Richard Gonzalez[1]
Psych 613
Version 3.0 (Nov 2021)

1 LECTURE NOTES #1

1.1 Reading assignment

- Review MD chs 1&2 or G chs 1-5. Acquaint yourself with notation and underlying logic.
- Familiarize yourself with the basics of a statistical package (e.g., SPSS, R)

1.2 Goals for Lecture Notes #1

- Review basic statistical concepts
  - two sample t test
  - role of statistical assumptions
  - hypothesis testing
  - descriptive statistics
  - exploratory data analysis
  - remedial measures
- Introduce SPSS and R syntax

[1]These lecture notes have benefited from feedback given by teaching assistants and students. Previous teaching assistants, or GSIs, in chronological order: Mark Calogero, Steve Holste, Kimi Yamagishi, Ericka Peterson, Henry Schmidt, Lisa Smith, Jane Swanson, Bill Fulton, Dave Waller, Ian Skurnick, William Brenneman, Ajita Gopikrishnan, Shane Mueller, Greg Dyson, David Fencsik, Jennifer Hu, Arran Caza, Vera Sacharin, Alexandra Atkins, Jonathon Kopecky, Julie Maslowsky, Igor Grossmann, Ryan Bremner, Bill Chopik, Josh Wondra, Brian Vickers, Kim Brink, Ben Blankenship, Anne Waldo, Nick Michalak, Esra Ascigil and Koji Takahashi. Current GSIs are Esra Ascigil, Megha Ghosh and Zac Zhang. Olena Sinkevich provided helpful comments on an earlier draft. I also want to thank my University of Washington colleague and friend, John Miyamoto. We had many conversations about how to teach statistics. These discussions usually occurred over food (most often pastries), which provides one of many explanations for why I've gained 30 pounds since I began teaching statistics in 1990.

1. Review syllabus

2. Study tips

   This is a demanding course because the pace is quick. There is a lot of material that needs to be covered during the term. Here are some study tips that will help you perform well in this course.

   - Stay current with the lecture materials

     An ideal strategy would be to attend lecture and **that same night** read the lecture notes from that day. If something isn't clear, be prepared to ask a question (e-mail, office hours, after class). It is important not to fall behind in the lecture material because lectures build on previous lectures. If you don't understand something today, things will get worse tomorrow.

   - Stay current with the textbook reading

     The textbooks supplement the lecture notes. I recommend that you read the assigned chapters each week. I know, reading statistics chapters probably sounds like one of the nerdiest things you could possibly do. But, I believe it helps reinforce your understanding of the material. You want to learn how to analyze and model your data; the lecture notes and textbook readings will give you an understanding of why things are done in particular ways. Such an understanding will be useful throughout your career as new techniques emerge and you will be able to learn them on your own.

     The textbooks (especially KNNL) go into much more detail than is needed in this class. Don't lose yourself in the detail when reading the textbooks. Treat these textbooks as guides that help you tie together loose ends and provide more detailed background. The textbooks are a good source when you want more information about a topic (even more than what you find in wikipedia). The textbooks also allow you to check your understanding by reading the words of different authors with different examples and sometimes different viewpoints.

   - Stay current with the problem sets

     The worst thing you can do is wait until the day before the due date to start the problem sets. If you wait until the night before, you are going to subject yourself to a very frustrating experience. Also, don't try to learn SPSS or R at the same time you are trying to learn the statistics underlying the question being asked. If you don't know the underlying statistics and are simultaneously trying to figure out how to get SPSS to do something (and you may not be sure what that something should be), then I guarantee that you will be frustrated and will learn little from the problem set.

     Here is my suggestion. The day the problem set is distributed take a look at each question. Make a note of what each question is asking you to do. For example, "Question

1 is asking me to do a one-way between-subjects ANOVA, followed by a planned contrast; it is also asking me to check the statistical assumptions". Be sure to write this down in your own words because this is essentially what you will turn in (with computer output sprinkled here and there). When you have each question outlined in terms of what it asks, you have completed the hard part of the homework assignment. The rest is just mechanical use of the computer to get the necessary output. This strategy will help separate your understanding of statistics from the use of the computer. The latter can lead to silly things like a forgotten period in SPSS syntax that, for novices, can take a relatively long time to figure out. If you don't understand what a question is asking, then seek clarification immediately before plunging into SPSS.

Feel free to work in study groups. A good way to conduct a study group is to meet the day after the problem set is distributed. Each member of the study group has done his/her outline (as described in the previous paragraph). The purpose of the group meeting is to go over each problem, to compare each member's outline, and learn from any differences. It is important that everyone do an outline and that the work not be divided among the group members (e.g., I do questions 1-3, you do questions 4-6) because that defeats the purpose of the problem set. You need to master all the material. Any questions or ambiguities that come up in the group that are not resolved can be sent to the e-mail list, brought up in class, or discussed with the instructor/GSIs. Once the outline is complete, then it is just a few routine runs of SPSS or R to get the needed output to fill in the blanks. Even though I am encouraging study groups, I still expect everyone to do their own work in the sense of writing their own answers in their own words and running the computer software themselves. It won't help you learn if you simply watch another person produce output—you don't learn a skill by watching someone else do it (my cello skills won't improve just by watching Yo Yo Ma). Many of the questions in the problem sets ask you to interpret the results. Those interpretations should be in your own words. It is fine to discuss interpretations in the study group, but it is unacceptable for every group member to have the identically worded interpretation.

3. Intuitive review of a few concepts from Introductory Statistics (you might find it helpful to review an introductory, undergraduate textbook)

Key definitions from Introductory Statistics

(a) Some definitions

i. **mean** (aka $\overline{x}$, μ, $\hat{\mu}$): take a sum of n scores and divide that sum by n; the intuition for the mean is "center of gravity" (illustrated by a histogram where the mean is the position of the fulcrum that balances the distribution); know the difference between sample and population mean

ii. **median**: the score at (or near, by some appropriate definition) the 50th percentile

iii. **variance** (aka s^2, σ^2, $\hat{\sigma}^2$): a measure of variability around the mean defined as the "average" of the squared deviations from the mean; the word average is in quotes because the denominator depends on degrees of freedom, e.g., if using a sample to estimate a population variance one divides the sum of squared deviations from the mean by n - 1 (intuitively, one degree of freedom is lost because the estimate of the variance depends on another estimated number, the mean); know the difference between the sample variance and the population variance

iv. **standard deviation** (aka s, σ, $\hat{\sigma}$: the square root of the variance): know the difference between the standard deviation and the standard error

v. **interquartile range** (aka IQR): the difference between the score corresponding to at (or near) the 75th percentile and the score corresponding to at (or near) the 25th percentile

CLT

(b) Central Limit Theorem (CLT)

The key idea is that the sampling distribution of the mean has known properties. In words, the theorem gives the properties of the mean of the means and variance of the means from repeated sampling. Put simply, take repeated samples of size n (with replacement) from some population and calculate the mean of each sample. Denote the mean of sample i by $\overline{Y}_i$. You can compute the mean of the means (i.e., the means of all the $\overline{Y}_i$'s), the variance of the sample means, and you can construct a histogram of the sample means. The central limit theorem states that as the size of the samples gets large, the distribution of the means becomes normally distributed with $E(\overline{Y}) = \mu$ and var = $\frac{\sigma^2}{n}$ (where the function E denotes expectation, or average). Thus, the CLT shows that the mean and variance of the sampling distribution of the mean is related to the mean and the variance of the original parent population, respectively. Figure 1-1 illustrates a normal distribution and the resulting sampling distribution when the sample size n = 5.

I'll present a small computer simulation in class to illustrate this concept.

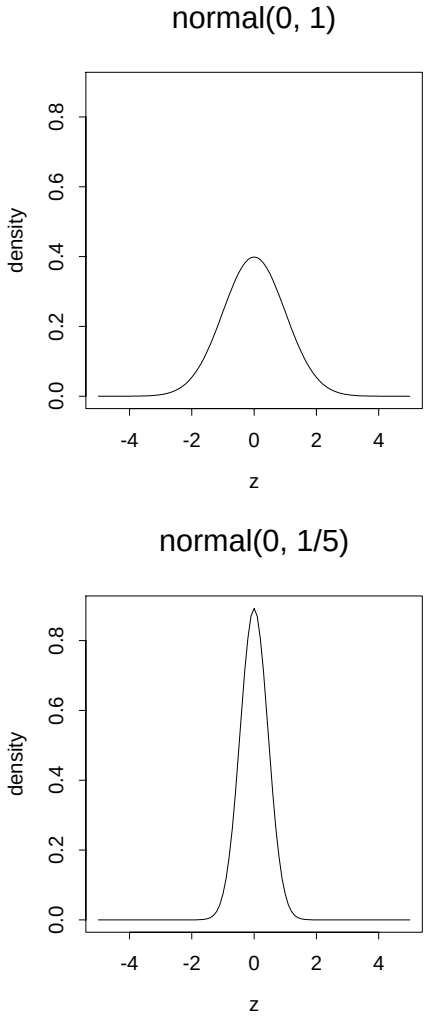
For a more detailed treatment, including mathematical proof, of the central limit theorem consult a mathematically-oriented, introductory textbook such as *A First Course in Probability* by Ross (but to understand this proof you need a good understanding of some advanced mathematical concepts).

(c) *t* distribution

Intuitive definition of t

In this course we will use the simple and elegant result that for normally distributed parameters

Figure 1-1: Illustration of the central limit theorem.
The upper panel shows a Normal distribution with population $\mu = 0$ and population $\sigma^2 = 1$. We take repeated samples of size 5 from this "parent" distribution. The lower panel shows the sampling distribution with $\mu = 0$ and $\sigma^2 = 1/5$. The sampling distribution corresponds to the distribution of means from samples with n = 5 taken from the parent distribution.



$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}} \quad (1\text{-}1)$$

(sometimes the denominator is called the estimated standard error). The right hand side of Equation 1-1 can be thought of as a standardized parameter estimate, where the parameter is divided by a measure of its sampling variability.

Equation 1-1 says that whenever a normally distributed estimate is divided by its estimated standard error, the resulting ratio follows a *t* distribution. The t distribution is very similar to the normal distribution except that, for small samples it has longer tails. As the sample size gets large, the t distribution approaches the normal distribution.

This is a useful result. Most of the estimators we deal with in this course are normally distributed (e.g., means, difference between two means, contrasts, regression coefficients, correlation coefficients when the population correlation is assumed to be $\rho = 0$). Thus, we can use the t distribution in many situations. The degrees of freedom corresponding to the denominator of Equation 1-1 will depend on the particular application, more on this later.

As with much of mathematics, the game we play in statistics is to develop a handful of tools that are useful across different situations. One such tool is the t distribution. Whenever we have a situation for which Equation 1-1 applies (that is, we have an estimate divided by its standard error), we can make use of the t distribution. So, the goal in many applications will be to tinker with the details of the problem so that we have an estimate and its standard error, then we can apply Equation 1-1; in other words, we will convert something we don't know how to handle into something we *do* know how to handle.

Excel and *t* values

It is convenient to use a *t* table to get the necessary values. For instance, with 20 degrees of freedom (explained later), the *t* value is 2.09 as given by the tabled values. If you want to compute your own tables, you can use a spreadsheet such as Microsoft's Excel. For instance, the excel function TINV gives the *t* value corresponding to a particular two-tailed α and a particular degrees of freedom. If you type "=TINV(.05,20)" (no quotes) in a cell of the spreadsheet, the number 2.085963 will appear, which is the *t* value corresponding to a two-tailed α of 0.05 with 20 degrees of freedom.

R and *t* values

For users of R the function qt() provides the *t* value for a given α-level and degrees of freedom. If you want the two-tailed *t* corresponding to α = 0.05 with 20 degrees of freedom, then the command

```
qt(.975,20)
```

produces 2.085963 just as the Excel command above.

Excerpt from the t table ($\alpha = .05$, two-tailed)

df	t value
1	12.71
⋮	⋮
20	2.09
⋮	⋮
100	1.98
⋮	⋮
∞	1.96

(d) Confidence Intervals (CI) and Hypothesis Testing

The CI is generally more informative than the result of a hypothesis test. Review an intro book if these concepts are not familiar. Here are two elementary examples of both the CI and the standard hypothesis test.

Intuitive definition of the CI

i. Example 1: One sample CI and *t* test

$$\text{estimate} \pm \text{margin of error} \tag{1-2}$$

$$\overline{Y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{1-3}$$

where s is the sample st. dev. and $\overline{Y}$ is the sample mean. The confidence interval yields a lower and an upper "bound" around the point estimate $\overline{Y}$. Note that the "margin of error" is composed of three ingredients: *t* at the desired α-level with degrees of freedom = n - 1 (this is merely a "table lookup"), the variability in the sample, and the sample size. Think about what happens to the "margin of error" as you vary these three parameters.

