Stat-403/Stat-650
Intermediate Sampling and Experimental Design and Analysis
2005

C. J. Schwarz
Department of Statistics and Actuarial Science, Simon Fraser University
cschwarz@stat.sfu.ca

September 11, 2005

Contents

2 Introduction to Statistics **5**

2.1 TRRGET - An overview of statistical inference 6
2.2 Parameters, Statistics, Standard Deviations, and Standard Errors 10
2.2.1 A review 10
2.2.2 Theoretical example of a sampling distribution 14
2.3 Confidence Intervals 17
2.3.1 A review 17
2.3.2 Some practical advice 24
2.3.3 Technical details 25
2.4 Hypothesis testing 26
2.4.1 A review 26
2.4.2 Technical details 29
2.4.3 Type I, Type II and Type III errors 30
2.4.4 Some practical advice 31
2.4.5 The case against hypothesis testing 33
2.4.6 Problems with p-values - what does the literature say? . . 35
Statistical tests in publications of the Wildlife Society . . 35
The Insignificance of Statistical Significance Testing . . . 36
Followups 37
2.5 Meta-data 37
2.5.1 Scales of measurement 38
2.5.2 Types of Data 39
2.5.3 Roles of data 40
2.6 Bias, Precision, Accuracy 40
2.7 Types of missing data 44
2.8 Transformations 45
2.8.1 Introduction 45
2.8.2 Conditions under which a log-normal distribution appears 47
2.8.3 ln vs log 47
2.8.4 Mean vs Geometric Mean 48
2.8.5 Back-transforming estimates, standard errors, and ci . . . 49
2.8.6 Back-transforms of differences on the log-scale 50
2.8.7 Some additional readings on the log-transform 50
2.9 Standard deviations and standard errors revisited 61

2.10 Other tidbits 63
2.10.1 Interpreting p-values 63
2.10.2 False positives vs false negatives 64
2.10.3 Specificity/sensitivity/power 64

Chapter 2

Introduction to Statistics

Statistics was spawned by the information age, and has been defined as the science of extracting information from data. Technological developments have demanded methodology for the efficient extraction of reliable statistics from complex databases. As a result, Statistics has become one of the most pervasive of all disciplines.

Theoretical statisticians are largely concerned with developing methods for solving the problems involved in such a process, for example, finding new methods for analyzing (making sense of) types of data that existing methods cannot handle. Applied statisticians collaborate with specialists in other fields in applying existing methodologies to real world problems. In fact, most statisticians are involved in both of these activities to a greater or lesser extent, and researchers in most quantitative fields of enquiry spend a great deal of their time doing applied statistics.

The public and private sector rely on statistical information for such purposes as decision making, regulation, control and planning.

Ordinary citizens are exposed to many 'statistics' on a daily basis. For example:

- "In a poll of 1089 Canadians, 47% were in favor of the constitution accord. This result is accurate to within 3 percentage points, 19 time out of 20."
- "The seasonally adjusted unemployment rate in Canada was 9.3%".
- "Two out of three dentists recommend Crest."

What does this all mean?

Our goal is not to make each student a 'professional statistician', but rather to give each student a subset of tools with which they can confidently approach many real world problems and make sense of the numbers.

2.1 TRRGET - An overview of statistical inference

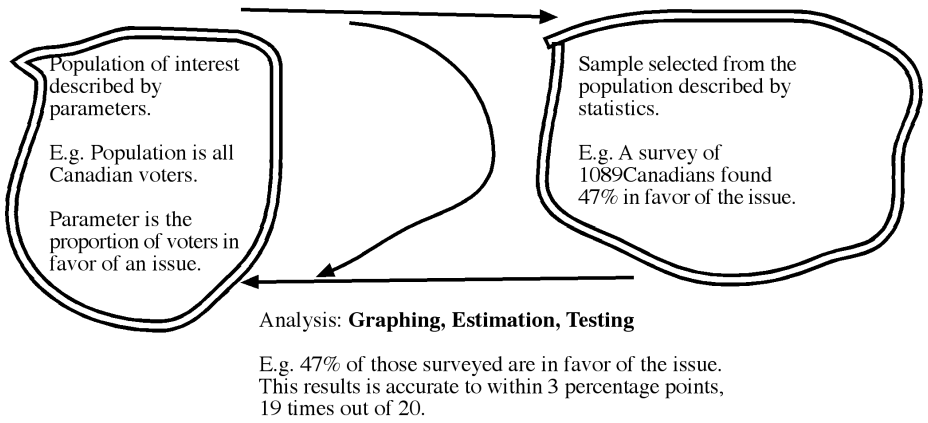
Section summary:

1. Distinguish between a population and a sample
2. Why it is important to choose a probability sample
3. Distinguish among the roles of randomization, replication, and blocking
4. Distinguish between an 'estimate' or a 'statistic' and the 'parameter' of interest.

Most studies can be broadly classified into either *surveys* or *experiments*.

In *surveys*, the researcher is typically interested in describing some population - there is usually no attempt to manipulate units within the population. In *experiments*, units from the population are manipulated in some fashion and a response to the manipulation is observed.

There are four broad phases to the survey or the experiment. These phases define the paradigm of Statistical Inference. These phases will be illustrated in the context of a political poll of Canadians on some issue as illustrated in the following diagram.



The four phases are:

1. What is the population of interest and what is the parameter of interest? This formulates the research question - what is being measured and what is of interest.

   In this case, the population of interest is likely all eligible voters in Canada and the *parameter* of interest is the proportion of all eligible voters in favor of the accord.

   It is conceivable, but certainly impractical, that every eligible voter could be contacted and their opinion recorded. You would then know the value of the parameter exactly and there would be no need to do any statistics. However, in most real world situations, it is impossible or infeasible to measure every unit in the population.

   Consequently, a sample is taken.

2. Selecting a sample

   We would like our *sample* to be as representative as possible - how is this achieved? We would like our answer from our sample to be as precise as possible - how is this achieved? And, we may like to modify our sample selection method to take into account known division of the population - how is this achieved?

   Three fundamental principles of Statistics are *randomization*, *replication* and *blocking*.

   Randomization This is the most important aspect of experimental design and surveys. It ensures that the sample is 'representative' of

the population by ensuring that, 'on average', the sample will contain a proportion of population units that is about equal, for *any variable that may influence the responses of the unit* as found in the population.

If an experiment is not randomized or a survey is not randomly collected, it rarely (if ever) provides useful information.

Many people confuse 'random' with 'haphazard'. The latter only means that the sample was collected without a plan or thought to ensure that the sample obtained is representative of the population. A truly 'random' sample takes surprisingly much effort to collect!

E.g. The Gallup poll uses random digit dialing to select at random from all households in Canada with a phone. Is this representative of the entire voting population? How does the Gallup Poll account for the different patterns of telephone use among genders within a household?

A random sample is an example of a equal probability sample. As you will see later in the notes, the assumption of equal probability not crucial - what is crucial is that every unit in the population have a known probability of selection.

Replication = Sample Size This ensures that the results from the experiment or the survey will be precise enough to be of use. A large sample size does not imply that the sample is representative - only randomization ensures representativeness.

Do not confuse replication with repeating the survey a second time.

In this example, the Gallup poll interviews about 1100 Canadians. It chooses this number of people to get a certain precision in the results.

Blocking (or stratification) In some experiments or surveys, the researcher knows of a variable that strongly influences the response. In the context of this example, there is strong relationship between the region of the country and the response.

Consequently, precision can be improved, by first *blocking* or *stratifying* the population into more homogeneous groups. Then a separate randomized survey is done in each and every stratum and the results are combined together at the end.

In this example, the Gallup poll often stratifies the survey by region of Canada. Within each region of Canada, a separate randomized survey is performed and the results are then combined appropriately at the end.

3. Data Analysis

Once the survey design is finalized and the survey is conducted, you will have a mass of information - *statistics* - collected from the population. This must be checked for errors, transcribed usually into machine readable form, and summarized.

The analysis is dependent upon **BOTH** the data collected (the sample) and the way the data was collected (the sample selection process). For example, if the data were collected using a stratified sampling design, it must be analyzed using the methods for stratified designs - you can't simply pretend after the fact that the data were collected using a simple random sampling design.

We will emphasize this point continually in this course - you must match the analysis with the design!

For example, 511 out of 1089 Canadians interviewed were in favor, i.e. 47% of our sample respondents were in favor.

4. Inference back to the Population

   Despite an enormous amount of money spent collecting the data, interest really lies in the population, not the sample. The sample is merely a device to gather information about the population.

   How should the information from the sample, be used to make inferences about the population?

   Graphing A good graph is always preferable to a table of numbers or to numerical statistics. A graph should be clear, relevant, and informative. Beware of graphs that try to mislead by design or accident through misleading scales, chart junk, or three dimensional effects.

   There a number of good books on effective statistical graphics - these should be consulted for further information. [1]

   Estimation The number obtained from our sample is an *estimate* of the true, unknown, value of the population parameter. How precise is our estimate? Are we within 10 percentage points of the correct answer?

   A good survey or experiment will report a measure of precision for any estimate.

   In this example, 511 of 1089 people were in favor of the accord. Our estimate of the proportion of all Canadian voters in favor of the accord is 511/1089=47%. These results are 'accurate to within 3 percentage points, 19 times out of 20', which implies that we are reasonably confident that the true proportion of voters in favor of the accord is between 47%-3%=44% and 47%+3%=50%.

   Technically, this is known as a 95% confidence interval - the details of which will be explored later in this chapter.

   (Hypothesis) Testing Suppose that in last month's poll (conducted in a similar fashion), only 42% of voters were in favor. Has the support increased? Because each percentage value is accurate to about 3 percentage points, it is possible that in fact there has been no change in support!.

[1] An "perfect" thesis defense would be to place a graph of your results on the overhead and then sit down to thunderous applause!

It is possible to make a more formal 'test' of the hypothesis of no change. Again, this will be explored in more detail later in this chapter.

2.2 Parameters, Statistics, Standard Deviations, and Standard Errors

Section summary:

1. Distinguish between a parameter and a statistic
2. What does a standard deviation measure?
3. What does a standard error measure?
4. How are estimated standard errors determined (in general)?

2.2.1 A review

DDTs is a very persistent pesticide. Once applied, it remains in the environment for many years and tends to accumulate up the food chain. For example, birds which eat rodents which eat insects which ingest DDT contaminated plants can have very high levels of DDT and this can interfere with reproduction. [This is similar to what is happening in the Great Lakes where herring gulls have very high levels of pesticides or what is happening in the St. Lawrence River where resident beluga whales have such high levels of contaminants, they are considered hazardous waste if they die and wash up on shore.] DDT has been banned in Canada for several years, and scientists are measuring the DDT levels in wildlife to see how quickly it is declining.

The Science of Statistics is all about measurement and variation. If there was no variation, there would be no need for statistical methods. For example, consider a survey to measure DDT levels in gulls on Triangle Island off the coast of British Columbia, Canada. If all the gulls on Triangle Island had exactly the same DDT level, then it would suffice to select a single gull from the island and measure its DDT level.

Alas, the DDT level can vary by the age of the gull, by where it feeds and a host of other unknown and uncontrollable variables. Consequently the average DDT level over ALL gulls on Triangle Island seems like a sensible measure of the pesticide load in the population. We recognize that some gulls may have

levels above this average, some gulls below this average, but feel that changes in the average DDT level are indicative of the health of the population.

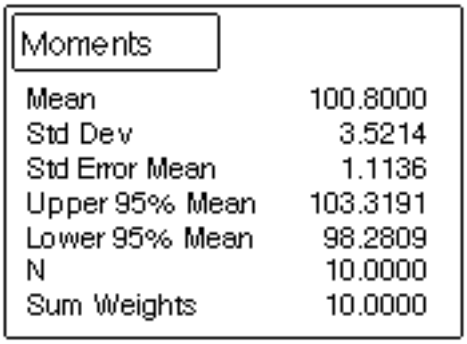
Population mean and population standard deviation. Conceptually, we can envision a listing of the DDT levels of each and every gull on Triangle Island. From this listing, we could conceivably compute the true population average and compute the (population) standard deviation of the DDT levels. [Of course in practice these are unknown and unknowable.] Statistics often uses Greek symbols to represent the theoretical values of population parameters. In this case, the population mean is denoted by the Greek letter *mu* (μ) and the population standard deviation by the Greek letter *sigma* (σ). The population standard deviation measures *the variation of individual measurements about the mean in the population.*

In this example, μ would represent the average DDT over all gulls on the island, and σ would represent the variation of values around the population mean. Both of these values are unknown.

Scientists took a random sample (how was this done?) of 10 gulls and found the following DDT levels.

```
100, 105, 97, 103, 96, 106, 102, 97, 99, 103.
```

Sample mean and sample standard deviation The sample average and sample standard deviation could be computed from these value using a spread sheet, calculator, or a statistical package. Here is the output from JMP:



A different notation is used to represent sample quantities. In this case the sample mean, denoted $\overline{Y}$ and pronounced Y-bar, has the value of 100.8 ppm, and the sample standard deviation, denoted using the letter s, has the value of 3.52 ppm. The sample mean is a measure of the middle of the sample data and the sample standard deviation measures the variation of the sample data around the sample mean.

What would happen if a different sample of 10 gulls was selected? It seems reasonable that the sample mean and sample standard deviation would also change among samples, and we hope that if our sample is large enough, that the change in the statistics would not be that large.

Here is the data from an additional 8 samples, each of size 10:

											Sample	
Set	DDT levels in the gulls										mean	std
1	102	102	103	95	105	97	95	104	98	103	100.4	3.8
2	100	103	99	98	95	98	94	100	90	103	98.0	4.1
3	101	96	106	102	104	95	98	103	108	104	101.7	4.2
4	101	100	99	90	102	99	105	92	100	102	99.0	4.6
5	107	98	101	100	100	98	107	99	104	98	101.2	3.6
6	102	102	101	101	92	94	104	100	101	97	99.4	3.8
7	94	101	100	100	96	101	100	98	94	98	98.2	2.7
8	104	102	97	104	97	99	100	100	109	102	101.4	3.7

Note that the statistics ($\overline{Y}$ - the sample mean, and s - the sample standard deviation) change from sample to sample. This is not unexpected as it highly unlikely that two different samples would give identical results.

