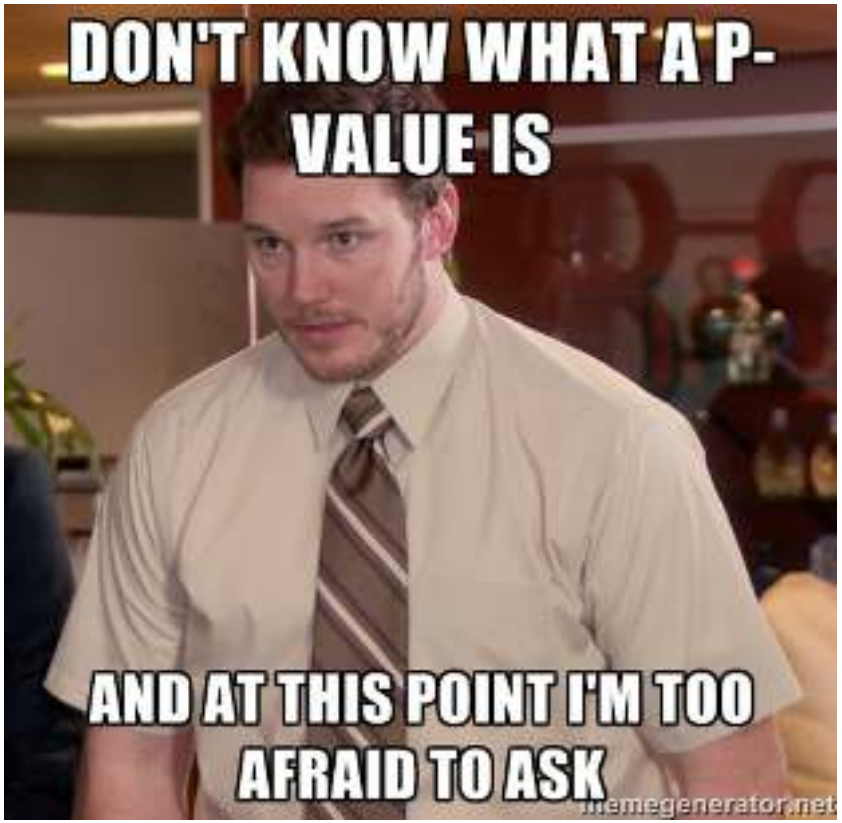
Basic Statistics Using R

University of Toronto Scientific Coders

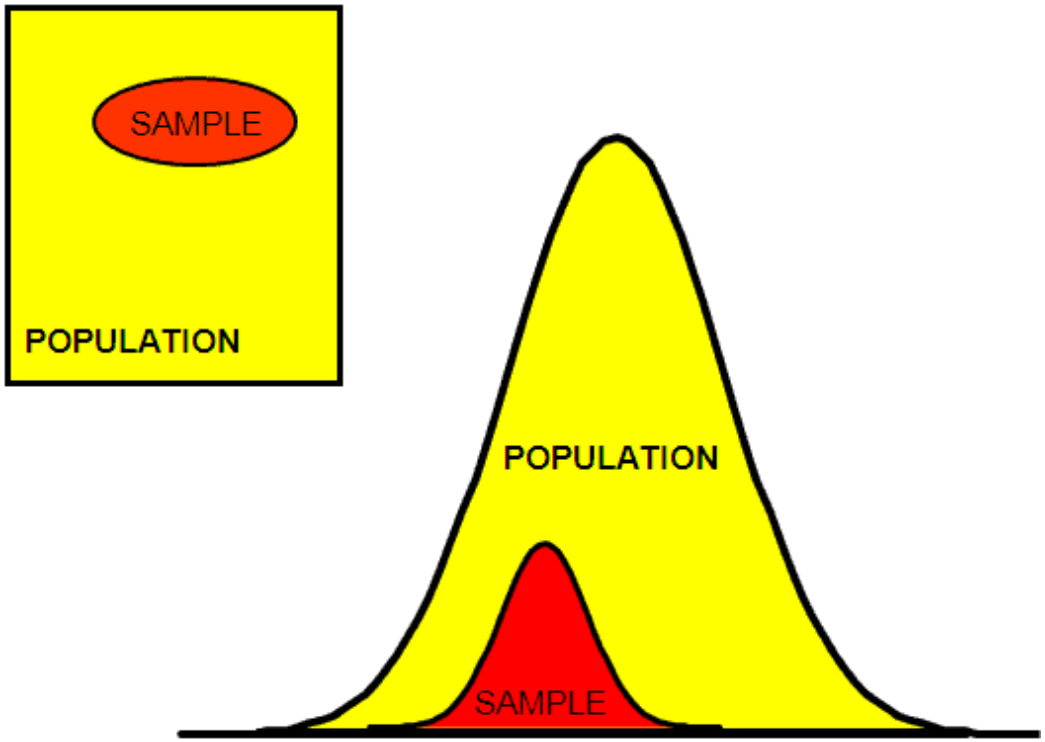
Prepared by Lindsay Coome



Populations and Samples

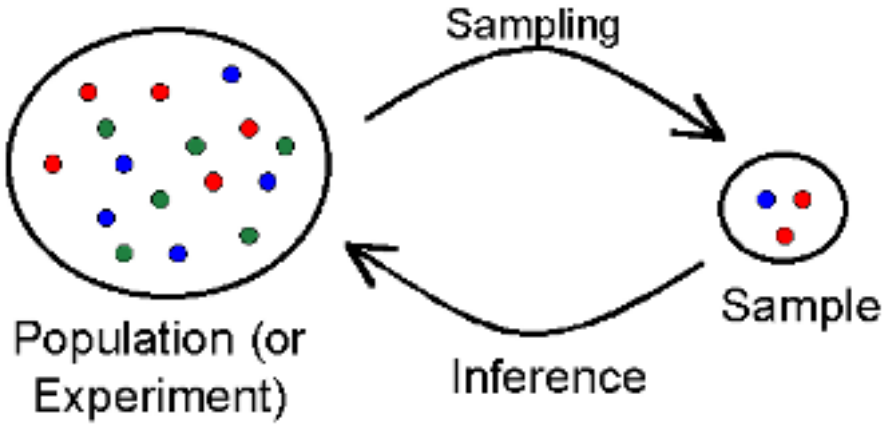
Population: A large collection of things we want to study (usually people in psychology; could be anything!)

Sample: A subgroup of some population, which can be random or nonrandom



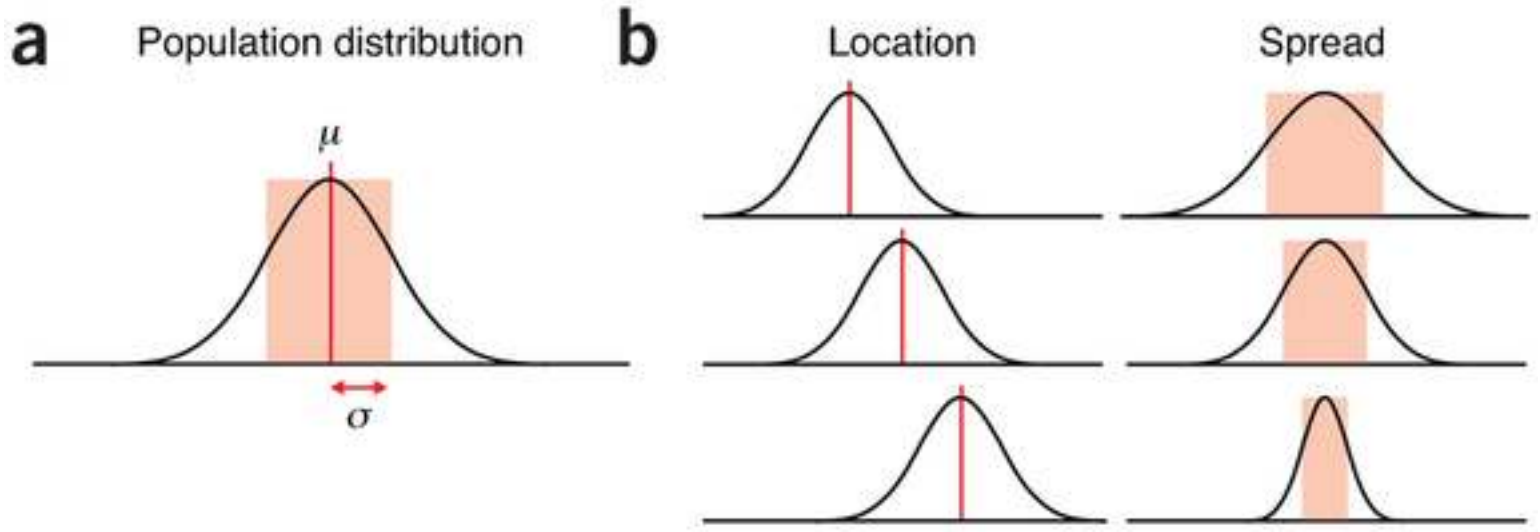
Statistics

- We use statistics to determine the values of a variable in the population
- We usually cannot measure the entire population
- So, we use a sample to measure the value of a variable, which is then used to estimate the population value



