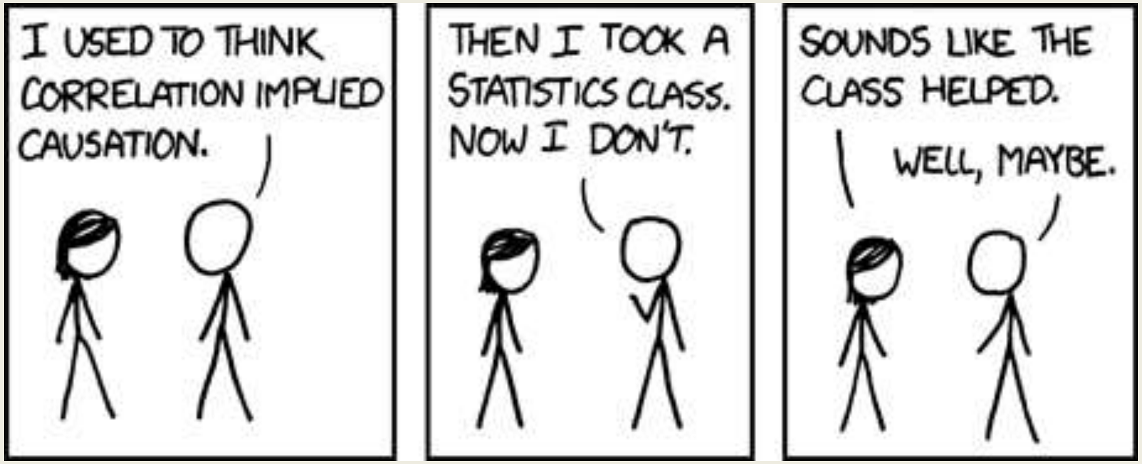
Inferential Statistics

Overview

- This is an experimental methods course, and as such you need to be familiar with how to understand methods and how to interpret the results of experimental studies.



Overview

- By the end of this unit you should be familiar with:
  - **t-tests**
  - **Alpha level**
  - **p-value**
  - **Chi-squared**

  - **Significance or Reliability**
  - **Type I and II Errors**
  - Signal:noise

What is a Statistic?

- A statistic is a one number summary of information about a distribution of a measure that is unbiased.

- Example 1: The percentage of women is 55%. N = 22 students.
- (Knowing the number of students we could use this to count how many men and how many women with reasonable accuracy. 22 * .55 = 12.1. People have to be whole numbers to we have to round this to an integer. There are 12 women and 10 men.

Example:

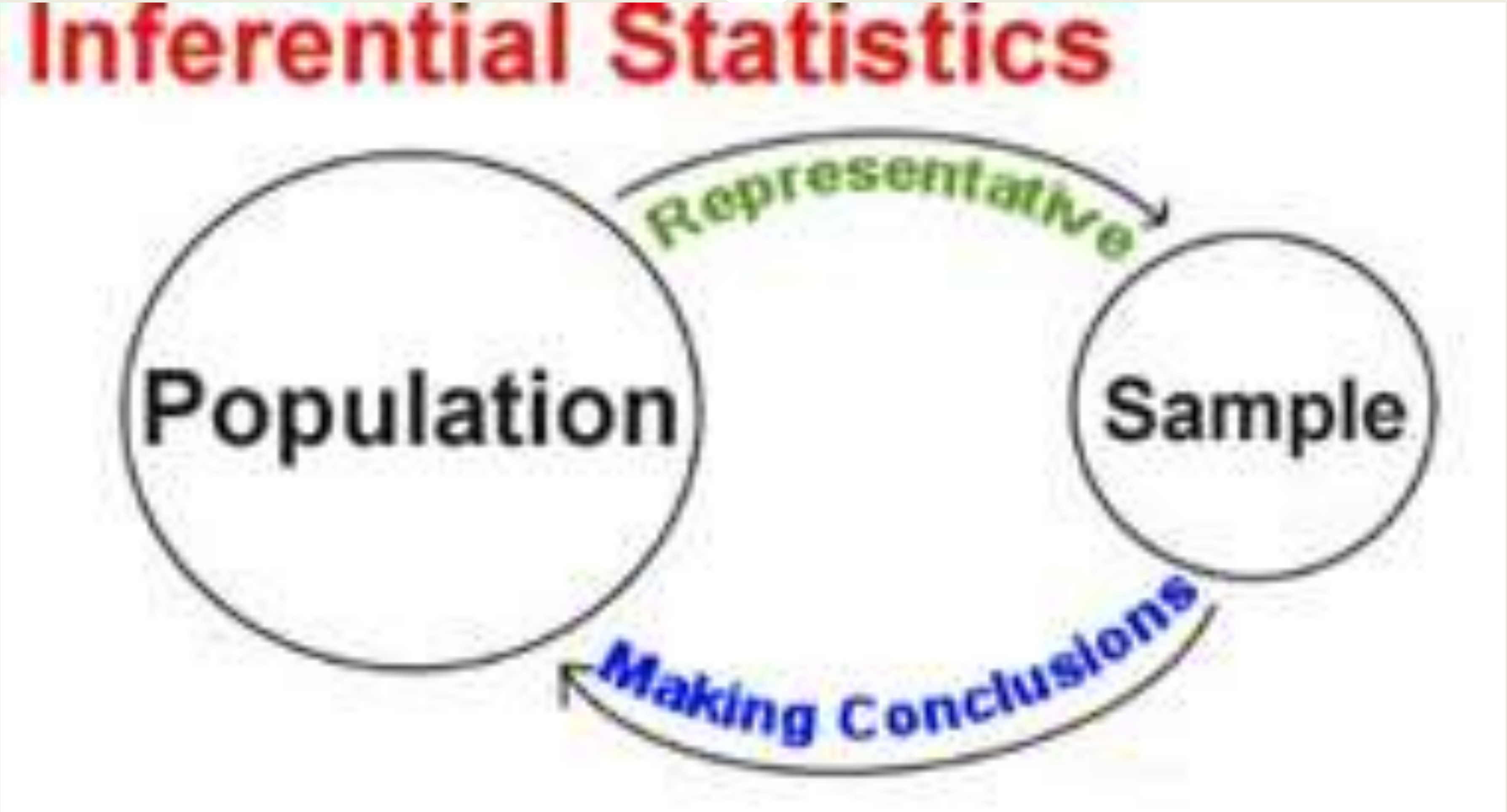
- The mean height in the class is 64 inches and the standard deviation is 5 inches. To compute these, I would need to use every single person's height and not just some people's, in order to be unbiased.

- Suppose I visually estimated the three tallest and three shortest people, and included only their heights to compute the mean and standard deviation. Which SD would be higher?

