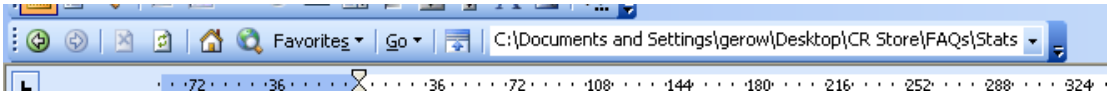
Frequently Asked Questions…

I've organized the questions by topic (some questions may appear in more than one place. Click below to get to pertinent questions…

You might wish to have a "go-back" option, so you can return from any link you follow. To do that, get the **Web Tools** toolbar to appear at the top of your Word file: follow **Tools** → **Customize**, select the **Toolbars** tab. Scroll down and select **Web** (not **Web Tools).** It should look (something like):



The little green arrow button on the left will act as a "return" button.

Table of Contents

Describing Data

1. Can you help me with "right-skew" and "left"? I keep getting them backwards. Answer.

2. What's the difference between $\mu_1 - \mu_2$ and $\bar{y}_1 - \bar{y}_2$? Aren't they saying the same thing? Answer

3. How can you know some samples are not random samples, if someone give you the data for analysis, and you did not do experiment by yourself? **Answer.**

4. I did not understand the notion of a confidence interval for standard deviation. **Answer.**

5. Can you clarify the difference between an estimator and an estimate, w/ examples of both? **Answer.**

6. How do you identify whether a single datum is an outlier, and what do you do with it? **Answer**

7. I don't understand why, if the population you are studying is huge, you don't perforce need a large sample size to do a good job. **Answer**

Distributions of Statistics and Normality

1. What exactly is the Central Limit Theorem? Answer.

2. How can we claim the distribution of the sample mean is Normal, without having to know the theoretical proofs of the Central Limit Theorem? Answer.

3. Can you give me some examples of assessing Normality of the population from which I've taken a sample? Answer.

4. Please connect the dots for me regarding, SE (of the mean), SD, and Normality… Answer.

5. We assess whether the mean has a Normal distribution, but we use the *t*-distribution for testing and confidence intervals. Why? And, while I'm asking, what are degrees of freedom? They are weird. Answer.

6. So what exactly is it that becomes more Normal as sample size (*n*) increases? Our data? The population? The distribution of the mean? Answer.

7. Is it true that the larger the mean, the smaller the sample size required to achieve Normality of the distribution of the mean? Answer

8. How do I know for sure that my sample size is big enough? Answer

9. How do you know that original distributions themselves are naturally more symmetric? Answer

10. If there is more variation in the data, would n = 100 still be a good sample size or would the variance decrease the level of confidence and increase the confidence interval? Answer

11. Where do the *z* (and *t*) distributions come from? Answer

Standard Error and Standard Deviation

1. What is standard deviation?
    a. Formula: Answer.
    b. Meaning: Answer.
    c. Uses: Answer.

2. Please connect the dots for me regarding, SE (of the mean), SD, and Normality… Answer.

3. Why does the formula for the SE of a mean have $\sqrt{n}$ in the denominator? (Caution: this question is relatively technical) Answer.

4. What is SE used for? Answer.

5. Tell me about SD and SE (last time, I promise)… Answer.

6. When I calculate the sample mean, I get a number (say, 17.4). No matter how many times I calculate it, I get the same number. How can the mean have a standard deviation (also called standard error)? It never changes. Answer

7. Why does the sample SD "underestimate" the true population SD? Answer

8. How do you know what standard error formula to use? Answer

9. How do you choose between reporting SD, or SE, or a C.I.? I see all three in the literature**.** Answer.

10. For one data set we studied in class, the sample SD was about 13, but the variance was huge (169). What does that roughly mean in presentation language? Answer.

11. Why is the "SD of the sampling distribution of the mean" also called the "SE of the mean?" Answer.

12. Why do you claim to use "range divided by three" as a data-free estimate of SD? Answer.

The *t*-tools (mostly) and *z*-tools

1. When do you use a paired *t*-tool? Answer.

2. What are degrees of freedom? They are weird. Answer.

3. When do we perform a t-test (or any other test) and why is it so important to do so?? Answer.

4. When do we ever actually need to know the degrees of freedom? Answer.

5. When should we choose the "use equal variances" option for the two-sample *t*? Answer.

6. I'm still a little confused if you want a one tailed test do you use a one sample t tool? Answer.

7. I'm still a little shaky with the issue of the validity of a *t*-tool. Answer.

8. For paired data, I can see that a one-sample t-test can be used but I need help articulating WHY that is. Answer.

9. I want to perform a t-test on Binomial proportions data. How do I calculate a SD when I'm working with a success/failure proportion? Answer.

10. Are there issues concerning equal variances between the two samples in a paired t-tool? That wasn't mentioned as a validity condition. Answer.

11. Are there many varieties of t-tables? z-tables? Answer

12. Where do the *t* (and *z*) distributions come from? Answer

13. Why is the *z* distribution called "the standard"? what is so special about it that is the "standard"? Answer

14. How is *t* calculated when there is no table? What is the formula? Likewise, is it important to know this formula? What happens when there is no table to help me? Answer.

15. How is *z* calculated (if there is no handy table)? What is the formula? Should I (would it be useful for me) know this formula? Answer.

16. How do we know when to use the *z* distribution, and when to use the *t*? Answer.

17. Does it matter if the sample sizes in our two groups are not the same? Answer.

18. Is the primary reason for looking for normality in the difference of the means so they can be compared? Answer

19. Can you ever know that the data is paired if you don't know the details of the study design? Answer.

20. If the difference of the means and the mean of the differences are the same numerically, why don't the paired and two-sample analyses end up with the same result? Answer.

Hypothesis Testing

1. What are the steps to take to complete a hypothesis test? **Answer.**

2. Are the alpha value and confidence level really as closely related as they seem? They seem to be almost the same thing. **Answer.**

3. In hypothesis testing, if you are using a complementary confidence level and alpha (95%, 0.05, say), are there situations when the parameter of interest can fall within the CI, but be rejected by the test (or vice versa)? **Answer.**

4. When would one choose alpha to be different than 0.05? **Answer.**

5. I am confused on the definition of a null hypothesis, and how to choose, or state it in any given situation. **Answer.**

6. Can you tell me the difference between "biological significance" and "statistical significance"? **Answer.**

7. On p-values…
   a. So if the p-value is larger than alpha, we accept the null. Correct? **Answer.**
   b. Suppose alpha is .05. If my test shows a p-value of .02 or .04 I would reject the null hypothesis, but if it was .06 or .07 then I would fail to reject the null. Is this right? **Answer.**
   c. Is the following correct? The smaller the p-value, the more evidence we have against the null, no matter what our alpha level is; and the larger the p-value, the less evidence we have against rejecting the null. **Answer.**
   d. I am still confused on the p-value. Small p-value means exactly what? Large p-value means? **Answer.**
   e. When is the p-value close enough to alpha to fail to reject the null? **Answer.**
   f. When our p value is just a bit bigger than say an alpha of 0.05 (say 0.051) what is done? **Answer.**

8. On one- and two-tailed tests…
   a. When would you use a one-tailed test? **Answer.**

b. Can a researcher avoid two-tailed tests altogether by guessing which way the mean is going to change? Does going with a one-tailed test introduce some risk? **Answer.**
c. I'm still a little confused: if you want a one tailed test do you use a one sample t tool? **Answer.**
d. Tell me again: when can I cut the p-value in half? **Answer.**

9. I don't feel that comfortable with rejecting or failing to reject the null. **Answer.**

10. I wish I had more exposure to the language scientists use in terms of "observed significance" and significant result". **Answer.**

11. I've read that a null hypothesis can always be rejected, if the sample size is big enough. So, does this mean that the P-value is also related to the sample size? **Answer.**

12. I'm having trouble wording hypotheses. Can you help? **Answer.**

13. What is the correct way to state the null hypothesis for a one-tailed test? **Answer.**

14. When you do a one-tailed test, do you divide alpha by 2? **Answer.**

15. When we got a p=0.00, is this a miscalculation on our part? **Answer.**

16. What are Type I and Type II errors? **Answer.**

17. I was reading some stuff on hypothesis testing and I ran across this statement: "If the null hypothesis is always false, what's the big deal about rejecting it?" (Cohen 1990) **Answer.**

18. Can you explain the limitation of hypothesis testing based on "probability of observed result"?? **Answer.**

19. Where do I find the t-critical value in Minitab output? Answer

Working with Binomial Proportions

1. I want to perform a t-test on Binomial proportions data. How do I calculate a SD when I'm working with a success/failure proportion? **Answer.**

2. For inference on a single proportion, when would you recommend using the Normal approximation method? **Answer.**

3. When we want to perform a proportion test, do we still care about the normality of the distribution (of the sample proportion)? **Answer.**

4. I don't understand the theory behind the "+4" method. Why can we add 2 to each category? What makes that legal? **Answer.**

5. I don't understand why exactly we use the +4 method. It just seems like a way to actually cheat. Besides that, what exactly at that point is wrong with using the simple method? **Answer.**

6. Can you reiterate the reason why we use the pooled data option for Binomial proportions? **Answer.**

7. I've heard folks describe the margin of error for a confidence interval in relative terms (e.g. "the margin of error is 10%). Why is that not recommended for confidence intervals for Binomial proportions? **Answer.**

On Confidence Intervals

1. What is the relationship between confidence level and margin of error? **Answer.**

2. What is meant by the lower boundary of the confidence interval? **Answer.**

3. What exactly is a "Confidence Bound" and when do I use it? **Answer.**

4. What exactly is the difference between a confidence limit and a confidence bound? **Answer.**

5. Could you explain exactly what a prediction interval is used for? I'm a little confused about that. **Answer.**

6. In class, you stated that the lower bound from a 90% confidence interval can be used as a 95% lower bound (one-sided interval). How is this so? **Answer.**

7. Are the alpha value and confidence level really as closely related as they seem? They seem to be almost the same thing. **Answer.**

8. How do you interpret a confidence interval for the mean? Is it like this (supposing 95% confidence level)? "We are 95% confident that the true population mean falls between 9.67 and 5.45." **Answer.**

9. How do we know whether to use the *z* or the *t* when making a confidence interval? **Answer.**

10. Will a 95% confidence interval contain the population parameter 95% of the time regardless of the size of n? Does the sample size have any influence over this property of the CI?

Answer.

11. I'm still not clear on the difference between "confidence level" and "confidence interval". **Answer.**

12. How do you choose between reporting SD, or SE, or a C.I.? I see all three in the literature. **Answer.**

13. Why do we care about confidence interval more than standard deviation since confidence interval seems to be calculated on hypothetically huge sampling while SD is based on actual data collected? **Answer.**

14. How do we know what confidence level to choose? **Answer.**

15. Why is 95% the most commonly chosen level in science? Wouldn't it be better to use 99% (more sure of your results)? **Answer.**

16. If there is more variation in the data, would n = 100 still be a good sample size or would the variance decrease the level of confidence and increase the confidence interval? Answer

Bootstrapping and Randomization Procedures

1. We use bootstrapping to make inferences on ratios (of medians or means). I get that. But when you make a ratio of (say) two medians, and get (say) 1.3, how do we get from that to saying one is 30% larger? Answer

2. If you do bootstrapping and find that your sampling distribution is not normal, can you not trust the confidence intervals created? Answer.

3. Why do means and the difference in means tend to come out normal in a t-distribution (with a large enough sample size), but medians and difference in medians would not necessarily be normal? Answer.

Regression Analysis

1. So, if the whole point of regression is to minimize the variance of the residuals, when do we stop? Do we keep squaring, cubing terms till we get diminishing returns on our r squared, then stop? Or is there a more cookbook approach? **Answer**

2. What is R^2-adj and how do we use it? **Answer**

3. Is the adjusted R^2 what the computer thinks the R^2 would be with more data? **Answer**

4. When we speak of a regression model, what do we mean by a "model"? **Answer**

5. In a scientific setting, wouldn't you always want to use the model that is the best predictor? **Answer**

6. If we look at the relation between two variables, can we say that log transformation is useful for ratio data but not for other forms of data? **Answer**

Why does the formula for the SE of a mean have $\sqrt{n}$ in the denominator?

This is a technical question; the answer is accordingly technical. First, the SE formula we use generates an *estimate* of the SD of the distribution of the sample mean. I want to start with the actual SD, and play with its formula, then come back to the estimation business. First the actual SD of the mean is $\frac{\sigma}{\sqrt{n}}$ or $\sqrt{\frac{\sigma^2}{n}}$, where σ denotes the population SD, and σ^2 (literally, the square of the SD) the population variance. I want to work directly with the stuff under the square root sign[1]: $\frac{\sigma^2}{n}$, which formally is called the variance of the mean.

I need to use two rules for deriving the variance of a statistic here; I won't prove that those rules are valid (that's quite beyond the scope of this course), but I will just share them with you.

(1) When you add independent random variables together, the variance of that sum is the sum of the variances: add the variables, add their variances.

(2) When you multiply a random variable by some chosen constant, the variance of the new creation is the variance of that random variable multiplied by the constant squared.

Let's recall the formula for the sample mean[2]: $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$. We have a random sample from some population whose variance we have denoted by σ^2. That is, the variance of y_i is[3] σ^2.

First step: add up all the observed values in the sample. By rule (1), the variance of that sum is $\sum_{i=1}^{n}\sigma^2 = n\sigma^2$.

