Research Methods I

Rob Semmens
CS 376

With significant input from Michael Bernstein and Dan Schwartz

Folks, do me a solid…

- Put your last names in the file name. And do this on every other assignment in your class-taking lives.
  - Your TA’s prefer .pdf files

Abstracts

- Project Abstract Grading is done
- Let’s talk about how to talk about related work…
  - “Previous research has shown that novice undergraduate students do not draw visualizations to help them in faux medical diagnosis problem solving task.”

  <here’s the key>

  “This study extends this work by demonstrating that they will also not do this in an authentic engineering project scheduling task found in many engineering textbooks. Furthermore…”

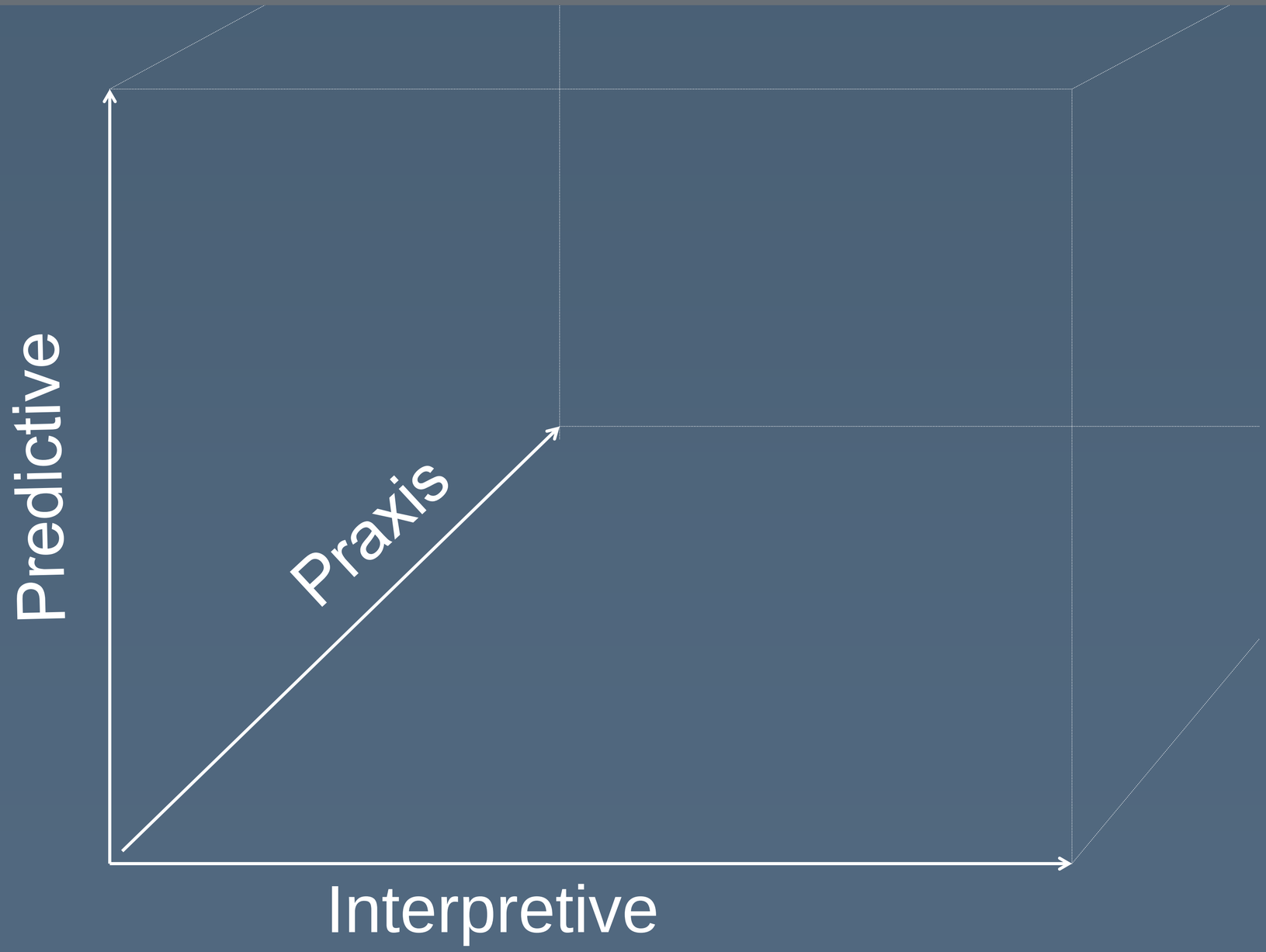
- Watch your language!
  - competence != efficacy != ability (perhaps)
  - Define the terms you need to, then stick with them throughout

Scoping

Goals of Research

- Social science differs from physical science.
- Argument over linear accelerators.
  - Creating a world that only exists in accelerator.
    - Therefore, it is not “real.”
  - A strange argument for human affairs.
  - We largely create our social world.
    - Legal system, economics, schools.
- More types of social-scientific knowledge.

Three Vectors of Social Scientific Knowledge



Predictive Knowledge

- Goal
  - Ascertain the regularities of social reality.
- Criterion
  - Identification of conditions that replicate a given outcome.
- Proto-typical instances:
  - Finding correlation between achievement and SES
  - Forecasting teacher retention
  - Determining robustness of an instructional treatment
  - Isolating a cause of autism
- Typical Issues
  - Hidden factors * causality * generalization

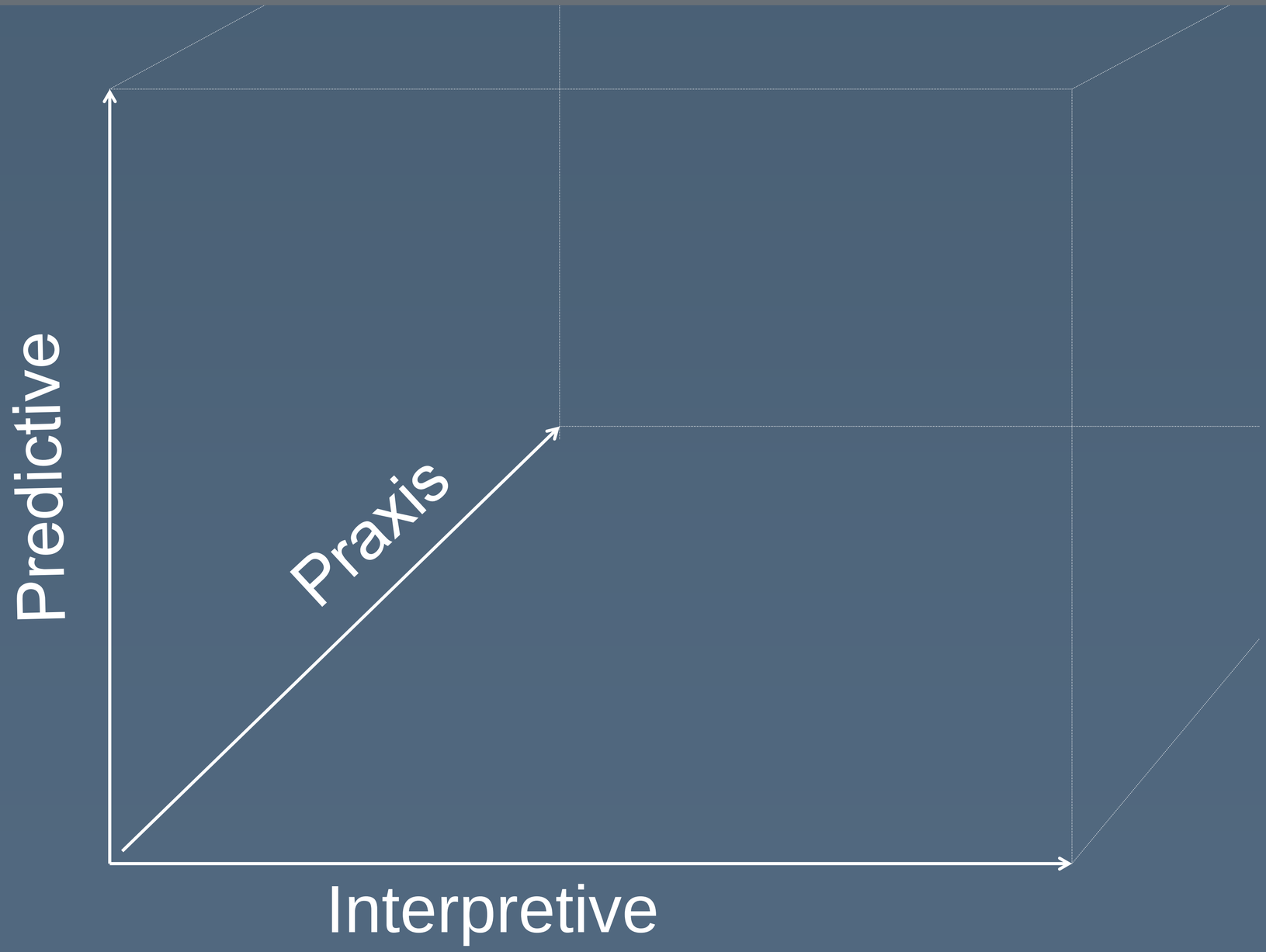
Interpretive Knowledge

- Geertz, 1973
  - *Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning.*
- Goal
  - Insight on the meanings and signs that organize social reality. (Like understanding a text rather than predicting an outcome.)
- Criterion
  - An account of words and deeds that could eventually be accepted by the subject.

