Comments for MEDB 5501, Week 15

Topics to be covered (1/2)

- What you will learn
  - Running SPSS
  - Data types
  - Descriptive statistics
  - Confidence intervals and hypothesis tests
  - Visualization
  - Assessing normality
  - Independent samples and paired samples t-tess

Topics to be covered (2/2)

- What you will learn
  - One factor ANOVA
  - REDCap and one sample confidence intervals
  - Chi-square tests
  - McNemar test
  - Nonparametric tests
  - Correlations
  - Linear regression

You learned how to run SPSS

- Advantages
  - Easy to learn
  - Covers all of the basics
- Disadvantages
  - Expensive
  - Legacy limitations
  - Difficult to customize output
  - Fewer advanced methods

Speaker notes

In the first week of class, you learned SPSS. It is a very popular program among researchers, especially those who do not have the time to learn a more complex system. Because almost all of your data management and analytical choices in SPSS come from the menus, there is less to memorize. SPSS also does a very good job at providing all of the basic approaches for data analysis.

It is an expensive program, and there are certain features that are held over from the days of the mainframe computer. SPSS offers few ways that you can customize your output. It is often difficult to suppresss statistics that you don’t really want and SPSS will also often fail to round their output to a reasonable number of significant digits.

While SPSS has some advanced methods, it does tend to lag a few years behind other software in incorporating new approaches.

Why not R or Python?

- Advantages
  - Easy to customize output
  - Fastest to incorporate new methods
  - No start-up cost (but TANSTAAFL)
  - Best tools for reproducible research
- Disadvantages
  - Inconsistent syntax
  - Must learn coding

Speaker notes

I am a big fan of R and am in the process of learning Python. Both systems have been very good at quickly incorporating new methods. You can control the output at a very fine level of detail. You can easily choose what to include or exclude and can change the number of digits to round to.

There is no start-up cost for either system. You can try out the programs without any initial investment. Keep in mind, however, the acronym, TANSTAAFL. This was first popularized by a science fiction writer, Robert Heinlein. It standds for “There Ain’t No Such Thing As A Free Lunch”. While you may not need any initial investment in Python or R, you will have to expend time and energy as these systems have a very steep learning curve. There’s a cost associated with the steep learning curve and it may outweigh the “free” initial cost.

Both R and Python have built in tools that make it easy to make your research reproducible. This is largely through the use of literate programming and integration with version control software.

The syntax is wildly inconsistent, unfortunately. There’s a joke about how a camel is a horse designed by a committee. Well both R and Python were created by a committee effort and new enhancements are added in a scattershot approach. This leads to a lot of inconsistency in how you code your program.

Corporate programs like SPSS, Stata, and SAS have strong internal controls. While this does not always guarantee perfect consistency, these companies do a lot better than R and Python.

You have to develop at least a limited amount of programming skills to effectively use R or Python. There is no comprehensive menu system that you can rely on.

Why not SAS/Stata?

- Advantages
  - Fast to incorporate new methods
  - Not as much programming as Python/R
  - Very strong customer support
- Disadvantages
  - Expensive start-up costs
  - Rigid output

Speaker notes

Both SAS and Stata are very good at incorporating new methods. Maybe not quite as fast as R and Python, but close. Both systems require a bit of programming. More than SPSS, but not as much as R and Python. Stata does have a pretty good menu system. It is not quite as user-friendly as SPSS but it comes close.

Both programs, but especially SAS, have very strong customer support. This at least partially offsets the initial start-up costs.

Like SPSS, the output from SAS and Stata is a bit rigid and difficult to modify.

Break #1

- What you have learned
  - Running SPSS
- What’s coming next
  - Data types

Data types

- Nominal
  - Summarize using proportions/percentages
  - Show both numerator and denominator?
- Ordinal
  - Summarize using median/interquartile range
- Interval/Ratio
  - Summarize using mean/standard deviation
  - SPSS calls this “scale”
- Controversy: can you add/average ordinal data?

Speaker notes

Nominal data has no natural ordering, and is usually summarized as a percentage/proportion. In general, I think it is a good idea to show the numerator and denominator along with the percentage. Both 2/8 and 100/400 produce a percentage of 25%. But one is quite different from the other.

