A Brief Introduction to Experimental Design With People

•••

Katherine Livins

Quant Hour April 21st 2016

Goals

1. To provide a basic understanding of the relationship between experiments and causality
2. To note the things you specifically need to consider when designing experiments to measure human behavior (and why)
3. To help you *think and talk* about your experiments

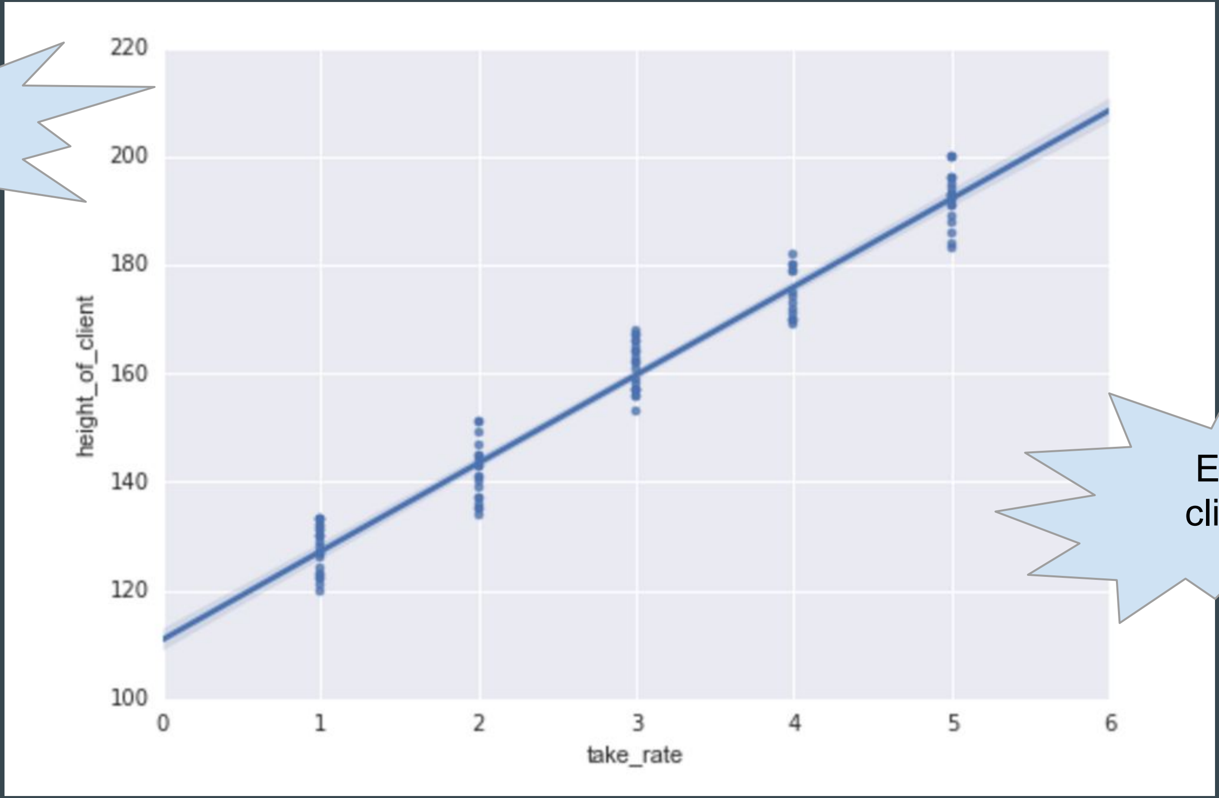
Disclaimer

- Not every question needs an experimental answer
    - E.g., “How many clients do we have?”
    - E.g., “What clients have tended to purchase a given piece of clothing?”
- Experiments are often time consuming, costly, and if done without careful testing/planning can lead to dangerous or inaccurate conclusions

Why Experiment Then?