Interpretive Knowledge

- Prototypical instances
  - Describing how different cultures experience school.
  - Detailing the construction of classroom identities.
  - Contrasting children’s views of mathematics.
  - Characterizing a moment of epiphany.
  - Detailing moment-to-moment interactions.
- Issues
  - Vantage/assumptions of researcher * Do readers of work “experience” same interpretation.* Another subtext?
- The interpretive endeavor takes a special knack:
  - *I never knew how badly you understood me until you started setting me up on blind dates.*

Three Vectors of Social Scientific Knowledge



Praxis Knowledge

- G. H. Mead, 1899
  - *In society, we are the forces that are being investigated, and if we advance beyond the mere description of the phenomena of the social world to the attempt at reform, we seem to involve the possibility of changing what at the same time we assume to be necessarily fixed.*
- Goal
  - Determine which aspects of social reality are fixed and which are mutable.
- Criterion
  - Evidence of precipitating a new social reality.

Praxis Knowledge

- An unusual view:
  - Praxis knowledge shows that what was assumed or thought to be fixed can be changed.
  - In other words, a theory is true to the extent that it can change the world to fit it.
- Examples
  - Proactive political theories (communist manifesto)
  - The contact hypothesis and busing
  - Demonstrations of excellence in downtrodden places.
  - School reform effort in New York City.
  - Design experiments
- Issues
  - Really new? * Really a change?

Praxis is explicitly value laden

- If the goal is change, praxis is most direct.
  - Interpretation and prediction leave change to others.
    - “I reveal interpretations. My papers will make others change.”
    - “I find the laws, let the engineers decide how to use them.”
    - These require an unstudied link between theory and change.

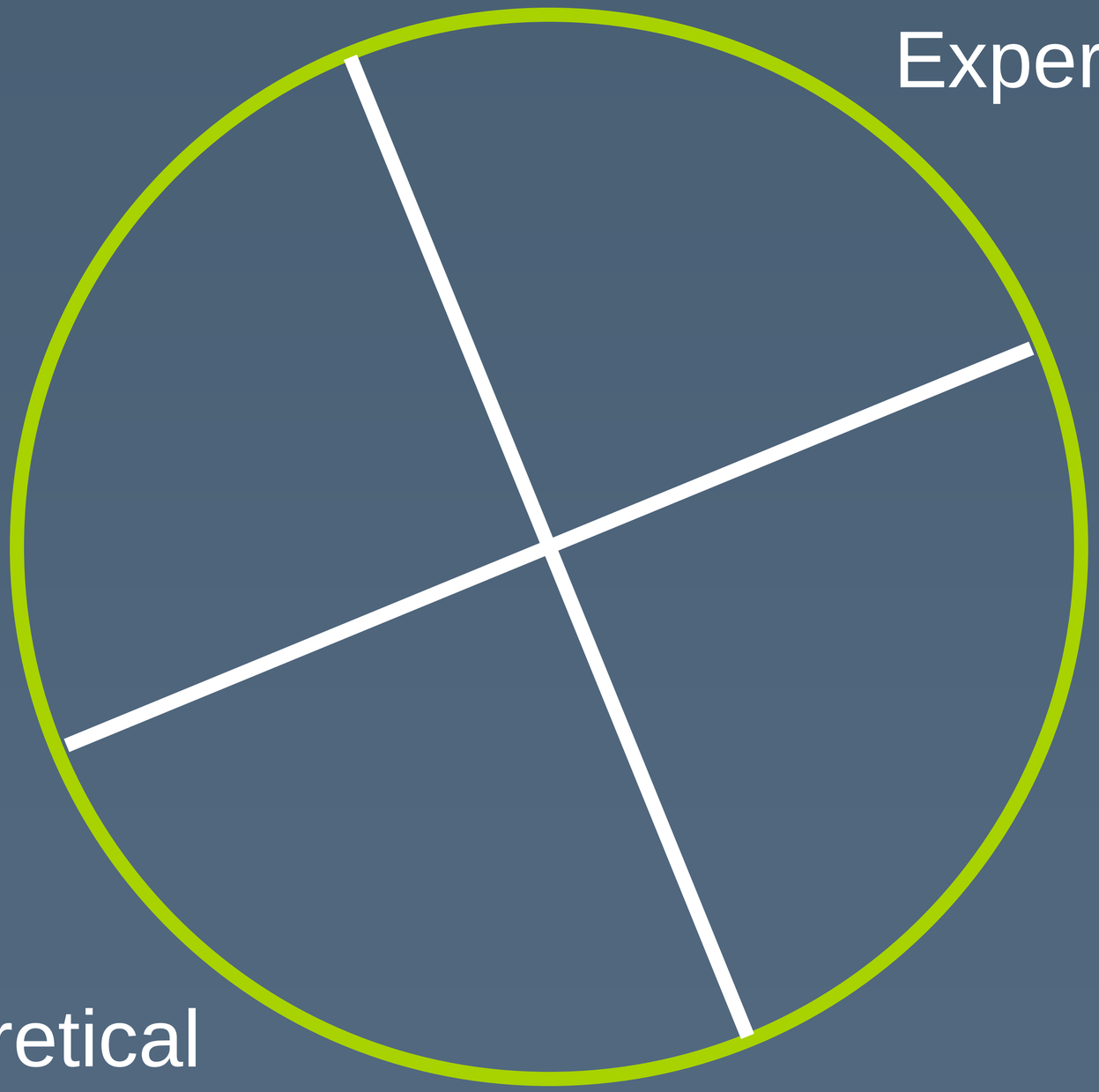
- If the goal of research is change, then it has the burden to decide what change to make.

Value Laden Research

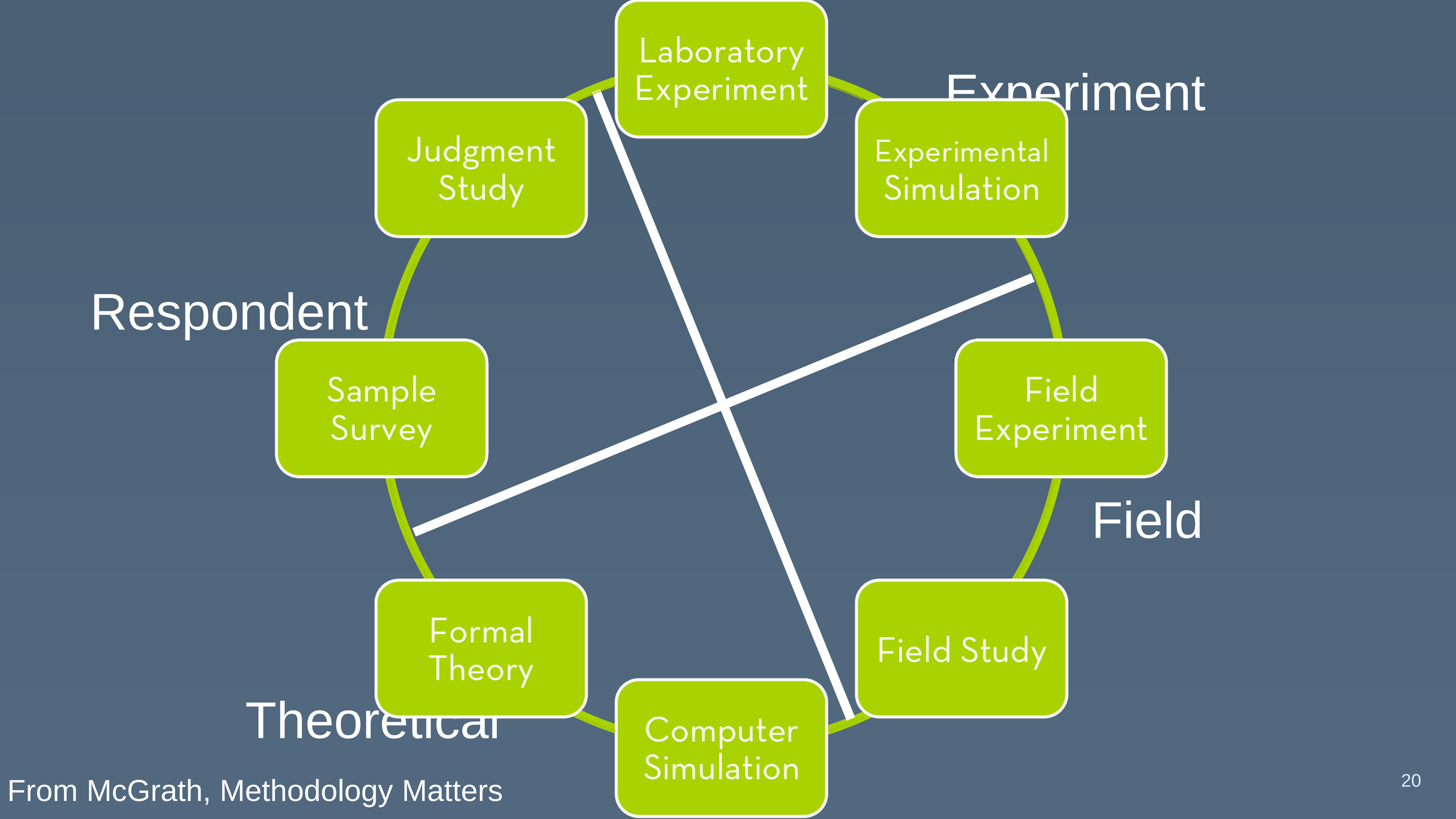
- Doesn’t wanting and trying to make a particular outcome violate scientific principles of being dispassionate and objective?
  - No. Asserting values and desired outcomes does not override **the requirements of truth and integrity**.