Ordinal data is categorical where the categories have a natural ordering, but the spacing between categories is ambiguous. When assigning grades, the difference between an A and a B is not quite the same as the difference between a B and a C. It certainly is quite different from the difference between a D and an F.

You should summarize ordinal data using a median and the interquartile range.

Interval and ratio data represent measurements where differences are considered meaningful. A red blood cell count of 4.2 million per ml differs from another count of 4.1 million by same amount as counts of 3.6 million per ml versus 3.5 million per ml.

There is quite a bit of controversy about ordinal data. Can you take two or more ordinal measurements and add them together? Can you average an ordinal measurement? There is no easy answer to this. I am fairly liberal in allowing this, but others who are smarter than I am disagree.

Other distinctions

- Discrete
  - Fractional values not allowed
- Continuous
  - Potentially any value in some interval
- Binary
  - Code as 0 and 1

Speaker notes

Some people take interval/ratio data make a further distinction between discrete and continuous. A discrete variable does not allow fractional values, only whole numbers. If you want to quibble, you could also say that something like shoe sizes where you can have a limited number of sizes like 9 1/2 that are not whole numbers.

Counts are a special case of discrete data. Rates are a count divided by a measure of time. There are some specialized statistical methods for counts and rates.

Break #2

- What you have learned
  - Data types
- What’s coming next
  - Descriptive statistics

Counts/proportions/rates

- Counts are special type of discrete data
- Proportion is count divided by bigger count
- Rates is count divided by time

Speaker notes

Counts are a special case of discrete data. It has a lower bound of zero and often no firm upper bound.

A proportion is a count divided by a bigger count. Typically the numerator represents a subset of the denominator. A proportion is always between zero and one.

Rates are a count divided by a measure of time. Rates can never be negative, but can sometimes be larger than one.

There are some specialized statistical methods for counts, proportions, and rates.

Demonstrate how to get counts and proportions in SPSS

Errors

- Measurement
- Validity
- Reliability
- Sampling

Speaker notes

There are several sources of error in Statistics. Measurement error occurs typically with physical measurements and reflects the fact that these measurements are not always as precise as we would like.

Finding ways to reduce measurement error are an important part of designing good research studies.

Errors in validity and reliability occur typically with constructs, though they can occur with physical measurements as well. A construct is a theoretical quantity that can only be measured indirectly. Validity errors occur when the indirect measurement is not an accurate representation of the theoretical construct.

Reliability errors occur when the measurements differ by observer or over short spans of time.

Sampling error is uncertainty caused by using a sample (a subset of the population) to represent the results in the entire population.

Descriptive statistics

- Mean, median
- Percentiles, quartiles
- Standard deviation
  - Special application with normal data

Speaker notes

The mean and the median are both measures of central tendency. You calculate the mean by adding up all the values in the data and dividing by the sample size. The median is the middle value or the average of the two middle values after sorting the data from low to high.

Percentiles are also computed after sorting the data. They represent the value so that a fixed proportion of the data is smaller and the rest is larger. The 10th percentile is the value where roughly 10% of the data is smaller and 90% of the data is larger.

The lower quartile is the 25th percentile and the upper quartile is the 75th percentile. These two values plus the median split the data into four groups, each having roughly 25% of the data.

The standard deviation is a measure of variation. If the data is roughly normally distributed then about 68% of the data lies within one standard deviation of the mean, about 95% of the data lies within two standard deviations of the mean and almost all of the data lies within three standard deviations of the mean.

Demonstrate how to get means and standard deviations in SPSS

Behavior of the mean versus an individual

- Central Limit Theorem
  - Sample mean is approximately normal
  - Even if individual observations are not
- $s.e.(\bar{X}) = \frac{s}{\sqrt{n}}$

Speaker notes

Standard deviation divided by the square root of the sample size.