Inferential Statistics

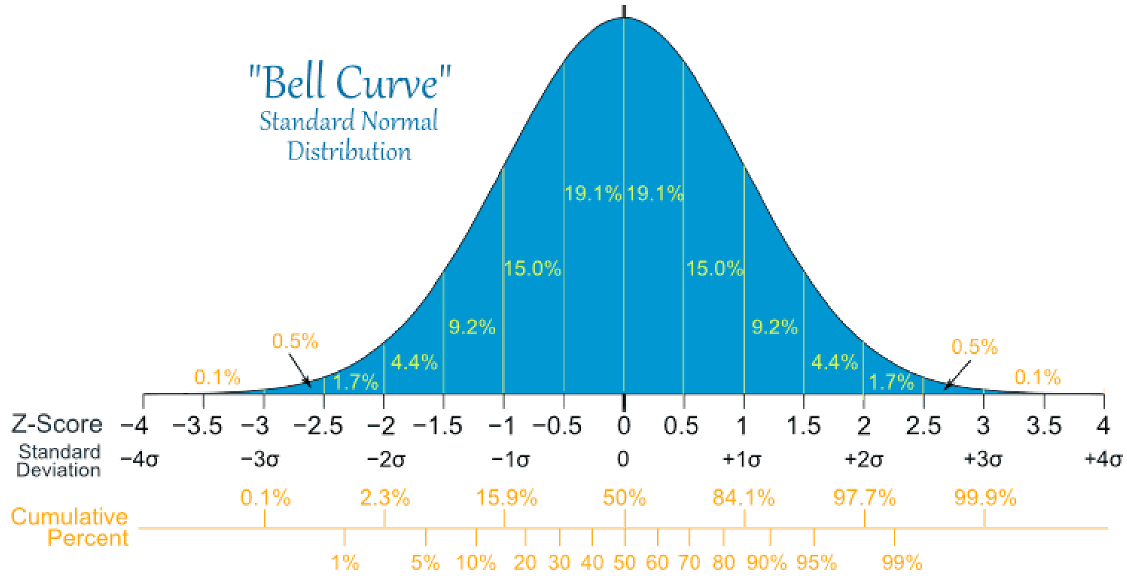
- Allow us to make probabilistic statements about whether hypotheses should be rejected.
- They can be used to test whether different sets of data *with the same measure* are essentially equivalent or different.
- They can also be used to test whether conditions in experiments produce different results *on the same measure*.
- There is no certainty in these inferences (100% confidence). Rather they give us a band of confidence and probability our inferences are wrong.

Population – Sample Comparison



Most statistical tests that assume that the dependent variable is normally distributed .

There are tests for that. (You don’t have to know about that).



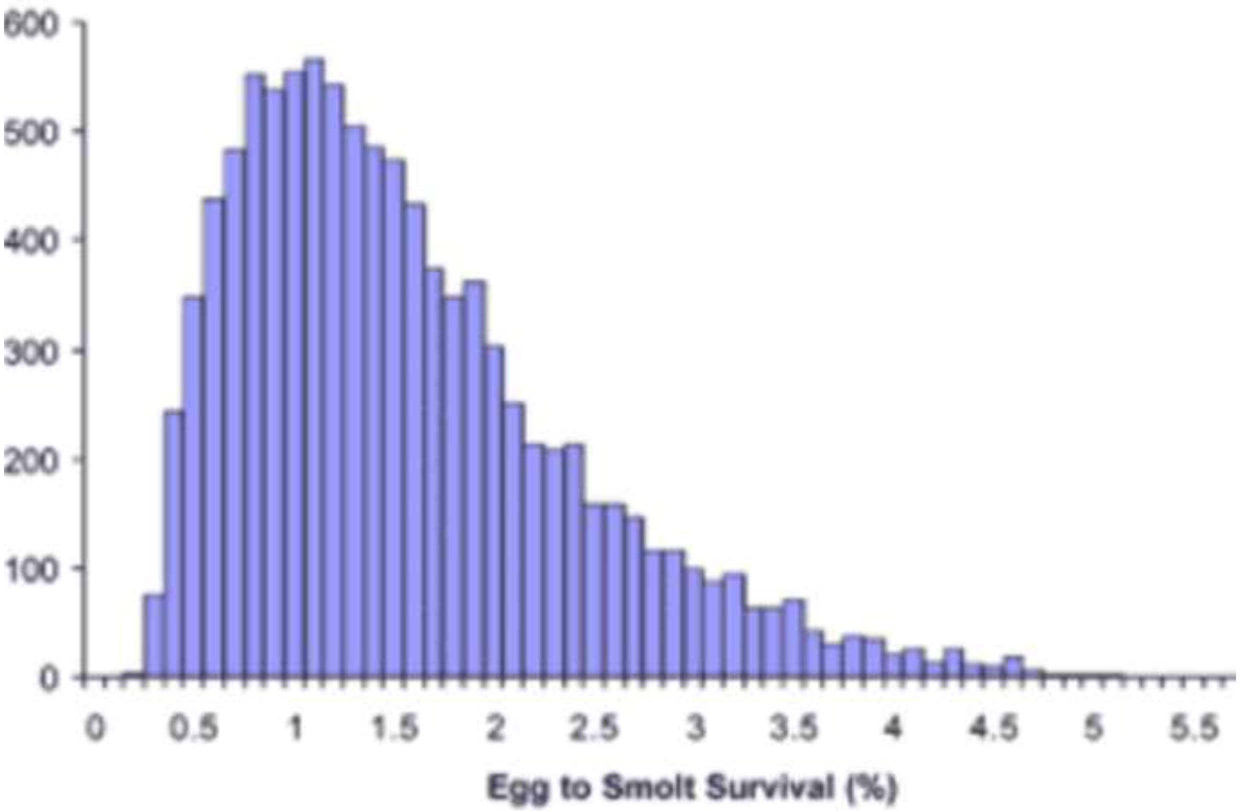
Normal or Gaussian Distribution

This is a frequency plot. It shows how many cases (y axis) had what value of survival (X axis).

There is only one variable here.

This is NOT a good example of a normal distribution because it is not very symmetric.

But it does have a central tendency (hump) like a normal distribution.