- Doesn’t this mean you are imposing your values?
  - Yes! And this is where vigilance is necessary.
    - The IRB helps
  - Know the setting
  - Return value to participants
  - Don’t waste people’s time—pilot!
  - Look for negative consequences

Scoping



From McGrath, Methodology Matters



From McGrath, Methodology Matters

Choosing a Method

The research question drives the method!

- Interpretive
  - Description of behavior
  - May build to theory
  - Variables may not be identified before data collection
  - Statistics often not used
- Predictive/Praxis
  - Leads to Operational Definitions of Hypotheses
  - Tests theory
  - Variables and levels identified before data collection
  - Statistics almost always used

- Methods are technical details not belief systems!

However, method triangulation

- All methods are flawed
- Thus, your argument becomes far stronger if you can demonstrate the same phenomenon using multiple methods
  - Complement your statistics with semi-structured interviews
  - Complement qualitative work with primary source evidence or log data

Becoming A Bartender

The Role of External Memory Cues
by King Beach

- selected an occupation which intuitively seemed to place heavy demands on memory
- Enrolled in the course!
  - Observations
  - Interview with Instructors
- This “motivated the construction of the experimental hypothesis”

Study Design

Dependent Variables

- Time to complete each drill
- Number of common ingredients poured at the same time
- Number of drink errors (wrong drink, but made correctly)
- Number of ingredient errors (right drink, but made incorrectly)
- Frequency of overt rehearsal
- Looks at mixology book
- Looks into glasses
- Glass position

Findings

- Instructors faster and more accurate than graduates, who were faster and more accurate than students
- Counting backwards caused accuracy problems for students, but not graduates or instructors
- Instructors poured the common ingredients more frequently
- Graduates made many more errors using the black glasses instead of the regular glasses, no difference for novices or experts
- Graduates looked in the glass much more when they were black, novices and experts did not

Variables

- Independent
  - What you manipulate
- Also called a predictor variable, or stimulus, or factor

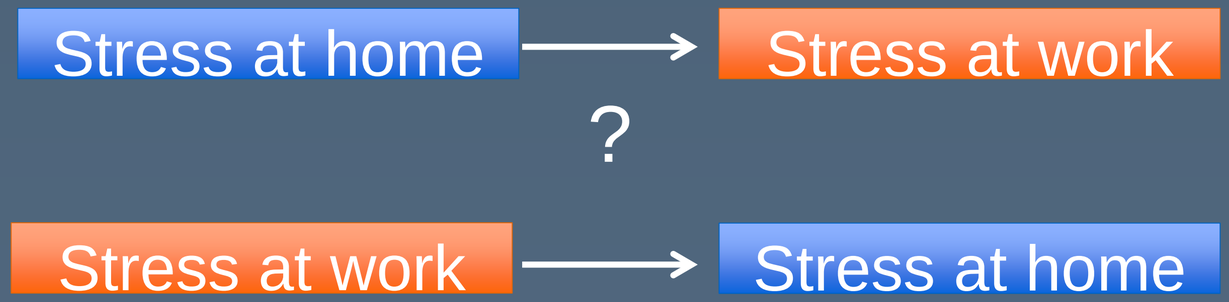
- Dependent
  - What happens as the result of the manipulation
- Also called the response variable

Variables & Operational Definitions

- Variable
  - Any event, situation, behavior, or characteristic that varies
  - Must have two or more *levels*
    - Continuous vs. Categorical
- Operational definition
  - A set of procedures used to measure or manipulate a variable
- Construct validity
  - Does the operational definition fit the variable in question?
  - Are you measuring what you think you're measuring?

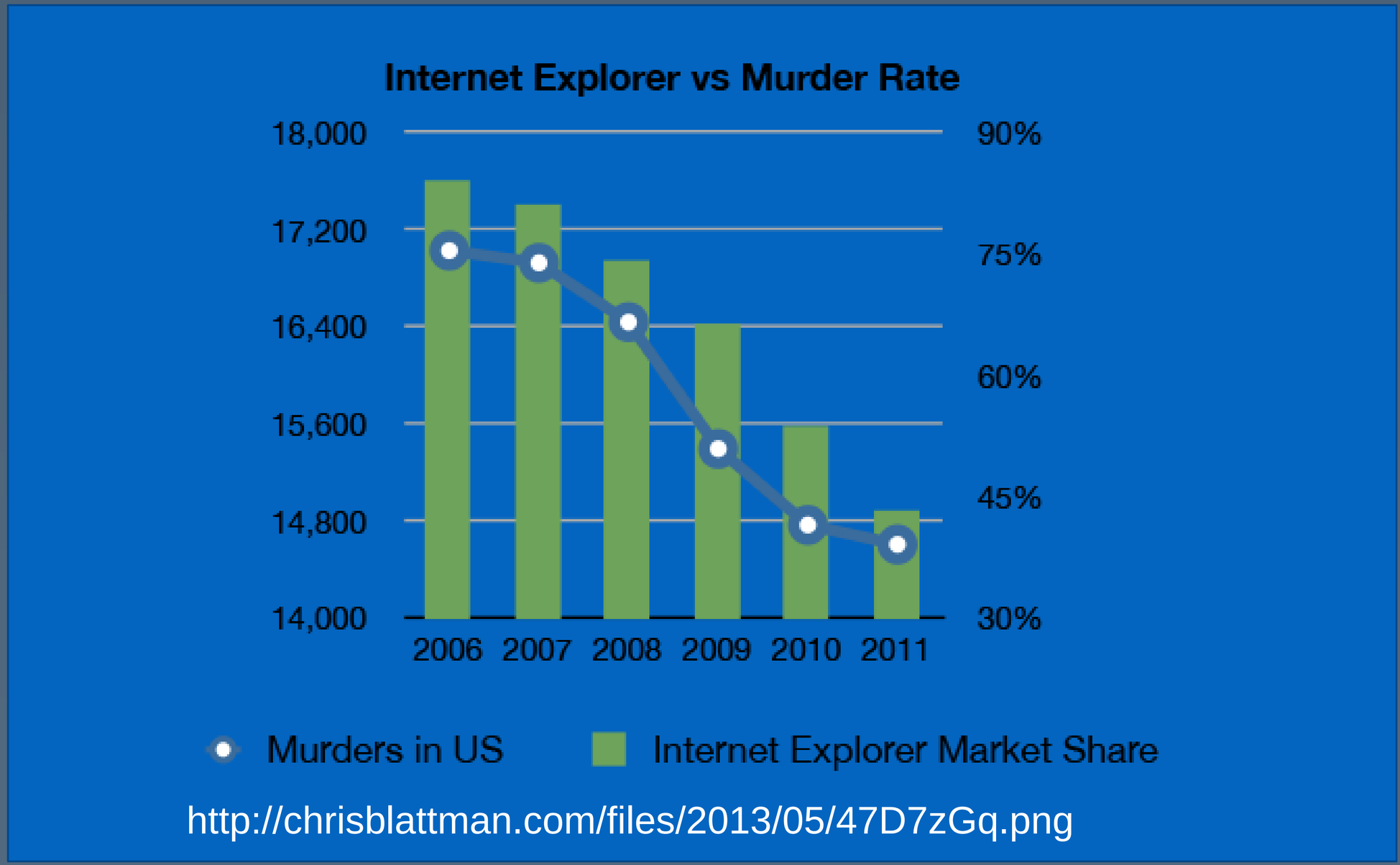
Nonexperimental methods

- Correlational studies: still predictive
- No claim direction of cause and effect



