11- A whirlwind tour of statistics

Lorrie Cranor, Hana Habib, and Jessica Colnago,

February 22, 2017

05-436 / 05-836 / 08-534 / 08-734 / 19-534 / 19-734
Usable Privacy and Security

Carnegie Mellon University
CyLab

institute for SOFTWARE RESEARCH

Engineering & Public Policy

Today! Statistics!

- The main idea and building blocks
- Hypothesis testing
- Major tests you'll see
- Non-independent data
- Practice!

Important Note

- In some cases in today's lecture, we will intentionally be imprecise (and sometimes not technically accurate) about certain concepts. We are trying to give you some intuition for these concepts without extensive formal background.

BUILDING BLOCKS

Statistics

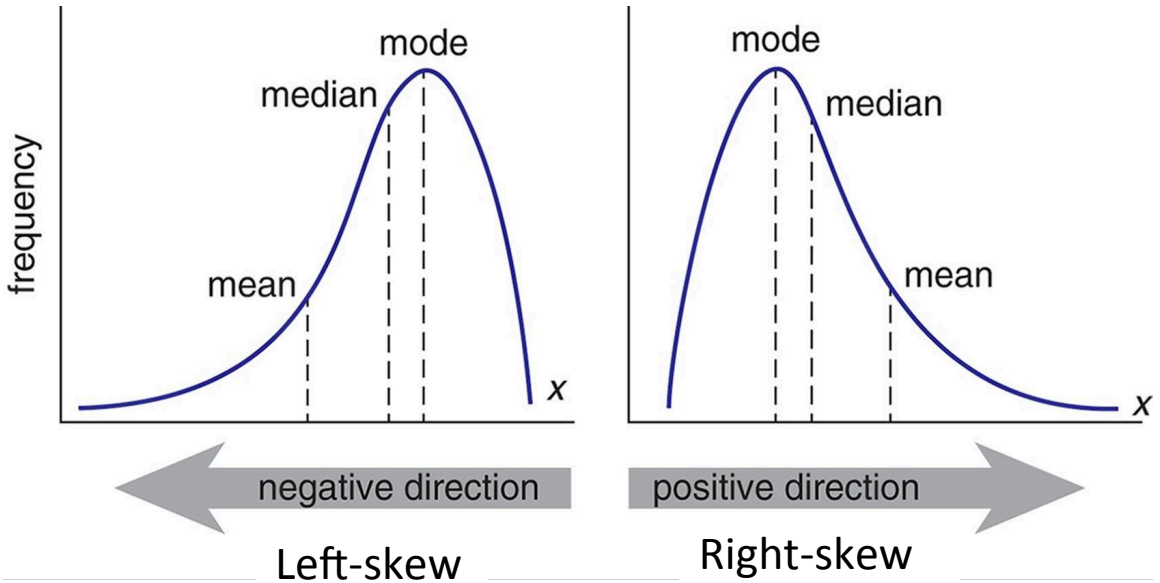
- In general: analyzing and interpreting data
- Statistical hypothesis testing: is it unlikely the data would look like this unless there is actually a difference in real life?
- Statistical correlations: are these things related?

What kind of data do you have?

- Quantitative
  - Discrete
  - Continuous
- Categorical
  - Nominal (no order)
  - Ordinal (ordered)

Exploratory Data Analysis (EDA)