What does the variation in the sample mean over repeated samples from the same population tell us? For example, based on the values of the sample mean above, could the true population mean DDT over all gulls be 150 ppm? Could it be 120 ppm? Could it be 101 ppm? Why?

If more and more samples were taken, you would end up with a a large number of sample means. A histogram of the **sample means** over the repeated samples could be drawn. This would be known as the **sampling distribution** of the sample mean.

The latter result is a key concept of statistical inference and can be quite abstract because, in practice, you never see the sampling distribution. The distribution of individual values over the entire population can be visualized; the distribution of individual values in the particular sample can be examined directly as you have actual data; the hypothetical distribution of a statistics over repeated samples from the population is always present, but remains one level of abstraction away from the actual data.

Because the sample mean varies from sample to sample, it is theoretically possible to compute a standard deviation of the statistic as it varies over all possible samples drawn from the population. This is known as the **standard error** (abbreviated *SE*) of the statistic (in this case it would be the standard error of the sample mean).

Because we have repeated samples in the gull example, we can compute the actual standard deviation of the sample mean over the 9 (the original sample, plus the additional 8 sample means). This gives an estimated standard error of 1.40 ppm. This measures the variability of the statistic ($\overline{Y}$) over repeated samples from the same population.

But - unless you take repeated samples from the same population, how can the standard error ever be determined? For example, refer back to the output from JMP. How was the value of 1.1136 determined for the standard error of the mean?

Now statistical theory comes into play. Every statistic varies over repeated samples. In some cases, it is possible to derive from statistical theory, how much the statistic will vary from sample to sample. In the case of the sample mean for a sample selected at random from any population, the *se* is theoretically equal to:

$$se(\overline{Y}) = \frac{\sigma}{\sqrt{n}}$$

Note that every statistic will have a **different theoretical formula for its standard error and the formula will change depending on how the sample was selected**.

But this theoretical standard error depends upon an unknown quantity (the theoretical population standard deviation σ). It seems sensible to estimate the standard error by replacing the value of σ by an estimate - the sample standard deviation s. This gives:

$$\textit{Estimated Std Error Mean} = s/\sqrt{n} = 3.5214/\sqrt{10} = 1.1136 \textit{ ppm}.$$

This number is an estimate of the variability of $\overline{Y}$ in repeated samples of the same size selected at random from the same population.

A Summary of the crucial points:

- **Parameter** The parameter is a numerical measure of the entire population. Two common parameters are the population mean (denoted by μ) and the population standard deviation (denoted by σ). The population standard deviation measures the variation of individual values over all units in the population. Parameters *always* refer to the population, never to the sample.
- **Statistics or Estimate:** A statistic or an estimate is a numerical quantity computed from the SAMPLE. This is only a guess as to the true value of the population parameter. If you took a new sample, your estimate computed from the second sample, would be different than the value computed form the first sample. Two common statistics are the sample

mean (denoted $\overline{Y}$), and the sample standard deviation (denotes s). The sample standard deviation measures the variation of individual values over the units in the sample. Statistics *always* refer to the sample, never to the population.

- **Sampling distribution** Any statistic or estimate will change if a new sample is taken. The distribution of the statistic or estimate over repeated samples from the same population is known as the sampling distribution.
- **Theoretical Standard error:** The variability of the estimate over all possible repeated samples from the population is measured by the standard error of the estimate. This is a theoretical quantity and could only be computed if you actually took all possible samples from the population.
- **Estimated standard error** Now for the hard part - you typically only take a single sample from the population. But, based upon statistical theory, you know the form of the theoretical standard error, so you can use information from the sample to estimate the theoretical standard error. Be careful to distinguish between the standard deviation of individual values in your sample and the estimated standard error of the statistic. The formula for the estimated standard error is different for every statistic and also depends upon the way the sample was selected. **Consequently it is vitally important that the method of sample selection and the type of estimate computed be determined carefully before using a computer package to blindly compute standard errors.**

The concept of a standard error is the MOST DIFFICULT CONCEPT to grasp in statistics. The reason that it is so difficult, is that there is an extra layer of abstraction between what you observe and what is really happening. It is easy to visualize variation of individual elements in a sample because the values are there for you to see. It is easy to visualize variation of individual elements in a population because you can picture the set of individual units. But it is difficult to visualize the set of all possible samples because typically you only take a single sample, and the set of all possible samples is so large.

2.2.2 Theoretical example of a sampling distribution

Here is more detailed examination of a sampling distribution where the actual set of all possible samples can be constructed. It shows that the sample mean is unbiased and that its standard error computed from all possible samples matches that derived from statistical theory.

Suppose that a population consisted of five mice and we wish to estimate the average weight based on a sample of size 2. [Obviously, the example is hopelessly simplified compared to a real population and sampling experiment!]

Normally, the population values would not be known in advance (because then why would you have to take a sample?). But suppose that the five mice had weights (in grams) of:

```
33, 28, 45, 43, 47.
```

The population mean weight and population standard deviation are found as:

- μ = (33+28+45+43+47) = 39.20 g and
- σ = 7.39 g.

The population mean is the average weight over all possible units in the population. The population standard deviation measures the variation of individual weights about the mean, over the population units.

Now there are 10 possible samples of size two from this population. For each possible sample, the sample mean and sample standard deviation are computed as shown in the following table.

Sample units		Sample Mean ($\overline{Y}$)	Sample std dev (s)
33	28	30.50	3.54
33	45	39.00	8.49
33	43	38.00	7.07
33	47	40.00	9.90
28	45	36.50	12.02
28	43	35.50	10.61
28	47	37.50	13.44
45	43	44.00	1.41
45	47	46.00	1.41
43	47	45.00	2.83
Average		39.20	7.07
Std dev		4.52	4.27

This table illustrates the following:

- this is a theoretical table of all possible samples of size 2. Consequently it shows the actual sampling distribution for the statistics $\overline{Y}$ and s. The sampling distribution of $\overline{Y}$ refers to the variation of $\overline{Y}$ over all the possible samples from the population. Similarly, the sampling distribution of s refers to the variation of s over all possible samples from the population.

- some values of $\overline{Y}$ are above the population mean, and some values of $\overline{Y}$ are below the population mean. We don't know for any single sample if we are above or below the true value of the population parameter. Similarly, values of s (which is a sample standard deviation) also varies above and below the population standard deviation.

- the average (expected) value of $\overline{Y}$ over all possible samples is equal to the population mean. We say such estimators are **unbiased**. **This is the hard concept!** The extra level of abstraction is here - the statistic computed from an individual sample, has a distribution over all possible samples, hence the sampling distribution.

- the average (expected) value of s over all possible samples is NOT equal to the population standard deviation. We say that s is a **biased** estimator. This is a difficult concept - you are taking the average of an estimate of the standard deviation. The average is taken over possible values of s from all possible samples. The latter is an extra level of abstraction from the raw data. [There is nothing theoretically wrong with using a biased estimator, but most people would prefer to use an unbiased estimator. It turns out that the bias in s decreases very rapidly with sample size and so is not a concern.]

- the standard deviation of $\overline{Y}$ refers to the variation of $\overline{Y}$ over all possible samples. We normally call this the **standard error** of a statistic. [The term comes from an historical context that is not important at this point.]. Do not confuse the standard error of a statistic with the sample standard deviation or the population standard deviation. The standard error measures the variability of a statistic (e.g. $\overline{Y}$) over all possible samples. The sample standard deviation measures variability of individual units in the sample. The population standard deviation measures variability of individual units in the population.

- if the previous formula for the theoretical standard error was used in this example it would fail to give the correct answer:

  i.e. $se(\overline{Y}) = 4.52 \neq \frac{\sigma}{\sqrt{n}} = \frac{7.39}{\sqrt{2}} = 5.22$ The reason that this formulae didn't work is that the sample size was an appreciable fraction of the entire population. A finite population correction needs to be applied in these cases. As you will see in later chapters, the se in this case is computed as:

$$se(\overline{Y}) = \frac{\sigma}{\sqrt{n}}\sqrt{(1-f)}\sqrt{\frac{N}{(N-1)}} = \frac{7.39}{\sqrt{2}}\sqrt{(1-\frac{2}{5}}\sqrt{\frac{5}{4}} = 4.52$$

  Refer to the chapter on survey sampling for more details.

2.3 Confidence Intervals

Section summary:

1. Understand the general logic of why a confidence interval works
2. How to graph a confidence interval for a single parameter
3. How to interpret graphs of several confidence intervals
4. Effect of sample size upon the size of a confidence interval
5. Effect of variability upon the size of a confidence interval
6. Effect of confidence level upon the size of a confidence interval

2.3.1 A review

The basic premise of statistics is that every unit in a population cannot be measured – consequently, a sample is taken. But the statistics from a sample will vary from sample to sample and it is highly unlikely that the value of the statistic will equal the true, unknown value of the population parameter.

Confidence intervals are a way to express the level of certainty about the true population parameter value based upon the sample selected. The formulae for the various confidence intervals depend upon the statistic used and how the sample was selected, but are all derived from a general unified theory.

The following concepts are crucial and will be used over and over again in what follows:

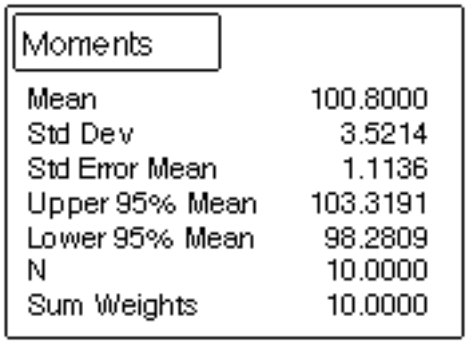
- **Estimate:** The estimate is the quantity computed from the SAMPLE. This is only a guess as to the true value of the population parameter. If you took a new sample, your estimate computed from the second sample, would be different than the value computed form the first sample. It seems reasonable that if you select your sample carefully that these estimates will sometimes be lower than the theoretical population parameters; sometimes it will be higher.
- **Standard error:** The variability of the estimate over repeated samples from the population is measured by the standard error of the estimate. It again seems reasonable that if you select your sample carefully, that the statistics should be 'close' to the true population parameters and that the standard error should provide some information about the closeness of the estimate to the true population parameter.

Refer back to the DDT example considered in the last section. Scientists took a random sample of gulls from Triangle Island (off the coast of Vancouver Island, British Columbia) and measured the DDT levels in 10 gulls. The following values were obtained (ppm):

```
100, 105, 97, 103, 96, 106, 102, 97, 99, 103.
```

What does the sample tell us about the true population average DDT level over all gulls on Triangle Island?

We use JMP to compute sample statistics:



Moments	
Mean	100.8000
Std Dev	3.5214
Std Error Mean	1.1136
Upper 95% Mean	103.3191
Lower 95% Mean	98.2809
N	10.0000
Sum Weights	10.0000

The sample mean, $\overline{Y}$, = 100.8 ppm, and the sample standard deviation, s=3.52 ppm measures the middle of the sample data and the spread of the sample data around the sample mean.

Based on this sample information, is it plausible to believe that the average DDT level over ALL gulls could be as high as 150 ppm? Could it be a low a 50 ppm? Is it plausible that it could be as high as 110 ppm? As high as 101 ppm?

Suppose you had the information from the other 8 samples.

										Sample	
Set	DDT levels in the gulls									mean	std
1	102	102	103	95	105	97	95	104	98	103	100.4	3.8
2	100	103	99	98	95	98	94	100	90	103	98.0	4.1
3	101	96	106	102	104	95	98	103	108	104	101.7	4.2
4	101	100	99	90	102	99	105	92	100	102	99.0	4.6
5	107	98	101	100	100	98	107	99	104	98	101.2	3.6
6	102	102	101	101	92	94	104	100	101	97	99.4	3.8
7	94	101	100	100	96	101	100	98	94	98	98.2	2.7
8	104	102	97	104	97	99	100	100	109	102	101.4	3.7

Based on this new information, what would you believe to be a plausible value for the true population mean?

It seems reasonable that because the sample means when taken over repeated samples from the same population seem to lie between 98 and 102 ppm that this should provide some information about the true population value. For example, if you saw in the 8 additional samples that the range of the sample means varied between 90 and 110 ppm – would your plausible interval change?

Again statistical theory come into play. A very famous and important (for statisticians!) theorem, the Central Limit Theorem, gives the theoretical sampling distribution of many statistics for most common sampling methods.

In this case, the Central Limit Theorem states that the sample mean from a simple random sample from a large population should have an approximate normal distribution with the $se(\overline{Y}) = \frac{\sigma}{\sqrt{n}}$. The *se* of $\overline{Y}$ measures the variability of $\overline{Y}$ around the true population mean when different samples of the same size are taken. Note that the variability of the sample mean is LESS than that of individual observations – does this makes sense?

Using the properties of a Normal distribution, there is a 95% probability that $\overline{Y}$ will vary within about $\pm 2se$ of the true mean (why?). Conversely, there should be about a 95% probability that the true mean should be within $\pm 2se$ of $\overline{Y}$! **This is the crucial step in statistical reasoning.**

Unfortunately, σ - the population standard deviation is unknown so we can't find the *se* of $\overline{Y}$. However, it seems reasonable to assume that s, the sample standard deviation, is a reasonable estimator of σ, the population standard deviation. So, $\frac{s}{\sqrt{n}}$, should be a reasonable estimator of $\frac{\sigma}{\sqrt{n}}$. This is what is reported in the JMP output and we have that the **Estimated Std Error Mean** $= s/\sqrt{n} = 3.5214/\sqrt{10} = 1.1136$ ppm. This number is an estimate of how variable $\overline{Y}$ is around the true population mean in repeated samples of the same size from the same population.