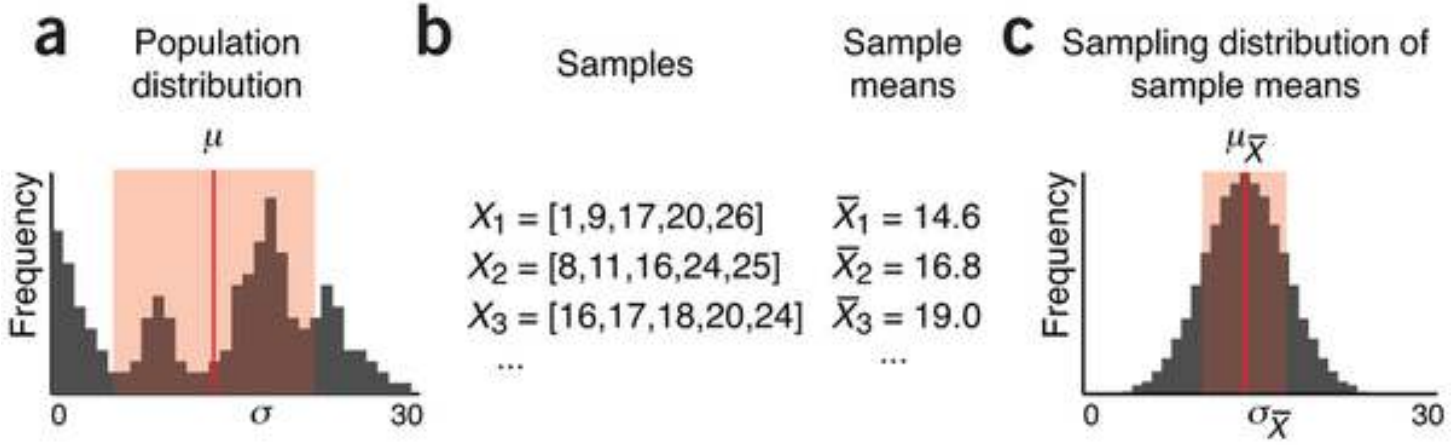
Statistics

- Statistics usually fall into two categories: descriptive and inferential
- Location (e.g. mean) and spread (e.g. standard deviation, variance) are important descriptive stats



Inferential Statistics

- We use sampling to make inferences about the population
- A sampling distribution is a distribution of the sample means of all possible samples of a certain size (i.e. your sample size) taken from the population
- The idea of a sampling distribution: take many samples from the same population, collect all the means from the samples, and display the distribution of means

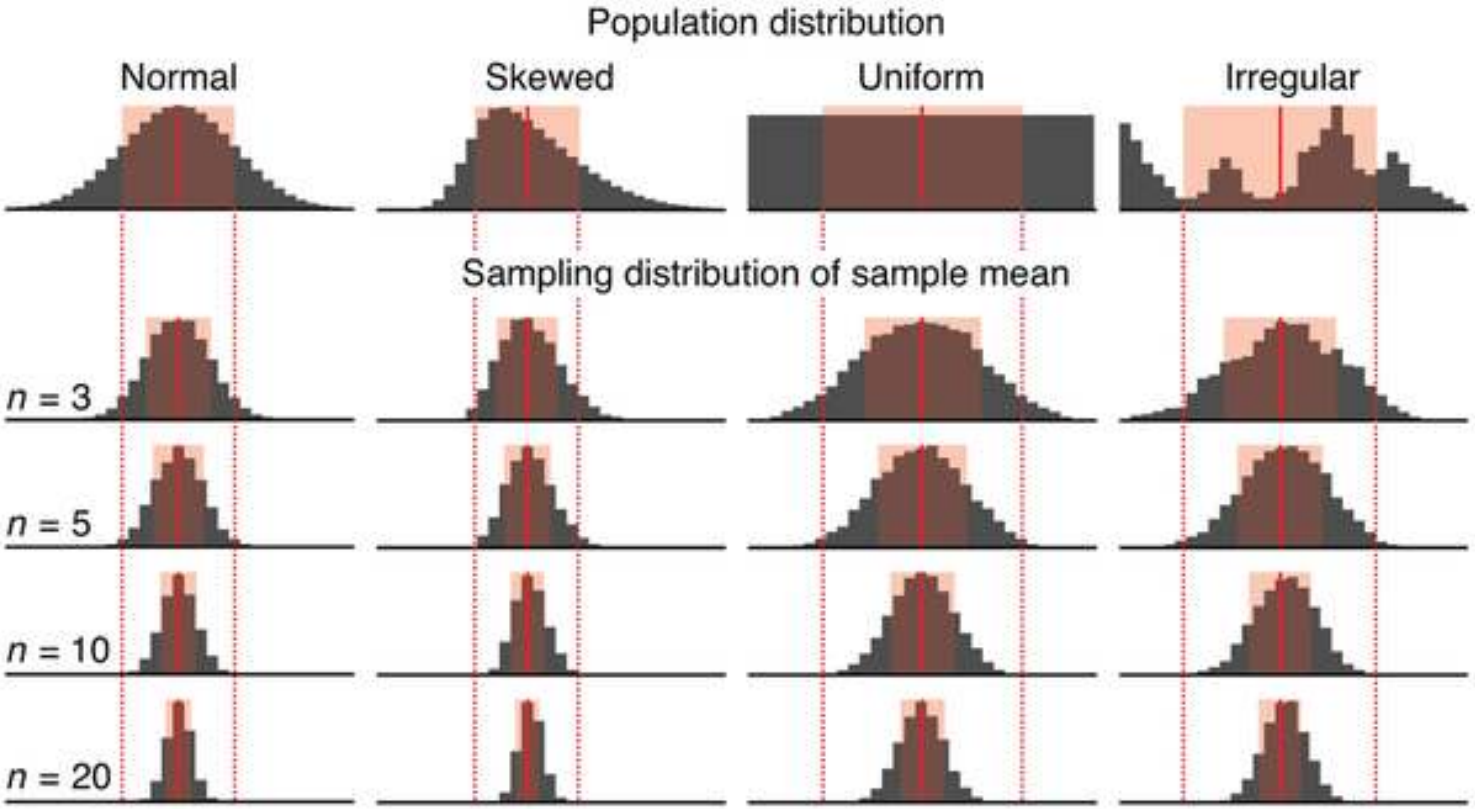


Inferential Statistics

Why do we care about sampling distributions?