Second step: multiply by[4] $\frac{1}{n}$. By application of rule (2) to the result of the first step, we get $\left(\frac{1}{n}\right)^2\left(n\sigma^2\right) = \frac{\sigma^2}{n}$.

Done. Take the square-root (result is $\frac{\sigma}{\sqrt{n}}$ or $\sqrt{\frac{\sigma^2}{n}}$), and we have the actual SD of the distribution of the sample mean. We estimate this by replacing (the forever unknown) σ by the sample standard deviation (denoted often by *s* or SD).

[1] I'll take the square-root later, to get us back to the SE…

[2] In words, add up all the numbers, and divide by the sample size.

[3] Of course, y_i is just a number from our sample. We say it has a variance in the sense that if we were to repeatedly do the study y_i would vary from repeat to repeat; the variability therein is measured by σ^2.

[4] That's just another way to say, divide by *n*.

On the *t* distribution and Degrees of Freedom…

I often get asked the following by students new to statistics:

> Why do you say, "The distribution of the means is approximately Normal", then we don't use the Normal distribution; we use the *t*-distribution. And, while I'm asking, what are "degrees of freedom"?

A tiny bit of background recall: to do tests or make confidence intervals on means, we use as a basis the "standardized" mean. For example, if we were testing the hypothesis that the mean equals, say 5.0 (percent alcohol, maybe, for Gossett (see below)), we calculate

$$\frac{\bar{y} - 5.0}{SD/\sqrt{n}}, \text{ where}$$

- $\bar{y}$ is the sample mean,
- 5.0 is (in this case) the hypothesized value of the mean,
- *SD* is the sample standard deviation, and
- *n* is the sample size.

It is indeed the case that the distribution of the sample mean (from a random sample) is approximately Normal[5]. But…

Back in the early part of the 20th century, William Gossett[6] noticed that standardized (as above) means from small samples were more variable than was predicted using the Normal distribution directly.

Gossett asked the question, "What is the standardized distribution of the sample means when we have the following?"

(a) The population the data came from is Normal, but
(b) We need to estimate the SD of that population (using the sample SD).

He arrived (after 200 sheets of paper, three pencils and several erasers)[7] at a formula for the distribution that we now call the *t*-distribution[8]. The formula for drawing the curve looks like this:

[5] True if either the sample itself comes from a Normal distribution (in which case the distribution of the sample mean is *exactly* Normal) or if there is a sufficient sample size.
[6] He statistically studied various measurements on beer for Guinness Brewery for quality-control purposes.
[7] Projection based on what it would have taken me; Gossett likely used less…
[8] Guinness Brewery forbade Gossett from publishing his statistical research findings in any manner that identified him (Guinness didn't want the competition to learn about their new and improved quality control tools); he published under the pseudonym "Student", and chose to label his distribution with the letter *t* (perhaps after Britain's other national beverage ☺). It is sometimes called "Student's *t*" distribution.

$$\frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{(n-1)}{2}\right)\sqrt{(n-1)\pi}}\frac{1}{\left(1+\frac{x^2}{n-1}\right)^{\frac{((n-1)+1)}{2}}},$$

where

- $\Gamma(\bullet)$ is a mathematical function called the gamma function, evaluated for whatever is in the parentheses (details of it are not important here);
- x is some value on the real number line.
- The function is generically often written with the symbol υ (Greek letter "nu") where I've written $n-1$ (but that would obscure an important point for me in my purpose here in explaining degrees of freedom).

This formula generates a bell-shaped curve, and has the property that as sample size gets bigger, it looks more and more closely like a perfect Normal distribution. For small samples, the curve is somewhat shorter and fatter than a perfect Normal.

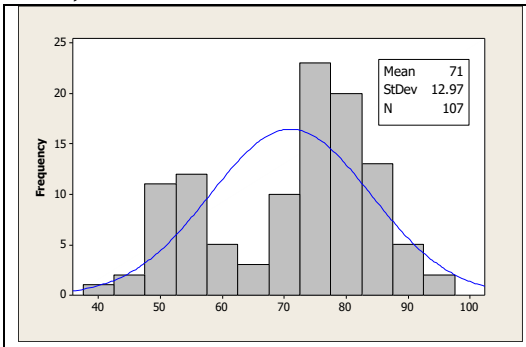
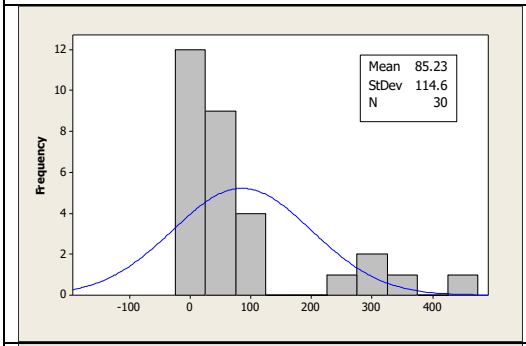
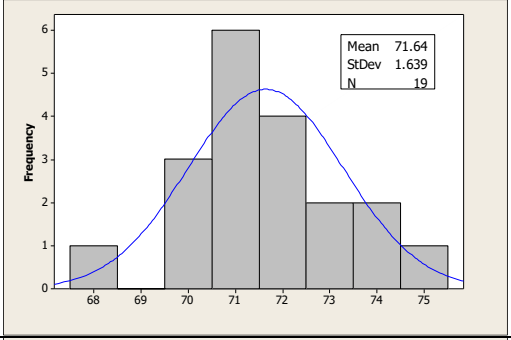
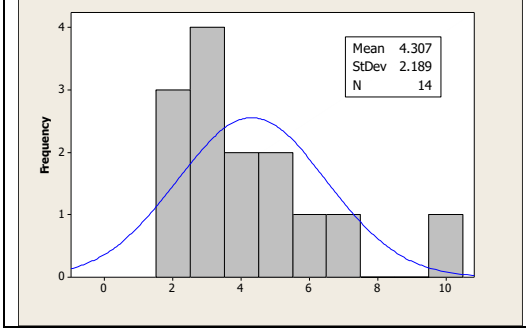
Spurred by Gossett's work, others realized that a similar thing happens in other circumstances (when you compare two means, work with the slope from a regression, etcetera); namely that the observed "standardized statistic" is more variable than expected if you used the Normal distribution directly. After some more paper-scratching (but with clues to the details provided by Gossett's work), the result in each case was Gossett's distribution again, but this time, wherever *n*-1 appeared in his original work another formula, unique to the statistical setting appeared. Examples of some of these formulae:

- *n*-2 for the two-sample comparison of means (assuming equal variances),
- *n*-1- *k* for a linear regression model with *k* predictors

All these formulae had the sample size in them in one way or another.

This created the desire to give this shape-shifting thing a name. The name chosen, "degrees of freedom" was inspired by thinking about theoretical aspects of the problem from a geometric perspective (as was the habit of theoreticians back then), and the name has to do with the dimensionality of the sampling space. For practical purpose, those details matter not: the point was that this thing needed a name, and one was given… degrees of freedom.

Can you give me some examples of assessing Normality of the population from which I've taken a sample? Here are some examples using the mean from a single sample. For each I will just show the histogram (with overlain Normal curve), and give you my assessment. In each case, let us assume that the data are a random sample from some population.

	Large sample size (107) means I trust the histogram to reasonably well represent the actual shape of the population of values. I am therefore quite certain the population is quite non-Normal (it is apparently bimodal[9]). Large sample size also means that I am quite comfortable asserting that the distribution of the sample mean will be quite nicely approximated by the Normal distribution.
	The sample size is reasonably large (30); it seems clear the population is quite skewed. The distribution of the sample means is likely "OK" as regards Normality, but I wouldn't bet a lot on it. For instance, p-values might be only "close" to correct, and confidence interval limits will also only be approximate. I'd likely use it, but also likely not feel really confident…
	Sample size is modest (19), but the histogram does not appear badly asymmetric. I'd feel comfortable asserting approximate Normality for the distribution of the sample mean.
	Nope. The histogram is quite skewed. I don't trust that fully: with a small sample size, a single sample could appear skewed even from a symmetric population. But between the apparent skewedness and the small sample size, I would not feel comfortable in asserting approximate Normality for the distribution of the sample mean.

Sometimes, you can make the assertion of Normality from sources beyond your data: observations from other researchers, for instance.

[9] Bimodal means having two humps; the Normal has only one…

Are the alpha value and confidence level really as closely related as they seem? They seem to be almost the same thing.

We usually choose the confidence level to be "complementary to alpha", meaning, for example, that alpha =0.05 is usually seen in the company of a 95% confidence level (and ditto for 0.10 and 90% and so on). The reason is story-telling consistency.

I'll illustrate with (artificial) examples. Let's keep alpha at .05. The data are times (in minutes between eruptions of Old Faithful Geyser. There are 107 observations, with a mean of 71, and sample SD of 13.

For this case, suppose the historical mean inter-eruption time is 68 minutes, and we want to test whether the current mean is the same.

Case 1: confidence interval supports 68; test rejects it.

Scientist A chooses to use a 99% confidence level while testing: $H_0 : \mu = 68$ versus $H_A : \mu \neq 68.$ The p-value is 0.018, so we reject the null: 68 is not, you believe, likely to be the true mean. A 99% CI is (67.7, 74.3). If you are pretty sure the truth is between 67.7 and 74.3, surely 68 is a possible value. The C.I. supports 68; the test rejects it (at alpha = 0.05).

Same setting: scientist B chooses the confidence level (95%) complementary to alpha. The 95% confident interval is now (68.5, 73.5). The interval says you are pretty sure 68 is not a plausible value, so does the test. They, in their own ways, tell the same tale.

For this case, suppose the historical mean inter-eruption time is 68.8 minutes, and we want to test whether the current mean is the same.

Case 2: confidence interval does not support 68.8; test fails to reject it.

Scientist A chooses to use a 90% confidence interval while testing: $H_0 : \mu = 68.8$ versus $H_A : \mu \neq 68.8.$ The p-value is 0.08 (fail to reject: 68.8 is possibly the true mean). A 90% CI is (68.9, 73.1). The C.I. (barely) considers 68.8 to be not a plausible value; the test fails to rejects it.

Same setting: scientist B chooses the complementary level: 95%. The interval is now (68.5, 73.5). The interval says 68.8 is a plausible value, so does the test. They, in their own ways, tell the same tale.

In summary: If you choose alpha and confidence level to not be complementary, you run the risk of having (apparently) different conclusions come from the tests as compared to the interval. Both are choices. You don't HAVE to choose them to work together, but you do otherwise at your own risk.

I am confused on the definition of a null hypothesis, and how to know what it is in any given situation.

An excellent analogy to hypothesis testing is a criminal court case. There, someone is brought to trial because there is a belief (supported by evidence) that he or she committed a crime. Once in court, however, that belief is nullified for purposes of carrying out the trial. We take as a starting point the hypothesis that the defendant is innocent, and examine the evidence to see if it is inconsistent with that hypothesis. If enough evidence accrues that is inconsistent beyond a reasonable doubt with the hypothesis of innocence, that initial hypothesis is rejected, and the defendant declared guilty. If there is not enough evidence to support that conclusion, the defendant is declared "not guilty"[10].

Just like the criminal court case, one rationale behind our hypothesis testing approach is that we choose the null so that we can predict consequences (assuming, as a starting point, that the null is true). We know what to expect of an innocent person (not at the scene of the crime at the time, no bloodstains, etcetera). By choosing the nullity of the research hypothesis, we get some predictability of our data if the null is true.

Below, I will give only illustrative "for-instances". Don't take lists here as exhaustive.
In order to state a null hypothesis in a given situation, several matters need to be decided.

(1) What is the statistical context? Is it a case of a single sample of data, two independent samples, or two paired samples? Are we to do a comparison of some sort, assess a relationship (i.e. regression)?
(2) What parameter are we being asked to assess (which leads pretty quickly to a choice of statistic)? Is it a mean? A proportion? The difference in two means (or proportions)? Is it the slope in a regression?
(3) Is there a sense of direction to the question? If so, consider a one-tailed test.[11]

Examples:

1. **Test the hypothesis that the mean humerus length among birds that perished is less than 730 thousands of an inch.** This is easy to read as a one-tailed single-sample test for a mean (*t*-test). Accordingly, $H_0 : \mu \geq 730$, and $H_A : \mu < 730$.
2. **Is there any evidence that the birds that survived are larger than the birds who did not survive?** Choice of parameter is not made explicit here; it would be safe in this case to assume the use of means (only since that is usual). We have means from two samples, and a sense of direction to the question. Thus: $H_0 : \mu_s - \mu_p \leq 0$, and $H_A : \mu_s - \mu_p > 0$

Notes:

(1) Hypotheses are stated in terms of parameters. Testing will use relevant statistics.
(2) The question or hypothesis being posed gets embedded in H_A. Its nullity (not opposite) is expressed in H_0.

[10] short-hand, really, for "not found to be guilty": we don't declare that we've shown them to be innocent.
[11] I note that some testing procedures don't have a "sense of direction". For instance, you cannot do a one-tailed Chi-Square test.

I'm having trouble wording hypotheses. Can you help? I'll try, and will do it using examples to illustrate the principles.
Principle One. Hypotheses are stated in terms of population parameters, not sample statistics. The actual values used can come from any of a variety of sources (historical data, for instance). They are usually posed as fixed numbers (so have no SD associated with them).
Principle Two. The research question or hypothesis gets embedded in the alternate hypothesis.
Strategic tips: (1) Identify the parameter(s) of interest. **(2)** State the alternate first. **(3)** Examine the research question or hypothesis for words that indicate a sense of direction (suggesting a one-tailed test).