Consequently, it seems reasonable that there should be about a 95% probability, that the true mean is within ± 2 estimated *se* of the sample mean, or, we state that **an approximate 95% confidence interval** is computed as: $\overline{Y} \pm 2(estimated se)$ or $100.8 \pm 2(1.1136) = 100.8 \pm 2.2276 =$ (98.6 -> 103.0) ppm.

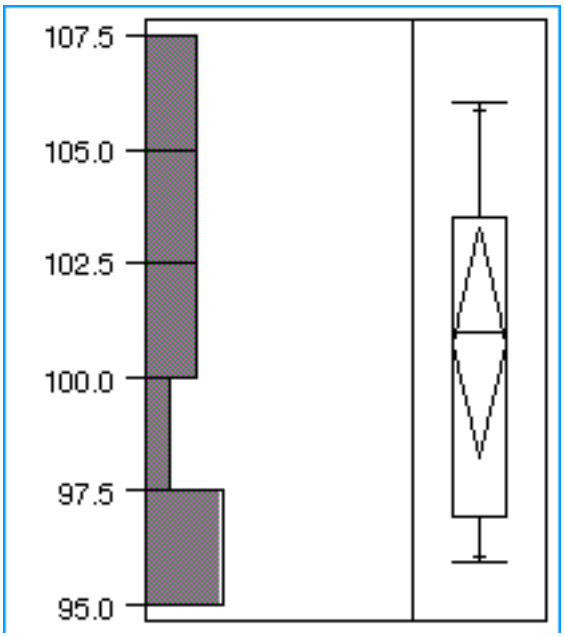
It turns out that we have to also account for the fact that s is only an estimate of σ (s can also vary from sample to sample) and so the estimated *se* may not equal the theoretical standard error. Consequently, the multiplier (2) has to be increased slightly to account for this and the interval computed by JMP of (98.3 -> 103.3) is slightly wider. For large samples (typically greater

than 30), there is virtually no difference in the interval because then s is a very good estimator of σ and no additional correction is needed.

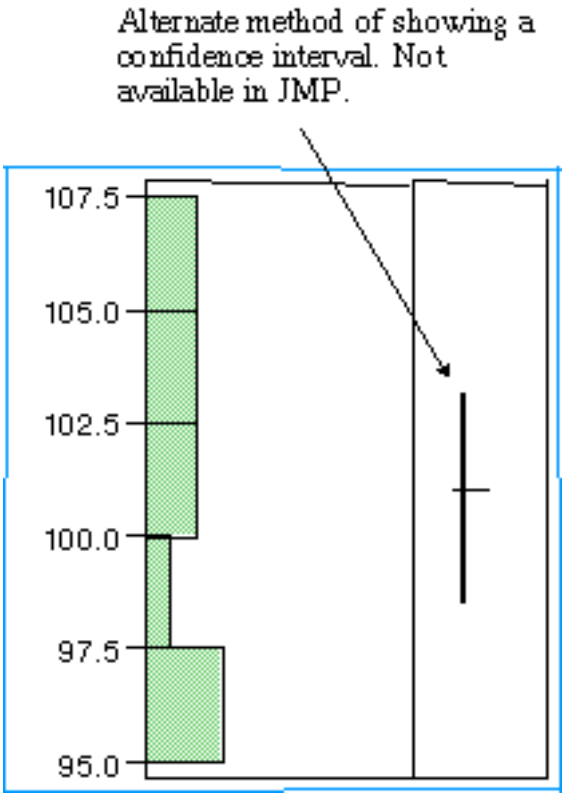
We say that we are 95% confident the true population mean (what ever it is) is somewhere in the interval (98.6 -> 103.0) ppm. What does this mean? We are pretty sure that the true mean DDT is not 110 ppm, nor is it 90 ppm. But we don't really know if it is 99 or 102 ppm. Plausible values for the true mean DDT for ALL gulls is any value in the range 98.6 -> 103.0 ppm.

Note that the interval is NOT an interval for the individual values in the population but rather for the true population mean μ. Also, it is not a confidence interval for the sample mean (which you know to be 100.8) but rather for the unknown population mean μ. **These two points are the most common mis-interpretations of confidence intervals**.

This interval can be graphed in JMP using diamonds (use the *Analyze->Distribution* platform.



The confidence interval is shown as the upper and lower limits of the diamond. Notice that a box-plot and the diamonds are telling you different attributes of the data. Many packages and published papers don't show diamonds, but rather simply show the mean and then either $\pm 1se$ or $\pm 2se$ from the mean as approximate 68% or 95% confidence intervals such as below:



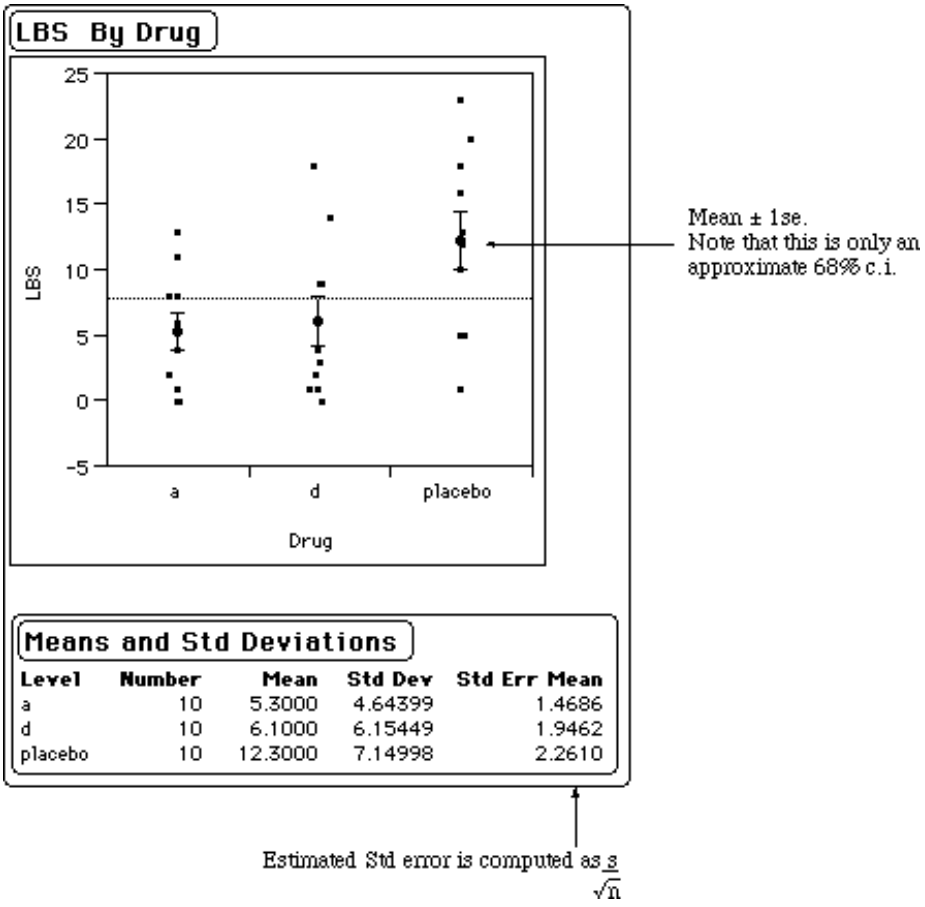
What are the likely effects of changing sample sizes, different amount of variability, and different levels of confidence upon the confidence interval width?

It seems reasonable that a large sample size should be 'more precise', i.e. have less variation over repeated samples from the same population. This implies that a confidence interval based on a larger sample should be narrower for the same level of confidence, i.e. a 95% confidence interval from a sample with $n = 100$ should be narrower than a 95% confidence interval from a sample with $n = 10$ when taken from the same population.

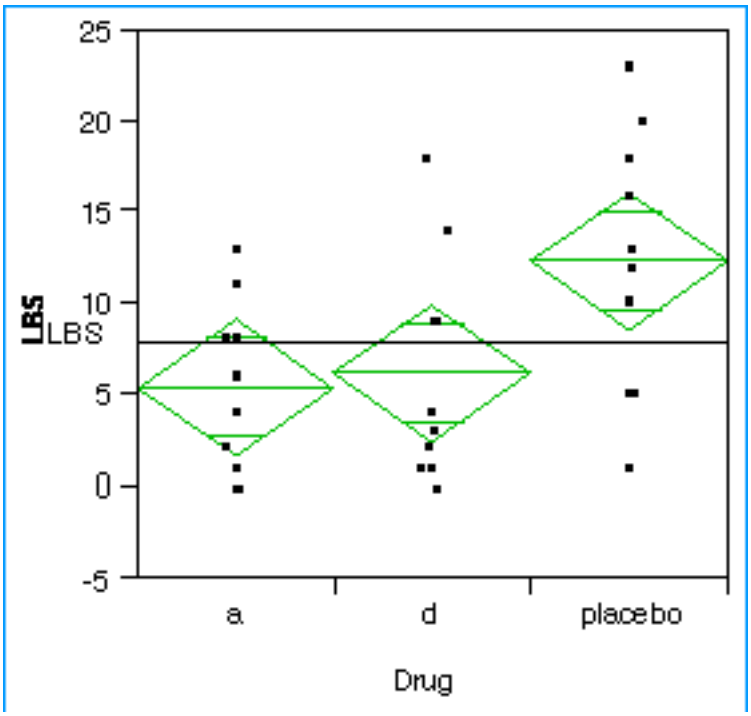
Also, if the elements in a population are more variable, then the variation of the sample mean should be larger and the corresponding confidence interval should be wider.

And, why stop at a 95% confidence level - why not find a 100% confidence interval? In order to be 100% confident, you would have to sample the entire population – not practical for most cases. Again, it seems reasonable that interval widths will increase with the level of confidence, i.e. a 99% confidence interval will be wider than a 95% confidence interval.

How are several groups of means compared if all were selected using a random sample? For now, one simple way to compare several groups is through the use of side-by-side confidence intervals. These can be generated using the *Analyze->Fit Y-by-X* platform in JMP (make sure the X variable is nominal or ordinal scale (see Section 2.5.1)), and then choosing the Means Diamonds from the special pop-up menu. For example, consider the dataset DRUG.JMP in the SAMPLE DATA directory of of JMP. This contains the change in weight of animals when given one of three different drugs (A, D, or Placebo). Make sure that the *drug* column is nominal scale, the *lbs* column (weight change in pounds) is continuous scale, and then use the *Analyze->Fit Y-by-X* platform.

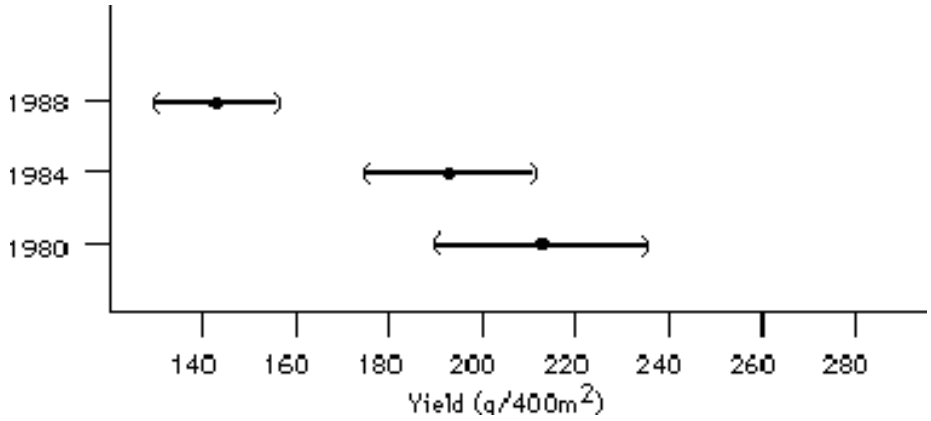


Or, confidence interval diamonds can also be displayed.



What does this show? Because the 95% confidence intervals for drug A and drug D have considerable overlap, there doesn't appear to be much of a difference in the population means (the same value could be common to both groups). However, the overlap between the Placebo and the other drugs is not very much. The population means may differ. [Note the distinction between the *sample* and *population* means in the above discussion.]

As another example, consider the following graph of barley yields for three years along with 95% confidence intervals drawn on the graphs. The data are from a study of crop yields downwind of a coal fired generating plant that started operation in 1985. What does this suggest?



Because the 95% confidence intervals for 1984 and 1980 overlap considerably, there really isn't much evidence that the true mean yield differs. However, because the 95% confidence interval for 1988 does not overlap the other two groups, there is good evidence that the population mean in 1988 is smaller than in the previous two years.

In general, if the 95% confidence intervals of two groups do not overlap, then there is good evidence that the group population means differ. If there is considerable overlap, then the population means of both groups might be the same.

2.3.2 Some practical advice

- In order for confidence intervals to have any meaning, the data must be collected using a probability sampling method. No amount of statistical wizardy will give valid inference for data collected in a haphazard fashion. Remember, haphazard does not imply a random selection.
- If you consult statistical textbooks, they are filled with many hundreds of formulae for confidence intervals under many possible sampling designs. The formulae for confidence intervals are indeed different for various estimators and sampling designs – but they are all interpreted in a similar fashion.
- A rough and ready rule of thumb is that a 95% confidence interval is found as $estimate \pm 2se$ and a 68% confidence interval is found as $estimate \pm 1se$. Don't worry too much about the exact formulae - if a study doesn't show clear conclusions based on these rough rules, then using more exact methods won't improve things.
- The crucial part is finding the *se*. This depends upon the estimator and sampling design – pay careful attention that the computer package you

are using and the options within the computer package match the actual data collection methods. **I can't emphasize this too much!** This is the most likely spot where you may inadvertently use inappropriate analysis!

- Confidence intervals are sensitive to outliers because both the sample mean and standard deviation are sensitive to outliers.
- If the sample size is small, then you must also make a very strong assumption about the population distribution. This is because the central limit theorem only works for large samples. Recent work using *bootstrap* and other resampling methods may be an alternative approach.
- The confidence interval **only tells you imprecision caused by sampling errors**. It does not cover potential imprecision caused by nonresponse, undercoverage, measurement errors etc. In many cases, these can be orders of magnitude larger - particularly if the data was not collected according to a well defined plan.