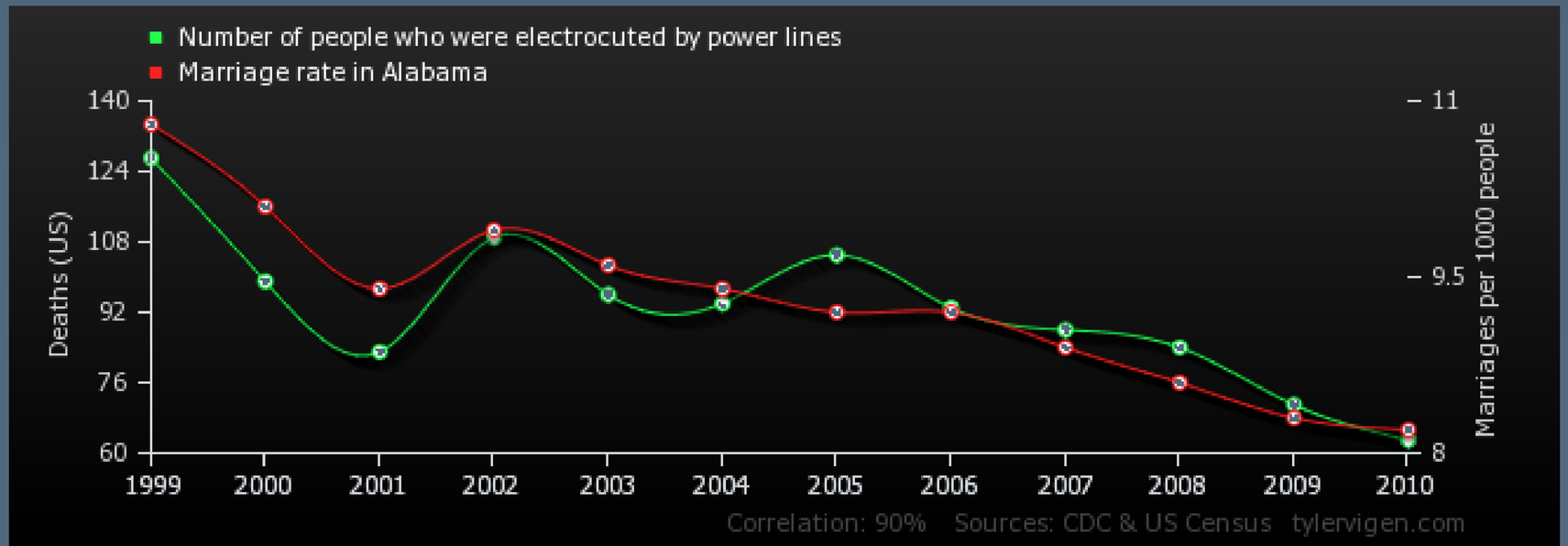
- There still some reason to think that the two variables are related…

Correlated variables are not causally related

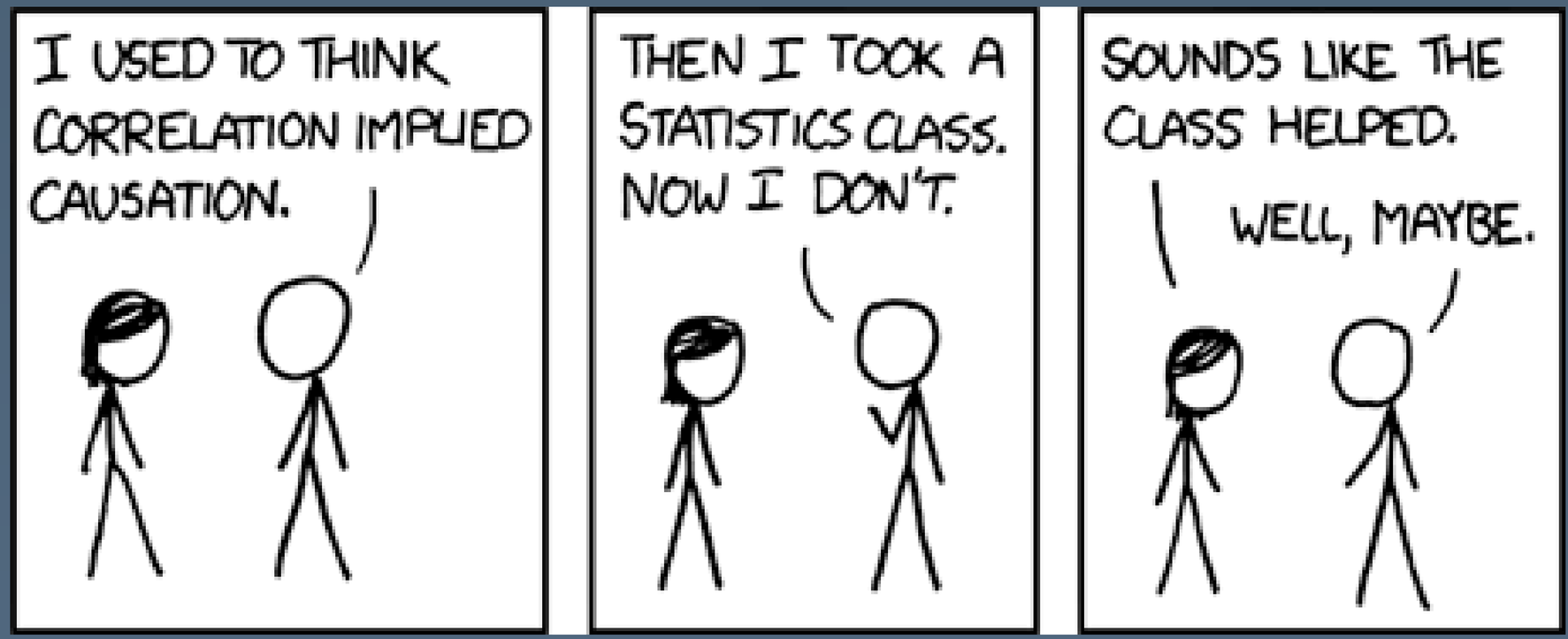


http://chrisblattman.com/files/2013/05/47D7zGq.png


US crude oil imports from Norway
Drivers killed in collision with railway train
Millions of barrels
Deaths (US)
140
120
100
80
60
40
20
0
90
81.67
73.33
65
56.67
48.33
40
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 95%
Sources: Dept. of Energy & CDC
tylervigen.com


Number of people who were electrocuted by power lines
Marriage rate in Alabama
Deaths (US)
Marriages per 1000 people
140
124
108
92
76
60
11
9.5
8
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Correlation: 90%
Sources: CDC & US Census
tylervigen.com

Correlated variables are not causally related



http://xkcd.com/552/


SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
DilbertCartoonist@gmail.com
Dilbert.com
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
11-28-11 © 2011 Scott Adams, Inc./Dist. by Universal Uclick
CORRELATION DOES NOT IMPLY CAUSATION.
SIGH

The third variable problem

- Another extraneous variable may be related to both variables in question.
- The third variable is a **confounding factor** or **confound**
- When we can identify one that will for sure have an effect on the DV, we **control** for it

Experimental methods

- Important difference between experimental and non-experimental studies?

- Randomization
  - Assigning participants to groups at random
  - Helps to alleviate the third variable problem because an extraneous variable is just as likely to affect one group as the other group
  - Random does not mean chaos! Random also does not mean the experimenter's best guess at creating random groups!

Framing an evaluation

- The difficulty: defining and isolating the construct that you are trying to maximize
- It is tempting to aim for something easy: time, task completion, number of clicks
- But, testing the easily quantifiable could miss the point.

Construct Validity—It pays the bills!

The construct validity of organizational commitment has recently been investigated in several studies. The authors of these studies have concluded that organizational commitment is a valid construct, sufficiently distinct from job satisfaction. Our re-analysis of data reported in these studies, however, suggests that the construct validity evidence is unconvincing. Analysis of meta-analytic results cast further doubt on the discriminant validity of organizational commitment as typically measured. Based on these findings, suggestions for future research are offered.

Types of Measures

- Counts, categorical, binomial
  - People who did it or didn’t do it
- Ordinal
  - 1st place, 2nd place, 3rd place
  - Likert scale
- Interval/ratio
  - Test score
  - Reaction time
  - Likert scale again (Oh, I hate you survey!)
- This is determined in study design, (not after data collection)!

Scoping

What Test to Run?