Simple numerical example. Imagine we have a group of 21 people and their average age is 30 with a standard deviation of 2. To compute the 95% CI we compute the standard error as $\frac{s}{\sqrt{n}} = \frac{2}{\sqrt{21}} = .436$. The corresponding t value with 20 degrees of freedom is 2.09 from the table so the margin of error is 2.09*.436 = .911. The confidence interval is then the mean plus or minus the margin of error, or [29.089, 30.911].

The standard frequentist interpretation of a 95% confidence interval is that 95% of

such intervals (under repeated sampling, much like we saw in the case of the central limit theorem) contain the true population value. That is, the population value is a fixed number, each study of sample size N provides an estimate and a CI around that estimate. The standard theory asserts that it is the information in the study that is variable—the true population value does not change. If you repeatedly performed study after study, 95% of these intervals would include the true population value and 5% would not.

The hypothesis test for this situation has an analogous form. Let's assume that the null hypothesis is that the population mean $\mu = 0$ (but, more generally, the null hypothesis can be $\mu =$ k where k is any real number).

$$t_{\text{observed}} = \frac{\text{estimate of mean}}{\text{st. error of mean}} \quad (1\text{-}4)$$

$$= \frac{\overline{Y}}{s/\sqrt{n}} \quad (1\text{-}5)$$

with degrees of freedom = n - 1. After computing the observed t from Equation 1-5, one would then compare the computed t to the tabled value at the desired α level and n - 1 degrees of freedom.

Simple example continued. Recall we have N = 21, mean = 30 and standard error = .436. Suppose we are testing whether the average age of this sample of 21 differs from a norm we have of, say, 28. In this case the null hypothesis is 28 and we want to test whether the observed mean of 30 is sufficiently different from the norm. The observed t is $\frac{30-28}{.436} = 4.59$. The critical t with 20 degrees of freedom is 2.09. The observed t exceeds the critical t (in absolute value sense) so we reject the null hypothesis and can claim that this sample of 21 is likely from a different population than the norm. The p value is .0002.

Hypothesis testing template

It is useful to introduce a template for the hypothesis test that I will use throughout the year. All hypothesis tests proceed in the general manner described in Figure 1-2. For the specific case of a one sample *t* test the template can be completed as shown in Figure 1-3.

The standard frequentist interpretation of $\alpha = .05$ means that if the null hypothesis is true, then under repeated sampling 5% of the studies would incorrectly reject the null hypothesis. That is, the null hypothesis is treated as a fixed number, each study provides a test of the null hypothesis (the information in the study is what is variable, not the fixed value of the null hypothesis). If the null hypothesis is true and you repeatedly perform study after study (under the same conditions), 5% of these studies will incorrectly reject the null hypothesis.

I hope you can see that there is a relation between the CI and the hypothesis test.

Figure 1-2: General Hypothesis Testing Framework

Null Hypothesis The null hypothesis H_o and its alternative H_a are stated.

Structural Model and Test Statistic One states the underlying model (more on this later), its parameters, and test statistics that will evaluate the performance of the model against data relative to the null hypothesis.

Critical Test Value Each hypothesis test as a critical value, or "a number to beat," in order to make a statistical decision. This number is dependent on other features such as degrees of freedom, number of tests made, one- or two-sided hypothesis test, etc.

Statistical decision Reject or fail to reject the null hypothesis.

Both are very similar in their setup. The hypothesis test is identical to checking whether the confidence interval includes the value of the null hypothesis. For example, if the CI around a sample mean has a lower bound of .8 and an upper bound of 12.1, I immediately know (without having to do any additional computation) that I would reject the null hypothesis that the population mean $\mu = 0$. The reason is that the CI does not include 0 within its interval. A second example: if the CI around a sample mean has a lower bound of -2.4 and an upper bound of 14.3, I immediately know that I would fail to reject the null hypothesis because the interval *does* include the value of the null hypothesis.

There is a sense in which the CI is more informative than the hypothesis test because the latter is included in the former. The CI also provides information about variability around the parameter estimate. One could criticize the null hypothesis test for confounding two pieces of information—effect and its variability get lumped into a single number. So a t ratio of 2 could arise in many different ways (e.g., the effect is small and its variability is small, the effect is large and its variability is large, etc). However, the CI keeps those two terms separate in that one sees the center of the interval (the estimate) separate from the width of the interval (its variability).

ii. Example 2: Two sample *t* test and CI.

This example involves two groups. We will use the trick of converting this new problem into something that we already know how to solve. Let $\overline{Y}_1 - \overline{Y}_2 = D$, denote the difference of two means. We know from section 3c above that

$$t \sim \frac{D}{\text{st. dev. of the sampling dist of D}} \tag{1-6}$$

For this application the degrees of freedom = $n_1 + n_2 - 2$.

Figure 1-3: Hypothesis test template for the one sample *t* test

Null Hypothesis

- H_o: $\mu = 0$
- H_a: $\mu \neq 0$ (two-sided test)

Structural Model and Test Statistic

The structural model is that the dependent variable Y consists of a grand population mean μ plus random noise ϵ. In symbols, for each subject i his or her individual observation Y_i is modeled as $Y_i = \mu + \epsilon_i$.

The test statistic operates on the population mean μ and specifies its sampling distribution. The test of the hypothesis will involve an estimate over the standard error of the estimate, therefore we make use of the definition of the *t* distribution (Equation 1-1)

$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

Using the statistical results stated above we write the specific details for this problem into the definition of the *t* distribution

$$t_{\text{observed}} = \frac{\overline{Y} - \mu}{s/\sqrt{n}}$$

Critical Test Value Because we are using a test statistic based on the *t* distribution, we use the *t* table to find the critical value of *t*, denoted t_{critical}. We decide the α level (such as $\alpha = 0.05$ two-tailed), then do a table lookup to find the critical value. For instance, if we want to perform a two-tailed test with $\alpha = 0.05$, the critical *t* value with 20 degrees of freedom will be 2.09 (that is, t_{critical} = 2.09). This acts as the cutoff in the next step; the theoretical gauntlet has been thrown.

Statistical decision If the observed *t* computed from the raw data (the second step) exceeds in absolute value terms the critical value t_{critical}, then we reject the null hypothesis. If the observed *t* value does not exceed the critical value t_{critical}, then we fail to reject the null hypothesis. In symbols, if $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

Assumptions of the t test

We need to find the sampling distribution of D (i.e., the denominator of Equation 1-6). If the two samples are **independent**, then we can use the following result:

$$\text{var}(\overline{Y}_1 - \overline{Y}_2) = \text{var}(\overline{Y}_1) + \text{var}(\overline{Y}_2) \quad (1\text{-}7)$$

$$= \frac{\text{var}(Y_1)}{n_1} + \frac{\text{var}(Y_2)}{n_2} \quad (1\text{-}8)$$

Equation 1-7 uses the independence assumption, which allows us to write the variance of a difference of two means as a sum of two variances. Equation 1-8 uses the CLT to write the variance of each mean as a function of the population variance divided by sample size.

If both groups have **equal population variances**, denoted

$$\text{var}(Y_1) = \text{var}(Y_2) = \text{var}(Y)$$

(note that there are no "bars" over the Y's because we are talking about the variance of the population Y not the variance of means $\overline{Y}$), then Equation 1-8 reduces to

$$\text{var}(\overline{Y}_1 - \overline{Y}_2) = \text{var}(Y)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad (1\text{-}9)$$

This equation states the original variance we need (the variance of the difference between two means) is, under the assumptions made, equal to the variance of the population times a factor that depends on the sample size. The right hand side of the last equation is not hard to work with. Recall that to get here we had to assume that (1) data were independent and (2) the two population variances were identical.

All we've done is change the problem from trying to find $\text{var}(\overline{Y}_1 - \overline{Y}_2)$ to finding the easier quantity $\text{var}(Y)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. We performed that move by assuming independence and equal population variances. To compute var(Y) we can pool our individual estimates of $\text{var}(Y_1)$ and $\text{var}(Y_2)$. We will denote the estimate of the pooled variance as

pooled error term

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)\text{var}(Y_1) + (n_2 - 1)\text{var}(Y_2)}{(n_1 - 1) + (n_2 - 1)} \quad (1\text{-}10)$$

$$= \frac{\sum_{j=1}(Y_{1j} - \overline{Y}_1)^2 + \sum_{j=1}(Y_{2j} - \overline{Y}_2)^2}{\text{degrees of freedom}} \quad (1\text{-}11)$$

The reason for this high level of detail is to highlight how crucial the assumptions of independence (Equation 1-7) and equality of population variances are in the two sample *t* test. The classic two-sample *t* test depends on these assumptions in a

fundamental way. If the assumptions do not hold, then the results of the t test are suspect.

p-values

The hypothesis testing template doesn't specifically mention p-values. That's by design. Technically, when we perform a hypothesis test the issue is only whether or not the observed test statistic (such as t observed) is more extreme than the critical value from the table lookup. If your t observed is equal to the critical value, then the p-value is .05. It turns out that the t critical (two-tailed, α = .05) can be stated in the p-value scale as .05. The p-value is another way to express the t observed, and we can compare our observed p-value to the critical α = .05 level. The decision to reject or fail to reject is identical regardless of whether we use t observed and t critical, or we use p-value and α.

A p-value can serve as a measure of how far one is from the critical test value. As your observed test statistic moves away from the critical value, then the p-value is less than .05. The decision (reject, fail to reject) remains the same. Traditionally, the view has been that there is relatively little info in the p-value. To make sports analogies, all that matters is that you cross the endzone, not how far into the endzone you go; a ball is over the line in tennis, how far over doesn't matter; when a pitch is outside the strike zone it is a ball, it doesn't matter how far outside the strike zone.

I wrote a paper years ago (Greenwald et al, 1995; we'll talk about it later in the semester) about what information one can extract from a p-value and the answer is relatively little. But if you do report p-values you might as well report the complete value (like p = .023 rather than p ¡ .05) so that you can extract that little bit of extra info, which we'll talk about later. We already know whether or not the test is significant (aka null hypothesis rejected), so also writing down p ¡ .05 in your paper is not necessary. To make things worse, as people began to realize that p-values didn't provide the info they thought, they started adding more things to the list of what to report. So some fields now report not just t observed, degrees of freedom, whether the difference between the two means is statistically significant, and confidence intervals (say in a graph with means and +/- 1 standard error around the means) but also p-values, effect sizes, and power. It turns out though that much of the extra information is completely redundant with info already reported (means, st dev, sample sizes, t observed), and if some of the info is provided the other missing pieces can be computed.

Now, we return to the original question of finding a confidence interval around the difference between two means. The CI is simply

$$\text{estimate} \pm \text{margin of error} \tag{1-12}$$

$$\text{D} \pm t_{\alpha/2}\ \text{s}_{\text{pooled}}\sqrt{1/\text{n}_1 + 1/\text{n}_2} \tag{1-13}$$

Figure 1-4: Hypothesis test template for the two sample *t* test

Null Hypothesis

- H_o: $D = 0$
- H_a: $D \neq 0$ (two-sided test)

Structural Model and Test Statistic

The structural model is that the dependent variable Y consists of a grand population mean μ plus a treatment effect α_j plus random noise ϵ (we'll come back to this later). In symbols, for each subject i his or her individual observation Y_i is modeled as $Y_i = \mu + \alpha_j + \epsilon_i$.

The test statistic operates on the difference between two population means D and specifies its sampling distribution. The test of the hypothesis will involve an estimate over the standard error of the estimate, therefore we make use of the definition of the *t* distribution (Equation 1-1)

$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

Using the statistical results stated above we write the specific details for this problem into the definition of the *t* distribution

$$t_{\text{observed}} = \frac{\text{D}}{\text{s}_{\text{pooled}}\sqrt{1/\text{n}_1 + 1/\text{n}_2}}$$

with df = $\text{n}_1 + \text{n}_2 - 2$.

Critical Test Value We use the *t* table to find the critical value of *t*, denoted t_{critical} for the specific degrees of freedom, two-sided, and $\alpha = 0.05$.

Statistical decision If $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

The hypothesis test that the population value of D = 0 has an analogous form,

$$t_{\text{observed}} = \frac{\text{D}}{\text{s}_{\text{pooled}}\sqrt{1/\text{n}_1 + 1/\text{n}_2}} \tag{1-14}$$

with df = $\text{n}_1 + \text{n}_2 - 2$.

Thus, the CI conveys the same information as the (two-tailed) hypothesis test. But, unlike the hypothesis test the CI does not confound the size of the effect with its variability (i.e., does not lump both numbers into a single ratio but keeps them separate). The CI format will become very useful later in the course. Several journals are moving in the direction of encouraging authors to report confidence intervals rather than *p*-values from inferential tests.