- We know certain things about the properties of sampling distributions that allow us to use statistical tests
- When you have a big enough sample size, the sampling distribution of the mean approximates a special distribution shape – the Normal distribution (AKA the bell curve) – no matter what your population looks like

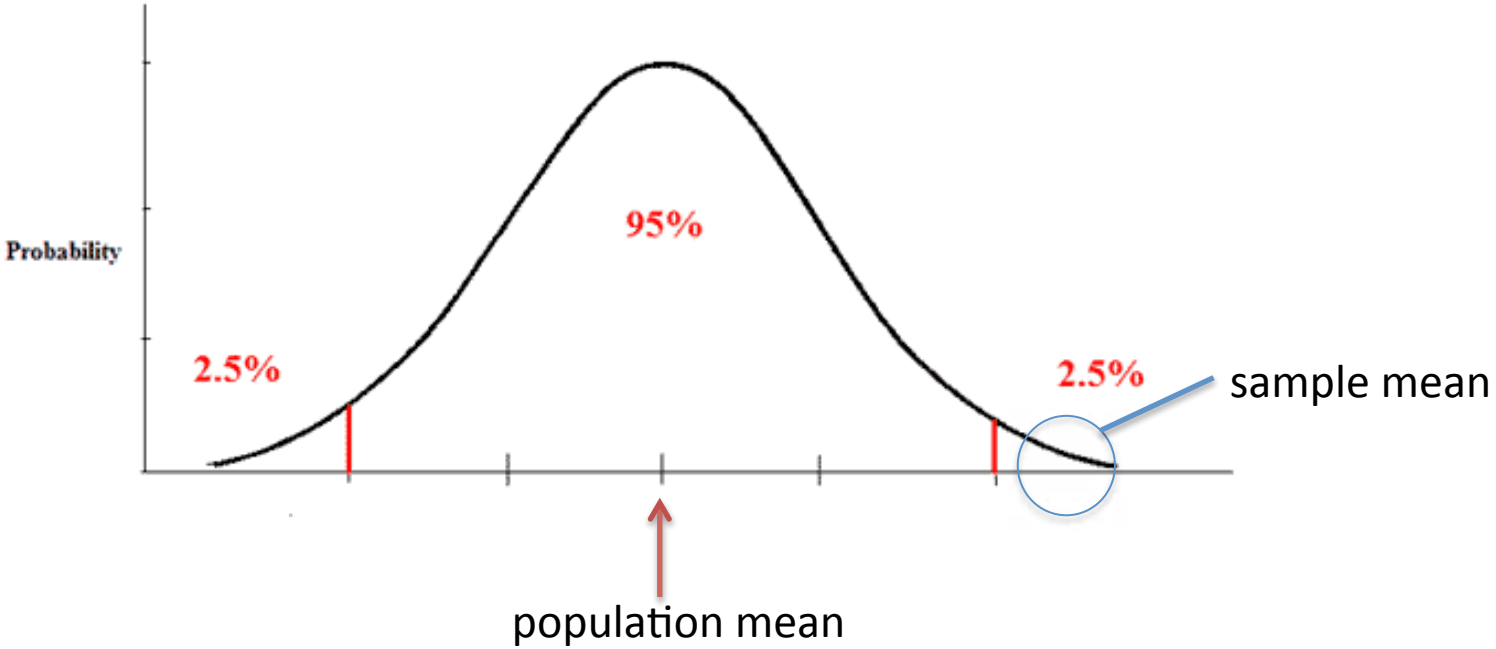
Inferential Statistics



Note that with larger sample sizes, the spread of the sampling distributions gets smaller – this means that with large sample sizes, the means of the samples you're picking from the population are closer to the true population mean!

Inferential Statistics

- One-sample T test: is my sample different from the population?
- Calculate the probability of getting your sample if the null hypothesis is true – can do this by hand using the T statistic, or a statistical package can do it for you
- The researcher makes a comparison between (1) the original singular (one-sample) statistical outcome and (2) a statistical model (sampling distribution) that gives the probability of observing all given outcomes of that statistic

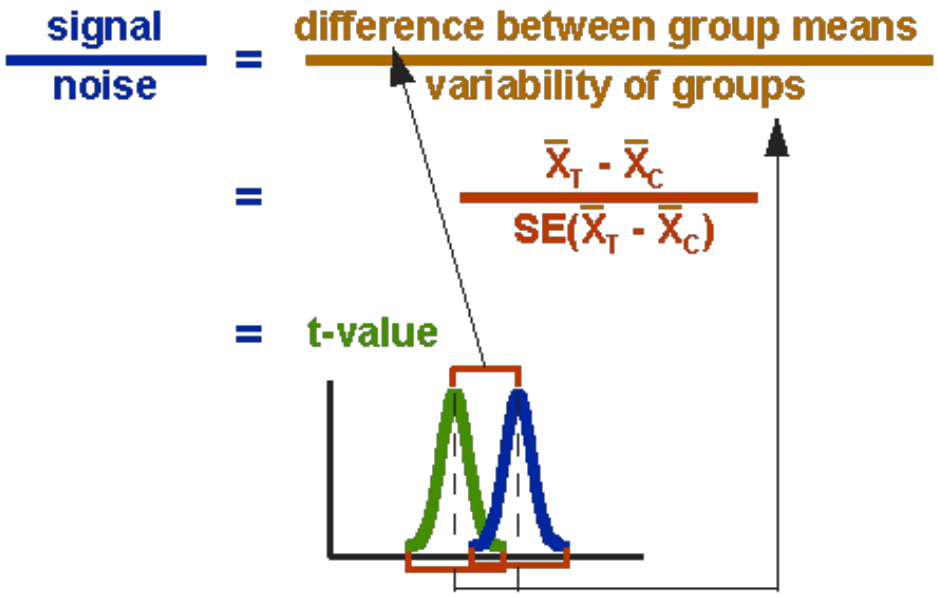


P-Value

- The results of your t-test have a p-value associated with it
- The p-value tells you how likely your result is if the null hypothesis is true
- A p-value of .05 means that if your null hypothesis is true, you'd still obtain your current sample about 5% of the time (or about 1 in 20 experiments) due to random sampling error
- Most people consider 5% to be an acceptable amount of uncertainty, and consider a p-value of .05 to be the standard cutoff (also known as alpha or α)
- This means that even if people are using statistics correctly 100% of the time, 1 in 20 statistically significant experiments are due to chance alone!

Two-Sample T-Test

- Usually, we want to see if we can infer differences between two populations by comparing two samples
- Can do this with a two-sample (AKA independent sample) t-test



Two-Sample T-test in R

A test to see if women score's on a math test are different than men's:

```
Women = c(99, 85, 78, 93, 95, 87)
Men = c(72, 75, 70, 68, 70, 80)
t.test(Women,Men,alternative="two.sided")
```

```
        Welch Two Sample t-test

data:  Women and Men
t = 4.7332, df = 7.959, p-value = 0.001498
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.71012 25.28988
sample estimates:
mean of x mean of y
     89.5      72.5
```

When sharing your p-value with others, report the t-statistic and degrees of freedom (shown here as df) as well

Two-Sample T-test in R

- Can perform other kinds of t-tests as well
- Example of one-sample T-test to see if women (on average) scored significantly lower than 95% on the math test:

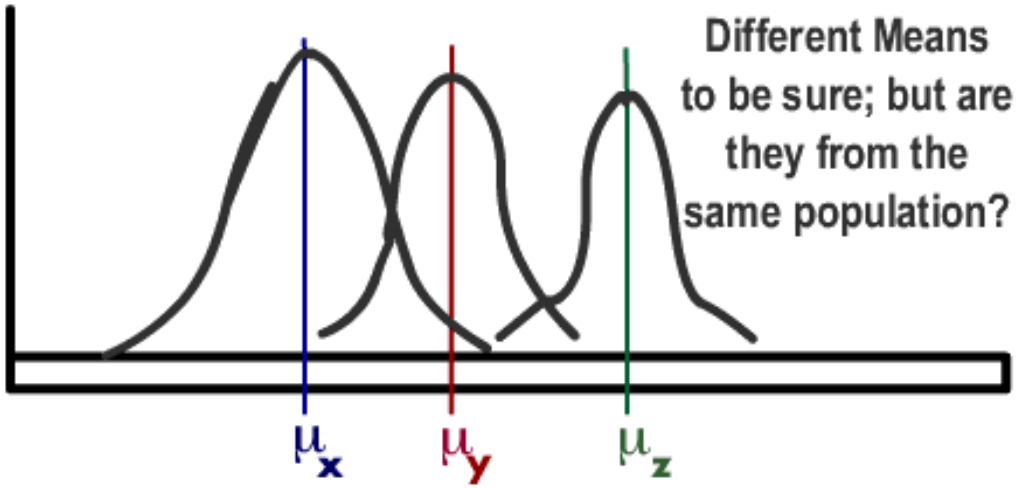
```
t.test(Women, alternative = "less", mu = 95)
```