2.3.3 Technical details

The technical details of a confidence interval for the population mean when the sample is collected using a random sample from a normal population are presented here.

The exact formula for a confidence interval for a single mean when the data are collected using a simple random sample from a population with normally distributed data is:

$$\overline{Y} \pm t_{n-1} \times se$$

or

$$\overline{Y} \pm t_{n-1} \frac{s}{\sqrt{n}}$$

where the t_{n-1} refers to values from a t-distribution with $(n-1)$ degrees of freedom. Values of the t-distribution are tabulated in Section **??**. For the above example for gulls on Triangle Island, $n = 10$, so the multiplier for a 95% confidence interval is $t_9 = 2.2622$ and the confidence interval was found as: 100.8 $\pm$ 2.262(1.1136) = 100.8 $\pm$ 2.5192= (98.28 -> 103.32) ppm which matches the results provided by JMP.

Note that different sampling schemes may not use a t-distribution and most certainly will have different degrees of freedom for the t-distribution.

This formula is useful when the raw data is not given, and only the summary statistics (typically the sample size, the sample mean, and the sample standard deviation) are given and a confidence interval needs to be computed.

What is the effect of sample size? If the above formula is examined, the primary place where the sample size n comes into play is the denominator of the standard error. So as n increases, the *se* decreases. This is sensible because as the sample size increases, $\overline{Y}$ should be less variable (and usually closer to the true population mean). Consequently, the width of the interval decreases. The confidence level doesn't change - we would still be roughly 95% confident, but the interval is smaller. The sample size also affects the degrees of freedom which affects the t-value, but this effect is minor compared to that change in the *se*.

What is the effect of increasing the confidence level? If you wanted to be 99% confident, the t-value from the table in Section **??** increases. For example, the t-value for 9 degrees of freedom increases from 2.262 for a 95% confidence interval to 3.25 for a 99% confidence interval. In general, a higher confidence level will give a wider confidence interval.

2.4 Hypothesis testing

Section summary:

1. Understand the basic paradigm of hypothesis testing
2. Interpret p-values correctly
3. Understand Type I, Type II, and Type III errors
4. Understand the limitation of hypothesis testing

2.4.1 A review

Hypothesis testing is an important paradigm of Statistical Inference, but has its limitations. In recent years, emphasis has moved away from formal hypothesis testing to more inferential statistics (e.g. confidence intervals) but hypothesis testing still has an important role to play.

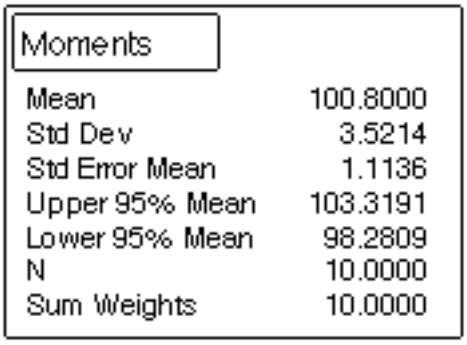
Again consider the example of gulls on Triangle Island introduced in previous sections.

Of interest is the mean DDT level in the gulls. Let (μ) represent the average DDT over all gulls on the island. This value is unknown.

Scientists took a random sample (how was this done?) of 10 gulls and found the following DDT levels.

```
100, 105, 97, 103, 96, 106, 102, 97, 99, 103.
```

Here again is the output from JMP:



Moments

Mean	100.8000
Std Dev	3.5214
Std Error Mean	1.1136
Upper 95% Mean	103.3191
Lower 95% Mean	98.2809
N	10.0000
Sum Weights	10.0000

Now suppose that the value of 98 ppm is a critical value for the health of the species. Is there evidence that the current mean level is different than 98 ppm?

First examine the 95% confidence interval presented above. The confidence interval excludes the value of 98 ppm, so one is fairly confident that the actual mean level differs from 98 ppm. Furthermore the confidence interval gives information about what the true mean DDT level could be. Note that the hypothesized value of 98 ppm is just outside the 95% confidence interval.

A hypothesis test is much more 'formal' and consists of several steps:

1. **Formulate hypothesis**. This is a formal statement of two alternatives. The null hypothesis (denoted as H_0 or H) indicates the state of ignorance or no effect. The alternate hypothesis (denoted as H_1 or A) indicates the effect that is to be detected if present.

   Both the null and alternate hypothesis can be formulated before any data are collected and are always formulated in terms of the population parameter, as this is the variable of interest.

   In this case:

   H:$\mu = 98$, i.e. the mean DDT levels for ALL gulls is 98 ppm.

   H:$\mu \neq 98$, i.e. the mean DDT levels for ALL gulls is not 98 ppm.

   This is known as a two-sided test because we are interested if the mean is either greater than or less than 98 ppm. [2]

[2] It is possible to construct what are known as one-sided tests where interest lies ONLY if the mean exceeds 98 ppm, or a test if interest lies ONLY if the mean is less than 98 ppm. These are rarely useful in ecological work.

2. **Collect data**. Again it is important that the data be collected using probability sampling methods. The form of the data collection will influence the next step.

3. **Compute a test-statistic and p-value**. The test-statistic is computed from the data and measures the discrepancy between the observed data and the null hypothesis, i.e. how far is the observed sample mean of 100.8 ppm from the hypothesized value of 98 ppm?

   At this point, use JMP to do a test of the hypothesis. You should obtain the output below:

```
Test Mean=value

Hypothesized Value      98
Actual Estimate        100.8
df                       9
Std Dev                  3.52136

                t Test
Test Statistic  2.5145      <- Test statistic
Prob > |t|      0.0331      <- Two-sided p-value
Prob > t        0.0165      <- One-sided p-value
Prob < t        0.9835
```

   One discrepancy measure is known as a T-ratio and is computed as:

$$T = \frac{(estimate - hypothesized\ value)}{estimated\ se} = \frac{(100.8 - 98)}{3.52136/\sqrt{10}} = 2.5145$$

   This implies the estimate is about 2.5 *se* above the null hypothesis value.

   Note that there are many measures of discrepancy of the data with the null hypothesis - JMP also provides a 'non-parametric' statistic suitable when the assumption of normality in the population may be suspect.

   How discrepant is the test-statistic? The unusualness of the test statistic is measured by the probability of observing the current test statistic assuming the null hypothesis is true. This is denoted the **p-value**.

   In this case, there are three possible p-values depending upon the form of the alternate hypothesis. Because we are interested if the mean DDT value is greater than or less than 98 ppm, the appropriate p-value is the one denoted '$Prob > |t|$'. Our p-value is found to be 0.0331.

4. **Make a decision**. How are the test statistic and p-value used? The basic paradigm of hypothesis testing is that unusual events provide evidence against the null hypothesis. Logically, rare events "shouldn't happen" if the null hypothesis is true. **This logic can be confusing! We will discuss it more in class**.

In our case, the p-value of 0.0331 indicates there is an approximate 3.3% chance of observing a sample mean that differs from the hypothesized value of 98 if the null hypothesis were true.

Is this unusual? There are no fixed guidelines for the degree of unusualness expected before declaring it to be unusual. Many people use a 5% cut-off value, i.e. if the p-value is less than 0.05, then this is evidence against the null hypothesis; if the p-value is greater than 0.05 then this not evidence against the null hypothesis. [This cut-off value is often called the **α-level**.] If we adopt this cut-off value, then our observed p-value of 0.0331 is evidence against the null hypothesis and we find that there is evidence that the true mean DDT level is different than 98 ppm.

2.4.2 Technical details

The example presented above is a case of testing a population mean against a known value when the population values have a normal distribution and the data is selected using a simple random sample.

- The null and alternate hypotheses are written as:

  H: $\mu = \mu_0$

  A: $\mu \neq \mu_0$

  where $\mu_0 = 98$ is the hypothesized value.

  The test statistic:

  $$T = \frac{(estimate - hypothesized\ value)}{estimated\ se} = \frac{\overline{Y} - \mu_0}{s/\sqrt{n}}$$

  then has a t-distribution with $n - 1$ degrees of freedom.

  The observed value of the test statistic is

  $$T_0 = \frac{(100.8 - 98)}{3.52136} = 2.5145$$

  The p-value is computed by finding

  $$Prob(T > |T_0|)$$

  and is found to be 0.0331.

- In some very rare cases, the population standard deviation σ is known. In these cases, the true standard error is known, and the test-statistics is compared to a normal distribution. This is extremely rare in practise.

- The assumption of normality of the population values can be relaxed if the sample size is sufficiently large. In those cases, the central limit theorem indicates that the distribution of the test-statistics is known regardless of the underlying distribution of population values.

2.4.3 Type I, Type II and Type III errors

Hypothesis testing can be thought of as analogous to a court room trial. The null hypothesis is that the defendant is "innocent" while the alternate hypothesis is that the defendant is "guilty". The role of the prosecutor is to gather evidence that is inconsistent with the null hypothesis. If the evidence is so unusual (under the assumption of innocence), the null hypothesis is rejected.

Obviously the criminal justice system is not perfect. Occasionally mistakes are made (innocent people are convicted or guilty parties are not convicted).

The same types of errors can also occur when testing scientific hypotheses. For historical reasons, two possible types of errors than can occur are labeled as Type I and Type II errors:

- **Type I error** Also known as a false positive. A Type I error occurs when the null hypothesis is erroneously rejected when, in fact, it is true. How can this occur? Well, the p-value measures the probability that the data could have occurred, by chance, if the null hypothesis were true. We usually reject the null hypothesis if the p-value is small, i.e. a rare event. However, rare events do occur and perhaps the data is just one of these rare events. The Type I error rate can be controlled by the cut-off value used to decide if a hypothesis is to be rejected. If you reject hypotheses when the p-value is less than the α=.05 level, then you are willing to accept a 5% chance of making a Type I error.
- **Type II error** Also known as a false negative. A Type II error occurs when the null hypothesis is not rejected, when, in fact, it is false. How can this occur? The usual reasons for a Type II error to occur are that the sample size is too small to make a good decision. For example, suppose that the confidence interval for the gull example, extended from 50 to 150 ppm. No value in the range of 50 to 150 would be rejected by a null hypothesis test.

There are two types of correct decision:

- **Power**. The power of a hypothesis test is the ability to correctly reject the null hypothesis when in fact is is false, i.e the ability to detect if the null hypothesis is false. This is controlled by the sample size.
- **Specificity**. The specificity of a test is the ability to correctly fail to reject the null hypothesis when it is true.

In any experiment, it is never known if one of these errors or a correct

decision has been made. The Type I and Type II errors and the two correct decision can be placed into a summary table:

		Action Taken	
		Reject the null hypothesis	Fail to reject the null hypothesis
True state of nature	Null Hypothesis true	Type I error= False positive error. This is controlled by the α-level used to reject the null hypothesis.	Correct decision. Also known as the specificity of the test.
	Null Hypothesis false	Correct decision. This is known as the power of the test or the sensitivity of the test. Controlled by the sample size with a larger sample size having greater power to detect a false null hypothesis.	Type II error= False negative error. Controlled by sample size with a larger sample size leading to fewer Type II errors.

Usually, a Type I error is more serious (convicting an innocent person) and so we want good evidence before we reject the null hypothesis. We measure the strength of the evidence by the p-value. Typically, we want the p-value to be less than about 5% before we reject the null hypothesis, but this can be varied depending on the problem.

Most experimental studies tend to ignore power (and Type II error) issues. However, these are important – for example, should an experiment be run that only has a 10% chance of detecting an important effect? We will explore issues of power and sample size in later chapters.

What is a Type III error? This is more whimsical, as it refers to a correct answer to the wrong question! Too often, researchers get caught up in their particular research project and spent much time and energy in obtaining an answer, but the answer is not relevant to the question of interest.

2.4.4 Some practical advice

- The p-value does **NOT** measure the probability that the null hypothesis is true. It measures the probability of observing the sample data assuming the null hypothesis were true. You cannot attach a probability statement to the null hypothesis in the same way you can't be 90% pregnant! The hypothesis is either true or false – there is no randomness attached to a hypothesis. The randomness is attached to the data.

- Users of statistics have often emphasized certain standard levels of significance such as 10%, 5%, or 1%. These reflect a time when it was quite impossible to compute the exact p-values, and only tables were available. In this modern era, there is no excuse for failing to report the exact p-value.

- A rough rule of thumb is to reject the hypothesis if the observed test statistic is more than 2 *se* away from the hypothesized value.

- The p-value is also known as the **observed significance level**. In some textbooks, you choose a prespecified significance level (known as the α level) and if the p-value is less than α, you reject the null hypothesis. For example, α is often set at 0.05 (denoted α =0.05). If the p-value $< \alpha = 0.05$, then you reject the null hypothesis; otherwise you fail to reject the null hypothesis. This method is old and falling out of favor in place of reporting the actual p-value.

- Some 'traditional nomenclature':

  - p-value < 0.05, we reject the null hypothesis (we say there is strong evidence against the null hypothesis)
  - p-value is between 0.15 and 0.05, we sit on the fence (indicate that we really do not wish to make a decision; perhaps more data is needed)
  - p-value is greater than 0.15, do not reject the null hypothesis. (we say that there is no evidence against the null hypothesis).

  However, the point at which we reject the null hypothesis will depend upon the situation at hand and the consequences of wrong decisions.

- Preferred terminology is to either **reject the null hypothesis** or **fail to reject the null hypothesis**. It is not proper to state things like:

  - accept the null hypothesis
  - accept the alternate hypothesis
  - the null hypothesis is true
  - the null hypothesis is false,

  The reason is that you haven't 'proved' the truthfulness of the hypothesis; rather you only got evidence against it. It is the same reasons that jury trials return verdicts of 'guilty' (rejected the hypothesis) or 'not guilty' (failed to reject the hypothesis). A jury trial does NOT return an 'innocent' guilty.