(e) Recap: assumptions in the two sample t test

i. **independent samples**

ii. **equality of population variances**

iii. **normality** (in practice what tends to be critical is that the distributions be symmetric)

SPSS syntax

4. SPSS syntax

Outline of SPSS syntax[2] for a two-sample t-test

```
t-test groups = GROUPING VARIABLE
 /variables = DEPENDENT VARIABLE.
```

You may want to graph each mean separately with its own standard error. The SPSS syntax outline for 95% confidence intervals is

```
graph
  /errorbar (CI 95) DV by GROUP.
```

Similarly, the SPSS syntax outline for plus or minus one standard error is

```
graph
  /errorbar (STERROR 1) DV by GROUP.
```

The error bar of plus or minus one standard error roughly corresponds to the hypothesis test, in the sense of whether or not the error bars overlap corresponds to the decision of whether or not to reject the null hypothesis. Another way to say this, is that the t_{critical} value is approximately 2 (the exact value depends on the degrees of freedom for the specific problem as well as

[2] I will adopt the convention that capitalized words are names chosen by the user. For instance, you might have a dependent variable in your dataset that you call "precall" for percent of words correctly recalled. In the syntax below, you would replace "DEPENDENT VARIABLE" with "precall".

other things such as how the standard error of the difference is define and whether or not we assume equal population variances). The number "to beat" is approximately two. Intuitively, by plotting one standard error from one mean and a second standard error from the other mean, we have the criterion that if the two error bars do not overlap this roughly corresponds to a significant t test (because the nonoverlapping error bars roughly corresponds to a t of 2).[3]

I don't know of an easy way to get SPSS to plot the 95% CI around the difference of the two means (where the difference is the estimate around which the CI is computed). The previous two commands give intervals around individual cell means not the difference between two means. The numerical confidence interval around the difference between two means is presented in the output of the T-TEST command, but unfortunately SPSS doesn't seem to plot it.

SPSS Example

Here is an excerpt of the example in Appendix 2. This example has an outlier that we will discuss a little bit later.

Data for an experiment with two independent groups:

Group 1	Group 2
3	4
4	5
5	6
4	5
3	4
4	5
5	6
4	5
3	4
4	11

Note that Group 2 has one outlier; otherwise, the scores in Group 2 are equal to Group 1 scores plus one.

[3] If you want to be exact in having overlapping bars correspond to the test of significance, you also need to take into account other factors. For example, for two groups with equal sample sizes using the classic equal variance t test, the denominator of the t observed is $s_p\sqrt{\frac{2}{n}}$, so there is a $\sqrt{2}$ factor that needs to be taken into account in defining the width of the interval around the mean. Also, instead of just saying the t critical is 2, one should use the exact critical value for t (e.g., for 20 df, the exact critical t is 2.086, so that means we need half of 2.086 for each of the two means). Being careful in this way, one can construct a graph where one can use the criterion of nonoverlapping intervals and reach the identical conclusions as the test of significance. Some people, like Geoff Cumming (2013, *Psychological Science*), have argued that we should just use 95% CIs around the mean and forget completely about doing hypothesis testing, and also not trying to scale the figure, like I did earlier in this long footnote, to connect with hypothesis testing results. There still is no convention about how to plot error bars so you should be clear what you are plotting such as stating "error bars reflect plus/minus one standard error" or whatever

The SPSS syntax to read the data and run the two sample t test:

```
data list free /group dv.
begin data.
1 3
1 4
1 5
1 4
1 3
1 4
1 5
1 4
1 3
1 4
2 4
2 5
2 6
2 5
2 4
2 5
2 6
2 5
2 4
2 11
end data.

 t-test groups=group(1,2)
   /variables=dv .
```

```
t-tests for Independent Samples of GROUP


                            Number
Variable                   of Cases       Mean          SD   SE of Mean
-----------------------------------------------------------------------
DV

GROUP 1                      10         3.9000        .738         .233
GROUP 2                      10         5.5000       2.068         .654
-----------------------------------------------------------------------

         Mean Difference = -1.6000

         Levene's Test for Equality of Variances: F= 2.204  P= .155


      t-test for Equality of Means                                   95%
Variances   t-value       df    2-Tail Sig     SE of Diff          CI for Diff
-------------------------------------------------------------------------------
Equal         -2.30       18          .033           .694       (-3.059, -.141)
Unequal       -2.30    11.25          .041           .694       (-3.124, -.076)
-------------------------------------------------------------------------------
```

5. R syntax

R syntax

One way to get two-sample *t* tests in R is through this command

```
t.test(dependentvar ~ groupingvar, var.equal=TRUE)
```

where dependentvar is the name of the dependent variable, groupingvar is the name of the group variable, and the var.equal subcommand denotes we assume equal variances.

Graphing error bars in R requires some knowledge of setting up plots, which we'll cover later. There are several libraries that offer error bar capabilities, including the packages gplots (function plotCI), Zelig (function plot.ci), plotrix (function plotCI), and the package ggplot2 (function geom_errorbar).

Here is the same example but using R. I put the data in file called example.dat and then read it into R. You'll need to edit the "PATH/TO/FILE/" part to where you saved the example.dat file and uncomment the setwd line by deleting the hashtag.

```
# setwd('PATH/TO/FILE/example.dat)
data <- read.table("example.dat", header = T)
data <- data.frame(data)
```

For completeness, here are the contents of the file example.dat.

```
group dv
1 3
1 4
1 5
1 4
1 3
1 4
1 5
1 4
1 3
1 4
```

```
2 4
2 5
2 6
2 5
2 4
2 5
2 6
2 5
2 4
2 11
```

Next we make the column called group a factor so that commands know to treat it as group 1 and group 2 rather than the numbers 1 and 2. Here I specify the column called group and convert it to a factor.

```
data$group <- factor(data$group)
```

I use the data argument and assign it the value data; this allows R to find the variables called group and dv because they "reside" in the data.frame called data.

```
t.test(dv ~ group, data = data, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dv by group
## t = -2.3041, df = 18, p-value =
## 0.03335
## alternative hypothesis: true difference in means between group 1 and grou
## 95 percent confidence interval:
##  -3.058927 -0.141073
## sample estimates:
## mean in group 1 mean in group 2
##             3.9             5.5
```

The rest of this section presents some R tips. The output looks ugly; R has several ways to "tidying" output, including packages to put output in relevant APA format. Here I illustrate by "piping" the output of t.test() into tidy then selecting the columns estimate to conf.high then creating a kable printing 2 significant digits.

```
library(tidyverse)
library(broom)
t.test(dv ~ group, data = data, var.equal = T) %>%
    tidy() %>%
    select(estimate:conf.high) %>%
    kable(digits = 2)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
-1.6	3.9	5.5	-2.3	0.03	18	-3.06	-0.14

Note that I've had to load in two packages (aka libraries) to take advantage of tidy and formatted table kable(). It can be a pain loading libraries throughout a session. I tend to put all libraries at the beginning of a session, using another package called pacman that checks if the requested packages are installed. The code below looks complicated because I first check if pacman is installed (if not, the code will install pacman), then load the pacman package, then list all the libraries I need in my session (I list the packages used to produce Lecture Notes #1). The p_load() command will automatically install requested packages if they are not already installed.

```
if(!("pacman" %in% installed.packages()[,"Package"]))
    install.packages("pacman")
library(pacman)
p_load("tidyverse","broom","knitr","formatR","brms",
       "bayesplot","boot")
```

6. Exploratory Data Analysis

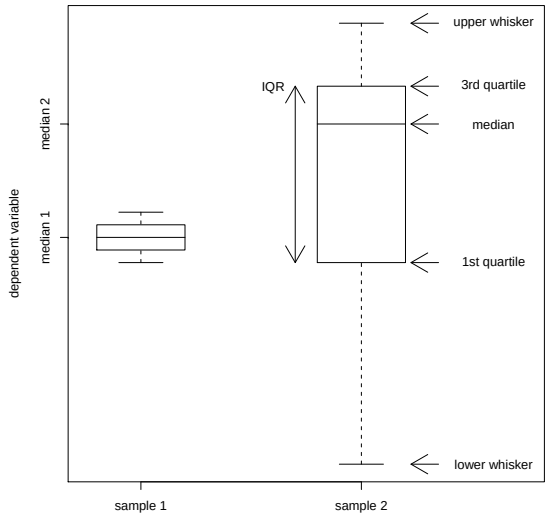
A boxplot is a "robust and resistant" graphical technique for examining assumptions such as normality (or symmetry) and equality of variances. These graphics also can provide a visual assessment of the effects of a transformation, which will become one tool we can use when dealing with violations of assumptions.

Boxplots require knowing the median, the quartiles, the smallest and largest values. Review an intro book for more information. The boxplot is depicted and summarized in Figure 1-5. The whiskers usually are drawn to the data point closest to but not exceeding 1.5 interquartile ranges. Data points that exceed the whiskers are plotted as individual points and may be potential outliers.

Boxplot in SPSS

The SPSS syntax for a boxplot is

Figure 1-5: The boxplot.




```
examine variables = DV by GROUP
 /plot boxplot
 /nototal.
```

Boxplot in R

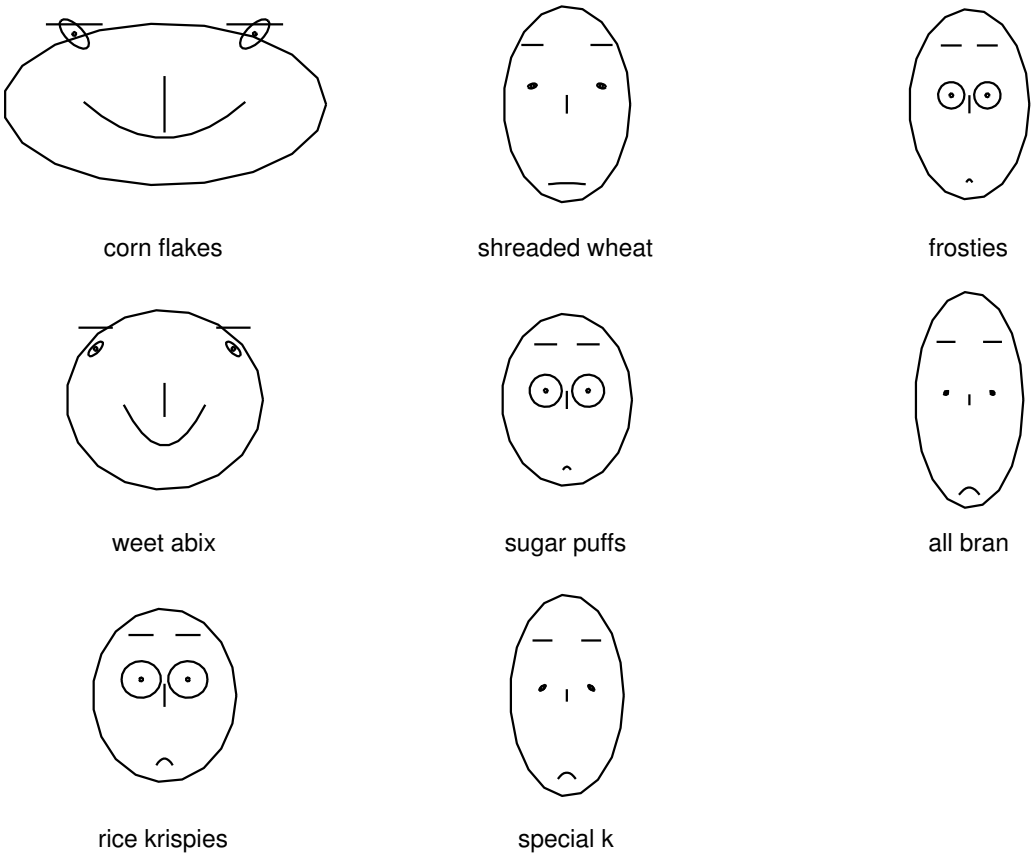
The standard way of plotting boxplots in R is through the boxplot command. The basic command is