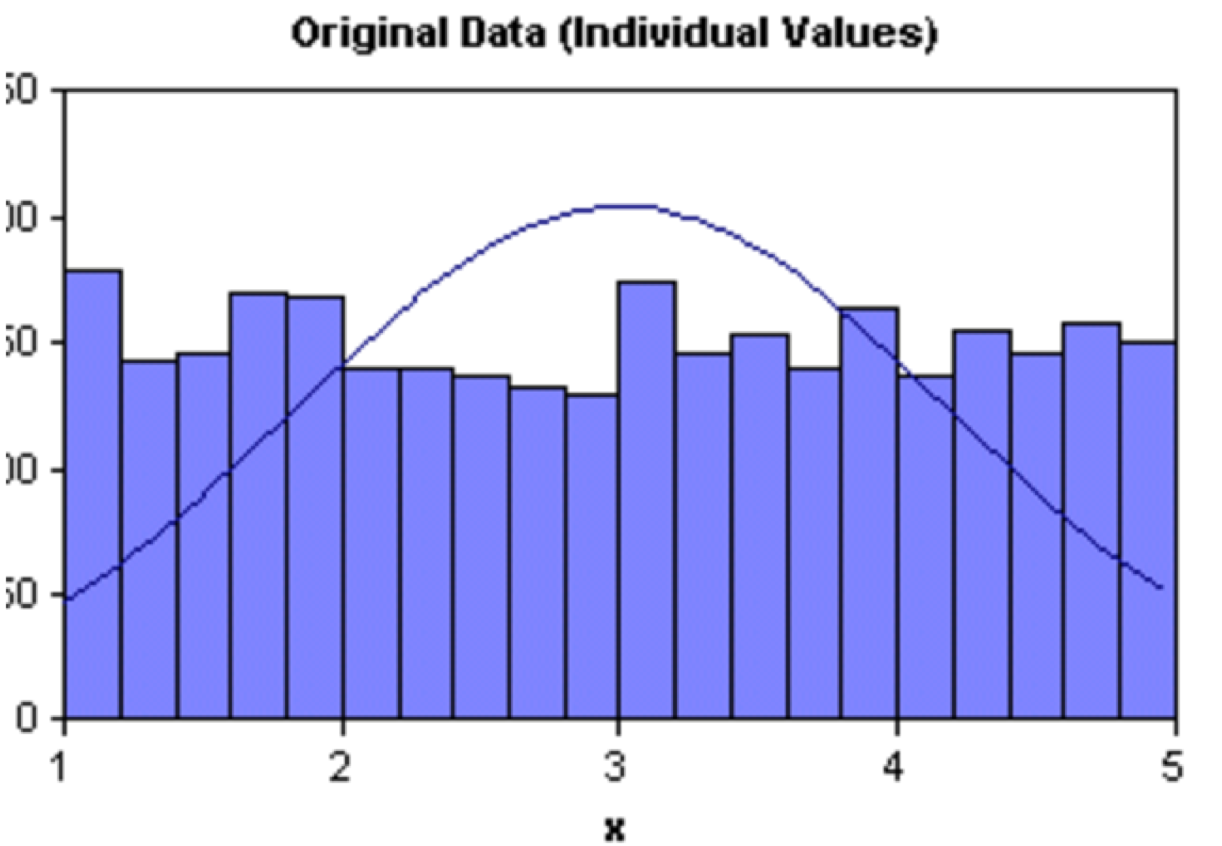
(Could be transformed to make more normal.)

This is also a poor example of a normal distribution.

There are approximately the same number of cases at each value of x.

There is no central tendency.

The standard deviation will be high because a median or mean does not describe much of the sample well at all.



No transformation will make this normal.

Why statistics need probabilities

- We have to figure out how seriously to take what the data we observe/record tell us about the state of the universe/
- i.e., how accurately do they inform us about that? How much confidence should we have in that information?
- So we compare observed data against hypothetical probability distributions.
- Conceptual example…

You roll two dice. We add up the sum. If you guess the number before the roll, you get $1.

What number would you bet?

How could you maximize the chance you would win?

Betting game

Tests whether data you collect appear to come from a known distribution or have a known value.

IOW, do the data fit a baseline expectancy?

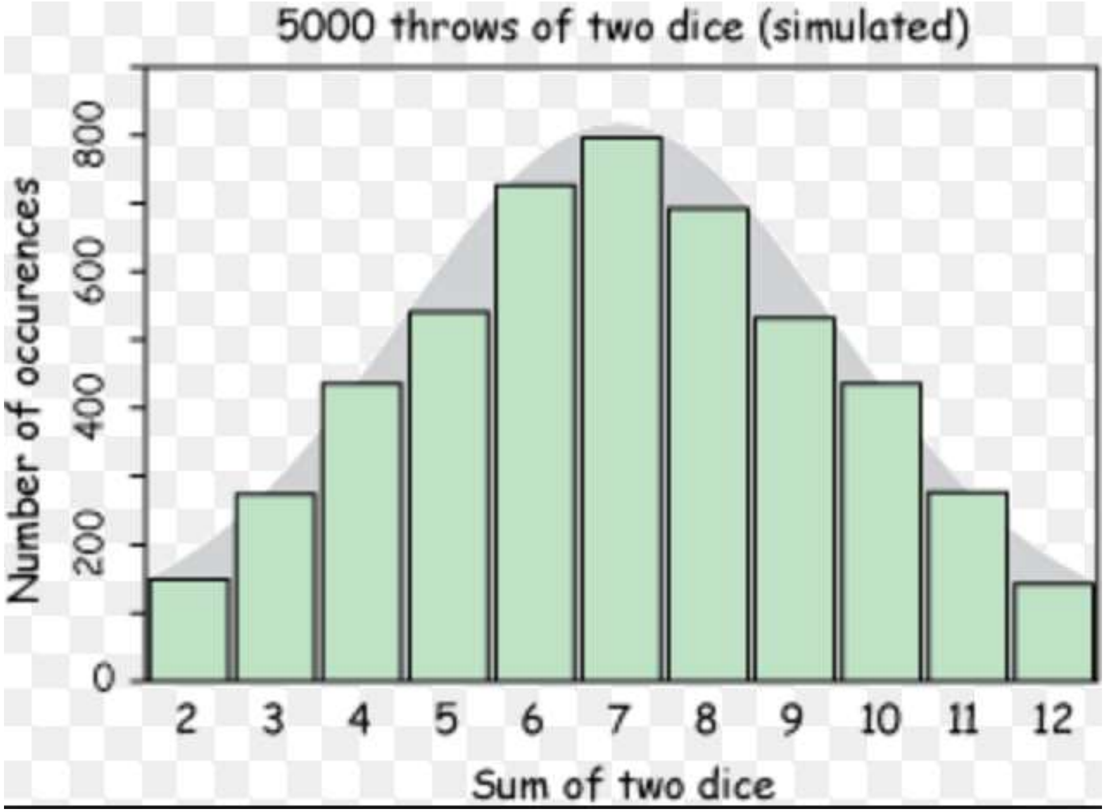
DV: Total of 2 dice

If total is 1 or under, 13 or more, MUST be measurement error.

Can observe here:

Central tendency is 7. *That would be your best bet.*

Lower chance of 4 than of 8, etc.

