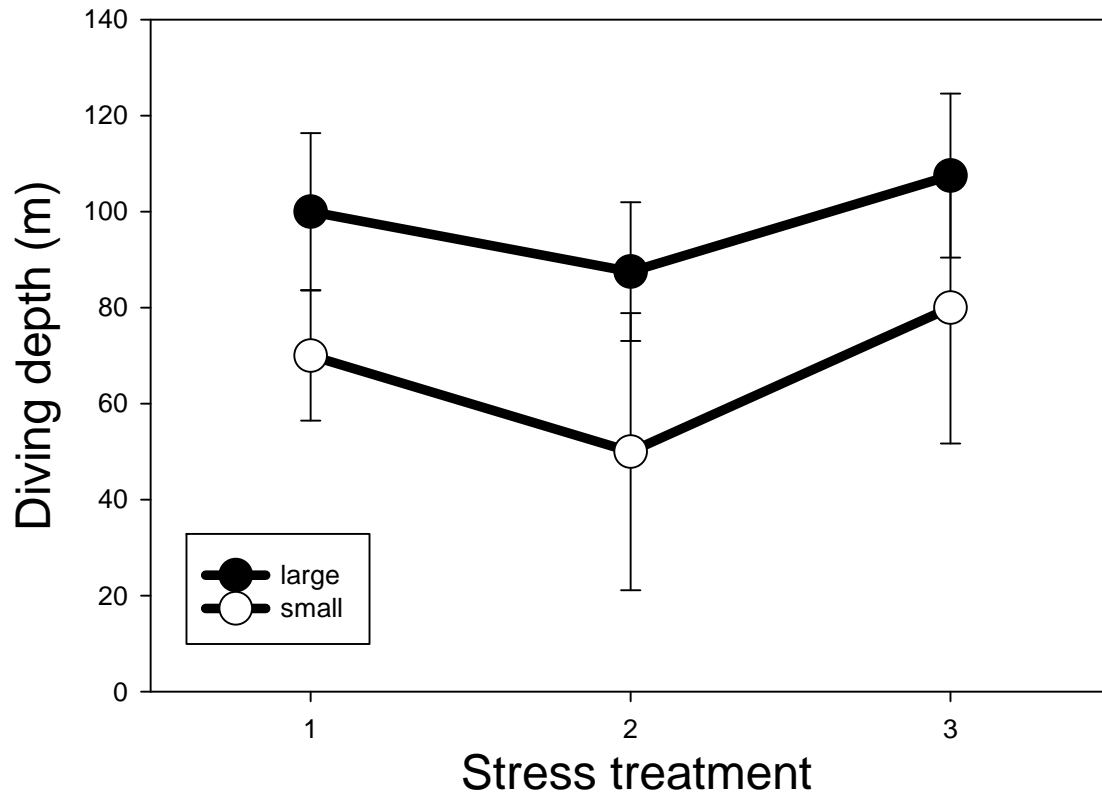
$$\text{Int SS}_{23} = (107.5 - 98.33 - 93.75 + 82.5)^2 * 4 = 17.36$$

Our Interaction SS is then = **108.33**

The 2-factor ANOVA table would appear as:

Source of variation	Sum of squares	df	Mean square (MS)	F
Body size	6016.67	1	6016.67	13.97
Stress	2575.00	2	1287.50	2.99
Interaction (A x B)	108.33	2	54.17	0.13
Within-groups	7750.00	18	430.56	
Total	16450.00	23		

The P-values are 0.002 for the body size effect, 0.076 for the Stress effect, and 0.883 for the interaction.



***Review example 12.2 in Zar here

Interpreting main effects and interactions

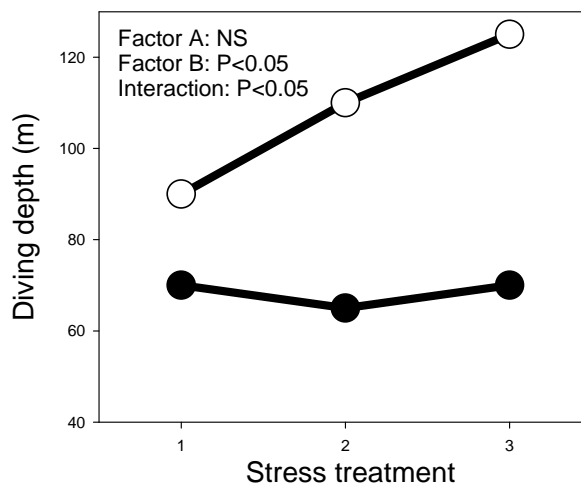
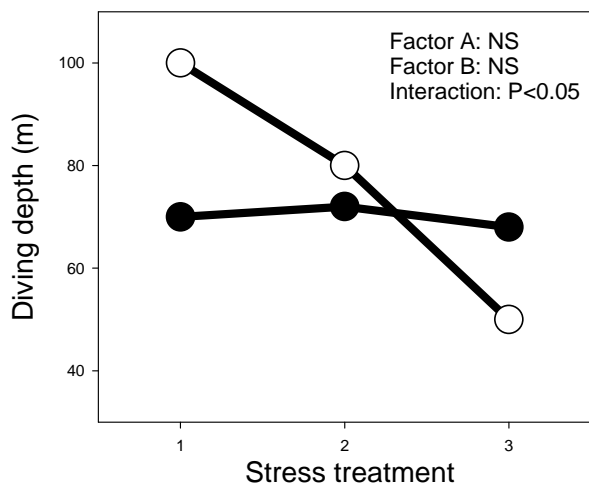
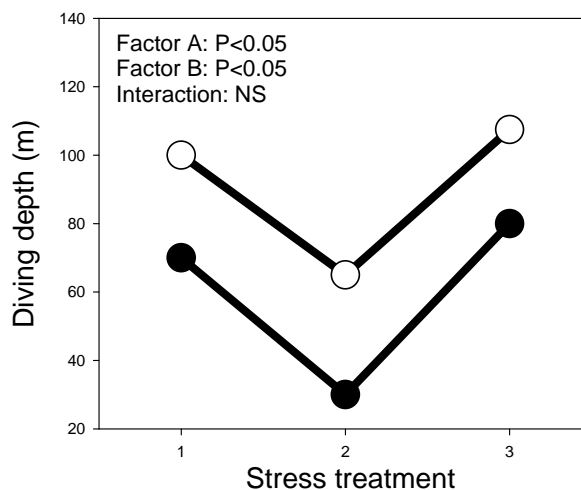
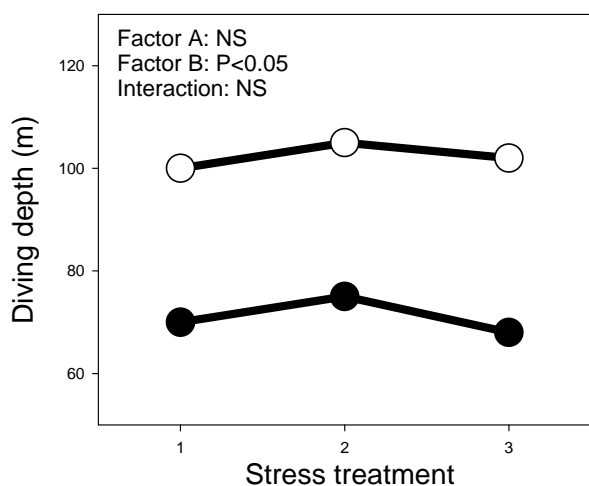
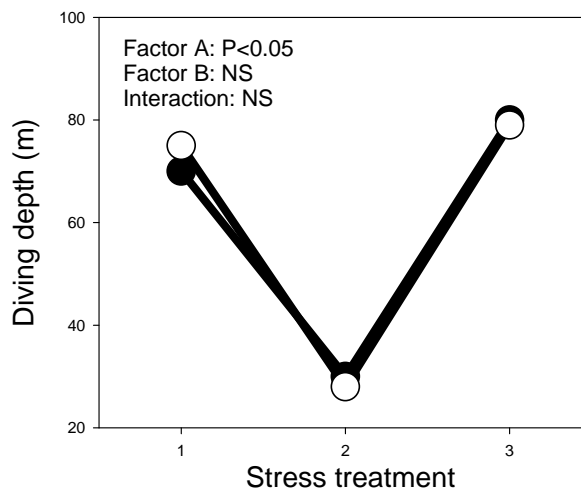
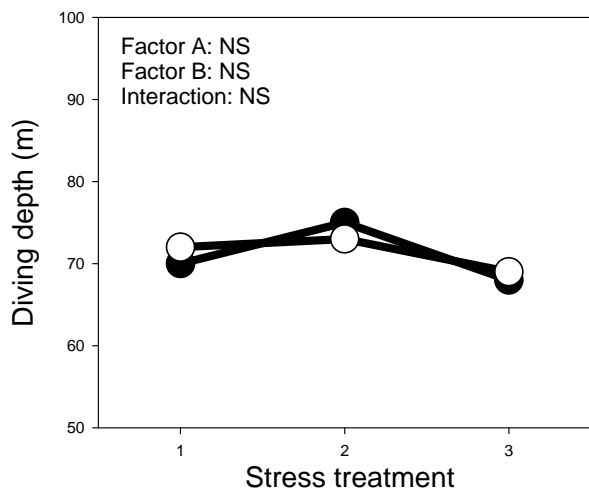
Many times you will read in your stats textbooks that if your factorial ANOVA detects a significant interaction effect, then you cannot interpret the main effects (Zar says something to this effect at the bottom of p. 242). In other words, the effect of each of your main factors will depend on the level of the other factor. This statement is generally, *but not always*, true. There are cases when you will find a significant interaction term, but are still able to draw inferences about the main effects. The best way to interpret your effects is to **plot them**. One-way ANOVA results can be easily interpreted using simple bar graphs to

represent the treatment means along with some measure of error (SD whiskers). For 2-factor ANOVA's, bar graphs don't work as well. Gotelli and Ellison (2004) recommend the following:

For plotting 2-way ANOVA results:

1. Plot the response variable on the y-axis and Factor A on the x-axis
2. For Factor B, plot the means of factor B at the corresponding level of factor A using symbols with different shapes or colors. Each symbol will represent a unique treatment combination (all $a \times b$)
3. Connect the symbols with lines
4. Add error bars as an option

Factor A is stress treatment (3 levels along the X-axis) and Factor B is body size represented by the two different lines (white and black symbols)



Calculating effect sizes after ANOVA

After we have conducted our ANOVA, completed any contrast or multiple comparison testing, and plotted the treatment means, we may wish to supplement the results of our hypothesis test by determining the proportion of the variance that can be attributed to each of our factors. We can further *partition our variance* after ANOVA by looking at the expected mean squares.

For a 1-way ANOVA, we know that our within-groups MS estimates the variation within groups (or residual error) and that our among-groups MS estimates the treatment effect plus the residual error:

$$\sigma_e^2 = MS_{\text{within-groups}}$$

$$\sigma_e^2 + n \sigma_A^2 = MS_{\text{among-groups}}$$

We can rearrange to isolate the treatment effect:

$$\sigma_A^2 = (MS_{\text{among-groups}} - MS_{\text{within-groups}})/n$$

One minor adjustment to account for the fact that we are estimating the variance from a finite sample when our factors are fixed and we have:

$$\sigma_A^2 = (MS_{\text{among-groups}} - MS_{\text{within-groups}}) * (a-1) / na$$

We can now estimate the **proportion of explained variance (PEV)** for our treatment using:

$$PEV_A = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

For a 2-factor ANOVA, the variance components are calculated using the following:

$$\text{Factor A} = (MS_{\text{factor A}} - MS_{\text{within-groups}}) * (a-1) / nab$$

$$\text{Factor B} = (MS_{\text{factor B}} - MS_{\text{within-groups}}) * (b-1) / nab$$

$$\text{Interaction} = (MS_{\text{interaction}} - MS_{\text{within-groups}}) * (a-1) * (b-1) / nab$$

We then calculate our PEV's by dividing each variance component by the sum of all variance components.

For our turtle diving depth example we would have:

$$\text{Factor A} = (6016.67 - 430.56) * (2-1) / 24 = \mathbf{232.75}$$

$$\text{Factor B} = (1287.50 - 430.56) * (3-1) / 24 = \mathbf{71.41}$$

$$\text{Interaction} = (54.17 - 430.56) * (2-1) * (3-1) / 24 = \mathbf{-94.10 (0)}$$

$$\text{Residual variance component} = \mathbf{430.56}$$

Then, our PEV's are calculated as:

$$\text{Factor A} = 232.75 / (232.75 + 71.41 + 0 + 430.56) = 0.317$$

$$\text{Factor B} = 71.41 / (232.75 + 71.41 + 0 + 430.56) = 0.097$$

$$\text{Interaction} = 0$$

$$\text{Residual} = 430.56 / (232.75 + 71.41 + 0 + 430.56) = 0.586$$

Thus, 31.7% of the variance is attributed to body size, 9.7% of the variance is attributed to stress level, and 58.6% of the variance remains attributed to random error. Overall, 41.4% of the variance can be attributed to our treatment effects.

Keep in mind that this variance partitioning only applies to the factors that we actually measured. Any variation in unmeasured (or uncontrolled) factors will be contained in the residual MS or the treatment MS if the unmeasured factors interact with our treatment factors. Also, the importance of any factor will depend on the *levels of that factor* that we included in our model. These limitations make it difficult to compare PEV's across studies.

Alternative ways to measure effect size

In performing ANOVA using statistical software packages, many will include some measure of effect size similar to the PEV's that we calculated above. I will introduce a couple of them of here just so you know what they are when you see them.

Eta squared (η^2) and **partial Eta squared (η_p^2)** are two measures of effect size that are routinely included in ANOVA output or available as an option in the output. The calculations for each are very simple. For Eta squared we have:

$$\eta^2 = \frac{SS_{factor}}{SS_{total}}$$

And, for partial Eta squared we have:

$$\eta_p^2 = \frac{SS_{factor}}{SS_{factor} + SS_{error}}$$

For our turtle diving depth example we would have:

Factor	SS_{factor}	SS_{error}	SS_{total}	η^2	η_p^2
Body size	6016.67	7750.00	16450.00	0.366	0.437
Stress	2575.00	7750.00	16450.00	0.157	0.249
Interaction	108.33	7750.00	16450.00	0.007	0.014

The Eta squared values are additive, meaning that we can interpret that about 53% of the total variance was due to our measured factors and their interaction (with about 47% remaining in unexplained residual error).

The partial Eta squared values are not additive, they should only be interpreted relative to each other (not the overall model). The partial in the name is because partial Eta squared values reflect effect size while controlling for all other variables in the model (in essence, effect in the absence of all other effects). Partial eta-squared values can thus be interpreted as the percent of variance in the dependent variable *uniquely attributable* to the factor. A general rule of thumb is that a partial Eta squared value above 0.2 is considered a large effect, above 0.1 (but less than 0.2) would be a moderate effect, and below 0.1 would be a small effect.

Omega squared (ω^2) is another measure of the proportion of variance in the response accounted for by a factor. It is estimated as:

$$\omega^2 = \frac{[SS_{factor} - (df_{factor} * MS_{error})]}{(MS_{error} + SS_{total})}$$

For our turtle diving depth example we would have:

Factor	SS_{factor}	SS_{error}	SS_{total}	ω^2
Body size	6016.67	7750.00	16450.00	0.331
Stress	2575.00	7750.00	16450.00	0.102
Interaction	108.33	7750.00	16450.00	-0.045

Omega squared (ω^2) is one of the most commonly used measures of the magnitude of the factor effects. Like partial Eta squared values, omega squared values are not additive. As a general rule, the effect can be interpreted as "large" when over 0.15, "medium" when 0.06 to 0.15, and "small" when 0.05 or less.

Note that our results for Omega squared values are more conservative than the Eta squared or partial Eta squared values that we obtained, and also agree more closely with our original PEV values. To be conservative and avoid overstating the results of the test, I tend to favor the use of Omega squared values for reporting effect sizes, but there is no single best approach.



Fixed vs. Random effects in ANOVA models

So far, all the tests we have covered have assumed that our factors are **fixed**. This is the standard (*or default*) assumption in statistical software packages (and also Excel) when we run a factorial analysis. However, we may also wish to analyze factors that we would define as **random**. The distinction lies in our interest (and thus, our scope of inference) about the levels of the factors we are testing. For a fixed factor, the levels that are being tested are the only ones of interest and our inferences are thus, restricted to those particular levels. For a random factor, the levels tested represent a random subset of many possible levels of that factor and we wish to be able to extend our inferences to all levels of the factor, not just the ones that we tested. In a **mixed model** design, some of our factors are fixed and some are random (e.g., one each in a 2-factor mixed model).

Determining whether a factor is fixed or random is not always simple, but as we will see, it has important implications for our sampling design and the calculation of our F-ratios. If our factor is a set of locations or times (usually randomly or systematically chosen), then it should generally be treated as a random factor. Similarly, if one of our factors represents a categorization of a continuous variable (e.g., body size), and we wish to extend our inferences across the range of that variable, we should also treat the factor as random. Alternatively, if the factor represents a defined set of categories that is limited in number (e.g., sex or species), we would treat it as a fixed factor. Gotelli and Ellison (2004) present a general 'rule of thumb' on p. 321 which suggests considering the ratio x/X where x is the number of levels of a factor you are testing and X is the number of possible levels of that factor. If the ratio is close to 0, then you probably should treat the factor as random. However, if the ratio is close to 1,

you can treat the factor as fixed. *There is no clear answer, it will depend on your question and the scope of inference you desire.*

The difference is critical because it affects how we calculate our F-ratios. Instead of dividing the factor MS by the error MS like we did in our fixed effects 2-factor model, a model with 2 random factors requires us to divide the factor MS values by the **interaction MS**. Similarly, in a mixed model, one of our factor MS terms is divided by the error MS and one is divided by the interaction MS. In both cases, the F-ratio for the interaction term is still calculated as it was for the fixed factor model, by dividing the interaction MS by the error MS. *This distinction only matters for factorial designs; the calculations in a 1-way ANOVA are the same whether the single factor is random or fixed.*

The ANOVA table for a two-factor **random effects** design:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A (random)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_i - \bar{X})^2$	a-1	$\frac{FactorA_{SS}}{a-1}$	$\frac{FactorA_{MS}}{Interaction_{MS}}$
Factor B (random)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_j - \bar{X})^2$	b-1	$\frac{FactorB_{SS}}{b-1}$	$\frac{FactorB_{MS}}{Interaction_{MS}}$
Interaction (A × B)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2$	(a-1) × (b-1)	$\frac{Interaction_{SS}}{(a-1)(b-1)}$	$\frac{Interaction_{MS}}{within_{MS}}$
Within-groups	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X}_{ij})^2$	ab(n-1)	$\frac{within_{SS}}{ab(n-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X})^2$	N-1		

The ANOVA table for a two-factor **mixed model** design:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A (fixed)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_i - \bar{X})^2$	a-1	$\frac{FactorA_{SS}}{a-1}$	$\frac{FactorA_{MS}}{Interaction_{MS}}$
Factor B (random)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_j - \bar{X})^2$	b-1	$\frac{FactorB_{SS}}{b-1}$	$\frac{FactorB_{MS}}{within_{MS}}$
Interaction (A × B)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2$	(a-1) × (b-1)	$\frac{Interaction_{SS}}{(a-1)(b-1)}$	$\frac{Interaction_{MS}}{within_{MS}}$
Within-groups	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X}_{ij})^2$	ab(n-1)	$\frac{within_{SS}}{ab(n-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X})^2$	N-1		

For our turtle diving depth example, body size should be considered a *random factor* since it is a continuous variable that we categorized. If we used a mixed model design, the F-ratio for body size would be calculated the same as for our fixed effects model, but the F-ratio for our stress factor would now use the interaction MS in the denominator. This results in a calculated $F = 23.77$, which is much higher than the $F = 2.99$ we had before. However, the critical F-value from Table B.4 is now based on the interaction df of 2, not the error df of 18, and is $F_{crit} = 19.0$. Therefore, we would reject the null hypothesis and conclude that our stress factor significantly affected diving depth, but the P-value would be about 0.045, not that much different than our P of 0.076 from the fixed effects model.