	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

http://yatani.jp/teaching/doku.php?id=hcistats:start

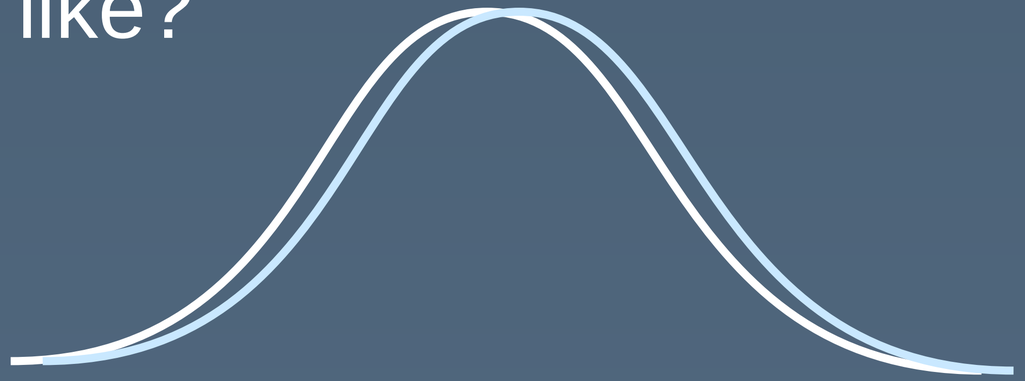
Always follow every step!

1. Visualize the data
2. Compute descriptive statistics (e.g., mean)
3. Remove outliers >2 standard deviations from the mean
4. Check for heteroskedasticity and non-normal data
   - Easiest to check by visualizing the data
   - If there’s a problem, try a log, square root, or reciprocal transform
   - Our tests are typically robust against non-normal data, but not against heteroskedasticity
5. Run statistical test
6. Run any posthoc tests if necessary

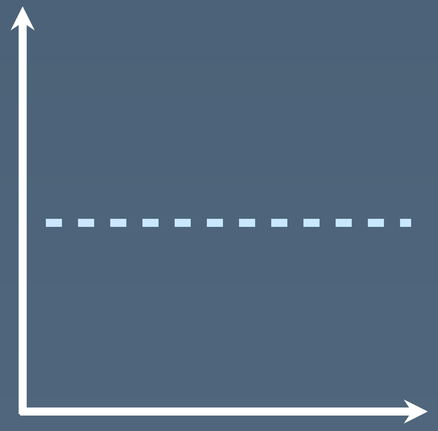
Hypothesis Testing

Anatomy of a statistical test

- If your change had no effect, what would the world look like?



No difference in means

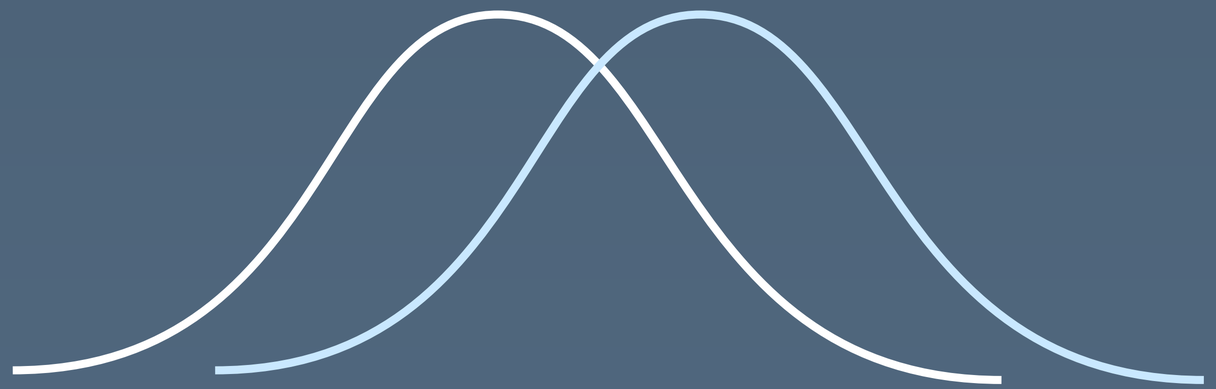


No slope in relationship

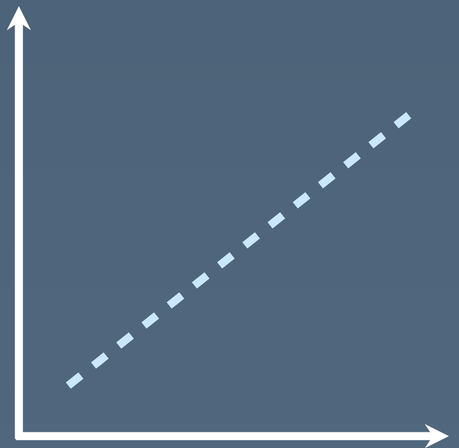
- This is known as the null hypothesis

Anatomy of a statistical test

- Given the difference you observed, how likely is it to have occurred by chance?



Probability of seeing a mean difference at least this large, by chance, is 0.012



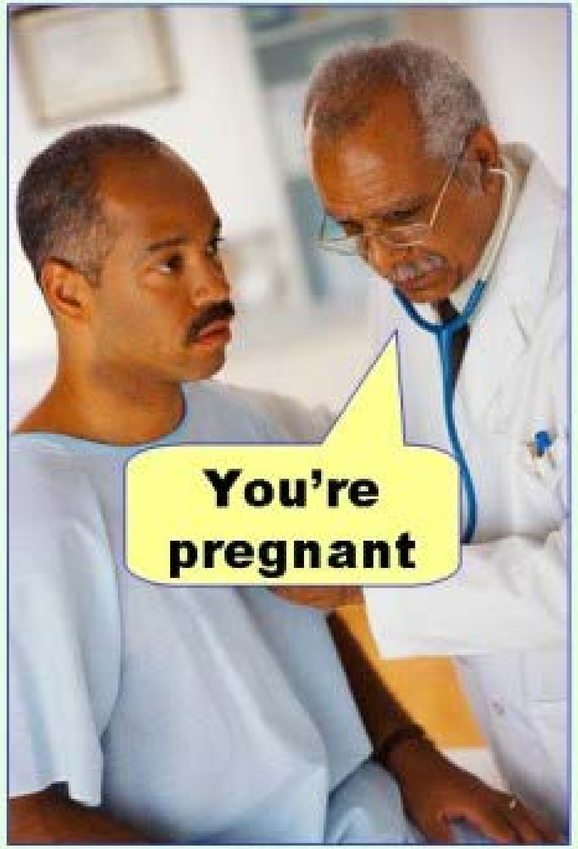
Probability of seeing a slope at least this large, by chance, is 0.012

Errors

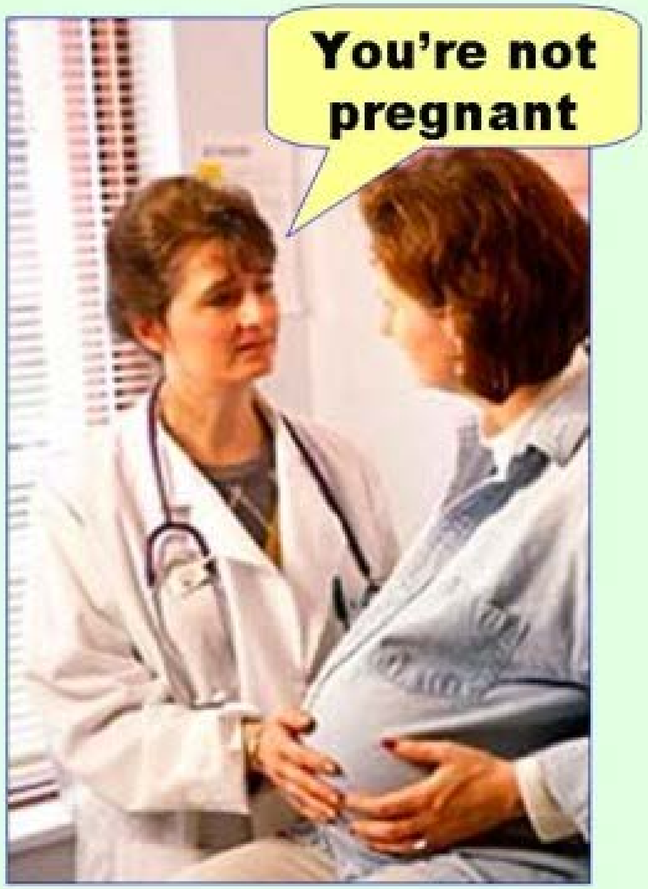
Difference detected? \ Difference exists?	Y	N
Y	True positive	Type 1 error publish false findings
N	Type 2 error get more data?	True negative

Errors

Type I error
(false positive)



Type II error
(false negative)



p-value

- The probability of seeing the observed difference by chance
  - In other words, P(Type I error)
- Typically accepted levels: 0.05, 0.01, 0.001

Comparing two populations: counts

Count or occurrence data

- “Fifteen people completed the trial with the control interface, and twenty two completed it with the augmented interface.”

	control	augmented
success	5	22
failure	35	18

Pearson’s chi-square test for independence

- Determine the expected number of outcomes for each cell

	control	augmented	total
success	5	22	27
failure	35	18	53
total	40	40	80

- Expected is (row total)*(column total) / overall total.
  - Upper left: expected is 27*40/80 = 13.5