- Can also do t-tests on paired data (AKA paired-samples t-test), for example measures of weight for a number of individuals before and after weight loss treatment
- See below link for more relevant R code for t-tests:

http://www.stat.columbia.edu/~martin/W2024/R2.pdf

Analysis of Variance

- What if we want to compare more than two groups?
- Use ANOVA – analysis of variance
- Can compare multiple groups to see if they are equivalent
- However, this isn't really a test of MEANS – it's called analysis of variance for a reason!
- ANOVA looks to see if the variance BETWEEN groups (or treatments, or whatever you're comparing) is bigger than the variance WITHIN groups



ANOVA in R

- First, we need to make sure we have coded our data so that the variable we want to analyze is matched with the appropriate factor (i.e. the variable that separates the data into groups)
- For example, a drug company wants to test three different drugs for pain relief. Subjects were instructed to take one of the three drugs during their next headache, and the company recorded subjects' pain on a scale of 1 to 10. Here are the scores:

Drug A	4	5	4	3	2	4	3	4	4
Drug B	6	8	4	5	4	6	5	8	6
Drug C	6	7	6	6	7	5	6	5	5

- To read this into R, we use the following code:

```
pain = c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5, 4, 6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
drug = c(rep("A",9), rep("B",9), rep("C",9))
```

- And then create a data frame from these two columns, where each score is associated with its drug

```
migraine = data.frame(pain,drug)
```

ANOVA in R

- Now, we run the ANOVA in R to see if there are differences between the three groups:

```
results = aov(pain ~ drug, data=migraine)
summary(results)
```

- You need to save the results of your ANOVA in R and ask for a summary, or else you won't get the precious p-value!
- So what do our results look like?

```
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  28.22  14.111   11.91 0.000256 ***
Residuals   24  28.44   1.185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA in R

- Cool! We have significant results! What does it mean?
- It means at least two of our three treatment means differ from each other – but which ones? All? Some?
- To figure this out, we need to do what's called post hoc comparisons (AKA a "test of means")
- We compare pairs or groups of means to see if they differ
- Today we are going to use a test called Tukey's Honestly Significant Difference (HSD) test because it's easy to do in R – but there are many, many more post hoc tests available, some more "conservative" than others

```
TukeyHSD(results, conf.level = 0.95)
```

- Here, "results" refers to the saved results from the ANOVA

ANOVA in R

- Tukey's test shows that drug A differs from both drugs B and C:

```
$drug
         diff        lwr      upr     p adj
B-A 2.1111111  0.8295028 3.392719 0.0011107
C-A 2.2222222  0.9406139 3.503831 0.0006453
C-B 0.1111111 -1.1704972 1.392719 0.9745173
```

- For a more detailed explanation of one-way ANOVAs in R:

http://www.stat.columbia.edu/~martin/W2024/R3.pdf

Two-Way ANOVA

- Wait! What if we want to look at another factor – like gender?
- For this we can do a two-way, or two-factor, ANOVA
- First, we need to add another column listing the gender of each patient