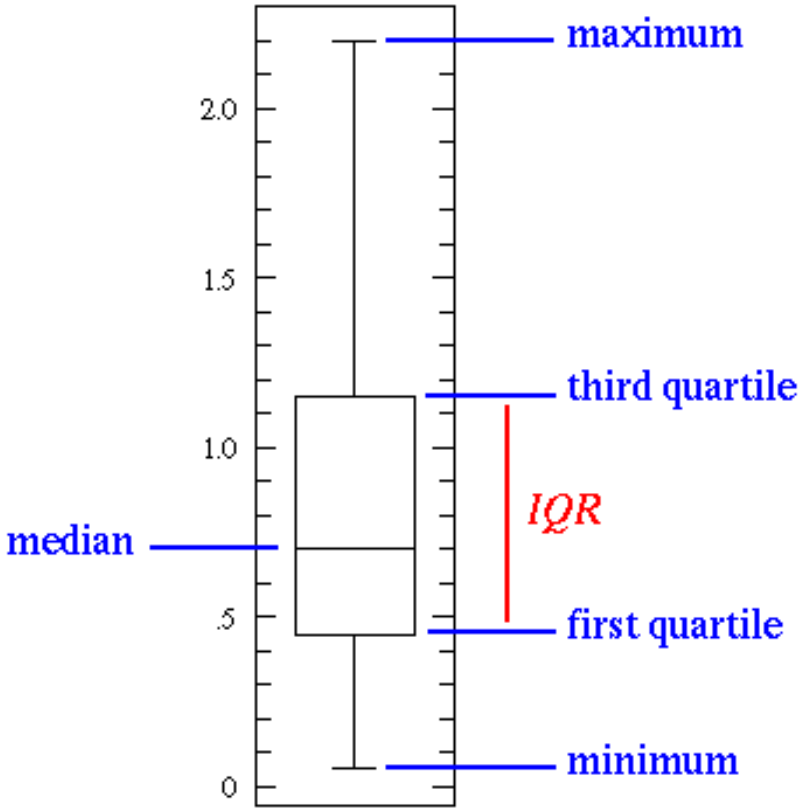
- Shape
  - Kurtosis
  - Skewness
- Center
  - Mean
  - Median
  - Mode



http://schoolbag.info/physics/physics_math/physics_math.files/image790.jpg

EDA Continued

- Spread
  - Standard Deviation
  - Variance
  - Interquartile range



http://www.physics.csbsju.edu/stats/simple.box.defs.gif

Variables

- **Independent**: Variables changed in an experiment that may impact the outcome
- **Dependent**: Variables that are being tested in an experiment
- **Moderator**: Variables that are not part of the experiment but might still impact the outcome

Hypotheses

- **Null hypothesis**: There is no difference
- **Alternative hypothesis**: There is a difference
- You generally either “reject the null hypothesis” (find evidence in support of the alternative hypothesis) or “fail to reject the null hypothesis” (do not find evidence in support of the alternative hypothesis)