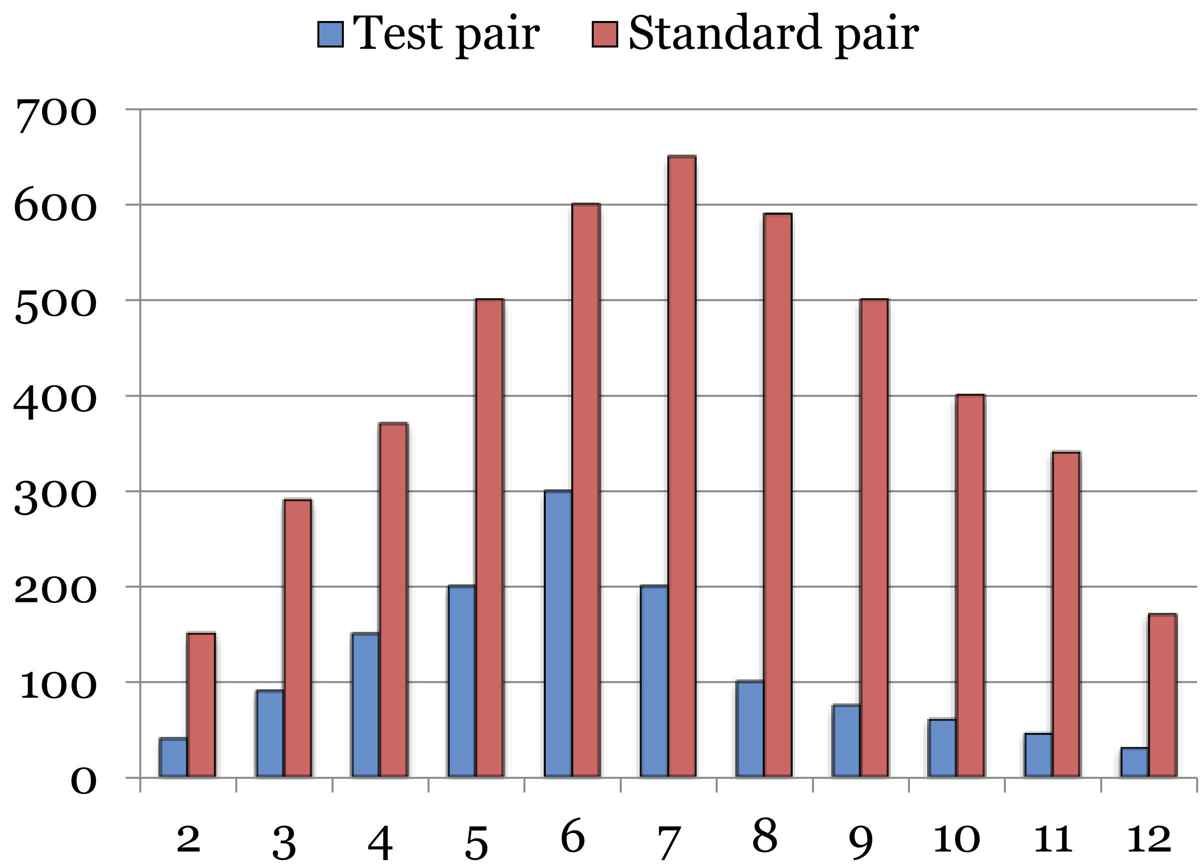


One-sample Student's t-test: Do the observed data appear to match expected or population distribution?

Suppose we did throw the dice 500 times and we got the blue distribution.

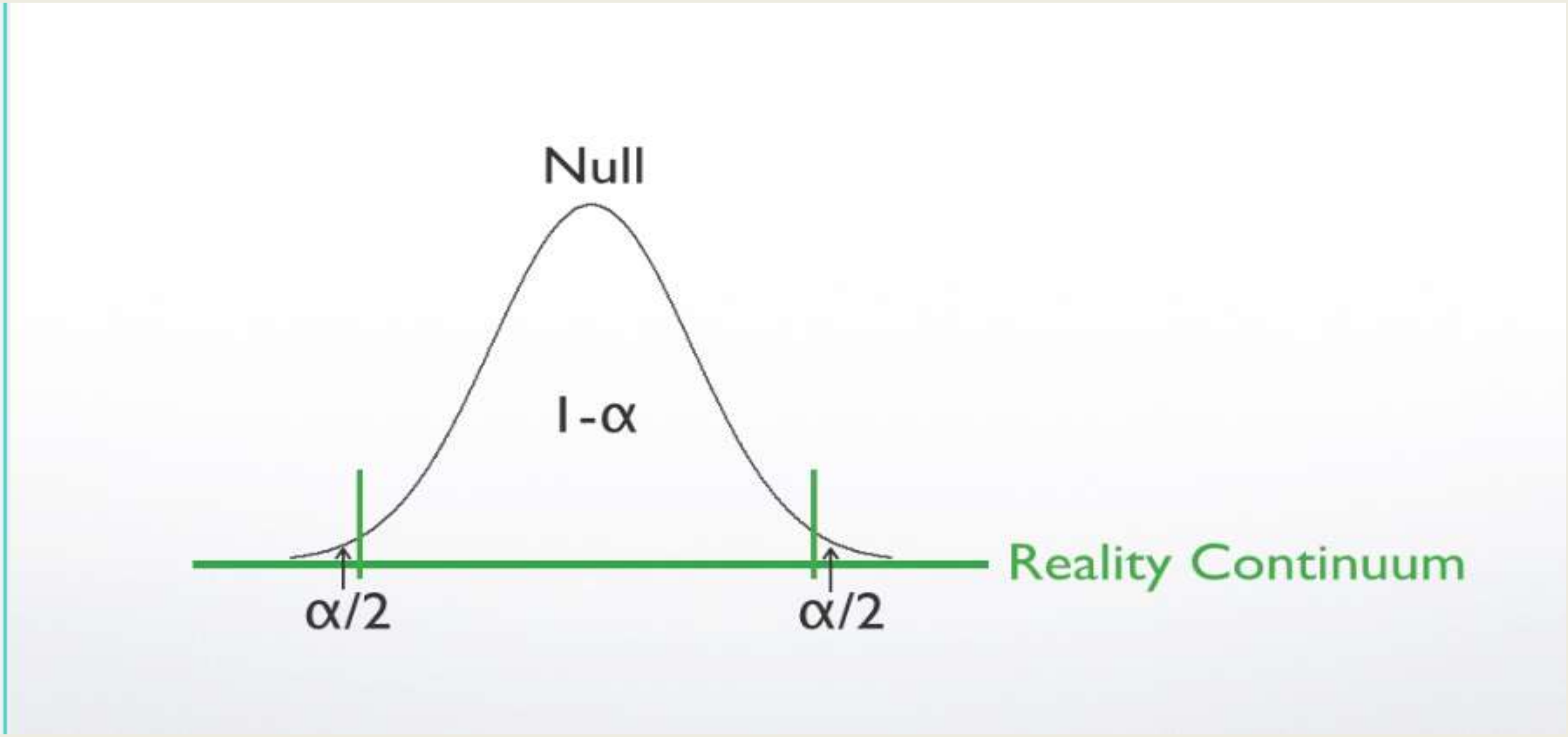
The pink distribution is from a pair of dice we trust.

Is the test pair of dice fair?



Comparing an observed data set against a known one (or could be a hypothetical one).

The p-value is the risk we are willing to take to reject the null hypothesis. The criterion for p we set at α. This is a two-tailed test:



Using statistics properly

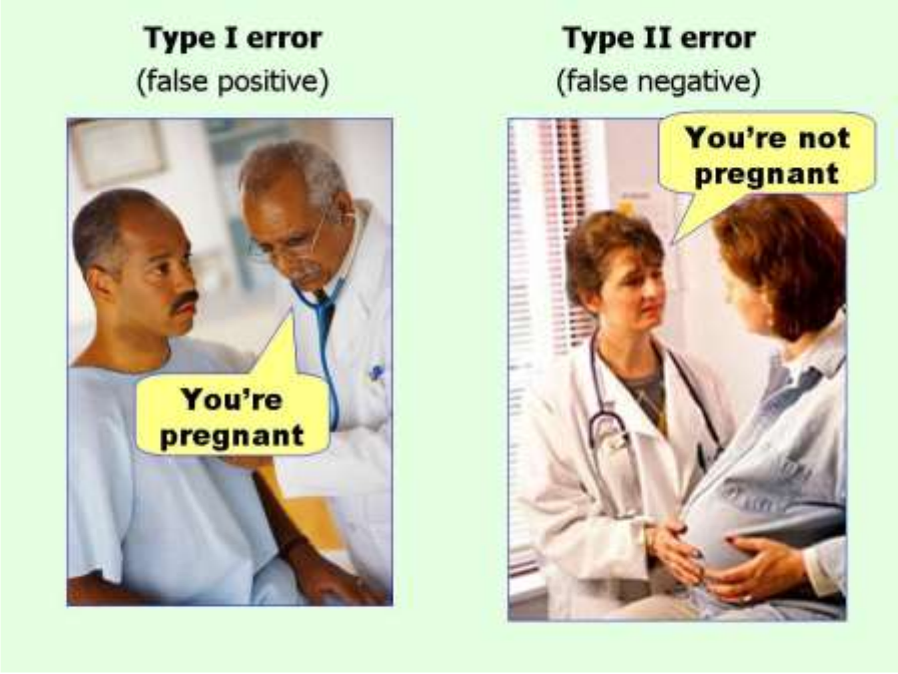
- When we have fewer observations, we should have less confidence in making inferences from the data.
- This means that larger ns and precise measurement are important!
- If we don't have enough data, we can't really make appropriate inferences from the data. For example, with an n=0 (the limit of "not enough data") we have nothing to make an inference from.
- Luck (i.e., measurement error) is always involved, so we can also believe we see an effect even though in reality (the rest of the possible cases we did not observe), it isn't.

