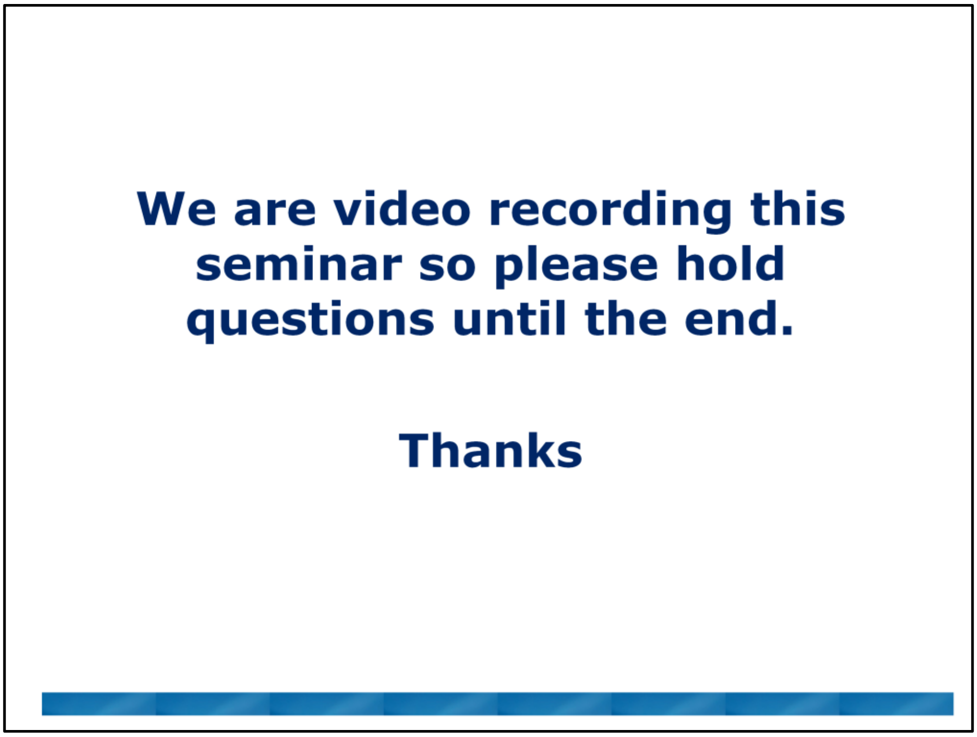
In Search of Significance: Hypothesis testing and exploratory data analysis

Sandra Taylor, Ph.D.
June 8 & 15, 2016

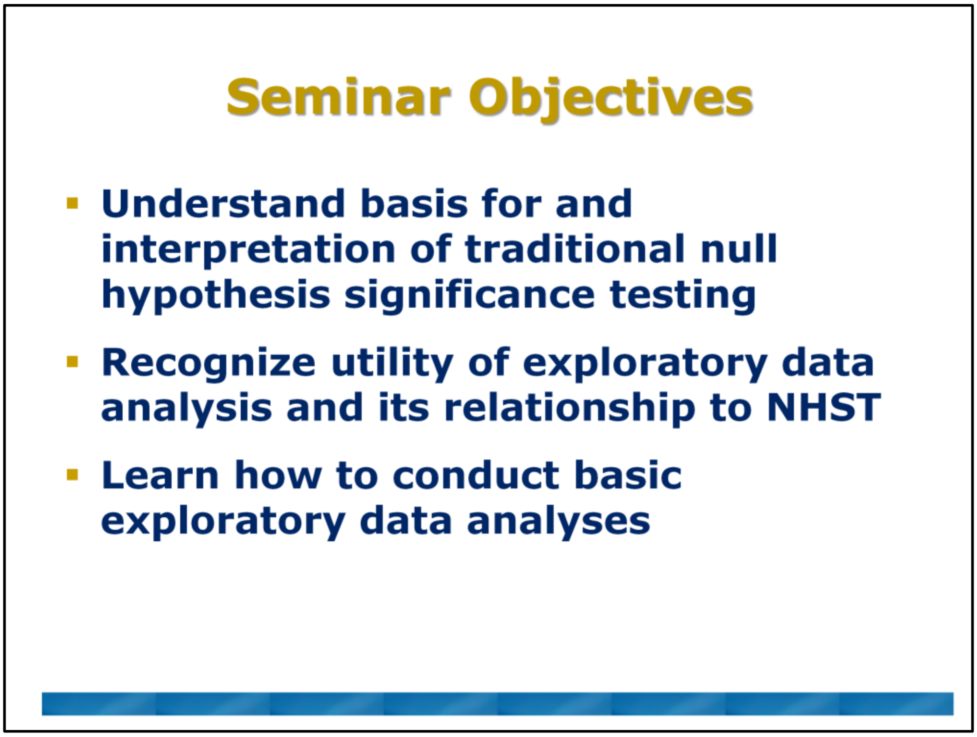
UCDAVIS CLINICAL AND TRANSLATIONAL SCIENCE CENTER
UCDAVIS MIND INSTITUTE
UCDAVIS COMPREHENSIVE CANCER CENTER
UCDAVIS ENVIRONMENTAL HEALTH SCIENCES CENTER

Hi, I’m Sandy Taylor. I am a statistician affiliated with the CTSC and MIND IDDRC Biostatistics Core. Welcome to the first talk in the inaugural year of a year-long seminar series on applied statistics. We hope that you will find this series valuable such that we can continue to expand and strengthen the program. This talk in particular is intended to be conceptual and general in the sense that I am not going to go into when and how to conduct specific statistical tests. Subsequent seminars will focus on specific tests and procedures.



We are video recording this seminar so please hold questions until the end.

Thanks



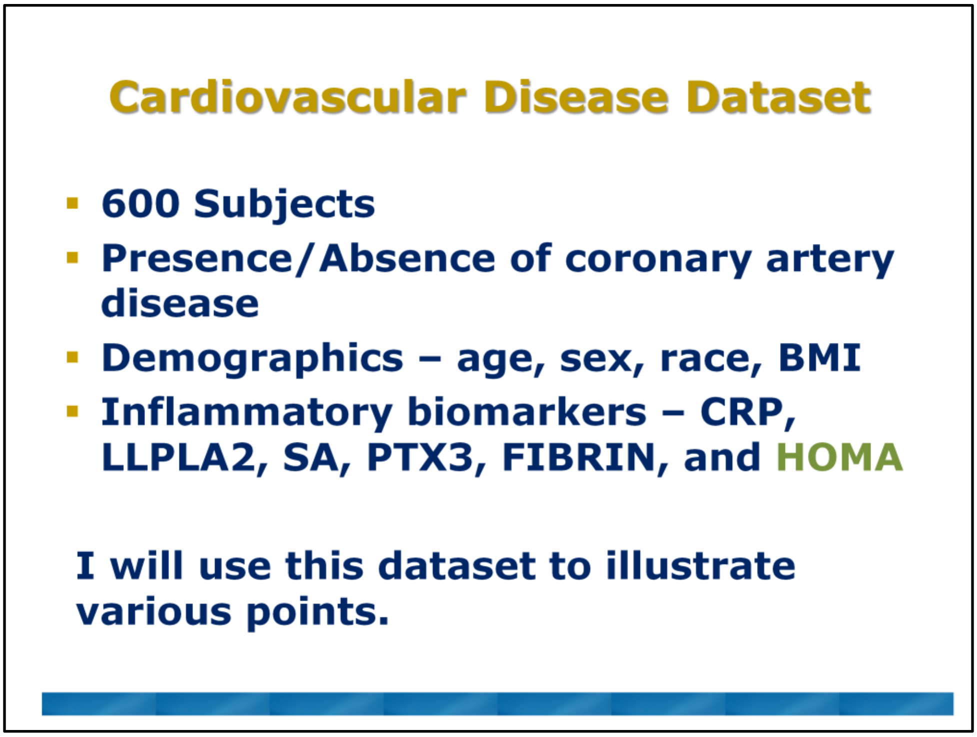
Seminar Objectives

- Understand basis for and interpretation of traditional null hypothesis significance testing
- Recognize utility of exploratory data analysis and its relationship to NHST
- Learn how to conduct basic exploratory data analyses

What this talk focuses on are critical fundamentals of statistics. As in every field, folks tend to be interested in the advanced skills and concepts but as every good coach will tell you, its fundamentals that wins games. In statistics, understanding and appreciating the fundamentals is similarly important, and will diminish the chance that you will do something stupid with your data.

My objectives for this seminar are for you to

- Understand the basis for and interpretation of traditional null hypothesis significance testing
- Recognize the utility of exploratory data analysis and its relationship to NHST
- Learn how to conduct basic exploratory data analyses

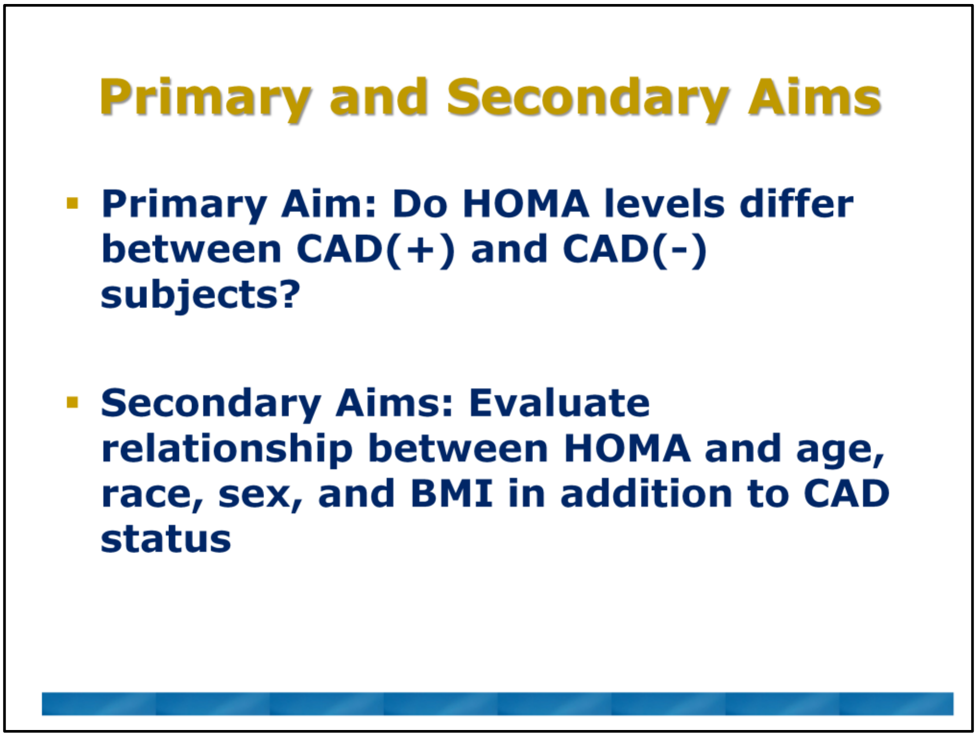


Cardiovascular Disease Dataset

- 600 Subjects
- Presence/Absence of coronary artery disease
- Demographics – age, sex, race, BMI
- Inflammatory biomarkers – CRP, LLPLA2, SA, PTX3, FIBRIN, and HOMA

I will use this dataset to illustrate various points.

I am going to use a data set on inflammatory biomarkers and cardiovascular disease as an example data set to illustrate various points. This data set has 600 subjects with approximately equal numbers of subjects with and without CAD. We have some basic demographics age, sex, race and BMI and we have values for 6 inflammatory biomarkers. For this talk, I am just going to focus on one of them HOMA. Homeostasis Model Assessment (HOMA) estimates steady state beta cell function (%B) and insulin sensitivity (%S), as percentages of a normal reference population.

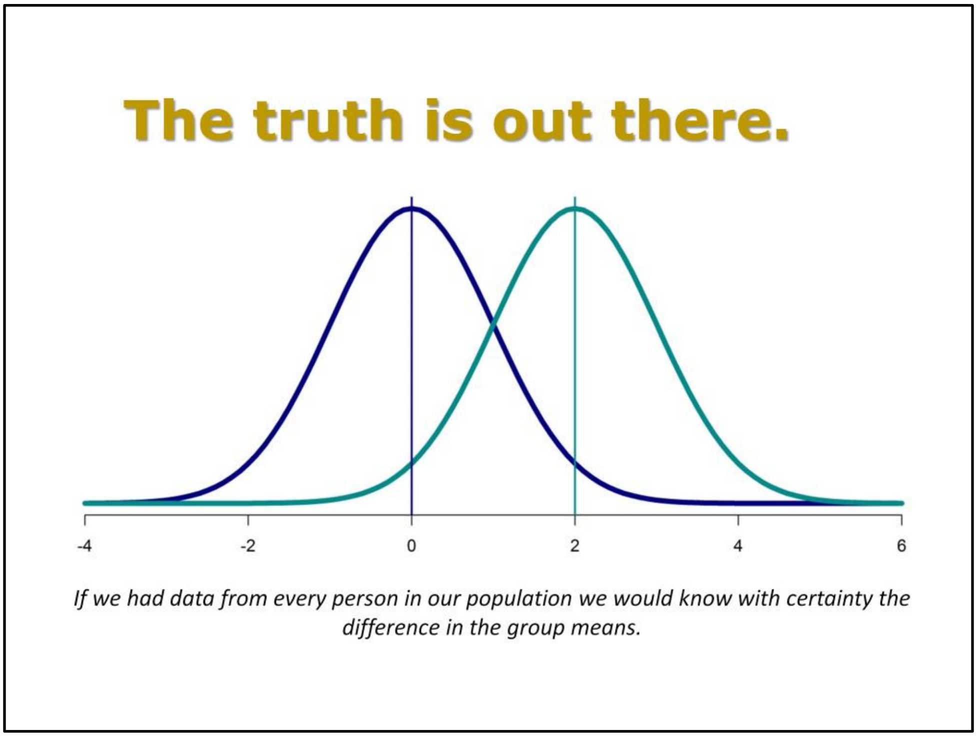


Primary and Secondary Aims

- Primary Aim: Do HOMA levels differ between CAD(+) and CAD(-) subjects?
- Secondary Aims: Evaluate relationship between HOMA and age, race, sex, and BMI in addition to CAD status

While I am sure we can think of a multitude of questions we might want to investigate with this data set, let’s suppose that are primary aim is in whether HOMA levels differ between CAD status. For secondary aims, we might want to further examine the relationship between HOMA and CAD status but with consideration of demographics as covariates.

I am going to start with the primary aim to illustrate null hypothesis significance testing and then use the secondary aims to illustrate exploratory data analysis.



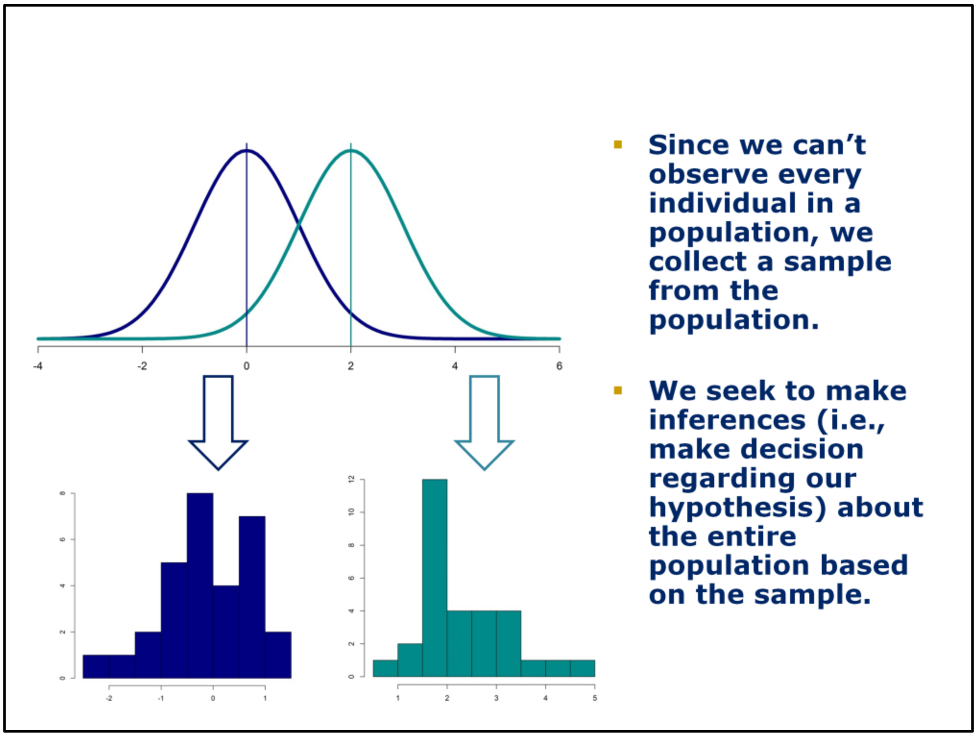
The truth is out there.