Case Study (1): A researcher has collected data on size (as measured by the length of the humerus bone) of sparrows from two groups: those that survived and those that perished in, a severe winter storm. Are the survivors larger?
Strategic thinking: The parameters of interest would be the means from the two groups. I see the word "larger", so I'll go one-tailed.
H_A: mean size of survivors is bigger than mean size of those that perished. Using the Greek letter μ to symbolize the population mean, and subscripts *s* and *p* to denote the two populations, the alternate can be compactly written as $H_A : \mu_S > \mu_P$. Although this is a natural enough statement; we are going to use the observed difference in the means $D = \bar{y}_S - \bar{y}_P$ as evidence, so another way to state the alternate that reflects the statistic better is $H_A : \mu_S - \mu_P > 0.$ That suggests the null as $H_0 : \mu_S - \mu_P \leq 0$ (technically correct) or $H_0 : \mu_S - \mu_P = 0$ (easier to use because it states the value actually used in the test).

Case Study (2): A researcher has collected data on the amount of pollen removed and length of visit, per visit by, bees. Is more pollen removed with longer visits?
Strategic thinking: The phrasing of the question suggests a relationship, which we usually examine using regression. I'll take the slope[12] of a simple regression relationship to be the parameter of interest. I see the word "more", so I'll go one-tailed.
H_A: The true slope of the relationship between amount (response) and length (predictor) is positive. Using the Greek letter β to symbolize the true slope, the alternate can be compactly written as $H_A : \beta > 0.$ Thus the null is $H_0 : \beta \leq 0$ (technically correct) or $H_0 : \beta = 0$ (easier to use).

Case Study (3): A researcher has collected data on closing force of crab pincers for two species (call them A and B). Do they show, on average, different pincer strength?
Strategic thinking: Again, it looks like means for two populations. I don't see the any direction suggested, so I'll go two-tailed.
I will use notational conventions already discussed for case 1. $H_A : \mu_{Species\,A} - \mu_{Species\,B} \neq 0.$ That suggests the null as $H_0 : \mu_{Species\,A} - \mu_{Species\,B} = 0.$

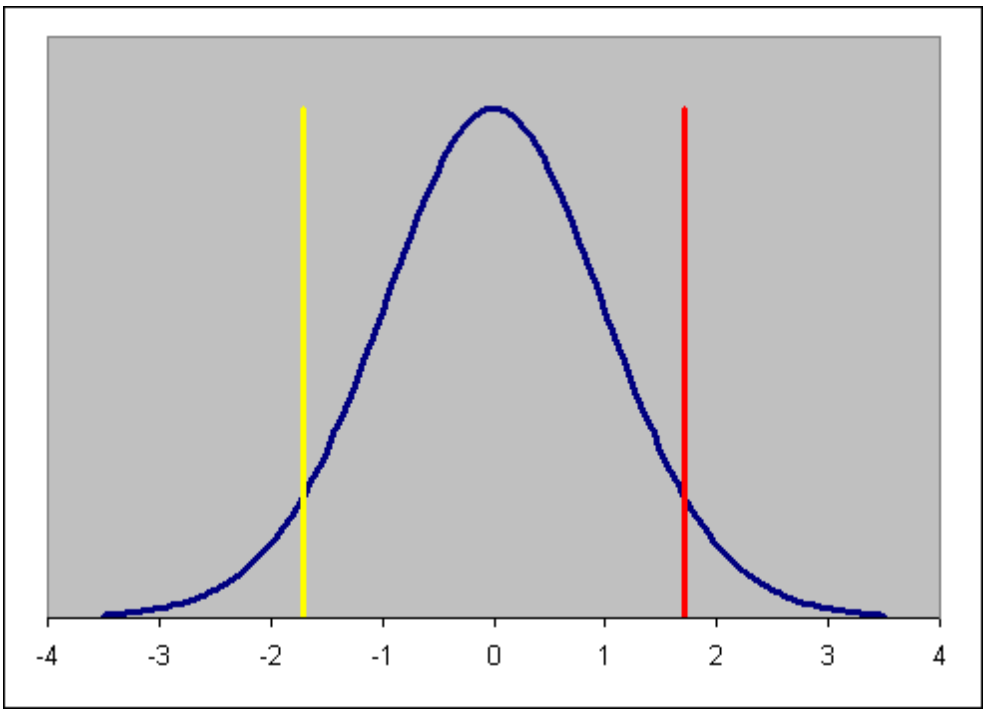
[12] This would be the slope of the relationship if all possible visits from all possible bees were observed. This is an example of a mythical population: it doesn't really exist.

In class, you stated that the lower bound from a 90% confidence interval can be used as a 95% lower bound (one-sided interval). How is this so?

(The following numbers have been concocted to facilitate mental arithmetic…) Suppose we have a sample of 25 values, with mean 10 and SD 15. The estimated SD of the sampling distribution of the mean is $\frac{\text{sample SD}}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3.$

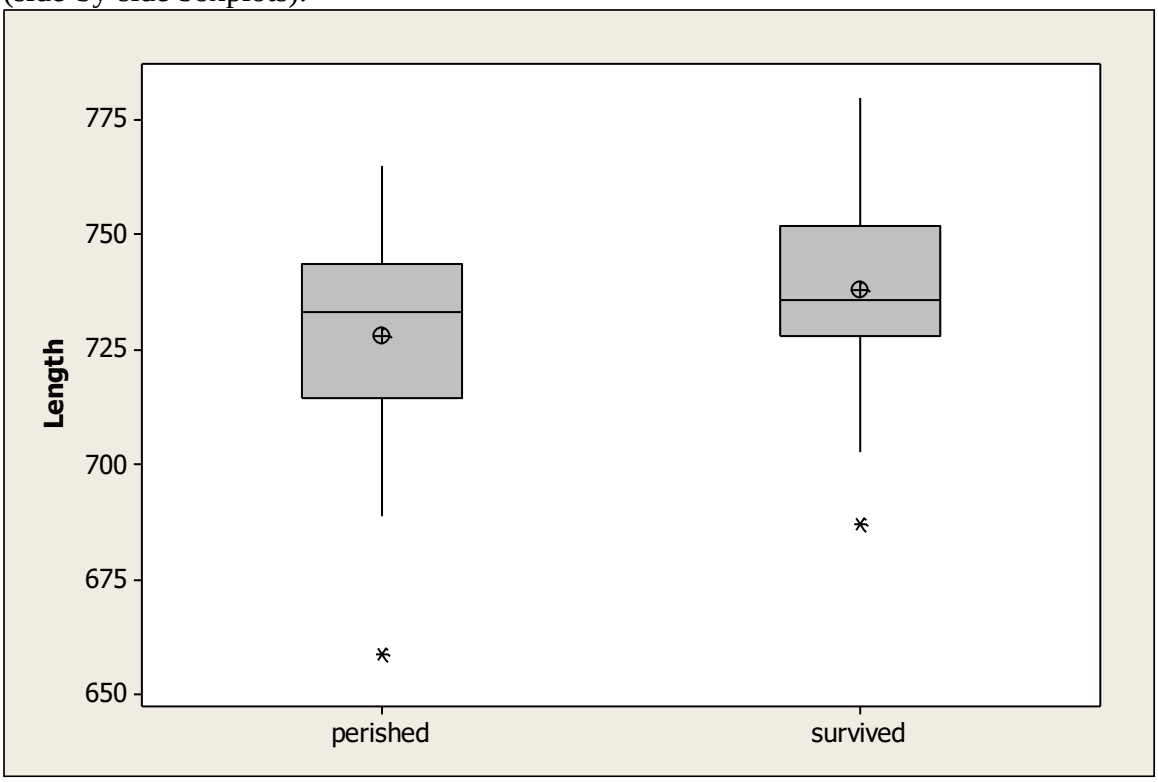
(1) To construct a 90% confidence interval: from a *t*-distribution with 24 *d.f.*, we see that the central 90% of the distribution is captured within 1.71 SDs (see Figure below: 5% is below -1.71; 5% is above 1.71, 90% is captured between). So we would make the confidence interval as[13] $10 \pm 1.71 \times 3 \Rightarrow (4.9, 15.13).$

(2) To construct a 95% Lower Bound: from a *t*-distribution with 24 *d.f.*, we see that the upper 95% of the distribution is bounded at -1.71 SDs below the mean (see Figure below: 5% is below -1.71; 95% is above). So we would make the lower bound as $10 - 1.71 \times 3 \Rightarrow 4.9.$

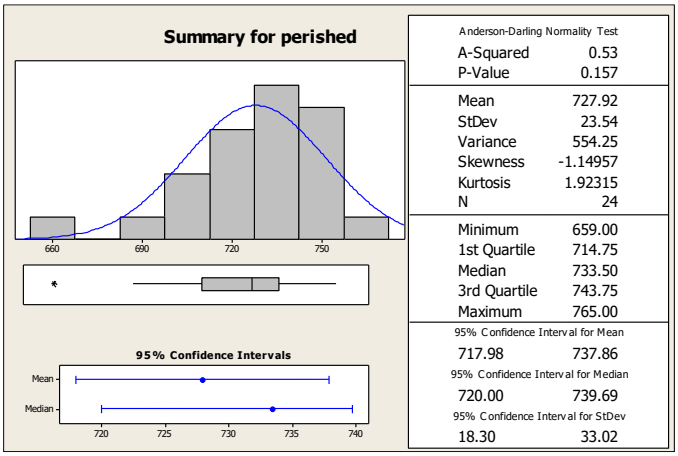


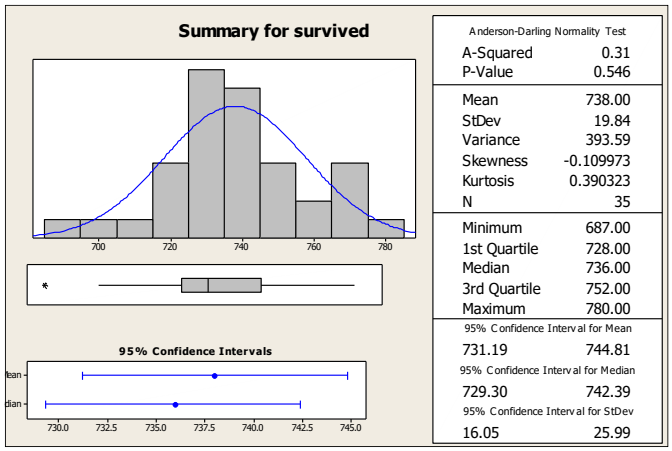
[13] Formula (in words) is mean plus and minus *t* times the SD of the mean (recall that SD of the sampling distribution of the mean is often called its SE).

Two-sample illustration. In the late 1800's, an observational study was done comparing the size (as measured by the humerus[14] length) of 24 sparrows that perished during a severe winter storm to the size of 35 sparrows that survived the storm. Here is a brief summary of the data (side-by-side boxplots).



The plus sign in the circle indicates the mean…
Source: "The Statistical Sleuth", by Dan Schafer and Fred Ramsey.
Here is a detailed summary of each sample:





[14] A wing-bone (analog of a bone in our arm); it was used because it (1) wouldn't have changed as a result of death (unlike, say, weight), and (2) is easy to measure in live birds).

What exactly is the Central Limit Theorem?
The CLT is essentially, a mathematical proof of what we have been observing in class: the distribution of the mean from a random sample becomes more and more Normal as sample size increases. There are actually a large number of variations of this theorem (statisticians need to publish, too!), proving the general result for a wide variety of circumstances.

How can we claim the distribution of the sample mean is Normal, without having to know the theoretical proofs of the Central Limit Theorem?
Consider
It is a (counter-intuitive, at first glance) law of nature (not just a statistical theorem) that the distribution of means from random samples will be more Normally distributed than the population that gave rise to the sample
Having observed that phenomenon, statisticians sought to understand the circumstances when this would occur (and when it wouldn't). That's when they realized that random sampling was a key, and began to realize, through very hard math, that skewness was a key factor: less skewed populations require smaller samples for Normality to appear for the distribution of the means; more skewed populations require larger.
This led to someone finally proving that algebraically (with random samples) the distribution of the sample mean will indeed be more Normal than the population the sample came from. It is called the Central Limit Theorem. That spurred others to try to prove the phenomenon more broadly (than for just a single mean), so actually there are now quite a collection of CLTs…

To use this, it is sufficient to simply gain practice in judging the combination of (1) non-Normality of the population (as evidenced by your sample) and (2) sample size.

Please connect the dots for me regarding, SE (of the mean), SD, and Normality…
Any distribution has variability, which we commonly measure by the SD[15] (or its square, the variance). In doing statistical inference (and I use here for simplicity the case of a mean from a single random sample), there are three distributions to consider:

(1) the population from which we have drawn our sample (population SD);
(2) the sample itself, the SD for which is used to estimate the SD in the population, and
(3) the distribution of our sample mean has its own SD, which we estimate with what we call the SE (of the mean). This last point is true no matter what shape its distribution takes on. *If* the population is Normal or (failing that) *if* the sample size is adequately large, the distribution of the sample mean will be approximately Normal.

What is SE used for?
Routinely, we assume the distribution of whatever statistic (a mean, difference in two means, a slope, a proportion, and so on) is Normal. We use that Normal distribution[16] to calculate p-values

[15] SD is common short-hand for standard deviation.
[16] Technically, we use the *t*-distribution as an altered version of the Normal distribution.

and set limits for confidence intervals. In order to use a Normal distribution for those purposes, we need to estimate its mean and its SD (those two establish its location on the real number line (mean) and scale (SD) (the shape is fixed by using the formula for a Normal distribution). The SE of whatever statistic is in fact an estimate of the SD of that distribution, so is critical to the calculations.