```
boxplot(dependentvar ~ independentvar)
```

where dependentvar is the name of the dependent variable and independentvar is the name of the independent variable.

An aside: Chernoff's faces. Cereal example. Consumers rated several cereals on several dimensions such as "fun for children to eat". Chernoff had the creative idea of taking a caricature and coding the features according to the value of the variables. For example, we can represent the value on the variable "a cereal you come back to" by the area of the face. Similarly, the shape of the face codes "taste nice", the length of the nose codes "popular with the family", location of mouth codes "very easy to digest" (yes there comes a time on one's

Figure 1-6: Chernoff's faces.

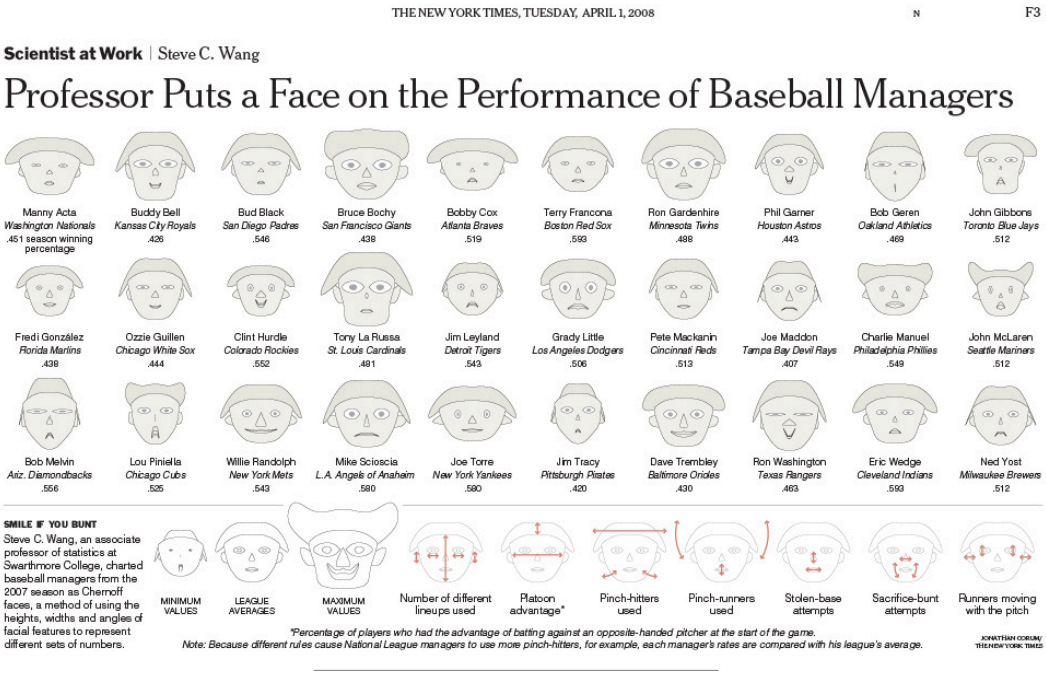


life when concern with digestion kicks in), curve of the smile codes "nourishing", width of the mouth codes "natural flavor", etc.

Well, I was surprised in 2008 to open up the *New York Times* and see Chernoff faces reported as though they were something new (Figure 1-7). The story made it seem that this data analysis technique represents more information than is really there. Anyway, it was fun to see something from statistics make it into the *Times*, despite the fact that it appeared on April Fools Day.

These techniques, generally called exploratory data analysis (EDA), allow you to understand your data. There are many more techniques in the EDA tradition that we will discuss throughout the course, especially when examining residuals.

Figure 1-7: Chernoff's faces.



THE NEW YORK TIMES, TUESDAY, APRIL 1, 2008 N F3

Scientist at Work | Steve C. Wang

Professor Puts a Face on the Performance of Baseball Managers

Manny Acta, *Washington Nationals*, .451 season winning percentage
Buddy Bell, *Kansas City Royals*, .426
Bud Black, *San Diego Padres*, .546
Bruce Bochy, *San Francisco Giants*, .438
Bobby Cox, *Atlanta Braves*, .519
Terry Francona, *Boston Red Sox*, .593
Ron Gardenhire, *Minnesota Twins*, .488
Phil Garner, *Houston Astros*, .443
Bob Geren, *Oakland Athletics*, .469
John Gibbons, *Toronto Blue Jays*, .512
Fredi González, *Florida Marlins*, .438
Ozzie Guillen, *Chicago White Sox*, .444
Clint Hurdle, *Colorado Rockies*, .552
Tony La Russa, *St. Louis Cardinals*, .481
Jim Leyland, *Detroit Tigers*, .543
Grady Little, *Los Angeles Dodgers*, .506
Pete Mackanin, *Cincinnati Reds*, .513
Joe Maddon, *Tampa Bay Devil Rays*, .407
Charlie Manuel, *Philadelphia Phillies*, .549
John McLaren, *Seattle Mariners*, .512
Bob Melvin, *Ariz. Diamondbacks*, .556
Lou Piniella, *Chicago Cubs*, .525
Willie Randolph, *New York Mets*, .543
Mike Scioscia, *L.A. Angels of Anaheim*, .580
Joe Torre, *New York Yankees*, .590
Jim Tracy, *Pittsburgh Pirates*, .420
Dave Trembley, *Baltimore Orioles*, .430
Ron Washington, *Texas Rangers*, .463
Eric Wedge, *Cleveland Indians*, .593
Ned Yost, *Milwaukee Brewers*, .512

SMILE IF YOU BUNT

Steve C. Wang, an associate professor of statistics at Swarthmore College, charted baseball managers from the 2007 season as Chernoff faces, a method of using the heights, widths and angles of facial features to represent different sets of numbers.

MINIMUM VALUES
LEAGUE AVERAGES
MAXIMUM VALUES
Number of different lineups used
Platoon advantage*
Pinch-hitters used
Pinch-runners used
Stolen-base attempts
Sacrifice-bunt attempts
Runners moving with the pitch

*Percentage of players who had the advantage of batting against an opposite-handed pitcher at the start of the game.
Note: Because different rules cause National League managers to use more pinch-hitters, for example, each manager's rates are compared with his league's average.

JONATHAN CORUM/THE NEW YORK TIMES

There are two packages in R that have functions for plotting Chernoff faces: aplpack and TeachingDemos.

Appendix 2 shows how to do boxplots for the example we just did two sample t tests.

7. Testing whether assumptions hold

There are procedures one can perform to test whether the assumptions such as equal variance hold. An example is the Levene test[4], which compares two (or more) variances to test whether the population null hypothesis of equal variances can be rejected.

But these tests make their own assumptions. There is something silly about performing a test to check an assumption in order to perform a test..., so I won't emphasize such statistical tests on assumptions. Further, statistical tests on assumptions behave like any other tests: with enough power any difference, no matter how small, will reject the null hypothesis. This is not a desirable feature when testing assumptions because many statistical tests can tolerate small deviations from assumptions, but if sample size is large even the smallest deviations will appear statistical significant.

SPSS prints out these tests on assumptions (such as the Levene test) automatically, so you may simply ignore them.

R doesn't provide these tests automatically. Instead, R has the ability to apply a different t test that doesn't assume equal variances, which we'll discuss later. You just change the var.equal argument to FALSE.

In Lecture Notes 2 I'll present more techniques for checking assumptions. These techniques will not rely on statistical tests and are relatively easy to perform and understand.

Our friend the boxplot

One useful technique for checking the equality of variance assumption involves the boxplot. To check for equality of variance, you can graph boxplots separately for each group. If the width of the box (i.e., the interquartile range, IQR) for one group differs dramatically from the width for the other group, and/or the length of the whiskers across groups differs dramatically, then that suggests a violation of equality of variance. "Dramatically" will be defined in more detail later. The boxplot can also be used to check for symmetry of the distribution (which is a necessary property of the normal distribution). A symmetric distribution should have it's median close to the middle of the "box" and it's two whiskers (indicating minimum and

[4]The Levene test is based on absolute deviations from the sample medians. There are several versions of the Levene test around–see Kutner et al for a discussion. Another test you might hear about is the Hartley test, which is a bad test for checking equality of population variances because it is very sensitive to departures from normality. The Hartley test amounts to taking a ratio of the two variances and using a standard F test, again see Kutner et al for discussion.

maximum) about equal length.

8. Effects of outliers; a simple example using an Excel spreadsheet and the two-sample *t* test.

What to do when assumptions are violated

9. Three alternative procedures when assumptions are violated

   (a) You can use a test that does not make the particular assumption you're violating.

   For example, Welch's separate variance two-sample *t* test does not make the equality of variances assumption. The Welch test is similar to the classical two sample *t* test but it does not make the assumption that both groups have equal population variances. The computation is similar to the classic test with two exceptions: 1) there is no pooling because variances are not assumed to be equal, and 2) the degrees of freedom are "adjusted" to take into account the discrepancy between the two variances.

Welch's t test

   FYI: here is the formula for Welch's two-sample *t* test. Most statistical packages compute this formula, however it is instructive to look at how the formula works. Similar to the typical t test we have

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1-15}$$

   The degrees of freedom are complicated and are defined as

$$\frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (n_1 - 1)(1 - c)^2} \tag{1-16}$$

   with

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{1-17}$$

   Quite ugly!

   To help us understand these degrees of freedom we'll consider the bounds. The degrees of freedom for Welch's test are greater or equal to the sample size minus one corresponding to the group having the greater variance. For example, if one group has size 5 and the other size 10, then Welch's df will be no smaller than either 5-1=4 or 10-1=9. Whether the lower bound is 4 or 9 in this example is driven by which group has the greater variance (e.g., if the group with size 5 has the greater variance, then the lower bound would be 4; however if the group with size 10 has the greater variance, then the lower bound would be 9). A conservative approximation to the lower bound is simply to

say that the lower bound is the smaller sample size minus one (denoted min(n1,n2) - 1, where min represents the minimum of the two sample sizes). This conservative approximation will be right when the group with the smaller sample size has the greater variance but not when the other group has the greater variance. The upper bound is less than or equal to the usual degrees of freedom (i.e., $df_{Welch} \leq n_1 + n_2 - 2$). The magnitude of change in degrees of freedom could be interpreted as an indication of how severely the equality of variance assumption is violated.

Note that if the two variances are equal, then both the classical two sample *t* test and Welch's separate variance *t* test yield identical values. So, a simple decision rule that many statisticians adopt is to always use Welch's test. The rationale for this decision rule should be obvious: when the population variances are equal one gets the same result as the classical test, but when the population variances differ Welch's test gives the correct result. Also, if the two sample sizes are equal ($n_1 = n_2$), then both the Welch and the observed values of the classical *t*s are equivalent (but note that the degrees of freedom could still differ due to the difference in how degrees of freedom are calculated, yielding different *p*-values for the two tests even though the values of the ts may be the same). The difference in degrees of freedom means the critical t value for the Welch will not be the same as the t critical in the classic test that assumes equal variances.

SPSS performs Welch's separate variance *t* test. But, not all psychologists are familiar with Welch's test and you may get funny looks if you present a test having noninteger degrees of freedom (e.g., $t(12.7) = 2.20$). Some people round down to the nearest integer, but if you do that you should label it clearly as a Welch's *t*-test or a separate variance test in your write-up. I don't mind having noninteger degrees of freedom.

R performs the Welch *t* test with the argument to the t.test() command
var.equal = FALSE.

Here is another heuristic for checking the violation of equality of variances (we've already talked a little about using boxplots for this purpose). If the original degrees of freedom and the Welch degrees of freedom are close to each other, then that suggests that the equal population variance assumption probably holds because the degree of freedom adjustment wasn't drastic. However, if there is a large discrepancy between the two degrees of freedom (i.e., Welch is penalizing in a big way), then that suggests the equal variance assumption may be violated.

(b) You can go to a nonparametric test such as Mann-Whitney U (aka Wilcoxon Rank Sum Test and Kruskal-Wallis)

The Mann-Whitney two sample test is a nonparametric version of the two sample t test. It is identical to performing the usual pooled t test on the ranks of the data (rather

than the raw data). That is, take two independent samples and transform the scores into ranks (the ranks are computed with respect to the both samples together, not ranking one sample separately and then ranking the second sample). Then compute a regular two sample t test on those ranks rather than the original observed scores. This t test on ranks is identical to the Mann-Whitney U test (if there are ties in rank then a special correction is needed for the equivalence to hold).

The ranking reduces the effects of outliers and tends to help with violations of equality of variance. I'll come back to this test and its generalization later when we cover ANOVA. It is remarkable that the classic t-test works well on data that have been transformed to ranks.

(c) Transform the dependent measure

Sometimes the dependent variable can be transformed in other ways besides ranks. On the new scale the assumptions might seem more reasonable. Appendix 1 shows how a simple transformation can improve (by visual inspection) the normality assumption. Appendix 2 shows a simple example of a two-sample *t* test that violates the homogeneity of variance assumption. Appendix 3 discusses some of the measurement concerns surrounding transformations.

Data analysis is a juggling act

Keep in mind that there are two sets of assumptions whenever you measure and perform statistical tests. One set of assumptions deals with the measurement part (see Appendix 3) and the other set of assumptions deals with the statistical tests (e.g., independence, equal variances, normal distributions). The art of data analysis involves juggling these two sets of assumptions, especially when remedial measures needed to deal with one set violate assumptions in the other set.

We will return to the issue of transformations later. At that time, I'll present some techniques that will help you select the best transformation given the particular properties of the data, and I will also spend some more time on exploratory data analysis.

10. What to report in a paper and what to pre-register

Comparing means of two groups is relatively straightforward so I won't go into as much detail here as I will in later lecture notes. Suppose we have an experiment where participants are randomly assigned to two groups: experimental and control. Here is a sample writeup for the results section with completely madeup numbers. "The mean reaction time for the experimental condition was 356ms (sd=40) and the mean for the control condition was 410ms (sd=45). This difference was statistically significant, Welch t (122.4) = 4.3, p=.02. The 95% confidence interval for the difference between the two means was (-64, -44)." Note that I

didn't start with the statistical test. I don't like results section that read like this: "A two sample t test rejected the null hypothesis, t (122.4) = 4.3, p=.02. See Table for means." Put your descriptive statistics front and center; the statistical test and CI merely play supporting roles and punctuate the sentence. The statistical tests are not your primary results, they provide modeling aspects and provide conventions like α levels and confidence intervals for making decisions.

The paper would have a data analysis section and that's where you would state how you test hypotheses with a two sample t-test, why you are using the Welch test, how you checked assumptions, any remedial measures you used (transformation or nonparameteric tests), etc.

I haven't talked about effect sizes yet and some journals require effect sizes. I defer effect size discussion until the next set of lecture notes.

In terms of pre-registration, it can get pretty ugly. For each hypthesis state the test you will run and provide the code you will use. State how you will check assumptions. What will you do if assumptions aren't met? If you use Baysian methods (short introduction below), which prior will you use on each of your parameters, what robustness checks will you use to assess sensitivity of analysis to the choice of prior, what criteria will you use to assess simulation convergence? Some preregistration protocols ask you to outline all your analytical steps including assumption checking and remedial measures. For some cases, like unequal variances in between subjects designs, you can make things easier by pre-registering a Welch test. In that case if you have unequal variances you are safe, if you meet the assumption you are still ok because if the assumption is met, the Welch converges to the classical equal variance t test. But you may not anticipate all the issues and you may have a more complex pattern of assumption violation, such as nonnormal data, outliers, missing data, etc. A Welch won't handle those and some ways of handling those issues may require equal variances. So pre-registration can get rather detailed. A good strategy is to pre-register the code you will use and include detailed comments about what you will look at and actions you will take depending on those results, e.g., "I will conduct a boxplot in R using the cammand presented below to evaluate whether distributions are symmetric, inspect outliers, and evalute the equality of variance assumption. If I see outliers as detected by the boxplot default parameters, I will also conduct a Mann-Whitney U for robustness and will report both the originally planned t test and the Mann-Whitney test in paper."

If you pre-register, then do what you said you would do. It is fine to deviate from the plan, but you should be transparent in the paper and point out which analyses were preregistered and which were not. There is nothing wrong with that; the key is to be transparent. What if a new paper with a new analytic method comes out as you are completing the study so you didn't know about it at the time of preregistration? You can still use the new method. "At the time of preregistration this technique was not available. We report the preregistered analysis as well as the newer analytic approach." There are several recent cases of authors waving the pre-registration flag in their paper, but when you compare the plan with what was reported in

the paper, there are major discrepancies and the authors didn't call them out. At least with pre-registration one can compare the plan with the implementation, but don't assume that just because a paper has a pre-registration badge that all analyses presented were preregistered (apparently not all reviewers and editors verify that the preregistered plan was followed).

11. Bayesian Approach

Bayesian statistics has been rediscovered. It was prominent in the 1940s to 1960s and, while much work was done to develop theoretical understanding, it did not catch on despite several attempts by proponents to convince entire disciplines such as psychology about its merits (e.g., Edwards, Lindman & Savage, 1963, Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242). One reason why it has come back is that computing power makes Bayesian statistics relatively easy to conduct, whereas in the past it required sophisticated mathematics to accomplish such as integrating out parameters and deriving conditional and marginal distributions. Throughout the term I'll present snippets of Bayesian thinking and approaches. If you want a readable introduction to the topic see the Feinberg & Gonzalez (2012) chapter that's in the articles section of Canvas.

For now, I'll just point out special case Bayesian tests of the one sample and two sample t-tests we reviewed here. Bayesian statistics makes the same assumptions as we have already made. The data are assumed independent, groups have equal population variances and the data are normally distributed. In addition, the Bayesian also introduces a prior distribution. This represents the uncertainty on every unknown parameter prior to seeing data . The output of the analysis is a posterior distribution that takes into account not just the data but the prior information, including the uncertainty of all the unknown parameters in the model. The important point about the Bayesian framework is that it provides probability information about the unknown parameters given the data. I'll explain through the one and two sample t-tests.

(a) One sample t-test with unknown variance.

In this case we have two unknown parameters: the mean and the variance. We hope to use data to reduce that uncertainty. In this illustration I'll use a pair of special priors, known as the noninformative priors, that says all values are equally likely (so a uniform distribution on the mean and a uniform distribution on log σ, a detail I don't want to get into now). We assume data are normally distributed and independent. Under this setup it is possible derive the posterior distribution for the unknown mean that incorporates the uncertainty of the unknown variance. The posterior distribution of the mean follows a t distribution with n - 1 degrees of freedom. We know the entire posterior distribution but we can pick out specific pieces, such as the two values corresponding to the middle

95% of the distribution. To compute those values we use

$$\overline{\mathrm{Y}} \quad \pm \quad t_{\alpha/2} \, \frac{\mathrm{s}}{\sqrt{n}} \tag{1-18}$$

where s is the sample st. dev. and $\overline{\mathrm{Y}}$ is the sample mean. Does this look familiar? It is identical to the CI we saw for the one sample t test. The interval will be the same under these assumptions and this choice of prior distributions. The interpretation differs however. The Bayesian interprets this interval as the 95% credible interval such that there is a 95% probability that the true unknown value of the mean is contained in the interval. Recall the frequentist interpretation is one of repeated sampling (95% of such intervals...), so the Bayesian interpretation is more natural and more in line with what researchers are hoping statistical tests provide.

Of course, if we choose different priors, then the Bayesian posterior distribution will not be the same as the frequentist. The prior is quite general, it can incorporate previous information, feasible values, etc. As the prior gets more complicated, then it isn't always possible to solve the Bayesian posterior with formulas and we have to resort to simulation methods in order to estimate the posterior distribution (such as MCMC and Gibbs sampling). The frequentist analysis of "all values are equally likely" prior to observing data is criticized by Bayesians. But the frequentists criticize the Bayesians for having to specify a prior; for them inference should be based on data, not on prior beliefs. I'll come back to this again several times in the year and go into the nuances.

(b) Two sample t test.

In a two sample t test there are 4 unknown parameters: two unknown means and two unknown variances. If we assume equal population variances we can reduce that to 3 unknown parameters. Further, assume independence and normally distributed data, and invoke three uniform priors, one on each of the two means and one on log σ. In this setup the posterior distribution for the difference between two means follows a t distribution with degrees of freedom n_1 + n_2 - 2. One can then probe this distribution, such as finding the middle 95% and the formula turns out to be identical to the classical computation of the 95% confidence interval we saw before, but the interpretation is in terms of the credible interval.

One can also derive a Bayesian test with unequal population variances. It also follows a t distribution and the degrees of freedom are almost identical to the Welch test. For derivations of the Bayesian t tests I presented in these lecture notes see Box and Tiao's 1973 book *Bayesian Inference in Statistical Analysis* and for a modern text book see Gelman et al, Bayesian Data Analysis. You'll need a good understanding probability theory and calculus.

I purposely selected special case prior distributions so that the Bayesian analysis would yield

the same result of as the classic frequentist CIs we saw earlier. Of course, the Bayesian is free to choose priors and every choice of prior would lead to different results than the frequentist approach. The Bayesian approach can also be extended quite easily. For example, Kruschke (2013) proposed a Bayesian two sample t test that has an additional parameter to handle outliers (so data that aren't normally distributed). Some extensions are not easy to derive, but usually they can be estimated easily with simulation methods. So there isn't really *one* Bayesian t-test because there are different priors and various extensions but a Bayesian framework. When working with Bayesian statistics it is important to do sensitivity analysis to verify that the conclusions are not unique to the particular choice of prior distributions, or at least give a sense for how much a prior has to change before the conclusions change.

12. Bootstrapping

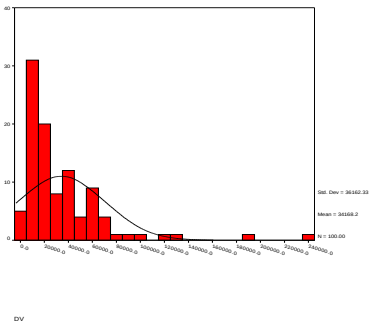
There is another method for computing standard errors that we will use periodically this year. It uses the same logic as the central limit theorem but instead of relying on an asymptotic theorem to give a formula for the standard error, the bootstrap method uses a similulation to compute the standard error. The most common form of the bootstrap for a mean looks a lot like the logic I presented justifying the central limit theorem: take a sample and compute the mean, take a second sample and compute a second mean, etc. Store all those means. You now have an estimate of the standard deviation of the means or other terms likes the 95% CI by taking the value of the sampled means corresponding to the 2.5% and the 97.5% levels. However, a key difference is that rather than repeatedly sample from the population the bootstrap treats the sample like the "population" and repeated samples from that with replacement to create bootstrapped samples. Unlike the central limit theorem, the bootstrap is not limited to means. You can compute anything you want for each sample (such as a median, a correlation, a regression slope, a latent variable, etc) and get a standard error or a CI for that computed value. There are other forms of the bootstrap that instead of sampling raw data they sample the residuals from a fitted model (we first have to learn about models and residuals before covering that method), and other bells and whistles like bias-corrected bootstrap (which doesn't always remove the bias so don't be fooled by the name). Appendix 2 illustrates how to run bootstrapping in SPSS and R using the same example we've been using throughout these lecture notes.

2 Appendix 1

To show the effects of a transformation on the assumption of normality, I generated some random data. The histogram below shows that the data are not normally distributed (they are skewed). I know that such a distribution can be "transformed" into something that looks more normal by performing a log transformation. The second histogram shows the "improvement" on the log scale.

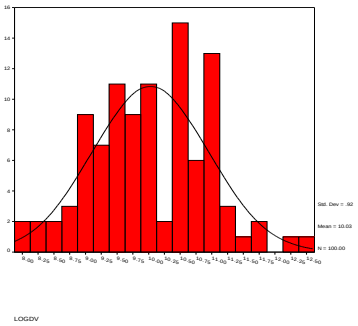
```
 data list free / dv.
begin data
[DATA GO HERE]
end data.