Break #3

- What you have learned
  - Descriptive statistics
- What’s coming next
  - Confidence intervals and hypothesis tests

Definitions

- Population, sample
- Parameter, statistic
- Null and alternative hypotheses
- Type I and Type II error rates
- P-values

General form of confidence interval

- statistic $\pm$ t or z $\times$ s.e.(statistic)
  - $\bar{X} \pm t(1-\alpha/2; n-1)\frac{S}{\sqrt{n}}$

General form of test

- T = (statistic-hypothesized value) / s.e.(statistic)
  - $\frac{\bar{X}-\mu 0}{S/\sqrt{n}}$

Demonstrate one sample t-test and confidence interval in SPSS

Break #4

- What you have learned
  - Confidence intervals and hypothesis tests
- What’s coming next
  - Visualization

Graphical displays

- Scatterplots
- Boxplots
- Bar charts

Demonstrate plots in SPSS

Break #5

- What you have learned
  - Visualization
- What’s coming next
  - Assessing normality

Problems caused by non-normality

- Outliers distort your perspective
- Inaccurate Type I error rates
- Confidence intervals too narrow or wide

Speaker notes

Sometimes, I think that researchers obsess too much about non-normality, but in fairness to them, it is an important issue. If your data does not follow a bell shaped curve, then several problems could happen.

First, you might see a greater degree of imprecision, reflected in very wide confidence intervals and loss of statistical power.

Second, you might have poor coverage probability. The 95% confidence interval might only reprent a 92% confidence level. The Type I error rate might be greater than 5%.

Third, and this point is not emphasized enough, is that non-normal data makes it difficult for you to extrapolate to future observations. Prediction of future events is a big part of Statistics. It is difficult even with normal data and becomes much more difficult with non-normal data.

Yes, you might say, but doesn't the Central Limit Theorem help us out? Well, yes, if the sample size is large, but it only helps with assuring accurate confidence levels and good control of the Type I error rate. Non-normal data will still often produce intervals that are too wide and tests that have too little power.

Is your data normally distributed, 1 of 2

- Best approaches
  - Q-Q plot
  - Histogram
  - Boxplot

Speaker notes

The easiest way to assess normality is visually. It is admittedly subjective, but objective tests like Shapiro-Wilk test are not good.

Is your data normally distributed, 2 of 2

- Non-visual approaches
  - Does mean differ only slightly from the median
  - Is the median halfway between the two quartiles
- For non-negative data
  - Is the standard deviation large relative to the mean
- Calculate skewness and kurtosis
- Do not rely on tests of normality

Speaker notes

If you are reading a paper and the authors do not provide a graphic visualization of normality (page limits are an issue!) then there are a few non-visual approaches that can sometimes help. They are not as good as a visual approach, though.

Demonstrate how to get Q-Q plots and histograms in SPSS

Break #6

- What you have learned
  - Assessing normality
- What's coming next
  - Independent samples and paired samples t-tess

Two sample t-test

- $H_0: \ \mu_1 = \mu_2$
- $H_1: \ \mu_1 \neq \mu_2$
- Assumptions
  - Normality
  - Equal variances
  - Independence

Speaker notes

Things started getting interesting around week 7. The two sample t-test (also known as the independent samples t-test) compares the population means of two groups.

There are three key assumptions.

Test statistic and confidence interval

- Standard error
  - $se = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  - $S_p = \sqrt{\frac{(n_1-1)S_1^2+n_2-1)S_2^2}{n_1+n_2-2}}$
- 1-α confidence interval
  - $(\bar{X}_1 - \bar{X}_2) \pm t(1-\alpha/2; n_1+n_2-2) \times se$
- Test statistic
  - $T = \frac{\bar{X}_1-\bar{X}_2}{se}$
  - Accept H_0 if T is close to zero, or if CI includes 0

Speaker notes

The confidence interval for the difference in population means adds and subtracts a t percentile times the standard error of the difference in means.

The test statistic is the difference in sample means divided by the standard error.

Demonstrate the two-sample t-test in SPSS

Three things you need for a power calculation

- Research hypothesis
- Standard deviation of your outcome measure
- Minimum clinically important difference

Speaker notes

When you are planning a research study, you should select a sample size that gives you reasonable power (at least 80%, but 90% is better). The power calculation requires three things, and getting the last two, the standard deviation and the minimum clinically important difference, can be tricky.

Demonstrate power calculation in SPSS

Paired t-test

- Natural pairing
  - Left eye and right eye
  - Before and after measures
  - Closely related familty members
  - Litters
- Designed pairing
  - Matching on age, race, gender, etc.