If we had data from every person in our population we would know with certainty the difference in the group means.

For our primary aim, we want to know if the mean levels of HOMA differ between CAD groups.

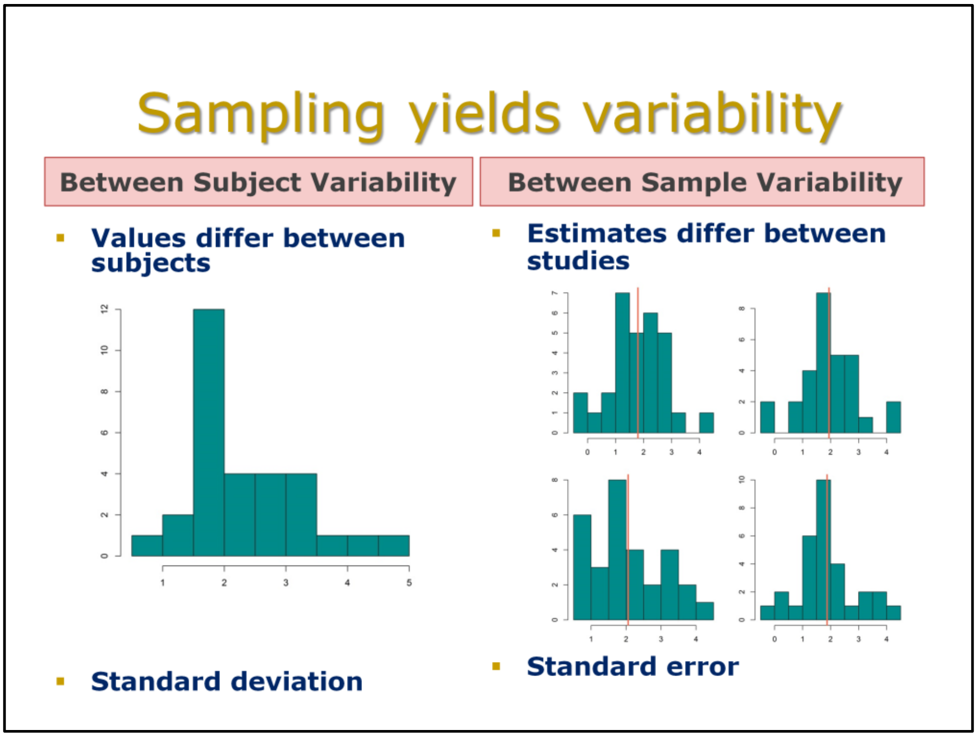
The truth is out there.

If we had data from every person in our population, we would know with certainty the difference in means between the groups.



- **Since we can't observe every individual in a population, we collect a sample from the population.**
- **We seek to make inferences (i.e., make decision regarding our hypothesis) about the entire population based on the sample.**

Clearly though we can't instantaneously measure HOMA levels and CAD status in every individual in our population of interest. Therefore, we collect a sample from the population and we seek to make inferences (i.e., make decision regarding our hypothesis) about the entire population based on a manageable sample.



Sampling yields variability

Between Subject Variability

- **Values differ between subjects**
- **Standard deviation**

Between Sample Variability

- **Estimates differ between studies**
- **Standard error**

When we take a sample from a population, we encounter two types of variability. One is the between subject variability. We are pretty tuned into this type of variability. People differ. The histogram on the left represents a sample taken from the population in the previous slide. This shows the distributions of the values obtained across subjects. Clearly there are differences between patients. This variability is summarized by the standard deviation.

The other type of variability we encounter is between study variability or between sample variability. We tend to be much less tuned into this type of variability. This is the variability that occurs between samples. The right hand side shows 4 different samples of the same size from the same distribution. The vertical line shows the sample mean. The true population mean is actually 2. What we see is variability in the shape of the distributions even though the source data is normally distributed and the means bounce around. I want you to really think about this and appreciate it. When you take a sample, you get one of these and you base your conclusions and estimate parameters on this one sample. With a different sample you would almost certainly have different parameter estimates (e.g., means) but could also draw a different conclusion.

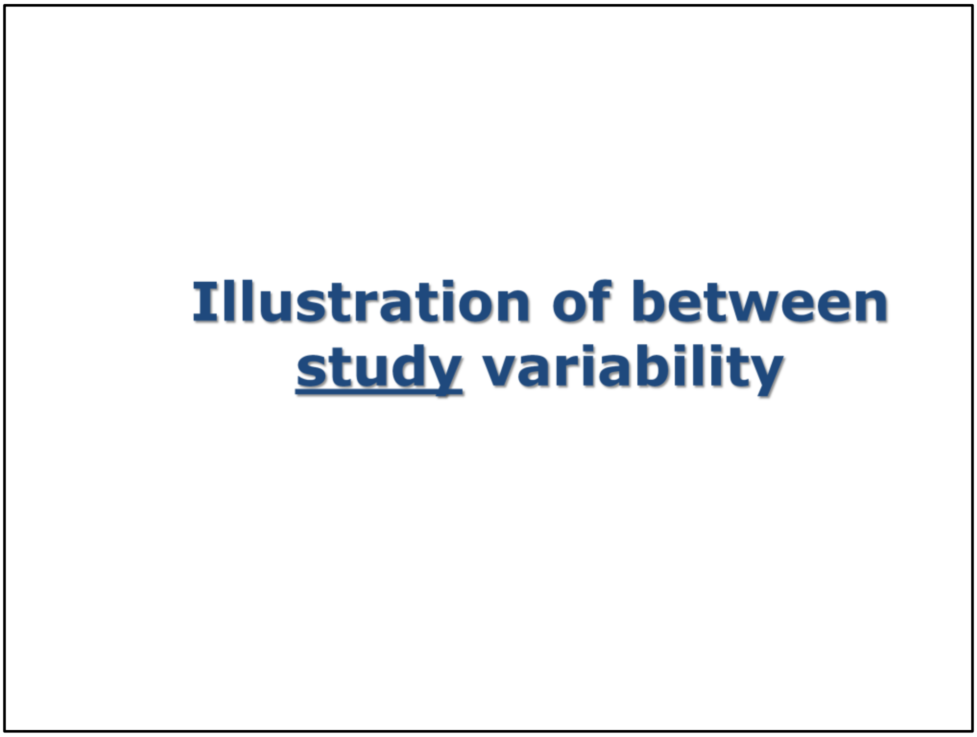


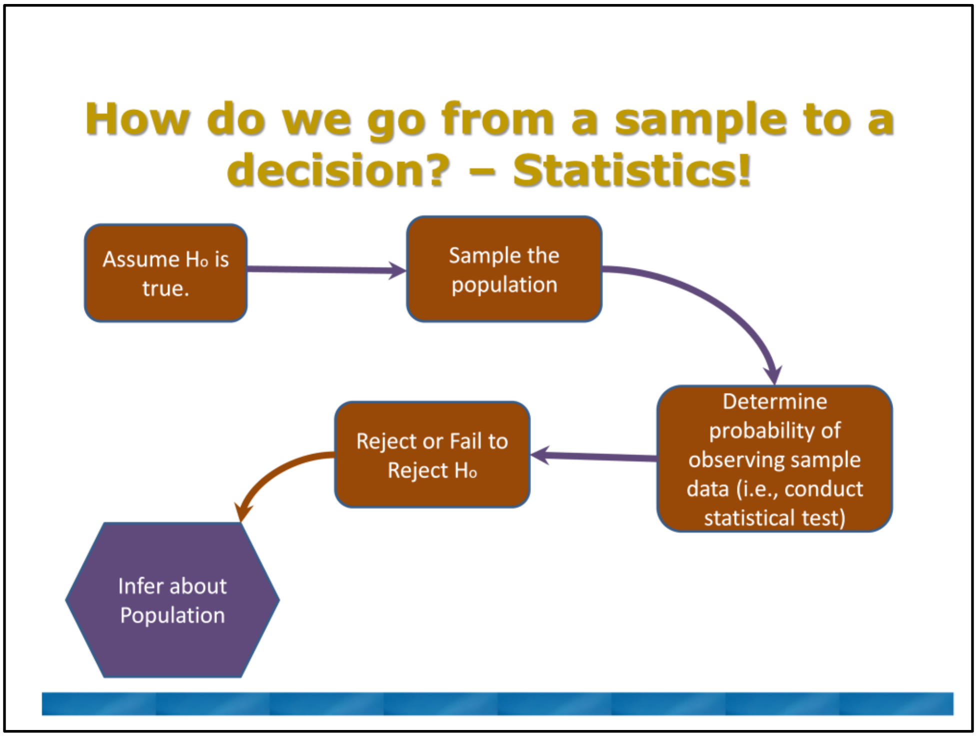
Illustration of between study variability

Hopefully I can get this simulation to work. What I want to do is further illustrate between study/sample variability.
Go to "Basics" "Distribution of means (continuous distribution)"

Here I simulate a distribution with a mean of 2 and standard deviation of 1 which is what was illustrated in the previous slides. Step through a couple samples. Then have it walk.

Points here are

1. Every sample is different
2. Means bounce all over the place but are largely centered on 2 the true mean
3. Mean of means converges to the true mean



How do we go from a sample to a decision? – Statistics!

Assume H_o is true.

Sample the population

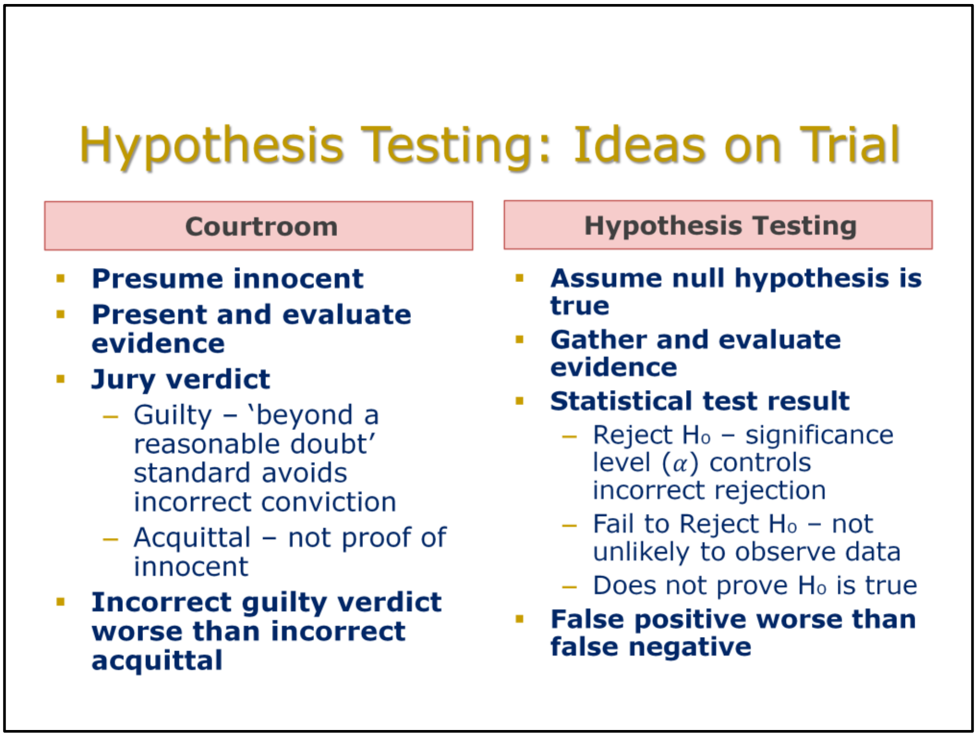
Determine probability of observing sample data (i.e., conduct statistical test)

Reject or Fail to Reject H_o

Infer about Population

Given the within and between sample variability, how to go from a sample to a decision? That is where statistics comes into play.

This diagram shows the steps in the process. First, we assume the null hypothesis is true. For example, we take the position that mean HOMA levels do not differ between CAD groups. Second, we collect a sample. Third, loosely we estimate the probability of observing the sample data if the null hypothesis is true. Fourth, based on that probability we make a decision of whether to reject the null hypothesis and fifth we extend our decision to the population of interest.



Hypothesis Testing: Ideas on Trial

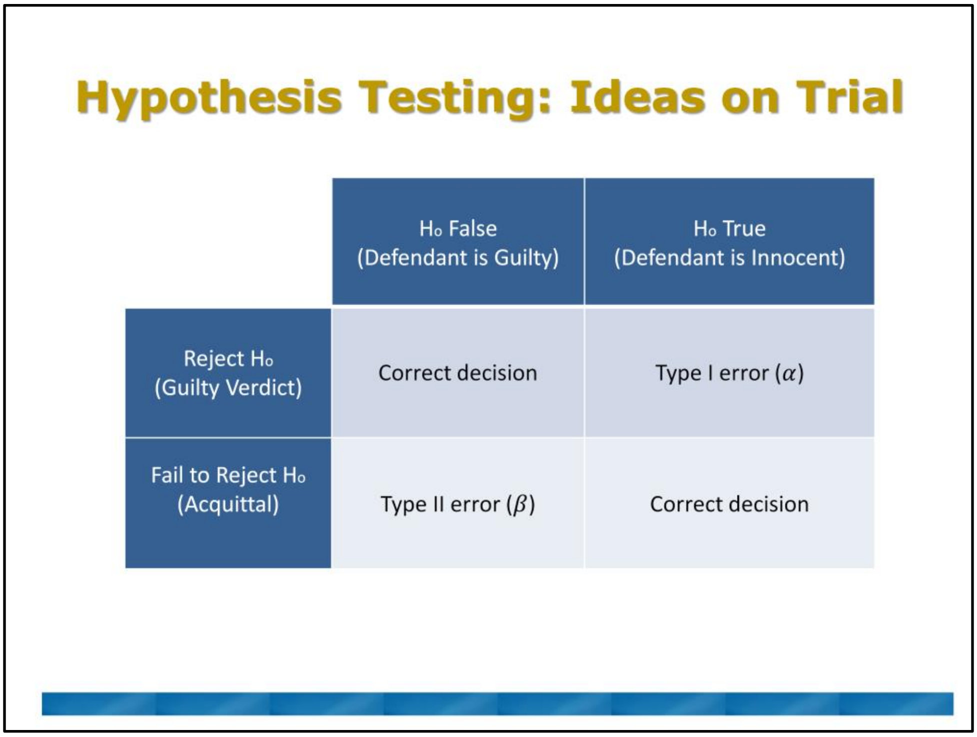
Courtroom

- Presume innocent
- Present and evaluate evidence
- Jury verdict
  - Guilty – 'beyond a reasonable doubt' standard avoids incorrect conviction
  - Acquittal – not proof of innocent
- Incorrect guilty verdict worse than incorrect acquittal

Hypothesis Testing

- Assume null hypothesis is true
- Gather and evaluate evidence
- Statistical test result
  - Reject H_0 – significance level (α) controls incorrect rejection
  - Fail to Reject H_0 – not unlikely to observe data
  - Does not prove H_0 is true
- False positive worse than false negative