```
migraine$gender <- c("F", "F", "F", "M", "M", "M", "M", "F", "F", "M", "M", "F", "F", "F", "M", "F", "M", "M", "M", "F", "F", "F", "F", "M", "F", "M", "M")
```

Two-Way ANOVA

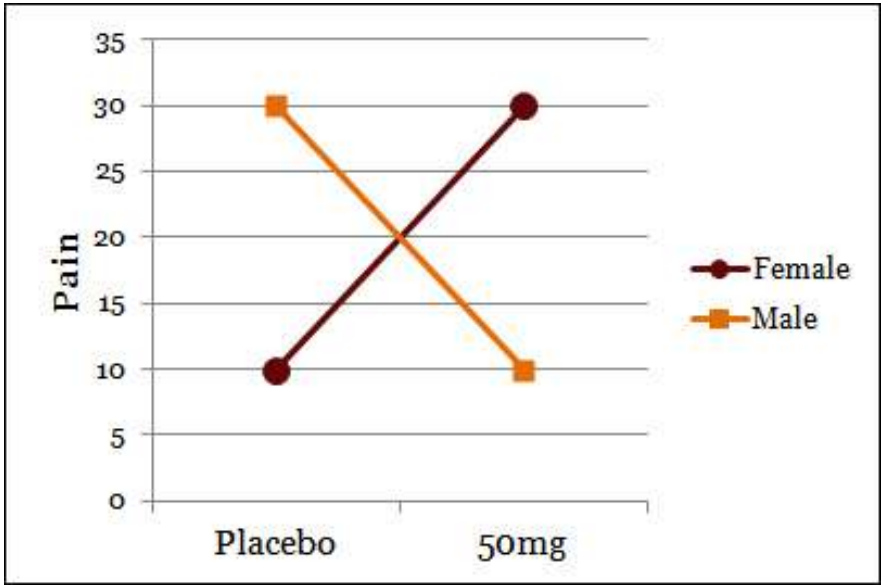
- Then, we run the ANOVA the same as before, but this time with our added factor of gender:

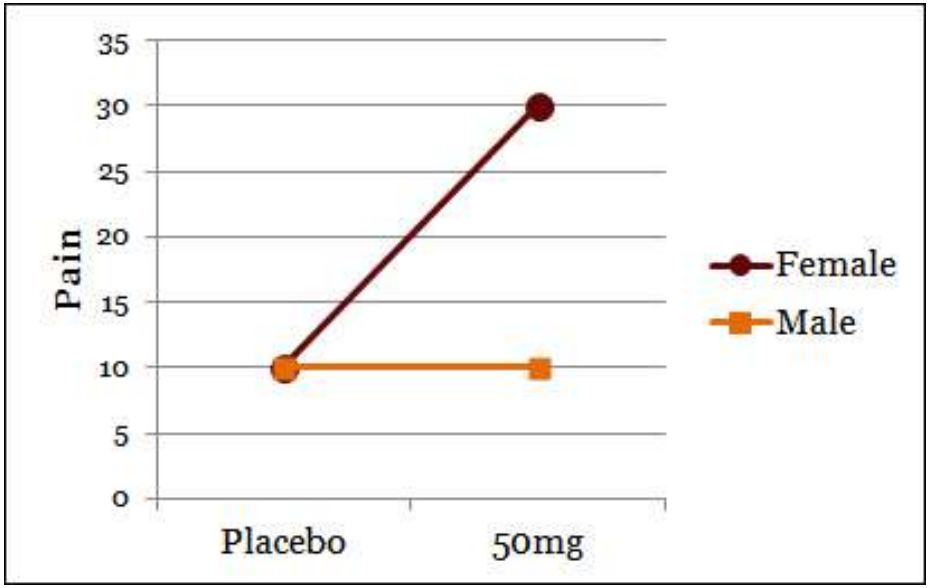
```
results = aov(pain ~ drug*gender, data=migraine)
summary(results)
```

- We add factors by attaching them with *
- This will also give us results for an interaction between the factors
- An interaction tells us whether the effect of one factor is different for different levels of the second factor
- For example, does the effect of each drug depend on whether you're a man or woman?

Two-Way ANOVA

- Two examples of interactions:





In this case, the effect of the drug depends on whether you are a man or woman

Two-Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
drug	2	28.222	14.111	28.088	1.16e-06	***
gender	1	0.002	0.002	0.004	0.952	
drug:gender	2	17.893	8.946	17.808	3.00e-05	***
Residuals	21	10.550	0.502			

- We still have that main effect of drug, but no main effect of gender
- However, we now have an interaction between drug and gender!
- To find out what this means, we usually proceed with post hoc tests as usual:

TukeyHSD(results, conf.level = 0.95)

- But this gives us a lot of comparisons – time to visualize the interaction

Two-Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
drug	2	28.222	14.111	28.088	1.16e-06	***
gender	1	0.002	0.002	0.004	0.952	
drug:gender	2	17.893	8.946	17.808	3.00e-05	***
Residuals	21	10.550	0.502			

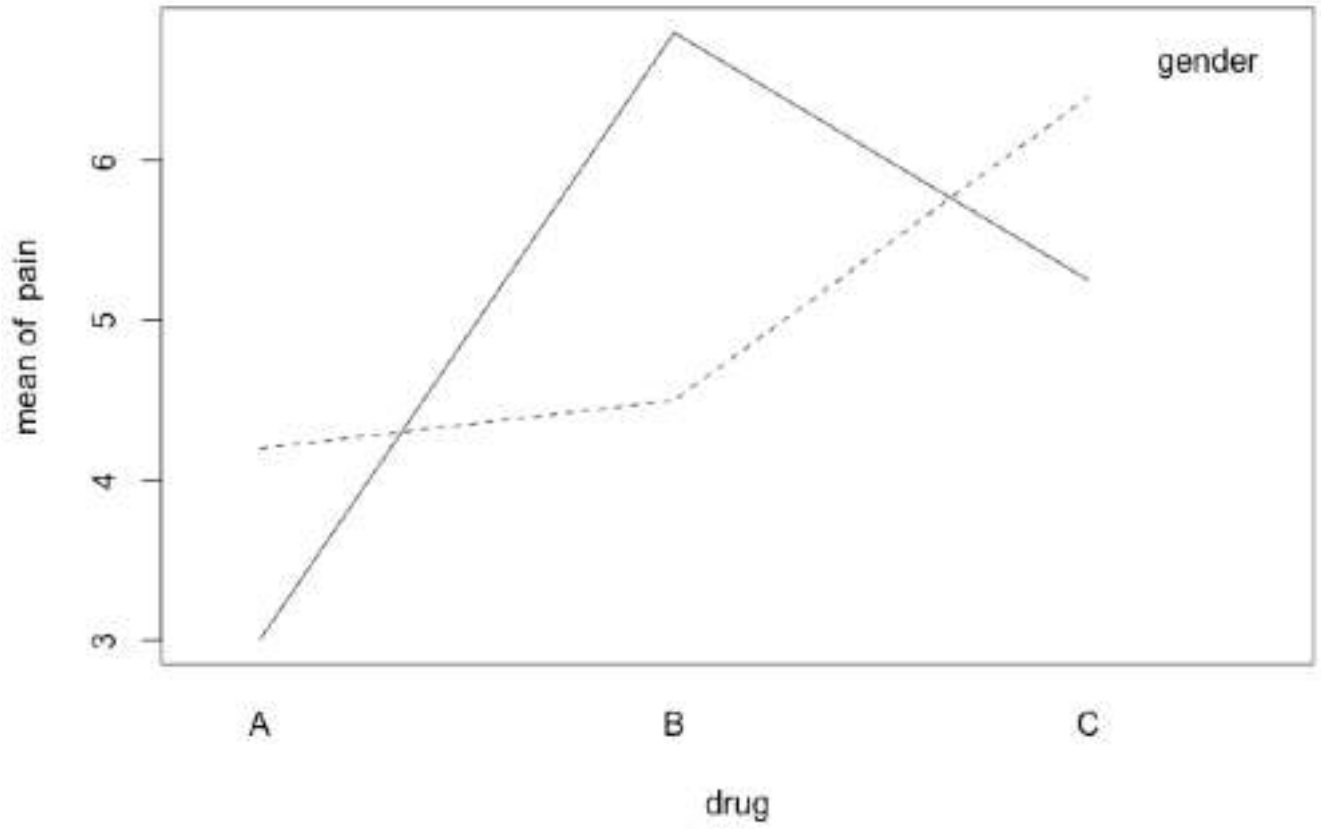
- We still have that main effect of drug, but no main effect of gender
- However, we now have an interaction between drug and gender!
- To find out what this means, we usually proceed with post hoc tests as usual:

TukeyHSD(results, conf.level = 0.95)

- But this gives us a lot of comparisons – time to visualize the interaction

Visualizing ANOVAs

interaction.plot(drug, gender, pain)



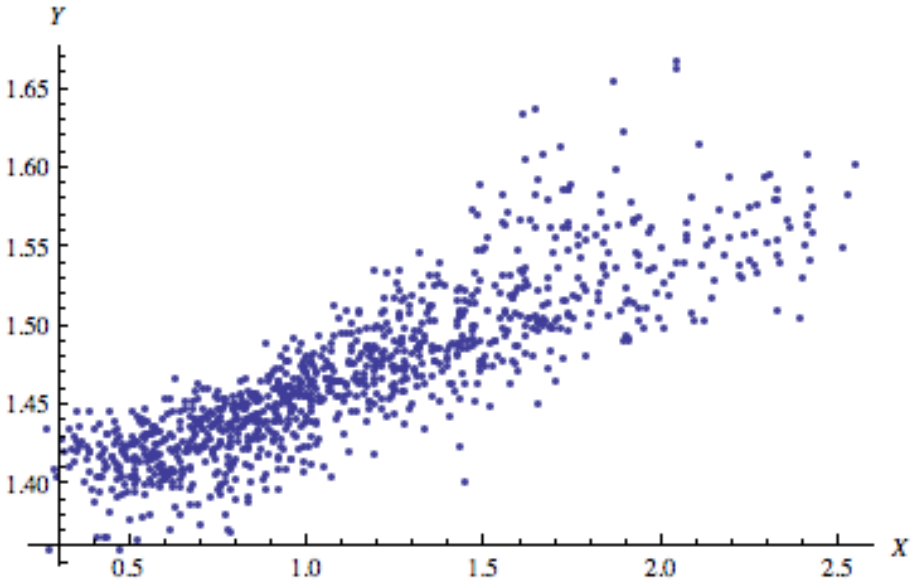
Two-Way ANOVA

- For more on two-way ANOVAs, including how to use a covariate in your analysis, see this link:

http://www.stat.columbia.edu/~martin/W2024/R8.pdf

Correlation and Regression

- Correlation/regression is actually very similar to ANOVA
- Instead of factors with discrete levels, we want to look at the relationship between two variables for "matched" or "paired" data
- Can look at strength or direction of the relationship



Correlation in R

- Let's create a new dataset – a column for the age of contestants on season 12 of the bachelor, and how many weeks they were on the show before they were sent home:

```
weeks = c(8,7,6,5,4,4,3,3,3,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1)
age =
c(24,23,25,24,23,22,27,28,23,30,27,30,24,27,29,30,27,29,25,3
4,27,22,25,28,28,23)
bachelor = data.frame(weeks, age)
```

- Then, we run the correlation on our two variables:

```
cor.test(bachelor$weeks, bachelor$age)
```

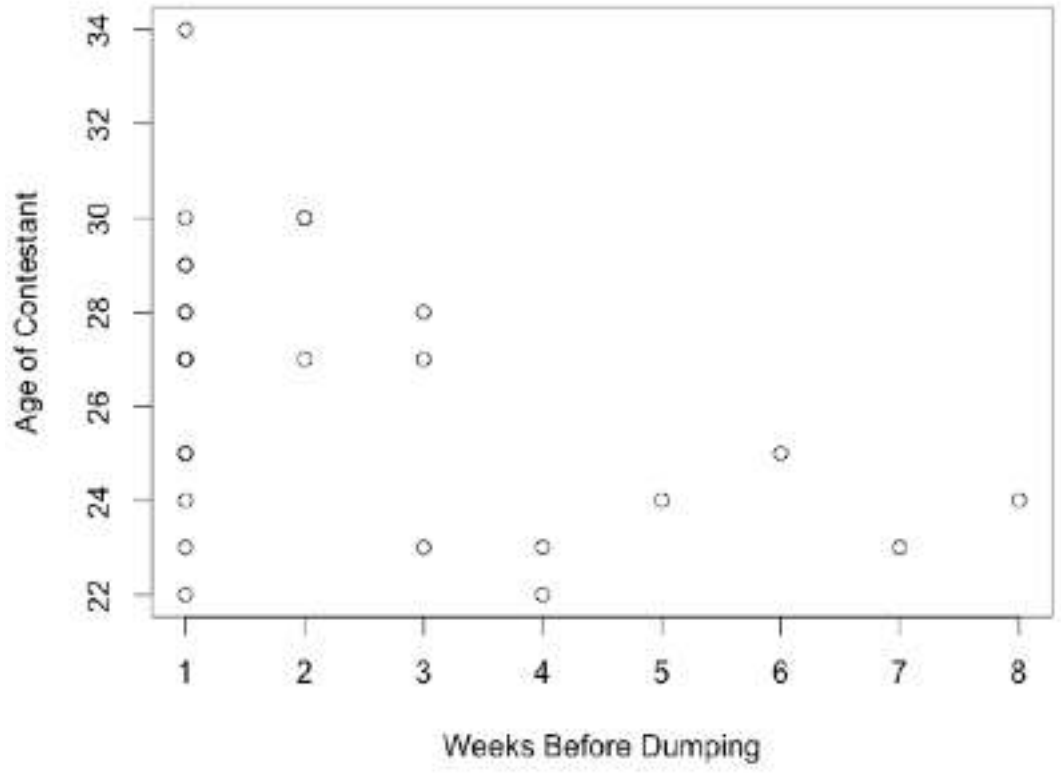
Correlation in R