Research hypothesis

- $\mu_D = \mu_1 - \mu_2$
- $H_0 : \ \mu_D = 0$
- $H_1 : \ \mu_D \neq 0$
- Assumptions on $D_i = X_{1i} - X_{2i}$
  - Normality
  - Independence

Speaker notes

The hypothesis can be written in terms of the difference in means. Note that the assumptions are about the sample differences being normal and being independent. You could have a non-normal distribution for X1 and a similar non-normal distribution for X2, and if these cancel out when you take the difference, great!

The independence assumption is that the differences are independent of one another. You do not assume that X1 and X2 are independent within a pair. In fact, you hope that they are not independent, but rather that they are positively correlated. When the pairs are positively correlated, you can get a difference with a lot less noise.

Confidence interval and test statistic

- Standard error
  - $se = \frac{S_D}{\sqrt{n}}$
- 1-α confidence interval
  - $\bar{D} \pm t(1-\alpha/2)se(\bar{D})$
- Test statistic
  - $T = \frac{\bar{D}}{se}$
  - Accept H_0 if T is close to 0 or if CI includes 0

Demonstrate paired t-test in SPSS

Break #7

- What you have learned
  - Independent samples and paired samples t-tess
- What’s coming next
  - One factor ANOVA

Bonferroni adjustment

- For m hypotheses
  - $P[E_1 \cup \ldots \cup E_m] \leq m\alpha$
- Test each hypothesis at α/m
  - Preserves overall Type I error rate
- Example, 3 simultaneous hypotheses
  - Reject H0 if p-value < 0.0133
- Should you be concerned about overall Type I error rates?

Speaker notes

If the probability of a Type I error rate is alpha for a series of m hypotheses, then the chances of making at least one Type I error is m times alpha. If you decrease the Type I error rate for each individual hypothesis to alpha divided by m, then you can safely state that the overall Type I error rate is less than alpha.

Should you try to control the overall Type I error rate? This is controversial, though most researchers agree that you should be concerned.

ANOVA research hypothesis

- $H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1: \ \mu_i \neq \mu_j$ for some i, j
  - Note: do not use $H_1: \ \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$

Speaker notes

If you have k groups with k population means, the null hypothesis is that all the population means are equal. Be careful, though. The alternative hypothesis is that there are some means that might not be equal. It is not the same as saying that all the population means are unequal.

If you have three groups, you should reject the null hypothesis if the first two population means were equal, but the third population mean was different than the first two.

Important assumptions

- Same as independent-samples t-test
  - Normality
  - Equal variances
  - Independence

How to check assumptions

- Boxplots
- Analysis of residuals, e_{ij}
  - e_{ij}= Observed - Predicted
  - $e_{ij} = Y_{ij} - \bar{Y}_i$
- Do not assess normality using Y_{ij}

Speaker notes

Boxplots of each group are useful for checking the assumptions of normality and equal variances. You can also compute the residuals. Residuals are defined in the section on linear regression, but they can be computed here. They represent what you observed minus what you would predict based on the model. This is essentially just subtracting off the group means.

If you tried to assess normality using the Y's, you might see a multi-modal distribution, but this is not a violation of the assumptions.

Tukey post hoc tests

- If you reject H0, which values are unequal?
  - With k groups, there are k(k-1)/2 comparisons
- Studentized range (Tukey test)
  - Requires equal sample sizes per group
  - Uses harmonic mean approximation for unequal sample sizes.
    - Do not use harmonic means if seriously different sample sizes.

Speaker notes

If you reject the null hypothesis, you can conclude that there are at least two groups out of k where the population means differ from one another. This is a rather vague and open-ended conclusion. You can use several procedures to identify where the differences lie. The most common is the studentized range test, more commonly called the Tukey post-hoc test.

Demonstrate ANOVA in SPSS

Break #8

- What you have learned
  - One factor ANOVA
- What’s coming next
  - REDCap and one sample confidence intervals

Bad joke

- Researcher gets 6 year, $10 million grant
- Writes in final report
  - “This is a new and innovative surgical technique and we are 95% confident that the cure rate is somewhere between 3% and 98%.”

Speaker notes

I tell this fictional story and it illustrates why sample size calculations are important. You want to assure the granting agency that if they spend a lot of money, you won’t produce a confidence interval so wide as to be meaningless.