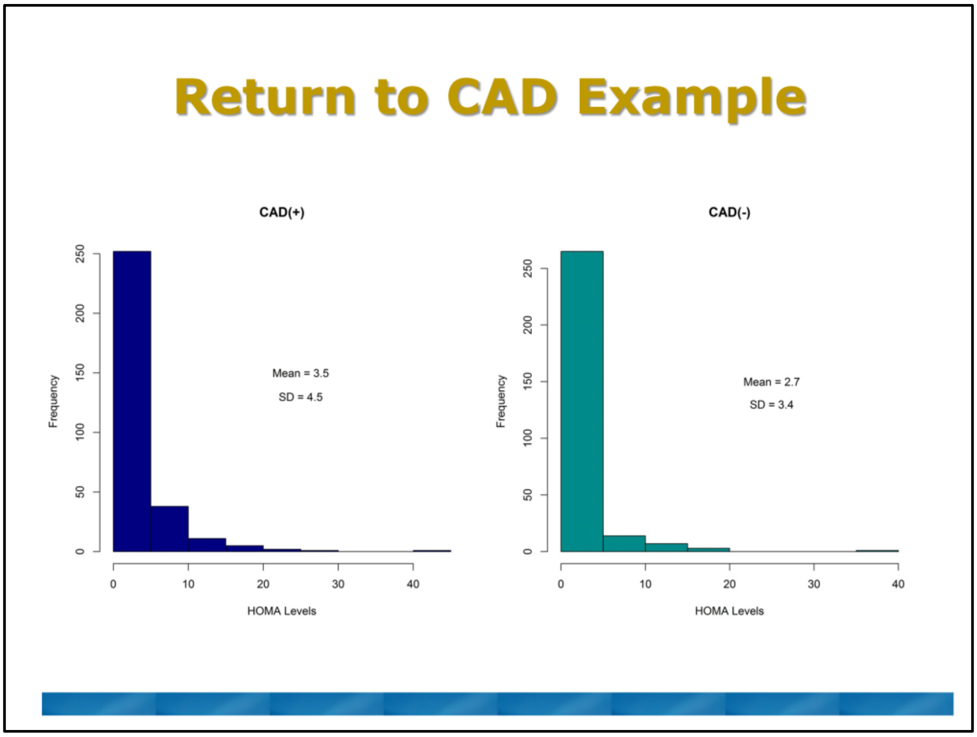
To firm up these ideas, let's take our judicial system as an analogy. In the US, when someone is put on trial, we presume they are innocent. This is analogous to assuming the null hypothesis is true. In a trial, the prosecution and defense present evidence and the jury evaluates that evidence. This is analogous to taking a sample from the population. The jury gives a verdict. The jury is instructed to find the defendant guilty if the evidence presented supports guilt "beyond a reasonable doubt" otherwise the jury is to acquit the defendant. The "beyond a reasonable doubt" standard is a high bar intended to minimize false convictions. An incorrect guilty verdict is deemed worse than incorrect acquittal. Analogously, in hypothesis testing, we reject Ho (i.e., find Ho guilty) if our p-value is < set significance level. We select a small significance level (i.e., 0.05) to guard against false positives given the perspective that false positives are worse than false negatives (i.e., not finding a difference). In both cases (courtroom and NHST) it is important to remember that acquittal/failure to reject is not proof of innocence/truth of null. It just means that there was not sufficient evidence to support the alternative conclusion.



Hypothesis Testing: Ideas on Trial

	H_o False (Defendant is Guilty)	H_o True (Defendant is Innocent)
Reject H_o (Guilty Verdict)	Correct decision	Type I error (α)
Fail to Reject H_o (Acquittal)	Type II error (β)	Correct decision

We can summarize the possible outcomes both of a trial and hypothesis testing. We make correct decisions if we convict a defendant who is guilty or acquit a defendant who is not guilty. We make an error if we convict an innocent person (false positive type I error) or acquit a guilty person (false negative type II error)

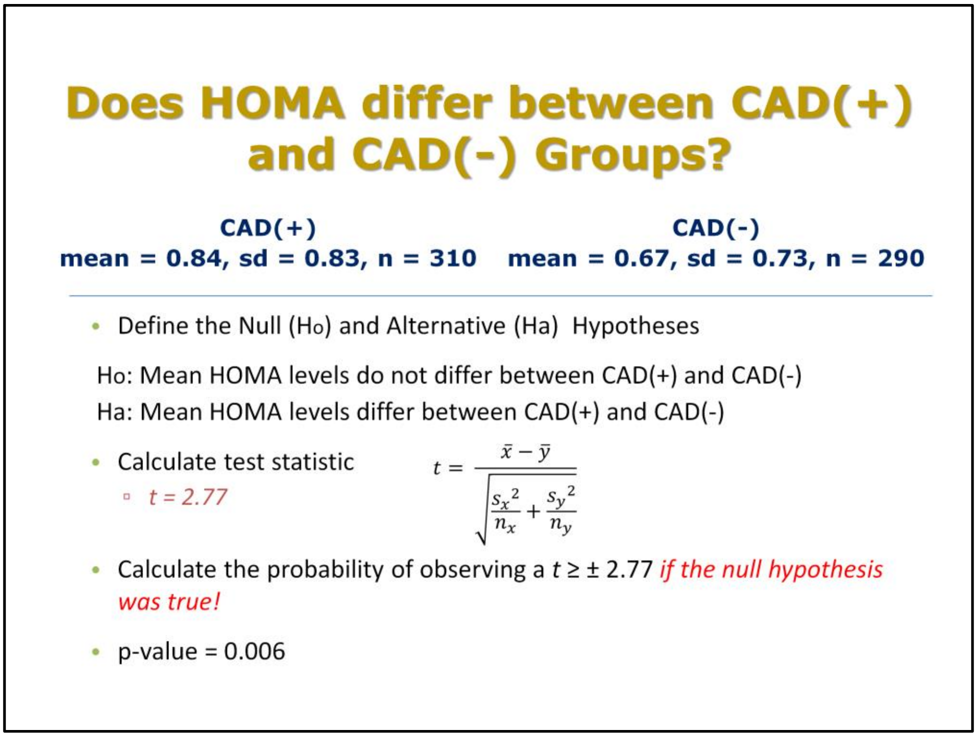


Return to CAD Example

With that background, let’s go back to the HOMA and CAD data.

Recall that we want to evaluate whether mean HOMA differs between CAD groups.

What does our data look like? We see that the shape of the distributions is similar between groups. Both are strongly right-skewed which we will deal with later. We can calculate the means and the standard deviations for each group which yields a mean of 3.5 in the CAD(+) and a mean of 2.7 CAD(-). So our sample data suggests that HOMA may be higher in CAD(+) subjects than CAD(-). We want to use this sample to make inferences, i.e., draw some conclusion about HOMA levels for CAD(+) and (-) people in our target population.



Does HOMA differ between CAD(+) and CAD(-) Groups?

CAD(+)	CAD(-)
mean = 0.84, sd = 0.83, n = 310	mean = 0.67, sd = 0.73, n = 290

- Define the Null (Ho) and Alternative (Ha) Hypotheses

Ho: Mean HOMA levels do not differ between CAD(+) and CAD(-)
Ha: Mean HOMA levels differ between CAD(+) and CAD(-)

- Calculate test statistic
  - *$t = 2.77$*

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{s_x}^2}{n_x} + \frac{{s_y}^2}{n_y}}}$$

- Calculate the probability of observing a $t \geq \pm 2.77$ *if the null hypothesis was true!*
- p-value = 0.006

Since the data are strongly right skewed, I log transformed the data to yield a more symmetric (e.g., bell shaped) distribution. With this transformation, we see of course that then mean of the CAD(+) group is larger than the CAD(-) group *IN THIS SAMPLE*

We define our null hypothesis as Mean HOMA levels do not differ between CAD(+) and CAD(-). We define in as the opposite of what we want to so because it is much easier to disprove something than to prove something.

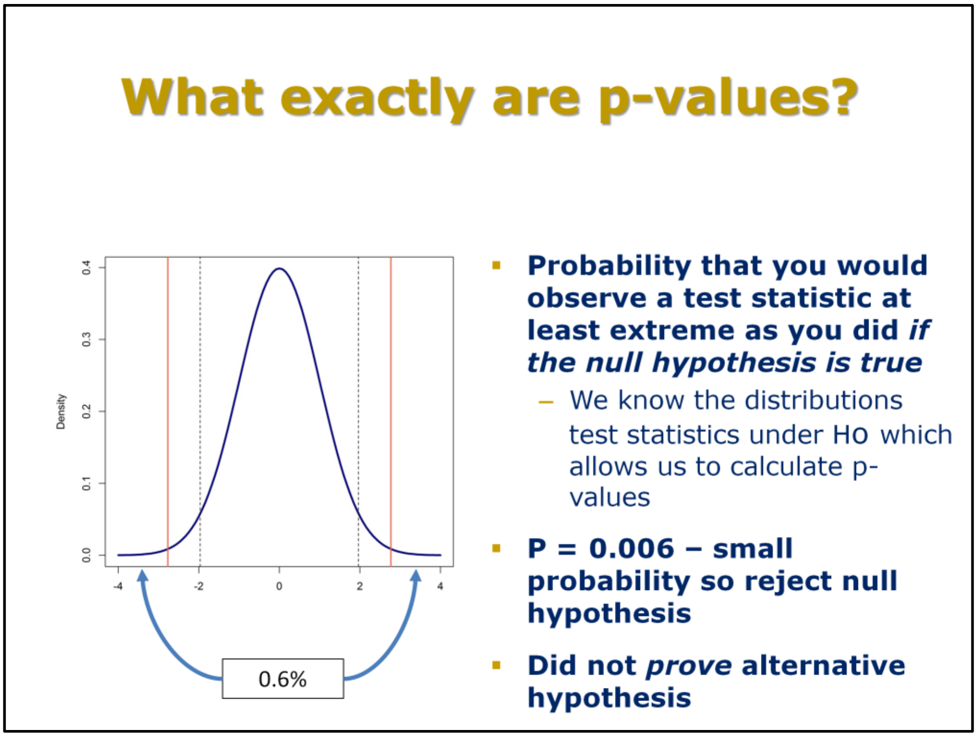
We then calculate a test statistic. In this case we use a t-test and calculate a t-statistic. The formula uses the difference between the observed means as a function of the data variability. We get a t statistic of 2.77.

Now, assume the null hypothesis is true, meaning that the means of the two groups are the same. If this were true then we would expect the value of the t-statistic to be 0 since the difference in the means would be 0.

So, if they were the same, how likely is it that we would get a t-statistics of 2.77 or greater. Remember the simulations I showed of different samples and how the means bounce around. We could have ended up with this difference simply by chance. Since we know how the statistic is distributed under the null hypothesis we can calculate the probability that we would get something this extreme simply by chance.

This turns out to be 0.006 or 0.6%. This is what a p-value is. Since it is pretty darn unlikely that we would see a difference like this if the null was true, we reject the null hypothesis.

In courtroom terms, we conclude the null is false “beyond a reasonable doubt”

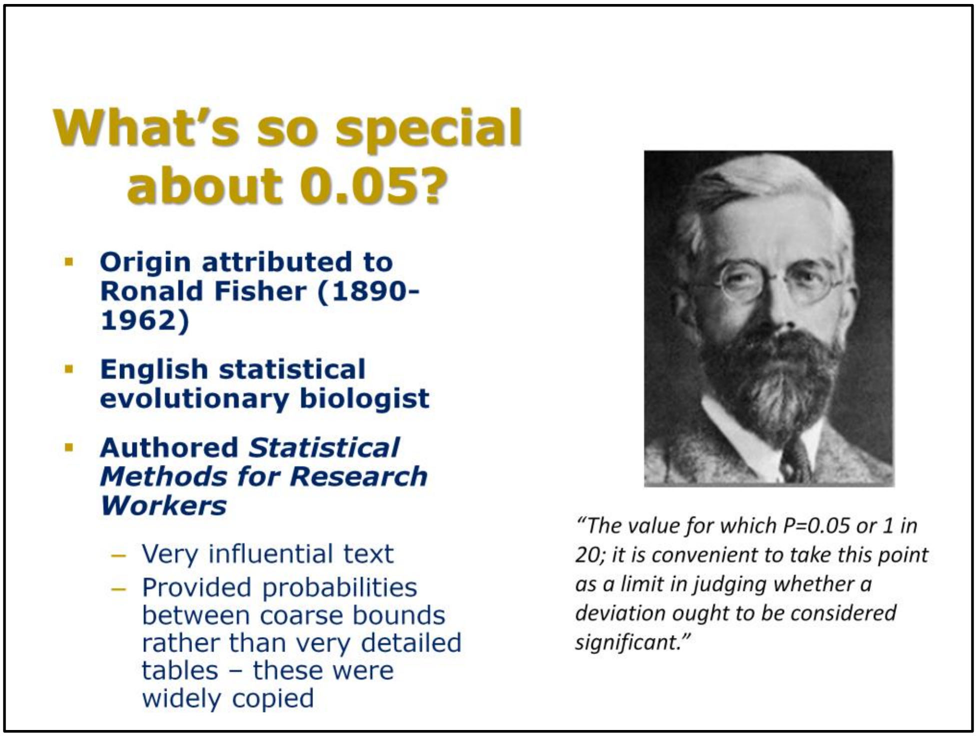


What exactly are p-values?

- **Probability that you would observe a test statistic at least extreme as you did *if the null hypothesis is true***
  - We know the distributions test statistics under Ho which allows us to calculate p-values
- **P = 0.006 – small probability so reject null hypothesis**
- **Did not *prove* alternative hypothesis**

I want to spend a little more time talking about p-values. I suspect you feel I am beating a dead horse but this is really important. There is a lot of misunderstanding about p-values. P-values are NOT the probability that the null hypothesis is true. It is the probability that you would observe a test statistic as or more extreme than the one you did IF the null was true.

Recall between study variability. With a different sample, we will get different mean values, different standard deviations hence a different test statistic and A DIFFERENT P-value.



What's so special about 0.05?

- **Origin attributed to Ronald Fisher (1890-1962)**
- **English statistical evolutionary biologist**
- **Authored *Statistical Methods for Research Workers***
  - Very influential text
  - Provided probabilities between coarse bounds rather than very detailed tables – these were widely copied

"The value for which P=0.05 or 1 in 20; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant."

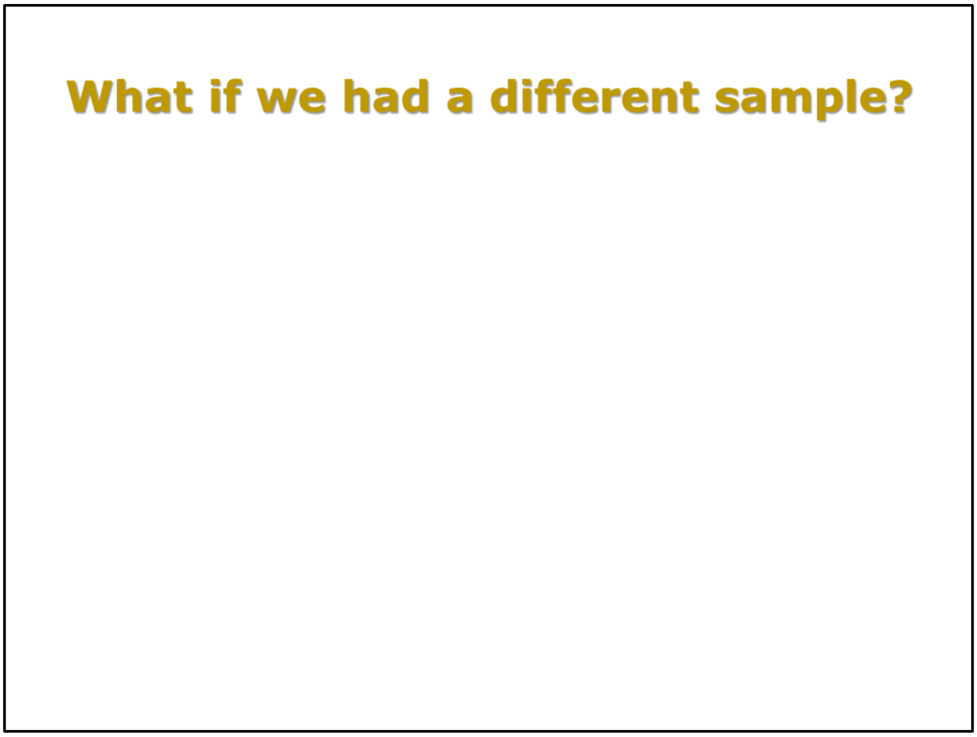
Everybody knows that a difference is "significant" if the p-value is less than 0.05. Right? So if you have p-value of 0.051 you have nothing but if it is 0.049 you can publish. Poppycock!