In addition to affecting the calculation of our F-ratios, models that include random factors require some additional thinking about our sampling design. For our fixed effects model, we wanted to replicate as much as possible for each factor level *to increase our degrees of freedom in the within-groups (error) MS*. This would increase our statistical power. For random factors, however, replicating within each factor level isn't as important as testing many factor levels. This is because the number of treatment levels determines the degrees of freedom for the interaction term. *If we boost the df for the interaction term, we increase our power to detect main effects.*

Variance components for random effects and mixed models

Whether we have fixed, random, or both types of factors in our design also affects how we calculate the proportion of explained variance (**PEV**) for each factor.

Component of variance	Fixed effects model (A fixed, B fixed)	Random effects model (A random, B random)	Mixed effects model (A fixed, B random)
Factor A	$\frac{(MS_A - MS_{error})(a - 1)}{abn}$	$\frac{(MS_A - MS_{AxB})}{bn}$	$\frac{(MS_A - MS_{AxB})(a - 1)}{abn}$
Factor B	$\frac{(MS_B - MS_{error})(b - 1)}{abn}$	$\frac{(MS_B - MS_{AxB})}{an}$	$\frac{(MS_B - MS_{AxB})}{an}$
A x B	$\frac{(MS_{AxB} - MS_{error})(a - 1)(b - 1)}{abn}$	$\frac{(MS_{AxB} - MS_{error})}{n}$	$\frac{(MS_{AxB} - MS_{error})}{n}$
Error	MS_{error}	MS_{error}	MS_{error}

Once we have computed the variance components, the PEV for each factor is then calculated as before (the ratio of the variance component of interest to the sum of all variance components).

PEV calculations

$$\text{Factor A} = \sigma_A^2 / (\sigma_A^2 + \sigma_B^2 + \sigma_{A \times B}^2 + \sigma_e^2)$$

$$\text{Factor B} = \sigma_B^2 / (\sigma_A^2 + \sigma_B^2 + \sigma_{A \times B}^2 + \sigma_e^2)$$

$$\text{Interaction} = \sigma_{A \times B}^2 / (\sigma_A^2 + \sigma_B^2 + \sigma_{A \times B}^2 + \sigma_e^2)$$

$$\text{Residual} = \sigma_e^2 / (\sigma_A^2 + \sigma_B^2 + \sigma_{A \times B}^2 + \sigma_e^2)$$

Other multifactor ANOVA designs

In addition to the standard 2-factor design, there are many other factorial ANOVA designs that can be employed. We will introduce a few of them here.

The **randomized block design** is employed when we place our factor levels within blocks, which are areas (space) or time periods within which the environmental conditions are relatively similar. Our blocks are generally arranged so that environments are more similar within a block than between blocks. Then, our factor levels are assigned randomly within the blocks. In a simple randomized block design, each block contains exactly 1 replicate of each factor level (i.e., no replication within blocks). Recall our example of flatfish recruitment in different habitat types. The different habitat types would represent different levels of the factor of interest and the block would represent a second factor.

The randomized block design is essentially a **mixed model 2-factor design with no replication**. The factor of interest is the *fixed factor* and the block is considered as a *random factor*. We will be able to estimate the factor effect, the block effect, and the error term, but no interaction term (since we have no replication within blocks). Note that we will lose some degrees of freedom for the error MS relative to a 1-way layout (*we have used these df to estimate the block effect*). If the block effect is large, we will have reduced the error SS enough to offset the loss of some df, and this design will be more powerful for detecting the effect of our factor of interest relative to a 1-way layout. However, if the block effect is weak, the reduction in our error SS will not offset the loss of the error MS df that we used to estimate the block effect, and we will have sacrificed some power.

The ANOVA table for a randomized block design:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A	$\sum_{i=1}^a \sum_{j=1}^b (\bar{X}_i - \bar{X})^2$	a-1	$\frac{FactorA_{SS}}{a-1}$	$\frac{FactorA_{MS}}{within_{MS}}$
Block	$\sum_{i=1}^a \sum_{j=1}^b (\bar{X}_j - \bar{X})^2$	b-1	$\frac{Block_{SS}}{b-1}$	$\frac{Block_{MS}}{within_{MS}}$
Within-groups	$\sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2$	(a-1)(b-1)	$\frac{within_{SS}}{(a-1)(b-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X})^2$	ab-1		

Note that although we consider the block as a random factor, we still construct our F-ratio for Factor A using the error MS in the denominator. This is because there is no interaction term (i.e., we assume no interaction between the block and our factor) since we have no replicates within each block. Keep in mind that we can test for the block effect, but it is usually not of interest. We expect there to be a large block effect or we wouldn't have blocked in the first place. The design allows us to adjust for the differences in our response across blocks to more clearly see the effects of our factor of interest.

Nested designs

Another type of ANOVA design that we have discussed is a **nested ANOVA model**. These designs involve subsampling within our replicates (i.e., we are not adding any more independent replicates, but are increasing the precision with which we estimate each replicate). The design allows us to test for differences among our replicates in addition to testing for the main effect of our factor. These types of designs also are appropriate when we wish to partition the variance in a hierarchical fashion (e.g., stations within sectors, sectors within regions, etc.). We need to identify where the nesting occurs to calculate the F-ratios properly (*we don't use the error MS in the denominator for all calculations*).

The ANOVA table for a nested design:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_i - \bar{X})^2$	a-1	$\frac{FactorA_{SS}}{a-1}$	$\frac{FactorA_{MS}}{replicates_{MS}}$
Among replicates within levels of Factor A	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (\bar{X}_{j(i)} - \bar{X}_i)^2$	a(b-1)	$\frac{replicates_{SS}}{a(b-1)}$	$\frac{replicates_{MS}}{subsamples_{MS}}$
Subsamples within replicates	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X}_{j(i)})^2$	ab(n-1)	$\frac{subsamples_{SS}}{ab(n-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n (X_{ijl} - \bar{X})^2$	abn-1		

As an example of a hierarchical nested design, suppose we were going to measure flatfish recruitment in an estuary. The estuary is divided spatially into two large regions, each with 3 sectors, and we randomly select 6 stations within each sector. We have 3 replicate cages to capture settling fish at each station. We can analyze the data using a nested design to partition variance at our station level, our sector level, and our region level. To calculate the F-ratios for the region effect and the sector(region) effect [*read sector nested within region effect*], we use the station(sector(region)) MS and df rather than the error MS and df in the denominator. Then, the station(sector(region)) effect is calculated using the error MS and df.

The split-plot design

Recall that in a split-plot design, we have a randomized block design and then a second treatment is applied to the blocks (our whole-plot factor). So, we have our within-plot or subplot factor, our blocks (or plots), and our whole-plot factor. If you remember our flatfish recruitment example, habitat type was our within-plot or subplot factor, we had blocks with each of the 3 habitat types in each block, then we applied a predation treatment (caged, uncaged, control caged) to the blocks.

The ANOVA table for a split-plot design:

Source of variation	Sum of squares	df	Mean square (MS)	F
Factor A (whole-plot)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (\bar{X}_i - \bar{X})^2$	a-1	$\frac{FactorA_{SS}}{a-1}$	$\frac{FactorA_{MS}}{FactorB(A)_{MS}}$
Factor B(A) (plots or blocks nested within A)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (\bar{X}_j - \bar{X})^2$	a(b-1)	$\frac{FactorB(A)_{SS}}{a(b-1)}$	
Factor C (within-plot)	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (\bar{X}_l - \bar{X})^2$	(c-1)	$\frac{FactorC_{SS}}{(c-1)}$	$\frac{FactorC_{MS}}{B(A)xC_{MS}}$
A x C Interaction	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (\bar{X}_{il} - \bar{X}_i - \bar{X}_l + \bar{X})^2$	(a-1)(c-1)	$\frac{AxC_{SS}}{(a-1)(c-1)}$	$\frac{AxC_{MS}}{B(A)xC_{MS}}$
B(A) x C Interaction	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (X_{ijl} - \bar{X}_{il})^2$	a(b-1)(c-1)	$\frac{B(A)xC_{SS}}{a(b-1)(c-1)}$	
Total	$\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^c (X_{ijl} - \bar{X})^2$	abc-1		



Note that there is no error term; that is because we can't isolate it since we have no replication within our blocks, and we assume we have no interaction between our within-plot factor and the plots nested within A. Therefore, the $B(A) \times C$ interaction MS serves as our measure of residual error for the model and is used as the denominator to calculate the F-ratios for Factor C and the $A \times C$ interaction. To test for the effects of the whole-plot treatment (Factor A), we use the MS for Factor B(A) in the denominator since the plots serve as independent replicates for the whole-plot factor.

Repeated measures designs

There are two main types of repeated measures ANOVA designs. The first is when each replicate is exposed to different experimental treatment levels applied at different times and generally in some randomized order. The second type is when the treatment is applied only once, but each replicate is measured for a response multiple times over the course of time. The first type of repeated measures design is simply analyzed as a randomized block design (see ANOVA table in previous section p. 162), and we assume no interaction between replicates and treatments. Randomization of the application of the different treatment levels reduces the effects of confoundedness between treatment and time. The second type of repeated measures design is analyzed as a split-plot design (see ANOVA table in previous section p. 165) with each replicate being equivalent to a plot. The treatment is the whole-plot factor and time is the within-plot factor. Again, we assume no interaction between replicates and time (our plot and within-plot factor). We can now estimate the effects of our treatment, time, and treatment \times time interaction (which is often the most interesting).

Three-factor or multi-factor designs

Theoretically, we can analyze as many factors simultaneously as we wish, but the data demands begin to get enormous for a fully crossed design. Three-factor designs are not uncommon to see in the literature, but interpretation can get complex, particularly for the interaction of all three factors. The test for the three-factor interaction term asks if the interaction between any two of the factors is the same at all levels of the third factor.

***Review example 14.1 in Zar here

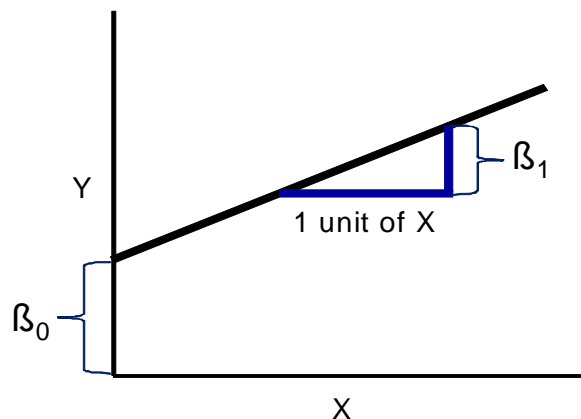
Regression and correlation

Recall from our previous introduction of different types of experimental designs that we use regression analysis to examine the relationship between continuous variables (i.e., our X variable is now continuous as opposed to categorical in ANOVA). **Simple linear regression** is used to determine the equation of the straight line that best describes the *functional relationship* between X and Y. **Correlation** describes the *association* between two random variables and its use implies that we don't know which variable is dependent on which, but rather that they simply vary together in some predictable way. **Regression** assumes that one variable is dependent on the other (Y dependent on X) in a *cause-and-effect relationship*. However, the same statistical models can be used in both cases.

The simplest function that describes the relationship between two variables is the linear model:

$$Y = \beta_0 + \beta_1 X$$

This equation describes a straight line with two parameters, the **intercept** (β_0) and the **slope** (β_1). The intercept is the value of the function when $X = 0$, and the slope measures the change in variable Y for each unit change in variable X.



Our data will consist of a series of paired observations that each includes an X value (X_i) and a Y value (Y_i). The model we use to fit our data is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

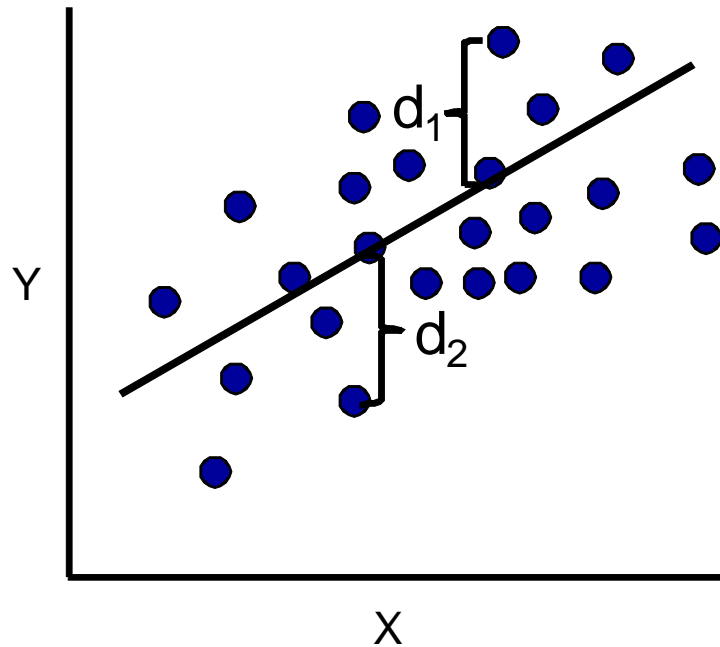
The term ε_i is the *error term*, which is a normal random variable with an expected value = 0 and a variance equal to σ^2 . The larger the value of σ^2 , the more noise there will be about the regression line. Once we have three or more data points, there will nearly always be some noise about the regression line.

Our first step when we plan to perform regression analysis should be to **plot our data**. We may be able to see a clear pattern. But how do we determine where the line is placed within our pattern? The regression line **must** pass through the single point defined by our mean X and our mean Y (\bar{X}, \bar{Y}). But, to determine which way the line pivots on this point, we need another quantity. We call this quantity the **residual**, d_i , and it is defined as the difference between the observed Y_i and the Y_i that is predicted by the regression equation ($Y_{\text{hat}} = \hat{Y}_i$). It is calculated as:

$$d_i = (Y_i - \hat{Y}_i)$$

We sum the squares of all of the residuals to create the **residual sum of squares (ResSS)**, and the best fit regression line is the one that minimizes the ResSS. Our line then results in the smallest average difference between each observed Y_i value and the predicted \hat{Y}_i value.

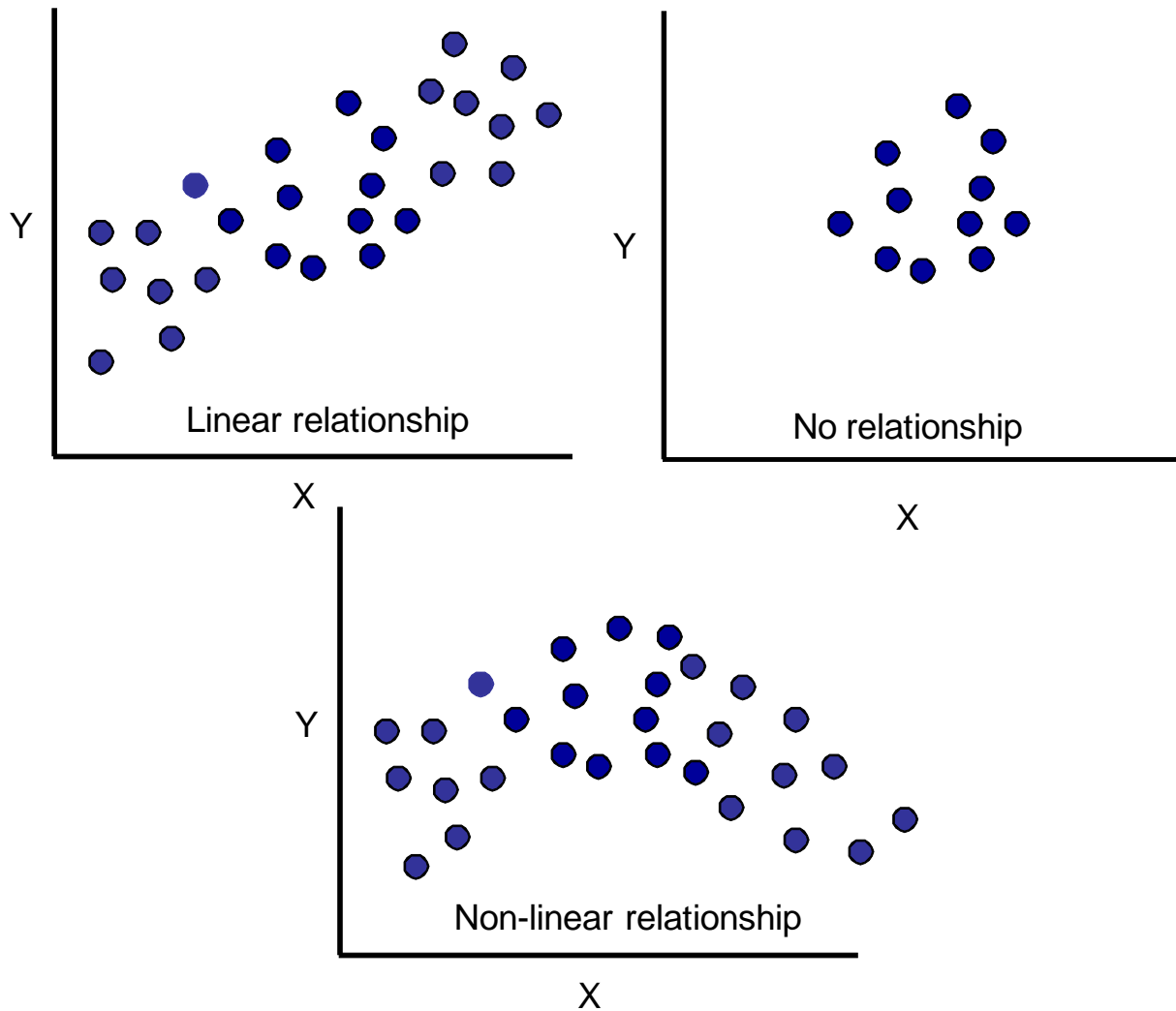
To illustrate this concept, we use what is called a **bivariate scatter diagram**:



Each observation from i to n will generate a residual. The model parameters (β_0 and β_1) are then chosen such that ResSS is minimized:

$$\sum_{i=1}^n d_i^2$$

***Keep in mind that bivariate scatter diagrams can take many shapes besides a linear relationship:



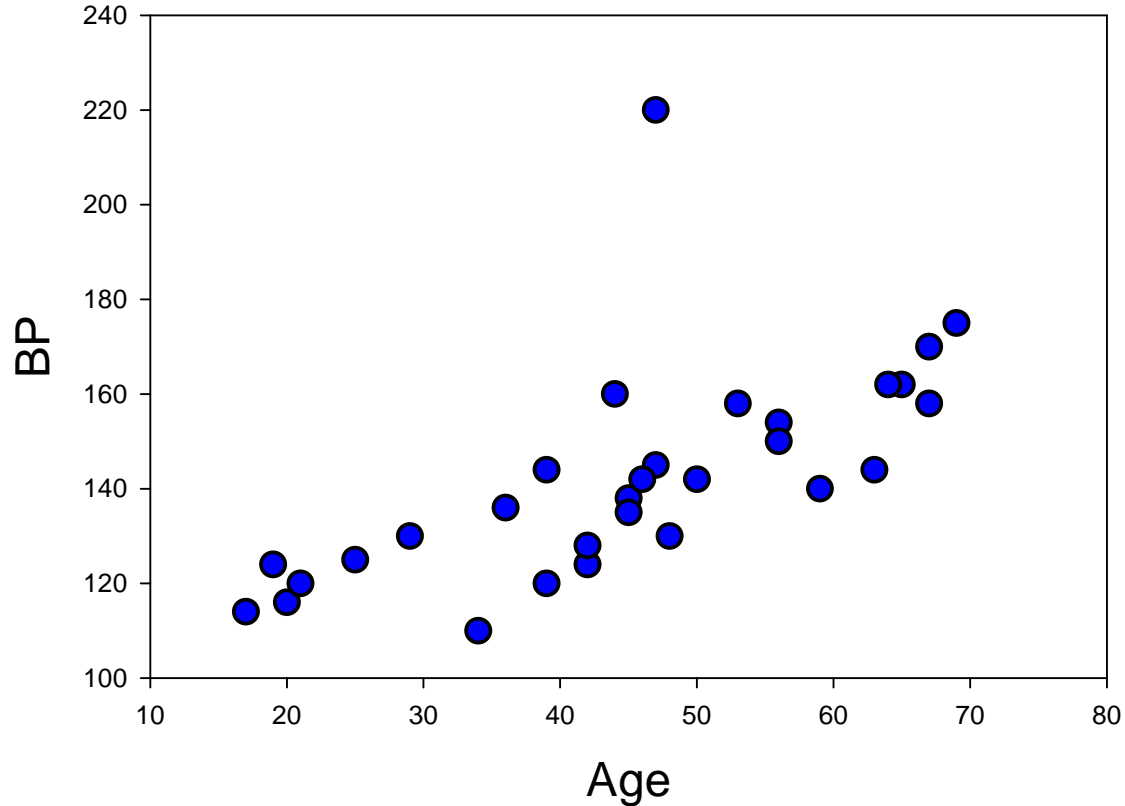
We will generally always start by assuming a simple linear relationship unless we have previous empirical or theoretical support for something more complex. We will deal with how to test for “**lack of fit**” of the linear model after we have finished describing the model and tested our hypothesis.