Type I and Type II Errors

- Type I error is a **false rejection** of the null hypothesis (i.e., you said you found support for your hypothesis when it was actually not there).

- Type II error is **failure to reject** the null (i.e., you said you did not find support for your hypothesis when it was indeed there)

Hypothesis Testing		Truth: The Null Hypothesis Is True	Truth: The Alternative Hypothesis Is True
Research	The Null Hypothesis Is True	Accurate	Type II Error
Research	The Alternative Hypothesis Is True	Type I Error	Accurate



Basic Principle

- If your radio has a LOT of static (noise), it can be hard or impossible to tell what it is really broadcasting (the true signal).
- Tuning, or having a more powerful receiver, should reduce the noise.
- The effect is: one hear the signal.
- All inferential (test) statistics are ratios of something like signal/noise.
- So, getting to measuring the effect accurately is a matter of increasing the signal (effect size) while decreasing the noise (a matter of increasing the N).

Student's t-test: one group
Compares one sample to a constant

- Mean is unweighted average of all cases,

  M or X-bar
- $M = \Sigma(X_i)/n$ for $i = 1, \ldots N$
- Standard deviation (SD)
- $SD = \mathrm{sqrt}(\Sigma(X_i-M)^2)/(N-1)$ for $i = 1, \ldots N$
- $t = (M-\mu/SD) * \sqrt{N}$

- N participants
- X is measured variable
- μ = expected value of X
- df = N-1 (we used up 1 df to estimate the mean)

t has a probability distribution function, from which we determine p-value for this t and DF and α

One-sample t-test

$$M = \frac{\sum x_i}{n}$$

$$SD = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

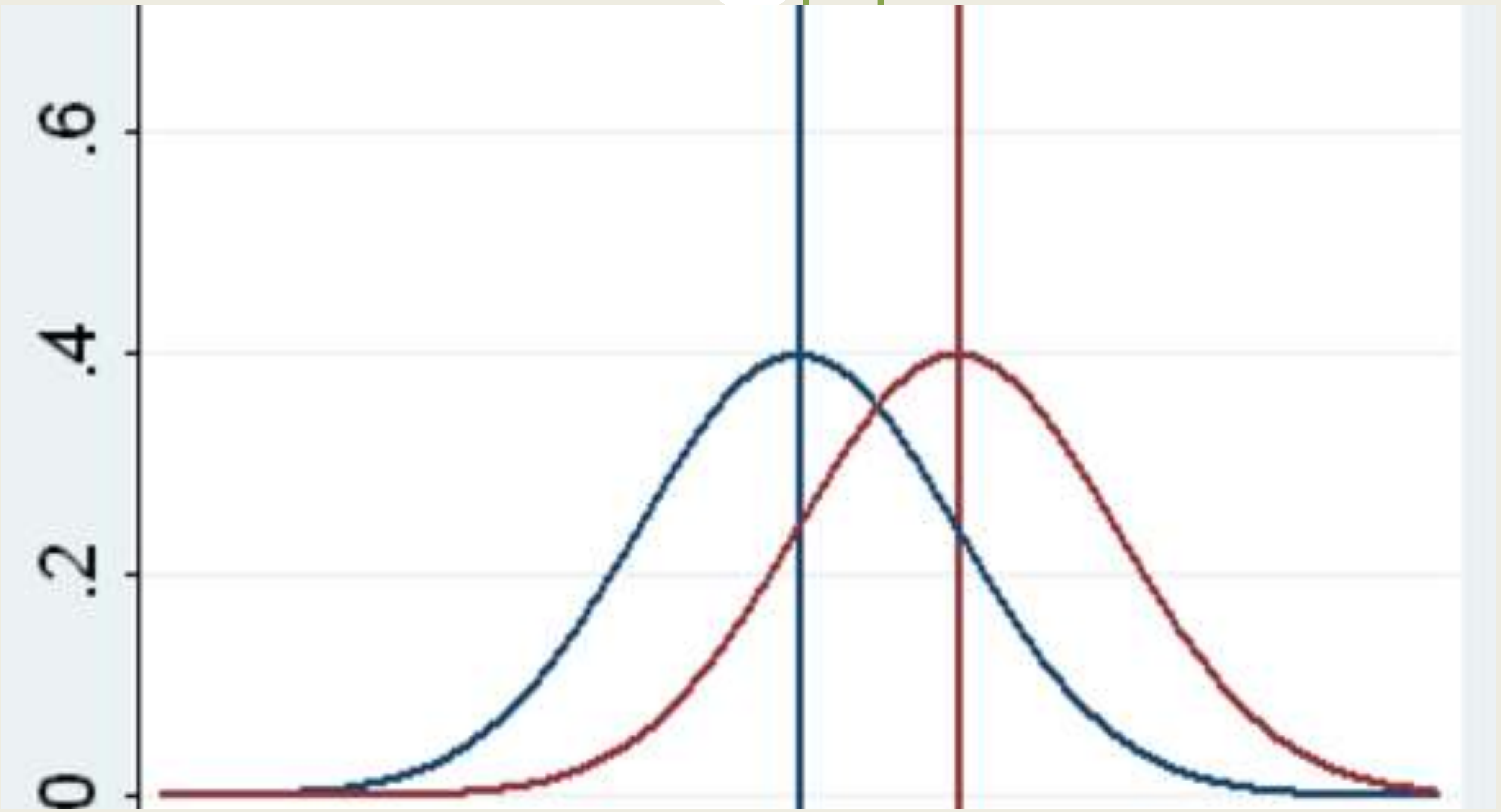
$$t = \frac{(M - \mu)}{SD}\sqrt{n}$$

- Calculate mean and SD
- Calculate t-value
- Compare t-value with df = n-1 to a table with alpha-level to check p-value
- Calculators are also available for free on line:
- https://www.socscistatistics.com/tests/tsinglesample/default.aspx

Student's t-test (due to Fisher): two groups

- Used to compare two samples to each other
- Assumes normal distribution; *t*-distributed
- Assumes equal group sizes (fyi: Levine's test)
    - **You are not required to know this—but if you see it reported alongside t-test results it's a test of homogeneity of variance between groups
- The point here is NOT to generalize from a sample to a population.
- Rather, the point is to test whether two distributions of data (e.g., from two experimental conditions) are equivalent, OR whether there is a reliable difference between them.