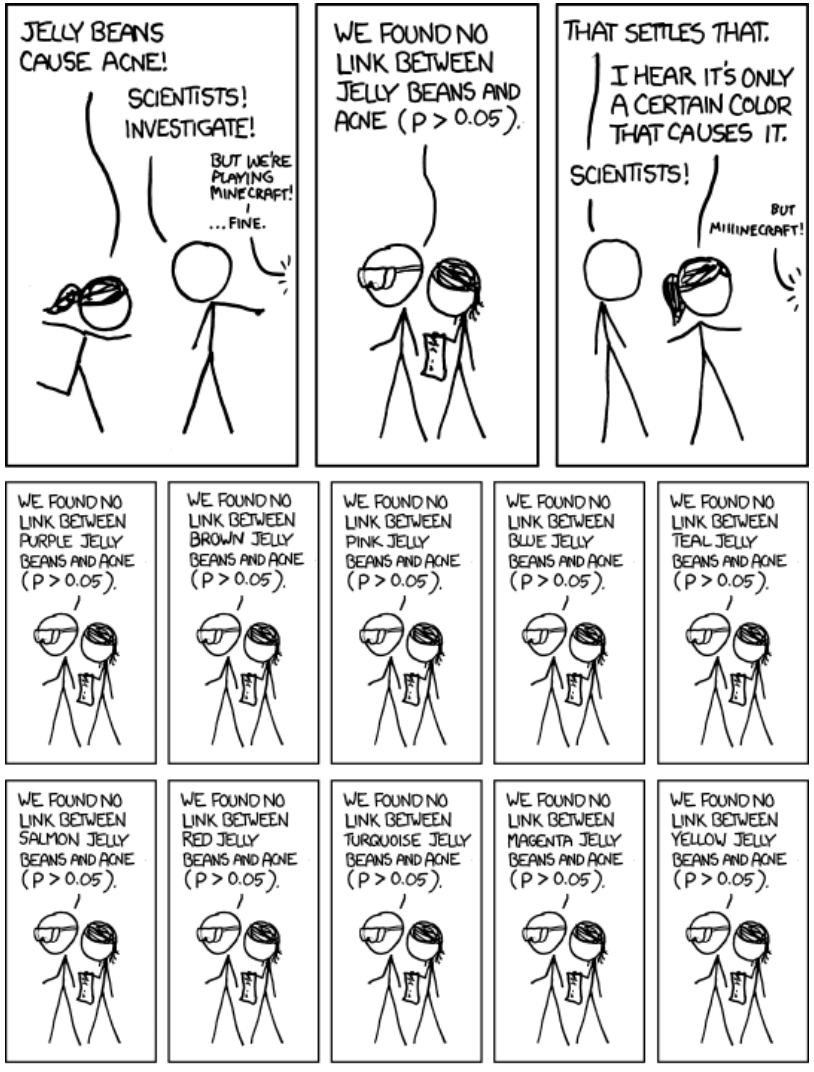
P values

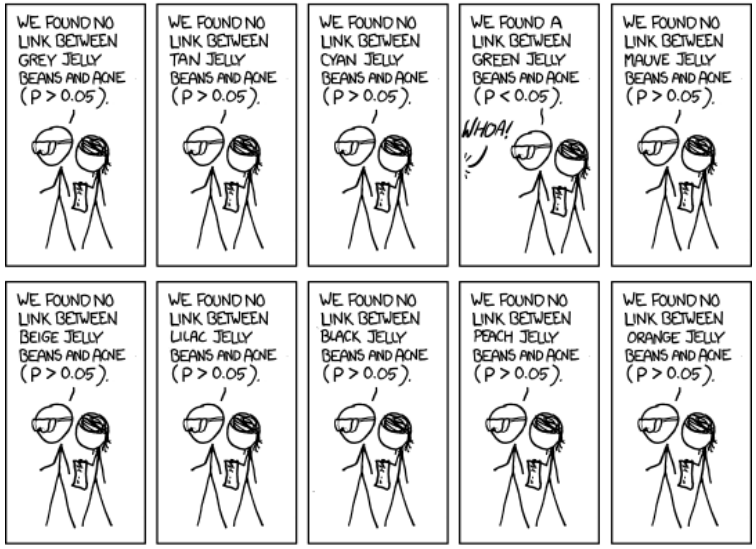
- What is the probability that the data would look like this if there's no actual difference?
- Most often, $\alpha = 0.05$
  - If $p < 0.05$, reject null hypothesis; there is a "significant" difference
  - You don't say that something is "more significant" because the p value is lower

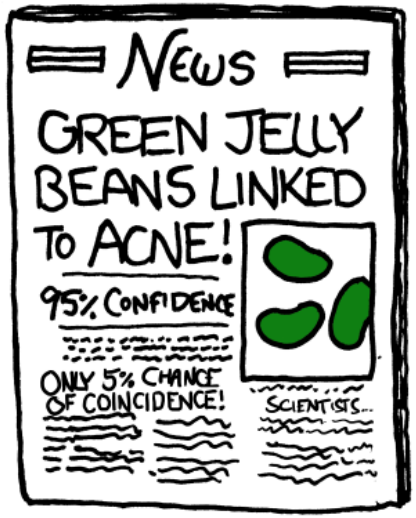
Type I Errors

- Type I error (false positive)
  - You would expect this to happen 5% of the time if $\alpha = 0.05$
- What happens if you conduct a lot of statistical tests in one experiment?

Contrasts







https://xkcd.com/882/

Contrasts

- If we determine that the variables are dependent, we may compare conditions
- Planned vs. unplanned **contrasts**
  - You have a limited number of planned contrasts (depending on the DF) for which you don't need to correct p values.
- Bonferroni correction (multiply p values by the number of tests) is the easiest to calculate but most conservative

Type II Errors

- Type II error (false negative)
  - There is actually a difference, but you didn't see evidence of a difference
- Statistical power is the probability of rejecting the null hypothesis if you should
  - You could do a **power analysis**, but this requires that you estimate the effect size

Threats to Your Experiment

- External validity (generalizability)
- Internal validity (confounding variables)
- Construct validity
- Incorrect Type I errors
- Power

Sources of Variation

- Measurement
- Environmental
- Treatment application
- Subject-to-subject

So much information!

PICKING THE RIGHT TEST

Not all tests are created equal

Different types of dependent and independent variables?