Let's illustrate these ideas using an example: Suppose we have some data on blood pressure and age for a random sample of 30 individuals.

Individual	BP	Age	Individual	BP	Age
(i)	(Y)	(X)	(i)	(Y)	(X)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

We have 30 pairs (X_i, Y_i) of observations that may be considered points in two dimensional space, so we can plot them as a bivariate scatter diagram.

Scatter diagram of blood pressure vs. age



***Note that we have one observation that is quite different from the rest, we will deal with the influence of this observation later.

In order to choose the parameters (intercept and slope) that generate the 'best fit' line, we need to introduce the idea of **covariance**. Recall our standard variance formula for one variable:

$$s^2_y = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

If we now consider two variables X and Y , we can define the **sum of the cross products** (SS_{XY}) as:

$$SS_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

And the sample **covariance** (S_{XY}) is then:

$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Unlike the variance for a single variable, the sample covariance can be either positive or negative depending on how the pairs of observations are organized (i.e., if large X 's are paired with large Y 's the covariance will be positive, but if large X 's are paired with small Y 's, the covariance will be negative). We can use our estimate of covariance to calculate our slope parameter (*which should make intuitive sense*).

$$\hat{\beta}_1 = \frac{S_{XY}}{s^2_X} = \frac{SS_{XY}}{SS_X}$$

To use the computational formula:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

Since our regression line must pass through (\bar{X}, \bar{Y}) , we can calculate the intercept as:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Lastly, we need to estimate the error term (ε_i) for our linear model. Recall that ε_i has a normal distribution with a mean = 0 and a variance = σ^2 . We use our estimate of the residual sums of squares (ResSS) to estimate the regression variance (σ^2).

$$\hat{\sigma}^2 = \frac{\text{ResSS}}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$$

which is simply:

$$= \frac{\sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2}{n-2}$$

***Note that the denominator is (n-2) instead of the usual (n-1). This is because we have used 2 degrees of freedom already to estimate the slope and the intercept.

The square root of the quantity above (σ) is referred to as the **standard error of the regression**, and is often reported as part of the regression output by many statistical software packages.

Returning to our blood pressure vs. age example:

Individual	BP	Age			
(i)	(Y)	(X)	(Y ²)	(X ²)	(XY)
1	144	39	20736	1521	5616
2	220	47	48400	2209	10340
3	138	45	19044	2025	6210
4	145	47	21025	2209	6815
5	162	65	26244	4225	10530
6	142	46	20164	2116	6532
7	170	67	28900	4489	11390
8	124	42	15376	1764	5208
9	158	67	24964	4489	10586
10	154	56	23716	3136	8624
11	162	64	26244	4096	10368
12	150	56	22500	3136	8400
13	140	59	19600	3481	8260
14	110	34	12100	1156	3740
15	128	42	16384	1764	5376
16	130	48	16900	2304	6240
17	135	45	18225	2025	6075
18	114	17	12996	289	1938
19	116	20	13456	400	2320
20	124	19	15376	361	2356
21	136	36	18496	1296	4896
22	142	50	20164	2500	7100
23	120	39	14400	1521	4680
24	120	21	14400	441	2520
25	160	44	25600	1936	7040
26	158	53	24964	2809	8374
27	144	63	20736	3969	9072
28	130	29	16900	841	3770
29	125	25	15625	625	3125
30	175	69	30625	4761	12075

(i)	(Y)	(X)	(Y ²)	(X ²)	(XY)
Sum	4276	1354	624260	67894	199576
Mean	142.53	45.13			
Variance	509.91	233.91			
Std Dev	22.58	15.29			
			n=	30	
			slope=	0.97	
			intercept=	98.75	

The slope is calculated using:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{199576 - \frac{(1354)(4276)}{30}}{67894 - \frac{(1354)^2}{30}} = 0.97$$

The intercept is then:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 142.53 - (0.97)(45.13) = 98.75$$

The equation for the straight line is:

$$\hat{Y} = 98.75 + 0.97X$$

For every X then, we can calculate a value for \hat{Y} and calculate our ResSS.

Individual	BP	Age			
(i)	(Y)	(X)	(\hat{Y})	(d_i)	(d_i) ²
1	144	39	136.54	7.46	55.65
2	220	47	144.30	75.70	5730.49
3	138	45	142.36	-4.36	19.01
4	145	47	144.30	0.70	0.49
5	162	65	161.76	0.24	0.06
6	142	46	143.33	-1.33	1.77
7	170	67	163.70	6.30	39.69
8	124	42	139.45	-15.45	238.70
9	158	67	163.70	-5.70	32.49
10	154	56	153.03	0.97	0.94
11	162	64	160.79	1.21	1.46
12	150	56	153.03	-3.03	9.18
13	140	59	155.94	-15.94	254.08
14	110	34	131.69	-21.69	470.46
15	128	42	139.45	-11.45	131.10
16	130	48	145.27	-15.27	233.17
17	135	45	142.36	-7.36	54.17
18	114	17	115.20	-1.20	1.44
19	116	20	118.11	-2.11	4.45
20	124	19	117.14	6.86	47.06
21	136	36	133.63	2.37	5.62
22	142	50	147.21	-5.21	27.14
23	120	39	136.54	-16.54	273.57
24	120	21	119.08	0.92	0.85
25	160	44	141.39	18.61	346.33
26	158	53	150.12	7.88	62.09
27	144	63	159.82	-15.82	250.27
28	130	29	126.84	3.16	9.99
29	125	25	122.96	2.04	4.16
30	175	69	165.64	9.36	87.61
Sum					8393.51

Our ResSS = 8393.51 and if we divide this by $n-2 = 28$, we get 299.77 for our estimate of the residual variance (mean square).

Our estimate of the **standard error of the regression** is then the square root of this quantity = **17.31**.



Partitioning the variance in regression

In the regression model, the total variation is defined by the sum of the squares of the Y variable (TotalSS or SS_Y):

$$TotalSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

We wish to partition this variance into its components. One of these components is pure (or random) error just due to random sampling from a normal distribution. This component of error is simply ε_i and we can estimate it using ResSS. The other component is not random error, but instead is systematic. The source of this variation is the regression relationship $Y_i = \beta_0 + \beta_1 X_i$ and is estimated as RegSS. We can estimate it from:

$$RegSS = TotalSS - ResSS$$

In our example, the TotalSS = 14787.53 and our ResSS was 8393.51, so our RegSS = 6394.02.

An index that reflects the relative contribution of regression versus residual variation to the overall variance is r^2 , the **coefficient of determination**, which is calculated as:

$$r^2 = \frac{RegSS}{TotalSS}$$

The r^2 value tells you the proportion of the variation in Y that can be attributed to variation in X, and it varies from 0 to 1. For our blood pressure vs. age example, $r^2 = 0.43$. The square root of $r^2 = r =$ **the product-moment correlation coefficient**. The sign of r can be either positive or negative, depending on the covariance between X and Y, and it can range between -1 and 1. It is calculated as:

$$r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}}$$

Hypothesis testing with regression models

Recall that we assume a cause-and-effect relationship between X and Y when we perform a regression analysis. The test of the existence of such a relationship is contained in the slope parameter, and whether it differs from 0. Therefore, our null hypothesis is that $\beta_1 = 0$. If we fail to reject this null, then we have no evidence for the functional dependence of Y on X.

First, we can organize our regression data into an ANOVA table.

Source	df	SS	MS	F-ratio	P-value
Regression	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\frac{SS_{reg}}{1}$	$\frac{MS_{reg}}{MS_{residual}}$	$F_{1,n-2}$
Residual	n-2	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\frac{RSS}{n-2}$		
Total	n-1	$\sum_{i=1}^n (Y_i - \bar{Y})^2$			

For our blood pressure vs. age data, we have:

Source	df	SS	MS	F-ratio	P-value
Regression	1	6394.02	6394.02	21.33	$F_{(0.05)1,28} = 4.20$
Residual	28	8393.51	299.77		$P < 0.0005$
Total	29	14787.53			

***Keep in mind that ANOVA tables for regression analyses are usually not included in published manuscripts. Generally, the F-ratio, its degrees of freedom, and the P-value are reported in the text of the paper.

The significance of both our slope and intercept can each be tested using our familiar t-test. Recall that $t = \text{our parameter estimate} - \text{our hypothesized value} / \text{the standard error of the parameter estimate}$. We need **standard error estimates** for our slope and intercept to calculate t .

For the slope, the variance is estimated from:

$$= \frac{\hat{\sigma}^2}{SS_X}$$

For the intercept, the variance is estimated from:

$$= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X} \right)$$

The standard errors are just the square roots of those terms.

For our example, the t-value for the slope is:

$$t = \frac{0.97 - 0}{\sqrt{\frac{299.77}{6783.47}}} = \frac{.097}{0.21} = 4.62$$

and the t-value for the intercept is:

$$t = \frac{98.75 - 0}{\sqrt{299.77 \left(\frac{1}{30} + \frac{2036.72}{6783.47} \right)}} = \frac{98.75}{9.99} = 9.88$$

The critical t-value ($t_{0.05(2), 28}$) = 2.048, so both null hypotheses would be rejected. The hypothesis test for the significance of the slope is the same as the overall F-test for the entire regression model.

Regression Confidence and Prediction Intervals

Recall that we used estimates of the standard error for the slope and intercept to calculate t-statistics and test hypotheses about those parameters. We use the same standard error estimates to generate **confidence intervals** for the regression.

For the slope we had:

$$= \sqrt{\frac{\hat{\sigma}^2}{SS_X}} = \sqrt{\frac{299.77}{6783.47}} = 0.21$$

For the intercept we had:

$$= \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X} \right)} = \sqrt{299.77 \left(\frac{1}{30} + \frac{2036.72}{6783.47} \right)} = 9.99$$

The confidence interval for the slope is then:

$$\hat{\beta}_1 - t_{(\alpha, n-2)} \hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{(\alpha, n-2)} \hat{\sigma}_{\hat{\beta}_1}$$

A 95% CI for the slope in our example is:

$$0.97 - 2.048 * 0.21 \leq \beta_1 \leq 0.97 + 2.048 * 0.21$$

$$0.540 \leq \beta_1 \leq 1.400$$

Similarly, the confidence interval for the intercept is:

$$\hat{\beta}_0 - t_{(\alpha, n-2)} \hat{\sigma}_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_{(\alpha, n-2)} \hat{\sigma}_{\hat{\beta}_0}$$

A 95% CI for the intercept in our example is:

$$98.75 - 2.048 * 9.99 \leq \beta_0 \leq 98.75 + 2.048 * 9.99$$

$$78.29 \leq \beta_0 \leq 119.21$$

For any value X , we can also create **confidence intervals for the mean of the fitted Y values** using the standard error for a fitted Y :

$$\hat{\sigma}_{(\hat{Y}/X)} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X} \right)}$$

The confidence interval is then calculated as:

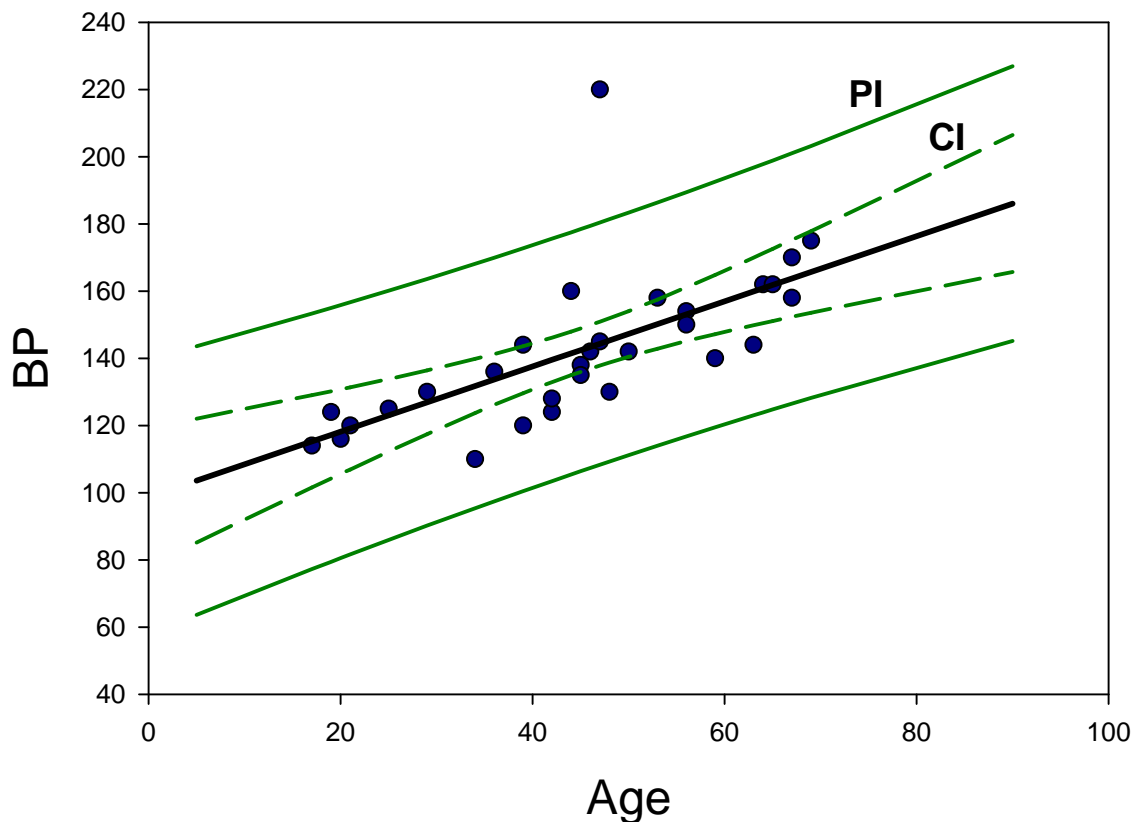
$$\hat{Y} - t_{(\alpha, n-2)} \hat{\sigma}_{(\hat{Y}/X)} \leq \hat{Y} \leq \hat{Y} + t_{(\alpha, n-2)} \hat{\sigma}_{(\hat{Y}/X)}$$

For any value X , we can also create what is referred to as a **prediction interval** using the standard error of the prediction for a single fitted Y :

$$\hat{\sigma}_{(\tilde{Y}/X)} = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X} \right)}$$

The prediction interval is then calculated as:

$$\hat{Y} - t_{(\alpha, n-2)} \hat{\sigma}_{(\tilde{Y}/X)} \leq \hat{Y} \leq \hat{Y} + t_{(\alpha, n-2)} \hat{\sigma}_{(\tilde{Y}/X)}$$



Note that prediction interval is wider than the confidence interval because we are predicting the variance we might expect in our predictions given new sample data (and the variance that is associated with that new data has to be accounted for). We can think of our confidence interval as measuring our confidence in estimates based on the available data, and the prediction interval as measuring confidence in estimates based on new data. Also note that both get wider the farther we get from \bar{X} . This should make sense, our confidence should decrease as we move away from the center of the distribution of our sample data.

Assumptions of regression

1. The functional relationship between X and Y is described by a linear model
 - *If this assumption is violated, the residual error will be inflated by a 'lack of fit' component.*
2. The X variable is measured without error
 - *If this assumption is violated, our estimates of the slope and intercept will be biased. We can use a Model II regression model if we suspect a severe violation.*
3. For any given X, the sampled Y values are independent and normally distributed
 - *The assumption of independence is always present, and normality allows use of parametric hypothesis tests. These are generally ignored unless violations are severe.*

4. Variances in Y are homogeneous for the range of X values

- *This assumption allows us to use the same estimate of residual error for the variance of the regression line. Ordinarily, some type of transformation is used to minimize heteroscedasticity, or we can use other regression models (e.g., nonlinear, quantile regression).*

Testing for 'lack of fit'

Up to this point, we have assumed that the relationship between X and Y is linear. However, we will often encounter bivariate relationships that are nonlinear. We can assess the appropriateness of the straight-line model using an ANOVA technique. Essentially, we will test for 'lack of fit' of our assumed straight-line model.

The key statistic is the **residual SS (ResSS)**. This quantity can be large for two reasons: (1) σ^2 is large (i.e., we have lots of variability in our sample data, or (2) the straight-line model is not appropriate. So, our residual SS can be partitioned into two components, one that describes pure error and the other that describes the extent of lack of fit of the straight-line model. We can estimate the pure error component by taking advantage of 'replicate observations' (i.e., multiple observations of Y taken at the same X).

Returning to our BP vs. Age example: We have two observations of Y at six values of X (X = 39, 42, 45, 47, 56, 67). The remaining 18 Y's are paired with a unique X value. We can then estimate a sum of squares (SS) for each of our X values that has more than a single Y.

$$\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \bar{Y}_{X_i})^2$$

For example, at $X = 42$, we have $Y = 124$ and 128 and $\bar{Y} = 126$.

$$\hat{\sigma}^2 = \sum_{i=1}^2 (124 - 126)^2 + (128 - 126)^2 = 4 + 4 = 8$$

If we do this for all six values of X we have:

X	Y	SS	df = $n_i - 1$
39	144; 120	288	1
42	124; 128	8	1
45	138; 135	4.5	1
47	220; 145	2812.5	1
56	154; 150	8	1
67	170; 158	72	1
		3193	6

So our $SS_{\text{pure error}} = 3193$

The $SS_{\text{lack of fit}}$ now is just = $ResSS - SS_{\text{pure error}}$

In our example we have,

$$SS_{\text{lack of fit}} = 8393.51 - 3193 = 5200.51$$

And $df_{\text{lack of fit}}$ now just = $Resdf - df_{\text{pure error}}$

$$df_{\text{lack of fit}} = 28 - 6 = 22$$

We then estimate the mean squares as:

$$\text{Pure error} = SS_{\text{pure error}}/df_{\text{pure error}} = 3193/6 = 532.17$$

$$\text{Lack of fit} = SS_{\text{lack of fit}}/df_{\text{lack of fit}} = 5200.51/22 = 236.39$$

$$\text{Then, } F = MS_{\text{lack of fit}}/MS_{\text{pure error}} = 236.39/532.17 = 0.44$$

$F_{0.05(22,6)} = 3.86$, so we cannot reject H_0 which stated that the straight-line model was appropriate.