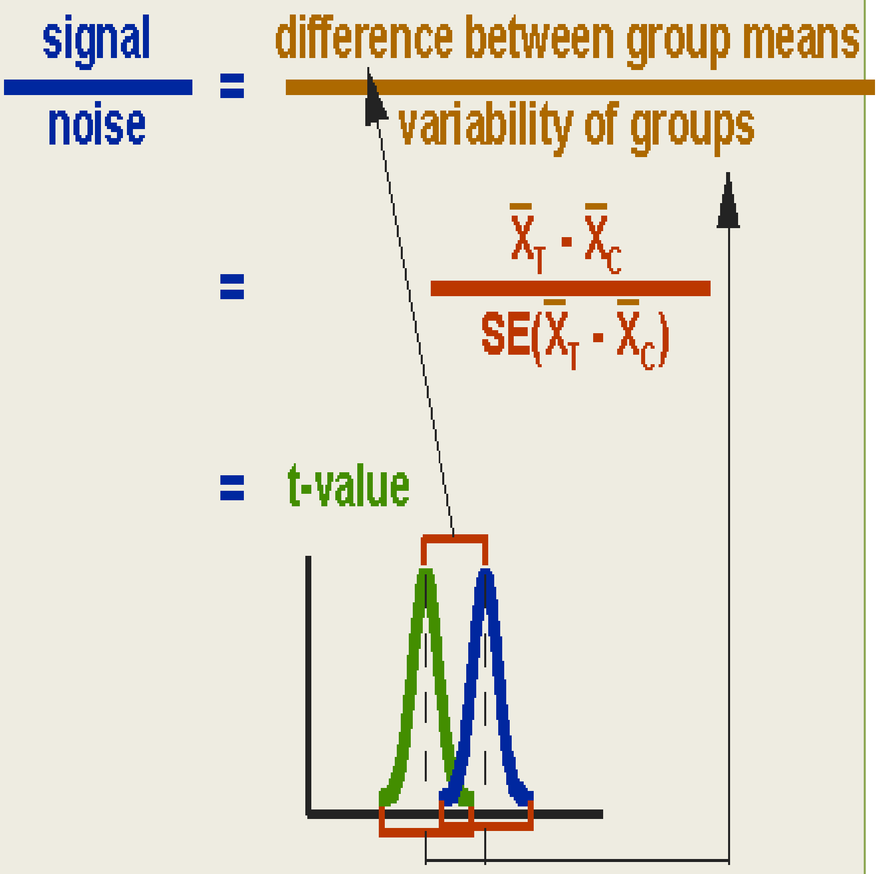
Question is: Do these two distributions overlap enough that we should consider them as from same population?



Two-group *t*-test (lower case *t*)

- To calculate this, you have to derive the means and standard deviations for both groups separately.

- Because you do that, you lose 2 df for this test.
- How can you make your SD smaller?



Interpreting Results

- Regardless of the inferential test used, you're going to come across *p*-values, or reliability ("significance")
  - Reliability (Statistical significance) means we have reason to believe that the effects are not due to chance. If we repeated the study again, we would have a high probability of finding a similar pattern. We reject the null hypothesis of no effect or no differences based on *a priori* significance criterion (by custon, p </= 0.05).
    - **In other words, the observed results were not due to random chance alone**
- Different statistics follow different distributions (*Z*, *t*, *F*, etc.) and have different assumptions, but <u>all tests will report a *p*-value.</u>

Reading Results: t-test

- "This study found that participants in the control group had statistically significantly higher weight (*M* = 150, *SD* = 11) than those in the experimental group (*M* = 120, *SD* = 5), $t(50) = 3.1$, $p < .01$. "

- Mean and Standard Deviation for each group
- t-statistic and df
  - 3.1 and 50, respectively
- p-value
  - < .01

Chi-Square

- χ^2 test of independence

- Chi (χ) -squared test is used to determine whether frequency distributions of two categorical variables significantly differ

- Conceptually similar to a t-test or ANOVA but does not compare continuous variables
- Compares frequency counts: how many in each category?

χ^2 test of independence

The question to test is whether this distribution shows no relationship between passing and training program.

If they are not related, the chance of passing should be the same for each training program

Overall the percent passing is (37+55)/(37+55+52+19) = or 56.44%

Is overall percent passing like the percent in Program 1? In program 2?

Condition	Number who passed	Number who failed
Training program 1	37	52
Training program 2	55	19

ANOVA

- **An**alysis **of Va**riance
- Used to compare three or more cells of factorial experimental design
  - How much, if at all, do the groups differ from each other? Is it a reliable difference?
- Assumes normal distribution of DVs
- Test statistic is F-distribution
- (An F with two cells equals a t^2)
- If there is a significant difference, you may see Tukey's HSD reported
  - This tells you which groups were different from each other and by how much
- η^2 (eta-squared)
  - Measure of effect size in ANOVA

In ANOVA, whatever the type, there is always only 1 Dependent Variable

Must be continuous (numerical/scale)

ANOVA is UNIVARIATE (1 Dependent Variable). If there are more than 1 Dependent Variables, use MANOVA

ANOVA can be:

- 1-way
  - 1 independent variable

Must be categorical (nominal/ordinal)

- 2-way
  - 2 independent variable

Must be categorical

- 3,4,etc-way
  - 3,4,etc independent variable

Must be categorical

Interactions

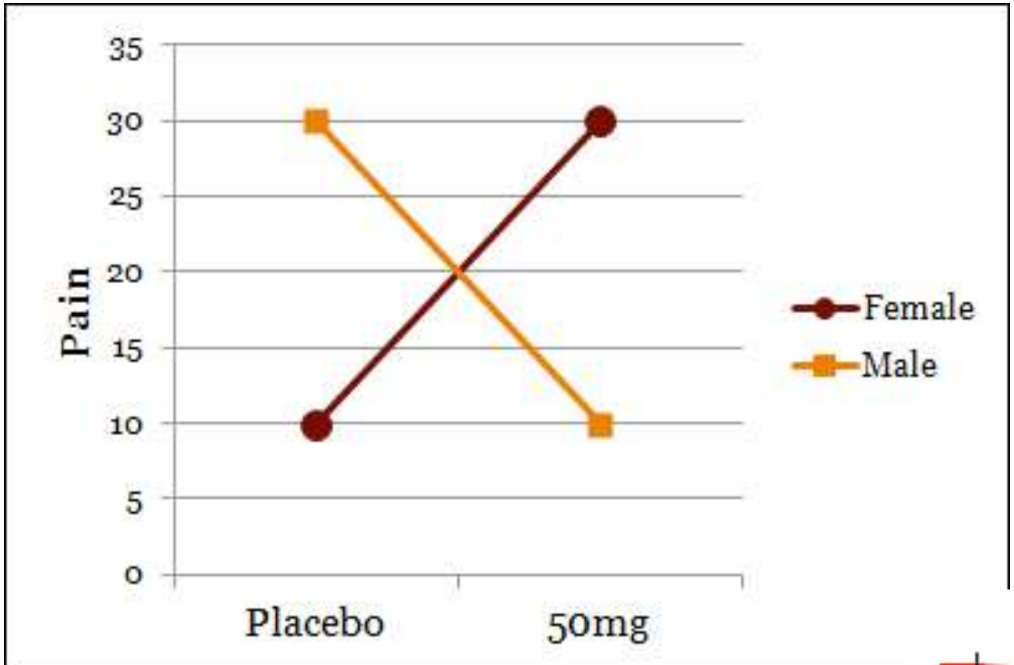
- The effect of IV_1 on the DV could be influenced by IV_2
- Factorial design (multiple factors)
  - ANOVA
- The interaction itself is NOT a variable, but a mathematical placeholder representing the relationship between IV_1 and IV_2 on the DV