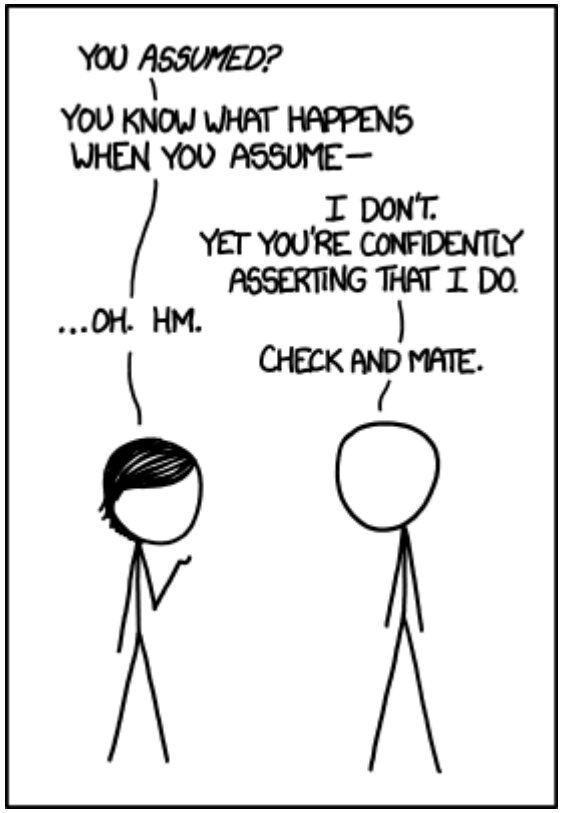
– Different tests!

Different data distributions?

– Different **assumptions**

→Different tests!!

Parametric vs non-parametric



Which tests are we learning about today?

Focusing on parametric tests!

		Dependent Variable	
		Categorical	**Quantitative**
Independent Variable	**Categorical**	Chi-Square Test Fisher's Exact Test	Logistic Regression
	Quantitative	t-Test (paired t-Tests) ANOVA (RM ANOVA)	Correlation Linear Regression

DV: CATEGORICAL
IV: CATEGORICAL

(Pearson's) Chi-squared (χ^2) Test

- Examples:
  - Does the gender (male, female) of the unicorn correlate with a unicorn's favorite color?
  - Does the type of food it eats correlate to its privacy concerns?
- H_0: Variable X factors are equally distributed across variable Y factors **(independence)**
- (Not covered today) Goodness of fit: Does the distribution we observed differ from a theoretical distribution?

Contingency tables

- Rows are one variable, columns the other

CreateAnnoying

Counts:

	0	1
0	161	32
1	165	33
2	168	34
3	170	30
4	164	32
5	161	35
6	167	32
7	129	60
8	128	61
9	154	40
10	153	40
11	154	38
12	142	42
13	121	67
14	124	76

Percentages:

	0	1
0	"83.42%"	"16.58%"
1	"83.33%"	"16.67%"
2	"83.17%"	"16.83%"
3	"85%"	"15%"
4	"83.67%"	"16.33%"
5	"82.14%"	"17.86%"
6	"83.92%"	"16.08%"
7	"68.25%"	"31.75%"
8	"67.72%"	"32.28%"
9	"79.38%"	"20.62%"
10	"79.27%"	"20.73%"
11	"80.21%"	"19.79%"
12	"77.17%"	"22.83%"
13	"64.36%"	"35.64%"
14	"62%"	"38%"

- $\chi^2 = 97.013$, df = 14, p = 1.767e-14

Chi-squared (χ^2) Notes

- Use χ^2 if you are testing if one categorical variable (usually the assigned condition or a demographic factor) impacts another categorical variable
  - If you have fewer than 5 data points in a single cell, use Fisher's Exact Test
- Do not use χ^2 if you are testing quantitative outcomes!

What are Likert-scale data?

- Respond to the following statement: Unicorns are magical.
  - 7: Strongly agree
  - 6: Agree
  - 5: Mildly agree
  - 4: Neutral
  - 3: Mildly disagree
  - 2: Disagree
  - 1: Strongly disagree

What are Likert-scale data?

- Some people treat it as continuous (meh!)
- Other people treat it as ordinal (ok!)
  - You can use Mann-Whitney U / Kruskal-Wallis **(non-parametric)**
- A simple way to compare the data is to "bin" (group) the data into binary "agree" and "not agree" categories (ok!)
  - You can use χ^2 **(parametric)**

DV: CATEGORICAL
IV: QUANTITATIVE

Choosing a numerical test

- Do your data follow a normal (Gaussian) distribution? (You can calculate this!)