What is the formula for SD?

The following is a copy of the explanation found in the "statistical formulae" section of ***Stats Alive***.

Click Escape to Exit

Some authors use the letter s for standard deviation. I prefer SD followed by a reference to the variable whose SD is being calculated (in parentheses).

To calculate a standard deviation, the final step is to take the square-root of the variance. So the term under the square-root here is the sample variance.

$$SD(Y) = \sqrt{\frac{\sum \left(Y_i - \bar{Y}\right)^2}{n-1}}$$

The variance is (almost) the average of squared differences between each observation and their mean. It turns out (for reasons we can ignore here) that dividing by (n-1) instead of n corrects for under-estimates of SD.

In characterizing variation, the size of differences (from the mean) is a natural choice. The mean of these will unfortunately always be precisely zero, because the negatives will perfectly cancel out the positives. To get around that, one could use the absolute value (size) of the differences or (as we have come to do) square the differences.

How can I get my head around SD without the formula?

I'll start by asking you a (rhetorical: don't bother to answer) question: what is a pound (as a unit of weight)? My bet: you can't answer precisely (and if you could I likely wouldn't understand it!). Most of us could not provide a definition of a pound, but we use the concept effortlessly, daily. How do we do this?

(3) We know its properties: a positive number that tells the weight (or mass) of an object; a bigger number means more weight; 0 means no weight at all.
(4) We have much experience using it.

Let's apply the same approach to SD. It is a positive number that tells of the variability in some numbers; a bigger SD means more variation; 0 means none at all. Now all you need is the experience part ☺.

How do we use SD?
If you tell me the SD, I can immediately get a sense of the range of values. For example, suppose the mean of some data was 100, and the SD 15. About half or more of the values are between 85 and 115 (mean ± 1 SD); most are between 70 and 130 (mean ± 2 SD), and virtually all (if not all) are between 55 and 145 (mean ± 3 SD). If the values come from a perfectly Normal distribution, I could pin these statement down more precisely (remember the 68-95-99.7 rule from your Intro Stat class?), but since most data *don't* come from Normal distributions, I won't sweat that.

So if the p-value is larger than alpha, we accept the null. Correct?
We don't ever accept the null, just as in a criminal court case, we never say, "innocent"; There, we stop at "not (found to be, based on the evidence at hand) guilty". There is a subtle, but important point in that choice of language, which I think is easier to appreciate when thinking about the court case setting. In science use, we reject or fail to reject the null hypothesis.

Is the following correct? The smaller the p-value, the more evidence we have against the null, no matter what our alpha level is; and the larger the p-value, the less evidence we have against rejecting the null.
Yes. Alpha comes in when we start writing adjectives. Smaller is indeed more evidence against; whether it is sufficient evidence to get us all excited depends on our choice of alpha. For instance a p-value of 0.02 connotes stronger evidence than a p-value of 0.04. If alpha was, say, 0.03, the p-value of 0.02 would lead to a "significant" result (i.e. we would reject the null); we would not for 0.04.

Suppose alpha is .05. If my test shows a p-value of .02 or .04 I would reject the null hypothesis, but if it was .06 or .07 then I would fail to reject the null. Is this right?
Technically, yes. Our science use would be more like "strong evidence against the null" (.02 or .04) and "weak evidence" (.06 or .07), relative to alpha = 0.05.

I am still confused on the p-value. Small p-value means exactly what? Large p-value means?
Let's discuss this for the case of testing a single mean (with some specified value of the mean as the null hypothesis). A large p-value is computed in association with (in this case) a mean that is quite close[17] to the hypothesized mean; a small p-value tells us the observed mean is quite far from it. In other words, the p-value measures (indirectly) how close we are to the hypothesized mean, in a way that allows us to put the alpha level into play easily (if the p-value is smaller than alpha, reject the null; otherwise do not).

When is the p-value close enough to alpha to fail to reject the null?
It's not a question of close or not; it's a question of larger or smaller. Technically, if p is bigger than alpha, fail to reject. But in science, we don't often have to do that (reject or not). So, for

[17] Close is defined relative to the variation we expect to see in the mean (if we repeated the study again and again).

instance, (suppose alpha is .05), for a p-value of 0.06 or 0.04, we would use similar language: modest evidence against the null, relative to alpha = 0.05, without ever having to commit to doing anything about it.

When our p value is just a bit bigger than say an alpha of 0.05 (say 0.051) what is done?
Not much. For science purposes we rarely have to actually do anything (like reject a truckload of computer parts, or adjust the fill rate on a stream of ketchup bottles) with our hypothesis tests. What we *actually* do is this: as the p-value gets smaller (relative to alpha), we get more daring and dramatic with our adjectives (I'll bet you didn't think the heart of science lay in creative writing…). Suppose alpha = 0.10. The following are descriptions of, not prescriptions for, usage):
p bigger than 0.20: no evidence against the null
p between 0.15 and 0.20: not much evidence against the null
p between 0.11 and 0.15: modest evidence against…
p between 0.08 and 0.11: decent evidence…
p between 0.04 and 0.08: good evidence…
p between 0.01 and 0.04: strong evidence…
p between 0.0001 and 0.01: very strong evidence (this will make me famous)
p less than 0.0001 (can someone find the phone number for the Nobel Prize Committee?)

I don't feel that comfortable with rejecting or failing to reject the null.
That's why you are in science, not in business ☺. Actually, in science, we don't often, strictly speaking, reject or fail to reject a null. More often, we express our results as having some degree of significance or some degree of evidence against the null. In science use, you'll see, instead of the reject/fail language, phrases like,
(1) "highly significant" or "strong evidence against the null" (p quite less than alpha), or
(2) "modestly significant" or "weak evidence against the null. (p smaller than alpha, but not by a lot).

When would you use a one-tailed test?
If your research question has "direction" to it, you use a one-tailed test. For example, "Due to excellent forage availability, we think that elk numbers (in a certain aspen stand) are now greater than the historical value of 20." Since I specified an interest in greater rather than just "not equal", it becomes a one-tailed test. See next question for more… Direction can be motivated by either an interest in one direction or another, or an understanding based on theory (or literature or colleagues or…) indicating such.

Can a researcher avoid two-tailed tests altogether by guessing which way the mean is going to change? Does going with a one-tailed test introduce some risk?
A one-tailed test is better in the sense that it will give a lower p-value[18], so will be more likely to reject a null (and most often, in science, the alternate is where the excitement lies). It is better

[18] Provided the evidence is in the direction of the alternate.

precisely because it is testing with a more sharply posed question. “Is this bigger?” is sharper than simply “Are they different?”. A more sharply posed question arises if
(1) your interests are in one direction or the other only (e.g. for medical treatments, we only care to see if they improve health) or
(2) you have previous reasons for posing a direction (e.g. food is scarce this year because of drought: are the antelope skinnier as a result?)

In fact, these are the only two legitimate reasons (interest and prior information) with which to justify a one-tailed test.

If you follow the strategy of detecting direction in your data, *then* creating hypotheses to fit, you will always report smaller p-values, to be sure, but you will actually be at risk for falsely rejecting the null at a rate far greater than the chosen alpha level. Suppose all the tests you do are actually two-tailed, but you do this little trick and use bogus one-tailed testing. Your actual alpha level will be double the one you claim to use. So if you do a one-tailed test, your colleagues will (and ought to) question the rationale for it. (and you should do the same whenever you see someone else doing one: there HAS to be a good reason)

I wish I had more exposure to the language scientists use in terms of “observed significance” and significant result”.
Reminder:
alpha is the chosen chance of falsely (i.e. “misled by random chance”) rejecting the null hypothesis.
P-value is chance (assuming null to be true) of observing a result “as or more extreme” than the one we actually got.

Language conventions (I am describing here, not necessarily promoting):

- alpha level often is called (chosen) significance level
- p-value often is called the observed significance level. (For example, if p-value is 0.043, it is true that one would reject the null for all choices of alpha down to 0.0430000000001 (essentially 0.043) (and you could put a lot more zeros in there if you want). So essentially the p-value is the smallest alpha that would coincide with rejection, which led to the name “observed significance level”
- If p-value is less than alpha (so we would reject the null), the foregoing language choices leave us saying “the test is significant” Properly, they should say “statistically significant”.
- If you fail to reject, you could say (continuing with this significance theme) that the results were not significant (just another way of saying the p-value was bigger than alpha).

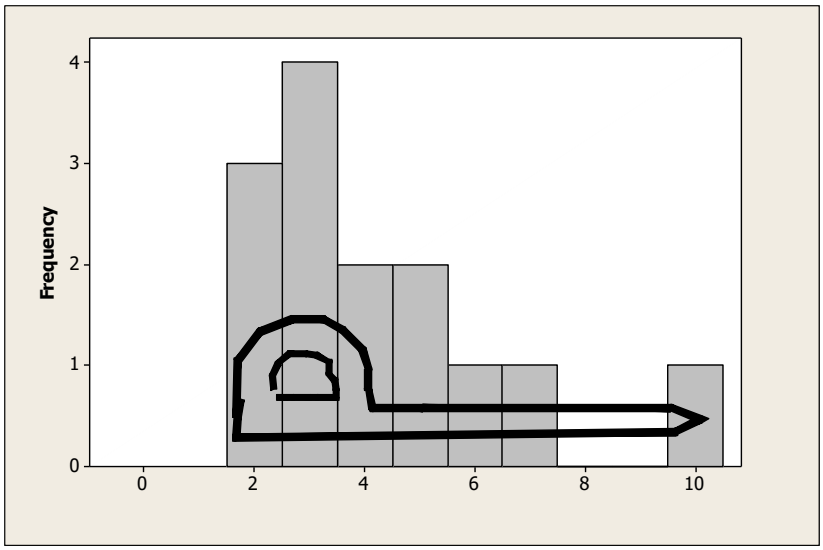
I’ve read that a null hypothesis can always be rejected, if the sample size is big enough. So, does this mean that the P-value is also related to the sample size?

The p-value itself is not related to sample size[19], but there is a valid point that with a large enough sample, one can find "statistical significance" where no biological significance exists.
Example: $H_0 : \mu = 71$, against $H_0 : \mu \neq 71$. Suppose SD is 10.
Let's suppose your sample mean is 71.5. (and that if the true mean is 71.5, it would be considered a meaninglessly small difference in real life). If that mean is from a sample size of 100 (SE of mean = $10/\sqrt{100} = 1$, the test will fail to reject (for most any usual alpha level) since the observed mean is only 0.5 SDs from the null.
Suppose sample size is 10,000. Now the SE of the mean is $10/\sqrt{10,000} = 0.1$, and that same mean (71.5) is now 5 (wow!!) standard errors from the null. Statistically significant (biologically meaningless) result gained by large sample size.
All it means is that you need to consider the statistical significance of your tests in the context of the biological meaningfulness.

When I calculate the sample mean, I get a number (say, 17.4). No matter how many times I calculate it, I get the same number? How can you say the mean has a standard deviation (also called standard error)? It never changes.
Understanding this point is central to doing statistics and having some sense of what it all means. Recall, briefly, that, when you say there is a 50% chance that a flipped coin is heads, the basis for so saying (whether you think about it explicitly or not) is that in the long run, 50% of fair coin flips are going to be heads. Similarly, when we do statistics[20], the basis for it is some intelligent guess for what means would look like **for samples drawn with the same random mechanism we used, from the same population.** Intuitively, you understand that means from such samples would vary randomly from repeat to repeat. Thus, in the context of imagining a (very) long run of repeating our experiment, we can say there is a distribution associated with our mean[21]. If there is a distribution of values, there must also be a standard deviation that can describe the variability in that distribution. So when we say "SD of the mean" or "SE of the mean" there is a slight shorthand implied. The longer version is "SD of the (never-to-be-seen) distribution of the mean.

Can you help me with "right-skew" and "left"? I keep getting them backwards. The problem with these terms is that the choice (made long, long ago, somewhere in the last millennium) of which to call right, and which to call left is pretty arbitrary. There are those who would argue it makes sense because… whatever. The fact is that it confuses lots of bright people; to me that is evidence that it doesn't make a whole lot of sense. Here, I offer a visual mnemonic for remembering. I've overlain the histogram below (using my highly refined art techniques) with a sword, used to *skewer* people. In which direction (left or right) is the skewer pointing? ☺

[19] Except in some technical sense that we can skip over here…
[20] using, here, the sample mean as a basis for making some conclusions about the mean in the population that the random sample came from
[21] and we usually pray for it to be at least approximately Normal in shape.



When do you use a paired *t*-tool?
You use the paired t-tool when you have two data sets (of same sample size, necessarily) wherein each pair of observations are biologically matched in some way. Examples: measurements on two sibs (twins, even); measurements taken before and after (some intervention) on a single person, etcetera. If you don't specifically have some matching criteria (i.e. you know from the study design that none was done, or none was mentioned in the setup of a question), assume they are independent.