Let's look at this graphically

$$\text{Recall that: } \text{TotalSS} = \text{RegSS} + \text{ResSS}$$

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$$

TotalSS = the total variability in Y before accounting for the effect of any independent variables (X)

RegSS = the reduction in variability in Y (e.g., variability explained) due to the inclusion of independent variables (X) in the model

ResSS = the amount of variability in Y left unexplained after accounting for the effect of any independent variables (X)

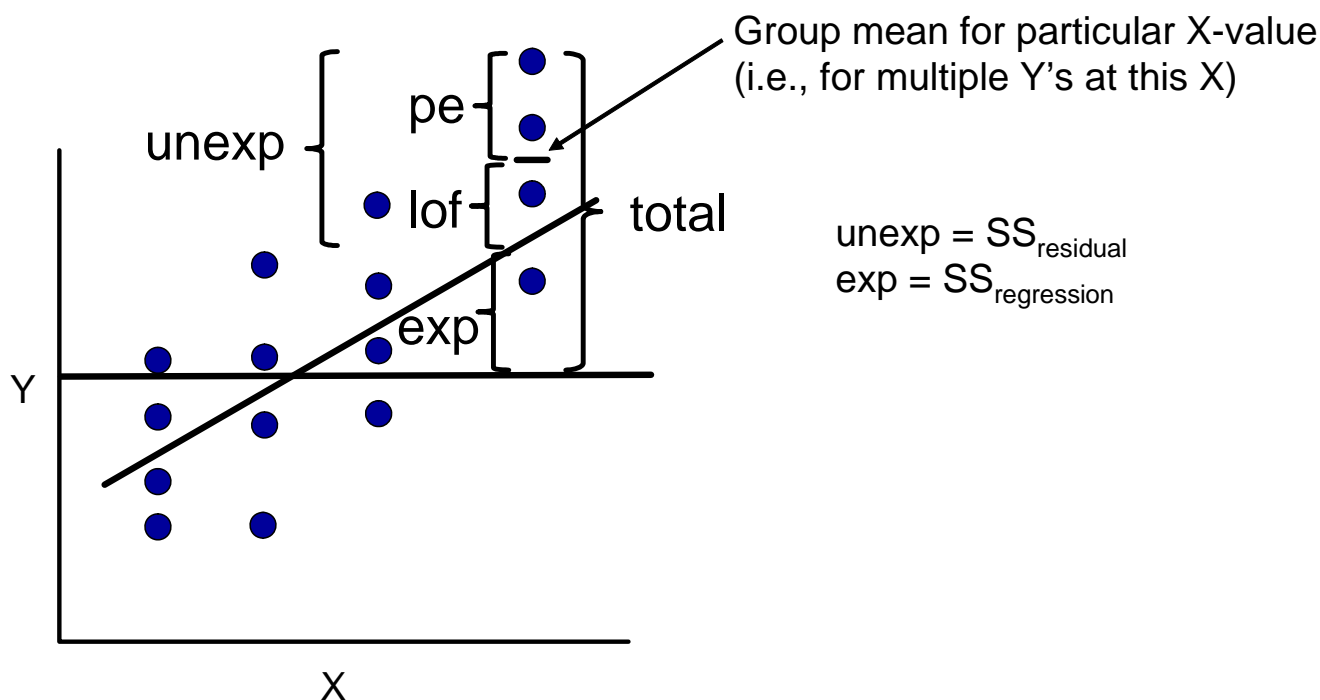
We can partition ResSS into pure error and lack of fit components

$$\text{ResSS} = \text{pure error} + \text{lack of fit}$$

$$(Y_i - \hat{Y}_i)^2 = (Y_i - \bar{Y}_{x_i})^2 + (\bar{Y}_{x_i} - \hat{Y}_i)^2$$



Differences between the Y observations and the mean \bar{Y} for that X arise from pure error. Differences between each mean \bar{Y} and the regression estimate for that X are due to lack of fit. So, we have a within-groups error and an among groups error.



We can display this in an ANOVA table:

Source	df	SS
Regression	1	$(\hat{Y}_i - \bar{Y})^2$
Lack of fit	k-2	$(\bar{Y}_{xi} - \hat{Y}_i)^2$
Pure error	n-k	$(Y_i - \bar{Y}_{xi})^2$
Total	n-1	$(Y_i - \bar{Y})^2$

***k = the number of unique groups

If H_0 is true then the $E(\bar{Y}_{xi} - \hat{Y}_i)^2 = 0$, but if the data isn't linear, then the deviations due to lack of fit will get increasingly large.

For our example:

Source	df	SS	MS	F
Regression	1	6394.02	6394.02	21.33
Lack of fit	22	5200.51	236.39	0.44
Pure error	6	3193.00	532.17	
Total	29	14787.53		

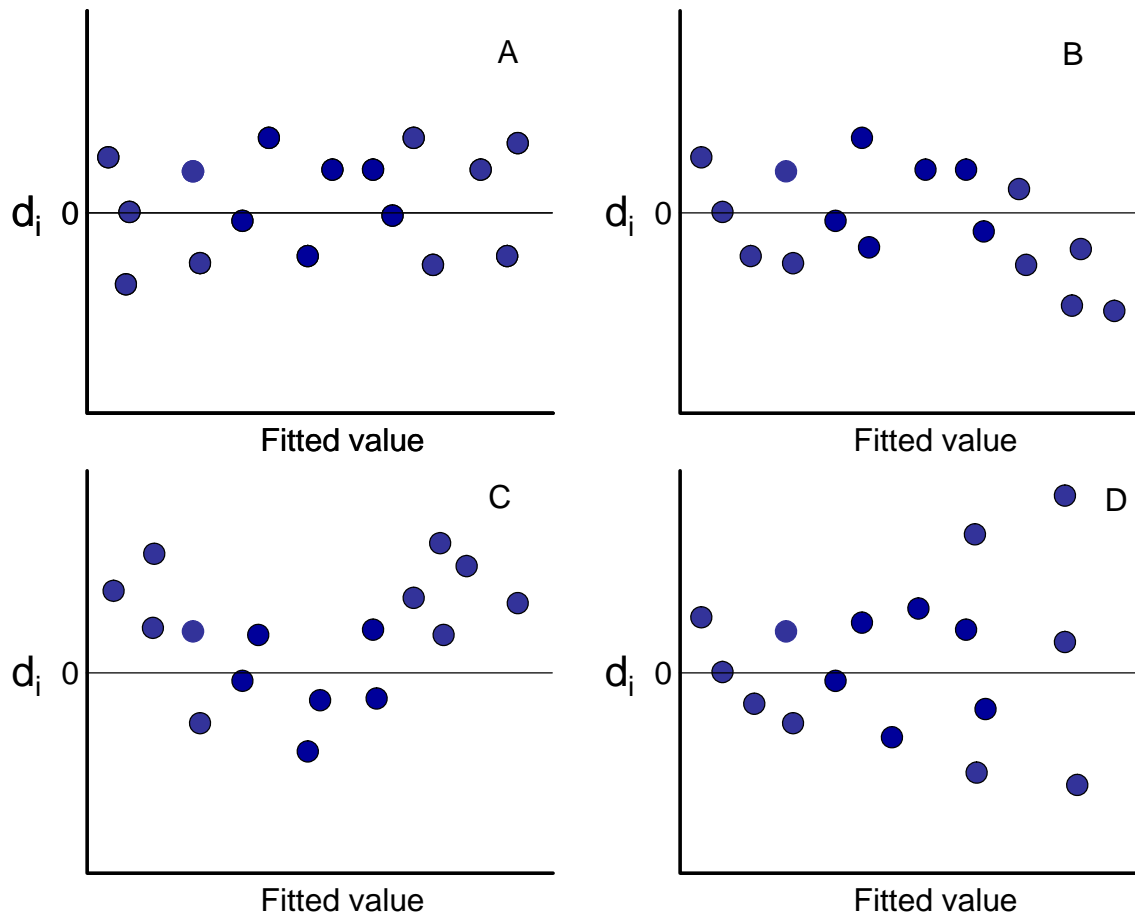
Another way to get an estimate of pure error is to arrange the data for all of your X 's as separate groups in a 1-way ANOVA (i.e., the Y -values for each X are placed in separate Excel columns). For our example data, we would have 24 groups (6 that each had 2 Y -values and 18 that each only had 1 Y -value). From the ANOVA, the within-groups SS will be your estimate of pure error for the regression model. Then you can return to your ANOVA table output from the regression model and estimate your error due to lack of fit by subtracting the pure error from the residual SS term, and then conduct your lack of fit F-test.

Regression diagnostics

The very first step in performing diagnostic evaluation of your fitted regression model should be to **plot the residuals** (d_i 's) versus the fitted values (\hat{Y} 's). Such a plot will provide clues about potential violations of several model assumptions including normality, homogeneity of variances, and the linearity of the model. The plot will also illustrate observations that may have considerable influence on the parameter estimates (e.g., outliers and leverage points).

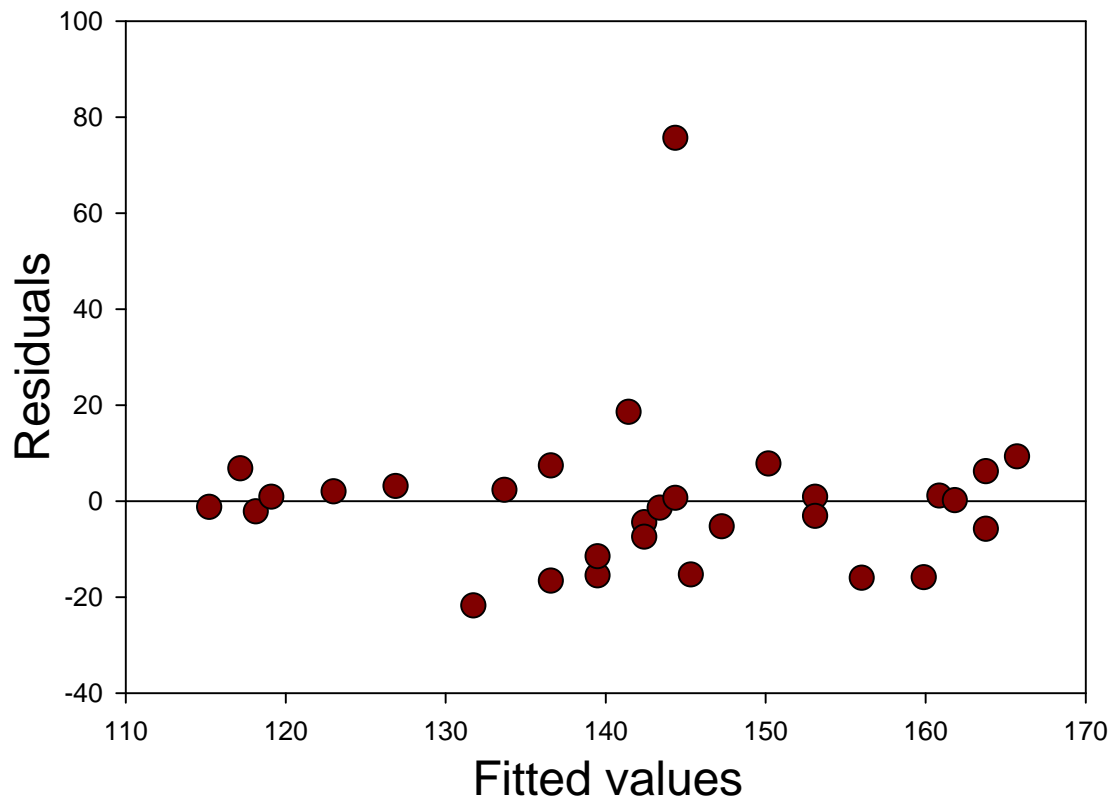
If the linear fit is appropriate and the basic assumptions of normality of errors and homogeneity of variances are not severely violated, the residuals should be centered on 0 and demonstrate no clear pattern.

Some example residual plots



Plot A above illustrates a pattern of residuals that indicates no violations of model assumptions or the presence of influential observations. Plots B and C both indicate some degree of *non-linearity* that would likely be detected during a 'lack of fit' test. Plot D indicates some degree of *heterogeneity of variances*, with the error increasing for larger fitted values. In many cases, **transformation** (e.g., logarithmic or square root, see Chapter 13 and sect. 17.10 in Zar) of the data may solve (or at least minimize) some of these problems and reduce the severity of violation of model assumptions. We can also evaluate the normality assumption graphically by plotting the distribution of residuals versus a standard normal.

Residual plot for BP vs. Age example



Identifying influential observations

A single observation that is substantially different from all other observations can make a large difference in the results of our regression analysis. If a single observation (or small group of observations) considerably changes our results, it would be good to know about this and investigate further. Keep in mind, there are **three ways** that an observation can be unusual.

Outliers: In linear regression, an outlier is an observation with a large residual. In other words, it is an observation whose response (Y-value) is unusual given its value (X) of the predictor

variable. An outlier could indicate a sample peculiarity or a data entry error or some other problem.

Leverage: An observation with an extreme value of the predictor variable (X) is called a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. Leverage points can often have large effects on the estimate of regression coefficients (slope and intercept).

Influence: An observation is said to be influential if removing the observation considerably changes the estimate of regression coefficients. Conceptually, influence can be thought of as the product (*not literally*) of leverage and outlierness.

A good way to identify potential outliers is to calculate what are called **standardized and studentized residuals**. Many times you will see these two terms used interchangeably, implying that they are the same. Alternatively, you may see them referred to as '*internally*' studentized and '*externally*' studentized residuals. For our purposes, we will refer to standardized residuals as those residuals that have been adjusted for the standard error of the regression and the leverage of the observation. We will refer to studentized residuals as a standardized residual that has been calculated using a regression standard error generated by a model with that observation *omitted*.

The standardized residual for each observation is:

$$\hat{e}_i = \frac{e_i}{s\sqrt{1-h_i}}$$

The studentized residual is then:

$$\hat{e}_i = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}$$

where e_i = residual, s and $s_{(i)}$ = standard error of the regression with and without the observation, and h_i = leverage.

We calculate leverage (h_i) as:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

We should generally inspect observations closely that have **leverage** $> (2k+2)/n$, where k = number of regression coefficients, and a **studentized residual with an absolute value** > 2 . The leverage essentially measures the influence of an observation due to its X-value, and the studentized residual measures its outlierness in the Y-direction.

We can combine these to generate an overall measure of **influence**. Some common statistics used to measure observation influence include **Cook's distance** and **DFITS**. Each truly represents a combination of our measures of leverage and outlierness.

We calculate DFITS as:

$$DFITS = r_i \sqrt{\frac{h_i}{1-h_i}}$$

where r_i = the studentized residual and h_i = leverage.

We calculate Cook's distance (D_i) as:

$$D_i = \frac{1}{k} \frac{s_{(i)}^2}{s^2} DFITS_i^2$$

Where k = the number of regression coefficients estimated, s = the standard error of the regression with the observation, and $s_{(i)}$ = the standard error of the regression without the observation.

As a general rule, we should examine observations carefully if they have DFITS values greater than **2 × the square root of k/n** , and if they have Cook's distance values greater than **4/n**.

Both DFITS and Cook's distance are general measures of influence. We can also consider more specific measures of influence that assess how each coefficient is changed by deleting individual observations. Typically, this is only done for the slope parameter in a simple linear regression. For convenience, we can call such a statistic **VarBeta**, and we should be concerned when removal of an observation causes greater than a **2 ÷ square root of n** absolute change in the parameter estimate.

The following table summarizes the general rules of thumb we can use for these statistics to identify observations worthy of further investigation (again, k is the number of coefficients (*2 for a simple linear regression*) and n is the number of observations).

Statistic	Value
leverage	$> (2k+2)/n$
abs(rstudent)	> 2
Cook's D	$> 4/n$
abs(DFITS)	$> 2*\text{sqrt}(k/n)$
abs(VarBeta)	$> 2/\text{sqrt}(n)$

Influential statistics for BP vs. Age example calculated in STATA

Age	BP	predicted Y	residual	leverage	standard	student	DFITS	Cooks D	VarBeta
17	114	115.2195	-1.2195	0.1500	-0.0764	-0.0750	-0.0315	0.0005	0.0278
19	124	117.1613	6.8387	0.1340	0.4245	0.4182	0.1645	0.0139	-0.1426
20	116	118.1321	-2.1321	0.1265	-0.1318	-0.1294	-0.0492	0.0013	0.0423
21	120	119.1030	0.8970	0.1192	0.0552	0.0542	0.0199	0.0002	-0.0169
25	125	122.9865	2.0135	0.0931	0.1221	0.1200	0.0384	0.0008	-0.0308
29	130	126.8700	3.1300	0.0717	0.1876	0.1844	0.0512	0.0014	-0.0375
34	110	131.7243	-21.7243	0.0516	-1.2884	-1.3045	-0.3043	0.0452	0.1811
36	136	133.6661	2.3339	0.0456	0.1380	0.1355	0.0296	0.0005	-0.0154
39	144	136.5787	7.4213	0.0389	0.4372	0.4308	0.0866	0.0039	-0.0327
39	120	136.5787	-16.5787	0.0389	-0.9767	-0.9759	-0.1963	0.0193	0.0741
42	124	139.4913	-15.4913	0.0348	-0.9107	-0.9079	-0.1723	0.0149	0.0352
42	128	139.4913	-11.4913	0.0348	-0.6756	-0.6689	-0.1270	0.0082	0.0259
44	160	141.4330	18.5670	0.0335	1.0908	1.0947	0.2039	0.0206	-0.0153
45	138	142.4039	-4.4039	0.0333	-0.2587	-0.2543	-0.0472	0.0012	0.0004
45	135	142.4039	-7.4039	0.0333	-0.4349	-0.4286	-0.0796	0.0033	0.0007
46	142	143.3748	-1.3748	0.0334	-0.0808	-0.0793	-0.0148	0.0001	-0.0008
47	220	144.3456	75.6544	0.0338	4.4455	8.0483	1.5064	0.3462	0.1856
47	145	144.3456	0.6544	0.0338	0.0385	0.0378	0.0071	0.0000	0.0009
48	130	145.3165	-15.3165	0.0345	-0.9003	-0.8972	-0.1697	0.0145	-0.0318
50	142	147.2582	-5.2582	0.0368	-0.3095	-0.3044	-0.0595	0.0018	-0.0183
53	158	150.1708	7.8292	0.0425	0.4621	0.4555	0.0959	0.0047	0.0445
56	154	153.0835	0.9165	0.0507	0.0543	0.0534	0.0123	0.0001	0.0072
56	150	153.0835	-3.0835	0.0507	-0.1828	-0.1796	-0.0415	0.0009	-0.0243
59	140	155.9961	-15.9961	0.0617	-0.9538	-0.9522	-0.2441	0.0299	-0.1655
63	144	159.8796	-15.8796	0.0804	-0.9564	-0.9549	-0.2823	0.0400	-0.2160
64	162	160.8504	1.1496	0.0858	0.0694	0.0682	0.0209	0.0002	0.0163
65	162	161.8213	0.1787	0.0915	0.0108	0.0106	0.0034	0.0000	0.0027
67	170	163.7630	6.2370	0.1038	0.3805	0.3746	0.1275	0.0084	0.1051
67	158	163.7630	-5.7630	0.1038	-0.3516	-0.3460	-0.1178	0.0072	-0.0970
69	175	165.7048	9.2952	0.1173	0.5714	0.5644	0.2058	0.0217	0.1741

Statistic	Value
leverage	> 0.200
abs(rstudent)	> 2
Cook's D	> 0.133
abs(DFITS)	> 0.516
abs(VarBeta)	> 0.365

For our example, we had no observations with high leverage. We did have one observation (47, 220) with a studentized residual greater than 2 (= 8.0483), a DFITS with an absolute value greater than 0.516 (= 1.5064), and a Cook's D greater than 0.133 (= 0.3462). However, note that the influence of this observation on our slope estimate was not that strong ($\text{VarBeta} = 0.1856 < 0.365$) and that there were other observations with just as much or more influence on our slope. The lesson from this finding is that leverage (distance from \bar{X}) contributes the most to the influence of an observation on parameter estimation. In our example, the large outlying observation had little leverage because its X-value was so close to \bar{X} .

Remember, that just because we identify observations with considerable influence on our model doesn't mean we discard them, it just means we should examine those observations closely for errors or some reason for its departure from the others. If we are unsure about an observation, we can always run the model with and without it and present the model output for both.