- If the hypothesis is rejected, a natural question to ask is 'well, what value of the parameter are plausible given this data'. This is exactly what a confidence interval tells you. Consequently, I usually prefer to find confidence intervals, rather than doing formal hypothesis testing.

- Carrying out a statistical test of a hypothesis is straightforward with many computer packages. However, using tests wisely is not so simple. Each test is valid only in circumstances where the method of data collection adheres to the assumptions of the test. Some hesitation about the use of significance tests is a sign of statistical maturity.

- Beware of outliers or other problems with the data. Be prepared to spend a fair amount of time examining the raw data for spurious points.

- **Hypothesis testing demands the RRR**. Any survey or experiment that doesn't follow the three basic principles of statistics (randomization, replication, and blocking) is basically useless. In particular, non randomized surveys or experiments CANNOT be used in hypothesis testing or inference. Be careful that 'random' is not confused with 'haphazard'.

2.4.5 The case against hypothesis testing

In recent years, there has been much debate about the usefulness of hypothesis testing in scientific research (see the next section for a selection of articles). There a number of "problems" with the uncritical use of hypothesis testing:

- **Sharp null hypothesis** The value of 98 ppm as a hypothesized value seems rather arbitrary. Why not 97.9 ppm or 98.1 ppm. Do we really think that the true DDT value is exactly 98.000000000 ppm? Perhaps it would be more reasonable to ask "How close is the actual mean DDT in the population to 98 ppm?"

- **Choice of α** The choice of α-level (i.e. 0.05 significance level) is also arbitrary. Should different decisions be made if the p-value is 0.0499 or 0.0501? Users of statistics have often emphasized certain standard levels of significance such as 10%, 5%, or 1%. These reflect a time when it was quite impossible to compute the exact p-values, and only tables were available. In this modern era, there is no excuse for failing to report the exact p-value.

  The value of α should reflect the costs of Type I errors, i.e. the costs of false positive results. In a murder trial, the cost of sending an innocent person to the electric chair is very large - we require a very large burden of proof, i.e. the p-value must be very small. On the other hand, the cost of an innocent person paying for a wrongfully issued parking ticket is not very large; a lesser burden of proof is required, i.e. a higher p-value can be used to reject the null.

  A similar analysis should be made for any hypothesis testing case, but rarely is done.

The tradeoffs between Type I and II errors, power, and sample size are rarely discussed in this context.

- **Obvious tests** In many cases, hypothesis testing is used when the evidence is obvious. For example, why would you even bother testing if the true mean is 50 ppm? The data clearly shows that it is not.

- **Interpreting p-values** P-values are prone to mis-interpretation as they measure the plausibility of the data assuming the null hypothesis is true, not the probability that the hypothesis is true. There is also the confusion between selecting the appropriate p-value for one- and two-sided tests.

  Refer to the Ministry of Forest's publication Pamphlet 30 on interpreting the p-value in the MOF library `MOFLibrary`.

- **Effect of sample size** P-values are highly affected by sample size. With sufficiently large sample sizes every effect is statistically significant but may be of no biological interest.

- **Practical vs statistical significance**. Just because the null hypothesis is rejected does not imply that the effect is very large. For example, if you were to test if a coin were fair and were able to toss it 1,000,000 times, you would reject the null hypothesis of fairness if the observed proportion of heads was 50.001%. But for all intents and purposes, the coin is fair enough for real use. **Statistical significance is not the same as practical significance**. Other examples of this trap, are the numerous studies that show cancerous effects of certain foods. Unfortunately, the estimated increase in risk from these studies is often less than 1/100 of 1%!

  The remedy for confusing statistical and practical significance is to ask for a confidence interval for the actual parameter of interest. This will often tell you the size of the purported effect.

- **Failing to reject the null vs no effect**. Just because an experiment fails to reject the null hypothesis, does not mean that there is no effect! A Type II error - a false negative error - may have been committed. These usually occur when experiments are too small (i.e. inadequate sample size) to detect effects of interest.

  The remedy for this is to ask for the power of the test to detect the effect of practical interest, or failing that, ask for the confidence interval for the parameter. Typically power will be low, or the confidence interval will be so wide as to be useless.

- **Multiple testing**. In some experiments, hundreds of statistical tests are performed. However, remember that the p-value represents the chance that this data could have occurred given that the hypothesis is true. So a p-value of 0.01 implies, that this event could have occurred in about 1% of

cases EVEN IF THE NULL IS TRUE. So finding one or two significant results out of hundreds of tests is not surprising!

There are more sophisticated analyses available to control this problem called 'multiple comparison techniques' and are covered in more advanced classes.

On the other hand, a confidence interval for the population parameter gives much more information. The confidence interval shows you how precise the estimate is, and the range of plausible values that are consistent with the data collected.

Modern statistical methodology is placing more and more emphasis upon the use of confidence intervals rather than a blind adherence on hypothesis testing.

2.4.6 Problems with p-values - what does the literature say?

There were two influential papers in the Wildlife Society publications that have affected how people view the use of p-values. Copies of these publications are available in the supplemental reading package.

Statistical tests in publications of the Wildlife Society

Cherry, S. (1998)
Statistical tests in publication of the Wildlife Society
Wildlife Society Bulletin, 26, 947-954.

> The 1995 issue of the *Journal of Wildlife Management* has >2400 *p*-value s. I believe that is too many. In this article, I will argue that authors who publish in the *Journal* and in he *Wildlife Society Bulletin* are over using and misunderstanding hypothesis tests. They are conducting too many unnecessary tests, and they are making common mistakes in carrying out and interpreting the results of the tests they conduct. A major cause of the overuse of testing in the *Journal* and the *Bulletin* seems to be the mistaken belief that testing is necessary in order for a study to be valid or scientific.

- What are the problems in the analysis of habitat availability.
- What additional information do confidence intervals provide that significance levels do not provide?

- When is the assumption of normality critical, in testing if the means of two population are equal?
- What does Cherry recommend in lieu of hypothesis testing?

The Insignificance of Statistical Significance Testing

Johnson, D. H. (1999)
The Insignificance of Statistical Significance Testing
Journal of Wildlife Management, 63, 763-772.

Also available on the web at[3]

> Despite their wide use in scientific journals such as The Journal of Wildlife Management, statistical hypothesis tests add very little value to the products of research. Indeed, they frequently confuse the interpretation of data. This paper describes how statistical hypothesis tests are often viewed, and then contrasts that interpretation with the correct one. He discusses the arbitrariness of p-values, conclusions that the null hypothesis is true, power analysis, and distinctions between statistical and biological significance. Statistical hypothesis testing, in which the null hypothesis about the properties of a population is almost always known a priori to be false, is contrasted with scientific hypothesis testing, which examines a credible null hypothesis about phenomena in nature. More meaningful alternatives are briefly outlined, including estimation and confidence intervals for determining the importance of factors, decision theory for guiding actions in the face of uncertainty, and Bayesian approaches to hypothesis testing and other statistical practices.

This is a very nice, readable paper, that discusses some of the problems with hypothesis testing. As in the Cherry paper above, Johnson recommends that confidence intervals be used in place of hypothesis testing.

So why are confidence intervals not used as often as they should? Johnson give several reasons

- hypothesis testing has become a tradition;
- the advantages of confidence intervals are not recognized;
- there is some ignorance of the procedures available;

[3] http://www.npwrc.usgs.gov/resource/1999/statsig/statsig.htm

- major statistical packages do not include many confidence interval estimates;
- sizes of parameter estimates are often disappointingly small even though they may be very significantly different from zero;
- the wide confidence intervals that often result from a study are embarrassing;
- some hypothesis tests (e.g., chi square contingency table) have no uniquely defined parameter associated with them; and
- recommendations to use confidence intervals often are accompanied by recommendations to abandon statistical tests altogether, which is unwelcome advice.

These reasons are not valid excuses for avoiding confidence intervals in lieu of hypothesis tests in situations for which parameter estimation is the objective.

Followups

In

> Robinson, D. H. and Wainer, H. W. (2002).
> On the past and future of null hypothesis significance testing.
> Journal of Wildlife Management 66, 263-271.

the authors argue that there is some benefit to p-value.s in wildlife management, but then

> Johnson, D. H. (2002).
> The role of hypothesis testing in wildlife science.
> Journal of Wildlife Management 66, 272-276.

counters many of these arguments. Both papers are very easy to read and are highly recommended.

2.5 Meta-data

Meta-data are data about data, i.e how has it been collected, what are the units, what do the codes used in the dataset represent, etc. It is good practice to store

the meta-data as close as possible to the raw data. For example, some computer packages (e.g. JMP) allow the user to store information about each variable and about the data table.

In some cases, data can be classified into broad classifications called scale or roles.

2.5.1 Scales of measurement

Data comes in various sizes and shapes and it is important to know about these so that the proper analysis can be used on the data.

Some computer packages (e.g JMP) use the scales of measurement to determine appropriate analyses of the data.

There are usually 4 scales of measurement that must be considered:

1. **Nominal Data**
   - the data are simply classifications, e.g. m/f
   - the data have no ordering, e.g. it makes no sense to state that m > f
   - the data values are arbitrary labels, e.g. m/f, 0/1, etc
2. **Ordinal Data**
   - the data can be ordered but differences between values cannot be quantified
   - e.g. political parties on left to right spectrum given labels 0, 1, 2
   - e.g. Likert scales, rank on a scale of 1..5 your degree of satisfaction
   - e.g. restaurant ratings
3. **Interval Data**
   - the data can be ordered, have a constant scale, but have no natural zero
   - this implies that differences between data values are meaningful, but ratios are not (e.g. 30C-20C=20C-10C, but 20C/10C is not twice as hot!
   - e.g. temperature (C,F), dates
4. **Ratio Data**
   - data can be ordered, have a constant scale, and have a natural zero

- e.g. height, weight, age, length

Some packages make no distinction between Interval or Ratio data, calling them both 'continuous'. However, this is, technically, not quite correct.

Only certain operations can be performed on certain scales of measurement. The following list summarizes which operations are legitimate for each scale. Note that you can always apply operations from a 'lesser scale' to any particular data, e.g. you may apply nominal, ordinal, or interval operations to an interval scaled datum.

- **Nominal Scale**. You are only allowed to examine if a nominal scale datum is equal to some particular value or to count the number of occurrences of each value. For example, gender is a nominal scale variable. You can examine if the gender of a person is F or count the number of males in a sample.
- **Ordinal Scale**. You are also allowed to examine if an ordinal scale datum is less than or greater than another value. Hence, you can 'rank' ordinal data, but you cannot 'quantify' differences between two ordinal values. For example, political party is an ordinal datum with the NDP to the left of the Conservative Party, but you can't quantify the difference. Another example is preference scores, e.g. ratings of eating establishments where 10=good, 1=poor, but the difference between an establishment with a 10 ranking and an 8 ranking can't be quantified.
- **Interval Scale**. You are also allowed to quantify the difference between two interval scale values but there is no natural zero. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C.
- **Ratio Scale**. You are also allowed to take ratios among ratio scaled variables. Physical measurements of height, weight, length are typically ratio variables. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards). This is because there is a natural zero.

2.5.2 Types of Data

Data can also be classified by its type. This is less important than the scale of measurement, as it usually does not imply a certain type of analysis, but can have subtle effects.

Discrete data Only certain specific values are valid, points between these values are not valid. For example, counts of people (only integer values allowed), the grade assigned in a course (F, D, C-, C, C+, ...).

Continuous data All values in a certain range are valid. For example, height, weight, length, etc. Note that some packages label interval or ratio data as continuous. This is not always the case.

Continuous but discretized Continuous data cannot be measured to infinite precision. It must be discretized, and consequently is technically discrete. For example, a person's height may be measured to the nearest cm. This can cause problems if the level of discretization is too coarse. For example, what would happen if a person's height was measured to the nearest meter.

As a rule of thumb, if the discretization is less than 5% of the typical value, then a discretized continuous variable can be treated as continuous without problems.

2.5.3 Roles of data

Some computer packages (e.g. JMP) also make distinctions about the role of a variable.

Label A variable whose value serves as an identification of each observation - usually for plotting

Frequency A variable whose value indicates how many occurrences of this observation occur. For example, rather than having 100 lines in a data set to represent 100 females, you could have one line with a count of 100 in the Frequency variable.

Weight This is rarely used. It indicates the weight that this observation is to have in the analysis. Usually used in advanced analyses.

X Identifies a variables as an 'independent' or 'predictor' variable. This will be more useful when actual data analysis is started.

Y Identifies a variable as a 'response' or 'dependent' variable. This will be more useful when actual data analysis is started.

2.6 Bias, Precision, Accuracy

The concepts of *Bias*, *Precision* and *Accuracy* are often used interchangeably in non-technical writing and speaking. However these have very specific statistical meanings and it important that these be carefully differentiated.

The first important point about these terms is that they CANNOT be applied in the context of a single estimate from a single set of data. Rather, they are measurements of the performance of an estimator over repeated samples from the same population. Recall, that a fundamental idea of statistics is that repeated samples from the same population will give different estimates, i.e. estimates will vary as different samples are selected. [4]

Bias is the difference between average value of the estimator over repeated sampling from the population and the true parameter value. If the estimates from repeated sampling vary above and below the true population parameter value so that the average over all possible samples equals the true parameter value, we say that the estimator is **unbiased**.

There are two types of bias - systemic and statistical. **Systemic Bias** is caused by problems in the apparatus or the measuring device. For example, if a scale systematically gave readings that were 10 g too small, this would be a systemic bias. Or is snorkelers in stream survey consistently only see 50% of the available fish, this would also be an example of systemic bias. Statistical bias is related to the choice of sampling design and estimator. For example, the usual sample statistics in an simple random sample give unbiased estimates of means, totals, variances, but not for standard deviations. The ratio estimator of survey sampling (refer to later chapters) is also biased.