Did you ever stop to think about what's so special about 0.05?

The origin is largely attributed to Ronald Fisher. Ronald Fisher was an evolutionary biologist with a statistical bent who worked during the early to mid 1900s. He published Statistical methods for research workers. In this text, he provided probabilities of standard distributions (e.g., normal, t, chi-square) between coarse bounds (e.g., 5 to 10%) rather than at every 1%. Much more compact. The text was very influential and the tables were widely copied.

In addition to the coarse tables, Fisher is also credited with saying "The value….

Although there are other passages from the text, that deviate from the 0.05, it seems that this was largely the impetus for the birth of 0.05 as denoting "significance"



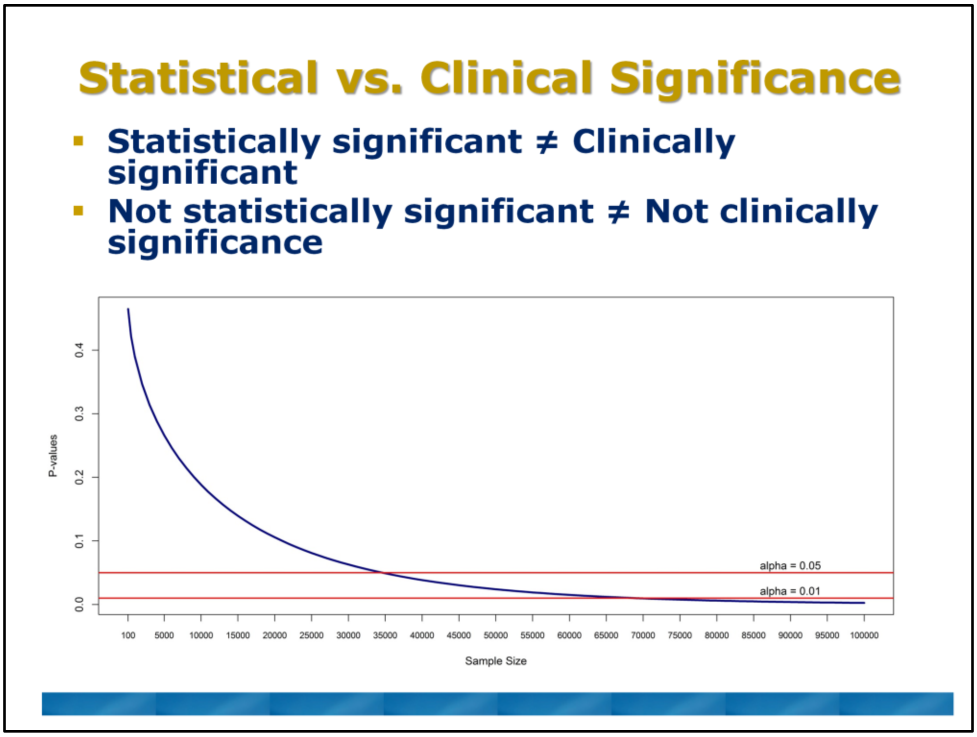
What if we had a different sample?

When we use for 0.05 to determine significance with this is our alpha/significance level. This value limits the Type I error rate – the rate at of false positives. It says if the Null hypothesis is TRUE, no more than 5% of the time we will reject it. So, if we repeatedly sampled, over and over again, and calculated the test statistic, if there truly is no difference in means, 5% of the time we would reject anyway.

Run two sample t-test with no difference

Note – most of the time (should be 95%) the test statistic is within the bounds but a few times it falls in the zone of "significance"

Remember – you don't know which of these samples you have or what the truth is.



Statistical vs. Clinical Significance

- **Statistically significant ≠ Clinically significant**
- **Not statistically significant ≠ Not clinically significance**

The next concept I want to touch on is statistical vs. clinical significance.
We found a p-value of 0.006 which is statistically significant and may well be of clinical significance as well.

This may not always be the case.
When we run a statistical test, we calculate a p-value and as we have seen the p-value gives the probability that we would observe a test statistic this extreme if the null hypothesis is true. In other words it is the probability of making a type I error if we chose to reject the null hypothesis. Note that this construct says absolutely nothing about clinical significance.

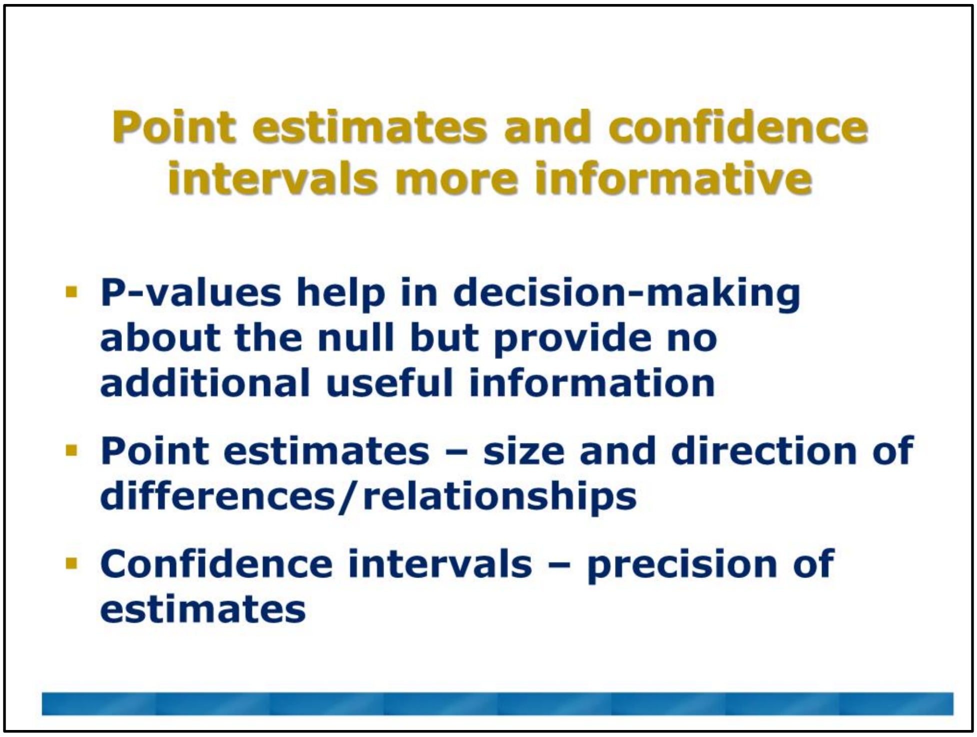
Just because something is statistically significant does not mean it is clinically significant and just because something is not statistically significant does not mean it is not clinically significant.

To illustrate this point, let's return to our HOMA data.

In our sample, the group without CAD has a mean of 0.67 on the log scale and the with CAD group has a mean of 0.84 yielding a difference of 0.17. What if the mean of the with CAD group was 0.68? Probably wouldn't consider this as clinically significant. Yet – with a large enough sample size this difference can become highly significant. What's happening here? Recall that our null hypothesis is that the means don't differ.

We don’t say anything about by how much. We just ask, “Are the means of these two groups different?” As our sample size gets larger we get more and more information about the true mean. Thus our estimates of the true means get more and more precise. Consider we have a difference of 0.01 and assume the standard deviation of the data is 0.8 similar to our sample. Let’s look at what happens to the p-value values as the sample size increases.

If we had a sample size of 35,000 a difference of 0.01 would give us a p-value < 0.05. A sample size of 70,000 for this same data would yield a p-value < 0.01. Statistically significant – Yes. Clinically significant – No. You may think that sample size like this are completely unrealistic but if you are using a national database, you can easily have sample sizes this large. In which case you are likely to find statistical significance but the magnitude of the difference might not be clinically significant. Which brings me to Point Estimates and Confidence Intervals.



Point estimates and confidence intervals more informative

- P-values help in decision-making about the null but provide no additional useful information
- Point estimates – size and direction of differences/relationships
- Confidence intervals – precision of estimates

Point estimates and confidence intervals are more informative than a p-value.

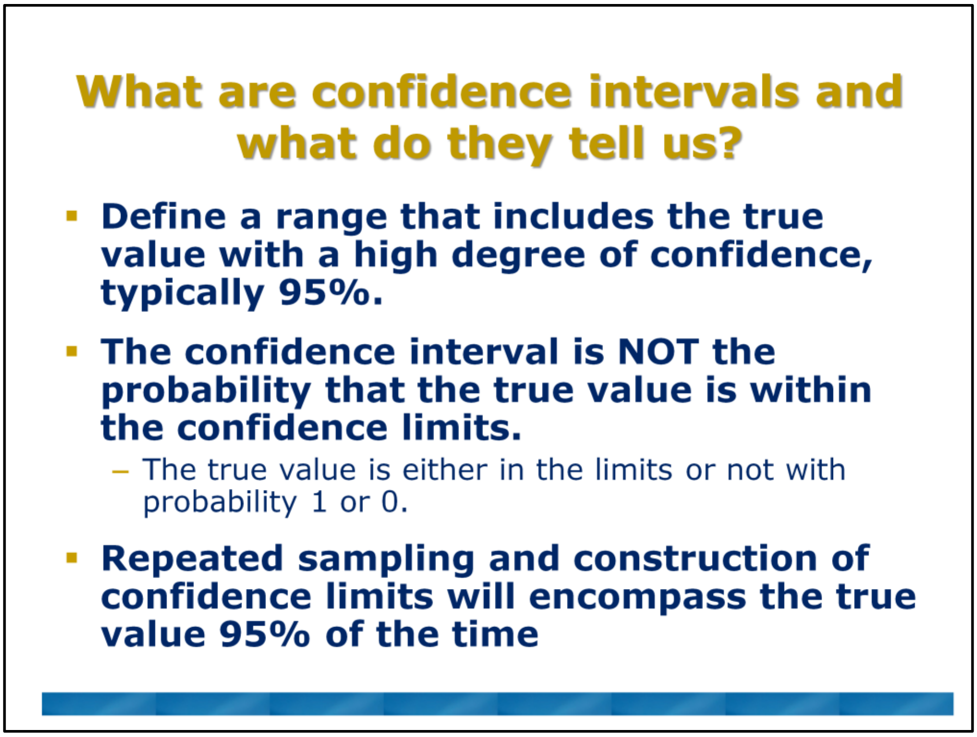
P-values help you in making a decision about the null hypothesis but provide no additional useful information.

Consider our HOMA example, if all I told you was that we found that the mean HOMA levels differ significantly between CAD(+) and CAD(-) subjects, what could you do with that information? Your next questions would or should be by how much and in which direction?

This is what you get with point estimates. They give you tangible information about the size and direction of differences or relationships. That is the fundamentally useful information from a study.

Then we have confidence intervals which should be equally important to you.

Confidence intervals give us a measure of the precision of our estimates. This is something we don’t use enough but is really important. It tells us how much our point estimates are likely to bounce around with different samples.



What are confidence intervals and what do they tell us?

- Define a range that includes the true value with a high degree of confidence, typically 95%.
- The confidence interval is NOT the probability that the true value is within the confidence limits.
  - The true value is either in the limits or not with probability 1 or 0.
- Repeated sampling and construction of confidence limits will encompass the true value 95% of the time

Confidence intervals tend to be misunderstood so I want to spend a little time on them.
What are confidence intervals and what do they tell us?

Confidence intervals define a range that encompasses the true population mean with a high degree of confidence. Recall that we have taken a sample of our populations (with and without CAD), and estimated the population difference in means between the two groups based on the sample. We know that if we took another sample that our estimate of the difference would be somewhat different. We want/need some idea of the range in the estimates that we would obtain given multiple samples and to define a range in which we have high confidence that it covers the true population value. Hence we define a 95% confidence interval for our estimate of the difference from our sample. Here again we need to recognize that our estimates and 95% confidence intervals will vary from sample to sample.

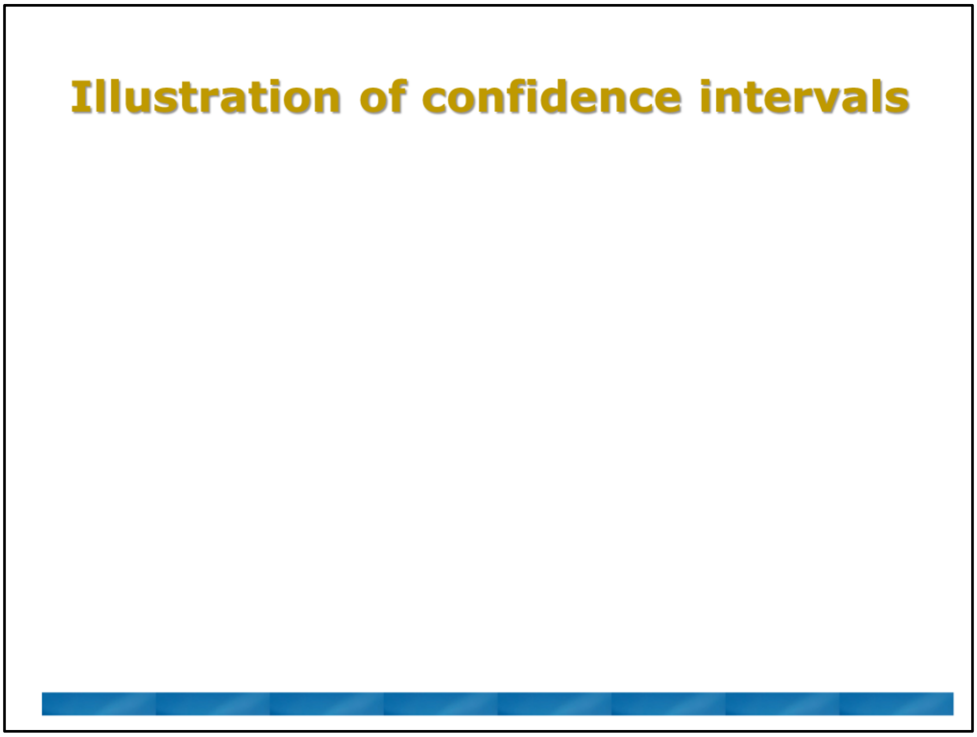


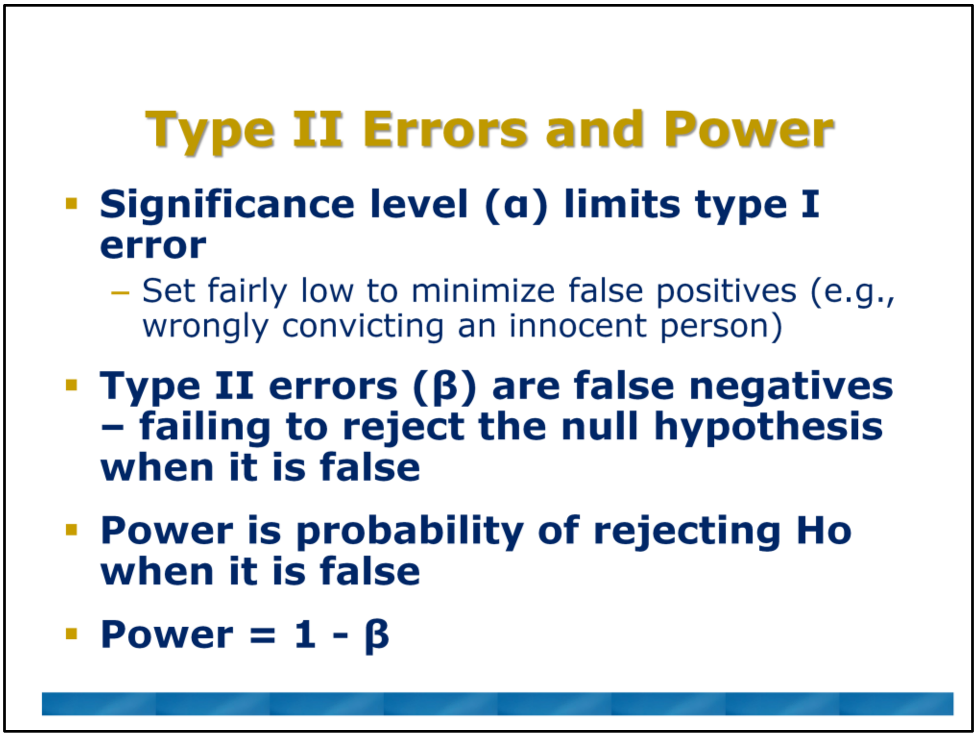
Illustration of confidence intervals

Use Test program.