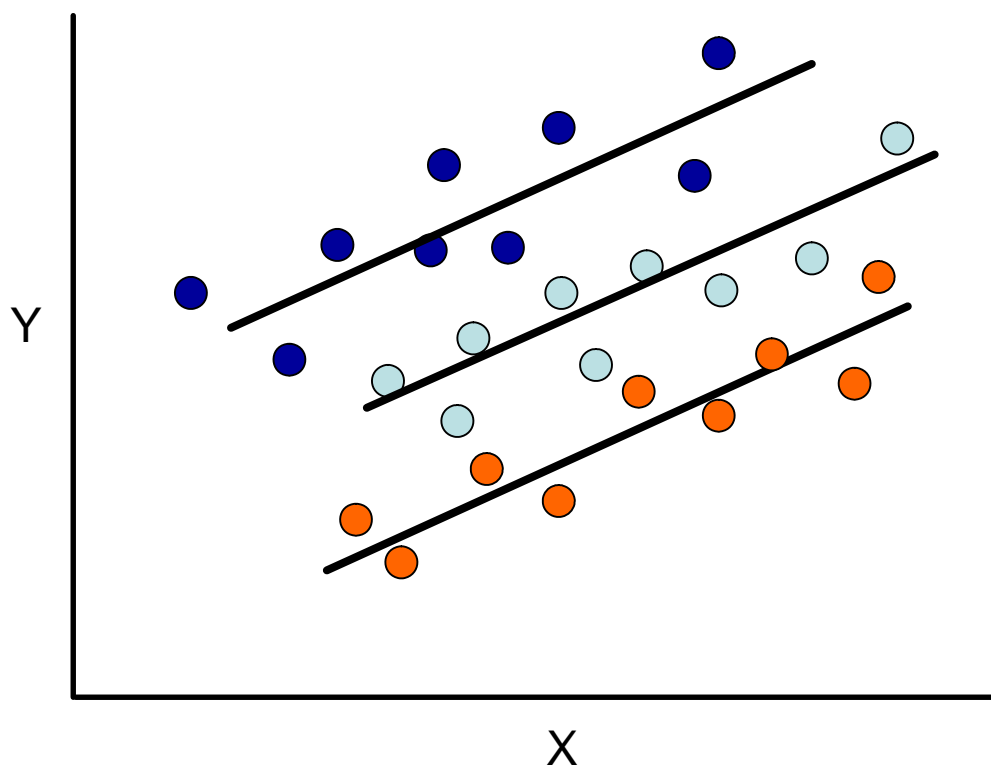


Analysis of Covariance (ANCOVA)

There will often be times when we calculate two or more regression lines from a data set and want to know whether the functional relationships described by the equations are the same. Since we now have an X-variable (predictor) in addition to our Y-variable (response) we need to use a type of joint analysis, **the analysis of covariance between X and Y**, to test our hypothesis.

The analysis of covariance (**ANCOVA**) tests a dependent variable Y for homogeneity among group means that have been *adjusted for the effect of the independent variable X*, now referred to as the **covariate**. A good way to visualize the structure of analysis of covariance is to picture each group being separately regressed on X, with all of the regression lines having a common slope (Sokal and Rohlf 1995, p. 499). The primary significance test of ANCOVA is a test of homogeneity of the adjusted means, which boils down to a test of homogeneity of elevations (**Y-intercepts**) so long as the slopes are not different. Thus, we need to test for a common slope first (*technically parallelism of slopes among the groups is an assumption of ANCOVA, though most consider it part of the test itself*).

We are hypothesizing that the value of our covariate (X) also contributes to the variation we observe in our response variable (Y) (i.e., if we don't measure the covariate, this source of variation would just be part of our pure error in the residual term). If our covariate does have an important effect, then by partitioning out this source of variation we will reduce the size of our residual error and achieve greater power to detect our treatment effects (differences among groups, which is the primary question of interest).



The model for ANCOVA is:

$$Y_{ij} = \mu + A_i + \beta_i (X_{ij} - \bar{X}_i) + \varepsilon_{ij}$$

where A_i is the treatment effect ($i = 1$ to a treatments or groups), β_i is the slope term from the regression, X_{ij} is the covariate value for observation Y_{ij} , and \bar{X}_i is the average value of the covariate for treatment group i . ε_{ij} is the error term. The above model implies that each treatment group is described by its own unique regression line (i.e., unique slope and intercept). If the slopes are not significantly different, then β_c (common slope) can be substituted for β_i . If the slope terms are not significantly different from **zero**, the model just collapses to a single-factor ANOVA. Similarly, if there is no effect of the treatment and no interaction, the model just collapses to a simple linear regression.

Let's begin with a simple test of homogeneity of slopes for the case of two regression lines. When we only have two slopes to compare, we can use a simple method that is analogous to a two-sample t-test. The test statistic is:

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}}$$

and the standard error of the difference between regression coefficients is:

$$s_{b_1 - b_2} = \sqrt{\frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_1} + \frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_2}}$$

where $s^2_{Y \cdot X} = \text{ResMS} = \sigma^2$ and the pooled ResMS is calculated as:

$$(s^2_{Y \cdot X})_p = \frac{\text{ResSS}_1 + \text{ResSS}_2}{\text{Resdf}_1 + \text{Resdf}_2}$$

As an example we'll use some data on gape height measurements for juveniles of two fish species, bluefish and striped bass. The height of the gape (when fully open) was measured and then regressed against body size for each species. We want to know if the slopes are significantly different. The ANOVA tables from the regression output for each data set are below along with some other quantities we will need.

For bluefish, we have:

Source	df	SS	MS	F-ratio	P-value
Regression	1	4769.51	4769.51	1863.93	$F_{(0.05)1,242} = 3.88$
Residual	242	619.24	2.56		$P < 0.0005$
Total	243	5388.76			

Other quantities:

$$\text{Slope} = 0.131 \quad \bar{X} = 121.72$$

$$\text{Intercept} = 2.428 \quad \bar{Y} = 18.37$$

$$\sum x^2 = SSx = 278143.05$$

$$\sum xy = SSxy = 36422.61$$

$$\sum y^2 = SSy = 5388.76$$

For striped bass, we have:

Source	df	SS	MS	F-ratio	P-value
Regression	1	1175.50	1175.50	1939.27	$F_{(0.05)1,20} = 4.35$
Residual	20	12.12	0.61		$P < 0.0005$
Total	21	1187.62			

Other quantities:

$$\text{Slope} = 0.129 \quad \bar{X} = 137.73$$

$$\text{Intercept} = -1.273 \quad \bar{Y} = 16.47$$

$$\sum x^2 = SSx = 70808.36$$

$$\sum xy = SSxy = 9123.34$$

$$\sum y^2 = SSy = 1187.62$$

So, it is clear to see that we have two highly significant regressions. Now we will compare the slopes.

Recall that we need a pooled residual MS:

$$(s^2_{Y \cdot X})_p = \frac{\text{Re } sSS_1 + \text{Re } sSS_2}{\text{Re } sdf_1 + \text{Re } sdf_2} = \frac{619.24 + 12.12}{242 + 20} = \frac{631.36}{262} = 2.41$$

Then, our standard error of the difference between slopes is:

$$s_{b_1 - b_2} = \sqrt{\frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_1} + \frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_2}} = \sqrt{\frac{2.41}{278143.05} + \frac{2.41}{70808.36}} = 0.0065$$

Our test statistic is then:

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}} = \frac{0.131 - 0.129}{0.0065} = 0.31$$

The critical value, $t_{0.05(2), 262} \approx 1.969$, so do not reject H_0 , $P > 0.50$.

We would conclude that the slopes are not significantly different. Given that a single common slope would adequately describe the rate of change in Y per unit X, we might next want to ask whether the elevations (Y-intercepts) of the lines are the same. Remember, this is really the main hypothesis test of ANCOVA. Since our regression lines are essentially parallel, we can compare Y for any single value of X to test our hypothesis, it is just simplest to use the Y-intercept ($X = 0$) since we have already calculated it.

Comparing two elevations

For our t-test, we will need to calculate some new quantities for what we shall refer to as our '**common**' regression line. These include the following:

$$\begin{aligned}\sum x_c^2 &= SSx_c = \left(\sum x^2\right)_1 + \left(\sum x^2\right)_2 \\ \sum xy_c &= SSxy_c = \left(\sum xy\right)_1 + \left(\sum xy\right)_2 \\ \sum y_c^2 &= SSy_c = \left(\sum y^2\right)_1 + \left(\sum y^2\right)_2\end{aligned}$$

The residual SS for our common regression =

$$\text{Res}SS_c = SSy_c - \frac{(SSxy_c)^2}{SSx_c}$$

The residual df for our common regression = $N - k - 1$, where k = the number the regressions we are comparing (2 in this case).

The residual MS is then $\text{Res}SS_c / \text{Res}df_c$

We can also now calculate our common slope since we've already found the slopes of our 2 groups to be statistically indistinguishable. The common slope is calculated as:

$$\beta_c = \frac{\left(\sum xy\right)_1 + \left(\sum xy\right)_2}{\left(\sum x^2\right)_1 + \left(\sum x^2\right)_2}$$

For our gape height example we have:

$$\sum x_c^2 = SSx_c = 278143.05 + 70808.36 = 348951.41$$

$$\sum xy_c = SSxy_c = 36422.61 + 9123.34 = 45545.95$$

$$\sum y_c^2 = SSy_c = 5388.76 + 1187.62 = 6576.38$$

The residual SS for our common regression =

$$\text{Res}SS_c = 6576.38 - \frac{(45545.95)^2}{348951.41} = 631.62$$

The residual df for our common regression = $266 - 2 - 1 = 263$

The **ResMS_c** is then $631.62/263 = 2.402$

Our common slope can now be calculated as:

$$\beta_c = \frac{45545.95}{348951.41} = 0.13052$$

The test statistic for comparing two elevations is calculated as:

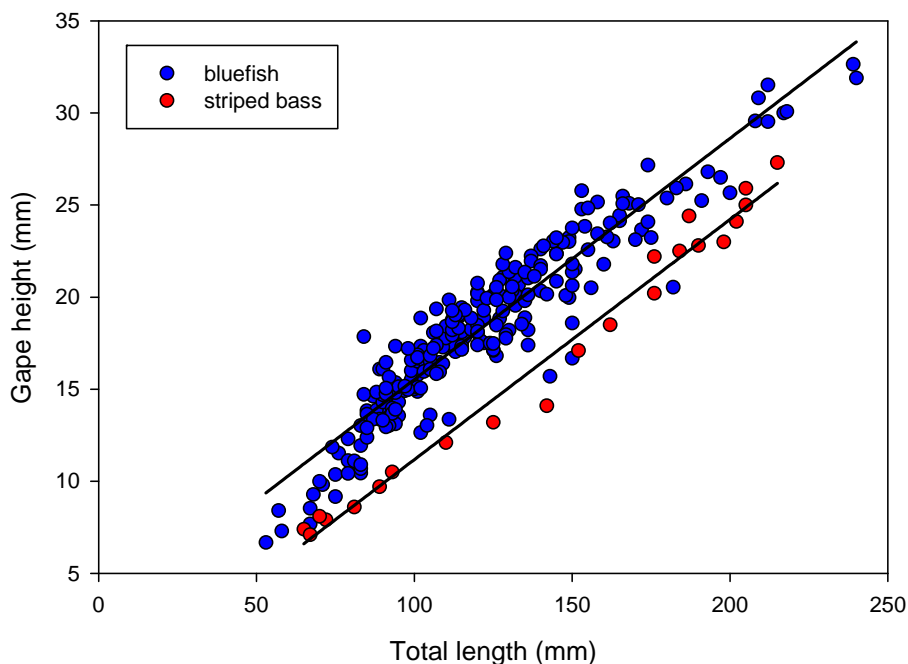
$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \beta_c (\bar{X}_1 - \bar{X}_2)}{\sqrt{\text{ResMS}_c \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{SSx_c} \right]}}$$

For our example:

$$t = \frac{(18.37 - 16.47) - 0.13052(121.72 - 137.73)}{\sqrt{2.402 \left[\frac{1}{244} + \frac{1}{22} + \frac{(121.72 - 137.73)^2}{348951.41} \right]}} = \frac{3.99}{0.348} = 11.47$$

The critical value, $t_{0.05(2), 263} \approx 1.969$, so reject H_0 , $P < 0.001$.

We would conclude that the elevations were significantly different, with bluefish having a larger gape height for a given body size compared to striped bass.

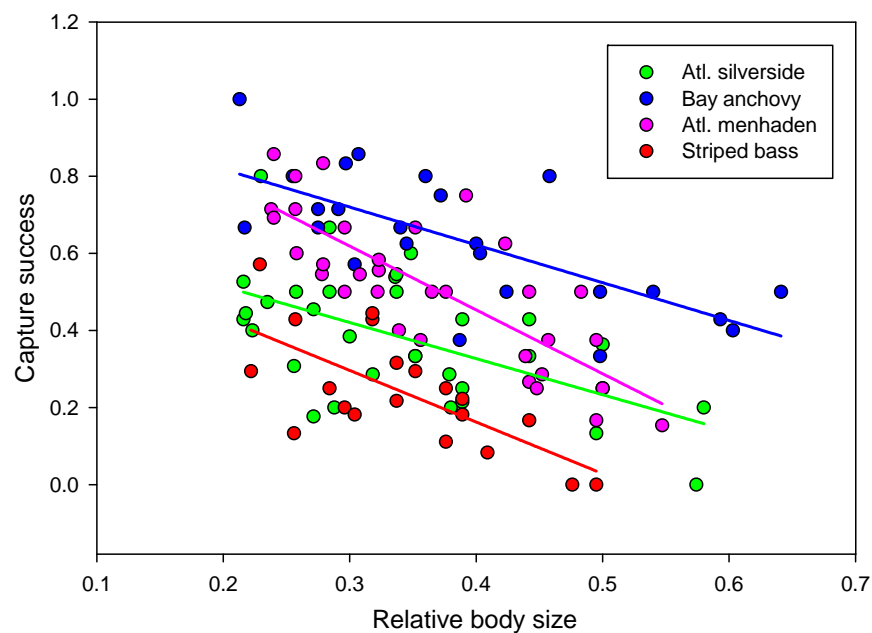


***Review examples 18.1 and 18.2 in Zar

Comparing more than two slopes and elevations

When we wish to compare more than two slopes or elevations, we can no longer use a t-test. Technically speaking, the ANCOVA procedures are only necessary once the number of groups we wish to compare exceeds 2. But, just like we can apply an ANOVA to compare 2 means, we can use the ANCOVA procedures to compare only 2 slopes or intercepts. We will use many of the quantities that we have already calculated above and then we will use an F-statistic instead of a t-test for hypothesis testing.

We will illustrate the basic calculations of ANCOVA using data for bluefish capture success when feeding on four different prey fish species. Capture success generally scales as a function of relative body size (prey size/predator size ratio), so relative body size will serve as our *covariate*. We have four significant regressions that relate bluefish capture success to relative body size for each of the different prey fish species. We wish to know if bluefish capture success differs among the prey types.



Here, we reproduce Table 18.1 from Zar to illustrate each of the quantities we need to carry out the ANCOVA procedures.

	$\sum x^2$	$\sum xy$	$\sum y^2$	Res SS	Res df
Reg 1	A_1	B_1	C_1	$SS_1 = C_1 - \frac{B_1^2}{A_1}$	$n_1 - 2$
Reg 2	A_2	B_2	C_2	$SS_2 = C_2 - \frac{B_2^2}{A_2}$	$n_2 - 2$
Reg k	A_k	B_k	C_k	$SS_k = C_k - \frac{B_k^2}{A_k}$	$n_k - 2$
Pooled Reg				$SS_p = \sum_{i=1}^k SS_i$	$N - 2k$
Common Reg	$A_c = \sum_{i=1}^k A_i$	$B_c = \sum_{i=1}^k B_i$	$C_c = \sum_{i=1}^k C_i$	$SS_c = C_c - \frac{B_c^2}{A_c}$	$N - k - 1$
Total Reg*	A_t	B_t	C_t	$SS_t = C_t - \frac{B_t^2}{A_t}$	$N - 2$

*The Total Reg row is computed after first combining the data from all k samples.

For our bluefish foraging example we have:

	$\sum x^2$	$\sum xy$	$\sum y^2$	Res SS	Res df
Atl. silverside	0.335	-0.314	0.885	0.591	31
Bay anchovy	0.346	-0.339	0.667	0.335	22
Atl. Menhaden	0.267	-0.442	1.155	0.423	31
Striped bass	0.113	-0.151	0.417	0.215	18
Pooled Reg				$SS_p = 1.564$	102
Common Reg	1.061	-1.246	3.124	$SS_c = 1.661$	105
Total Reg	1.092	-1.011	5.137	$SS_t = 4.201$	108

For our slope comparisons, we calculate a $\text{ResMS}_{\text{common}}$ and a $\text{ResMS}_{\text{pooled}}$ to use in our F-test.

$$\text{ResMS}_{\text{common}} = \frac{SS_c - SS_p}{k - 1} = \frac{1.661 - 1.564}{4 - 1} = 0.0323$$

$$\text{ResMS}_{\text{pooled}} = \frac{SS_p}{df_p} = \frac{1.564}{102} = 0.0153$$

Our F-ratio is then $0.0323/0.0153 = 2.11$ and our critical value is $F_{0.05(1), 3, 102} \approx 2.70$. We would not reject H_0 : the slopes are homogeneous ($0.10 < P < 0.25$).

For our elevation (Y-intercept) comparisons, we calculate a $\text{ResMS}_{\text{total}}$ and a $\text{ResMS}_{\text{common}}$ to use in our F-test (*note that this $\text{ResMS}_{\text{common}}$ is not the same as the one we calculated above for the slope comparisons).

$$\text{ResMS}_{\text{total}} = \frac{SS_t - SS_c}{k - 1} = \frac{4.201 - 1.661}{4 - 1} = 0.8467$$

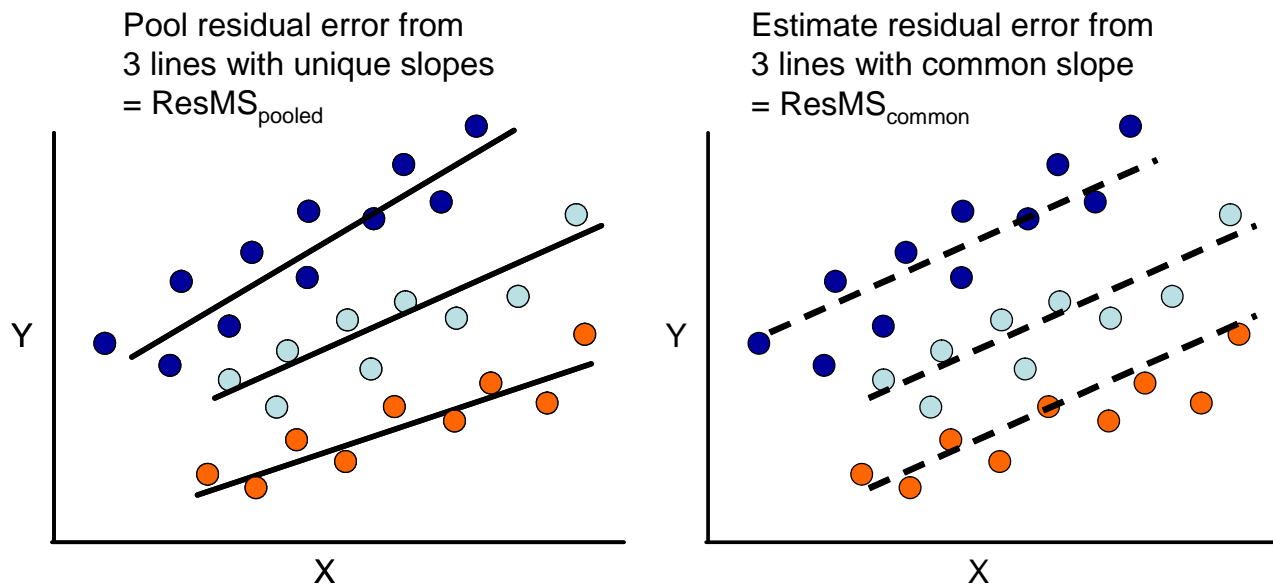
$$\text{ResMS}_{\text{common}} = \frac{SS_c}{df_c} = \frac{1.661}{105} = 0.0158$$

Our F-ratio is then $0.8467/0.0158 = 53.59$ and our critical value is $F_{0.05(1), 3, 105} \approx 2.70$. We would reject H_0 : the elevations are homogeneous ($P < 0.0005$). Bluefish capture success differs significantly among the four prey fish species. We could then proceed with multiple comparison tests (e.g., Tukey's HSD) to determine pair-wise differences between elevations.