- Experiments tell you about *causal relationships*
  - E.g., Does the style rainbow cause better styling decisions?
  - E.g., Does a certain layout cause faster reaction times?
  - E.g., Does a certain ad cause greater conversion?
- Clearly state the causal relationship that you want to study whenever talking about an experiment!

Correlation ≠ Causality



Better inventory?

Easier clients?

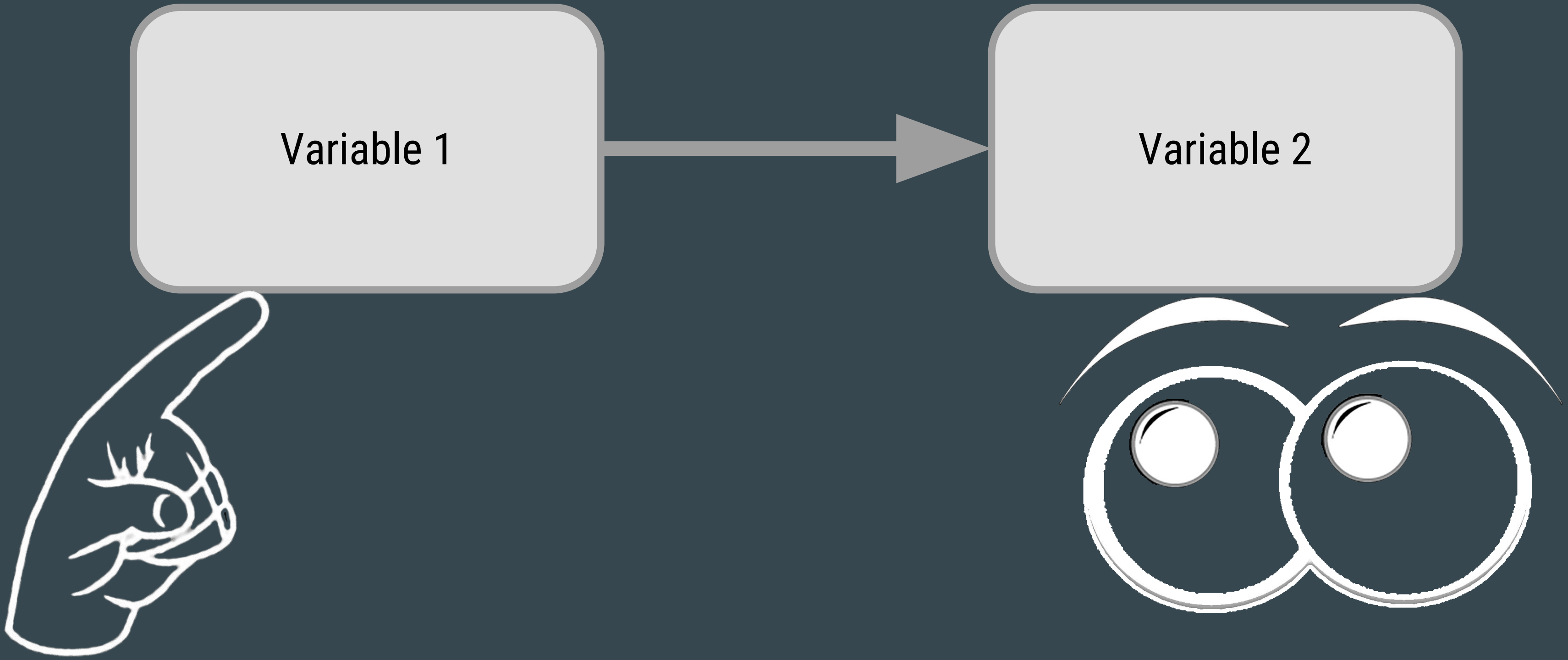
Causality

- Temporal precedence (duh)
- Covariation of cause and effect (easy enough)
- Elimination of all other explanations (surprisingly hard)

I know this sounds obvious to a group of PhDs, but EVERYTHING you do in an experiment will be to establish these three requirements. They are what give your conclusions legs and should be thought about when describing what you studied.