Let’s assume that we are interested in whether the means of two groups are different. Let’s assume that the means of two groups do in fact differ by 0.6 considering the whole population and that the standard deviations are 1.

Now let’s draw repeated samples from these groups, and estimate the mean difference and 95% CI.

There are a couple things I want to point out here. First, notice all the red intervals. These are 95% from a few samples from our population and they don’t cover the true difference of 0.6. On average 5% of the CI will not contain the true value. Think about this. I think we all have a tendency to think our point estimate is right on the money or at least close and we usually give little regard to the Cis but these simulations show that 5% of the time, the confidence interval isn’t even going cover the true value. Second, look at all the confidence intervals that cover 0. For those samples, we would not have concluded that the groups had different means even though they do. We would have failed to reject the Ho when we should have. This brings us to Type II errors and power.

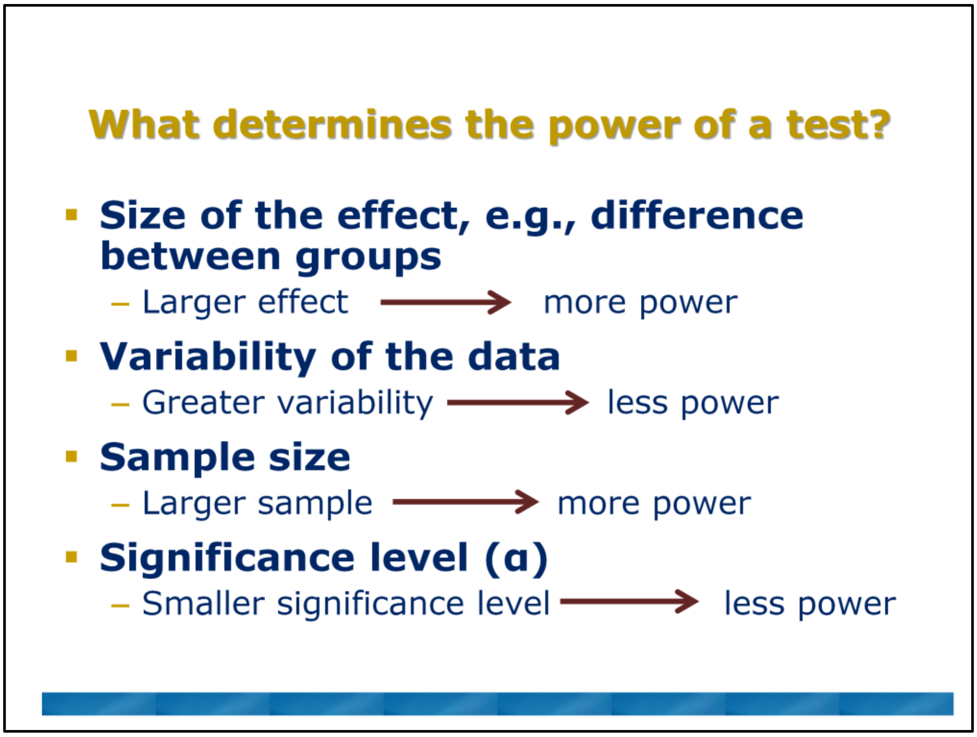


Type II Errors and Power

- **Significance level (α) limits type I error**
  - Set fairly low to minimize false positives (e.g., wrongly convicting an innocent person)
- **Type II errors (β) are false negatives – failing to reject the null hypothesis when it is false**
- **Power is probability of rejecting Ho when it is false**
- **Power = 1 - β**

Up until this point we have been talking mostly about type I errors – the probability of rejecting the null when it is true. We have wanted to minimize making this type of error and so we have set our significance level (alpha) low.

The other type of error is failing to reject the null when it is false. This is the type II error. We want to minimize this as well. We usually talk about Power which is 1 – the probability of a type II error and is defined as the probability of rejecting Ho when it is false. Clearly, we want power to be high.



What determines the power of a test?

- **Size of the effect, e.g., difference between groups**
  - Larger effect → more power
- **Variability of the data**
  - Greater variability → less power
- **Sample size**
  - Larger sample → more power
- **Significance level (α)**
  - Smaller significance level → less power

The power of a statistical test depends on 4 things – the size of the effect, the variability of the data, the sample size and the significance level.

The size of the effect is how different are the groups. For sample size calculations this is why we ask you how a large a difference you want to detect. The larger the difference or the effect the more power a procedure has. Intuitively this says that it is easier to spot big differences than small differences.

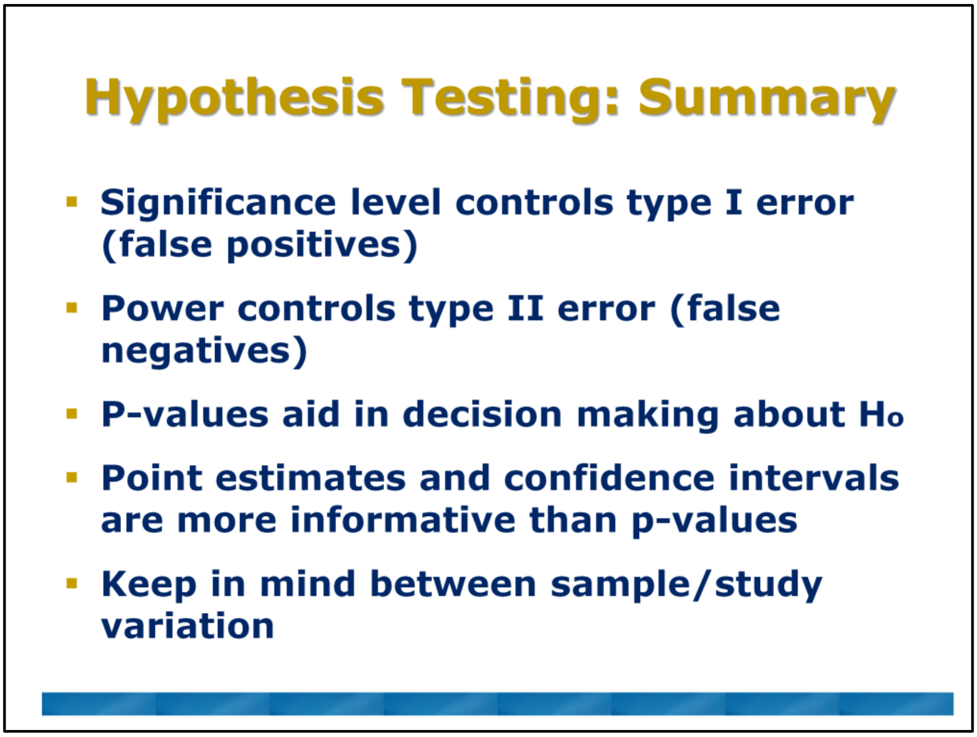
How variable is the data? As we have seen, our estimates vary with our sample. The more variable the data/population is the more variable our estimates will be between samples and the more difficult it will be to determine if two groups are different. Hence, more variability leads to less power.

Sample size. The larger the sample the more power we have to detect differences. As we get a larger and larger sample, we get more and more certain about what the true values which gives us increasing power to distinguish groups.

Finally, the significance level we set impacts power. Suppose we set alpha at 1.0, and rejected the null under all circumstances? Then our power would be 100% because we would ALWAYS reject the null when it was false but we would also reject the null when it was true and hence our type I error would be 100% as well.

Clearly we strike a balance here.

I also want to emphasize/point out that power and sample size calculations are part science and part art. It all “depends” on the assumptions on the assumptions you make in the calculations.



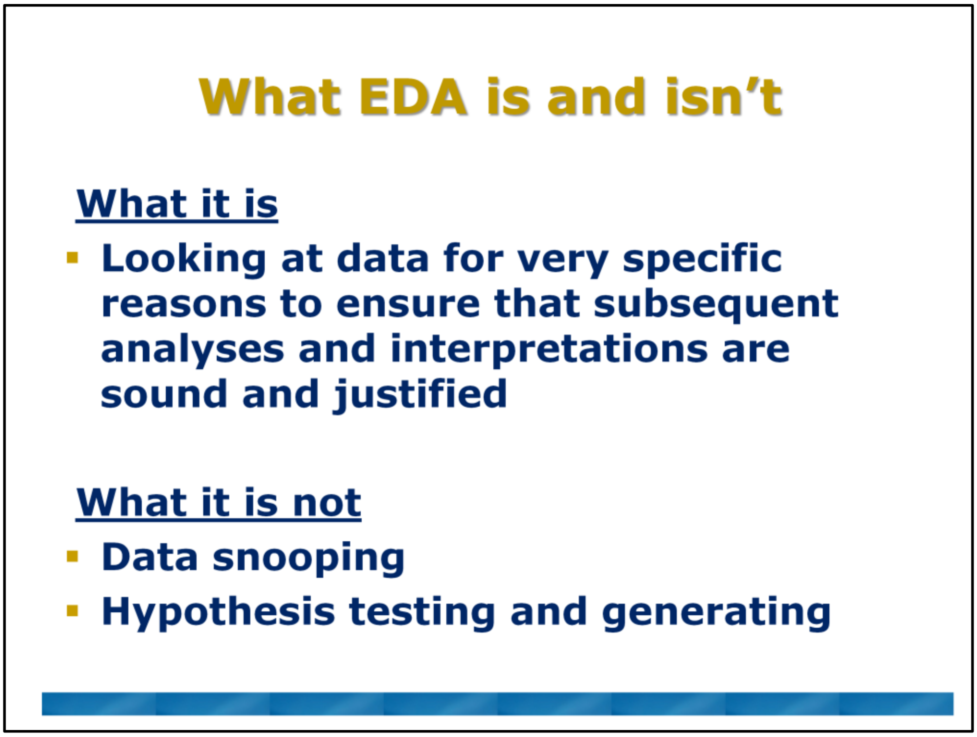
Hypothesis Testing: Summary

- Significance level controls type I error (false positives)
- Power controls type II error (false negatives)
- P-values aid in decision making about H_o
- Point estimates and confidence intervals are more informative than p-values
- Keep in mind between sample/study variation

This brings us to the close of the first portion of this talk which focused on the basics of hypothesis testing. In summary,

Exploratory Data Analysis

Now I want to shift gears a bit and talk about exploratory data analysis.



What EDA is and isn’t

What it is

- Looking at data for very specific reasons to ensure that subsequent analyses and interpretations are sound and justified

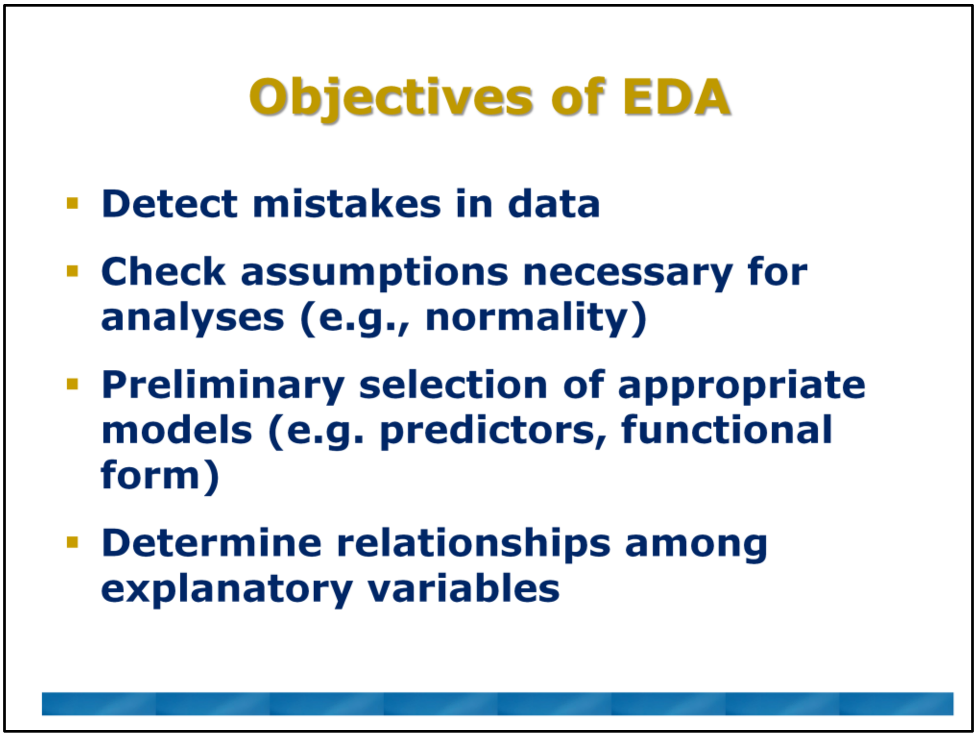
What it is not

- Data snooping
- Hypothesis testing and generating

Exploratory data analysis (aka EDA) is looking at your data for very specific reasons prior to formal statistical tests to ensure that subsequent analyses and interpretations are sound and justified.

You most definitely should look at your data before jumping into statistical analyses, BUT

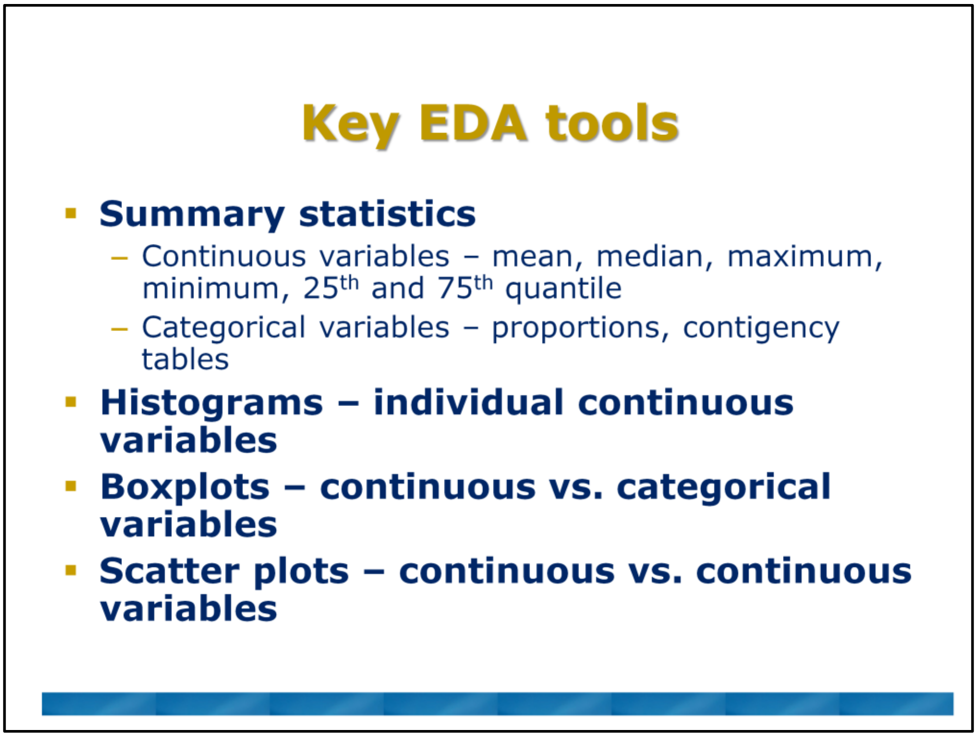
It is not data snooping or hypothesis testing and generating. It is not “let’s just see what we can find”



Objectives of EDA

- Detect mistakes in data
- Check assumptions necessary for analyses (e.g., normality)
- Preliminary selection of appropriate models (e.g. predictors, functional form)
- Determine relationships among explanatory variables