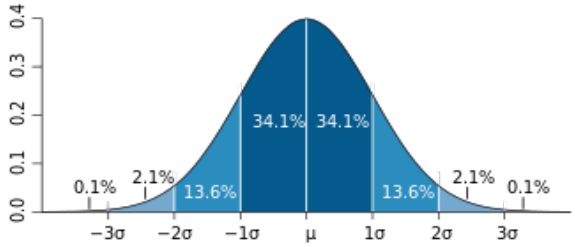


Image from http://www.wikipedia.org

  - If so, use parametric tests. If not, use non-parametric tests
- Does the data set have equal variance?
- Are your data independent?
  - If not, repeated-measures, mixed models, etc.

Independence

- Why might your data in UPS experiments not be independent?
  - Non-independent sample (bad!)
  - The inherent design of the experiment (ok!)
- If you have two data points of unicorns' race completion times (before and after some treatment), can you actually do a single test that assumes independence to compare conditions?

Numerical data

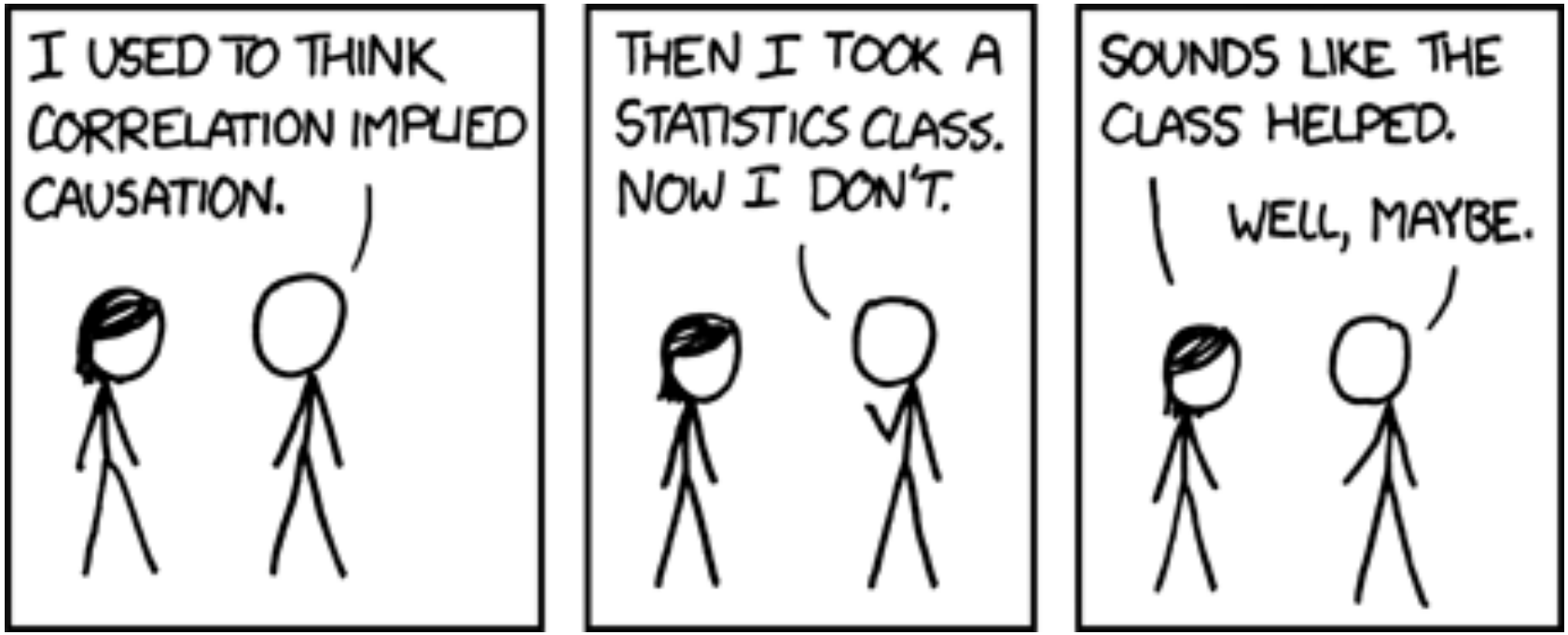
- Are values bigger in one group?
- Normal, continuous data (compare mean):
  - H_0: There are no differences in the means.
    - 2 conditions: t-test
    - 3+ conditions: ANOVA
- Non-normal data / ordinal data:
  - H_0: No group tends to have larger values.
    - 2 conditions: Mann-Whitney U (AKA Wilcoxon rank-sum test)
    - 3+ conditions: Kruskal-Wallis

DV: QUANTITATIVE

Correlation

- Usually less good: Pearson correlation
  - Requires that both variables be normally distributed
  - Only looks for a linear relationship
- Often preferred: Spearman's rank correlation coefficient (Spearman's ρ)
  - Evaluates a relationship's monotonicity
    - always going in the same direction or staying the same

Correlation **DOES NOT** imply causation



Choosing a numerical test

Check the assumptions!