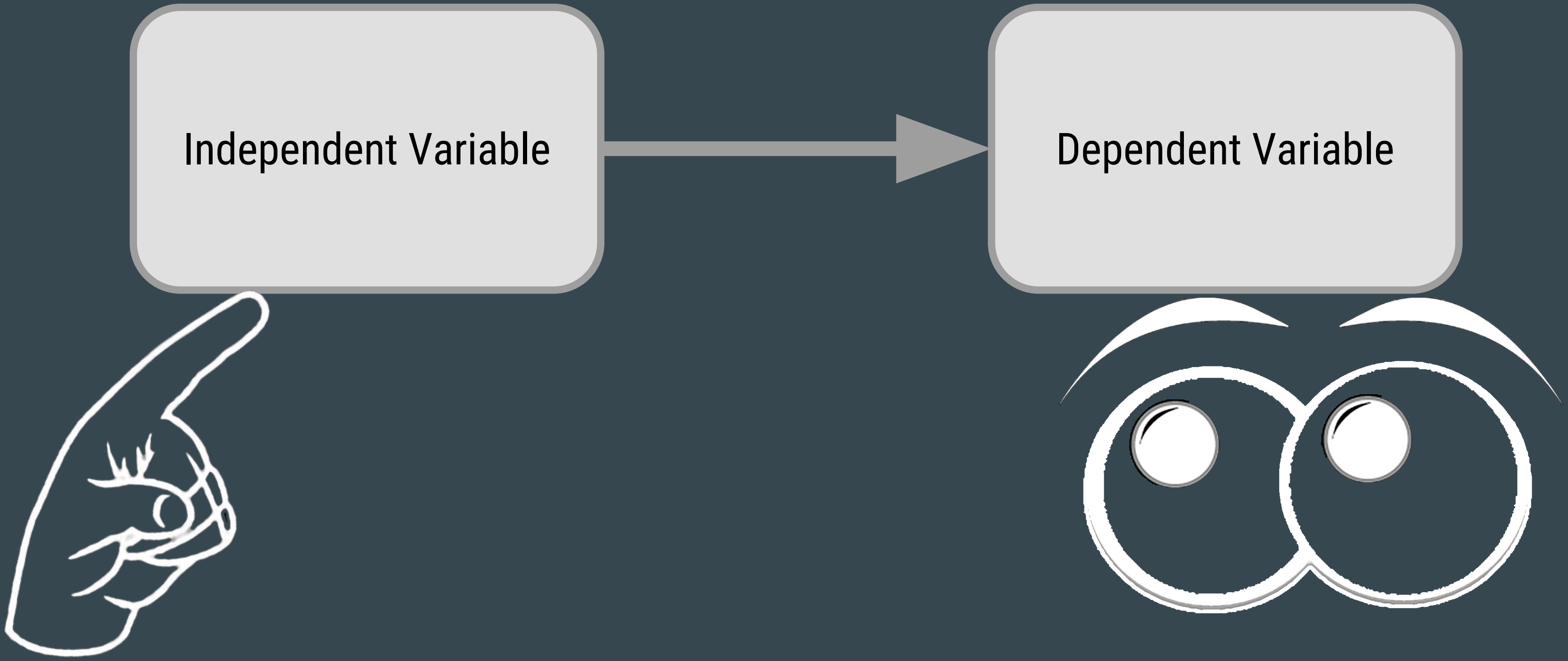
What Is An Experiment Then?

A setting in which you act like a complete control freak to determine whether it is likely that one variable causes some effect on another, in hopes that whatever you find in your experiment can be generalized to the real world.