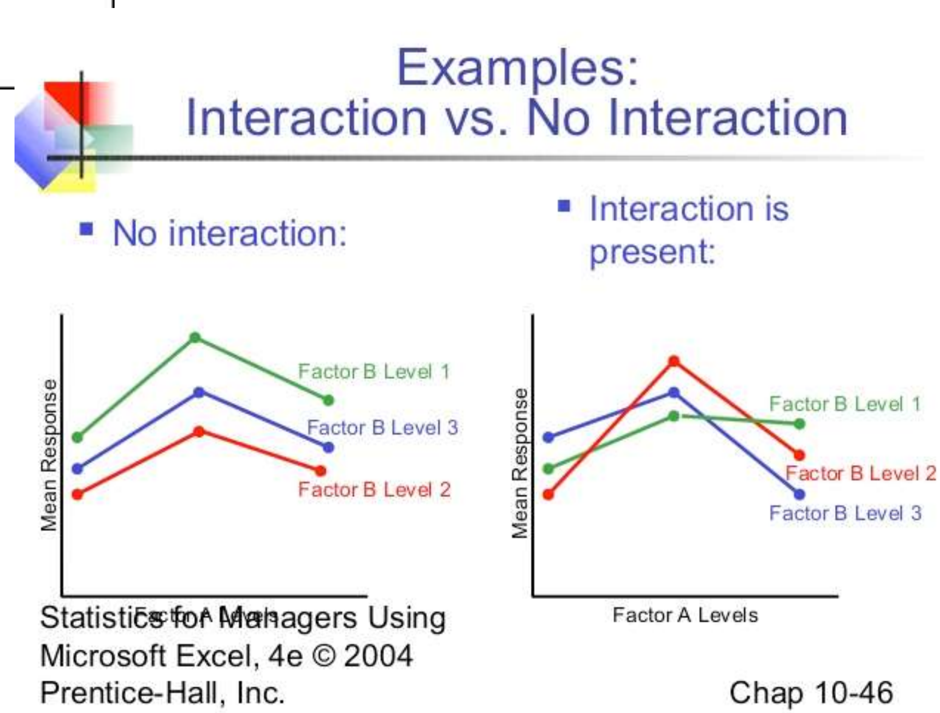
- A reliable interaction shows that there is a condition to when a statement is true. This can also be known as a dissociation, or one can say that IV_2 *moderates* the influence of IV_1 on the dependent variable X.

Reading Results: ANOVA

- "There was a statistically significant difference between groups as determined by a one-way ANOVA, $F(2, 30) = 5$, $p = .0003$. Tukey's HSD indicated that participants' reported self-esteem was statistically significantly lower when presented with sad images ($M = 4$, $SD = .25$) and neutral images ($M = 7$, $SD = 1.2$) compared to positive images ($M = 11$, $SD = 1.9$)."

- This statement (that I made up) tells us that there is an overall difference among the three conditions, the the Tukey's test tells us which differs from which. Note also the very small SD for the sad images. They apparently influenced everyone strongly.

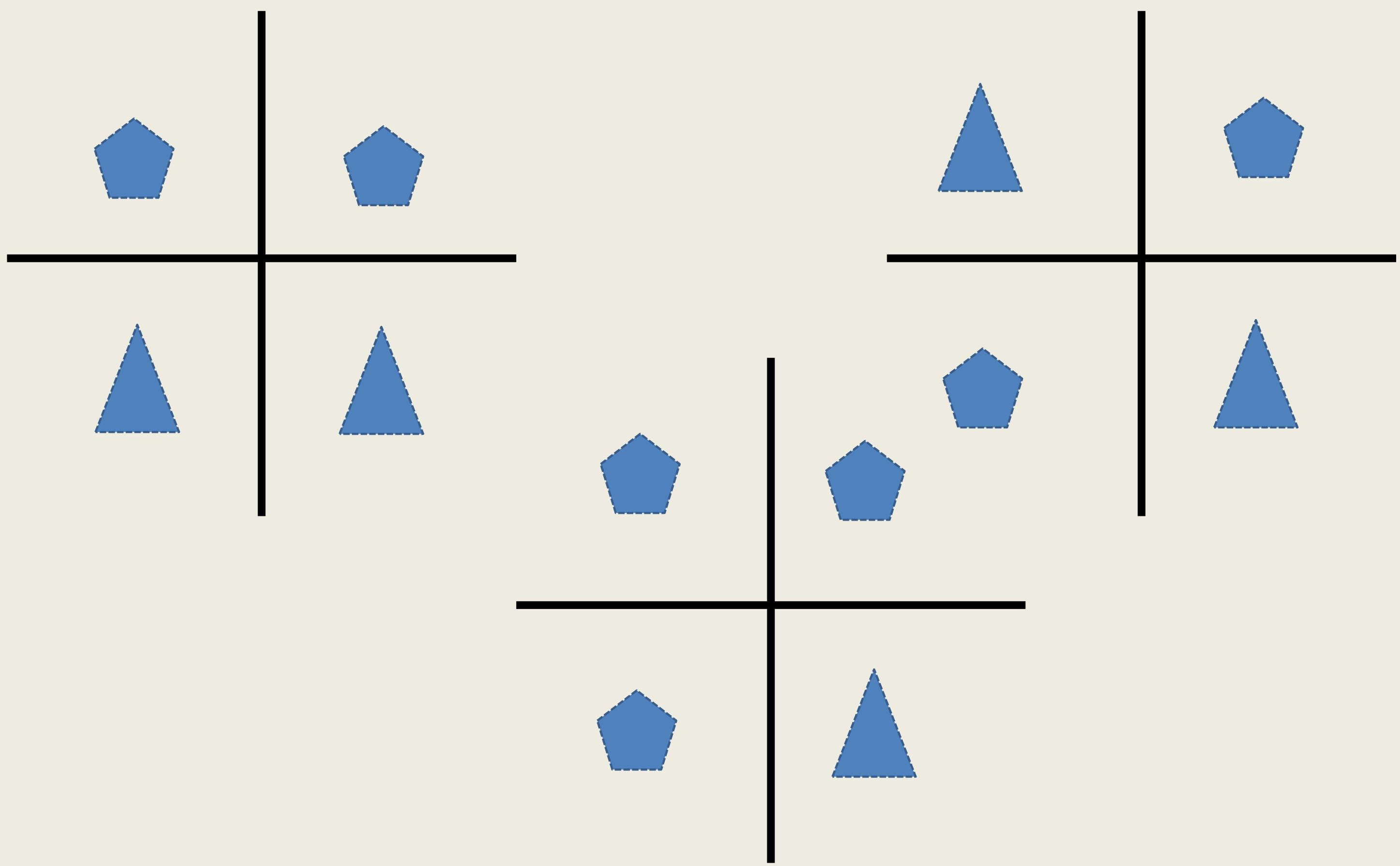

Pain
35
30
25
20
15
10
5
0
Placebo
50mg
Female
Male


Examples:
Interaction vs. No Interaction
No interaction:
Interaction is present:
Mean Response
Factor B Level 1
Factor B Level 3
Factor B Level 2
Factor A Levels
Mean Response
Factor B Level 1
Factor B Level 2
Factor B Level 3
Factor A Levels
Statistics for Managers Using Microsoft Excel, 4e © 2004 Prentice-Hall, Inc.
Chap 10-46