So what exactly is it that becomes more Normal as sample size (*n*) increases? Our data? The population? The distribution of the mean?
(1) In principle, the population never changes. It is what it is, and we take samples of it.[22]. For populations of biological data, the distribution is frequently skewed (right), and quite non-Normal. In principle, changing the sample size cannot affect the population.[23]
(2) As the sample size increases for a random sample, the sample histogram will look more and more like the population itself (what*ever* its shape). If the population is severely skewed, then the sample will look more surely severely skewed for large samples.
(3) As the sample size increases for a random sample, the distribution of the mean will look more and more Normal.
There is a certain paradox in this situation. If we have a large sample size, we trust the sample histogram to well represent the population, so can use it to test whether the population is Normal

[22] Ironically, in practice, this is almost never true. In some cases, the act of sampling removes values (permanently) from the population. (Example: dry weight of some plant. To measure it, we have to pick the plant and dry it in an oven). In many other cases, the population itself continually shifts. (Example: average weight of people living in the U.S.. People are constantly dying and being born, and those of us currently doing neither are changing in weight (some of us more than we'd like, some of us less)). So it is often a pretence (usually without dire consequences) to say there exists some fixed population (with a fixed, forever unknown mean) that we sample from. Oh, well.
[23] As I was saying, some forms of sampling remove elements from the population, so in that sense larger sample sizes remove more of it than small ones…

or not. But if we have a large sample size[24], we don't care, because it is quite likely that the distribution of the sample mean will be quite Normal anyway. On the other hand, if the sample size is small[25], in order to safely assume the distribution of the mean is Normal, we need to have the population be (nearly) Normal. The problem is we can't trust the sample histogram very much (so if even if it looks Normal, it could just be a fluke).

Is it true that the larger the mean, the smaller the sample size required to achieve Normality of the distribution of the mean? That turns out to be true for count data[26], but the bottom line is not quite that, in general. If you want a general rule, it might be "The more symmetric the population, the smaller the sample size requirements to achieve approximate Normality for the distribution of the mean". There are two important (because common) special cases. Count data is one. There, populations of count data that have larger means tend to be more symmetric. Binomial proportions data is another. There, if the population proportion (of whichever you choose of two categories) is 0.5 or close, the population is symmetric. If the proportion is extreme (small or large, the population is quite asymmetric. Hence, for that kind of data, the smallest sample size requirements occur when the population proportion is 0.5.

What is the correct way to state the null hypothesis for a one-tailed test?
Let's use a one-tailed test for the difference in two means as illustration. Suppose we have two groups (labeled *S* and *P* (for survived and perished) and we have already determined that we want $H_A : \mu_S - \mu_P > 0$. That suggests the null as $H_0 : \mu_S - \mu_P \le 0$. This is technically correct, but it doesn't tell you what number to enter for doing the test (there are an awful lot of numbers that fit the description "less than or equal to zero). When we do the test, we use 0 for the test value because, among all the possible values described by the null, 0 is the one that is the most conservative. That is, 0 is the value that will demand the most evidence before we will decide to reject. Thus, even though it is technically not correct, I tend to not get bent if you say $H_0 : \mu_S - \mu_P = 0$.

When do we perform a t-test (or any other test) and why is it so important to do so?? I'll stick with your choice of *t*-test as illustration.
The *t*-test is just a tool for testing a scientific hypothesis. From one perspective (why test at all), the importance of it is entirely beyond me: if the scientist thinks (for his or her reasons) that something is important to test, who am I to argue?
From another perspective (why do any specific test, like the *t*-test), try this. Suppose you were on trial (and innocent, of course!) for some dastardly deed. I guess you would feel really upset if you deduced that the trial procedure was flawed in a way that made it more likely for you to be found guilty. So from a procedural perspective, when and if you do a test (and do it correctly), it is important to do the correct test precisely because any other test is likely to not work as well (artificially changing (and taking out of your hands)) the chances of making mistakes.

[24] I'll stick my neck out and say 20 or more…
[25] Here I go again: I'll say 10. If you need to cite me on these, use Don Quixote, or John Lennon (anybody but me…)
[26] such things as "number of eggs per nest" or "number of fish in a stretch of stream"

What is the relationship between confidence level and margin of error?
The confidence level refers to how certain we want to be that our interval contains the parameter being estimated. The margin of error shows us how close to the parameter we are likely to be.
In short:
How close are you? margin of error
How sure do you want to be? confidence level…
The margin of error in a classically constructed confidence interval is a multiplier (from the *t* or *z* distribution) times the standard error of your statistic. The SE you have to live with; it reflects the variation inherent in your data (and is larger or smaller according to your sample size). Your choice of confidence level affects the margin of error because it will dictate the size of the multiplier (larger level → larger multiplier),

When do we ever actually need to know the degrees of freedom?
It's useful to understand, at some level, what the term means. As long as you know which tool to use in a given situation, Minitab (or whatever stats package) does it for you. For instance, given a one-sample setting, with a C.I. or test for the mean, Minitab knows to use a t-distribution with n-1 degrees of freedom.

What is meant by the lower bound of the confidence interval?
If you make a two-sided 95% CI, you are 95% sure the truth is between LL and UL (whatever the Lower and Upper bounds (or "limits" are for the interval). The lower bound is the smaller of the two numbers defining the interval With a one-sided interval, you say you are 95% sure the truth is at least LL (if a lower bound is given) or 95% sure the truth is at most UL (if an upper bound is given. Minitab will select lower or upper (or two-sided) to be consistent with your hypothesis test.

Tell me about SD and SE (last time, I promise)…
The formula (and number) routinely called the SE of the mean estimates the SD of the (never-to-be seen) sampling distribution of the mean. That's quite a mouthful, so I often shorten it to SD of the mean. I don't prefer (the much more common) "SE of the mean" because, since SE is a different "symbol" than SD, it must perforce (so says the mind of a reader) be some different thing. It's not. It is an SD, just of a certain distribution.
Note that this discussion has no bearing on Normality. The foregoing is true, simply. Normality (that is, a bell-shaped distribution) for the mean (see: I skipped that long phrase again) occurs with sufficiently large sample size.
In short:
SE of the mean is synonym for SD of the mean is (short-hand) synonym for SD of the sampling distribution of the mean.

Why does the sample SD "underestimate" the true population SD?
Imagine we have the numbers 10, 25, and 40, and we select one at random. On average, our selected number will be 25 (the actual average of the three numbers). Now suppose our goal is to estimate the square root of the average of these numbers (that would be 5). Our strategy: pick one at random, and take its square root. What are we going to get on average? The square roots are 3.16, 5, and 6.32, respectively. The average of these is 4.83, slightly less than the true value. What does this have to do with anything? Well, to calculate the SD, you first calculate the variance, then *take the square root*. It turns out that the variance formula, on average, estimates unbiasedly the true population variance, but by dint of the number-tweaking I just showed you, the sample SD tends to underestimate the true SD. This tendency shrinks quickly with increasing sample size.

When you do a one-tailed test, do you divide alpha by 2? Well, there is a "divide-by-2" connection between one- and two-tailed tests, but that isn't it. Let me explain, using the two-sample sparrow size example. Let's use alpha = 0.05, just to be conventional.

Let me start with a two-tailed test. We have $H_A: \mu_S - \mu_P \neq 0$ and $H_0: \mu_S - \mu_P = 0$. The observed difference is $\bar{y}_S - \bar{y}_P = 10.08$, with SE =5.86, which yields a *p*-value of 0.092. There is not enough evidence (at alpha = 0.05) with which to reject the null.
One-tailed test (1). Suppose, for whatever weird reason, the researcher had hypothesized that the perished birds would be bigger. Then $H_A: \mu_S - \mu_P < 0$ and $H_0: \mu_S - \mu_P \geq 0$. The observed difference is $\bar{y}_S - \bar{y}_P = 10.08$, so we stop. Dead. The evidence (difference in sample means) goes in a direction *opposite* to our alternate. Clearly there is NO evidence here in favor of it. (Had you gone ahead, the *p*-value would be greater than 0.50).
One-tailed test (2). The actual hypothesis was that the survivors would be bigger. Thus $H_A: \mu_S - \mu_P > 0$ and $H_0: \mu_S - \mu_P \leq 0$. The *p*-value is 0.046, (barely) enough evidence with which to reject the null, at alpha = 0.05.
The p-value for this test is precisely ½ that of the two-tailed test.

When should we choose the "use equal variances" option for the two-sample *t*?
Back in the old days, equal variances (equivalently, equal standard deviations) in the two populations was an assumption for inference on the difference between two means. That approach has fallen into disfavor. An approach that doesn't make that assumption is gaining consensus support among statisticians. For sake of differentiating them by name, I'll refer to the "equal variances" method as the *pooled-t* method, and the other as the *two-sample-t* method. The names aren't mine; I like the latter for identifying itself as the method of choice for two independent samples, and the former for its chief identifying characteristic, the calculation of a common standard deviation, done by pooling the data from both samples. Welch's *t*-test is a name in some use for the two-sample-*t*, recognizing the statistician who first developed it.

Its' never wrong to use the two-sample-*t*, and it gives almost identical answers as the pooled-*t* if in fact the variation is the same in the two populations (and the samples agree by having almost identical sample standard deviations). If the variances are not equal, the pooled-*t*

gives results that are not correct. For more details, see the notes on two-sample *t*-procedures in ***Stats Alive***.

Here is a table summarizing the situation.

Suppose the population variances are equal	**Suppose the population variances are not equal**
The "equal variances" *t*-tool will give the same answer (not perfectly so, but very close) as the "don't assume equal variances" version.	The "equal variances" *t*-tool will give the wrong answer; the "don't assume equal variances" version is closer to perfection.

What exactly is a "Confidence Bound" and when do I use it? I'll use the perished sample from the sparrow data as an illustration. Here are three different 95% confidence type statements that can be made from this data:

(1) I am 95% sure the population mean is between 717.98 and 737.86.
(2) I am 95% sure the population mean is at least 719.68.
(3) I am 95% sure the population mean is no more than 736.15.

The first statement is the usual confidence interval (which here I could also call a two-sided interval). The latter two are also valid confidence statements, called "one-tailed" or "one-sided" confidence intervals; the term "95% 'lower' or 'upper' confidence bound" is becoming standard. These days, if you ask a stats package for

- a two-tailed test, a classical two-sided C.I. gets produced;
- a one-sided (greater than) test, a lower confidence bound gets produced; and
- a one-sided (less than) test, an upper confidence bound gets produced.

Note: You could call the lower and upper limits of a classical C.I. "bounds" (and not be wrong, but (1) the word "bound" is becoming the norm when only a single value is offered, and (2) the one-sided intervals get described with a sense of direction: "greater than" or "less than" rather than "between.

Note further: The choice (of one-sided or two) is *usually* connected to choice of one- or two-tailed hypothesis tests, but it doesn't have to be. Simply, if you want to say, "I'm sure it's at least this big", or "I'm sure it's no bigger than this", you would use a one-sided interval for that job.

I'm still a little confused: if you want a one tailed test do you use a one sample t tool? One sample, two samples, three samples, tells you how many samples you have. One-or two tailed is a choice of whether a test has a "sense of direction" to it.

And also how do you get Minitab to distinguish between the two different tests? In almost all cases (one-sample, two-samples, and so on), there is an Options button you use to tell it about the test.

How do I know for sure that my sample size is big enough? Note: "Big enough" here means big enough that we trust in the Normality of the distribution of the mean (or difference in two means, if we are in a two-sample situation). There is no one, simple answer. That said, I will stick my neck out and make one up. Thirty is usually big enough for a single sample; as it gets smaller, I pay more attention to the shape of the sample histogram. Once it gets below, say, 15, I start sending off anxiety hormones. For comparing two means, 20 in each sample is "big enough", and nervous starts around 10…

How do you state the confidence interval for the mean? Is it like this (supposing 95% confidence level)? "We are 95% confident that the true population mean falls between 9.67 and 5.45." Yes, except that conventionally, one names the lower limit first.

When we got a p=0.00, is this a miscalculation on our part? No. The actual value is just something very small (say, 0.003), which, when rounded to two decimals appears as 0.00.

I'm still a little shaky with the issue of validity of a *t*-tool. Here's the deal.

(1) When you use a *t*-tool, *p*-values and confidence interval limits are calculated using a *t*-distribution. This is the correct distribution to use if (among other things) the distribution of your chosen statistic (the mean, the difference in two means, etcetera) is approximately Normal.
(2) If that (the Normality) is not true, the calculations are all bogus, making your results all bogus.

What's the difference between $\mu_1 - \mu_2$ and $\bar{y}_1 - \bar{y}_2$? Aren't they saying the same thing? $\mu_1 - \mu_2$ is the *true* difference in the *population* means; $\bar{y}_1 - \bar{y}_2$ is the *estimated* difference using the *sample* means. $\mu_1 - \mu_2$ is presumed to be some fixed but forever unknown number; $\bar{y}_1 - \bar{y}_2$ has variability (from sample to sample), which we use to refine our estimates of the former.

For paired data, I can see that a one-sample t-test can be used but I need help articulating WHY that is. Because a paired t-tool computationally reduces to a one-sample tool (but you don't need to do that by hand; just use the paired tool, and let Minitab do it for you).

If I want to perform a t-test on Binomial proportions data, how do I calculate a SD when I'm working with a success/failure proportion? The cool thing is that you don't need to (and by the way, the test and interval will get done using a *z*-distribution, not the *t*). It turns out that variation in binomial proportions in intimately connected to the proportion itself, so once you specify a proportion (or a sample-based estimate for it), the SE is automatic. The formula (given a value of *p*) for the standard error of the sample proportion is $\sqrt{\frac{p(1-p)}{n}}$.

Are there issues concerning equal variances between the two samples in a paired t-tool? That wasn't mentioned as a validity condition. Good question. It is not relevant (since, in the end, it really truly is a single sample analysis).

For inference on a single proportion, when would you recommend using the Normal approximation method? If the statistics package you are using has the exact distribution results, I'd say not to bother ever. On the other hand, for confidence intervals, the "+4" method (with the *z*-tool) will give results that are about as good as the exact ones. If you *don't* have access to exact methods, keep the "10:10" rule (10 "successes" *and* 10 "failures") in mind for valid use of the *z* (but note that the "+4" method for confidence intervals works reasonably well, even if the "10:10" rule is violated).