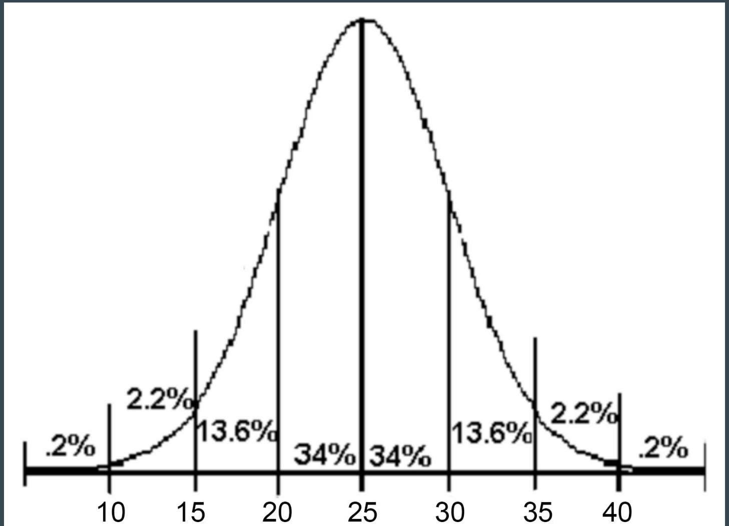
Variable 1

Variable 2

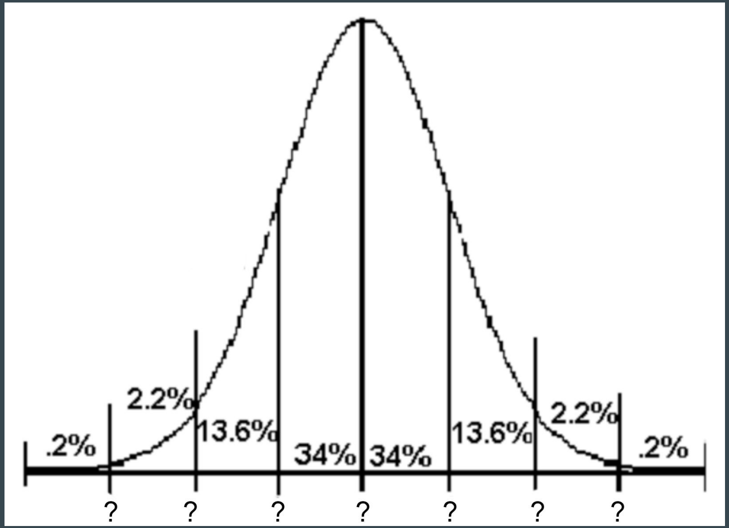


Independent Variable

Dependent Variable



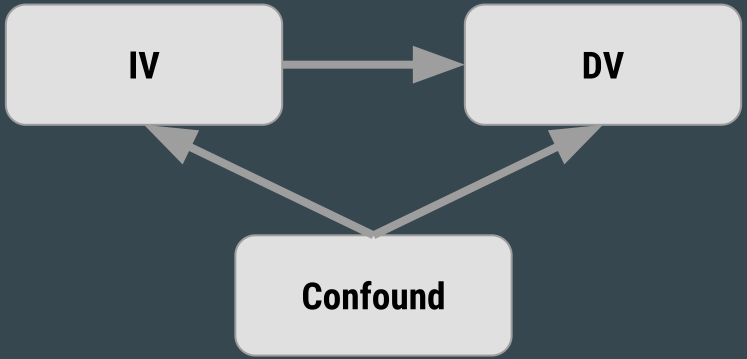
DV Distribution with one amount of IV



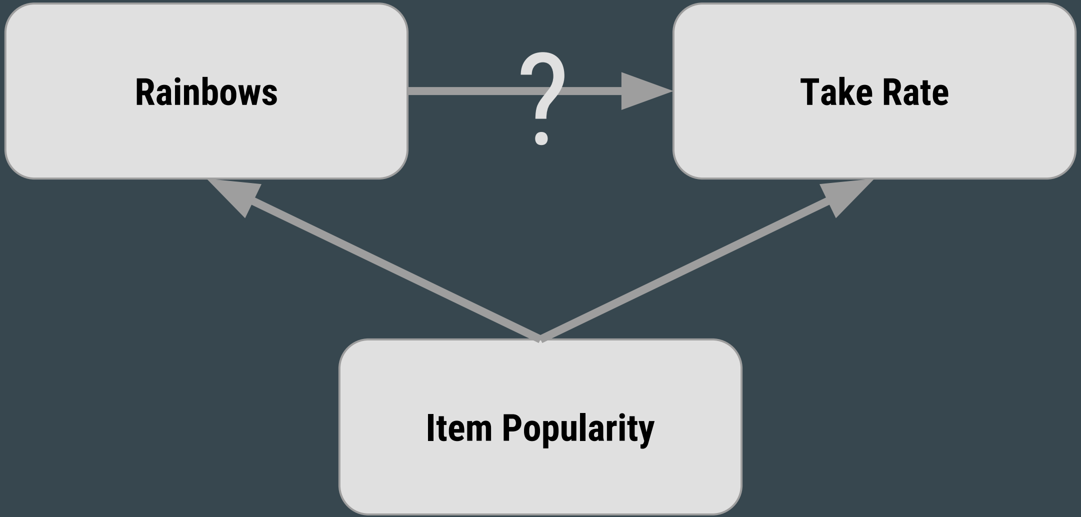
DV Distribution with another amount of IV -
Does it change?

Confounds

- Variables that can change/co-vary with the DV, which also co-vary with the IV
- Can *cause* a change in the DV = *alternative explanations for your findings!*
- Can lead you to perceive a “spurious relationship”



Confounds



General Approach

1. Figure out what you want to study
    - Research question (abstract and concrete)
    - Variables
    - Controls for your confound
2. Get some participants and divide them into groups
3. Give different treatments (amounts/types) of the IV(s) to each group
4. Measure the DV(s)
5. Compare how the groups performed on the DV (enter statistics)
6. If there’s a difference between your groups, conclude that the IV likely caused that difference

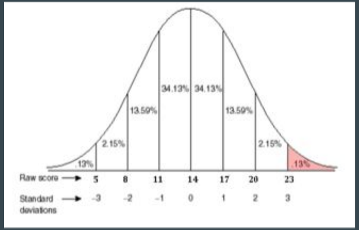
This will hold in almost every case!

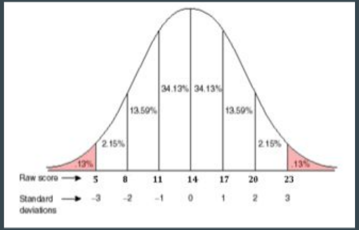
Note!

- Experimental design and statistics go hand-in-hand
    - NEVER start running an experiment without knowing what analyses you’ll need to use in order to interpret your results
    - Your business partners will be pretty unimpressed if you run a study and don’t know how to interpret your results, so imagine all possible outcomes and figure out what they might indicate

The Actual Design - Hypotheses

- Our *Null Hypothesis* is that each level of the IV will result in the same performance on the DV
- Our *Alternative Hypothesis* is that different levels of the IV will result in different performance on the DV
  - Can be “different” (resulting in 2-tailed statistics later-on)
  - Can be “greater than” or “less than” (resulting in a 1-tailed test later-on)





- Subsequent analyses will try simply compare DV performance distributions and either reject or fail to reject the null

Example

Note 2

- Experiments are typically based on inference - we observe a bunch of cases (but not all) and claim that a given outcome is *likely* the case
- We can therefore never prove that some variable makes a difference