GRAPH
  /HISTOGRAM(NORMAL)=dv .
```

The SPSS syntax above calls for a histogram with a normal curve superimposed on the histogram. The normal curve does not provide a reasonable approximation to the observed histogram.



Now, I'll transform those same data using a log transformation. Notice that the histogram appears more symmetric and bell-shaped after the transformation; the bell-shaped curve also represents the observed histogram better.

```
 compute logdv = ln(dv).
execute.

GRAPH
  /HISTOGRAM(NORMAL)=logdv .
```



The logic of a transformation to improve symmetry is that one is trying to affect scores differentially at the two ends of the distribution. For example, if there is a long tail on the right side, then a transformation that exerts more influence on those numbers may do the trick to transform the data into a more symmetric distribution. For example, if there are many scores between 1 and 5 but a few scores between 25 and 100, then a sqrt transformation may help with symmetry because the sqrt exerts relatively little effect on the low numbers but relatively more effect on the high numbers (i.e., 25 goes into 5, 100 goes into 10).

I'll redo this example below using R (go to page 1-39).

3 Appendix 2

Example of a Two-Sample T-Test With Violation of Homogeneity of Variance
First SPSS, then R

Data:

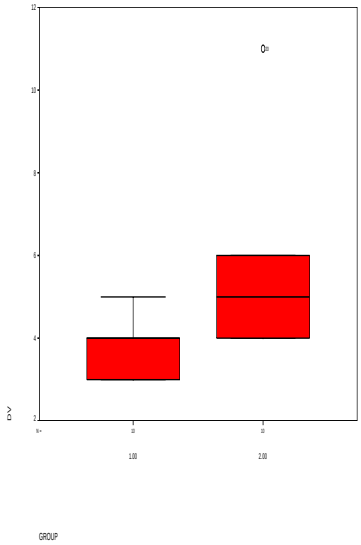
Group 1	Group 2
3	4
4	5
5	6
4	5
3	4
4	5
5	6
4	5
3	4
4	11

Note that Group 2 has one outlier; otherwise, the scores in Group 2 are equal to Group 1 scores plus one.

SPSS **Example using SPSS**

```
data list free /group dv.
begin data.
1 3
1 4
1 5
1 4
1 3
1 4
1 5
1 4
1 3
1 4
2 4
2 5
2 6
2 5
2 4
2 5
2 6
2 5
2 4
2 11
end data.

examine variables=dv by group
 /plot = boxplot .
     DV
By   GROUP          1.00

Valid cases:        10.0   Missing cases:        .0   Percent missing:      .0


Mean         3.9000  Std Err      .2333  Min          3.0000  Skewness     .1660
Median       4.0000  Variance     .5444  Max          5.0000  S E Skew     .6870
5% Trim      3.8889  Std Dev      .7379  Range        2.0000  Kurtosis    -.7336
```

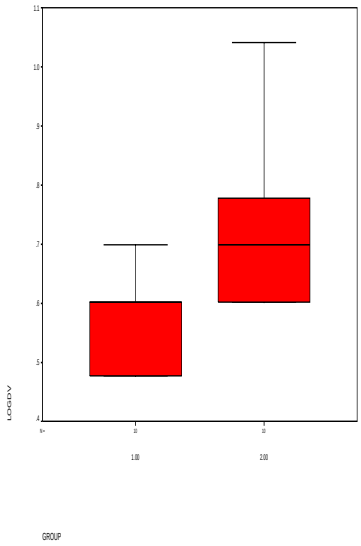
```
                                        IQR         1.2500  S E Kurt    1.3342

    DV
By  GROUP          2.00

Valid cases:         10.0   Missing cases:         .0   Percent missing:      .0


Mean         5.5000  Std Err      .6540  Min         4.0000  Skewness    2.4489
Median       5.0000  Variance    4.2778  Max        11.0000  S E Skew     .6870
5% Trim      5.2778  Std Dev     2.0683  Range       7.0000  Kurtosis    6.7601
                                         IQR         2.0000  S E Kurt    1.334
```



compute logdv = lg10(dv).
execute.

examine variables=logdv by group
/plot = boxplot .