We do EDA for some specific reasons, including …



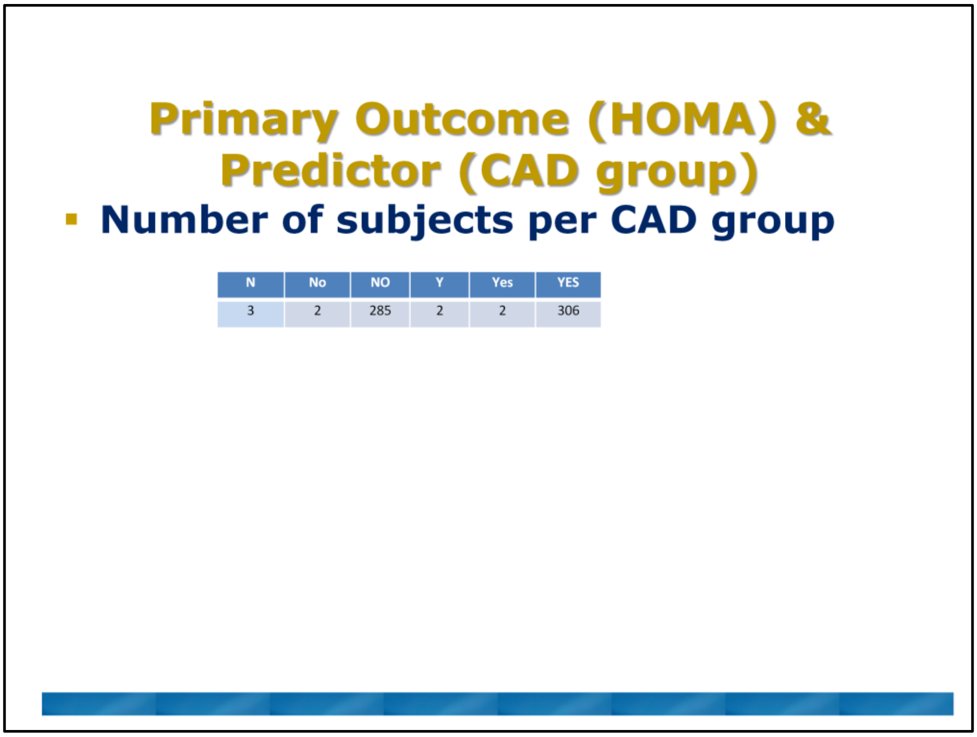
Key EDA tools

- **Summary statistics**
  - Continuous variables – mean, median, maximum, minimum, 25th and 75th quantile
  - Categorical variables – proportions, contigency tables
- **Histograms – individual continuous variables**
- **Boxplots – continuous vs. categorical variables**
- **Scatter plots – continuous vs. continuous variables**

The tools used for EDA are not very complicated or sophisticated. You are probably familiar with most if not all of them but may never have thought about using these tools as I suggest to do so here.

These tools are.

Now I am going to run through where to use these, how to use them and what to look for in analyzing your data.



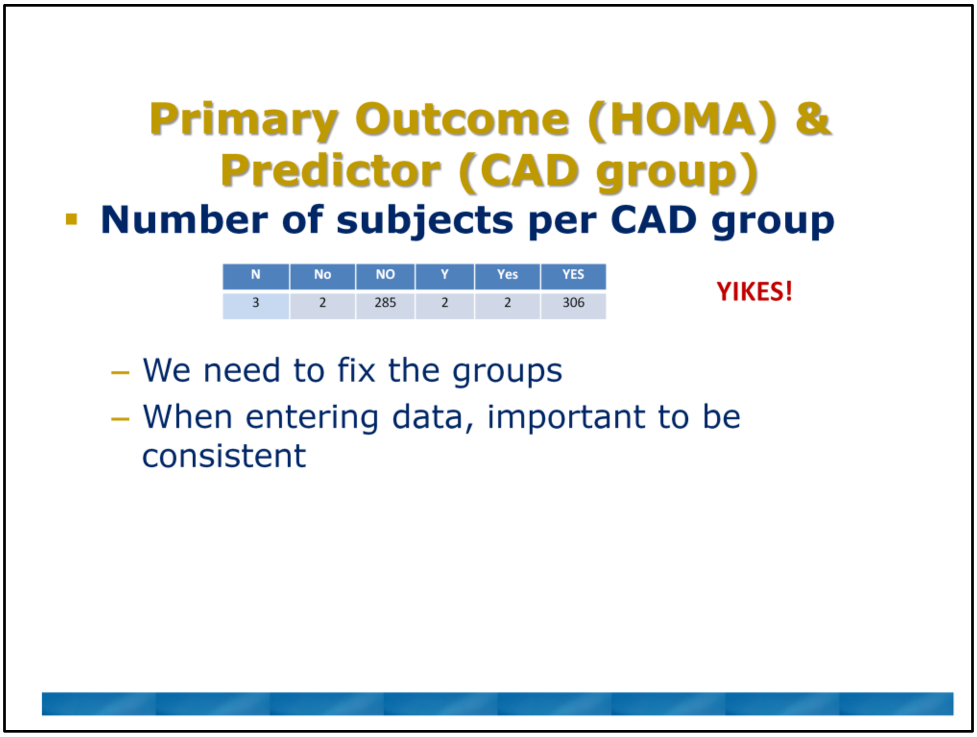
Primary Outcome (HOMA) & Predictor (CAD group)

- Number of subjects per CAD group

N	No	NO	Y	Yes	YES
3	2	285	2	2	306

Let’s start with our primary predictor – CAD group and primary outcome – HOMA levels.
First thing to do for CAD groups is tabluate the number in each group.
Suppose you got something that looked like this.

This happens ALL THE TIME



Primary Outcome (HOMA) & Predictor (CAD group)

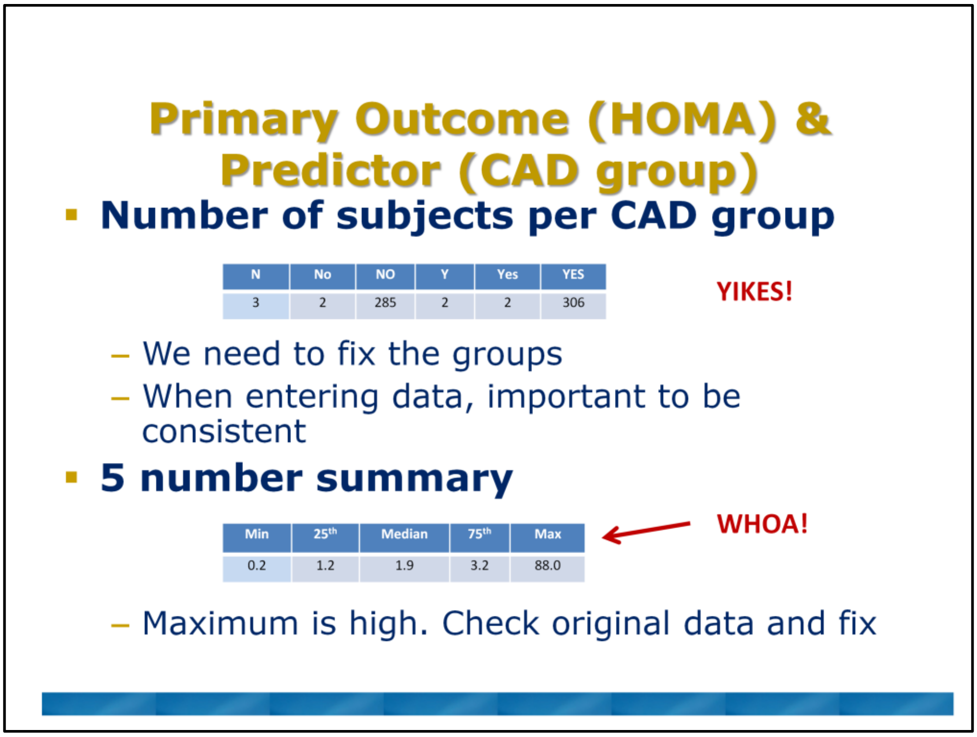
- **Number of subjects per CAD group**

N	No	NO	Y	Yes	YES
3	2	285	2	2	306

YIKES!

  - We need to fix the groups
  - When entering data, important to be consistent

So – we need to fix the groups. This also brings up the importance of being consistent when entering data.



Primary Outcome (HOMA) & Predictor (CAD group)

- Number of subjects per CAD group

N	No	NO	Y	Yes	YES
3	2	285	2	2	306

YIKES!

  - We need to fix the groups
  - When entering data, important to be consistent

- 5 number summary

Min	25th	Median	75th	Max
0.2	1.2	1.9	3.2	88.0

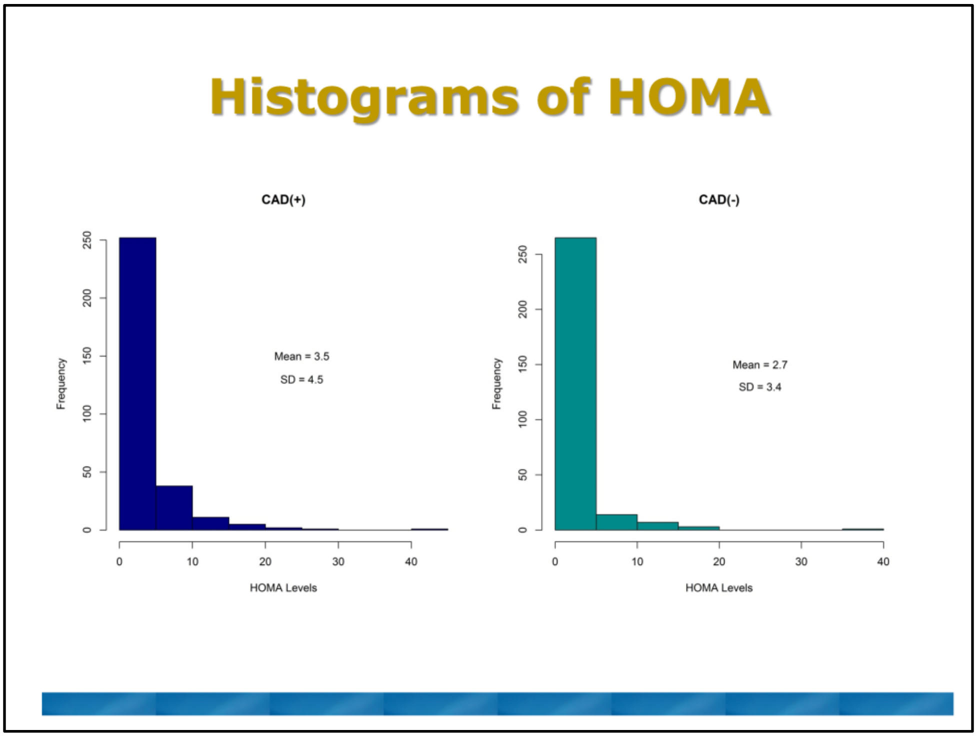
WHOA!

  - Maximum is high. Check original data and fix

Now let’s look at HOMA.

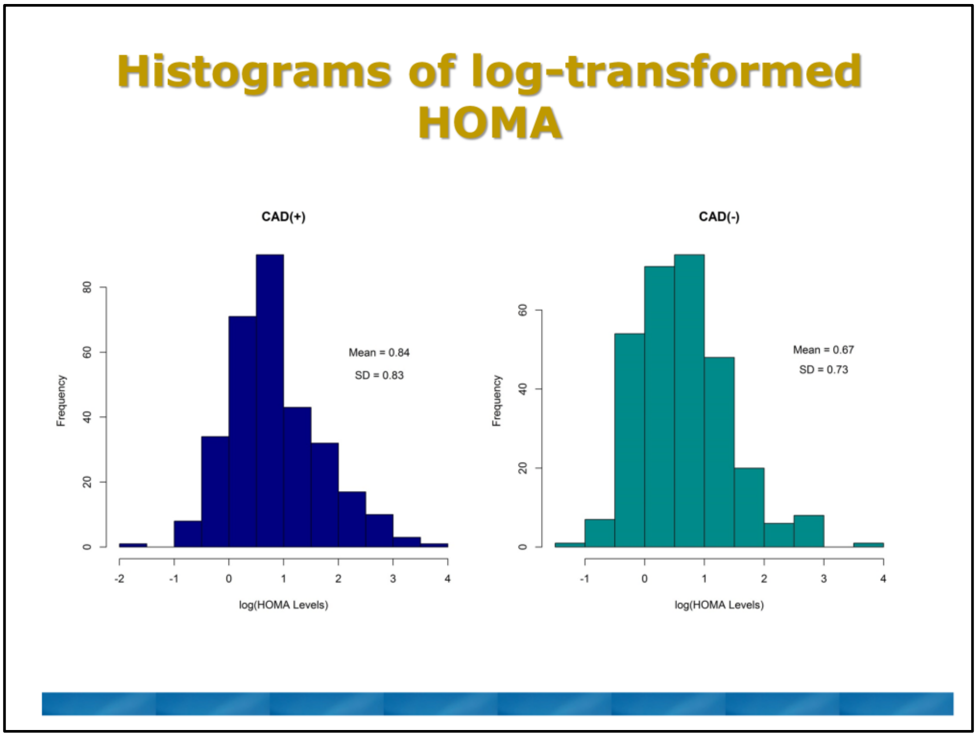
We can do a 5-number summary consisting of the min, max, median, 25th and 75th quantiles for a quick check.

Here we see the max is 88 which is super high. Hopefully we can go back to the original data, check the value and correct if wrong.



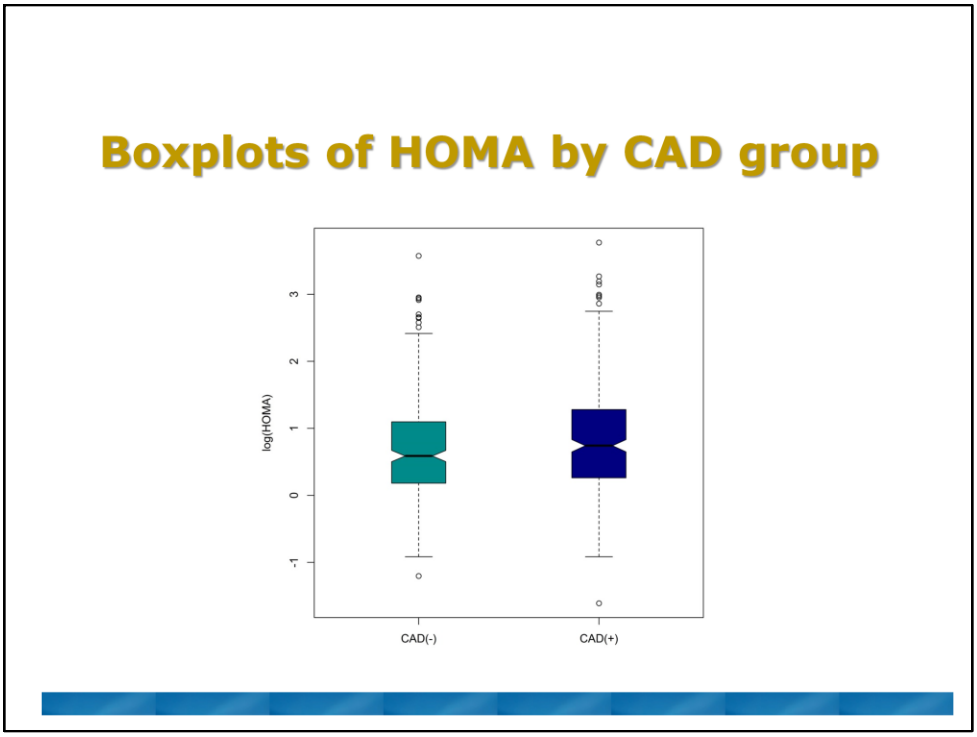
Histograms of HOMA

Histograms are a great way to get a quick look at the distributions of continuous variables. We have seen this before. Here we note that the distributions have similar shapes. That’s good but we note that they are strongly right skewed which would violate normality assumptions. So we try log transforming the data.