The Actual Design - Variables

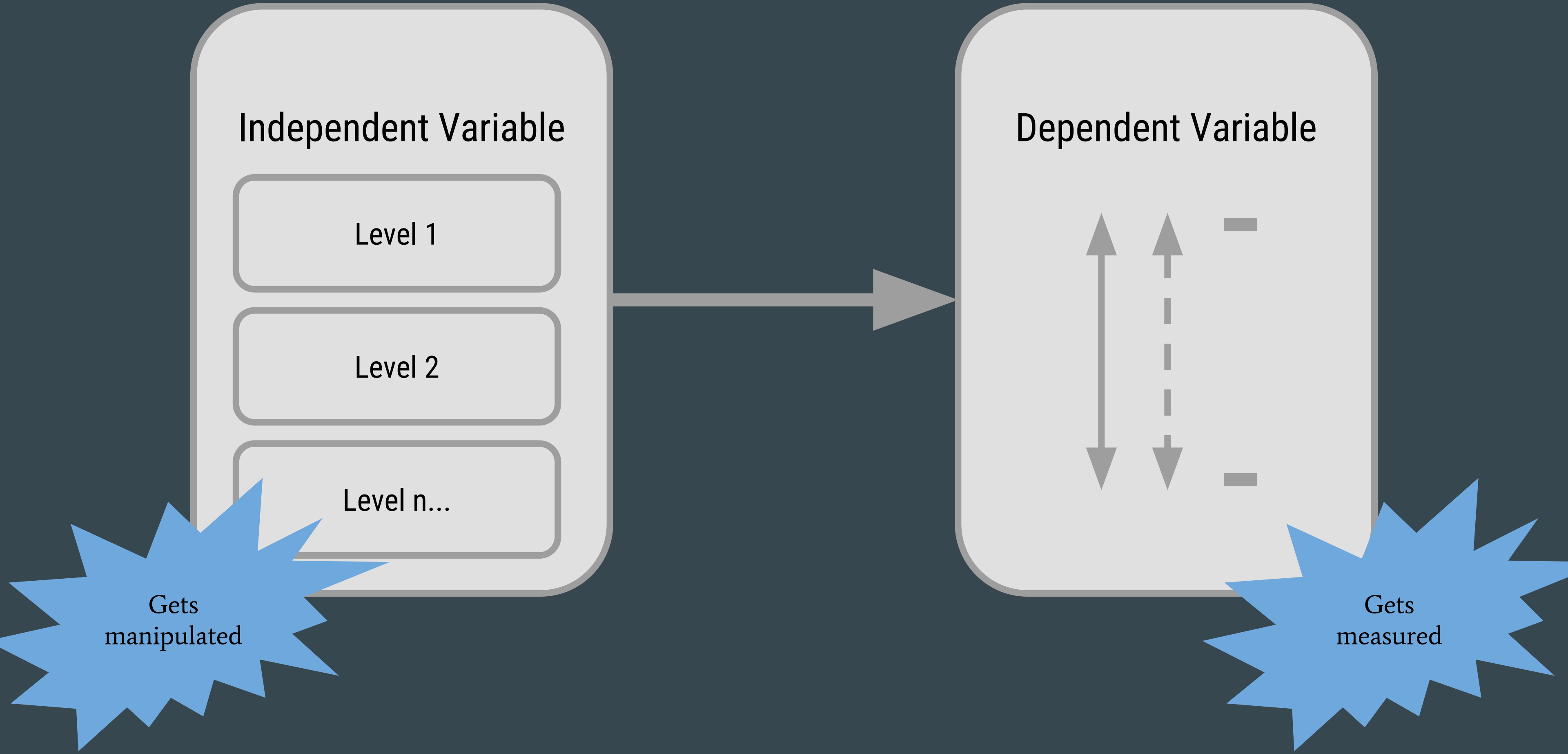
- When dealing with people we often want to measure abstract behaviors
  - E.g., How the heck do you measure “styling performance”?
    - Accuracy?
    - Reaction times?
    - A survey?
  - We have to define our variables in a way that is **measurable**
    - We call this “operationalizing them”

The IV

- We need to manipulate the IV (i.e., change it)
  - We are typically limited in our participant pool, so we want to select a limited number of *levels* for our IV
    - E.g., Selfie, No Selfie
- Think about the strength of manipulation you think you need
  - Careful with task difficulty though because you could see *ceiling effects* or *floor effects*
    - These happen when the task is too easy or too hard
    - They will make it look like your IV isn’t doing anything

The DV

- Since we're just measuring the DV it can take any value possible
- If we want to have more than one DV we need to think about the order in which they're measured
    - Does measuring one affect later measures?