Review confidence interval for a single mean

- If n < 30,
  - $\bar{X} - t(\alpha/2; n-1)se(\bar{X})$ and
  - $\bar{X} + t(\alpha/2; n-1)se(\bar{X})$
- If n $\geq$ 30,
  - $\bar{X} - z(\alpha/2)se(\bar{X})$ and
  - $\bar{X} + z(\alpha/2)se(\bar{X})$

Speaker notes

One of the simplest way to justify sample sizes, especially for a descriptive study is to specify a desired width for your confidence interval and show that it can be produced with the sample size you expect to collect.

It is a bit arbitrary because you won’t have any data on your standard deviation, but just make a good faith effort by reviewing the existing literature and any previous work you yourself may have done in the area.

Confidence interval for a single proportion

- Define standard error
  - $se(p) = \sqrt{\frac{p(1-p)}{n}}$
- $1 - \alpha$ confidence interval
  - $p - Z(\alpha/2)se(p)$ and
  - $p + Z(\alpha/2)se(p)$

Speaker notes

If your key statistic is a proportion rather than a mean, you can still look at the confidence interval as a way to justify sample size. You do need a rough estimate of the sample proportion, and again this requires a literature review and possibly any data you may have yourself from previous studies.

The widest confidence interval occurs when the sample proportion equals 0.5. So if you don’t have data from the literature review or your earlier studies, plug in 0.5 in this formula for a worst case scenario.

This confidence interval uses z rather than t, regardless of the sample size.

REDCap

- Highly recommended for data entry
- Requires careful pre-planning

Speaker notes

Dave Walsh provided a nice overview of REDCap. It does require some pre-planning prior to data collection, but this is good. You should think carefully about how you enter data, and REDCap helps you focus on this.

Demonstrate one sample confidence intervals in SPSS

Break #9

- What you have learned
  - REDCap and one sample confidence intervals
- What’s coming next
  - Chi-square tests

The two by two table

		Outcome	
		Yes	No
Exposure	Yes	O_{11}	O_{12}
	No	O_{21}	O_{22}

Speaker notes

One of the most common statistics you will encounter are counts in a two by two table.

Titanic data

	Alive	Dead	Total
Female	308	154	462
Male	142	709	851
Total	450	863	1,313

Speaker notes

Both the odds ratio and the relative risk compare the likelihood of an event between two groups. Consider the following data on survival of passengers on the Titanic. There were **462 female passengers: 308 survived and 154 died**. There were **851 male passengers: 142 survived and 709 died**.

Clearly, a male passenger on the Titanic was more likely to die than a female passenger. But how much more likely? You can compute the odds ratio or the relative risk to answer this question.

Odds ratio calculation

	Alive	Dead	Total	Odds
Female	308	154	462	154/308 = 0.5
Male	142	709	851	709/142 = 4.993
Total	450	863	1,313	

OR = 4.993 / 0.5 = 9.986

Speaker notes

The odds ratio compares the relative odds of death in each group. For females, the odds were exactly **2 to 1 against dying (154/308=0.5)**. For males, the odds were almost **5 to 1 in favor of death (709/142=4.993)**. The odds ratio is **9.986 (4.993/0.5)**. There is a **ten fold greater odds of death for males than for females**.

Relative risk calculation

	Alive	Dead	Total	Probability
Female	308	154	462	154/462 = 0.3333
Male	142	709	851	709/851 = 0.8331
Total	450	863	1,313	

RR = 0.8331 / 0.3333 = 2.499

Speaker notes

The relative risk (sometimes called the risk ratio) compares the probability of death in each group rather than the odds. For females, the probability of death is **33% (154/462=0.3333)**. For males, the probability is **83% (709/851=0.8331)**. The relative risk of death is **2.5 (0.8331/0.3333)**. There is a **2.5 greater probability of death for males than for females**.