```
t = -2.3978, df = 24, p-value = 0.02463
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.70663528 -0.06298623
sample estimates:
       cor
-0.4396126
```

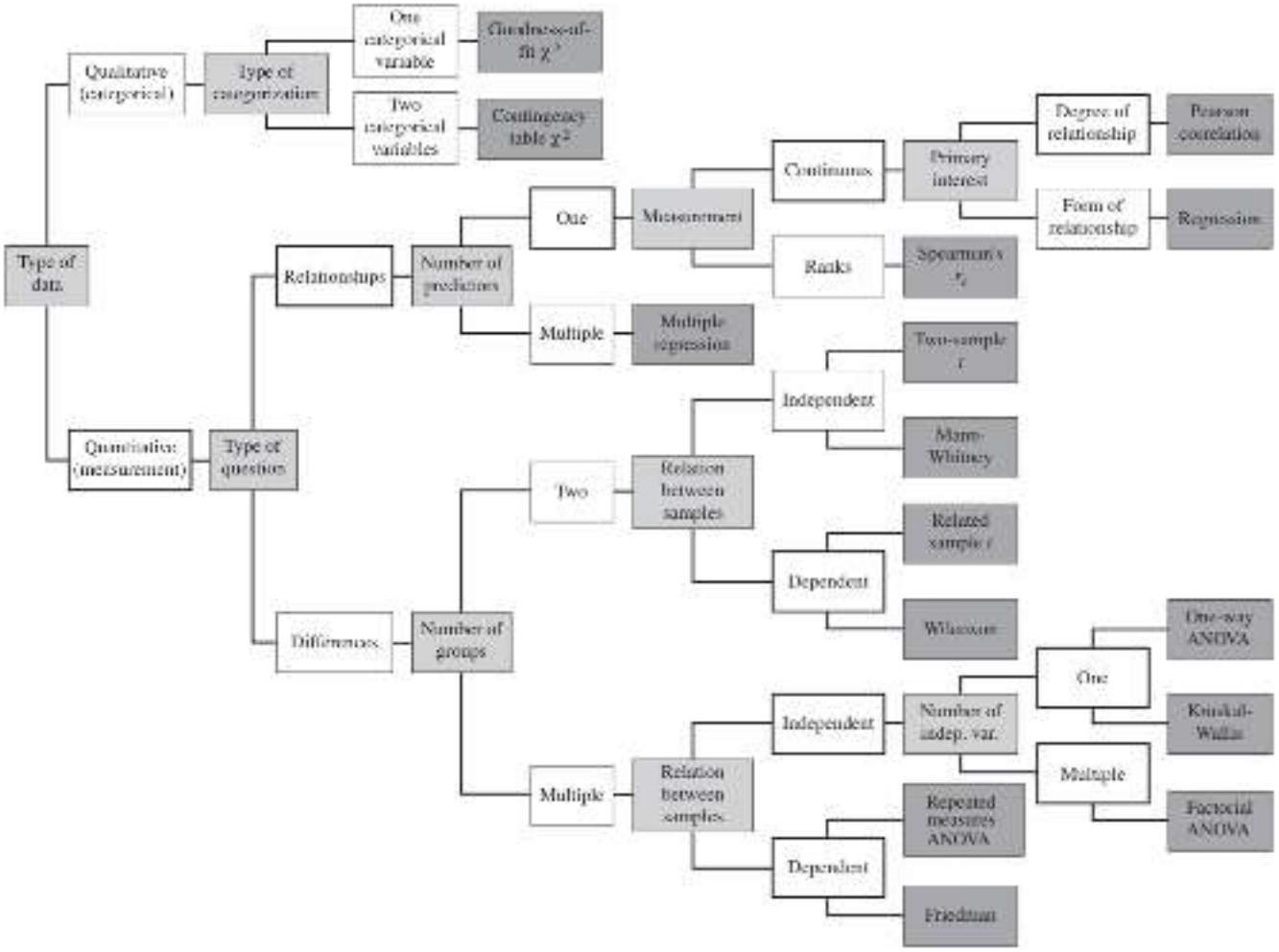
- These results tell us that there is a significant negative correlation between the age of a contestant and how long she stays on the show
- The older the contestant, the earlier she is sent home
- We know that it is a significant negative correlation because of the significant p-value and the sign of the correlation

Visualizing Correlations

plot(bachelor$weeks, bachelor$age, xlab="Weeks Before Dumping", ylab="Age of Contestant")



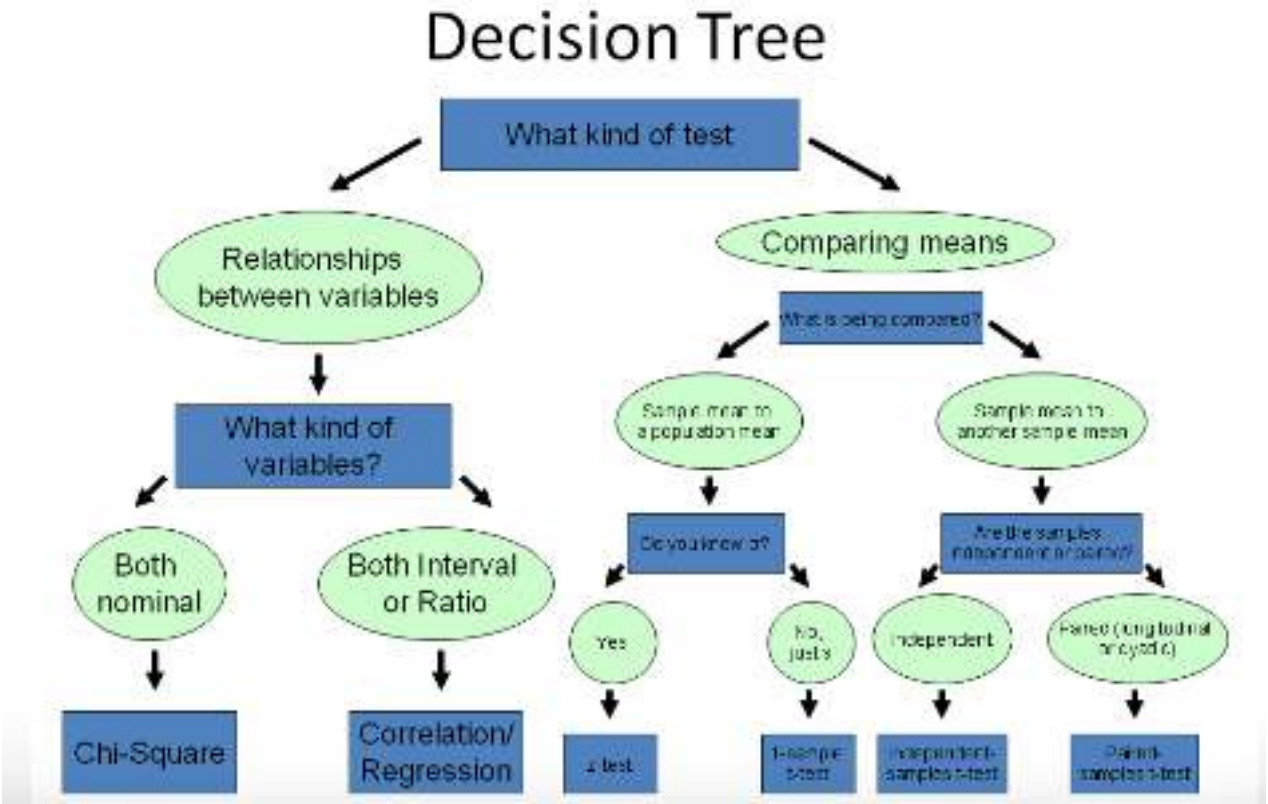
Lots More Ways to Analyze Data



From Howell, D. (2009). *Statistical methods for psychology*. Cengage Learning.

Lots More Ways to Analyze Data

- Repeated measures ANOVA, Split Plot ANOVA, Chi-Square, Multiple Regression and Correlation, Factor Analysis



More Resources on Statistics and R

- **Textbooks on statistics:**
  - Howell, D. (2012). *Statistical methods for psychology*. Cengage Learning.
  - McClave, J. T., & Sincich, T. (2013). *Statistics: Pearson New International Edition*. Pearson Higher Ed.
  - Moore, D. S., Notz, W., & Fligner, M. A. (2015). *The Basic Practice of Statistics*. WH Freeman and Company.
- **Online resources for stats and R:**
  - http://www.nature.com/collections/qghhqm/pointsofsignificance
  - http://www.cookbook-r.com/Statistical_analysis/
  - http://www.statmethods.net/stats/
  - http://www.gardenersown.co.uk/education/lectures/r/correl.htm#nav
  - http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ
- **Choosing the right statistical test:**
  - http://www.graphpad.com/support/faqid/1790/
  - http://www.biostathandbook.com/testchoice.html
- **Reporting basic statistics:**
  - http://my.ilstu.edu/~jhkahn/apastats.html
- **Dangers of and alternatives to null hypothesis testing (NHST):**
  - http://www.stats.org.uk/statistical-inference/Johnson1999.pdf

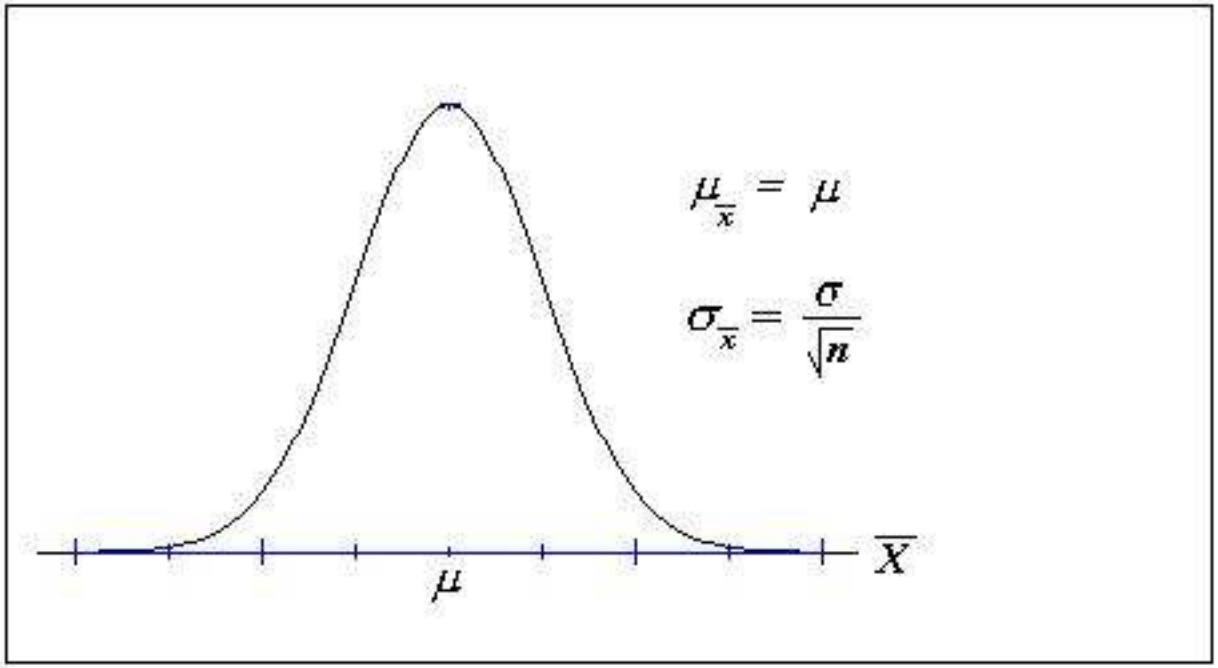
Supplemental - Notation

- Statistic: value of the variable in the sample
- Parameter: value of the variable in the population

	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

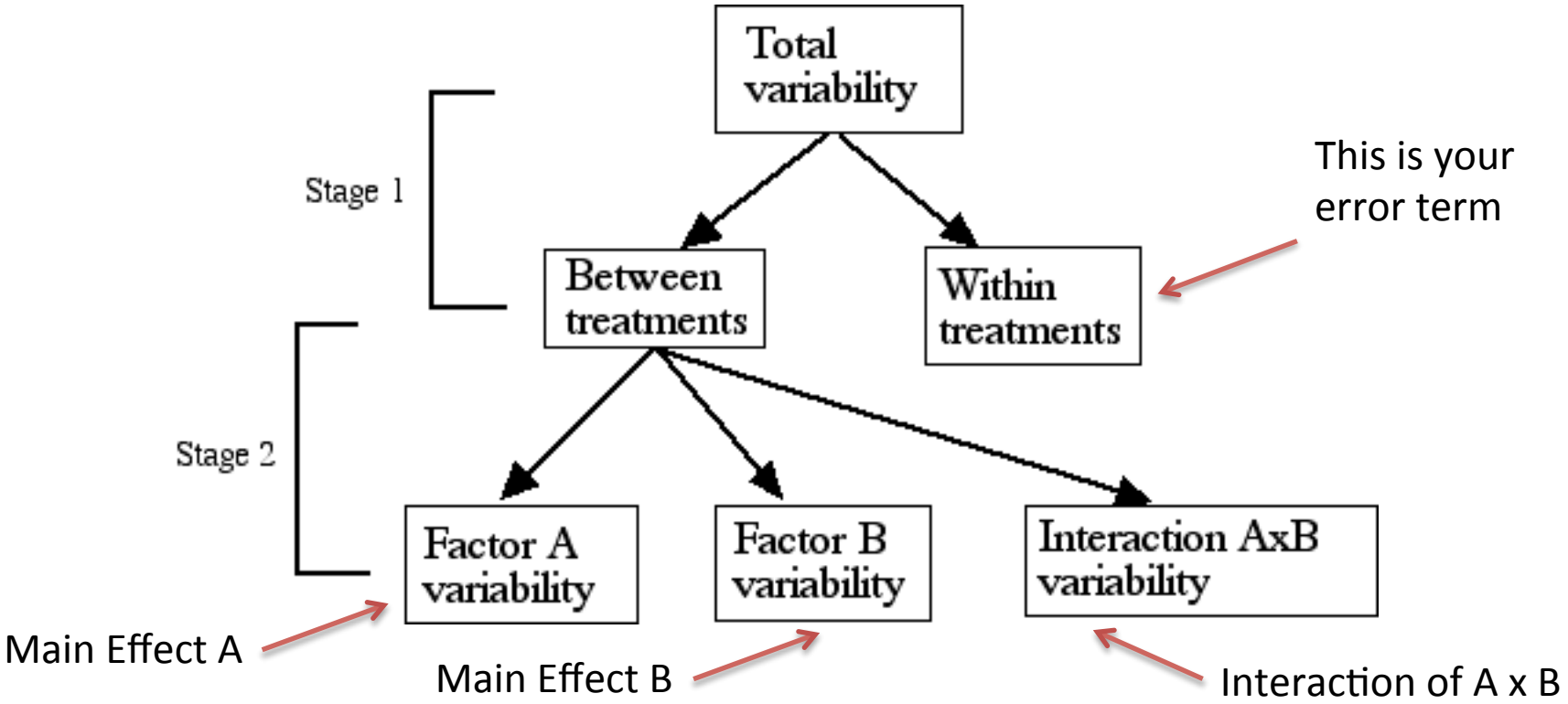
Supplemental – Sampling Distributions

- The mean of the sampling distribution = the mean of the population ($\mu_{\bar{X}} = \mu$) where $\mu_{\bar{X}}$ is the mean of the sampling distribution
- The standard deviation of the sampling distribution is also called the standard error of the mean (SEM)
- The relationship between them is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, where $\sigma_{\bar{X}}$ is the standard deviation of the sampling distribution



Supplemental - ANOVA

- Breakdown of how variance gets partitioned in ANOVA:



Supplemental – Doing a t-test when you've coded for more than two factors

- What if I want to only select two groups in my data for a t-test?
- On the migraine data, to only compare drugs A and B:

```
with(migraine, t.test(pain[drug == "A"],
pain[drug == "B"]))
```

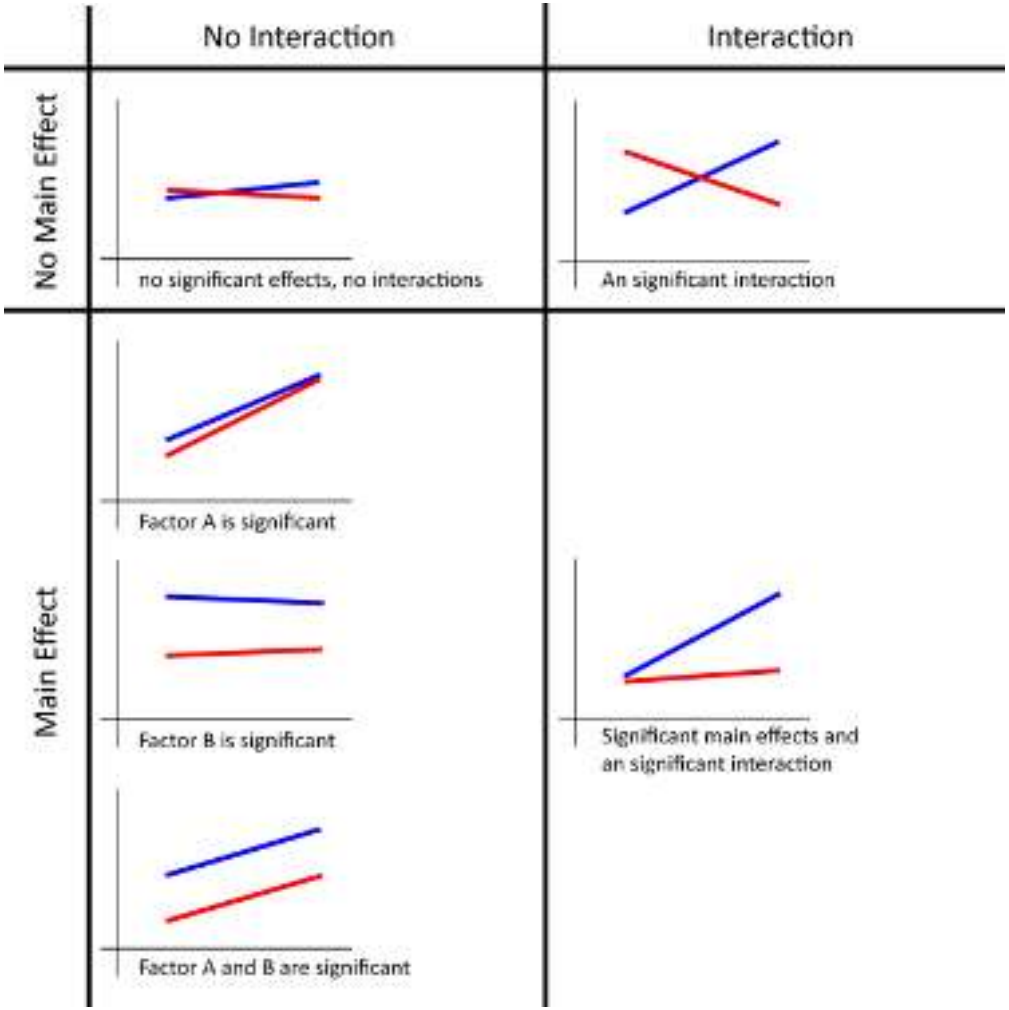
Supplemental – More post hoc tests

- Another popular method for post hoc analyses is the "Bonferroni method"