Independent Variable

Level 1

Level 2

Level n...

Gets manipulated

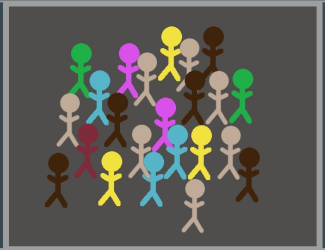
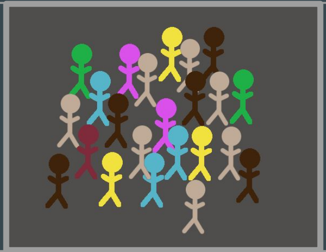
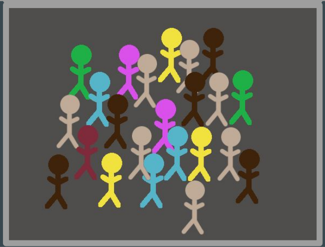
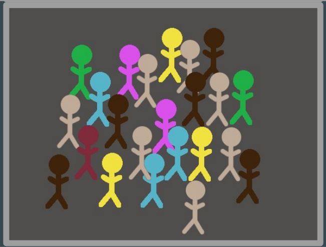
Dependent Variable

Gets measured

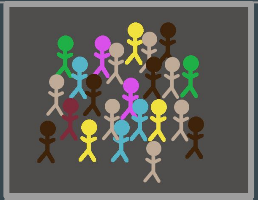
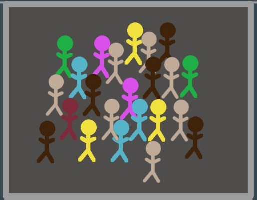
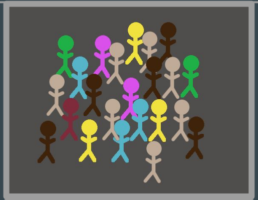
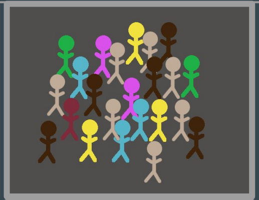
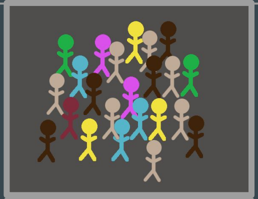
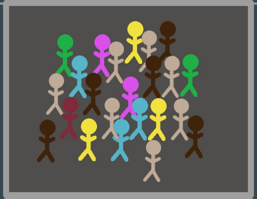
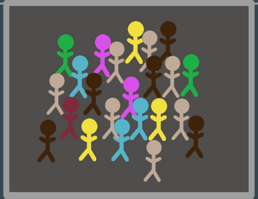
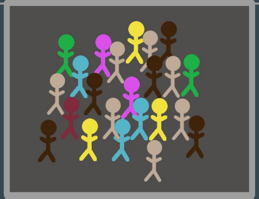
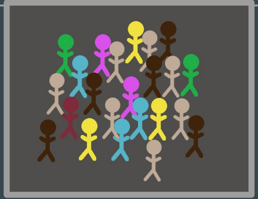
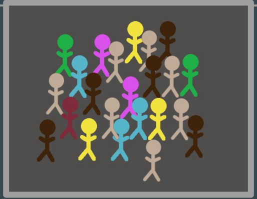
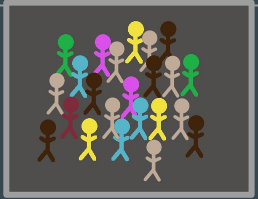
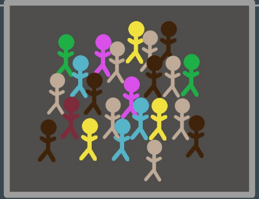
More Complex Designs