Expected deaths

P[Alive] = 450/1313 = 0.3427 P[Dead] = 863/1313 = 0.6572

	Alive	Dead
Female	462*0.3427	462*0.6572
Male	851*0.3427	851*0.6572

Observed versus expected

Observed

	Alive	Dead
Female	308	154
Male	142	709

Expected

	Alive	Dead
Female	158.3	303.6
Male	291.6	559.3

General form of Chi-squared test

- $\sum \frac{(O-E)^2}{E}$
  - O = Observed, E= Expected

Demonstrate Chi-squared test of association in SPSS

Chi-square test of goodness of fit

- Categorical variable with k levels
- $H_0: _1=_2=…=_k=

Example, clinic recruitment

Observed

Clinic	A	B	C	D	E	Total
Patients recruited	17	29	37	15	27	125

Expected

Clinic	A	B	C	D	E	Total
Patients recruited	25	25	25	25	25	125

Demonstrate Chi-squared test of goodness of fit in SPSS

Break #10

- What you have learned
  - Chi-square tests
- What’s coming next
  - McNemar test

Paired binary data

- Before and after intervention
- Left eye and right eye
- Control matched on age, sex, race
- Family member/same litter

Data layout for McNemar’s test

$a \quad b$

$c \quad d$

- $T =

Speaker notes

To compute McNemar’s test, you first need to calculate the two by two crosstabulation of the “before” and “after” measurements.

it is arbitrary whether the rows represent the “before” measurements and the columns represent the “after” measurements or the reverse. The order of the rows and columns is also arbitrary, but you must be consistent. If the binary variable is yes/no, then you can have the first row and the first column being “yes” or you can have the first row and column being “no” but you shouldn’t have the first row be “yes” and the first column be “no”.

The test statistic is the difference in the off-diagonal terms squared divided by the sum of the off-diagonal terms.

Demonstrate McNemar’s test in SPSS

Break #11

- What you have learned
  - McNemar test
- What's coming next
  - Nonparametric tests

Nonparametric = distribution free

- No need for normality assumption
- Most tests involve ranking

Mann-Whitney-Wilcoxon test, 1 of 2

- Two independent samples
  - $X_1, X_2, \ldots, X_n$
  - $Y_1, Y_2, \ldots, Y_m$
  - n, m do not have to be equal

Speaker notes

Add note.

Mann-Whitney-Wilcoxon (MWW) test, 2 of 2

- $U = \Sigma_i R(X_i)$
  - Accept H0 if U is close to $n \times \frac{n+m+1}{2}$
- $W = \Sigma_i \Sigma_j H[X_i - Y_j]$
  - H is a counting function
    - = 1 for zero or positive values
    - = 0 for zero or negative values
  - Accept H0 if W is close to $\frac{nm}{2}$

Demonstrate Mann-Whitney-Wilcoxon test in SPSS

Break #12

- What you have learned
  - Nonparametric tests
- What’s coming next
  - Correlations

Correlations

- Pearson correlation
  - $Cov(X, Y) = \frac{1}{n-1}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$
  - $Corr(X, Y) = \frac{Cov(X,Y)}{S_X S_Y}$
- Spearman correlation
  - $Corr(R(X), R(Y))

Demonstrate Pearson and Spearman correlations in SPSS

Break #13

- What you have learned
  - Correlations
- What’s coming next
  - Linear regression

Linear regression interpretation of a straight line

- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.

Speaker notes

In linear regression, we use a straight linear to estimate a trend in data. We can’t always draw a straight line that passes through every data point, but we can find a line that “comes close” to most of the data. This line is an estimate, and we interpret the slope and the intercept of this line as follows:

Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

The population model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, N$
  - ϵ_i is an unknown random variable
    - Mean 0, standard deviation, σ
    - Often assumed to be normal
  - β_0 and β_1 are unknown parameters
  - b_0 and b_1 are estimates from the sample

Violations of this model

- Nonlinearity
- Heterogeneity
- Non-normality
- Lack of independence

ANOVA table for linear regression

	SS	df	MS	$F-ratio$
$Regression$	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
$Error$	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
$Total$	SST	$n-1$		

Demonstrate linear regression in SPSS

Summary (1/2)

- What you have learned
  - Running SPSS
  - Data types
  - Descriptive statistics
  - Confidence intervals and hypothesis tests
  - Visualization
  - Assessing normality
  - Independent samples and paired samples t-tess

Summary (2/2)

- What you have learned
  - One factor ANOVA
  - REDCap and one sample confidence intervals
  - Chi-square tests
  - McNemar test
  - Nonparametric tests
  - Correlations
  - Linear regression