```
pairwise.t.test(pain, drug, p.adjust="bonferroni")
```

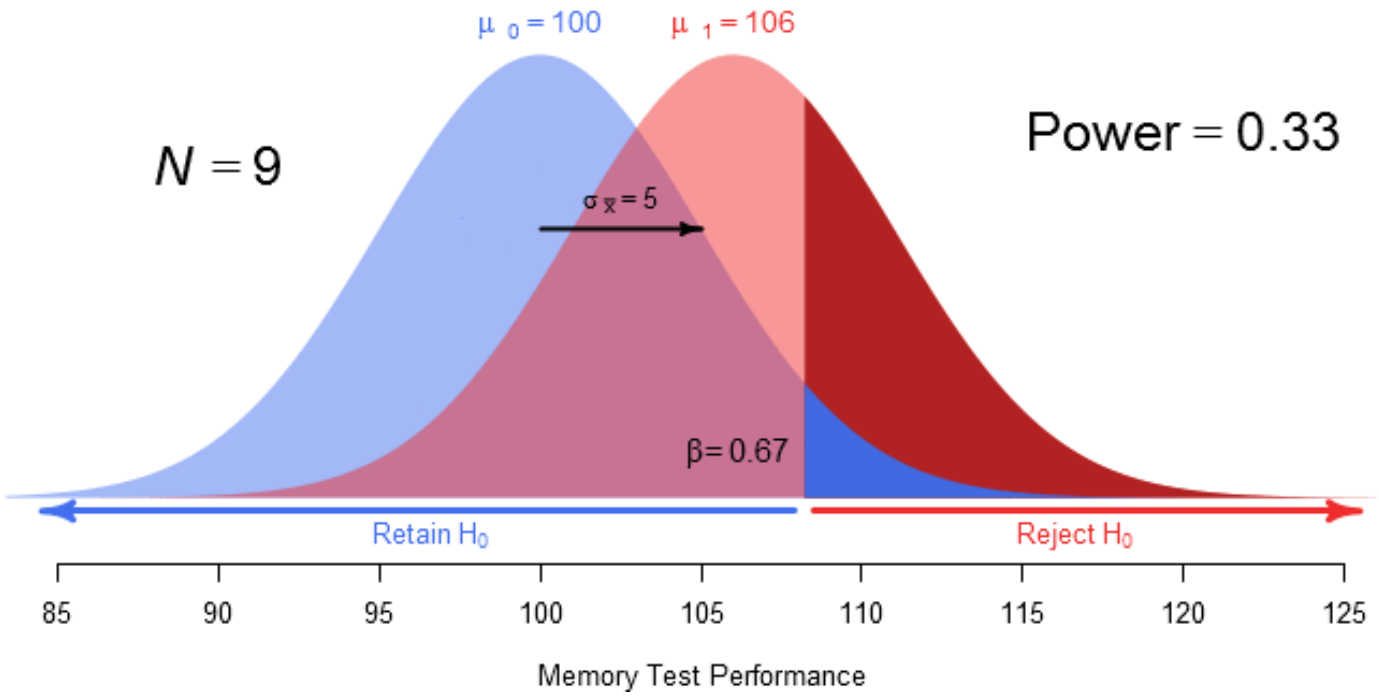
- Basically a bunch of t-tests on all the sub groups, but uses a correction that makes it harder to get a significant p-value (why the hell would I want to do that..??)
- Controls for familywise error – based on the idea that the more comparisons/tests you do, the more likely you are to find a false result (remember our 1 in 20 figure for a p-value cutoff of .05?)
- In this case, gives similar results to Tukey's HSD

Supplemental – More on interactions and main effects



Supplemental – Relationship Between Sample Size and Power

- Statistical power is the probability of correctly rejecting a false null hypothesis
- Check out this link for some great visualizations on the factors that affect power (also explains effect sizes!): http://my.ilstu.edu/~wjschne/138/Psychology138Lab14.html



Supplemental – Error Bars

- Error bars in graphs are usually based on the standard error of the mean (SEM)
- SEM is directly related to both the standard deviation (spread) of the sample and the size of the sample: SEM is calculated as s.d./√*n*
- Unlike standard deviation estimates alone, SEM bars aren't just telling us about the spread of the underlying population – SEM is telling us something about the amount of error in measurement
- Error bars tend to shrink as our sample size increases
- Good to plot at least positive errors bars when creating a bar graph of your ANOVA data
- More resources on error bars (including calculating them in R):
  - http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_semandsdnotsame.htm
  - http://www.biostathandbook.com/standarderror.html
  - http://rcompanion.org/rcompanion/c_03.html