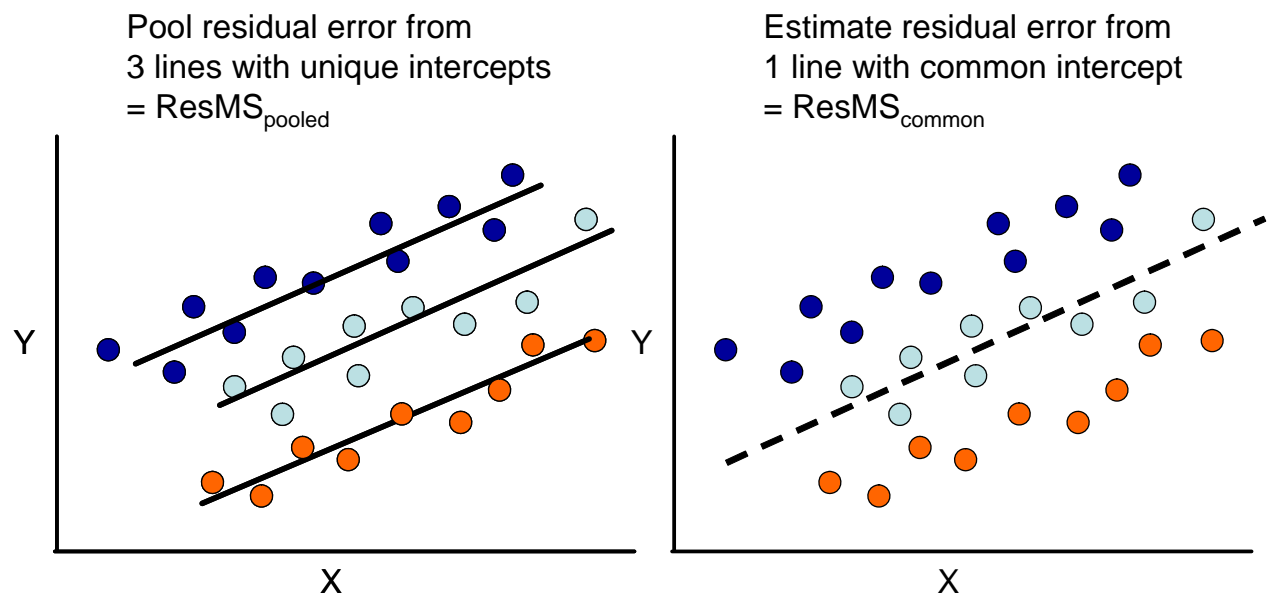
***Review example 18.4 in Zar

Visualizing the ANCOVA



Residual error around a common slope will always be greater than the pooled residual error around unique slopes.

The more the slopes differ, the greater this difference will be.



Residual error around a common intercept will always be greater than the pooled residual error around unique intercepts.

The more the intercepts differ, the greater this difference will be.

Alternative Regression Models

Polynomial Models

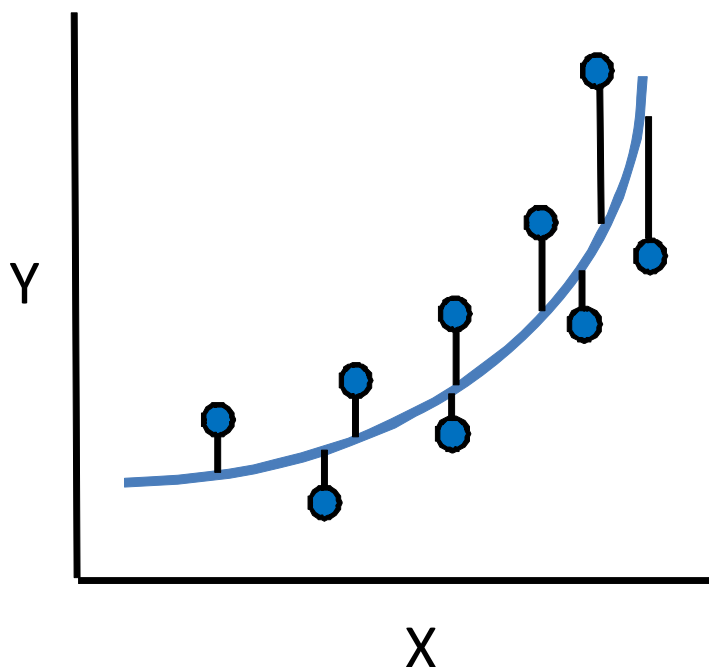
Beyond the simple linear model, the next group of models includes additional forms of the independent variable (X). This class of models is referred to as polynomial models and includes familiar parabolic curves as well as other higher order functions. The models are formed by the addition of one or more X terms that result from taking the original X term to successive powers. The basic form of the model is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_k X^k + \varepsilon$$

The simplest is the 2nd order polynomial (parabola), written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

As with our straight line model, we can use the same Least Squares methods to minimize the Residual Sums of Squares (ResSS) and estimate the parameters β_0 , β_1 , and β_2 .

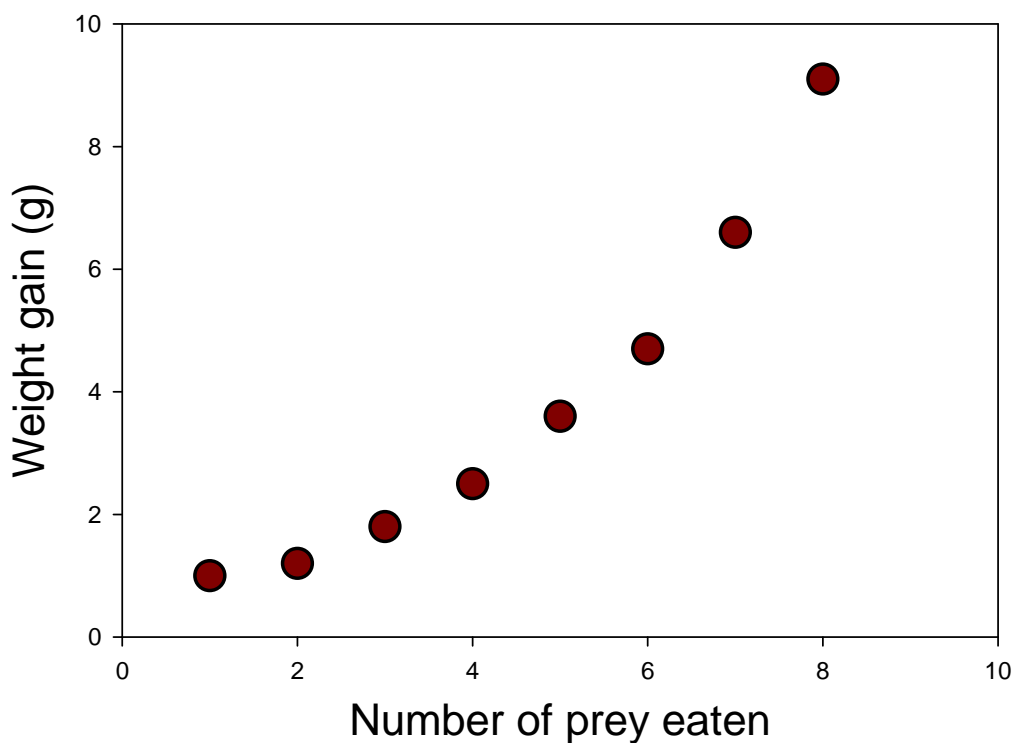


We can illustrate how the addition of a second X term affects the parameter estimates and the overall explanatory power of the model using an example.

We have some data on salamander weight gain (in grams) as a function of prey consumption (in numbers of prey eaten).

X (prey eaten) =	1	2	3	4	5	6	7	8
Y (weight gain) =	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Bivariate scatter plot



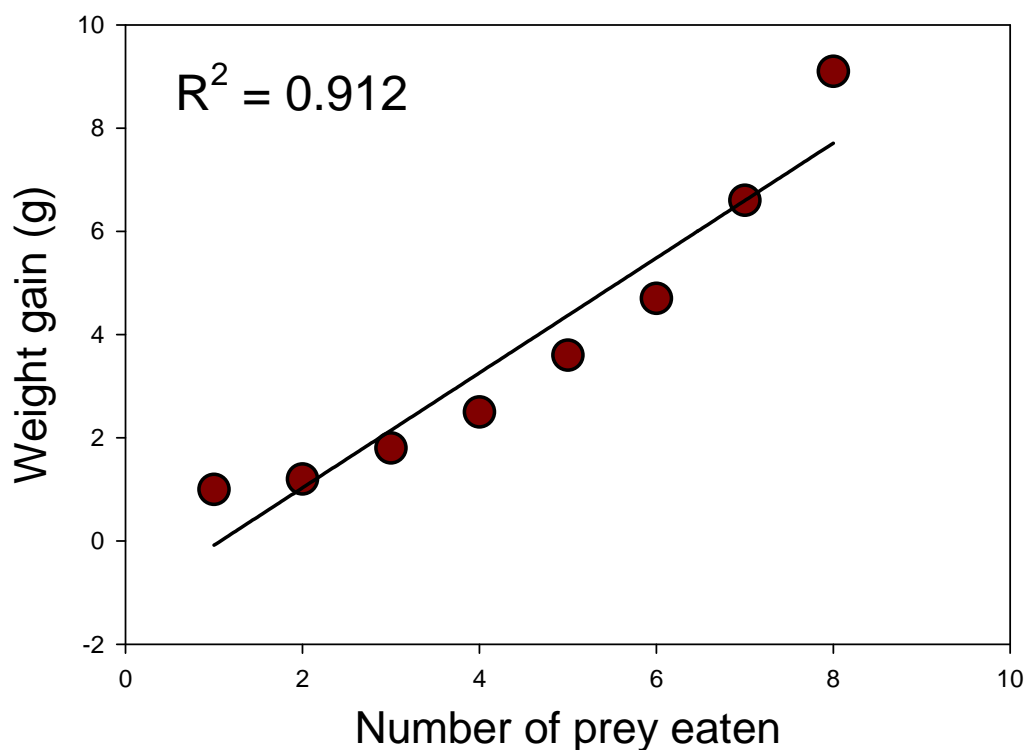
If we fit a straight-line model, we would obtain:

$$\hat{Y} = -1.20 + 1.11X$$

The ANOVA table would be:

Source	df	SS	MS	F-ratio	P-value
Regression	1	52.04	52.04	61.95	$F_{(0.05)1,6} = 5.99$
Residual	6	5.03	0.84		$P < 0.0005$
Total	7	57.07			

The R^2 value for the straight-line model = 0.912



Not bad, but the data appear curvilinear rather than forming a straight line. Since we don't have multiple Y 's for any X , we cannot formally test for lack of fit. However, we can fit a 2nd order polynomial and then test whether the additional X term significantly improves the predictive ability of the model.

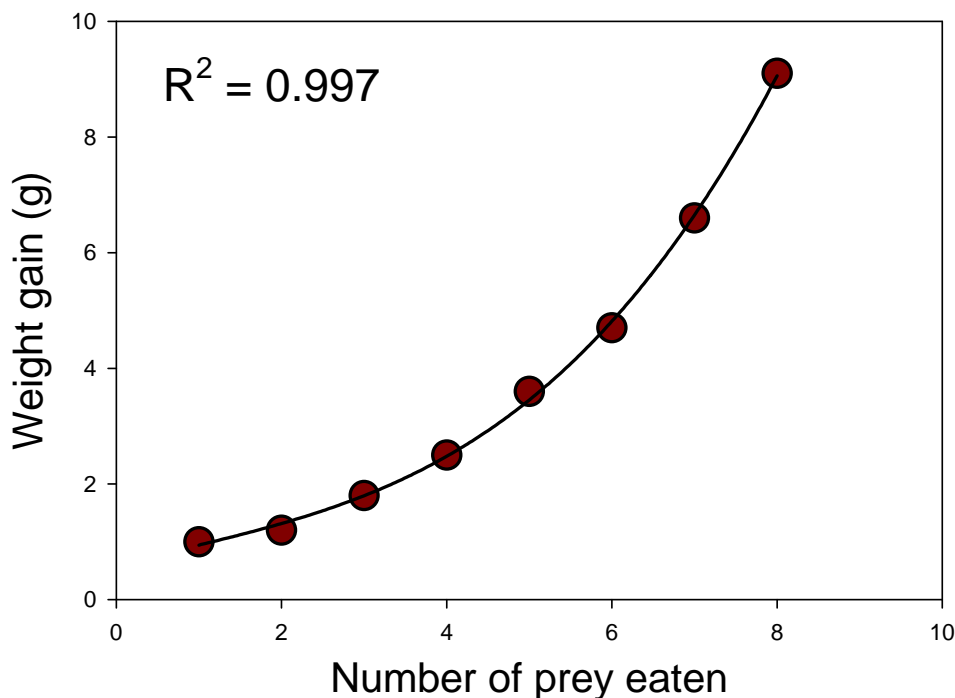
The 2nd order equation is:

$$\hat{Y} = 1.35 - 0.41X + 0.17X^2$$

The ANOVA table would now be:

Source	df	SS	MS	F-ratio	P-value
Regression	2	56.87	28.44	710.88	$F_{(0.05)2,5} = 5.79$
Residual	5	0.20	0.04		P<0.0001
Total	7	57.07			

The R^2 value for the 2nd order polynomial = 0.997



Much better, but the question now becomes whether the increase in R^2 value from 0.912 to 0.997 represents a significant increase in the predictive ability of the model. To answer this question, we can conduct what is known as a partial F-test.

First, we restructure the ANOVA table:

Source	df	SS	MS	F
Regression (X alone)	1	52.04	52.04	61.95
Regression (addition of X^2)	1	4.83	4.83	120.75
Residual	5	0.20	0.04	
Total	7	57.07		

This partitions the SS explained by the regression (RegSS) into the SS explained by X alone and the additional SS explained by adding the X^2 term to the model.

The partial F-test is conducted as:

$$F = \frac{\text{(extra SS due to adding } X^2\text{)}/1}{\text{MS residual for } 2^{\text{nd}} \text{ order model}}$$

$$F = \frac{4.83}{0.04} = 120.75$$

$F(0.05, 1, 5) = 6.61$, so we would reject H_0 with $P < 0.0005$. Hence, the addition of the X^2 term to the model significantly improves the predictive ability of the model. This means that this model is better than the straight-line model.

We can do the same test to see if adding another X term (X^3) improves the model further. In our case it actually does slightly, but there are other reasons to stick with the 2^{nd} order model (potential multicollinearity, the scatter plot suggests a 2^{nd} order model, and it's always best to use the simplest model to ease interpretation).

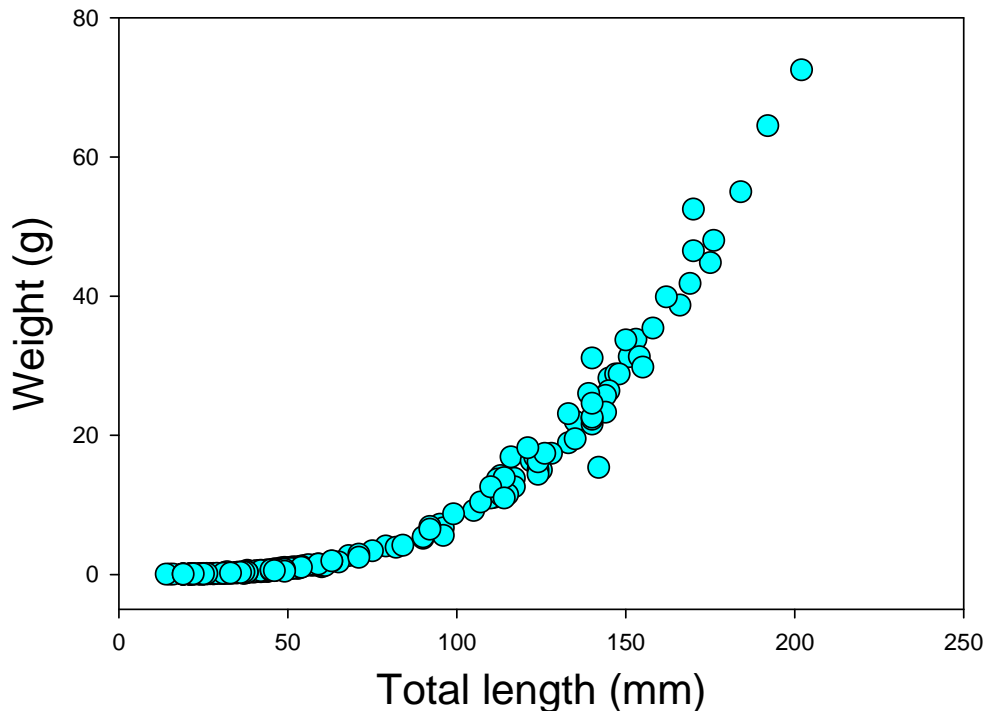
Nonlinear models

Often times, you may find that the relationship between X and Y in your data is even more complex. In other words, neither the straight-line model nor a simple polynomial model is going to be appropriate. A lack of fit test after fitting the more simple models can certainly help point you in another direction, or you may have empirical or theoretical reasons to fit another model. When computer availability was limited, great lengths were taken to try to fit nonlinear data using regression models that could be converted to a straight line model, typically by transforming the data (X, Y, or both). We still do this when we can because it lets us use fairly simple models that are easy to explain biologically. However, since computer power is no longer an issue, any number of more complex models can be fit. In addition to using the Least Squares fitting criteria, many modern techniques also use Maximum Likelihood fitting criteria. Regardless of the fitting criteria, parameter estimates for more complex models cannot be solved for algebraically, and instead require iterative approaches (essentially trial and error until the residual MS has been minimized or the likelihood function has been maximized).

Let's look at an example. We have data on weight (in grams) as a function of total length (in millimeters) for a fish (big surprise!). Some of the data is below and our $n = 231$.

<u>Total length (mm)</u>	<u>Weight (g)</u>
14	0.03
15	0.04
16	0.04
19	0.04
19	0.06
...	...