- Equal variance
- Normality
- Independence of errors
- Linearity
- Fixed-x

Regressions

- What is the relationship among variables?
  - Generally one outcome (dependent variable)
  - Often multiple factors (independent variables)
- The type of regression you perform depends on the outcome
  - Binary outcome: logistic regression
  - Ordinal outcome: ordinal / ordered regression
  - Continuous outcome: linear regression

Interactions in a regression

- Normally, outcome = $ax_1 + bx_2 + c + \ldots$
- Interactions account for situations when two variables are not simply additive. Instead, their interaction impacts the outcome
  - e.g., Maybe silver unicorns, and only silver unicorns, get a much larger benefit from eating pop-tarts before a race
- Outcome = $ax_1 + bx_2 + c + d(x_1x_2) + \ldots$

Example regression

- Outcome: completed unicorn race (or not)
- Independent variables:
  - Age
  - Number of prior races
  - Diet: hay or pop-tarts
  - (Indicator variables for color categories)
  - Etc.

WHAT IF THERE IS NO INDEPENDENCE?

Non-independence

- Repeated measures (multiple measurements of the same thing)
  - e.g., before and after measurements of a unicorn's time to finish a race
- Paired t-test (two samples per participant, two groups)
- Repeated measures ANOVA (more general)
  - Extra assumption! → Sphericity

Non-independence

- For regressions, use a mixed model
  - "Random effects" based on hierarchy/group
- Case 1: Many measurements of each unicorn
- Case 2: The unicorns have some other relationship. e.g., there are 100 unicorns each trained by one of 5 trainers. The identity of the trainer might impact a whole class of unicorns' performance.

```
----------
model- binary
**********

Cumulative Link Mixed Model fitted with the adaptive Gauss-Hermite
quadrature approximation with 20 quadrature points

formula: correct ~ gender + chosen + programming + age + alreadydid +

          experiment + chosen * programming + alreadydid * programming +      alreadydid *

          experiment + chosen * experiment + (1 | uid)
          data:    data

           link  threshold nobs logLik  AIC     niter     max.grad cond.H
           logit flexible  1832 -745.62 1565.24 53(13128) 1.23e-05 6.2e+05

          Random effects:
                 Var Std.Dev
          uid 0.7885   0.888
          Number of groups:  uid 223

          Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
          genderI prefer not to answer  0.475650   1.308540   0.363 0.716234
          genderMale                   -0.017708   0.205080  -0.086 0.931192
          chosenb                      -1.739132   0.472334  -3.682 0.000231 ***
          chosenc                       0.644282   0.630716   1.022 0.307014
          chosend                       0.571554   0.600672   0.952 0.341339
          chosene                       1.541800   0.778734   1.980 0.047717 *
          chosenf                      -0.481121   0.510956  -0.942 0.346393
          choseng                      -3.726763   0.503302  -7.405 1.32e-13 ***
          chosenh                      -1.706179   0.479596  -3.558 0.000374 ***
          choseni                      -0.280454   0.530171  -0.529 0.596813
          chosenj                      -0.348918   0.521329  -0.669 0.503313
          programming1                 -0.208038   0.580828  -0.358 0.720213
          age                          -0.017786   0.008671  -2.051 0.040242 *
          alreadydid                    0.173464   0.041030   4.228 2.36e-05 ***
          experiments                   0.139865   0.534377   0.262 0.793527
          chosenb:programming1          0.485281   0.656680   0.739 0.459913
          chosenc:programming1          0.278906   0.893211   0.312 0.754849
          chosend:programming1          1.243753   0.958374   1.298 0.194365
          chosene:programming1         -0.060274   1.029811  -0.059 0.953327
```

Time to review!

KAHOOT!

In groups:

- What statistical analysis would you do?
  - You randomly assign unicorns to have private stalls or public stalls. Does this assignment impact whether they finish their next race?
  - …and does this impact their finishing time?
  - You are analyzing interviews of 10 unicorn trainers and are reporting what these trainers think unicorns say (“neigh,” “ring-ding-ding,” etc.)
  - Do gender, state of residence, and education level impact unicorns’ level of privacy concern?