More on Interactions

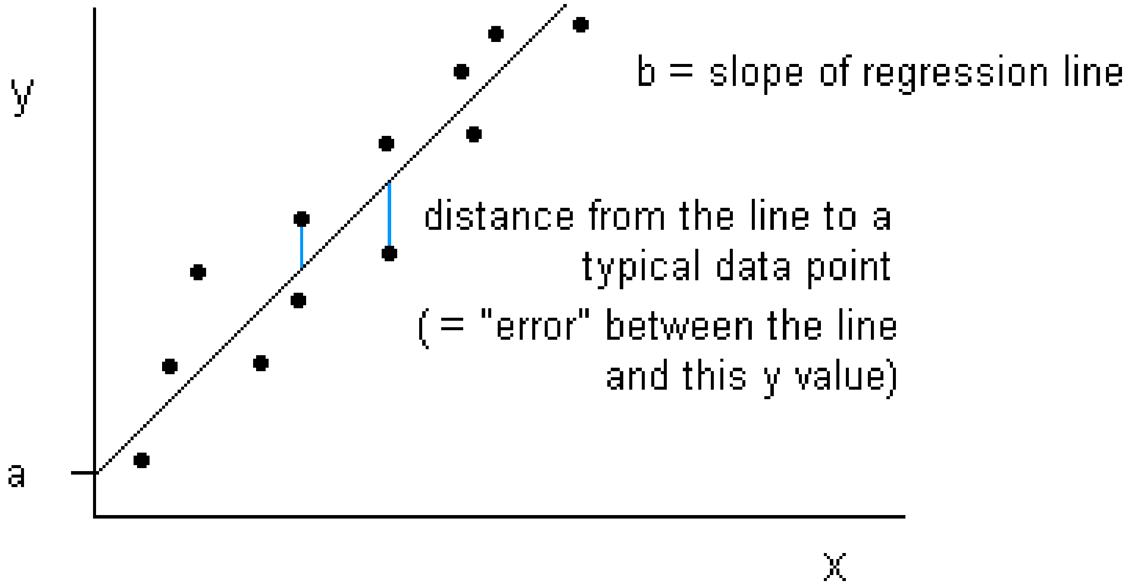
- Because they are *contingencies*, they can be hard to think about at once.
- An interaction means at least 2 different things happened:
- When someone has to describe ANOVA results with an "IF" in them, they might have an interaction.
- Interactions are also called "moderation" (because one variable "moderates" the effect on another one.
- Interactions are also called "dissociation" in experimental psychology, because one effect get unassociated with the other.

Interaction: One of these things is not like the other

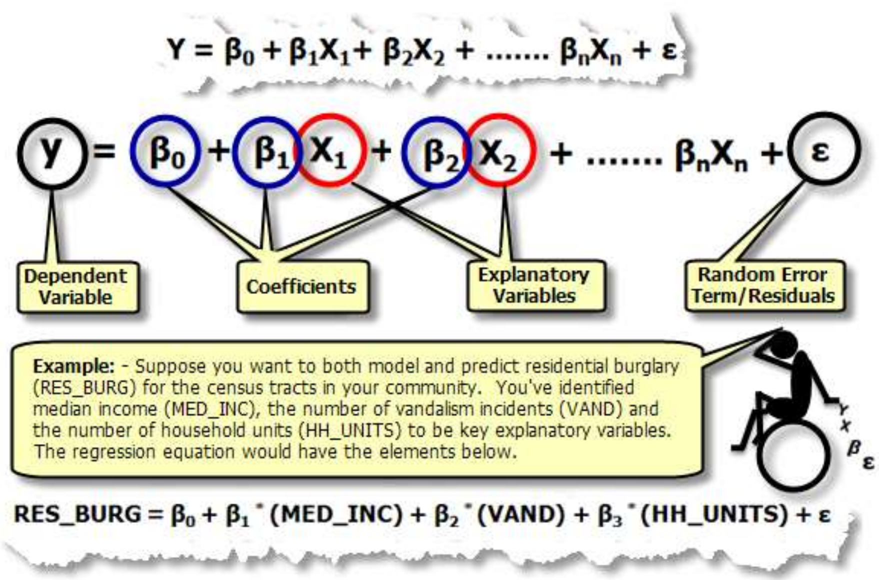


Regression

- Simple Linear
  - What's the relationship between one IV and one DV?
    - As the name implies, assumes a linear relationship between variables. The DV has to be continuous. The IV could be dichotomous or continuous
    - Examples: DV: how warm do you feel towards Democrats from 1 to 10? IV: Are you a Republican or not? IV: How warm do you feel towards Republicans from 1 to 10?
    - Beta, R^2
- Multiple Linear
  - What's the relationship between 2+ IVs and one DV?
- Logistic
  - DV is categorical, so a log transform must be utilized
    - Example: Are you Democrat, Republican, Green, Working Families, independent, unregistered?



When reading articles, the authors may explain their models like this:



Reading Results: Regression

- Constant is the mean.
- Each predictor variable will have an estimated weight, known as β.
- There is a t-test for each β that tests whether it differs reliably from 0.
- β can be positive or negative. β close to zero means there is no relation between X and Y.
- R^2 indicates how much variance in the outcome variable was accounted for by the whole set of predictor variables.

Other Helpful Resources

- These resources are easy to navigate and explain both simple and complex statistical concepts well. They may not use examples relevant to this course, but if you're having trouble deciphering results or want to know more about experimental/analytical methods check these out!

- UCLA stats help page (http://www.ats.ucla.edu/stat/)
- Wolfram Alpha (https://www.wolframalpha.com/examples/Statistics.html)
- Texas A&M Stats

(http://bobhall.tamu.edu/FiniteMath/Module8/Introduction.html)

- Laerd Stats

(https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php) *click around to navigate to another test—there was not a homepage to link to*

Mediation & Moderation

- Mediation
  - One variable explains the relationship between two other variables
    - E.g., Stress → **Rumination** → Depression
- Moderation
  - One variable affects the strength of the relationship between two other variables
    - Suppose the results look like this:
    - Stress → Rumination→ Depression for women
    - Stress does not make for more rumination for men.
      - **Gender** moderates the strength relationship between rumination and depression, so it's the moderator

Mediation & Moderation, cont.

- It's likely that you'll encounter *moderation* in this course more than mediation, but know the conceptual difference and whether the results make sense!
  - Do they fit with the hypothesis?
  - What about other theories?
  - Is the result surprising or novel for any reason?

Moderation is also a "statistical interaction" like in ANOVA and also called "dissociation" in experimental psychology.