```
    LOGDV
By  GROUP          1.00
Valid cases:         10.0   Missing cases:         .0   Percent missing:      .0

Mean          .5840  Std Err      .0263  Min          .4771  Skewness    -.1413
Median        .6021  Variance     .0069  Max          .6990  S E Skew     .6870
5% Trim       .5835  Std Dev      .0832  Range        .2218  Kurtosis    -.9666


    LOGDV
By  GROUP          2.00
Valid cases:         10.0   Missing cases:         .0   Percent missing:      .0

Mean          .7200  Std Err      .0413  Min          .6021  Skewness    1.7851
Median        .6990  Variance     .0171  Max         1.0414  S E Skew     .6870
5% Trim       .7087  Std Dev      .1306  Range        .4393  Kurtosis    4.1153
                                         IQR          .1761  S E Kurt    1.3342
```



t-test groups=group(1,2)
/variables=dv .

t-tests for Independent Samples of GROUP

Variable	Number of Cases	Mean	SD	SE of Mean
DV				
GROUP 1	10	3.9000	.738	.233
GROUP 2	10	5.5000	2.068	.654

Mean Difference = -1.6000

Levene's Test for Equality of Variances: F= 2.204 P= .155

	t-test for Equality of Means				95%
Variances	t-value	df	2-Tail Sig	SE of Diff	CI for Diff
Equal	-2.30	18	.033	.694	(-3.059, -.141)
Unequal	-2.30	11.25	.041	.694	(-3.124, -.076)

TRY A T-TEST ON THE TRANSFORMED SCORE
t-test groups=group(1,2)
/variables=logdv .

t-tests for Independent Samples of GROUP

Variable	Number of Cases	Mean	SD	SE of Mean

```
-----------------------------------------------------------------------
LOGDV

GROUP 1                      10          .5840        .083         .026
GROUP 2                      10          .7200        .131         .041
-----------------------------------------------------------------------

         Mean Difference = -.1360

         Levene's Test for Equality of Variances: F= .504   P= .487


      t-test for Equality of Means                                   95%
Variances   t-value       df    2-Tail Sig     SE of Diff          CI for Diff
-------------------------------------------------------------------------------
Equal         -2.78       18          .012           .049        (-.239, -.033)
Unequal       -2.78    15.27          .014           .049        (-.240, -.032)
-------------------------------------------------------------------------------
```

Here is the Mann-Whitney U test:

```
npar tests m-w = dv by group(1,2).
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

    DV
 by GROUP


    Mean Rank    Cases

         7.25       10  GROUP = 1.00
        13.75       10  GROUP = 2.00
                    --
                    20  Total

                              Exact             Corrected for ties
         U            W      2-Tailed P           Z      2-Tailed P
        17.5         72.5       .0115         -2.5800       .0099
```

SUMMARY:

We tested the hypothesis that the population means are equal, that is, $\mu_1 = \mu_2$. But, we observed that the variances were not equal (due to a single outlier). We tried the classic *t* test, the separate variance (Welch's) *t* test, a classic *t* test on the transformed variable, and a nonparametric test.

Normally, one does not do all these tests. I just show them together so you can compare them. You should begin thinking about how you want to handle violations of the assumptions. This is a choice based purely on aesthetics. Each method has its advantages and disadvantages.

One thing we did not do is omit the outlier. Rarely is throwing out a bad data point a good thing.

Bayesian analysis in SPSS is very limited. Version 25 introduced a few basic Bayesian analyses such as t tests, one way ANOVA, and simple regression. For now, better to use other Bayesian programs such as ones implemented in R or standalone ones like BUGS or STAN, which are mature and allow complete flexibility. But here goes for Bayesian methods in SPSS. The syntax for the two sample t test in Bayesian form:

```
BAYES INDEPENDENT
  /INFERENCE DISTRIBUTION=NORMAL VARIABLES=dv ANALYSIS=BOTH GROUP=group SELECT=LEVEL(1 2)
  /PRIOR EQUALDATAVAR=FALSE VARDIST=DIFFUSE
  /ESTBF COMPUTATION=ROUDER.
```

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
dv	= 1.00	10	3.9000	.73786	.23333
	= 2.00	10	5.5000	2.06828	.65405

Bayes Factor Independent Sample Test (Method = Rouder)[a]

	Mean Difference	Pooled Std. Error Difference	Bayes Factor[b]	t	df	Sig.(2-tailed)
dv	1.6000	.69442	.466	2.304	18	.033

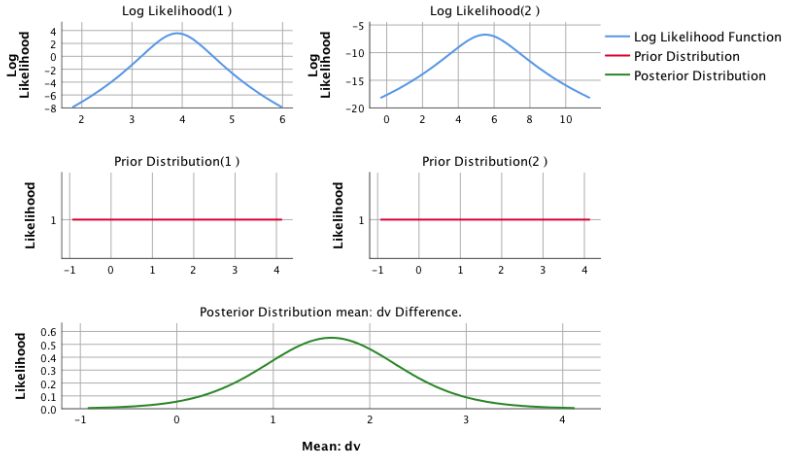
a. Assumes unequal variance between groups.

b. Bayes factor: Null versus alternative hypothesis.

Posterior Distribution Characterization for Independent Sample Mean[a]

	Posterior			95% Credible Interval	
	Mode	Mean	Variance	Lower Bound	Upper Bound
dv	1.6000	1.6000	.620	.0312	3.1688

a. Prior for Variance: Diffuse. Prior for Mean: Diffuse.



Finally, here is how to do bootstrapping in SPSS. We first run the bootstrap command to inform SPSS that subsequent command will be use bootstrap; here I use the default SPSS bootstrap samples of 1000.

```
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=group INPUT=dv
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000

T-TEST GROUPS=group(1 2)
  /VARIABLES=dv
  /CRITERIA=CI(.95).
```

Group Statistics

	group		Statistic	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] 95% Confidence Interval Lower	Bootstrap[a] 95% Confidence Interval Upper
dv	1.00	N	10				
		Mean	3.9000	.0134	.2319	3.5000	4.3636
		Std. Deviation	.73786	-.05266	.13596	.40825	.92796
		Std. Error Mean	.23333				
	2.00	N	10				
		Mean	5.5000	-.0030	.6231	4.5556	6.9167
		Std. Deviation	2.06828	-.29466	.85073	.50011	3.09225
		Std. Error Mean	.65405				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

		Levene's Test for Equality of Variances: F	Levene's Test for Equality of Variances: Sig.	t-test for Equality of Means: t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper
dv	Equal variances assumed	2.204	.155	-2.304	18	.033	-1.60000	.69442	-3.05893	-.14107
	Equal variances not assumed			-2.304	11.254	.041	-1.60000	.69442	-3.12421	-.07579

Bootstrap for Independent Samples Test

		Mean Difference	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] Sig. (2-tailed)	Bootstrap[a] 95% Confidence Interval Lower	Bootstrap[a] 95% Confidence Interval Upper
dv	Equal variances assumed	-1.60000	.01644	.65649	.116	-3.08619	-.52567
	Equal variances not assumed	-1.60000	.01644	.65649	.114	-3.08619	-.52567

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The output for the boostrapped difference between the two means is a separate table after the usual independent samples test, and also the table of descriptive statistics for each group also contains information from the bootstrap.

Repeat example with R

SWITCHING TO R: Data file is saved as a two column text file called "example.dat". I assume first row of example.dat has the column names and subsequent rows have the data, which look just like the two columns that are between begin data/end data in the SPSS example above. In this case the argument header=T is used. If the file example.dat doesn't column names, then use header=F and add the column names later with the names() command.

I like to use data.frames so after reading in the data file I convert it to a data.frame.

You'll need to edit the "PATH/TO/FILE/" part to where you saved the example.dat file.

```
# setwd('PATH/TO/FILE/example.dat)
data <- read.table("example.dat", header = T)
data <- data.frame(data)
```

For completeness, here are the contents of the file example.dat.

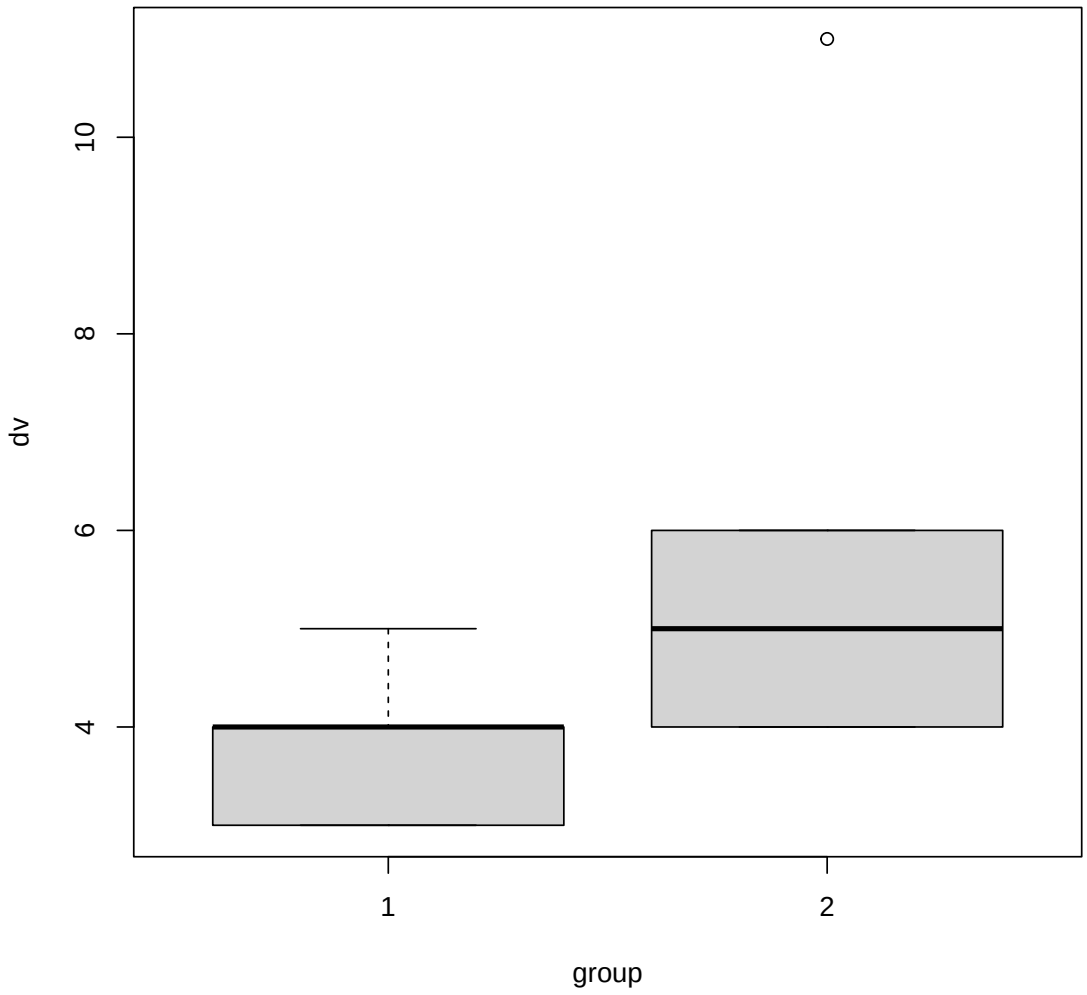
```
group dv
1 3
1 4
1 5
1 4
1 3
1 4
1 5
1 4
1 3
1 4
2 4
2 5
2 6
2 5
2 4
2 5
2 6
2 5
2 4
2 11
```

Next we make the column called group a factor so that commands know to treat it as group 1 and group 2 rather than the numbers 1 and 2. Here I specify the column called group and convert it to a factor. Also, add a column of log transformed dv to the data.frame data (use base 10 to be consistent with the SPSS output). Now the data.frame data will have three columns.