When we want to perform a test for a proportion, do we still care about the normality of the distribution (of the sample proportion)? If the stats package you are using does the exact method (which Minitab does by default for a single proportion), then no. For testing the difference between *two* proportions, exact methods are not commonly available (then, yes).

I don't understand the theory behind the "+4" method. Why can we add 2 to each category? What makes that legal? The two statisticians who studied improving the C.I. methods came up with a very complicated approach that had much better performance than the standard simple method.. Somewhere along the line, deep in the heart of Binomial darkness, they noticed that if they did the add 2, add 2 thing (for a 95% confidence interval), and used the regular tool on the altered data, this very simple method came very close (almost always) to their fancy approach. So, by a stroke of luck or genius, they hit on a very simple way to lead to better intervals. In more detail: Let z be the multiplier for a classical interval. The A&C method adds $0.5\times z^2$ to the number of "events" and z^2 to the sample size. This addition approximates a way more technical method based on something called the score statistic. When the confidence level is 95%, z is 1.96 so $0.5\times z^2$ is approximately 2, and z^2 is approximately 4. So the so-called "+4" method is an approximation to an approximation.

I don't understand why exactly we use the +4 method. It just seems like a way to actually cheat. Besides that, what exactly at that point is wrong with using the simple method? The confidence interval using the "simple" method doesn't behave well for small samples (in the long run, it won't hit the target as often as you declare (95% or whatever)). The +4 method is more accurate in that sense. Remember, it is just an arithmetic procedure; give me the data, and I'll give you a C.I., done according to whatever recipe. It's not cheating at all.

Can you reiterate the reason why we use the pooled data option for Binomial proportions? When we do a hypothesis test, we start out as though we believe in the null. For the case of two proportions, if that is so, the pooled estimate of p is the best single estimate of the true parameter. The idea is that the two different proportions we have are in fact two estimates of the same thing

(according to the null). Pooling the data averages them, which will give the best possible estimate of *p* (which in turn will give the best SE calculations, and (ultimately) correct p-values.

Tell me again: when can I cut the p-value in half?

7. The computer has given you the p-value from a two-tailed test; you wish to do a one-tailed test (that is, you did not choose the "one-tailed" option for that test).
8. You observe the data agree with your alternate hypothesis (for example, suppose you are testing for the difference in two means, $H_A : \mu_1 - \mu_2 > 0$ and $\bar{y}_1 > \bar{y}_2$ (i.e. $\bar{y}_1 - \bar{y}_2 > 0$. The p-value for this test will be exactly ½ that for the two-tailed test, so you could either select the one-tailed test option in the statistics package or take the two-tailed p-value and cut it in half.

How do we know whether to use the *z* or the *t* when making a confidence interval?
First, keep in mind that, 9 times out of 10, you'll not need to make that choice; it will get made for you by the stats package. For sake of adding to your conceptual understanding, here are (roughly) the circumstances when each is chosen.

Intervals (and *p*-values for tests) are calculated using the *z* distribution (the Normal distribution, scaled to have mean zero, and SD 1) for Binomial proportions (including their use as the response variable in regression models, which is called logistic regression). A *t*-distribution is used for most other situations. The difference (and the reason for the two choices (*z* and *t*) is that for Binomial proportions, the variation in the data is predicted just by knowing (or estimating) the proportion in question. For other (measured) data, knowing the mean does not (without further knowledge or assumptions) tell you anything about the variation. It has to be estimated independently of the mean. That extra bit of estimation causes the appropriate sampling distribution to be slightly different than the Normal (fatter a bit in the tails).
There is one other oddball class of cases where the *z* is used: in some relatively complicated estimation settings no-one knows what the appropriate degrees of freedom should be for using the *t*, so we shrug and just use the *z* instead. This doesn't occur in the classical statistical tools.

Will a 95% confidence interval contain the population parameter 95% of the time regardless of the size of n? Does the sample size have any influence over this property of the CI? It has absolutely none. What does change is that for larger sample sizes, the interval will be narrower. Thus, you can choose *how sure* you want to be (95% or whatever); sample size will determine *how close* you will get.

How do you know what standard error formula to use? I'll answer by asking you a question: what is it you are estimating and how did you choose your sample? Whatever you are estimating (a mean, difference in two means, proportion, difference in two proportions, a slope from a regression model, *whatever*) from whatever study design (simple random sample, stratified random sample, *whatever*) has its own standard error formula. Once you've stipulated the design

and choice of statistic, either the statistics package will immediately use the correct formula or you can, if required, look it up in a text somewhere.

I'm still not clear on the difference between "confidence level" and "confidence interval". A confidence interval is a range (specified by the lower limit and upper limit (e.g.[27] (10, 25)) such that you have a certain amount of confidence (specified by your choice of confidence level (e.f. 95% or whatever) that the parameter you are estimating is actually in that interval (in my little example, between 10 and 25).

How do you choose between reporting SD, or SE, or a C.I.? I see all three in the literature. The three are equivalent: if you tell me one, I can tell you the other two (I assume the report also includes the sample size and study design information, both of which are necessary for my claim). For instance, suppose we are estimating the mean of some population with a sample mean from a simple random sample. The SE of a mean is related to the sample SD and the sample size by its formulation: $SE(mean) = \frac{sample\ SD}{\sqrt{n}}$. Tell me one (SE or SD, and I can tell you the other (the meaning of equivalent). If someone reports a C.I., they surely report the confidence level; it is also relatively easy to get a C.I. given SE, *n*, and confidence level. I won't do the details of that here; I suppose you've seen the formula for a confidence interval, so will believe me. The specific choice (SE or SD of C.I.) is just one of habit. Some disciplines and journals use SE routinely, etcetera. The choice is of no import statistically.

Why do we care about confidence interval more than standard deviation since confidence interval seems to be calculated on hypothetically huge sampling while SD is based on actual data collected? This is actually a pretty deep question. Let's see if I can do it justice. If I flip a coin, it would be commonplace for you to declare, "There is a 50% chance[28] it came up heads." The rationale for this is that in a very long run of (fair) coin flips, there will indeed be 50% heads. You intuitively apply your knowledge of "the long run" to the current circumstance. We do the same thing when, for instance, estimating a mean. If we can predict what sample means from samples like ours would look like "in the long run[29]", we can apply that knowledge to the current (and only) sample mean. In particular, the confidence interval is based on that projection. It is very useful to be able to say (something like) "I am 95% sure the true population mean[30] is between this (lower limit) and that (upper limit). The SD, all by itself does tell us something concrete (it estimates the variability among individual data values in the population from whence we drew our sample), but it doesn't get us very far in estimating the population mean. For

[27] I just made up those numbers for no good reason.
[28] Technically, the word "chance" is a lay-person's term and should not be used informally when speaking of a *past* event. It would be more proper (but sound odd) to say, "I am 50% confident…"
[29] What we call the sampling distribution…
[30] (which I am estimating with my sample mean)

instance, suppose I estimated the mean number of fish per 100m of stream to be 10, and report the SD to be 8. What do you know regarding the precision of my estimate (the 10)? Nada. If you know the sample size is 4, you intuitively wouldn't place much faith in that estimate (and the SE[31] would bear that out (it would be equal to $\frac{8}{\sqrt{4}} = 4,$ yielding a 95% confidence interval of (approximately) (0, 16). On the other hand, if you knew the sample size was 900[32], your faith in that estimate would be higher (SE would be $\frac{8}{\sqrt{900}} = 0.27,$ for which a 90% C.I. is approximately (7.5, 8.5).

How is *z* calculated (if there is no handy table)? What is the formula? Should I (would it be useful for me) know this formula? To calculate *z* requires being able to integrate the Normal distribution to find the point (which we'll call *z*) that divides the Normal distribution into the appropriate sub-areas. For instance the correct value for making a 95% confidence interval is the value that has 2.5% of the curve being to the right (i.e. greater than *z*), and 97.5% to the left (i.e. less than). By the symmetry in the distribution, 95% of the distribution will be between *z* and –*z*. It's not possible to solve this calculus problem in closed form[33] (i.e. there is no formula), which is why *z* tables were created in the first place. So your choices are either a table, or (better yet) get a stats package to do it for you.

How is *t* calculated (if there is no handy table)? What is the formula? Should I (would it be useful for me) know this formula? A *t* distribution is just as difficult to work with computationally as the *z* distribution. See my answer to the *z* calculation question above…

How do we know what confidence level to choose? It is indeed a choice. For most scientific uses, 95% is chosen. Confidence level is usually chosen to be complementary to the alpha level of any tests associated with that same analysis (e.g. 95% is complementary to alpha = 0.05; 90% is complementary to alpha = 0.10, and so on). (By the way, there is a good reason for that: **Read why.)** Ironically, if you choose to use alpha = 0.05 (and 95% confidence level), no-one will ever notice. Should you choose some other level, someone is sure to challenge you on it, and ask, "Why?" I say ironically because you ought to have a reason for the choice no matter what.

Why is 95% the most commonly chosen level in science? Wouldn't it be better to use 99% (more sure of your results)? First, it is usual for the confidence level and the alpha level (if tests are done) to be complementary. **Read why.** Decades ago, when statistical testing procedures were being worked out (using pencil and paper), the researchers realized that life would be much simpler if a single choice (of alpha and C.L) were made[34]. Ronald Fisher (one of the founders of modern statistical practice) decided that a 5% chance of "false significance" (i.e.

[31] As soon as I invoke the SE, I am using that hypothetically huge sampling you refer to (i.e. the sampling distribution of the mean)
[32] Any fisheries biologist who knows how long it takes to sample a 100m reach of stream is laughing at me now…
[33] To actually do it, should you be sufficiently masochistic, would require sophisticated approximation methods.
[34] This is because it literally could take hours to do a single calculation; so to make a table that folks could look up answers might take months.

alpha = 0.05) was a reasonable risk to take. We've stuck with that level ever since. It is not unreasonable, but it *is* arbitrary.

How can you know some samples are not random samples, if someone give you the data for analysis, and you did not do experiment by yourself? You can't. Sometimes you have to think very hard about the data (and about whatever you know regarding the sampling protocol) in order to decide. There are some instances where a non-random sample can be treated as though it were random (effectively). For instance (using change in pockets again), if, each day at 8:00am, I select the first four students I see walk through the doors of the Student Union, I have NOT done a formal random sample. It is a convenience sample. Yet I find it easy to argue that the amount of change in their pockets would not be different than a random sample. For a different question, with that same sample, I might be in trouble. If I were to record, for instance, what their area of study was, the problem would be that it is at least somewhat likely that four people coming into the Union at about the same time might know each other or be coming from the same class, or whatever. Anyway, it can get difficult….

How do you know that original distributions themselves are naturally more symmetric? If you have a large sample, the histogram of the sample will be informative; failing that, you have to dig into examples of similar variables in your literature, and/or consult with others familiar with same.

If there is more variation in the data, would n = 100 still be a good sample size or would the variance decrease the level of confidence and increase the confidence interval? The sample size has two (unrelated) impacts: **(1)** if it is big enough (not to be defined by any one number, as you now know), the sampling distribution of the mean will be approximately Normal. **(2)** As sample size grows, so shrinks the SE of the mean (recall that the sample size is in the denominator of the SE formula). Consequently, a large sample size will shrink the width of a confidence interval. But **(3)** it will not affect one tiny bit the confidence level. You simply choose that, and the choice is based on whether you are a go-along-with-the-crowd scientist (in which case, choose 95%) or you choose it to reflect how sure you need to be with your interval. In a situation with more variation, any confidence interval will be wider than if you had less variation. Is n = 100 enough? It depends on how narrow you need the C.I. to be.

For one data set we studied in class, the sample SD was about 13, but the variance was huge (169). What does that roughly mean in presentation language? Not much. It is the square root of the variance (namely the SD) that connects meaningfully (in our minds) to the variation in the data. For purposes of interpretation, I think of the variance as simply a partial calculation on the way to the SD.

Are there many varieties of t-tables? z-tables? In two senses, yes. First, the electronic t-table I've given you changes (is actually a different table in that it will give different answers) for every difference d.f. you indicate. Second, the t-tables in the back of text books come in several different "presentation" styles. There are varieties of z-tables in this second sense.

Why is the, "SD of the sampling distribution of the mean," also called the "SE of the mean?" Because long, long ago, some statistician decided that having two things running around, both called SD, was confusing. So he named one of them (which is the SD of the distribution of the mean) the SE of the mean; the other (SD of the population of values, which is estimated by the so-called sample SD) he left the same. The choice of names burns me every semester, so I resent him for that choice.