The scatter plot looks like this:

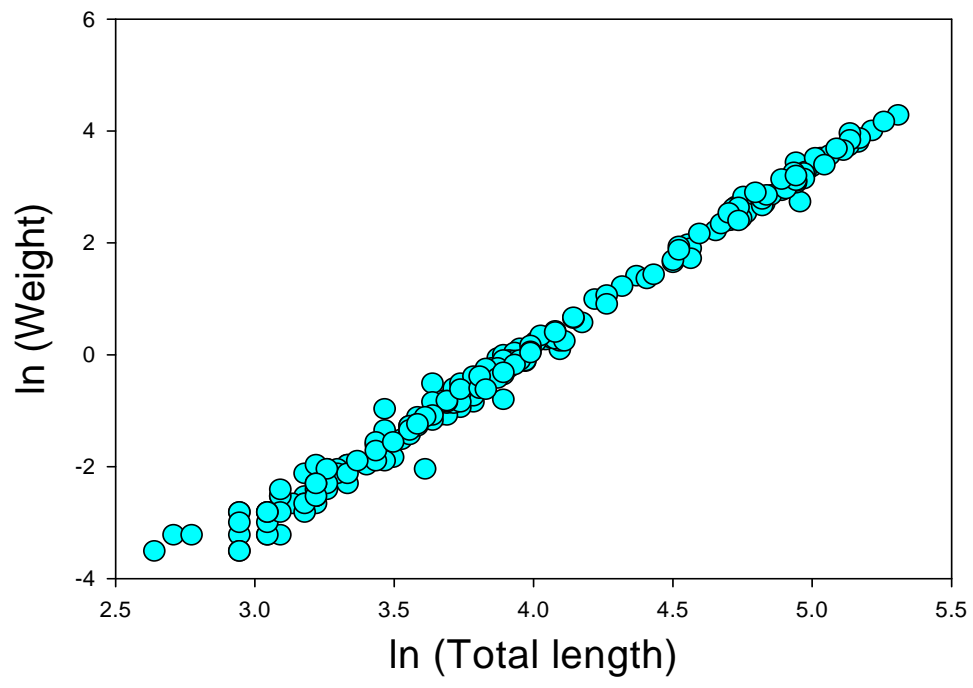


The data are clearly curvilinear, and a 2nd order polynomial probably wouldn't be a bad fit. However, empirical observations and allometric scaling theory tell us that we should fit a power function of the form:

$$\hat{Y} = \beta_0 X^{\beta_1}$$

We can fit this model in one of two ways. One way is to log-transform both the X and Y variables to convert the power function to a straight-line relationship:

$$\ln \hat{Y} = \ln \beta_0 + \beta_1 \ln X$$



Then we fit a straight-line model to estimate β_0 and β_1

$$\ln \hat{Y} = -12.64 + 3.19 \ln X$$

And our power function would then be:

$$Weight = 0.00000323 * TL^{3.19}$$

where $\beta_0 = e^{-12.64} = 0.00000323$ and $\beta_1 = 3.19$

The ANOVA table is:

Source	df	SS	MS	F-ratio	P-value
Regression	1	1019.04	1019.04	28897.66	$F_{(0.05)1,228} = 3.88$
Residual	228	8.04	0.04		$P < 0.0001$
Total	229	1027.08			

The R^2 value = 0.992

Clearly, a very strong relationship between length and weight!

The second way to obtain parameters for this relationship is to fit the raw untransformed data to the power function using a stats package with a nonlinear fitting algorithm. Recall that this will involve an iterative process to minimize the ResSS or maximize the likelihood function (I'll define this function later). Usually, the default in most stats packages is to minimize ResSS, but you can often change the fitting criteria. Also, when you use an iterative fitting process, the model needs to be seeded with starting parameter values that are in the ballpark of their actual values. The stats package will have default starting values, but the model will converge faster if it starts looking in the right place.

Here is some output from STATA, Inc. when fitting the nonlinear power function for our weight-length data:

```
nl (wt = {b0=0.000001}*tl^{b1=3})
(obs = 230)
```

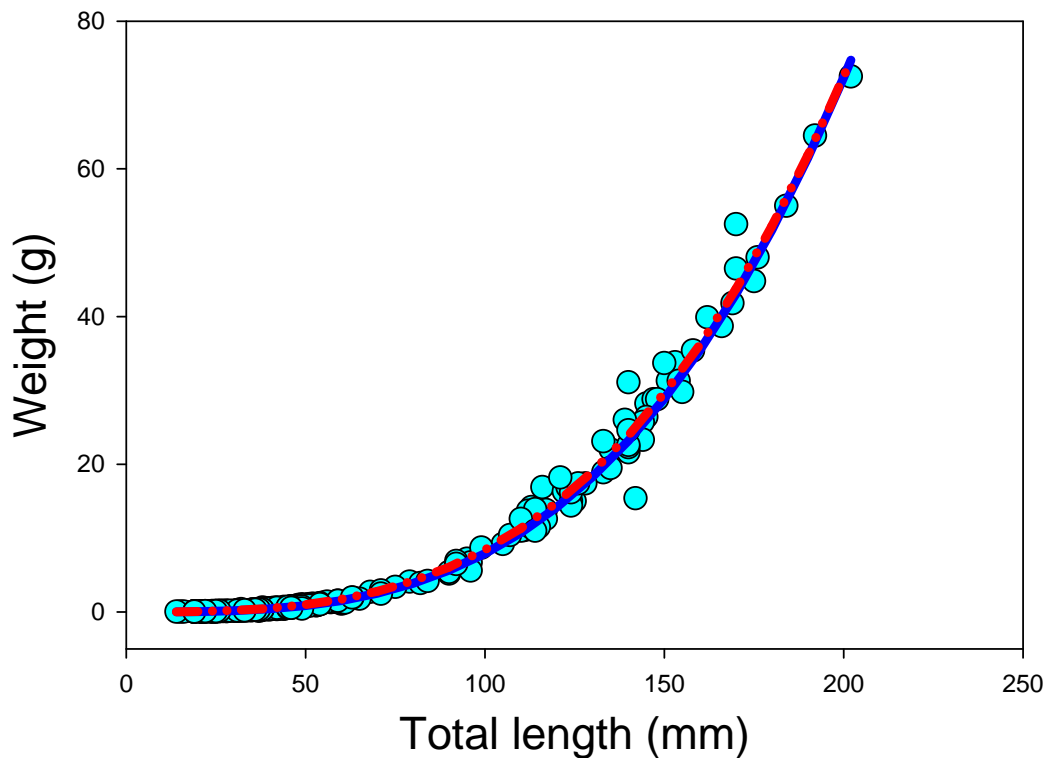
```
Iteration 0: residual SS = 5812.907
Iteration 1: residual SS = 4745.663
Iteration 2: residual SS = 665.886
Iteration 3: residual SS = 393.205
Iteration 4: residual SS = 392.7327
Iteration 5: residual SS = 392.7327
```

Source	SS	df	MS	
Model	48857.9216	2	24428.9608	Number of obs = 230
Residual	392.732703	228	1.72251185	R-squared = 0.9920
Total	49250.6543	230	214.13328	Adj R-squared = 0.9920
				Root MSE = 1.312445
				Res. dev. = 775.7732

wt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/b0	5.12e-06	8.87e-07	5.78	0.000	3.38e-06	6.87e-06
/b1	3.107703	.0341015	91.13	0.000	3.040508	3.174897

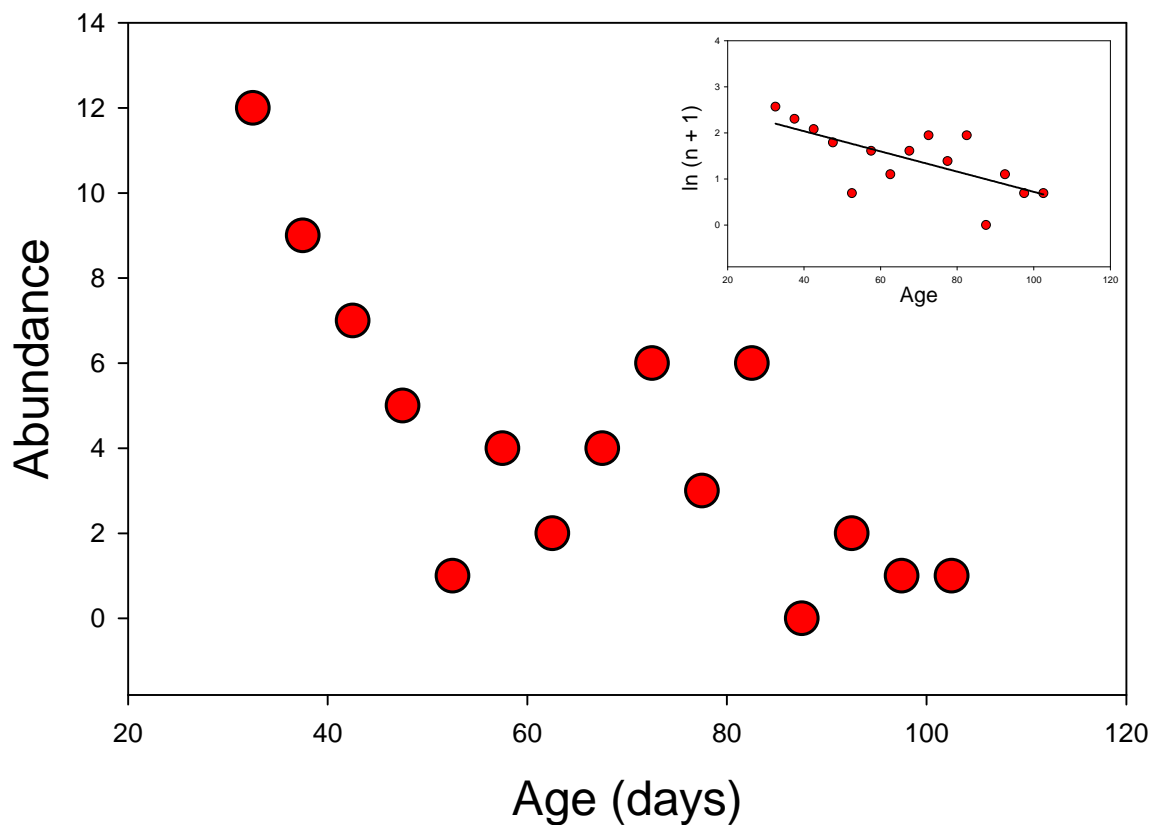
Our estimated power function is now:

$$Weight = 0.00000512 * TL^{3.11}$$



It's clear that both models fit the data very well in this example and generate nearly identical results. This won't always be the case. When sample sizes are small and variances are high, the results from the two fitting approaches can diverge to a considerable degree.

Let's look at another example. This one involves some abundance data for a fish as a function of age. We wish to estimate the mortality rate (decline in n with increasing age). The plot of the raw data is below.



Based on previous empirical observations and population dynamics theory, we expect mortality to be best described by an exponential decay function of the form:

$$\hat{Y} = \beta_0 e^{\beta_1 X}$$

If we log-transform the abundance data (Y-values), we can linearize this function:

$$\ln \hat{Y} = \ln \beta_0 + \beta_1 X$$

When we fit a straight-line model to estimate β_0 and β_1 we get:

$$\ln \hat{Y} = 2.91 - 0.022X$$

And our exponential decay function would then be:

$$Abundance = 18.41 * e^{-0.022X}$$

where $\beta_0 = e^{2.91} = 18.41$ and $\beta_1 = -0.022$

The ANOVA table is:

Source	df	SS	MS	F-ratio	P-value
Regression	1	3.36	3.36	11.70	$F_{(0.05)1,13} = 4.67$
Residual	13	3.73	0.29		P=0.0045
Total	14	7.09			

The R^2 value = 0.474

Alternatively, if we fit the exponential decay function to the untransformed data using a nonlinear iterative approach we get:

$$Abundance = 30.21 * e^{-0.033X}$$

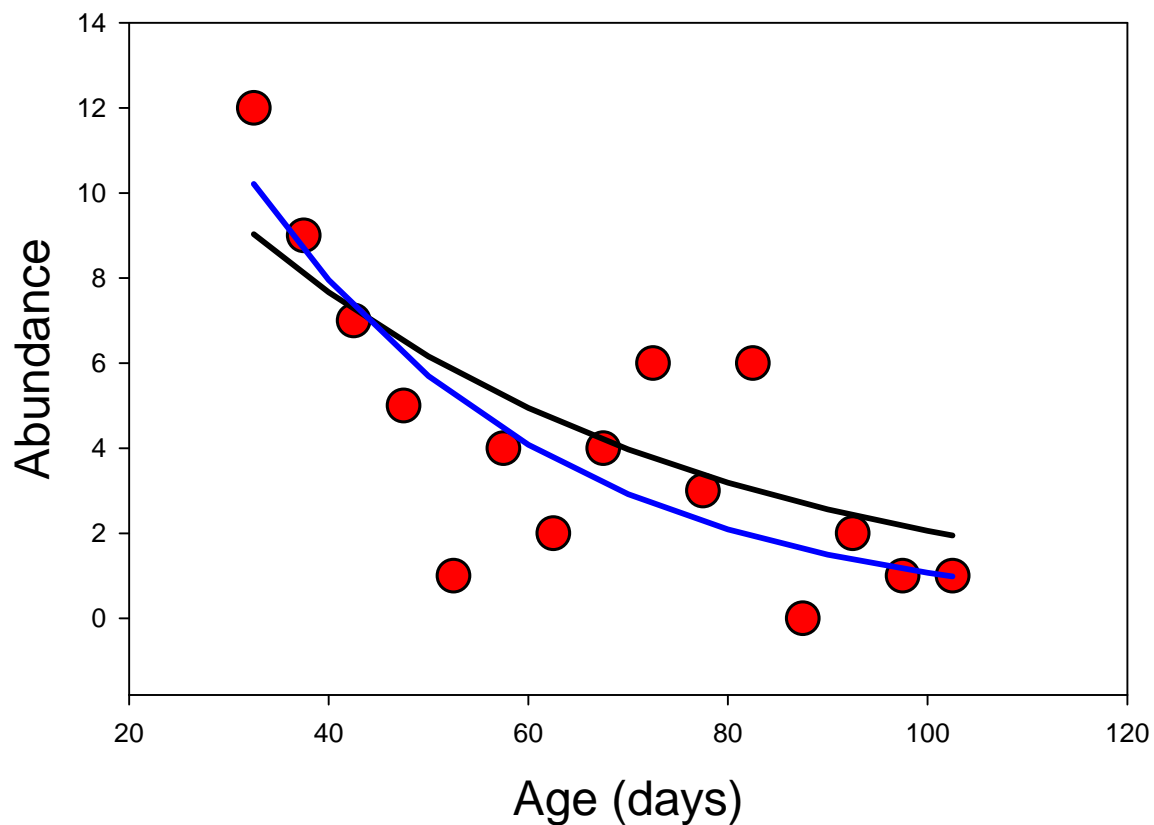
The output from STATA, Inc. is below

```
nl (n = {b0}*exp({b1}*age))
(obs = 15)
```

```
Iteration 0: residual SS = 158.4
Iteration 1: residual SS = 137.6356
Iteration 2: residual SS = 73.27781
Iteration 3: residual SS = 58.04145
Iteration 4: residual SS = 57.91858
Iteration 5: residual SS = 57.91852
Iteration 6: residual SS = 57.91852
Iteration 7: residual SS = 57.91852
Iteration 8: residual SS = 57.91852
Iteration 9: residual SS = 57.91852
```


Source	SS	df	MS	
Model	365.081478	2	182.540739	Number of obs = 15
Residual	57.9185216	13	4.4552709	R-squared = 0.8631
Total	423	15	28.2	Adj R-squared = 0.8420
				Root MSE = 2.110751
				Res. dev. = 62.83296

n	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/b0	30.20731	11.46542	2.63	0.021	5.437777	54.97684
/b1	-.03337	.0081388	-4.10	0.001	-.0509529	-.0157871



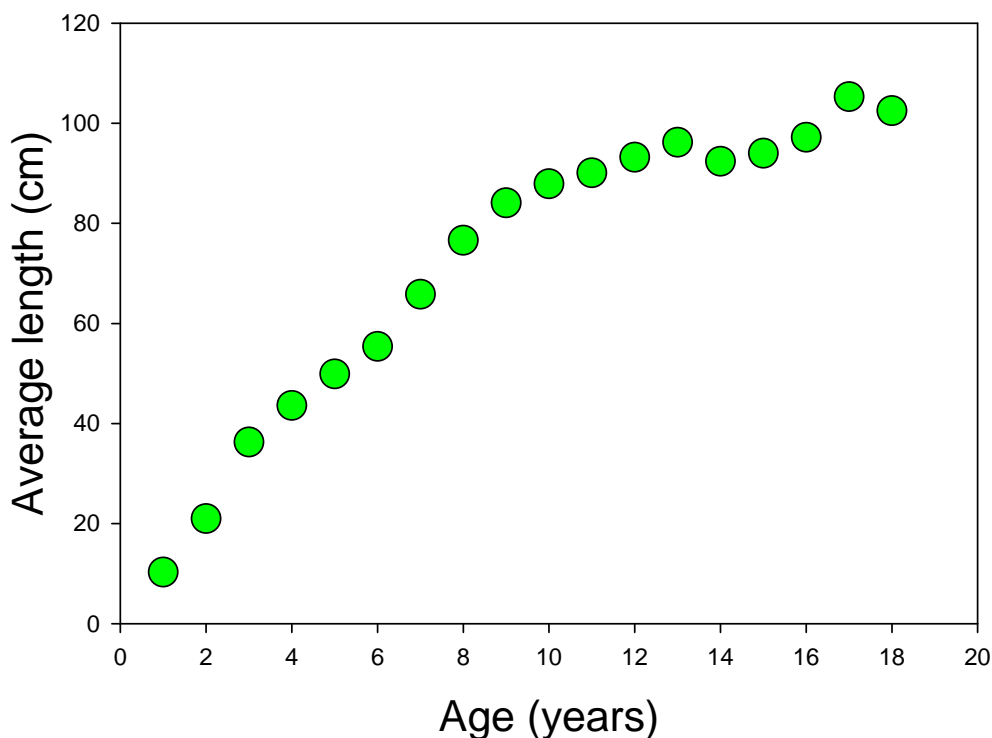
Using the nonlinear iterative approach results in a better model fit based on the elevated R^2 value, but the differences are mainly due to the combination of low sample size and high variance.

One last example for a situation when it's not possible to linearize the function we are interested in fitting.