```
data$group <- factor(data$group)
data$logdv <- log10(data$dv)
```

Boxplot for this example

```
boxplot(dv ~ group, data = data)
```



T test on both raw and log data. I use the data argument and assign it the value data; this allows R to find the variables called group and dv because they "reside" in the data.frame data.

```
t.test(dv ~ group, data = data, var.equal = T)


##
##  Two Sample t-test
##
## data:  dv by group
## t = -2.3041, df = 18, p-value =
## 0.03335
## alternative hypothesis: true difference in means between group 1 and group 2
```

```
## 95 percent confidence interval:
##  -3.058927 -0.141073
## sample estimates:
## mean in group 1 mean in group 2
##             3.9             5.5
```

```
t.test(logdv ~ group, data = data, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  logdv by group
## t = -2.7771, df = 18, p-value =
## 0.01243
## alternative hypothesis: true difference in means between group 1 and group 2
## 95 percent confidence interval:
##  -0.23891295 -0.03311734
## sample estimates:
## mean in group 1 mean in group 2
##       0.5839604       0.7199755
```

Perform Mann-Whitney test, which is called wilcox.test in R.

```
wilcox.test(dv ~ group, data = data, correct = T)
```

```
##
##  Wilcoxon rank sum test with
##  continuity correction
##
## data:  dv by group
## W = 17.5, p-value = 0.01108
## alternative hypothesis: true location shift is not equal to 0
```

Bayesian Analysis: To anticipate later lectures, I'll present a snippet of a Bayesian analysis. Here we can use it to verify the claim that if the priors are specified in a particular way the Bayesian analysis 95% interval mimics the 95% confidence interval from the classical test.

```
library(brms)
# so results are same sign as t test, assign
# group 2 as reference group
data$group.relevel <- relevel(data$group, ref = "2")
```

```
# set prior close to the specification above;
# probably better to stick with default prior for
# noninformative prior
bayes.prior <- prior(normal(0, 1e+06), class = b) +
    prior(normal(0, 1e+06), class = sigma) + prior(normal(0,
    1e+06), class = Intercept)
out.bayes <- brm(dv ~ group.relevel, data = data, prior = bayes.prior,
    iter = 20000, thin = 5)
summary(out.bayes)
plot(out.bayes)

# can also do a plot with 95% shaded; see also
# the tidybayes package
library(bayesplot)
mcmc_areas(as.matrix(out.bayes), regex_pars = "group.relevel",
    prob = 0.95)
```

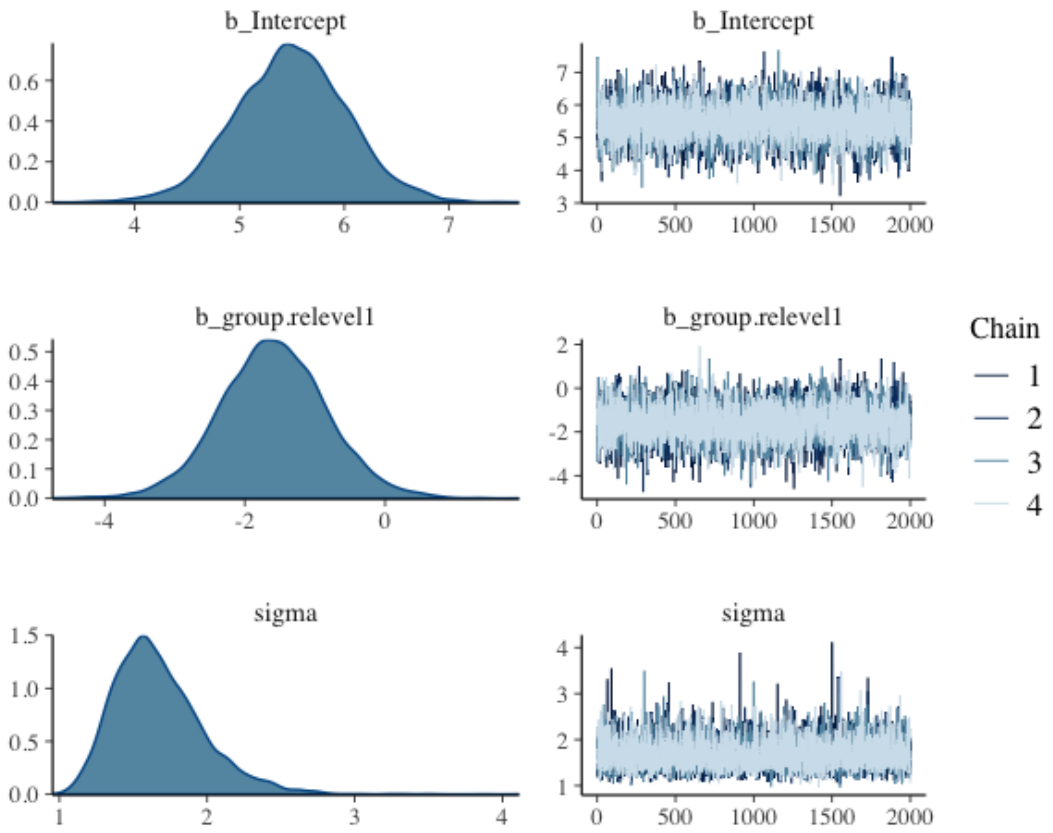
The snippets of the output include

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: dv ~ group.relevel
   Data: data (Number of observations: 20)
Samples: 4 chains, each with iter = 20000; warmup = 10000; thin = 5;
         total post-warmup samples = 8000

Population-Level Effects:
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept         5.49      0.54     4.45     6.55       7802 1.00
group.relevel1   -1.59      0.76    -3.11    -0.12       7801 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     1.67      0.30     1.20     2.38       8000 1.00
```

The estimate of the difference between the two means is -1.59 and the CI is [-3.11, -0.12], very close to the CI from the t test command under equal variances (small discrepancy between the t test and Bayes is in part due to sampling and the prior settings allowed by the brm program). I set the iterations of the simulation at 20,000. The first 10,000 are dropped to account for the simulator "settling in" and then thin means keep every fifth value of the remaining 10,000 sample draws so 2000 samples remain. Chain = 4 means this was done 4 times so a total of 8000 samples in the entire simulation.

Here is a plot of the Bayesian analysis, which yields a distribution for the difference between the two groups (middle row), the distribution of the residual standard deviation (bottom row) and the

Figure 1-8: Example Bayesian analysis plots

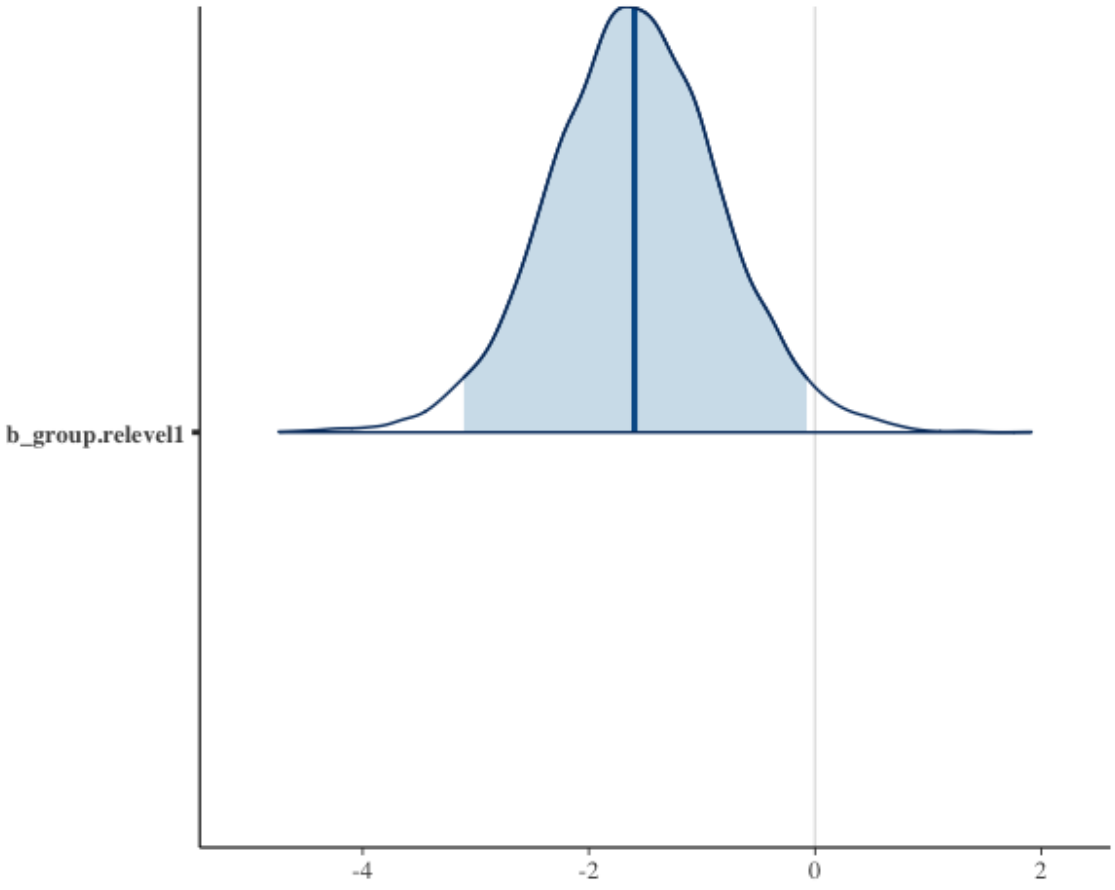


distribution of the mean of one group 2 (top row), which is nicely normal due to the central limit theorem in action. The second column are traceplots, which I'll cover later in the term.

There are many ways of accomplishing the boostrap in R. The most common approach is to use the general boot package in R. Other approaches include specific functions people have written and contributed to various packages such as the boot.t.est in the nonpar package. I prefer the boot package because it is general and can be used across many types of models.

```
library(boot)

# function to compute mean difference;
# illustrated using the t.test function to do the
# heavy lifting
mean.diff <- function(formula, data, indices) {
    boot.sample <- data[indices, ]
    means <- t.test(formula, data = boot.sample)$estimate
    return(means[1] - means[2])
```

Figure 1-9: Density plot of difference between means with shaded 95% interval



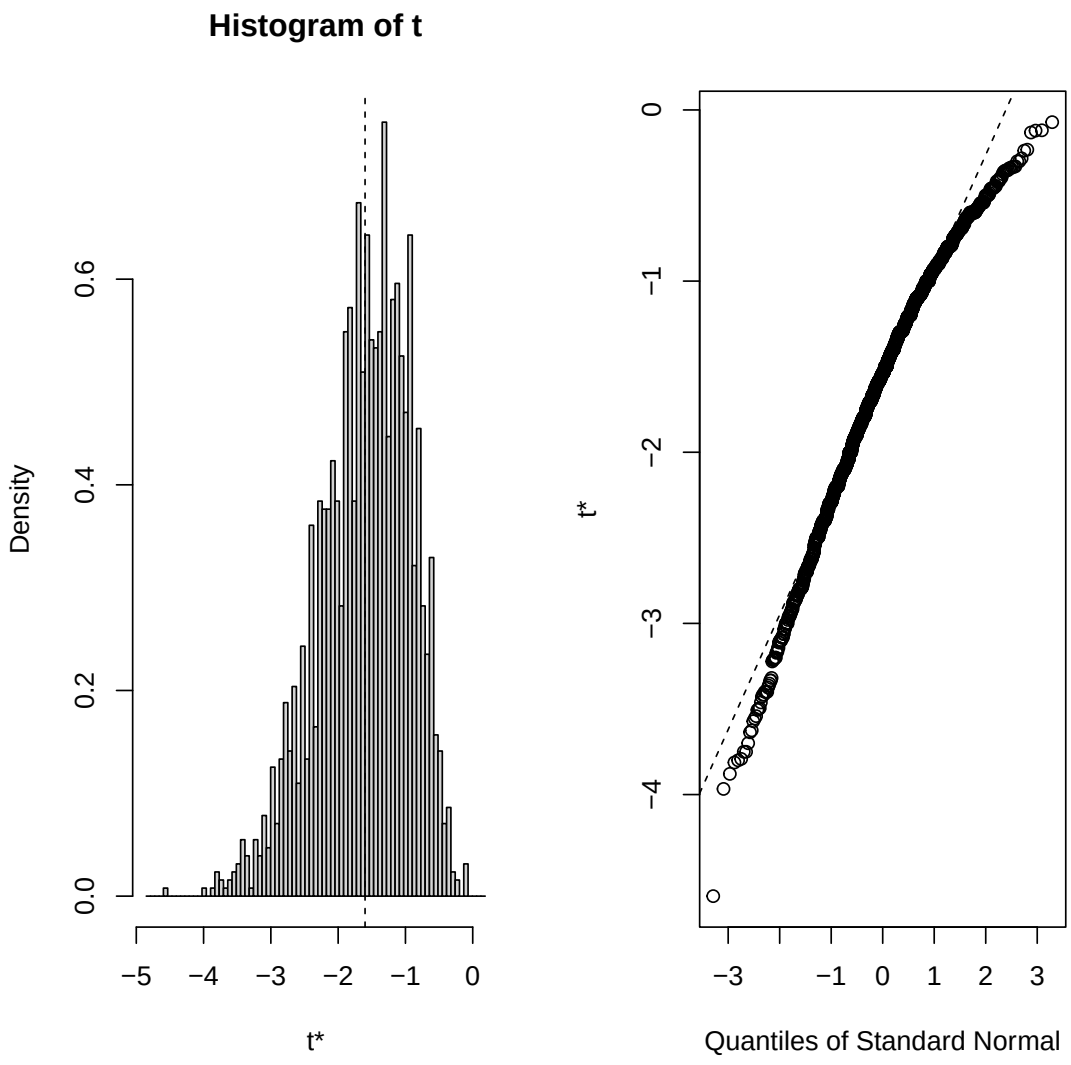
```
}

boot.results <- boot(data = data, statistic = mean.diff,
    R = 2000, formula = dv ~ group)
boot.ci(boot.results, type = c("norm", "basic", "perc",
    "bca"))


## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.results, type = c("norm", "basic", "perc",
##     "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   (-2.910, -0.278 )   (-2.658, -0.117 )
##
## Level     Percentile            BCa
## 95%   (-3.083, -0.542 )   (-3.717, -0.700 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable


# studentized needs more info about variance so I
# did not list type='all'
```

Note the slightly skewed distribution of the difference between the two means most likely due to the outlier and relatively small sample. This would slightly throw off inferences of the classic test. The normal probability plot (introduced later in the course) illustrates the long tails typical of the t distribution and also a slight asymmetry related to the skewness.