There is no way from the data at hand to detect systemic biases - the researcher must examine the experimental apparatus and design very carefully. For example, if repeated surveys were made by snorkeling over sections of streams, estimates may be very reproducible (i.e. very precise) but could be consistently WRONG, i.e. divers only see about 60% of the fish (i.e. biased). Systemic Bias is controlled by careful testing of the experimental apparatus etc. In some cases, it is possible to calibrate the method using "known" populations, – e.g. mixing a solution of a known concentration and then having your apparatus estimate the concentration.

Statistical biases can be derived from statistical theory. For example, statistical theory can tell you that the sample mean of a simple random sample is unbiased for the population mean; that the sample VARIANCE is unbiased for the population variance; but that the sample standard deviation is a biased estimator for the population standard deviation. [Even though the sample variance is unbiased, the sample standard deviation is a NON-LINEAR function of the variance (i.e. square-rooted) and non-linear functions don't preserve unbiasedness.] The ratio estimator is also biased for the population ratio. In many cases, the statistical bias can be shown to essentially disappear with reasonably large sample sizes.

[4] The standard error of a estimator measures this variation over repeated samples from the same population.

Precision of an estimator refers to how variable the repeated estimates will be over repeated sampling from the same population. Again recall that every different sample from the same population will lead to a different estimate. If these estimates have very little variation over repeated sample, we say that the estimate is **precise**. The standard error (SE) of the estimator measures the variation of the estimator over repeated sampling from the same population.

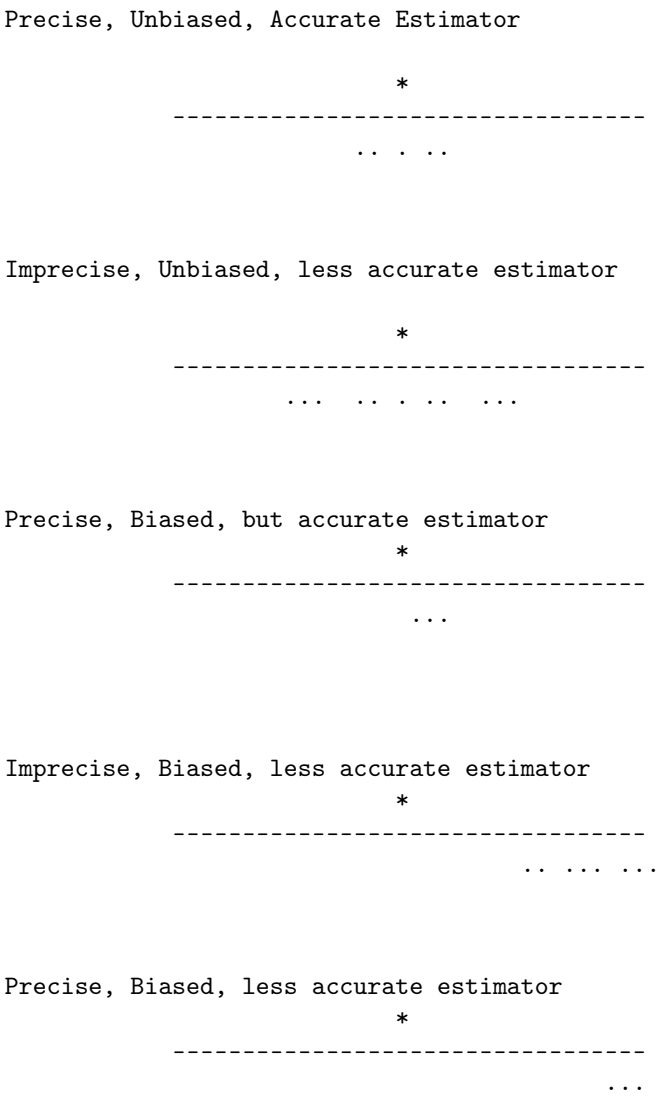
The precision of an estimator is controlled by the sample size. In general, a larger sample size leads to more precise estimates than a smaller sample size.

The precision of an estimator is also determined by statistical theory. For example, the precision (standard error) of a sample mean selected using a simple random sample from a large population is found using mathematics to be equal to $\frac{pop\ std\ dev}{\sqrt{n}}$. A common error is to use this latter formula for all estimators that look like a mean – however the formula for the standard error of any estimator depends upon the way the data are collected (i.e. is a simple random sample, a cluster sample, a stratified sample etc), the estimator of interest (e.g. different formulae are used for standard errors of mean, proportions, total, slopes etc.) and, in some cases, the distribution of the population values (e.g. do elements from the population come from a normal distribution, or a Weibull distribution, etc.).

Finally, **accuracy** is a combination of precision and bias. It measures the "average distance" of the estimator from the population parameter. Technically, one measure of the accuracy of an estimator is the **Root Mean Square Error (RMSE)** and is computed as $\sqrt{(Bias)^2 + (SE)^2}$. A precise, unbiased estimator will be accurate, but not all accurate estimators will be unbiased.

The relationship between bias, precision, and accuracy can be view graphically as shown below. Let * represent the true population parameter value, and periods (.) represent values of the estimator over repeated samples from

the same population.



Statistical theory can tell if an estimator is statistically unbiased, its precision, and its accuracy if a probabilistic sample is taken. If data is collected haphazardly, the properties of an estimator cannot be determined. Systemic biases caused by poor instruments cannot be detected statistically.

2.7 Types of missing data

Missing data happens frequently. There are three types of missing data and an important step in any analysis is to thing of the mechanisms that could have caused the missing data.

First, data can be **Missing Completely at Random (MCAR)**. In this case, the missing data is unrelated to the response variable nor to any other variable in the study. For example, in field trials, a hailstorm destroys a test plot. It is unlikely that the hailstorm location is related to the the response variable of the experiment or any other variable of interest to the experiment. In medical trials, a patient may leave the study because of they win the lottery. It is unlikely that this is related to anything of interest in the study.

If data are MCAR, most analyses proceed unchanged. The design may be unbalanced, the estimates have poor precision than if all data were present, but no biases are introduced into the estimates.

Second, data can be **Missing at Random (MAR)**. In this case, the missingness is unrelated to the response variable, but may be related to other variables in the study. For example, suppose that in drug study involving males and females, that some females must leave the study because they became pregnant. Again, as long as the missingness is not related to the response variable, the design is unbalanced, the estimates have poorer precision, but no biases are introduced into the estimates.

Third, and the most troublesome case, is **Informative Missing**. Here the missingness is related to the response. For example, a trial was conducted to investigate the effectiveness of fertilizer on the regrowth of trees after clear cutting. The added fertilizer increased growth, which attracted deer, which ate all the regrowth! [5]

The analyst must also carefully distinguish between values of 0 and missing values. They are NOT THE SAME! Here is a little example to illustrate the perils of missing data related to 0-counts. The Department of Fisheries and Oceans has a program called ShoreKeepers which allows community groups to collect data on the ecology of the shores of oceans in a scientific fashion that could be used in later years as part of an evironmental assessment study. As part of the protocol, volunteers randomly place 1 m^2 quadrats on the shore and count the number of species of various organisms. Suppose the following data

[5] There is an urban legend about an interview with an opponent of compulsory seat belt legislation who compared the lengths of stays in hospitals of auto accident victims who were or were not wearing seat belts. People who wore seat belts spent longer, on average, in hospitals following the accident than people not wearing seat belts. The opponent felt that this was evidence for not making seat belts compulsory!

were recorded for three quadrats:

Quadrat	Species	Count
Q1	A	5
	C	10
Q2	B	5
	C	5
Q3	A	5
	B	10

Now based on the above data, what is the average density of species A? At first glance, it would appear to be $(5 + 5)/2 = 5$. However, there was no data recorded for species A in Q2. Does this mean that the density of species A was not recorded because people didn't look for species A, or the density was not recorded because the density was 0? In the first instance, the value of A is Missing at Random from Q2 and the true estimated density of species A is indeed 5. In the second case, the missingness is informative, and the true estimated density is $(5 + 0 + 5)/3 = 3.33$.

The above example may seem simplistic, but many database programs are set up in this fashion to "save storage space" by NOT recording zero counts. Unfortunately, one cannot distinguish between a missing value implying that the count was zero, or a missing value indicating that the data was not collected. Even worse, many database queries could erroneously treat the missing data as missing at random and not as zeros giving wrong answers to averages!

The moral of the story is that 0 is a valid value and should be recorded as such! Computer storage costs are declining so quickly, that the "savings" by not recording 0's soon vanish when people can't or don't remember to adjust for the unrecorded 0 values.

If your experiment or survey has informative missing values, you could have a serious problem in the analysis and expert help should be consulted.

2.8 Transformations

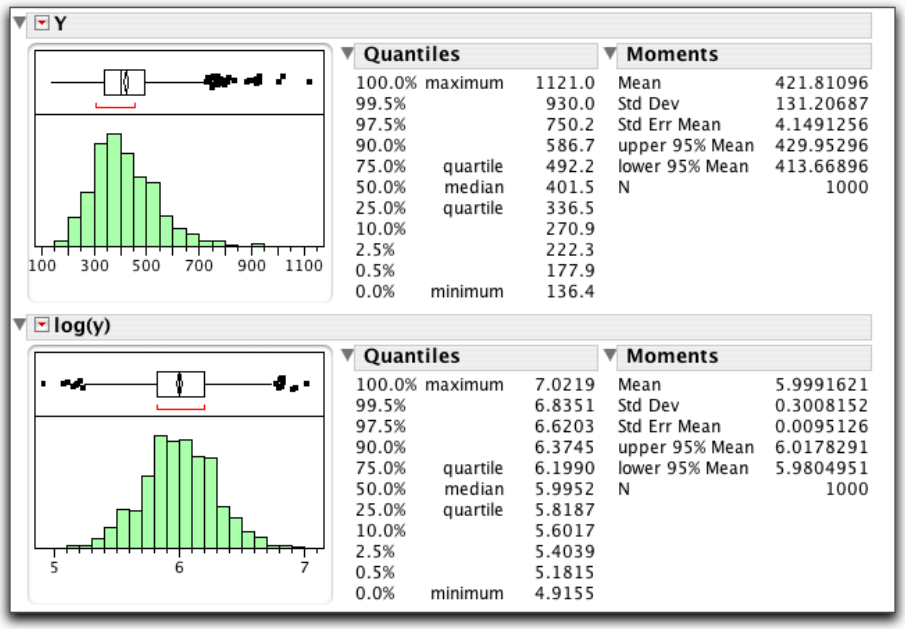
2.8.1 Introduction

Many of the procedures in this course have an underlying assumption that the data from each group is normally distributed with a common variance. In some cases this is patently false, e.g. the data are highly skewed with variances that change, often with the mean.

The most common method to fix this problem is a transformation of the data and the most common transformation in ecology is the logarithmic transform, i.e. analyze the $log(Y)$ rather than Y. Other transformations are possible - these will not be discussed in this course, but the material below applies equally well to these other transformations.

If you are unsure of the proper transformation, there are a number of methods than can assist including a Box-Cox Transform and an applicaiton of Taylor's Power Law. These are beyond the scope of this course.

The logarithmic transform is often used when the data are positive and exhibit a pronounced long right tail. For example, the following are plots of (made-up) data before and after a logarithmic transformation:



There are several things to note in the two graphs.

- The distribution of Y is skewed with a long right tail, but the distribution of $log(Y)$ is symmetric.
- The mean is the right of the median in the original data, but the mean and median are the same in the transformed data.
- The standard deviation of Y is large relative to the mean ($cv = \frac{std\ dev}{mean} = \frac{131}{421} = 31\%$) where as the standard deviation is small relative to the mean on the transformed data ($cv = \frac{std\ dev}{mean} = \frac{.3}{6.0} = 5\%$).

- The box-plots show a large number of potential "outliers" in the original data, but only a few on the transformed data. It can be shown that in the case of a a log-normal distribution, about 5% of observations are more than 3 standard deviations from the mean compared to a normal distribution with less than 1/2 of 1% of such observations.

The form of the Y data above occurs quite often in ecology and is often called a log-normal distribution given that a logarithmic transformation seems to "normalize" the data.

2.8.2 Conditions under which a log-normal distribution appears

Under what conditions would you expect to see a log-normal distribution? Normal distributions often occur when the observed variable is the "sum" of underlying processes. For example, heights of adults (within a sex) are fit very closely by a normal distribution. The height of a person is determined by the "sum" of heights of the shin, thigh, trunk, neck, head and other portions of the body. A famous theorem of statistics (the Central Limit Theorem) says that data that are formed as the "sum" of other data, will tend to have a normal distribution.

In some cases, the underlying process act multiplicatively. For example, the distribution of household income is often a log-normal distribution. You can imagine that factor such as level of education, motivation, parental support act to "multiply" income rather than simply adding a fixed amount of money. Similarly, data on animal abundance often has a log-normal distribution because factors such as survival act multiplicatively on the populations.

2.8.3 ln vs log

There is often much confusion about the form of the logarithmic transformation. For example, many calculators and statistical packages differential between the *common logarithm* (base 10, or *log*) and the *natural logarithm* (base e or ln). Even worse, is that many packages actually use *log* to refer to natural logarithms and $log10$ to refer to common logarithms. IT DOESN'T MATTER which transformation is used as long as the proper back-transformation is applied. When you compare the actual values after these transformations, you will see that $ln(Y) = 2.3log_{10}(Y)$, i.e. the log-transformed values differ by a fixed multiplicative constant. When the anti-logs are applied this constant will "disappear".

In accordance with common convention in statistics and mathematics, the use of $log(Y)$ will refer to the natural or $ln(Y)$ transformation.

2.8.4 Mean vs Geometric Mean

The simple mean of Y is called the arithmetic mean (or simply) the mean and is computed in the usual fashion.

The anti-log of the mean of the $log(Y)$ values is called the geometric mean. The geometric mean of a set of data is ALWAYS less than the mean of the original data. In the special case of log-normal data, the geometric mean will be close to the MEDIAN of the original data.