Where do the *z* (and *t*) distributions come from? It was understood very early in the development of statistical methods that, for instance, the sampling distribution of the mean is approximately Normal[35]. Given that, the sample mean itself stood in for an estimate of the mean of that sampling distribution (reasonable if the samples were drawn randomly from the population), and the formula we now know as the SE of the mean estimated the SD of that distribution. Let's symbolize those by $\bar{y}$ and *SE* respectively. Then, for instance, to make a 95% confidence interval, all they had to do was find the points (equally far above and below the mean, due to the symmetry of the curve) in that distribution such that 2.5% of the values lie below the lower number and 2.5% above the upper (the lower and upper confidence limits, respectively). The mathematical formula that represents the Normal distribution[36] (with mean and SD as estimated from the sample) is $\frac{\exp\left(-\frac{1}{2(SE)^2}\left(\bar{y}_f-\bar{y}\right)^2\right)}{\sqrt{2\pi(SE)}}$, where $\bar{y}_f$ (standing here for "future" sample means we have not yet seen) ranges all along the real number line, exp(•) represents $e^{(\bullet)}$, and *SE* is the standard error of the mean, which estimates the SD of the sampling distribution of the mean. Looks a tad daunting, no? With good reason: it can't be done in closed

[35] In fact, in the early days, researchers artificially (and to some extent, often unrealistically) forced this to be so, by assuming that the population from which one has sampled is in fact Normal. You already know that the more symmetric the population distribution, the smaller the sample size required to achieve (at least approximate) Normality in the sampling distribution of the mean. It turns out (and this shouldn't surprise you) that if the sample comes from a Normal distribution, then the means are perfectly Normally distributed, for any sample size. Until recently, when computer simulation studies released us from it, researchers often tortured their data (by doing transformations of various sorts) to make the distribution Normal (or closer than in the original) in order to make this so. These days, most textbooks still honor that history by making the assumption (that the data come from a Normal population) part of the initial presentation of most methods. Then, some pages later (usually), they'll tell you that it isn't a necessity, that sufficiently large samples will get you there.

[36] And what is here is not quite technically correct, but (for those in the know) be patient. I'll attend to the correction in due time.

form. Doing the calculations requires sophisticated calculus-based approximation methods, which complications would send the researcher off into hours of error-ridden tedium[37] every time a confidence interval or p-value from a test was required.

Using some standard statistical theory, someone realized that if we shift the location of the distribution (by subtracting $\bar{y}$ from all the values) and then re-scaled it (by dividing by *SE*), we would arrive at a distribution with the exact same shape, but a different mean (0) and a simpler standard deviation (1). The transformation looks like[38] $z = \frac{\bar{y}_f - \bar{y}}{SE}$ The formula looks like this. $\frac{\exp\left(-\frac{1}{2}(z)^2\right)}{\sqrt{2\pi}}$. This is the formula for the so-called standard Normal distribution, which formula is not a whole lot easier to deal with, but it did enable statistics to become practical. Someone took the time to do a large number of calculations for this distribution, and wrote up the results in a table (the so-called *z*-table). People used (and still do) this table to do calculations for the standard Normal distribution. They could get back and forth from this one to the one they really wanted (the sampling distribution of the mean) by back translating: $\bar{y}_f = (z - \bar{y}) \times SE$ .For instance, in the *z*-distribution, the central 95% of the distribution is between -1.96 and 1.96. This would be a 95% C.I. for that distribution. Take this and re-scale: $-1.96 \times SE$ and $1.96 \times SE$. Then re-locate by adding the estimated mean back in: $\bar{y} - 1.96 \times SE$ and $\bar{y} + 96 \times SE$. Thus was born the use of the *z*-distribution for making confidence intervals.

This simple state of affairs did not last for long (here is where I correct a technical problem with the approach, the solution of which is the *t*-distribution.

William Gossett, while working doing statistical analyses for Guinness Brewery, arrived at the sense that his sample means (from small samples) were a tad more variable than the Normal distribution formula would suggest. He came to realize that the problem was caused by the fact the SE of the mean (denoted in the formulae above simply by *SE*) was itself just an estimate. After some work he arrived at the correct formula for the standardized version of the test statistic (sticking to the game of getting a distribution with mean zero and SD 1, then back-calculating) seems sensible still), which he named the *t*-distribution. I discuss that discovery (including some discussion of the degrees of freedom) in more detail here.

How do we know when to use the *z* distribution, and when to use the *t*?

Generally, speaking,

(1) the *z*-distribution is used for inferences on Binomial proportions (including their use as response variables in regression models (called logistic regression)

[37] Error-ridden if it were me trying to do it by hand, tedious for certain.

[38] Whoever did this chose the letter *z*, perhaps after his Great Aunt Zelda.

(2) a *t*-distribution is used for inferences on means (including in regression models, with try to predict the mean of some response).
(3) As you know, to use a *t*-distribution, one needs to know the appropriate degrees of freedom formula. In some situations (not just simple means), that is not known, and the *z*-distribution is used for simplicity. It's not quite correct, but for decent sample sizes, the two distributions work quite similarly.

When would one choose alpha to be different than 0.05? For conventional scientific purposes, it has become common to accept a 5% chance of false significance[39]. Use of an alpha larger than 0.05 (say 0.10 or 0.15) is fairly common in pilot studies, where the consequences of a false significance are small[40]. Smaller (0.02 or 0.01, say) is used when the consequences of false significance are large[41].

What exactly is the difference between a confidence limit and bound? There isn't actually. You could refer to the lower limit as a lower bound and the upper limit as an upper bound. It has become (somewhat) a convention to use the phrase "confidence bound" for a one-sided interval (so the word "bound" is becoming bound up with that particular beast).

What are Type I and Type II errors? First, the word "error" does not mean that the biologist made a mistake. It means the biologists conclusion (led there by, if you will, bad luck with their data) is not consistent with the state of nature (with regard to the null and alternate hypotheses). The following table is a summary.

		State of Nature	
		Null is True	**Null is False**
Biologist Decision (based on data)	**Rejects Null**	With (chosen) probability α **Names:** significance level, Type I error, alpha-level	With probability β (depends on actual state of nature) **Names:** power of test, beta
	Fails to Reject Null	With (chosen) probability 1-α	With probability 1-β (depends on actual state of nature) **Names:** Type II error

[39] Meaning that one rejects the null hypothesis when in fact it is true…
[40] Usual consequence is that one "wastes" a bit of ones research effort chasing something that turns out to not be true. That's (1) pretty minor compared to publishing a "false" result, and (2) it essentially happens in private.
[41] For instance, suppose declaring a species to be declining in number triggered a $100,000,000 management effort. That's a lot of money (in reality, of course, it has to be balances against the consequences of the opposite error: missing a decline that has actually occurred).

I've heard folks describe the margin of error for a confidence interval in relative terms (e.g. "the margin of error is 10%). Why is that not recommended for confidence intervals for Binomial proportions?
Two reasons (suppose you've estimated 10% "successes and (therefore) 90% failures, with a margin of error of 5% for a 95% C.I.):

(1) "Percents" are commonly the units of measurement for Binomial proportions; in that case, the relative margin of error stated in the same units can be confusing. Further,
(2) The relative m.e. is not very meaningful. With your 10% successes, and m.e. of 5%, you would claim a relative m.e. of 50%. That same margin of error, applied to the failures, looks (relative to the estimate) more like 5.5% of the estimate. So the relative m.e. will change depending on whether you count "successes" or "failures". Not very meaningful.

I was reading some stuff on hypothesis testing and I ran across this statement: "If the null hypothesis is always false, what's the big deal about rejecting it?" (Cohen 1990) If you wish, go to http://www.indiana.edu/~stigtsts/quotsagn.html for more quotes and the reference. Imagine comparing two mice populations (regular diet and reduced calorie diet) for lifespan. Of course the two population means are not equal (equal meaning precisely the same down to the ten trillionth decimal). So in that strict interpretation sense, one could say the null is NEVER true, and, on those grounds, decry its use. I reject that. Anybody who gets hauled to court is unlikely to be 100% innocent, but it is a convenient stance to take for purposes of the trial. On the other hand, if ALL you report is the p-value from a test, it begs the question, "So what IS the effect size?" If ALL you report is the confidence interval, I can easily decide for myself whether the result is significant (in the statistical testing sense). So, if you do a test, please do still report the C.I. On the other hand, the C.I. pretty nicely (without attaching numbers like a p-value) tells the tale.

Can you explain the limitation of hypothesis testing based on "probability of observed result"?? When things are measured on a continuum, the probability assigned to any single point is zero.
Example 1. Suppose someone records their weight at 125 pounds. It's not really 125: it has been rounded. Suppose it is a really cheap scale, and all you can do is say it is between 120 and 130. What proportion of people on the planet has that weight? Just for sake of argument, let's say it is 1 person in 1000 (0.1%). That let's us say that a randomly selected person has probability 0.001 of being in that category. Then they get on a better scale (weighs to the nearest pound) and it is still 125 (so we know it is between 124.5 and 125.4999). That's about 1/10th of the original range (120 to 130); let's suppose the proportion is 1/10th of the previous. Now the probability is down to 0.0001. Do it again with a better scale (measures to the nearest 1/10th of a pound). Down goes the probability to 0.00001. Now switch to a scientific scale which measures to the nearest gram (which, say, is 50 times more accurate than the previous one). The probability drops to 0.0000002. And on and on: the more precisely you measure the person's weight, the smaller will be the number of people in that weight category. Theoretically, if you push the interval to be so narrow as to approach being infinitely small, so will the probability drop to virtually zero.
Example 2: Consider a Normal distribution with mean 10, SD 5 (for no good reason). What is the probability of observing 9.5? Strictly speaking, 9.5 is actually
9.5000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000

00000000000000000000000000000000000000000000000000000000000000000000000000 ... (and onward forever).
The chance of getting that exact value is zero. In reality, of course, we ALWAYS round our values (i.e. record weight to the nearest pound, ounce, gram, milligram, whatever). When we do that, we actually are just recording an interval in which the value falls. So if, say, we were recording things rounded to one decimal, then 9.5 COULD occur (with nonzero probability), because it really represents a value from 9.45 to 9.54999999999. (i.e. an interval). But even then, the probability of any single value, even one very close to the mean (or even the mean itself) would be quite low, so seeking a "low probability of observed result" as a basis for rejecting the null hypothesis would never work.

Could you explain exactly what a prediction interval is used for? I'm a little confused about that. A prediction interval is conceptually similar to a confidence interval, but (and I will use here a single sample, with the mean, as an example)

(1) With a C.I. you are trying to predict the mean of the population, with the sample mean as an estimate, and with precision coming from the sampling distribution of the mean. The C.I. says (with chosen confidence) where you think the mean of the population might be.
(2) In a P.I. you are trying to predict (with confidence) a range of values within which **a randomly chosen individual** might lie. The point estimate is the same: sample mean, but the interval is constructed differently, reflecting its different purpose. A P.I. will typically be much wider than a C.I. because it has to reflect not only the precision of your estimate of the mean, but also needs to build in the variation from individual to individual in the population (usually estimated by the sample SD).

It is not *very* often of interest to create a P.I., but here is an example where it might be. Consider a sample of times between eruptions of Old Faithful geyser in Yellowstone National Park. When folks ask, "How long until the next eruption?" a P.I. is in order. It is of little use to them (aggravating even) if you were to report a very precise estimate of the average waiting time because that doesn't give them any sense of the possible variation from that average that they might be dealing with.

A cautionary note: You understand that in making a C.I. it is important to be confident that you know (at least approximately) that the sampling distribution of the mean is Normal. That, as you know, will happen with a large enough sample size in most situations. A classically constructed P.I. has a much stronger (not oft' met) assumption: **that the population of individual values is approximately Normal**. That is, it is constructed assuming a Normal distribution for the data[42].

Does it matter if the sample sizes in our two groups are not the same? It does not (much). Equal sample sizes tend to produce results that are more precisely valid (i.e. the *t*-distribution is a closer approximation to the actual sampling distribution, but unequal sample sizes (as long as they are not wildly different (like one is triple the other) do no great harm. Some of the desire for equal sample sizes is historical: calculations get really easy with equal sample sizes and

[42] This is yet another reason you don't see P.I.s too often: their use would be more often than not invalid because that assumption is often not valid.

(assumed) equal SD. So there was an emphasis on that in years gone by, an emphasis that is in diminuendo now.

Is the primary reason for looking for normality in the difference of the means so they can be compared? The reason we look for normality for the sampling distribution of the difference is because the p-values and C.I.s created by a typical stats package are predicated on assuming Normality for that sampling distribution. If you decide for some reason that the sampling distribution is NOT Normal, then the usual tools (t-tools) are not valid.

In hypothesis testing, if you are using a complementary confidence level and alpha (95%, 0.05, say), are there situations when the parameter of interest can fall within the CI, but be rejected by the test (or vice versa)? No, simply. That's the strength of choosing complementary values.

Where do I find the t-critical value in Minitab output? You don't (need to): it is 100% synonymous to report the p-value being less than alpha (the critical t-method is the old-fashioned way, long out of regular practice.

Can you ever know that the data is paired if you don't know the details of the study design? It is usually by the details of the design that the pairing is revealed fully. Were subjects (whatever they are) chosen in pairs? Was the treatment randomly applied *within* pairs? "Yes" to those questions indicate pairing.

If the difference of the means and the mean of the differences are the same numerically, why don't the paired and two-sample analyses end up with the same result? As a point estimate, yes: the difference of the means is indeed the same value as the mean of the differences. Standard error is where they differ. In the paired case, the SE of the mean difference is (almost always) smaller than the (two-sample) SE of the differences of the means. That implies that the observed value is seen to be *statistically* further from the null in a properly done paired analysis. Here is an artificial example. Suppose mean diff (equals difference in means) is 3. Suppose SE of difference in means (two-sample style) = 2, and suppose the SE of the mean of the differences (paired style) = 1. In the two-sample analysis we observe that the difference in means is 3/2 = 1.5 SEs from zero (not all that far). In the paired analysis, we get that the mean of the differences is 3/1 = 3 SEs from zero.

When you make a ratio of (say) two medians, and get (say) 1.3, where do we get from that to saying one is 30% larger? We made a ratio of the two medians: $\frac{median_{reduced}}{median_{normal}}$ which equaled 1.3. That tells us the median in the numerator was 30% bigger than the median in the denominator. For instance, if the lower one was 50, the numerator one would be 65 (15 being 30% of 50).