The boot.ci function provides several types of confidence intervals and these are heavily debated in the field. The normal CI is similar to the regular CI in computation but uses the standard error computed from the bootstrap and a correction for bias, the percentile CI simply reports the 2.5% and 97.5% values of the bootstrap samples, the basic CI is like the percentile with a type of bias correction, and BCa is a different bias-correction imposed on the percentile method. There are many more types of CIs used in the bootstrapping literature. See, e.g., Efron and Tibshirani (1993), *An Introduction to the Bootstrap*; Davison and Hinkley (1997), *Bootstrap Methods and Their Application*, Chapter 5; various websites like www.datacamp.com/community/tutorials/bootstrap-r.

Figure 1-10: Distribution of Mean Differences and QQ plot

```
plot(boot.results)
```



4 Appendix 3: Measurement Issues v. Statistical Issues.

I mentioned the advantages of transforming data to meet statistical assumptions. However, statistical assumptions are not the only assumptions made when dealing with data. In addition, there are measurement concerns. In data analysis one must frequently juggle measurement concerns and statistical assumptions. A failure to appreciate this point has led to several misunderstandings in the literature. We now turn to a brief and elementary discussion of measurement issues.

Measurement deals with the assignment of numbers to objects. Statistics deals with what you do after you've collected the numbers and the error around those numbers. Obviously, one would want to assign numbers in some principled way (this is the fundamental problem of measurement). There are three primary issues in measurement:

1. under what conditions can one assign numbers to objects?

2. how unique is the scale? are there different numerical assignments that are just as good?

3. is the scale meaningful? once the scale is defined, then what are the numbers really telling us? what operations are permissible on the numbers?

For our purposes, we will only consider uniqueness and meaningfulness (2 and 3); the existence of numerical scales (1) is beyond our scope and a full treatment requires a strong background in abstract algebra. Uniqueness refers to permissible transformations that can be made on a scale that still preserve it's inherent properties. Examples:

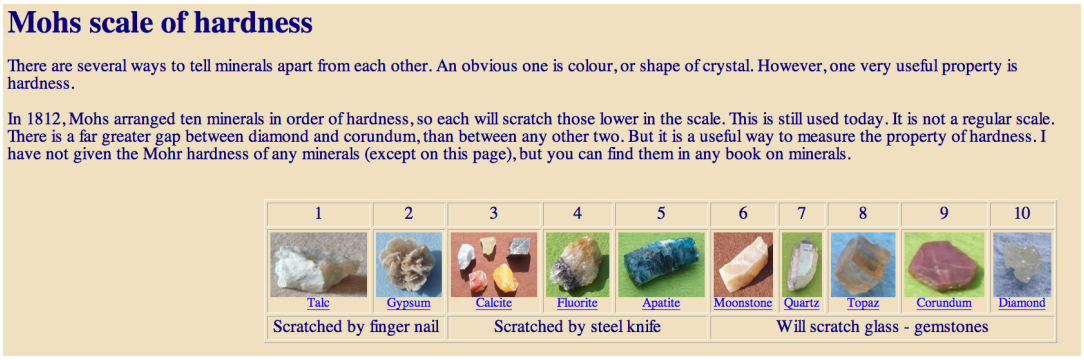
1. A ratio scale preserves ratios of scale values. Thus, the only permissible transformation is multiplication by a positive number (why?). Example: length, weight, and time.

2. An interval scale has the property of "equal intervals" throughout the scale. A linear transformation will preserve this property. Example: temperature.

3. An ordinal scale only has the property that it preserves order. Thus, any transformation f that preserves order (i.e., is monotonic in the sense that if $x > y$, then $f(x) > f(y)$) is permissible. Examples: Moh's hardness scale, academic grades (A, B, C, etc), Beaufort Wind Scale (0=calm, 1=light air, 2=light breeze, ..., 12=hurricane). See Figure 1-11.

4. A nominal scale uses numbers as labels. The word nominal means name. Thus, any one-to-one mapping is permissible.

Figure 1-11: Mohs Hardness Scale



Meaningfulness has been a rather tricky concept to define mathematically. Intuitively, meaningfulness deals with the following: once you have a scale and know its uniqueness properties, you can start using the scale (i.e., the numbers) to make statements about the original objects. That is, of course, the purpose in constructing a scale. Meaningfulness concerns the kinds of statements that can be made about the objects given the numbers. For example, it makes sense to say that Joe is twice as heavy as John if we measure their weights on a balance and see that Joe's numerical weight is twice the weight of John. But, it doesn't make sense to say that Florida is twice as hot as Seattle if we observe that the temperature in Florida is $80°$F and the temperature in Seattle is $40°$F (this example and the ones below were adapted from Roberts, 1979).

In general,

1. Ordinal scales can be used to make comparisons of order, like $f(x) > f(y)$

2. Interval scales can be used to make comparisons of difference, like

$$f(a) - f(b) > f(x) - f(y)$$

3. Ratio scales can be used to make comparisons such as

$$f(a) = 2f(b)$$

Consider the process of taking the mean of a bunch of scale values (a common practice in psychology). Taking a mean is simply a transformation of the scale values. For example, for four objects where f is the function that assigns a number to an object

$$\text{mean}[f(a), f(b), f(c), f(d)] = \tfrac{1}{n}[f(a) + f(b) + f(c) + f(d)]$$

What complications does having a "transformation on a scale" produce? Consider the dependent variable of weight. You have two groups of animals, one group is given some treatment and the other group is the control. You want to see whether the average weight of the animals in the treatment group (group a) is greater than the average weight of the animals in the control group (group b). In symbols, you want to know whether

$$\frac{1}{n_a}\sum f(a_i) \quad > \quad \frac{1}{n_b}\sum f(b_i) \tag{1-19}$$

Note that every animal is assumed to have the same scale in the sense that weight is measured in the same units.

What other scales different from f would also be meaningful? Ratio and interval scales are okay because the transformation permitted under meaningfulness (multiplication by a constant and linear transformation, respectively) will preserve the order of the means. However, taking means of an ordinal scale is meaningless because the ordering of the means may not be preserved when permissible transformations are performed. For example, taking square roots of all the numbers may alter the ordering of the means. So, even though a transformation may be called for to satisfy statistical assumptions (say to deal with the violation of equality of variance), an inappropriate transformation may violate the property of meaningfulness.

Another example. Take a group of subjects. Each subject watches two different movies and rates each movie on its level of aggressiveness using a seven point scale. The researcher wants to see whether movie a is rated as more aggressive than movie b. Can we take means? The knee-jerk answer is to say "yes". Let's look at this problem more closely. We have

$$\frac{1}{n}\sum f(a) \quad > \quad \frac{1}{n}\sum f(b) \tag{1-20}$$

which represents the statement "the average aggressiveness score for movie a is greater than the average aggressiveness score for movie b." Here the average is defined over subjects. The difference between Equations 1-19 and 1-20, in the latter there are no subscripts on a and b because they are always the same two movies.

It is more plausible to think that each subject has his or her own scale f_i leading to

$$\frac{1}{n}\sum f_i(a) \quad > \quad \frac{1}{n}\sum f_i(b) \tag{1-21}$$

But, it is this more plausible case that kills us as far as meaningfulness goes. If we take averages on the raw scores, not even a ratio scale is meaningful because each subject has his or her own scale, denoted by α_i. That is, even if we make the strong assumption that aggressiveness ratings are ratio scale (implausible as it may be), the ordering of the means will not necessarily be preserved. Thus, if each subject is allowed an arbitrary positive constant α_i (because we assume each f_i is a ratio scale) we have

$$\frac{1}{n}\sum \alpha_i f_i(a) \quad \text{and} \quad \frac{1}{n}\sum \alpha_i f_i(b) \tag{1-22}$$

The two terms in Equation 1-22 need not be ordered the same way as in Equation 1-20 due to the different α_is.

One way to guarantee that the ordering of the means observed in Equation 1-20 remains the same regardless of the arbitrary positive α's is to perform the log transformation before computing the mean. Recall that log(xy) = log(x) + log(y).

$$\frac{1}{n}\sum \log[f_i(a)] > \frac{1}{n}\sum \log[f_i(b)] \quad (1\text{-}23)$$

$$\frac{1}{n}\sum \log[\alpha_i f(a)] > \frac{1}{n}\sum \log[\alpha_i f(b)] \quad (1\text{-}24)$$

$$\frac{1}{n}\sum \log \alpha_i + \frac{1}{n}\sum \log f(a) > \frac{1}{n}\sum \log \alpha_i + \frac{1}{n}\sum \log f(b), \quad (1\text{-}25)$$

the terms containing the α_i's cancel out giving Equation 1-25 the same ordering as Equation 1-20. So, taking means of logs of different ratio scales is a meaningful operation[5]. I suggest you construct an example with made up numbers that produces such a reversal on the raw data but not on the log scale.

The reason I present this example is to illustrate that sometimes a transformation is justified because of measurement concerns (in the previous example the log played a useful role in making the mean a "meaningful" operation). So there are cases in which a transformation is legitimate on measurement grounds. One feature that makes a transformation legitimate is that it can force parameters from "permissible transformations" to cancel out.

A deep understanding of measurement issues requires much background knowledge in mathematics (in particular, abstract algebra and topology). I encourage you to learn more about measurement issues. But this is a course on statistics rather than measurement, so we won't spend much time on these issues.

Now we are ready to evaluate (and make some sense of) the following statement: "A *t* test doesn't know where the numbers came from, so any transformation is legitimate."

Yes, this is true. But, the statement only deals with the statistical assumptions of the *t* test (independence, equality of population variances, and normality). After all, we can even compute a *t* test on ranks and get a legitimate result with respect to *p*-values (i.e., the Mann-Whitney U test, which many consider to be a legitimate test). What the above statement doesn't consider is the notion of meaningfulness. If we want our descriptive statistics (i.e., the transformations we perform on our scales) to say something meaningful about the objects in question, then the scale type becomes very important.

A sermon

When analyzing data we usually want to say something about both the objects in question (that is why we collected data in the first place) and the statistical properties (that's why we are computing a

[5]Once you compute the mean on the log scale it is a good idea to transform the means back to the "original" scale; in this example, take the exponential of each mean. This is mainly for clarity when presenting your data to others. This transformed mean, i.e., mean on logs and then taking the exponential, is equivalent to what is known as the "geometric mean".

p value–in the classical sense, to compute the chances that the observed data could have come from the conditions specified in the null hypothesis). In other words, statistics helps us make sense of the numbers in a study, but statistics only operates on the numbers. To figure out what those numbers mean, i.e., how they relate to the objects in question, we need measurement theory. Statistics deals with making inferences from data; measurement theory deals with the connection between data and reality. Clearly, a good researcher is skillful in both statistics and measurement.

Let's return to the question of which test to use when the assumptions of the two sample *t* test are violated. Unfortunately, there is no best recommendation that works "across the board." The choice of test depends on the particular situation. The route we will take in this course will be to give more weight to the statistical assumptions and transform, or, to use Tukey's more neutral term, "re-express," the data. If we use transformations appropriately, we can get much mileage from the classical statistical techniques. Of course, as we go along I will point out situations where remedial procedures other than transformations may be more appropriate. The reason for emphasizing transformations is that they are quite general and can be used in just about any situation. As our models become more complicated we will not have options like tests that don't make a specific assumption (such as the Welch test) or nonparametric tests (like the Mann-Whitney).