For example, look that the data above. The mean of Y is 421. The mean of $log(Y)$ is 5.999 and $exp(5.999) = 4.03$ which is close to the median of the original data.

This implies that when reporting results, you will need to be a little careful about how the back-transformed values are interpreted.

It is possible to go from the mean on the transformed scale to the mean on the original scale. For log-normal data, [6] it turns out that

$$\overline{Y}_{original} = exp(\overline{Y}_{transformed} + \frac{s^2}{2})$$

. In this case:

$$\overline{Y}_{original} = exp(5.999 + \frac{(.3)^2}{2}) = 422$$

.

Unfortunately, the back-transformation of the standard deviation does NOT work! There is somewhat complicated formula available in many reference books, but a close approximation is that:

$$s_{untransformed} = s_{transformed} \times exp(\overline{Y}_{transformed})$$

For the data above we see that:

$$s_{untransformed} = .3 \times exp(5.999) = 121$$

which is close, but not exactly on the money.

[6] Other transformation will have a different formula

2.8.5 Back-transforming estimates, standard errors, and ci

Once inference is made on the transformed scale, it is often nice to back-transform and report results on the original scale.

For example, a study of turbidity (measured in NTU) on a stream in BC gave the following results on the log-scale:

Statistics	value
Mean on log scale	5.86
Std Dev on log scale	.96
SE on log scale	0.27
upper 95% ci Mean	6.4
lower 95% ci Mean	5.3

How should these be reported on the original NTU scale?

The estimated MEDIAN (or GEOMETRIC MEAN) on the original scale is found by the back transform of the mean on the log-scale, i.e. the estimated median $= exp(5.86) = 350$ NTU. The 95% confidence interval for the MEDIAN is found by doing a simple back-transformation on the 95% confidence interval for the mean on the log-scale, i.e. from $exp(5.3) = 196$ to $exp(6.4) = 632$ NTUs. Note that the confidence interval on the back-transformed scale is no longer symmetric about the estimate.

There is no direct back-transformation of the standard error from the log-scale to the original scale, but an approximate standard error on the back-transformed scale is found as $se_{original\ scale} = se_{log-scale} \times exp(5.86) = 95$ NTUs.

If the MEAN on the back-transformed scale is needed, recall from the previous section that

$$Mean_{original\ scale} = exp(Mean_{\text{log-scale}} + \frac{\text{std dev}^2_{\text{log-scale}}}{.5})$$

$$Mean_{original\ scale} = exp(Mean_{\text{log-scale}}) \times exp(\frac{\text{std dev}^2_{\text{log-scale}}}{.5})$$

$$Mean_{original\ scale} = \text{Median} \times exp(\frac{\text{std dev}^2_{\text{log-scale}}}{.5})$$

$$Mean_{original\ scale} = \text{Median} \times exp(\frac{.96^2}{2}) = \text{Median} \times 1.58$$

. Hence multiply the median, standard error of the median, and limits of the 95% confidence interval all by 1.58.

2.8.6 Back-transforms of differences on the log-scale

Some care must be taken when back-transforming differences on the log-scale. The general rule of thumb is that a difference on the log-sacle corresponds to a $log(ratio)$ on the original scale. Hence a back-transform of a difference on the log-scale corresponds to a ratio on the original scale. [7]

For example, here are the results from a study to compare turbidity before and after remediation was completed on a stream in BC.

Statistics	value on log-scale
Difference	-0.8303
Std Err Dif	0.3695
Upper CL Dif	-0.0676
Lower CL Dif	-1.5929
P-value	0.0341

A difference of -.83 units on the log-scale corresponds to a ratio of $exp(-.83) = .44$ in the NTU on the original scale. In otherwords, the median NTU after remediation was .44 times that of the median NTU before mediation. Of the median NTU before remediation was $exp(.83) = 2.29$ times that of the median NTU after remediation. Note that 2.29=1/.44.

The 95% confidence intervals are back-tranformed in a similar fashion. In this case the 95% confidence interval on the RATIO of median NTUs lies between $exp(-1.59) = .20$ to $exp(-.067) = .93$, i.e. the median NTU after remediation was between .20 and .95 of the median NTU before remediation.

If necessary you could also back-transform the standard error to get a standard error for the ratio on the original scale, but this is rarely done.

2.8.7 Some additional readings on the log-transform

Here are some additional readings on the use of the log-transform taken from the WWW. The URL is presented at the bottom of each page.

[7] Recall that $log(\frac{Y}{Z}) = log(y) - log(Z)$

Stats >> Model >> Log transformation

Dear Professor Mean, I have some data that I need help with analysis. One suggestion is that I use a log transformation. Why would I want to do this? -- Stumped Susan

Dear Stumped

Think of it as employment security for us statisticians.

Short answer

If you want to use a log transformation, you **compute the logarithm of each data value and then analyze the resulting data**. You may wish to transform the results back to the original scale of measurement.

The logarithm function tends to squeeze together the larger values in your data set and stretches out the smaller values. This squeezing and stretching can correct one or more of the following problems with your data:

1. **Skewed data**
2. **Outliers**
3. **Unequal variation**

Not all data sets will suffer from these problems. Even if they do, the log transformation is not guaranteed to solve these problems. Nevertheless, the log transformation works surprisingly well in many situations.

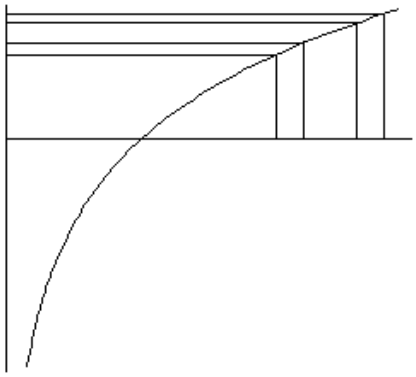
Furthermore, a log transformation can sometimes simplify your statistical models. Some statistical models are multiplicative: factors influence your outcome measure through multiplication rather than addition. These multiplicative models are easier to work with after a log transformation.

If you are unsure whether to use a log transformation, here are a few things you should look for:

1. **Is your data bounded below by zero?**
2. **Is your data defined as a ratio?**
3. **Is the largest value in your data more than three times larger than the smallest value?**

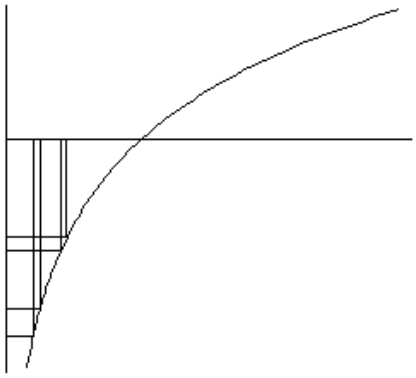
Squeezing and stretching

The logarithm function squeezes together big data values (anything larger than 1). The bigger the data value, the more the squeezing. The graph below shows this effect.



The first two values are 2.0 and 2.2. Their logarithms, 0.69 and 0.79 are much closer. The second two values, 2.6 and 2.8, are squeezed even more. Their logarithms are 0.96 and 1.03.

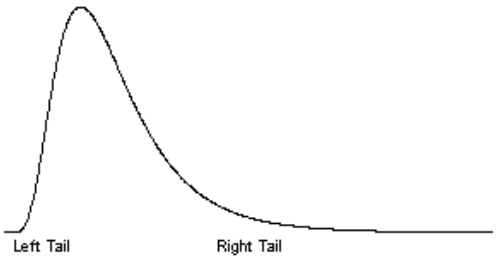
The logarithm also stretches small values apart (values less than 1). The smaller the values the more the stretching. This is illustrated below.



The values of 0.4 and 0.45 have logarithms (-0.92 and -0.80) that are further apart. The values of 0.20 and 0.25 are stretched even further. Their logarithms are -1.61 and -1.39, respectively.

Skewness

If your data are skewed to the right, a log transformation can sometimes produce a data set that is closer to symmetric. Recall that in a skewed right distribution, the left tail (the smaller values) is tightly packed together and the right tail (the larger values) is widely spread apart.



The logarithm will squeeze the right tail of the distribution and stretch the left tail, which produces a greater degree of symmetry.

If the data are symmetric or skewed to the left, a log transformation could actually make things worse. Also, a log transformation is unlikely to be effective if the data has a narrow range (if the largest value is not more than three times bigger than the smallest value).

Outliers

If your data has outliers on the high end, a log transformation can sometimes help. The squeezing of large values might pull that outlier back in closer to the rest of the data. If your data has outliers on the low end, the log transformation might actually make the outlier worse, since it stretches small values.

Unequal variation

Many statistical procedures require that all of your subject groups have comparable variation. If you data has unequal variation, then the some of your tests and confidence intervals may be invalid. A log transformation can help with certain types of unequal variation.

A common pattern of unequal variation is when the groups with the large means also tend to have large standard deviations. Consider housing prices in several different neighborhoods. In one part of town, houses might be cheap, and sell for 60 to 80 thousand dollars. In a different neighborhood, houses might sell for 120 to 180 thousand dollars. And in the snooty part of town, houses might sell for 400 to 600 thousand dollars. Notice that as the neighborhoods got more expensive, the range of prices got wider. This is an example of data where groups with large means tend to have large standard deviations.

With this pattern of variation, the log transformation can equalize the variation. **The log transformation will squeeze the groups with the larger standard deviations more than it will squeeze the groups with the smaller standard deviations**. The log transformation is especially effective when the size of a group's standard deviation is directly proportional to the size of its mean.

Multiplicative models

There are two common statistical models, additive and multiplicative. **An additive model assumes that factors that change your outcome measure, change it by addition or subtraction**. An example of an additive model would when we increase the number of mail order catalogs sent out by 1,000, and that adds an extra $5,000 in sales.

A multiplicative model assumes that factors that change your outcome measure, change it by multiplication or division. An example of a multiplicative model woud be when an inch of rain takes half of the pollen out of the air.

In an additive model, the changes that we see are the same size, regardless of whether we are on the high end or the low end of the scale. Extra catalogs add the same amount to our sales regardless of whether our sales are big or small. In a multiplicative model, the changes we see are bigger at the high end of the scale than at the low end. An inch of rain takes a lot of pollen out on a high pollen day but proportionately less pollen out on a low pollen day.

If you remember your high school algebra, you'll recall that the logarithm of a product is equal to the sum of the logarithms.

$$\log(a \times b) = \log(a) + \log(b)$$

Therefore, a logarithm converts multiplication/division into addition/subtraction. Another way to think about this in a multiplicative model, large values imply large changes and small values imply small changes. The stretching and squeezing of the logarithm levels out the changes.

When should you consider a log transformation?

There are several situations where a log transformation should be given special consideration.

Is your data bounded below by zero? When your data are bounded below by zero, you often have problems with skewness. The bound of zero prevents outliers on the low end, and constrains the left tail of the distribution to be tightly packed. Also groups with means close to zero are more constrained (hence less variable) than groups with means far away from zero.

It does matter how close you are to zero. If your mean is within a standard deviation or two of zero, then expect some skewness. After all the bell shaped curve which speads out about three standard deviations on either side would crash into zero and cause a traffic jam in the left tail.

Is your data defined as a ratio? Ratios tend to be skewed by their very nature. They also tend to have models that are multiplicative.

Is the largest value in your data more than three times larger than the smallest value? The

relative stretching and squeezing of the logarithm only has an impact if your data has a wide range. If the maximum of your data is not at least three times as big as your minimum, then the logarithm can't squeeze and stretch your data enough to have any useful impact.

Example

The DM/DX ratio is a measure of how rapidly the body metabolizes certain types of medication. A patient is given a dose of dextrometorphan (DM), a common cough medication. The patients urine is collected for four hours, and the concentrations of DM and DX (a metabolite of dextrometorphan) are measured. The ratio of DM concentration to DX is a measure of how well the CYD 2D6 metabolic pathway functions. A ratio less than 0.3 indicates normal metabolism; larger ratios indicate slow metabolism.

Genetics can influence CYP 2D6 metabolism. In this set of 206 patients, we have 15 with no functional alleles and 191 with one or more functional alleles.

The DM/DX ratio is a good candidate for a log transformation since it is bounded below by zero. It is also obviously a ratio. The standard deviation for this data (0.4) is much larger than the mean (0.1).

Descriptive Statistics

	N	Mean	Std. Deviation
DM/DX ratio	206	.104298	.426019
Valid N (listwise)	206		

Finally, the largest value is several orders of magnitude bigger than the smallest value.

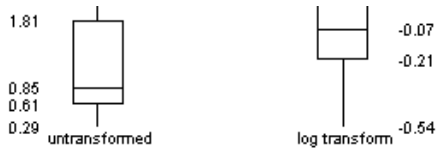
Descriptive Statistics

	N	Minimum	Maximum
DM/DX ratio	206	.0001	3.5541
Valid N (listwise)	206		

Skewness

The boxplots below show the original (untransformed) data for the 15 patients with no functional alleles. The graph also shows the log transformed data. Notice that the untransformed data shows quite a bit of skewness. The lower whisker and the lower half of the box are much packed tightly, while the upper whisker and the upper half of the box are spread widely.

The log transformed data, while not perfectly symmetric, does tend to have a better balance between the lower half and the upper half of the distribution.



Outliers

The graph below shows the untransformed and log transformed data for the subset of patients with exactly two functional alleles (n=119). The original data has two outliers which are almost 7 standard deviations above the mean. The log transformed data are not perfect, and perhaps there is now an outlier on the low end. Nevertheless, the worst outlier is still within 4 standard deviations of the mean. The influence of outliers is much less extreme with the log transformed data.

Unequal variation

When we compute standard deviations for the patients with no functional alleles and the patients with one or more functional alleles, we see that the former group has a much larger standard deviation. This is not too surprising. The patients with no functional alleles are further from the lower bound and thus have much more room to vary.

Report

DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	1.272	15	1.036
One or more functional alleles	.013	191	.025
Total	.104	206	.426

After a log transformation, the standard deviations are much closer.