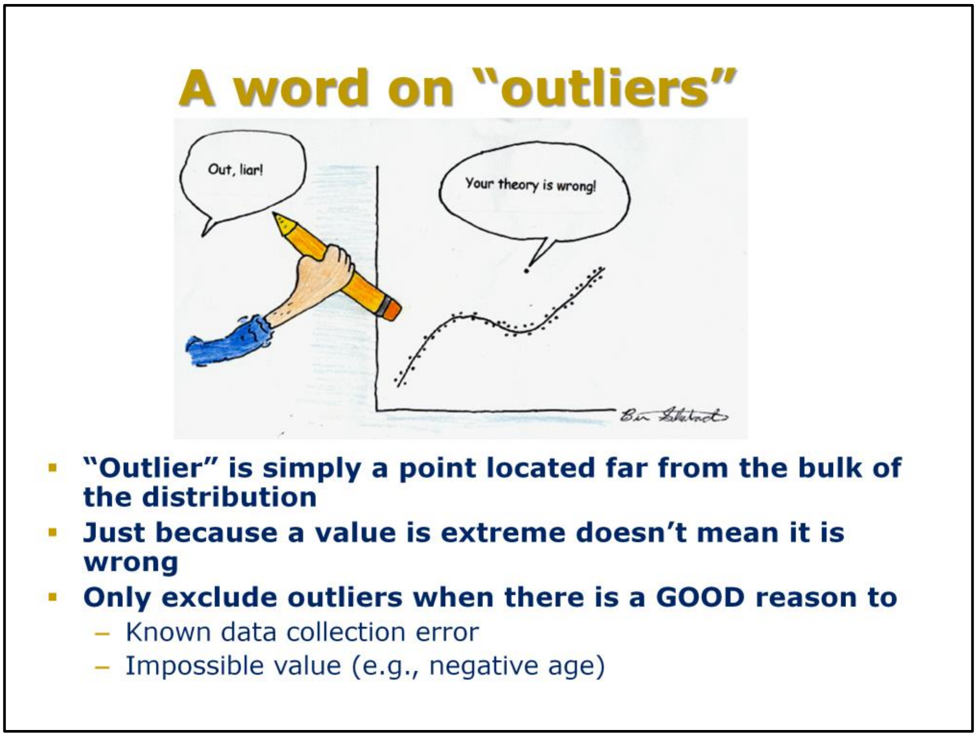
Histograms of log-transformed HOMA

That looks better and based on these then I log-transformed the data for analysis.



Boxplots of HOMA by CAD group

Boxplots are another great type of plot. In case you are not familiar with boxplots, the pinched piece in the middle is the median and the edges of the box are the 25th and 75th. The limits of the whiskers can vary with the program. Here they are the quartile +/- 1.5*IQR. The circles are anything beyond that sometimes referred to as outliers. These don't show the distribution of the data quite as clearly as the histograms do but they are much better for seeing group differences.



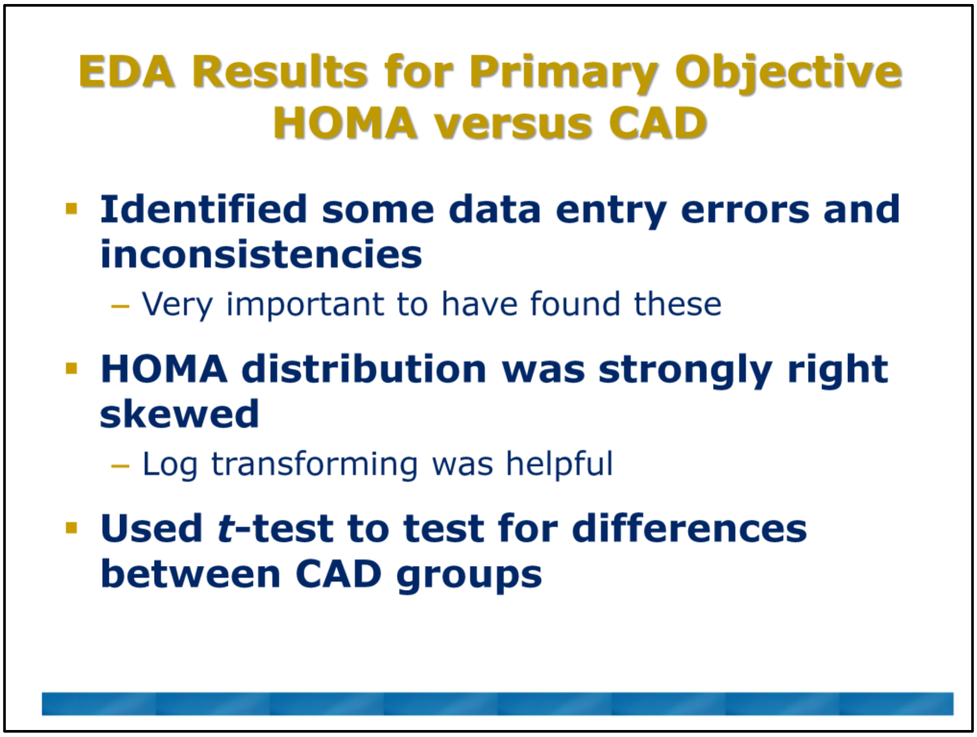
A word on "outliers"

Out, liar!
Your theory is wrong!

- "Outlier" is simply a point located far from the bulk of the distribution
- Just because a value is extreme doesn't mean it is wrong
- Only exclude outliers when there is a GOOD reason to
  - Known data collection error
  - Impossible value (e.g., negative age)

I mentioned that the points on the boxplot graphs are often called "outliers".

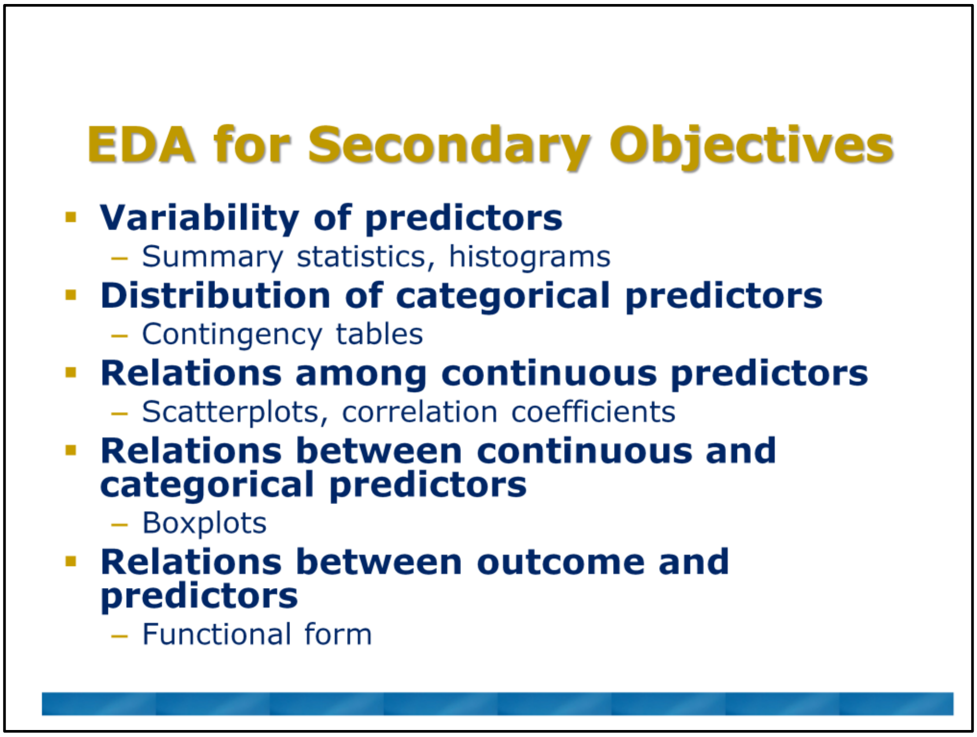
Statisticians frown on excluding "outliers". Only exclude outliers when there is a GOOD reason to do so.



EDA Results for Primary Objective HOMA versus CAD

- **Identified some data entry errors and inconsistencies**
  - Very important to have found these
- **HOMA distribution was strongly right skewed**
  - Log transforming was helpful
- **Used *t*-test to test for differences between CAD groups**

We have finished a quick EDA for our primary objective. As a result we…



EDA for Secondary Objectives

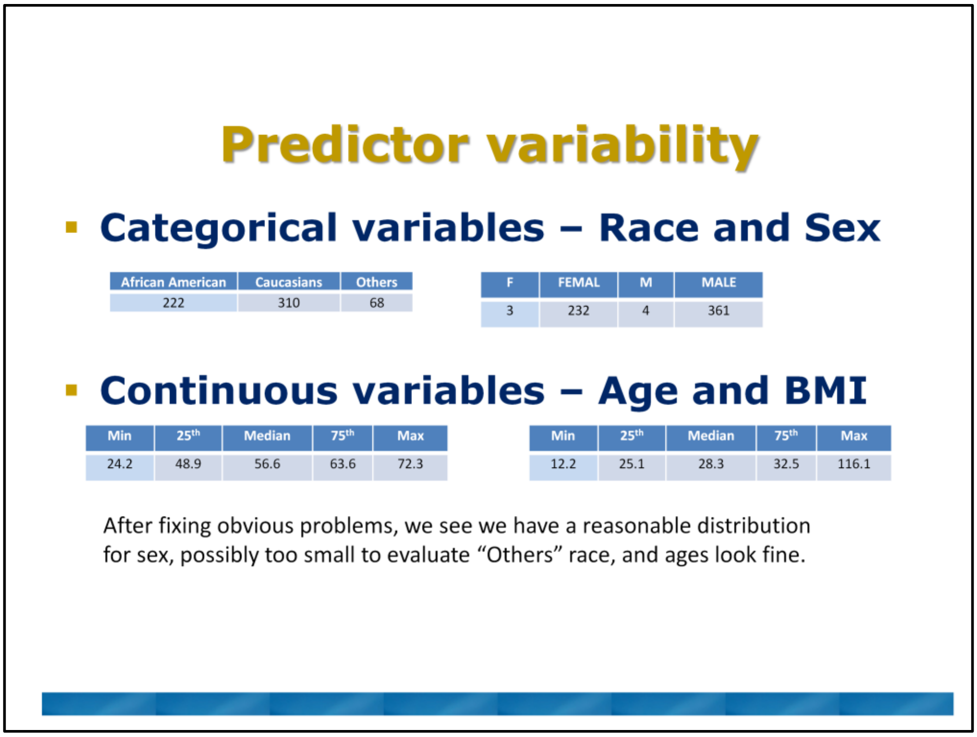
- **Variability of predictors**
  - Summary statistics, histograms
- **Distribution of categorical predictors**
  - Contingency tables
- **Relations among continuous predictors**
  - Scatterplots, correlation coefficients
- **Relations between continuous and categorical predictors**
  - Boxplots
- **Relations between outcome and predictors**
  - Functional form

Let’s turn now to EDA for the secondary objectives. Our secondary objective is to investigate the relationships between HOMA and age, race, sex, and BMI in addition to CAD status.

So, we are looking at these other predictors as confounders or effect modifiers which requires a multivariate analysis.
Before diving into a fitting multivariate models it is beneficial to look at the inter-relations among the predictors as well univariate relations between the outcome and predictors. This will help us interpret our results and build appropriate models. It also accomplishes data checking.

We should look at the variability of the predictors through summary statistics and histograms.
We then need to look at relationships among the predictors.
For categorical predictors, we can do this through contingency tables. For continuous predictors we can do this through scatter plots, and correlation coefficients. And boxplots are useful for continuous versus categorical predictors.

Lastly, we want to do some univariate analyses of looking at the relationships between the outcome and the predictors.



Predictor variability

- **Categorical variables – Race and Sex**

African American	Caucasians	Others
222	310	68

F	FEMAL	M	MALE
3	232	4	361

- **Continuous variables – Age and BMI**

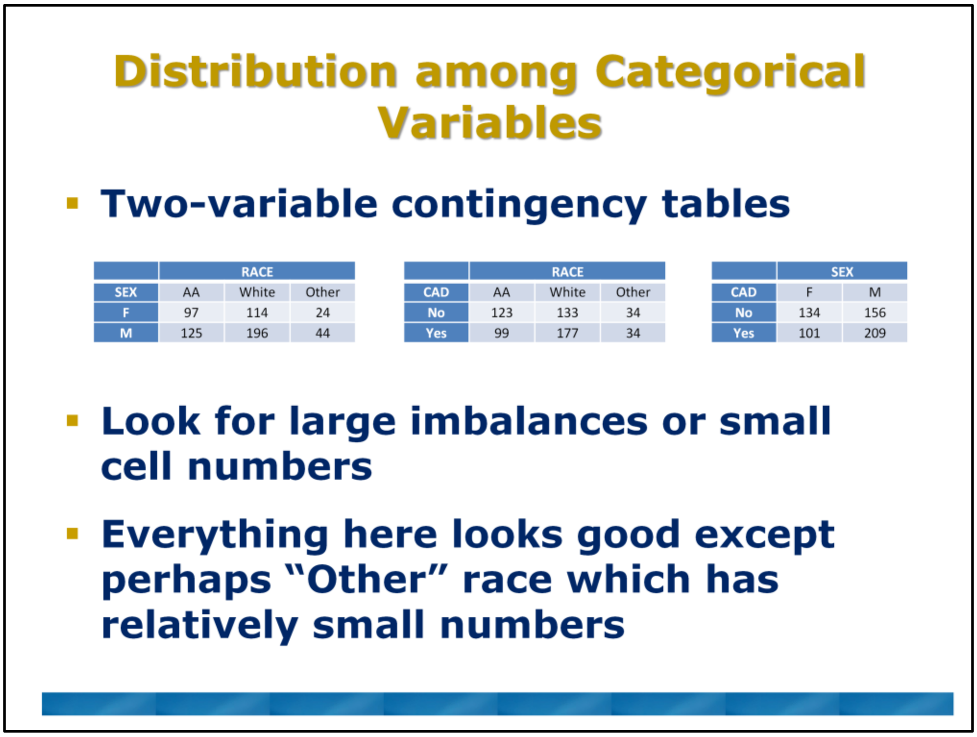
Min	25th	Median	75th	Max
24.2	48.9	56.6	63.6	72.3

Min	25th	Median	75th	Max
12.2	25.1	28.3	32.5	116.1

After fixing obvious problems, we see we have a reasonable distribution for sex, possibly too small to evaluate “Others” race, and ages look fine.

Predictor variability – we have 2 categorical variables and 2 continuous. For the categorical variables, we notice some coding discrepancies for sex and so we need to fix that. Race we notice relatively small numbers in the “Others” category so we might be limited on analyzing this group.

For continuous variables, we see a woefully excessive BMI value and fix that. Age – big range. No obvious issue data quality issues.



Distribution among Categorical Variables

- Two-variable contingency tables

	RACE		
SEX	AA	White	Other
F	97	114	24
M	125	196	44

	RACE		
CAD	AA	White	Other
No	123	133	34
Yes	99	177	34

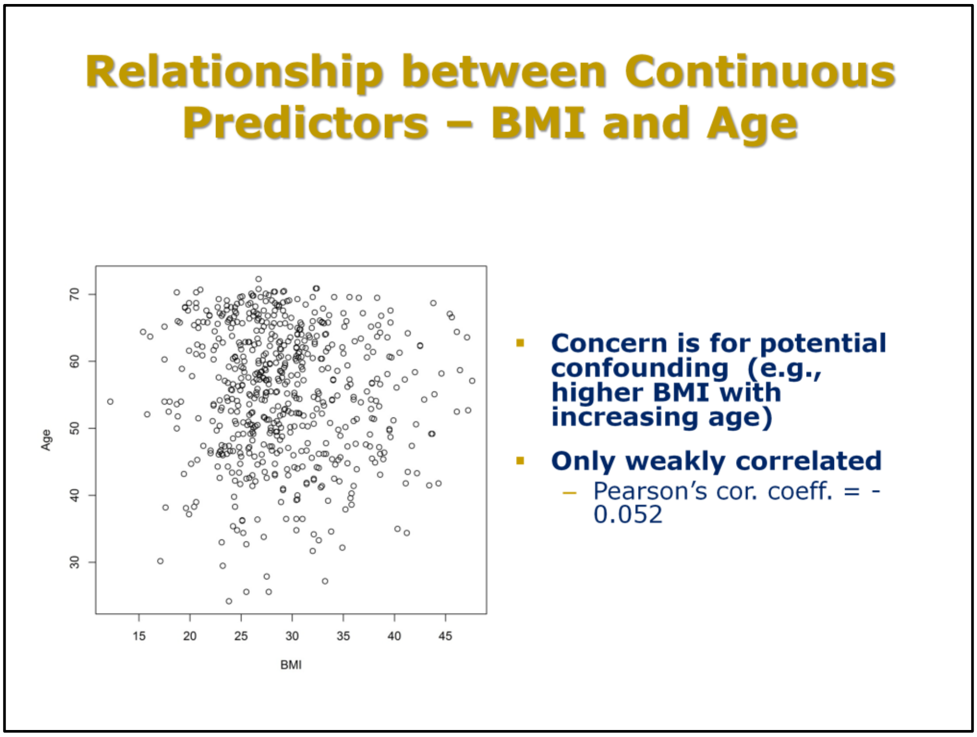
	SEX	
CAD	F	M
No	134	156
Yes	101	209

- Look for large imbalances or small cell numbers
- Everything here looks good except perhaps "Other" race which has relatively small numbers

For categorical variables, you should generate two-variable contingency tables.