The data are size at age date for striped bass. We wish to fit a theoretical growth curve that let's us obtain estimates of maximum size (in a lifetime) and the rate at which fish approach this size (growth coefficient). The function is called the von Bertalanffy growth curve:

$$Length = Length_{\infty} * \{1 - e^{-K(t-t_0)}\}$$

The plot looks like:



We seed the nonlinear model with starting values for each of our parameters (L_{inf} , K , and t_0) and obtain the STATA, Inc. output:

nl (basslength = {b0=100}*(1-exp(-{b1=0.05}*(bassage-{b2=0}))))
(obs = 18)

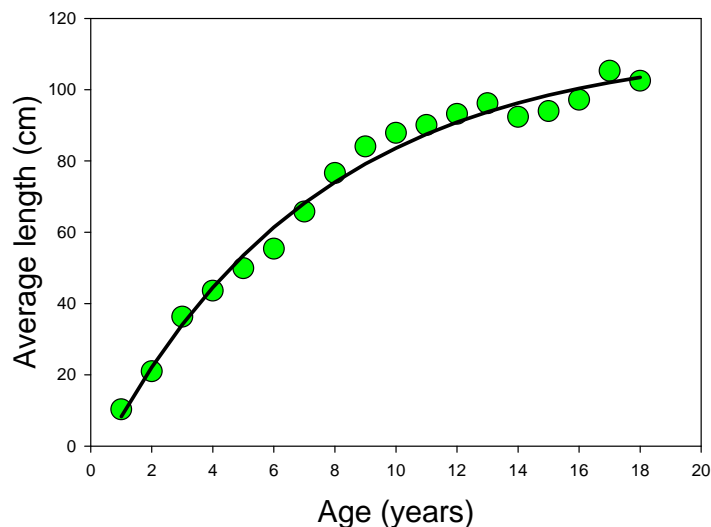
Iteration 0: residual SS = 26492.32
 Iteration 1: residual SS = 619.911
 Iteration 2: residual SS = 377.1966
 Iteration 3: residual SS = 189.1039
 Iteration 4: residual SS = 188.8788
 Iteration 5: residual SS = 188.8786
 Iteration 6: residual SS = 188.8786
 Iteration 7: residual SS = 188.8786

Source	SS	df	MS	
Model	108607.081	3	36202.3602	Number of obs = 18
Residual	188.878592	15	12.5919061	R-squared = 0.9983
Total	108795.959	18	6044.21996	Adj R-squared = 0.9979
				Root MSE = 3.548508
				Res. dev. = 93.39498

basslength	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/b0	112.8633	4.098082	27.54	0.000	104.1284	121.5981
/b1	.141702	.0147401	9.61	0.000	.1102843	.1731198
/b2	.461623	.210167	2.20	0.044	.0136628	.9095833

The parameterized model is:

$$Length = 112.86 * \{1 - e^{[-0.142(t-0.462)]}\}$$



Multiple Regression

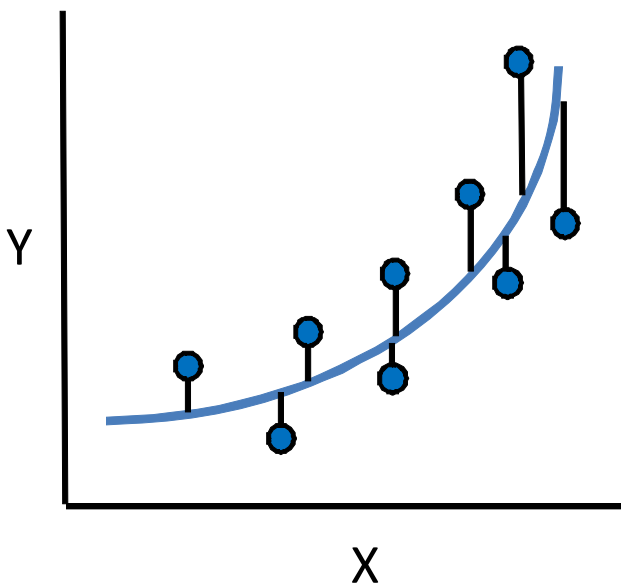
Many researchers collect data on numerous potential predictors for a single response variable and wish to analyze their overall and independent effects simultaneously (i.e., in one model). For these models, we will predict one response variable (Y) from k independent variables (X_1, X_2, \dots, X_k). The general regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_k X_k + \varepsilon$$

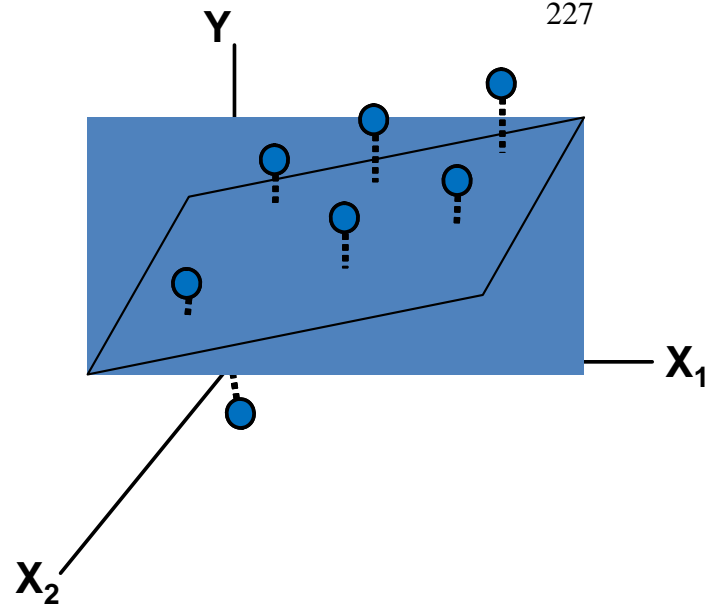
We need to note up front that, similar in some respects to the polynomial and nonlinear models we have looked at:

- 1) It is sometimes difficult to determine the best choice of model
 - Often, many reasonable candidate models will emerge
- 2) It is difficult (many times impossible) to visualize what the fitted model looks like
 - Can't plot the data when $k \geq 3$
- 3) The computations can't be done by hand (Thank God!)
 - Statistical software packages are required

When we had a single independent variable (even in the case of the polynomial models), graphical interpretation was mostly straight-forward. However, when $k \geq 2$ we are no longer dealing with a 2-dimensional line or curve, but rather with a hypersurface in $k + 1$ dimensional space. If $k \geq 3$, we can't even make a plot of the data.



2-D plot (line or curve)



3-D plot (plane or surface)

Assumptions of multiple regression

1. For each specific combination of X_1, X_2, \dots, X_k , Y is a random variable with a certain probability distribution
2. The Y observations are statistically independent
3. The mean value of Y is a linear function of X_1, X_2, \dots, X_k
4. Variances in Y are homogeneous for the range of X values
5. For any fixed X_1, X_2, \dots, X_k , Y is normally distributed

We'll start with an example: We have data on weight, length, and age for a group of 12 gray wolves. We want to build a model that will allow us to predict weight from having knowledge of length and age. This will help us to evaluate whether any particular wolf might be nutritionally deficient, through a comparison of its observed weight relative to expected weight.

The data are:

Individual	Weight (Y)	Length (X_1)	Age (X_2)
1	64	57	8
2	71	59	10
3	53	49	6
4	67	62	11
5	55	51	8
6	58	50	7
7	77	55	10
8	57	48	9
9	56	42	10
10	51	42	6
11	76	61	12
12	68	57	9

There are many possible models. For instance,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where X_1 = length and X_2 = age

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where $X_3 = X_2^2$

Or any number of more complex models. We can evaluate models, starting with a simple linear model and progressing to more complex models using our familiar partial F-test. Recall that using the 'least squares' approach, we will attempt to minimize the residual error (ResSS) for each model. We will discuss other model selection approaches shortly.

For our example, say we chose to fit the following model:

$$Weight = \beta_0 + \beta_1 Length + \beta_2 Age + \beta_3 Age^2 + \varepsilon$$

The least squares parameter estimates are:

$$\beta_0 = 3.438, \beta_1 = 0.724, \beta_2 = 2.777, \beta_3 = -0.042$$

and the corresponding model would be:

$$Weight = 3.438 + 0.724 Length + 2.777 Age - 0.042 Age^2 + \varepsilon$$

The ANOVA table for the model is:

Source	df	SS	MS	F-ratio	P-value
Regression	3	693.06	231.02	9.47	$F_{(0.05)3,8} = 4.066$
Residual	8	195.19	24.40		P=0.0052
Total	11	888.25			

The R^2 value for this three-variable model = 0.780

Although this model is significant overall, we might wish to evaluate whether each of the X variables is contributing in a meaningful way. We would start by fitting a single X variable model, then sequentially add X variables, conducting our familiar partial F-test's along the way. This approach would allow us to address the following questions:

- 1) Does knowing length significantly aid in predicting weight?
- 2) Does the addition of age significantly contribute to the prediction of weight after accounting for the contribution of length?
- 3) Does the addition of age² significantly contribute to the prediction of weight after accounting for the contribution of length and age?

Let's start with question 1:

The ANOVA table for the straight-line model with $X_1 = \text{length}$ is:

Source	df	SS	MS	F-ratio	P-value
Regression	1	588.92	588.92	19.67	$F_{(0.05)1,10} = 4.96$
Residual	10	299.33	29.93		P=0.0013
Total	11	888.25			

The R^2 value for the straight-line model = 0.663

Thus, the answer to question 1 is yes.

Moving on to question 2:

The ANOVA table for the model with $X_1 = \text{length}$ and $X_2 = \text{age}$ is:

Source	df	SS	MS	F-ratio	P-value
Regression	2	692.82	346.41	15.95	$F_{(0.05)2,9} = 4.26$
Residual	9	195.43	21.71		P=0.0011
Total	11	888.25			

The R^2 value for this two-variable model = 0.780

Also a significant overall model, but we need to conduct a partial F-test to answer the question.

The additional RegSS due to the addition of age to a model that already included length = $692.82 - 588.92 = 103.90$.

The partial F-test is conducted as:

$$F = \frac{\text{(extra SS due to adding } X_2\text{)}/1}{\text{MS residual for the two-variable model}}$$

$$F = \frac{103.90}{21.71} = 4.78$$

$F(0.05, 1, 9) = 5.12$, however $F(0.10, 1, 9) = 3.36$. Our exact P-value for this test = 0.056. Thus, our significance level is marginal, but the evidence suggests that the addition of the X_2 term (age) to the model substantially improves the predictive ability of the model, relative to the using length alone. *Note: *This is the reason that many statisticians don't like using P-values coupled with a fixed rejection threshold**

Let us keep age in the model and continue to question 3:

The ANOVA table for the model with $X_1 = \text{length}$ and $X_2 = \text{age}$ and $X_3 = \text{age}^2$ is the original table given above, which represents a significant overall model, but we need to conduct another partial F-test to answer question 3.

The additional RegSS due to the addition of age^2 to a model that already included length and age = $693.06 - 692.82 = 0.24$.

The partial F-test is conducted as:

$$F = \frac{\text{(extra SS due to adding } X_3\text{)}/1}{\text{MS residual for the three-variable model}}$$

$$F = \frac{0.24}{24.40} = 0.01$$

$F(0.05, 1, 8) = 5.32$ and our exact P-value for this test = 0.923. Thus, the evidence strongly suggests that the addition of the X_3 term (age^2) adds nothing substantial to the model.

The partitioned ANOVA table for this example would be:

Source	df	SS	MS	F
Regression (length alone)	1	588.92	588.92	19.67
Regression (addition of age)	1	103.90	103.90	4.78
Regression (addition of age^2)	1	0.24	0.24	0.01
Residual	8	195.19	24.40	
Total	11	888.25		

The best model is probably the one that includes both length and age as predictors of wolf weight, although a model that only used length as a predictor would be useful as well.

Another way to interpret the relative predictive capability of multiple independent variables is to compute standardized regression coefficients. The unstandardized regression coefficients for the full model that we presented above cannot be

compared to each other because the raw data are measured in different units (weight in kg, length in cm, and age in years). We can standardize the coefficients in two ways. One way is to standardize the data prior to running the model. For each variable, we calculate the mean and standard deviation. Each observation is standardized by subtracting it from the mean and dividing by the standard deviation. Then the regression is run on the standardized data to generate standardized coefficients. Alternatively, we can standardize the coefficients after running the model on the raw data. We simply calculate the ratio of standard deviations between the response variable (Y) and each X variable. Then multiply these ratios by the unstandardized regression coefficients to generate the standardized coefficients.

In our example, for the best model, which included just length and age (not the age² term), the standard deviations from the raw data were:

Weight (Y) = 8.986

Length (X₁) = 6.824

Age (X₂) = 1.899

<u>X variable</u>	<u>coeff.</u>	<u>Stdev ratio</u>	<u>standardized coeff.</u>
Length	0.722	0.759	0.548
Age	2.050	0.211	0.433

The standardized regression coefficients can be interpreted in a relative sense. Thus, in our data set, length was a somewhat stronger predictor of weight compared with age. If, for instance, one coefficient had been 0.50 and the other was only 0.10, you could say that one had five times the predictive capability as the other.



Alternatives to using P-values

All of the approaches that we have described in this course are based on the statistical methodology put forth by Fisher and contemporaries almost 100 years ago. The hypothesis testing methods involve stating a null hypothesis, determining an appropriate test statistic based on a theoretical sampling distribution, α -levels that are set arbitrarily, and the calculation of our P-value. We then evaluate statistical significance as yes/no decision (even though the probability distribution is continuous).

Many statisticians (and biologists) will argue that null hypothesis testing is relatively uninformative. Essentially, we discard a lot of useful information in exchange for our yes/no interpretation of a P-value. Rather, they argue, we should be evaluating multiple working hypotheses simultaneously and making inferences based on the relative weight of evidence supporting alternative models. Our question should be which hypothesis is best supported by empirical data? In other words, what is the evidence for hypothesis i ? These questions aren't new, per se, but some new theories and methods have built upon this theme. These new methods have been termed 'information-theoretic' approaches and they are based largely on Kullback-Leibler Information theory published in a seminal paper in 1951. It involves a lot of mathematical theory, but boils down to estimating the amount of information lost when using a model to approximate reality.

In 1973, Akaike discovered an efficient way to link information theory to statistical theory. He came up with a single quantity that could be used to estimate the expected Kullback-Leibler Information. The quantity is known as Akaike's Information Criterion and is abbreviated AIC. Calculation of AIC allows one to

rank multiple models from best to worst and to make inferences based on all of the models (Multimodel Inference).

Information-theoretic approaches provide:

- 1) Quantification of information loss, Δ_i .
 - This allows ranking of hypotheses, based on the data.
 - Information loss, Δ , is scaled to the best model.
- 2) Estimation of $P(H_i|X)$. The probability of a hypothesis H_i , given the data X .
- 3) Evidence ratios of model probabilities.
- 4) A framework where rigorous statistical inference can be based on all of the models in the set (multimodel inference).

Recall that for null hypothesis testing we calculate the probability of the observed data given the null, $P(X|H_0)$. Information-theoretic approaches calculate the probability of a specified hypothesis (model) given the observed data, $P(H_i|X)$. The information-theoretic approach, by its nature, places much more emphasis on hard thinking (we've heard of this before, right?) to identify and justify *a priori* a set of candidate models to evaluate.

How the information-theoretic approach works

Akaike discovered a formal relationship between Kullback-Leibler Information and the statistical Maximum Likelihood function. The Akaike Information Criterion (AIC) = $-2 \log_e(\text{ML}) + 2K$, where ML = the maximum likelihood function and K = the number of parameters estimated in the model. In practice, we generally use another term for bias correction:

AICc = $-2 \log_e (\text{ML}) + 2K + [2K*(K+1)]/n-K-1$
 where n = sample size. This corrects for bias when n is small relative to K.

For models fit using the 'Least Squares' approach, the $\log_e (ML) =$

$$\log_e (ML) = -\frac{n}{2} \log(\sigma^2)$$

where $\sigma^2 = \text{ResSS}/n$. Thus,

$$AICc = n \log_e \left(\frac{\text{ResSS}}{n} \right) + \left(\frac{2K(K+1)}{n-K-1} \right)$$

***Keep in mind that there are lots of likelihood functions; the one above is specific to least squares regression models.

Once, the AICc scores have been computed, we calculate simple differences ($\Delta_{i's}$) as $AICc_i - AICc_{\min}$

These values represent the expected Kullback-Leibler Information between the best model in your set and all other models. They are additive and make ranking the models easy.

Then, the likelihood of any particular model i , given the data can be expressed as:

$$L(\text{model } i|X) = e^{(-1/2\Delta_i)}$$

We can also measure the strength of evidence for any particular model by comparing its likelihood as a ratio of the sum of all likelihoods:

$$w_i = \frac{e^{\left(-\frac{1}{2}\Delta_i\right)}}{\sum e^{\left(-\frac{1}{2}\Delta\right)}}$$

These are referred to as Akaike weights.

If we return to our example, we had three models:

Model	ResSS	K	n
Length only	299.33	2	12
Length + Age	195.43	3	12
Length + Age + Age ²	195.19	4	12

Using the information-theoretic approach, we can calculate:

Model	AICc	Δ AICc	L	w _i
Length only	39.93	3.45	0.178	0.124
Length + Age	36.48	0	1	0.696
Length + Age + Age ²	39.18	2.70	0.259	0.180

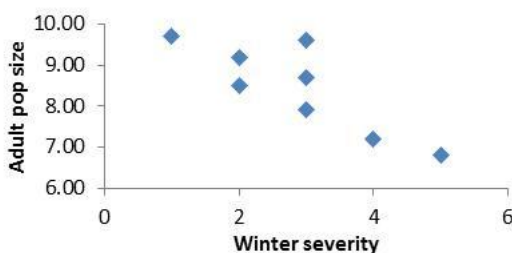
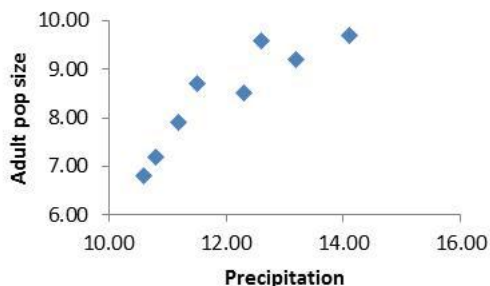
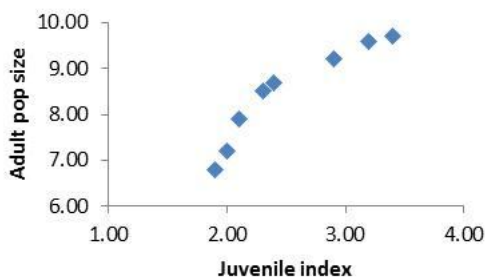
Thus, the model that included length and age has the greatest weight of evidence and is the most supported among the set of models we tested. Again, the set of models being considered is up to you, so it's best to do some hard thinking about your models rather than dropping all possible models into the 'meat grinder' and hoping that the computer will do the thinking for you.

Another example:

Below is some data for deer populations, specifically adult population size, a juvenile abundance index, annual precipitation, and winter severity. We might want to build a model to predict adult population size based on the other factors we have measured.

<u>adult pop size</u>	<u>juvenile index</u>	<u>precipitation</u>	<u>winter severity</u>
9.20	2.90	13.20	2
8.70	2.40	11.50	3
7.20	2.00	10.80	4
8.50	2.30	12.30	2
9.60	3.20	12.60	3
6.80	1.90	10.60	5
9.70	3.40	14.10	1
7.90	2.10	11.20	3

First, we take a quick look at how each of our predictor variables relates to adult population size:



Not knowing which variables will contribute most to explaining variability in adult population size, we choose to fit several models and compare them.

Here is the output that we'll need to calculate AIC scores:

<u>Model</u>	<u>RSS</u>	<u>K</u>	<u>AIC</u>	<u>delta AIC</u>
Juv index	0.961143	2	-10.5526	0
Precip	1.501297	2	-6.98489	3.567695
Winter	2.433563	2	-3.12068	7.43191
Juv + Precip	0.892516	3	-5.54522	5.007374
Juv + winter	0.597548	3	-8.7549	1.797689
Precip + winter	1.477971	3	-1.51017	9.042416
Juv + Precip + Winter	0.379909	4	-3.04479	7.5078

We see that the model which includes only the juvenile index has the lowest AIC score (least amount of information loss), and thus a delta AIC = 0. The model which includes both the juvenile index and winter severity also shows limited information loss ($\Delta AIC < 2$). Here are the model likelihoods and Akaike weights:

<u>Model</u>	<u>Likelihoods</u>	<u>Akaike weight</u>
Juv index	1.000	0.583
Precip	0.168	0.098
Winter	0.024	0.014
Juv + Precip	0.082	0.048
Juv + winter	0.407	0.237
Precip + winter	0.011	0.006
Juv + Precip + Winter	0.023	0.014
<i>sum</i>	<i>1.715</i>	

We see that the model including only juvenile index receives the most support and that the model which includes the juvenile index and winter severity is also moderately supported.

Keep in mind that these relative weights of evidence are confined to the set of models that we tested (i.e., if we added other models, these weights would change).

If no single model received overwhelming support, or if we end up with multiple models that receive good support, there are approaches to make inferences based on a set of multiple models. This is known as multimodel inference and the most common technique is model averaging, which is simply a weighted average of predictions or model parameters across all the models in our set. The Akaike weights (i.e., the individual model probabilities) are generally used for weighting.

Model validation

After we have selected a model, we would be well served to know something about its performance (i.e., its predictive capabilities). To get at this, we need an independent data set (not the one we used to construct our model). If one isn't available, modelers will often partition the data set they do have into training and validation pieces in order to quantify model performance.

A common approach is to use what is known as cross validation. One type is referred to as K-fold cross validation. This is the when the original data set is partitioned into k equal-sized subsamples. One of the subsamples is retained as the validation data set and models are trained using the other subsamples. This is repeated using each subsample as the validation data set

once. The model performances are usually averaged for each validation data set.

Two-fold cross validation is when our data can be neatly partitioned into 2 subsamples ($k = 2$). One subsample is used for training and one is used for validation, and vice versa. This works well for situations when an investigator has data from two years or two locations.

Leave-one-out cross validation is akin to a jackknife resampling approach. Each observation is left out of the model one at a time, and the model is fit using the remaining observations. Each single observation then serves as the validation data. This is the same as K-fold validation with $K = n$. This allows an investigator to quantify the performance of a model at predicting the response for each single observation. The performance is generally averaged across all of the observations.