Report

log DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	-.018	15	.335
One or more functional alleles	-2.281	191	.531

Total	-2.116	206	.785

Summary

Stumped Susan wants to understand **why she should use a log transformation for her data**. Professor Mean explains that a log transformation is often useful for **correcting problems with skewed data, outliers, and unequal variation**. This works because the **log function squeezes the large values of your data together and stretches the small values apart**. The log transformation is also useful when you believe that factors have a mutliplicative effect. You should consider a log transformation when your data are bound below by zero, when you data are defined as a ratio, and/or when the largest value in your data is at least three times as big as the smallest value.

Related pages in Stats

Stats: Geometric mean

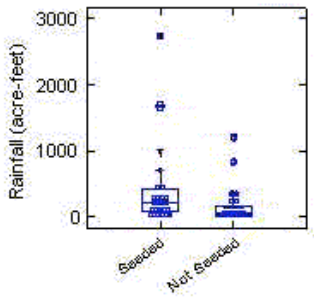
Further reading

The log transformation is special. Keene ON. Stat Med 1995: 14(8); 811-9. [Medline]

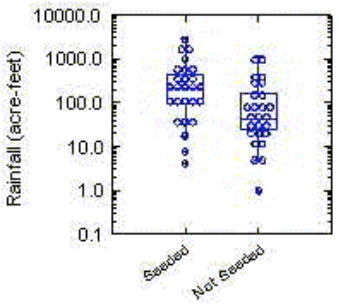
Stats >> Model >> Log transformation

This page was last modified on 01/10/2005 . Send feedback to ssimon at cmh dot edu or click on the email link at the top of the page.

Confidence Intervals Involving Data to Which a Logarithmic Transformation Has Been Applied



These data were originally presented in Simpson J, Olsen A, and Eden J (1975), "A Bayesian Analysis of a Multiplicative Treatment effect in Weather Modification," Technometrics, 17, 161-166, and subsequently reported and analyzed by Ramsey FL and Schafer DW (1997), *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxbury Press. They involve an experiment performed in southern Florida between 1968 and 1972. An aircraft was flown through a series of cloud and, at random, seeded some of them with massive amounts of silver iodide. Precipitation after the aircraft passed through was measured in acre-feet.

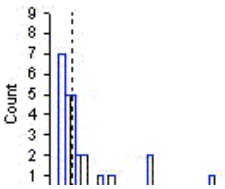


The distribution of precipitation within group (seeded or not) is positively skewed (long-tailed to the right). The group with the higher mean has a proportionally larger standard deviation as well. Both characteristics suggest that a logarithmic transformation be used to make the data more symmetric and homoscedastic (more equal spread). The second pair of box plots bears this out. This transformation will tend to make CIs more reliable, that is, the level of confidence is more likely to be what is claimed.

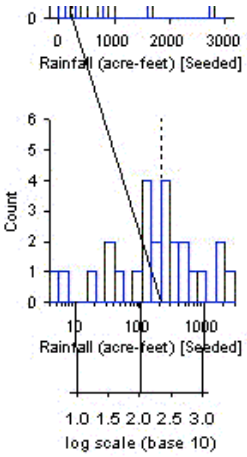
		N	Mean	Std. Deviation	Median
Rainfall	**Not Seeded**	26	164.6	278.4	44.2
	Seeded	26	442.0	650.8	221.6

		N	Mean	Std. Deviation	Geometric Mean
LOG_RAIN	**Not Seeded**	26	1.7330	.7130	54.08
	Seeded	26	2.2297	.6947	169.71

	95% Confidence Interval for the Mean Difference Seeded - Not Seeded (logged data)	
	Lower	**Upper**
Equal variances assumed	0.1046	0.8889
Equal variances not assumed	0.1046	0.8889



Researchers often transform data back to the original scale when a logarithmic transformation is applied to a set of data. Tables might include *Geometric Means*, which are the anti-logs of the mean of the logged data. When data are positively skewed, the geometric mean is invariably less than the arithmetic mean. This leads to questions of whether the geometric mean has any interpretation other than as the anti-log of the mean of the log transformed data.



The geometric mean is often a good estimate of the original median. The logarithmic transformation is monotonic, that is, data are ordered the same way in the log scale as in the original scale. If a is greater than b, then log(a) is greater than log(b). Since the observations are ordered the same way in both the original and log scales, the observation in the middle in the original scale is also the observation in the middle in the log scale, that is,

the log of the median = the median of the logs

If the log transformation makes the population symmetric, then the population mean and median are the same in the log scale. Whatever estimates the mean also estimates the median, and vice-versa. The mean of the logs estimates both the population mean and median in the log transformed scale. If the mean of the logs estimates the median of the logs, its anti-log--the geometric mean--estimates the median in the original scale!

The median rainfall for the seeded clouds is 221.6 acre-feet. In the picture, the solid line between the two histograms connects the median in the original scale to the mean in the log-transformed scale.

One property of the logarithm is that "the difference between logs is the log of the ratio", that is, log(x)-log(y)=log(x/y). The confidence interval from the logged data estimates the difference between the population means of log transformed data, that is, it estimates the difference between the logs of the geometric means. However, the difference between the logs of the geometric means is the log of the ratio of the geometric means. The anti-logarithms of the end points of this confidence interval give a confidence interval for the ratio of geometric means itself. Since the geometric mean is sometime an estimate of the median in the original scale, it follows that a confidence interval for the geometric means is approximately a confidence interval for the ratio of the medians in the original scale.

In the (common) log scale, the mean difference between seeded and unseeded clouds is 0.4967. Our best estimate of the ratio of the median rainfall of seeded clouds to that of unseeded clouds is $10^{0.4967}$ [= 3.14]. Our best estimate of the effect of cloud seeding is that it produces 3.14 times as much rain on average as not seeding.

Even when the calculations are done properly, the conclusion is often misstated.

> The difference 0.4967 does **not** mean seeded clouds produce 0.4967 acre-feet more rain that unseeded clouts. It is also improper to say that seeded clouds produce 0.4967 log-acre-feet more than unseeded clouds.
>
> The 3.14 means 3.14 times as much. It does **not** mean 3.14 times more (which would be 4.14 times as much). It does **not** mean 3.14 acre-feet more. It is a ratio and has to be described that way.

The a 95% CI for the population mean difference (Seeded - Not Seeded) is (0.1046, 0.8889). For reporting purposes, this CI should be transformed back to the original scale. A CI for a **difference** in the log scale becomes a CI for a **ratio** in the original scale.

The antilogarithms of the endpoints of the confidence interval are $10^{0.1046} = 1.27$, and $10^{0.8889} = 7.74$.

Thus, the report would read: "The geometric mean of the amount of rain produced by a seeded cloud is 3.14 times as much as that produced by an unseeded cloud (95% CI: 1.27 to 7.74 times as much)." If the logged data have a roughly symmetric distribution, you might go so far as to say,"The median amount of rain...is approximately..."

Comment: The logarithm is the only transformation that produces results that can be cleanly expressed in terms of the original data. Other transformations, such as the square root, are sometimes used, but it is difficult to restate their results in terms of the original data.

Last modified: Mon Sep 30 2002 14:15:42.

2.9 Standard deviations and standard errors revisited

The use of standard deviations and standard errors in reports and publications can be confusing. Here are some typical questions asked by students about these two concepts.

> I am confused about why different graphs in different publication display the mean ± 1 standard deviation; the mean ± 2 standard deviations; the mean ± 1 *se*; or the mean ± 2 *se*. When should each graph be used?

> What is the difference between a box-plot; ± 2 se; and ± 2 standard deviations?

The foremost distinction between the use of standard deviation and standard errors can be made as follows:

> Standard deviations should be used when information about INDIVIDUAL observations is to be conveyed; standard errors should be used when information about the precision of an estimate is to be conveyed.

There are in fact, several common types of graphs that can be used to display the distribution of the INDIVIDUAL data values. Common displays from "closest to raw data" to "based on summary statistics" are:

- dot plots
- stem and leaf plots
- histograms
- box plots
- mean ± 1 std dev. NOTE this is NOT the same as the estimate ± 1 *se*
- mean ± 2 std dev. NOTE this is NOT the same as the estimate ± 2 *se*

The dot plot is a simple plot of the actual raw data values (e.g. that seen in JMP when the *Analyze->Fit Y-by-X* platform is invoked. It is used to check for

outliers and other unusual points. Often jittering is used to avoid overprinting any duplicate data points. It useful for up to about 100 data points.

Stem and leaf plots are more detailed histograms. These and histograms first start by creating 'bins' representing ranges of the data (e.g. 0-4.9999, 5-.9999, 10-15.9999, etc.). Then the number of data points in each bin is tabulated. The display shows the number or the frequency in each bin. The general shape of the data is examined (e.g. is it symmetrical, or skewed, etc).

The box-plot is an alternate method of displaying the individual data values. The box portion displays the 25th, 50th, and 75th percentiles [8] of the data. The definition of the extent of the whiskers depends upon the statistical package, but generally stretch to show the "typical" range to be expected from data. Outliers may be "indicated" in some plots.

The box-plot is an alternative (and in my opinion a superior) display to a graph showing the mean ± 2 standard deviations because it conveys more information. For example, a box plot will show if the data are symmetric (25th, 50th, and 75th percentiles roughly equally spaced) or skewed (the median much closer to one of the 25th or 75th percentiles). The whiskers show the range of the INDIVIDUAL data values.

The mean ± 1 STD DEV shows a range where you would expect about 68% of the INDIVIDUAL data VALUES assuming the original data came from a normally distributed population. The mean ± 2 STD DEV shows a range where you would expect about 95% of INDIVIDUAL data VALUES assuming the original data came from a normally distributed population. The latter two plots are NOT RELATED to confidence intervals! This plot might be useful when the intent is to show the variability encountered in the sampling or the presence of outliers etc. It is unclear why many journals still accept graphs with ± 1 standard deviation as most people are interested in the range of the data collected – ± 2 standard deviations would be more useful.

I generally prefer the use of dot plots and bax-plots are these are much more informative than stem-and-leaf plots, histograms, or the mean ± some multiple of standard deviations.

Then there are display showing precision of estimates: Common displays are:

[8] The p^{th} percentile in a data set is the value such that at least p% of the data are less than the percentile; and at least (100-p)% of the data values are greater than the percentile. For example, the median=.5 quantile = 50th percentile is the value such that at least 50% of the data values are below the median and at least 50% of the data values are above the median. The 25th percentile=.25 quantile = 1st quartile is the value such that at least 25% of the data values are less than the value and at least 75% of the data values are greater than this value.

- mean ± 1 SE
- mean ± 2 SE
- lower and upper bounds of confidence intervals
- diamond plots

These displays do NOT have anything to do with the sample values - they are trying to show the location of plausible values for the unknown population parameter - in this case - the population mean. A standard error measures how variable an estimate would likely be if repeated samples/experiments from the same population were performed. Note that a *se* says NOTHING about the actual sample values! For example, it is NOT correct to say that a 95% confidence interval contains 95% of INDIVIDUAL data values.

The mean ± 1 SE display is not very informative as it corresponds to an approximate 68% confidence interval. The mean ± 2 SE corresponds to an approximate 95% confidence interval IN THE CASE OF SIMPLE RANDOM SAMPLING.

Graphs showing ± 1 or 2 standard errors are showing the range of plausible values for the underlying population mean. It is unclear why many journals still publish graphs with ± 1 *se* as this corresponds to an approximate 68% confidence interval. I think that a 95% confidence interval would be more useful corresponding to ± 2 *se*.

Caution. Often these graphs (e.g. created by Excel) use the simple formula for the *se* of the sample mean collected under a simple random sample even if the underlying design is more complex! In this case, the graph is in error and should not be interpreted!

Both the confidence interval and the diamond plots (if computed correctly for a particular sampling design and estimator) correspond to a 95% confidence interval.

2.10 Other tidbits

2.10.1 Interpreting *p*-values

> I have a question about *p*-values. I'm confused about the wording used when they explain the *p*-value. They say 'with p=0.03, in 3 percent of experiments like this we would observe sample means as

> different as or more different than the ones we got, if in fact the null hypothesis were true.' The part that gets me is the 'as different as or more different than'. I think I'm just having problems putting it into words that make sense to me. Do you have another way of saying it?

The p-value measures the 'unusualness' of the data assuming that the null hypothesis is true. The 'confusing' part is how to measure unusualness.

For example; is a person 7 ft (about 2 m) unusually tall? Yes, because only a small fraction of people are AS TALL OR TALLER.

Now if the hypothesis is 2-sided, both small and large values of the sample mean (relative to the hypothesized value) are unusual. For example, suppose that null hypothesis is that the mean amount in bottles of pop is 250 mL. We would be very surprised if the sample mean was very small (e.g. 150 mL) or very large (e.g. 350 mL).

That is why, for a two-sided test, the unusualness is 'as different or more different'. You aren't just interested in the probability of getting exactly 150 or 350, but rather in the probability that the sample mean is < 150 or > 350 (analogous to the probability of being 7 ft or higher).

2.10.2 False positives vs false negatives

> What is the difference between a false positive and a false negative

A false positive (Type I) error occurs if the hypothesis of interest is rejected when, in fact, it is true. For example, in a pregnancy test, the null hypothesis is that the person is NOT pregnant. A false positive reading would indicate that the test indicates a pregnancy, when in fact the person is not pregnant. A false negative (Type II error) occurs if the hypothesis of interest is NOT rejected when, in fact, it is false. In the case of a pregnancy test, a false negative would occur if the test indicates not pregnant, when in fact, the person is pregnant.

2.10.3 Specificity/sensitivity/power

> Please clarify specificity/sensitivity/power of a test. Are they the same?

The power and sensitivity are two terms for the ability to reject the null hypothesis when, in fact, it is false. For example, a pregnancy test with a 99that

the test correctly identifies a pregnancy when in fact the person is pregnant.

The specificity of a test indicates the ability to not reject the hypothesis when it is true - the opposite of a Type I error. A pregnancy test would have high specificity if it rarely declares a pregnancy for a non-pregnant person.