You are looking for large imbalances or small cell numbers. Small cell numbers impact the choice of statistical test or whether you can even evaluate some questions. Large imbalance could impact interpretation and your ability to test certain questions. For example, suppose we had 350 men with CAD instead of 209. That would leave very few men without CAD in our sample such that we probably couldn't test for an interaction effect between Sex and CAD. What if at the same time that most of the men in the sample had CAD that few of the women did. For example, if only 20 of the women had CAD. In that case it would be difficult to disentangle the effects of CAD from potential effects of sex on our outcome.

This data is a very nice data set and doesn't have these types of issues. Biggest potential concern is the relatively small numbers of Others race.

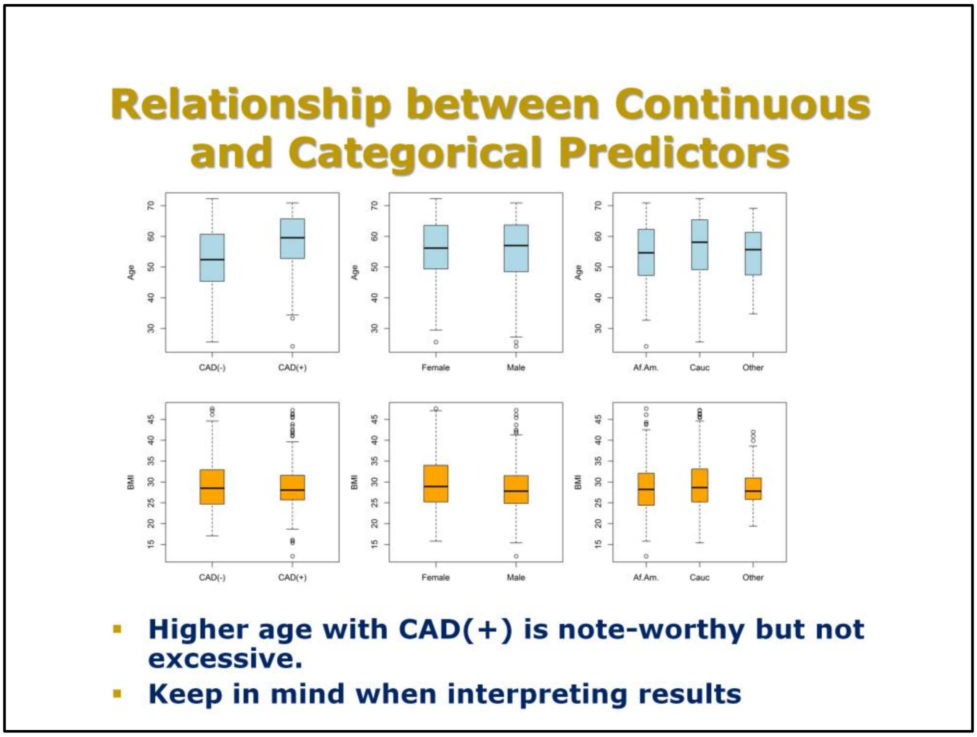


Relationship between Continuous Predictors – BMI and Age

- **Concern is for potential confounding (e.g., higher BMI with increasing age)**
- **Only weakly correlated**
  - Pearson's cor. coeff. = -0.052

How do the continuous predictors look? Simple scatter plots and correlation coefficients will do the trick. For age and BMI the scatter plot does not reveal any kind of relationship and the correlation coefficient is very low.

What we are looking for is the potential for these two to be related. If BMI and age were correlated, say positively related with BMI increasing with age, we would want to know this for later modeling. If very strongly correlated, best to include only one in the model. Even with weaker correlation though, it is important to understand these relationships as you build models and interpret results.

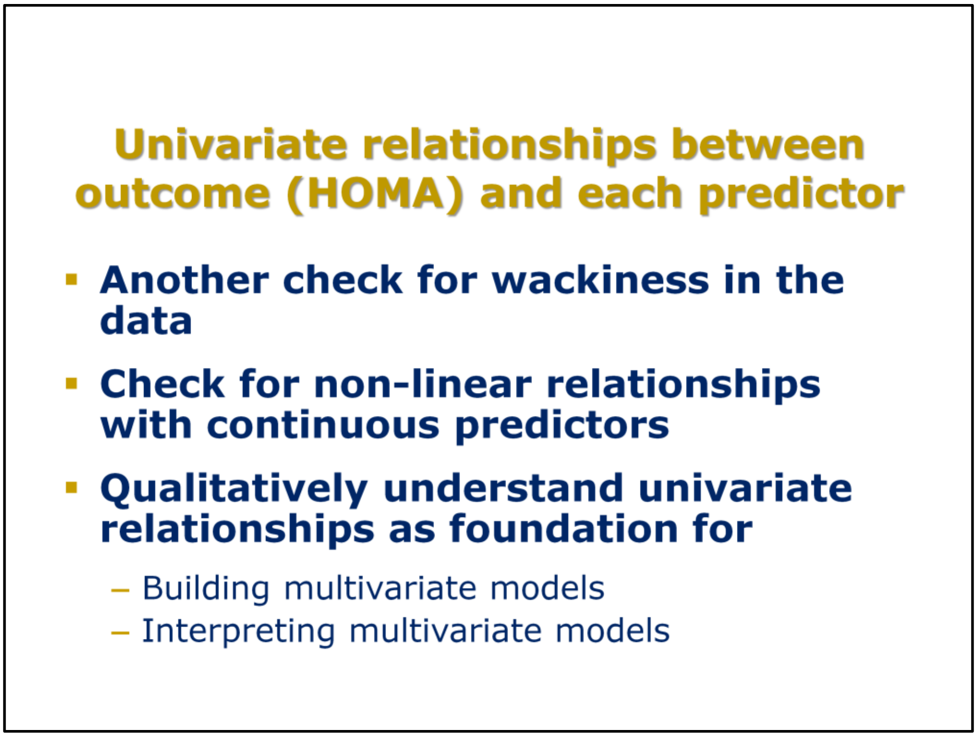


Relationship between Continuous and Categorical Predictors

- Higher age with CAD(+) is note-worthy but not excessive.
- Keep in mind when interpreting results

For the predictors, the last thing to look at are the continuous versus the categorical predictors. Again we are looking for big differences. Here, the higher age with CAD(+) patients is noteworthy and worth remembering but not something that would alter how I did the analysis. If there was a large difference with little overlap that could be a problem because then we would age and CAD status confounded and would have difficulty determining if differences in HOMA levels were from age or CAD status.

Yes – it is possible to adjust for age effects. That’s what we do with a multivariate model but if these is near complete separation, you can’t adjust for it.



Univariate relationships between outcome (HOMA) and each predictor

- Another check for wackiness in the data
- Check for non-linear relationships with continuous predictors
- Qualitatively understand univariate relationships as foundation for
  - Building multivariate models
  - Interpreting multivariate models

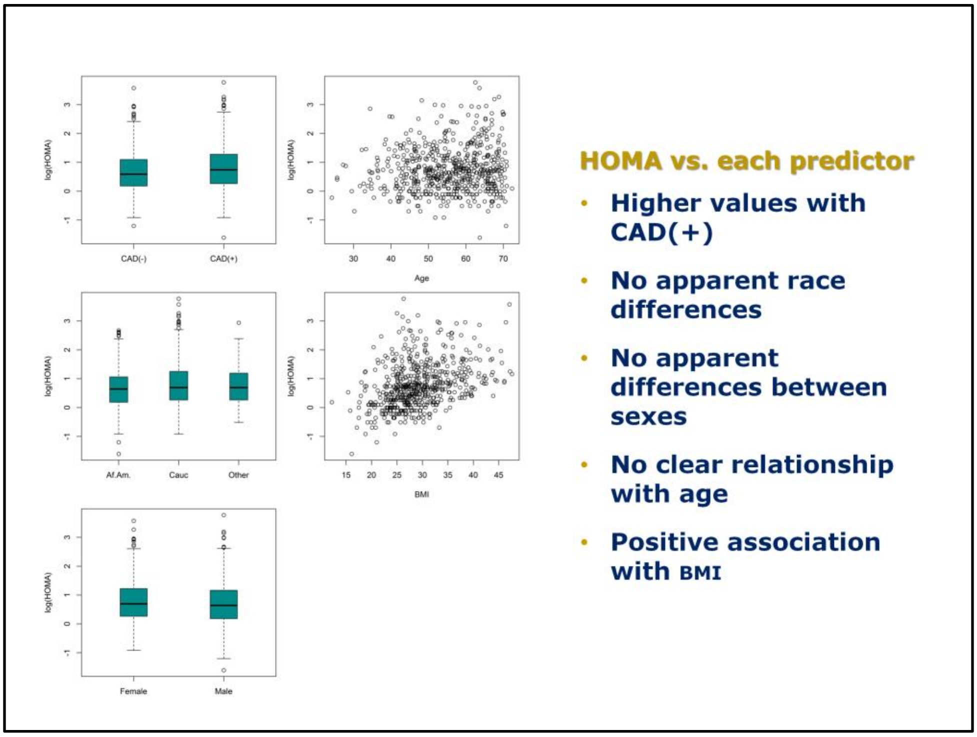
Now that we have gotten through the predictors, let’s look at HOMA versus each of the predictors.

This gives us, yet another check for wackiness in the data.

We also want to check for non-linear relationships with continuous predictors

Finally, these univariate looks, allow us to qualitatively understand what’s going on with each individual predictors which is important when we start building models and subsequently how we interpret models.

For example, in building models you don’t want to include two highly correlated predictors.



HOMA vs. each predictor

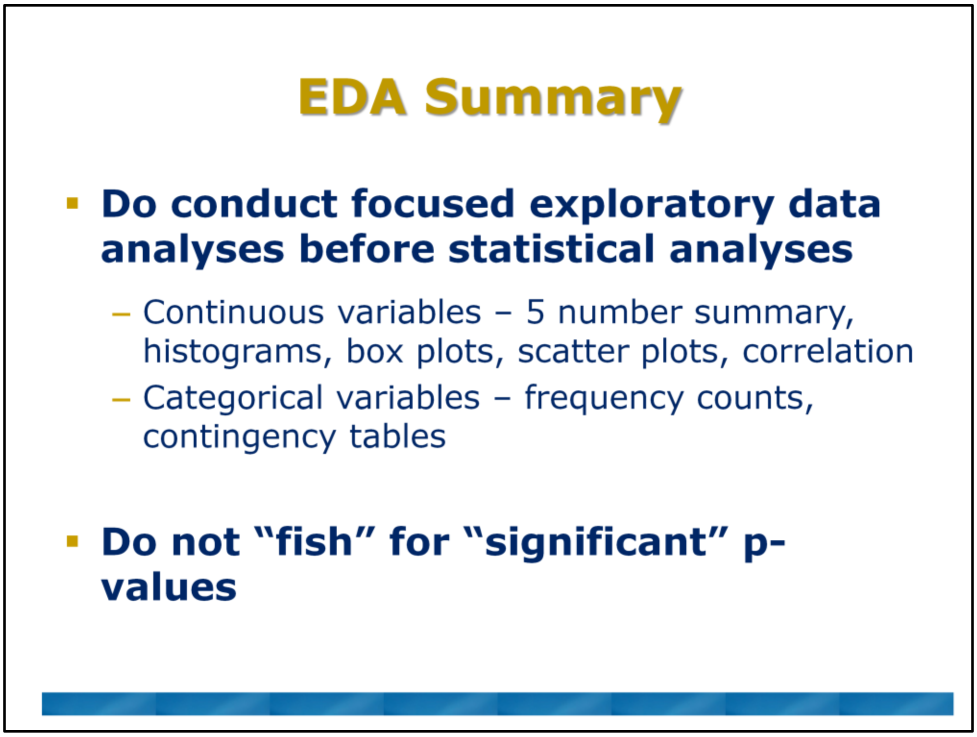
- Higher values with CAD(+)
- No apparent race differences
- No apparent differences between sexes
- No clear relationship with age
- Positive association with BMI

For continuous outcomes such as HOMA, we can use boxplots with categorical predictors and scatter plots for continuous predictors

Here we see higher values with CAD(+) as we knew but there doesn’t appear to be any race or sex effects.

Age doesn’t seem to have much of an effect if any but there does appear to be a positive association with BMI.

So, if I was modeling this, I would be inclined to put all these in the model but would only expect CAD and BMI to be significant predictors.



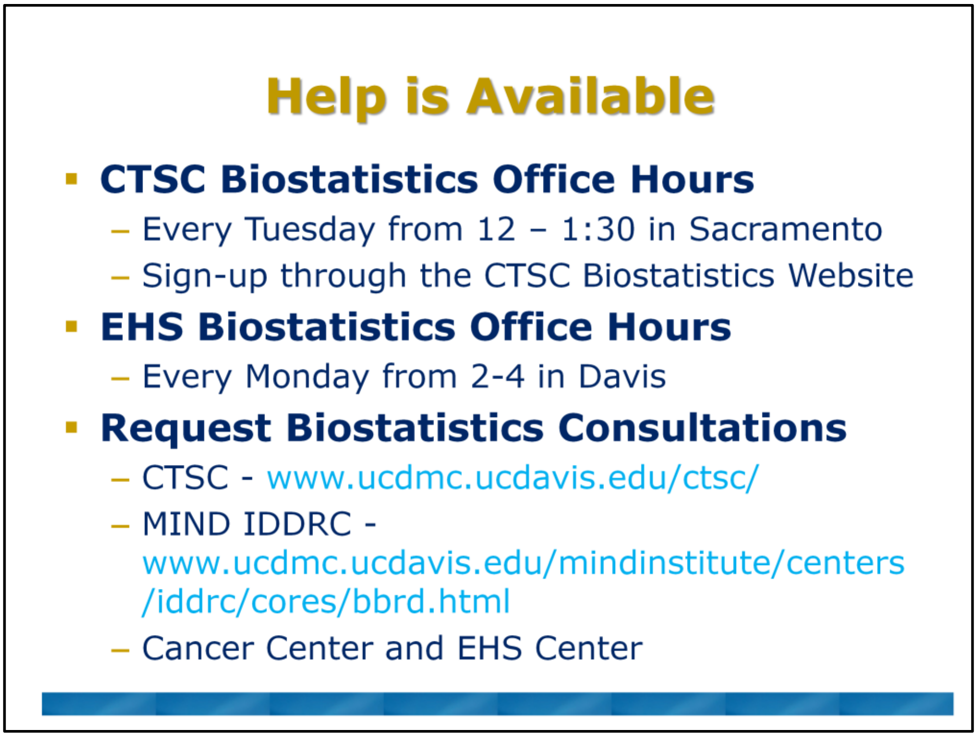
EDA Summary

- **Do conduct focused exploratory data analyses before statistical analyses**
  - Continuous variables – 5 number summary, histograms, box plots, scatter plots, correlation
  - Categorical variables – frequency counts, contingency tables
- **Do not "fish" for "significant" p-values**

In summary, it is good practice to conduct exploratory data analysis prior to conducting formal statistical tests. If you have a lot of predictors, I hope you can appreciate that EDA can be rather labor intensive.

EDA is really helpful for understanding your data and serving as the basis for interpreting your results.

That said, do not "fish" for significance.



Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30 in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4 in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center