Calculating a chi-square statistic

$$\chi^2 = \frac{(observed - expected)^2}{expected}$$

e.g., $(5\text{-}13.5)^2 / 13.5 = 5.35$

Sum this value over all possible outcomes

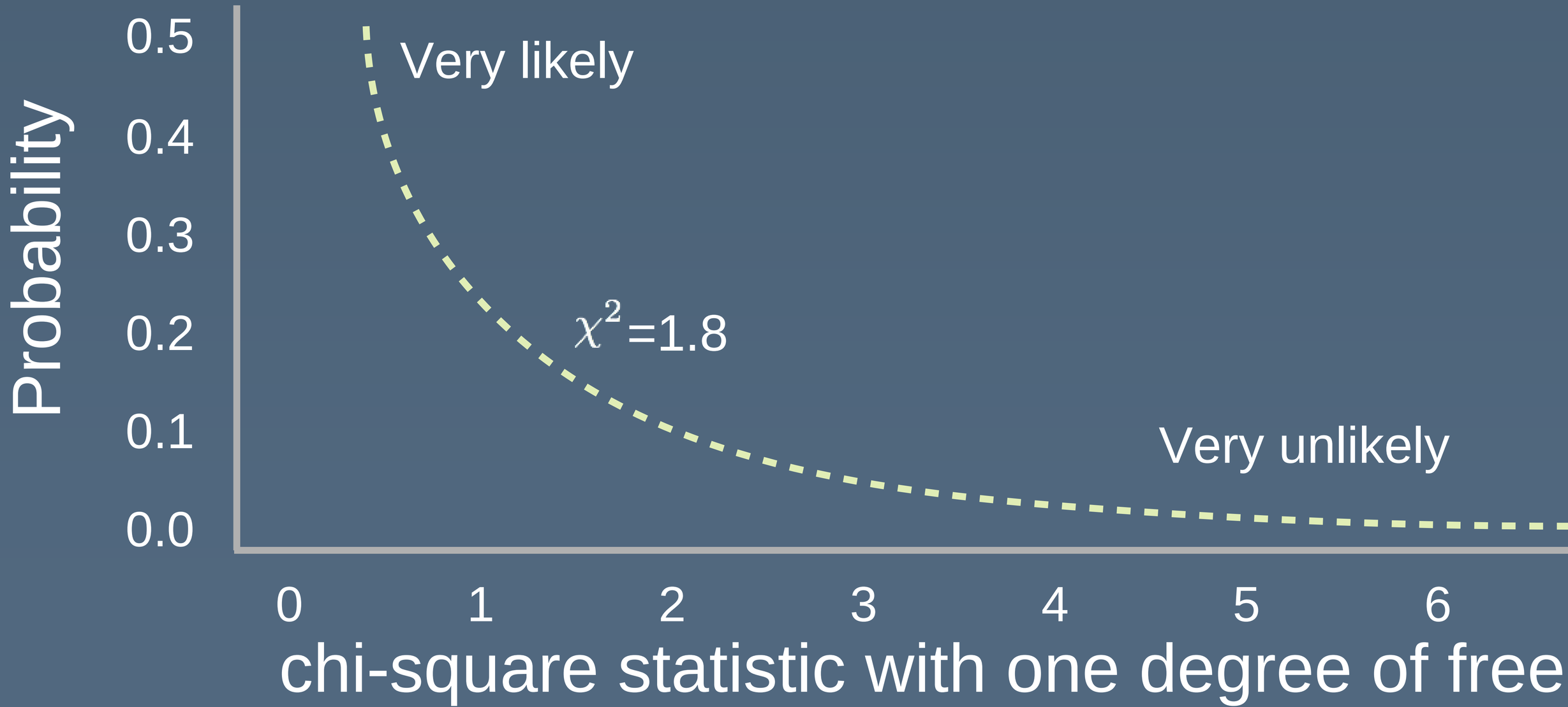
How many degrees of freedom?

- If we know there are a total of 40 participants…

5	???
???	18

- We get (rows - 1) * (columns -1) degrees of freedom.
  So, if it’s a two-by-two design, one degree of freedom.

Result: chi-square distribution



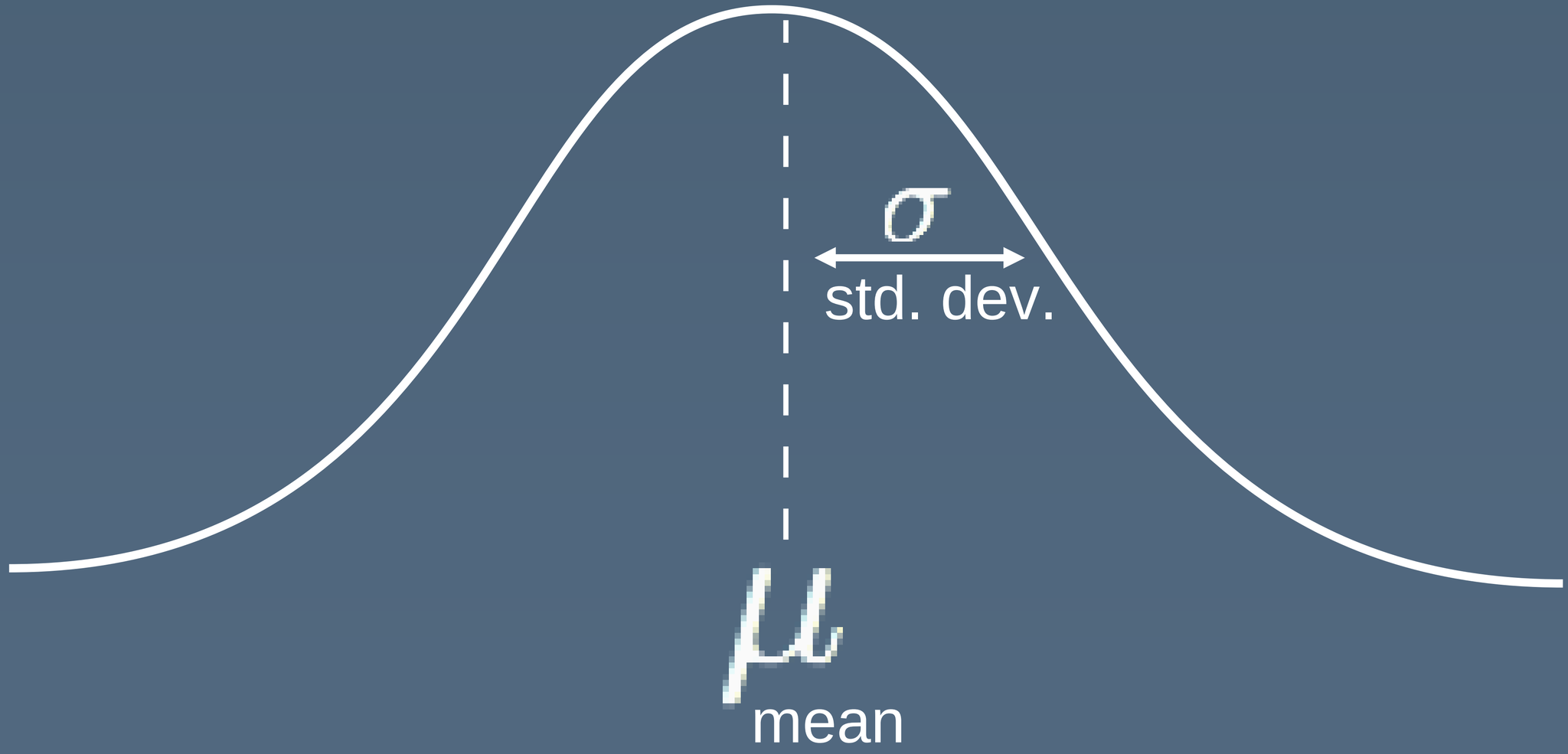
Pearson’s chi-square test for independence

`chisq.test` (HCI R tutorial at http://yatani.jp/HCIstats/ChiSquare`

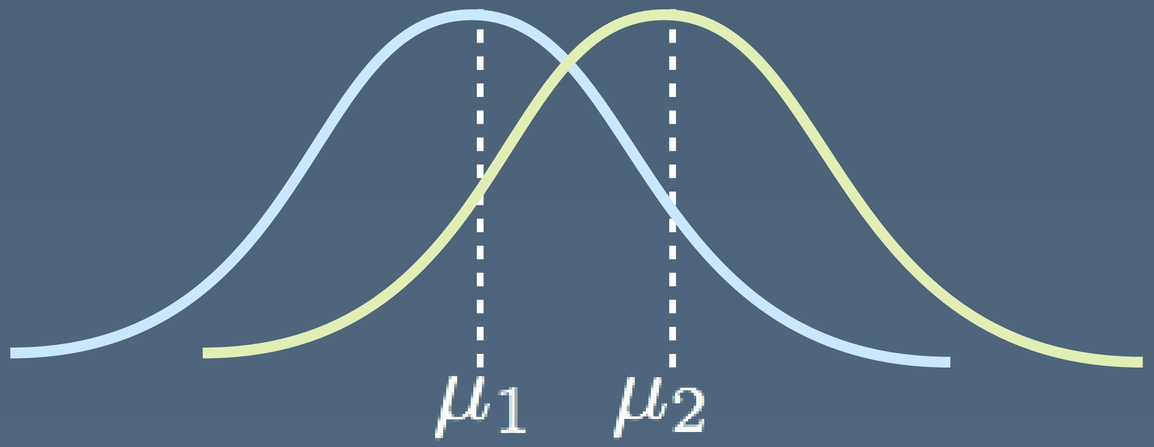
```
     [,1] [,2]
[1,]    5   22
[2,]   35   18
> chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  data
X-squared = 14.3117, df = 1, p-value = 0.0001549
```