If you do bootstrapping and find that your sampling distribution is not normal, can you not trust the confidence intervals created? On the contrary, the C.I. created by the bootstrap is trustworthy no matter what the sampling distribution, as long as

- the sample size is big enough that it (the sample) is a trustworthy representation of the population, and
- we have a random sample.

The idea is this: we need to have some estimate (think visual, like a histogram) of the sampling distribution of our chosen statistic in order to do inference with it. For only a very small collection of statistics (means, and differences in means being among them) is there any reason to believe the sampling distribution will be approximately Normal. Statistical theory tells us that (the Central Limit Theorem[43]). For many other choices of statistic, we don't know well enough what the sampling distribution might be. Bootstrapping (simulating the process of repeating our study a large number of times) gives us a way to get that estimate.

Why do means and the difference in means tend to come out normal in a t-distribution (with a large enough sample size), but medians and difference in medians would not necessarily be normal? In fact, of ALL the statistics you might choose to study, ONLY means (and differences therein) (and a very few others) are guaranteed to (with sufficient sample size) have a Normal sampling distribution. It's a wonderful, freaky fluke that we have been living off of (for doing analyses) for decades…

Can you tell me the difference between "biological significance" and "statistical significance"? In the context of a hypothesis test (with some chosen alpha level), a "statistically significant" result is a synonym for saying "my observed *p*-value is less than alpha". Conceptually, it says, "Given my stated willingness to reject incorrectly (that's alpha), the data provide sufficient evidence for me to feel comfortable rejecting the null hypothesis." Suppose for simplicity the hypothesis is a test for a difference in two population means. Rejecting the null means that you are reasonably sure the difference is not zero. That does not in and of itself say the difference is big enough to be biologically interesting or important or (hence) significant. It just means it is not likely zero. It could still be pretty small (depending on the confluence of variation in the data and the sample size).

I did not understand the notion of a confidence interval for standard deviation. The sample SD is a statistic (it estimates the population SD). As such, it varies from sample to sample (and if the data come from a Normal population) has a predictable sampling distribution. If estimating the variability in the population is considered important in and of itself, a C.I. might be considered useful. Cautionary note: the validity of a C.I. for the SD (at least that which is routinely available in a statistics package) is quite heavily dependent on the aforesaid Normality.

Can you clarify the difference between an estimator and an estimate, w/ examples of both? It's a technical point, important to statisticians and perhaps not so directly important to others. When you tell me about your estimator, you are telling me what statistic you are choosing. Mean, median, SD, for instance. An estimate is a number, arrived at by applying an estimator to

[43] Actually a large collection of related theorems, but we can think of it as one big idea…

some data. Example: I will use the sample mean as my estimator for the population mean. Given a particular data set, my estimate is _____ (fill in whatever number the sample mean happens to be).

How do you identify whether a single datum is an outlier, and what do you do with it?

Identification: In a single sample of values, an outlier is going to be simply an unusually small or unusually large observation. In more complicated setting (e.g. regression modeling) one has to look at the residuals[44], not the original data themselves. The same notion applies: an outlier is a residual that is unusually small or unusually large. Conventionally, "unusual" is defined as two or three (authors vary on that point) standard deviations above or below the mean. The actual number used (two or three) is not all that important in and of itself.

What then? OK, suppose you've identified an outlier (by whatever mechanism you are comfortable with). First, check the data entry to make sure it is not an entry error. Then it might be worth repeating whatever analysis with and without the outlier. If the conclusions don't change (qualitatively; of course, the actual numbers in your summary will change), no worries. Leave it in. If they *do* change in an important way (your call), then you have a problem. It is not a given that it should be removed. The occasional stray small or large value may be an important part of the process or population you are studying. You might decide to make your inferences to all but those unusual values: "Excepting *giant* three toed sloths, we found that…" Some folks present the results both with and without the odd value, and leave that choice to the reader.

Bottom line: As a statistician, I can help you identify an outlier; I cannot tell you what you should do about (or with) it.

I don't understand why, if the population you are studying is huge, you don't perforce need a large sample size to do a good job. Let's start by defining "good job" as a small standard error (that would imply a precise estimate). And, for simplicity, suppose we want simply to estimate the mean of some variable from a population of 300,000,000. There is a certain amount of natural variation from subject to subject (or plot to plot or whatever) in the population; let's characterize that by some standard deviation, which I will symbolize simply as SD. The formula for the standard error of a mean is $SD/\sqrt{n}$. Suppose we have a sample of 1,600 (which is clearly only a very small fraction of the population). The SE of the mean will then be $SD/\sqrt{1600} = SD/40$. Thus no matter the size of the SD, the SE of the mean is smaller by a factor of 40. A specific example: a guess for the SD of weight (among all people, from tiny babies on up) in the U.S. is about 100 pounds (weights range from around 5 pounds to over 300; a range of about 300 divided by 3 yields a conservative (i.e. large) guess of about 100). A random sample of only 1600 people will yield an SE of about 100/40 = 2.5. A 95% C.I. using this will have a margin of error of about 5 pounds!
The sample size itself dictates the precision, not its relation to the population size. The foregoing statement is *not* true for small populations, interestingly. Consider a population of size 23 (say, a class of graduate students taking a statistics class). A random sample of only 22 of them is

[44] I'll assume for now that you understand what a residual is, and will remove that assumption soon by adding that info into a pertinent section of the FAQs…

almost a census, and, accordingly, the precision is quite better than $n = 22$ would imply. Technically, this gets accounted for by something called the finite population correction, which I won't go into here, as I've already strayed from the main question…

What are the steps to take to do a hypothesis test?

Steps for *any* statistical analysis:

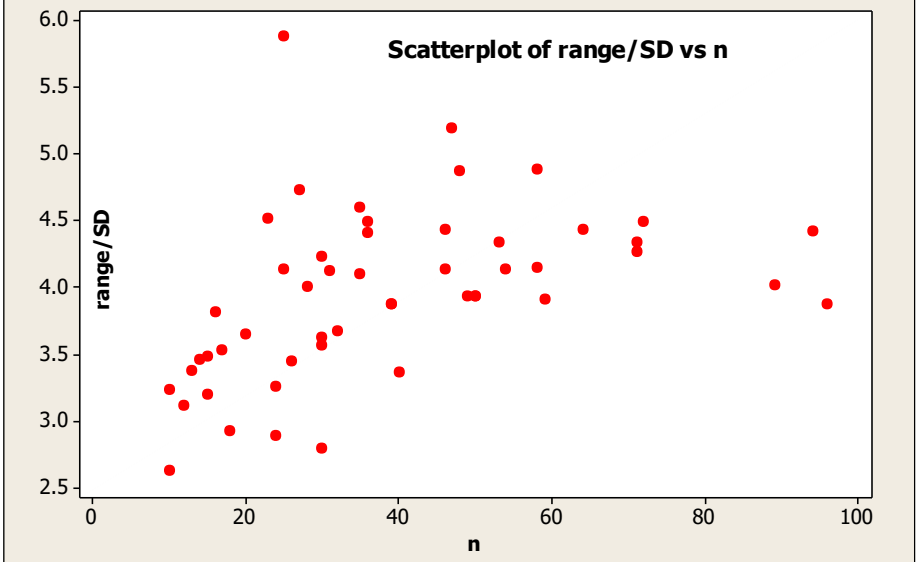
1. Identify the context:
    i. Design: one single sample, two samples, paired samples, …
    ii. Population parameter of interest: a mean, a difference in means, a proportion…
2. Choose a statistical tool:
    i. Identity a candidate tool (*t*-tool, *z*-tool, regression tools, …)
    ii. Be aware of the validity conditions and how to assess them (role of Normality, if any, variance conditions, …)

Steps for Hypothesis Test

1. Choose alpha level (.05 is conventional, but there are often reasons to choose differently)

2. Express the research hypothesis in terms of the parameter of interest
    a. I say "parameter" in the singular intentionally. Notice that, for instance, $\mu_1 - \mu_2$ (difference between two population means) is in fact a single number (arrived at using arithmetic on two numbers).
    b. If appropriate, decide whether the research hypothesis is one-tailed or two-tailed.
    c. The research hypothesis gets formalized as the alternate hypothesis for testing purposes.
3. Given the alternate hypothesis, write the null hypothesis.
4. Do the test, and get a p-value. Details differ (i.e. options to choose) depending on whether the test is one- or two-tailed, since, by default stats packages do two-tailed tests.
5. Make your conclusion by comparing the p-value to alpha (p-value smaller leads to a so-called "statistically significant" finding. Make your conclusion in the language of the research problem.

Why do you claim to use "range divided by three or four" as a data-free estimate of SD?
The following scatterplot comes from over 50 samples (of varying sample size) of data (on a wide variety of measured variables) contributed by students in a statistics class. For each, I computed range (maximum value minus minimum) divided by sample SD. In a case where you need to guess the range, imagine taking a sample of 20 or so, and guess what the largest and smallest might *typically* be. By *typically* I mean to not consider the "largest possible ever" nor the "smallest possible ever" since really unusual values are not commonly seen in samples of size 20 or so.



The graph illustrates that for modest sample sizes (20-30 say) the range is from about 3 to 4.5 or so. Hence my rule: take a guess for the range, and divide by 3 or 4 to get an estimate of the SD. Dividing by 3 will tend to be conservative.

So, if the whole point of regression is to minimize the variance of the residuals, when do we stop? Do we keep squaring, cubing (etcetera) terms till we get diminishing returns on our r squared, then stop? Typically folks don't go beyond quadratic (look in the literature) in that regard. The problem with higher order polynomials is that the higher you go, the more the model is at risk to fitting itself to peculiarities in your specific data set as opposed to expressing some sort of generalized result. So, even though it might work "technically" (i.e. get an improved fit to the data), it often backfires practically.

When we speak of a regression model, what do we mean by a "model"? A model is an abstract simplification of (here) a relationship between the mean of a response variable and changing values of a predictor. The simplest model we use is that of a straight line to predict how the mean of the response ("Y") changes with the predictor ("X").

In a scientific setting, wouldn't you always want to use the model that is the best predictor? Maybe. If you insist on that as your criterion (that is a choice), then yes. What if you got to choose between two predictors (of depression, say). One is a post-mortem brain analysis (which predicts REALLY well); the other is a blood-level of some hormone (which, say, does not predict nearly as well). Which model do you choose (assume the model is to be applied to me)?

What do you check the st. residuals for again? I thought it was outliers. By seeing if they were >2 SD away from the fitted line? The "st" part tells me the residuals are being measured in "number of SDs" above or below the line. "2" or more does indicate a point that is rather far from the line. BUT… if you had, say n = 107, I'd bet on at least several being that far away just by random chance. So the fact of being 2 or more SDs away from the line does not in and of itself say the point is dangerous in any way.

If we look at the relation between two variables, can we say that log transformation is useful for ratio data but not for other? Good point. It doesn't make sense to say "10 times as much" unless the variables are ratio-scale.

What is R^2-adj and how do we use it? If it is largely different (i.e. lower) than R^2, it warns you that you may have too many predictors for the amount of data you have (technically, that you are at risk of having a model that is over-fitting to the oddities of your particular data set, and may be lacking generalizability.

Is the adjusted R^2 what the computer thinks the R^2 would be with more data? Technically, that may not be correct (but a colleague and I have made a note to study it), but it is a beautiful heuristic that describes its behavior quite well. May I borrow it?

Relativity, or using log trans, is a way to get linear regression (the goal is to remove curvature or unequal scatter)? That does indeed happen (whenever a log-transformation is appropriate), but I view it as the appropriate way to capture relativity in a relationship.

Why do we need to choose which method of transforming that is most appropriate if all different ways of logging the variables give us the same relative relationship between x and y? It's not choosing a "method" – it's choosing a base. And the key point is that it matters not (except be sure to do the math correctly). And *that's* important because not everyone knows that and they think you "should" use *e* or "should" use 10.

How does the paradigm of say that "variances are equal" in a two sample test just change to "variances are not equal" I didn't say "variances not equal"; I said "don't assume they are" (in other words, it is still OK if they happen to be, but we don't *require* it) just because of technology. I guess this is a case where we introduce our feelings into our statistical testing. Not at all. In the beginning, no-one even knew how to deal with the sampling distribution without that assumption. If someone started out saying, "OK, let's drop that assumption", all that would happen is that they would wear out a bunch of pencils and erasers trying to figure it out. A guy (last name of Satterthwaite) finally pinned it down in 1946 (by which time the "equal variances" approach had several decades of acceptance, and you know how hard it is to get folks to change!) In those olden (post-Satterthwaite) days, the computational consequences of relaxing that assumption were sufficiently daunting that folks just simply didn't bother. Now all the bother is gone, and the "don't assume equal variances" approach (for the two-sample *t*) is here to stay.

Why is the *z* distribution called "the standard"? What is so special about it that is the "standard"? It is not "the standard" actually. Here is an instance of the meaning of words not being the same in everyday English as in statistics, as applied to distributions. In everyday English, "standard" can mean "usual"; it can also mean "a benchmark: that against which others are compared". In the statistical phrase, "standard Normal distribution", it means neither. All it means is this: a Normal distribution with mean = 0 and SD = 1. The act of "standardizing" (applied here to a Normal distribution, but applied also in many other circumstances) refers only to re-scaling (and sometimes centering at 0) to some pre-chosen convenient scale. As we work with the *t* and *z* distributions, you will see why it is convenient to have them scaled so their SD equals 1.