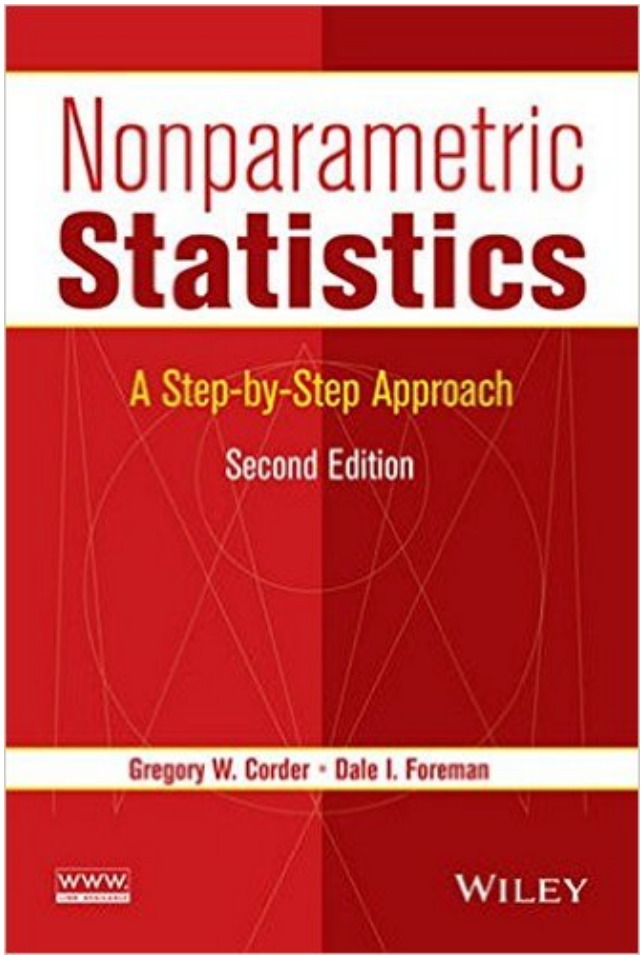
Picking a test

- http://webspace.ship.edu/pgmarr/Geo441/Statistical%20Test%20Flow%20Chart.pdf
- http://abacus.bates.edu/~ganderso/biology/resources/statistics.html
- http://med.cmb.ac.lk/SMJ/VOLUME%203%20DOWNLOADS/Page%2033-37%20-%20Choosing%20the%20correct%20statistical%20test%20made%20easy.pdf

Picking a test/good (basic) reference

https://www.amazon.com/Nonparametric-Statistics-Step--Step-Approach/dp/1118840313/ref=asap_bc?ie=UTF8

First edition available in electronic format from our own library!



(Pearson's) Chi-squared (χ^2) Test

- Examples:
  - Does the gender (male, female) of the pony correlate with a pony's favorite color?
  - Does the type of food it eats correlate to its privacy concerns?
- H_0: Variable X factors are equally distributed across variable Y factors **(independence)**
- (Not covered today) Goodness of fit: Does the distribution we observed differ from a theoretical distribution?

What if you have lots of questions?

- If we ask 40 privacy questions on a Likert scale, how do we analyze this survey?
- One technique is to compute a "privacy score" by adding their responses
  - Make sure the scales are the same (e.g., don't add agreement with "privacy is dumb" and "privacy is smart"… reverse the scale)
  - You should verify that responses to the questions are correlated!

What if you have lots of questions?

- Another option: factor analysis, which evaluates the latent (underlying) factors
  - You specify *N*, a number of factors
  - Puts the questions into N groups based on their relationships
  - You should examine factor loadings (how well each latent factor correlates with a question)
  - Generally, you want questions to load primarily onto a single factor to be confident

Practical questions/associated tests

I split unicorns into living in the forest or living in the clouds, and they each indicated whether or not they liked their new living environment.

Does the assigned system impact whether or not they liked it?

Practical questions/associated tests

I measured how long unicorns stay in their assigned living space from each unicorn.

Do unicorns that live in the cloud stay more time at home than those that live in the forest?

Practical questions/associated tests

I measured how long unicorns stay home and their magicalness score (1 to 100).

Are these values related to each other?

Practical questions/associated tests

I measured how long unicorns stay home and their magicalness score, age and weight.

I’m curious what input factors (if any) impact the output.