Comparing two populations: means

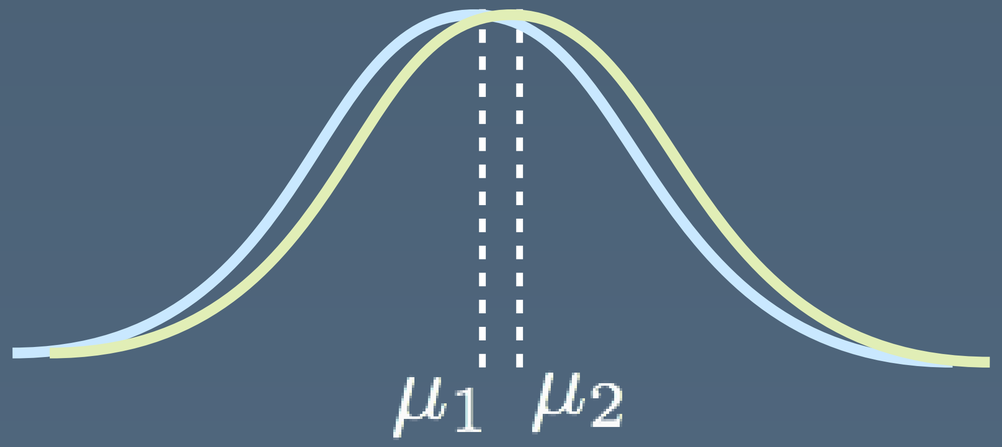
Normally distributed data



t-test: do they have the same mean?



likely have different means



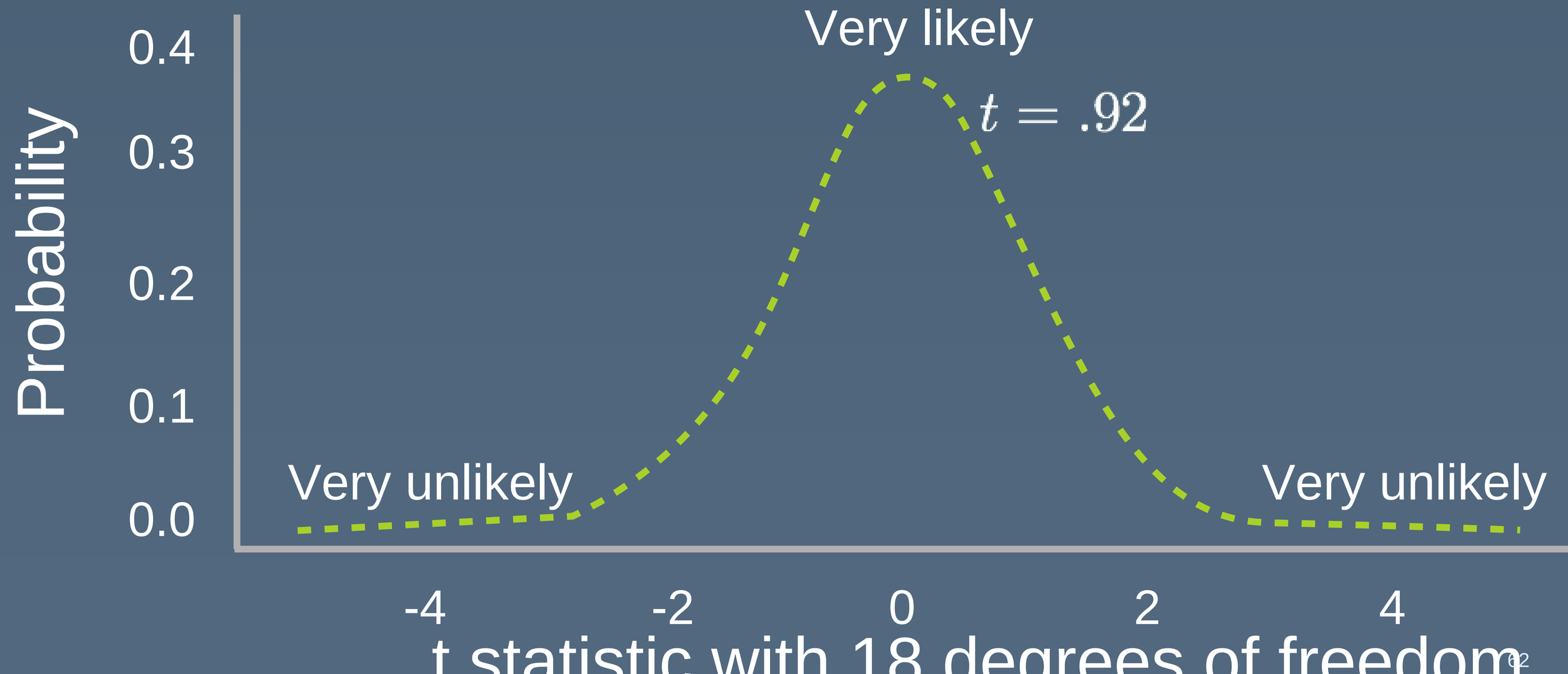
likely have the same mean
(null hypothesis)

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Numbers that matter:

- Difference in means
  larger means more significant
- Variance in each group
  larger means less significant
- Number of samples
  larger means more significant

Example t distribution



How many degrees of freedom?

- If we know the mean of N numbers, then only N-1 of those numbers can change.
- We have two means, so a t-test has N-2 degrees of freedom.

Running the test in R

- Use `t.test` (HCI R tutorial at http://yatani.jp/HCIstats/TTest)

```
> data
       group result
1    control      1
2    control      1
3    control      2
4    control      3
5    control      1
6    control      3
7    control      2
8    control      4
9    control      1
10   control      2
11 augmented      6
12 augmented      5
13 augmented      1
14 augmented      3
```

```
> t.test(data[data["group"] == "control", 2], data[data["group"]
 == "augmented", 2], var.equal=T)

        Two Sample t-test

data:  data[data["group"] == "control", 2] and data[data["group"
] == "augmented", 2]
t = -2.2014, df = 18, p-value = 0.04099
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 -2.73610126 -0.06389874
sample estimates:
mean of x mean of y
      2.0       3.4
```

Presenting the result

- “A t-test comparing the expert-rated scores of designs with the control (mean=2.0, std. dev=0.5) to the designs with the augmented condition (mean=3.4, std. dev=0.4) is significant ($t(18)=2.2$, $p<.05$).”

Within-subjects study designs

- It can be easier to statistically detect a difference if the participants try both alternatives.
- Why?

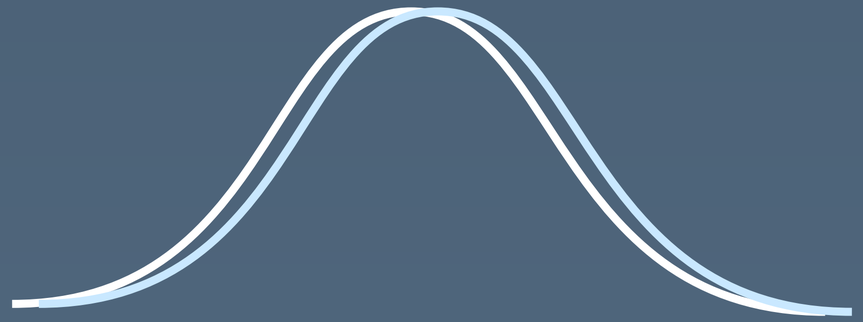
Paired t-test

Control	Augmented
1	6
2	
1	5
2	1
3	3
1	5
3	1
2	2
4	3

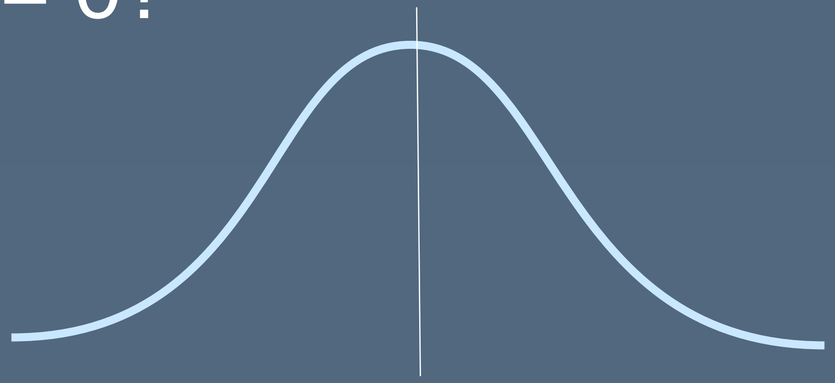
A paired test controls for individual-level differences.

Unpaired vs. paired t-test

- Do two normal distributions have the same mean?



- Paired t-test: does the distribution of (after - before) have mean = 0?



Paired t-test

$$t = \frac{\mu - 0}{\sqrt{\frac{\sigma^2}{N}}}$$

- Is the mean of that difference significantly different from zero?

Running a paired t-test in R

```
> t.test(data[data["group"] == "control", 2], data[data["group"]
 == "augmented", 2], paired=T)

        Paired t-test

data:  data[data["group"] == "control", 2] and data[data["group"
] == "augmented", 2]
t = -1.7685, df = 9, p-value = 0.1108
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 -3.1907752  0.3907752
sample estimates:
mean of the differences
                  -1.4
```

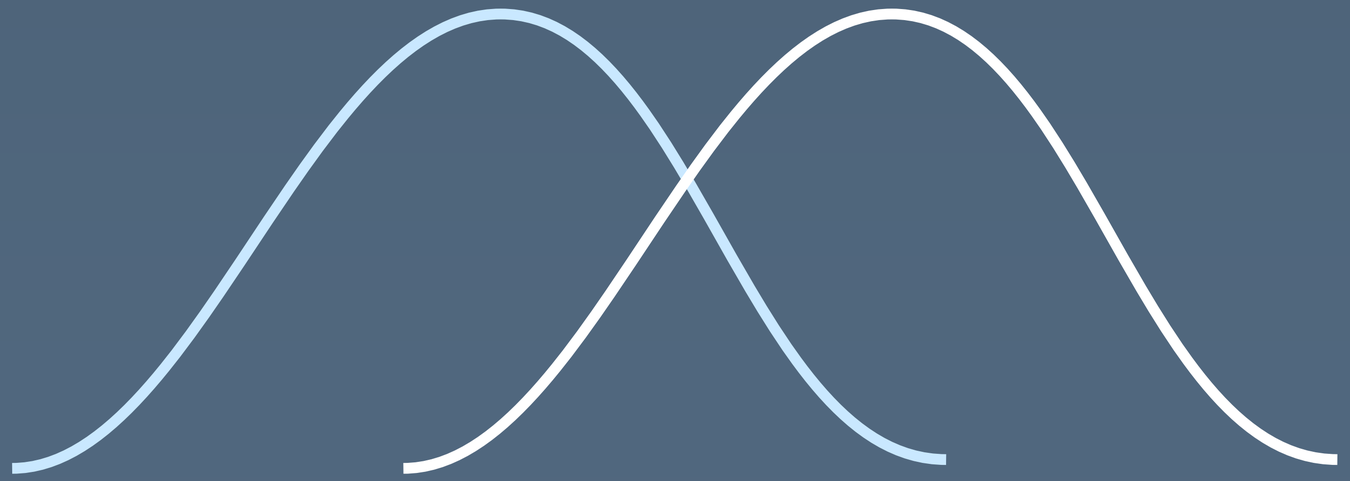
Why no longer significant? (Hint: look at the degrees of freedom “df”)

Ten participants. If we had twenty rows like before, much

ANOVA

t-test: compare two means

- “Do people fix more bugs with our IDE bug suggestion callouts?”

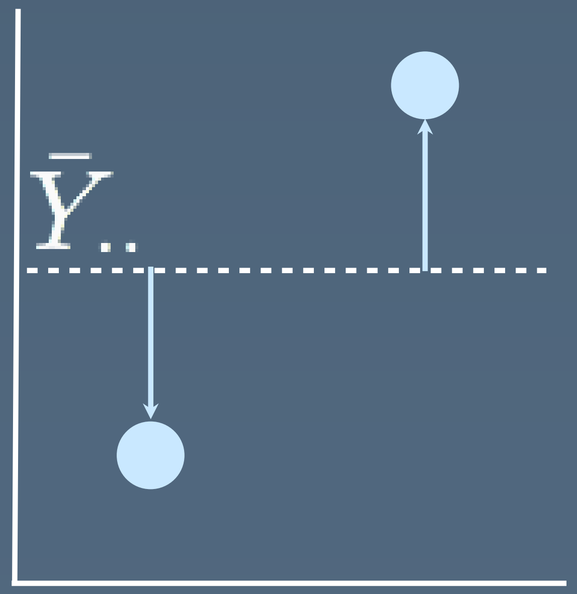


ANOVA: compare N means

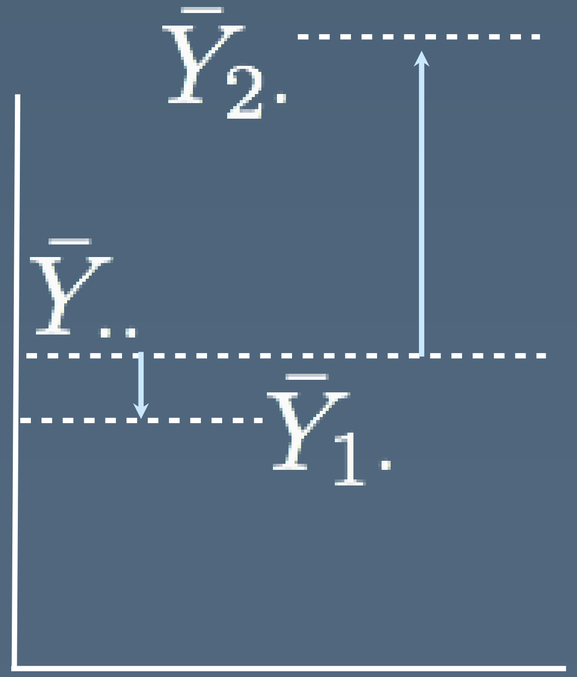
- “Do people fix more bugs with our IDE bug suggestion callouts, with warnings, or with nothing?”

Rough intuition for ANOVA test

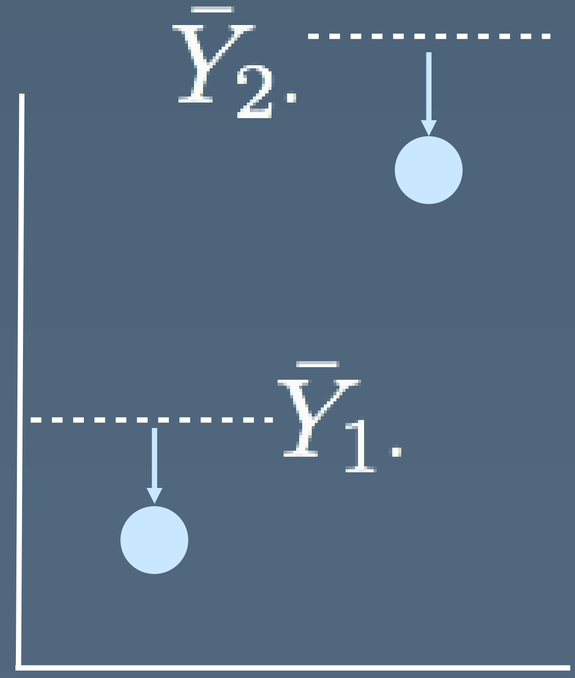
How much of the total variation can be accounted for by looking at the means of each condition?



total deviation
from grand mean



deviation of factor
mean from grand mean



deviation of
response from
factor mean

ANalysis Of VAriance (ANOVA)

- Degrees of freedom: how many values can vary? (Using n and r)

  Degrees of freedom in individual data points: n - 1
  Degrees of freedom in factor level averages: r - 1
  Combined: n - r

Finally: run the test!

- How large is the value we constructed from the F distribution?
- Test if $F^* > F(1-\alpha; r-1, n-r)$

```
> aov <- aov(value ~ group, data)
> summary(aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
group         2  22.75   11.38    12.1 0.00032 ***
Residuals    21  19.75    0.94
```

factor

error (“what’s left”)

3 factor levels hopefully F(2,21) p < .001

24 observations top >> bottom

Reporting an ANOVA

- “A one-way ANOVA revealed a significant difference in the effect of news feed source on number of likes (F(2, 21)=12.1, p<.001).”

```
> aov <- aov(value ~ group, data)
> summary(aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
group         2  22.75   11.38    12.1 0.00032 ***
Residuals    21  19.75    0.94
```

Summary

- p-values encode our desired probability of a false positive
- Chi-square test compares count or rate data
- t-test compares two means
- Paired t-test compares means within subjects
- ANOVA compares more than two means