- Sometimes you want more bang for your buck
- *Factorial designs* allow you to study multiple IVs
  - Can be really useful
  - Combine all levels of each IV to create groups
  - Warning: be careful not to make your experiment too big or too complex

Example

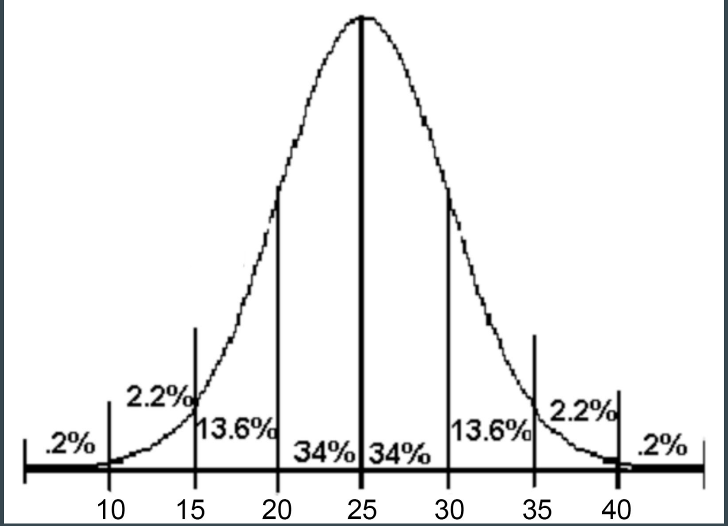
	IV 1 Level 1 - Advertisement 1	IV 1 Level 2 - No Advertisement
IV 2 Level 1 - Red Background		
IV 2 Level 2 - Blue Background		

Example

	IV 1 Level 1 - Advertisement 1	IV 1 Level 2 - Advertisement 2	IV 1 Level 3 - Advertisement 3	IV 1 Level 4 - No Advertisement
IV 2 Level 1 - Red Background				
IV 2 Level 2 - Blue Background				
IV 2 Level 3 - Green Background				

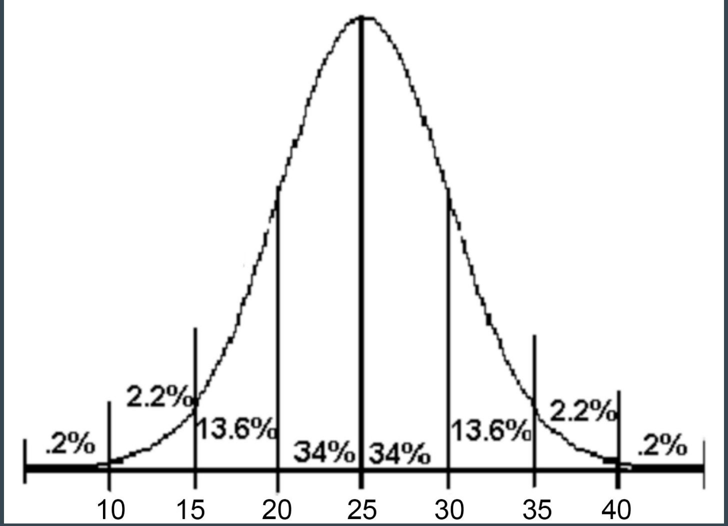
The Actual Design - Sampling

- Everyone you might be interested in studying is your *population*
  - E.g., All stylists
  - E.g., All Americans
  - E.g., All human beings
- Sometimes it's hard to get to everyone in your population so we *sample*
- You expect that what's going on in your sample is *representative* of what's going on in the population



Population

=



Sample

blood test

Sampling and Inference

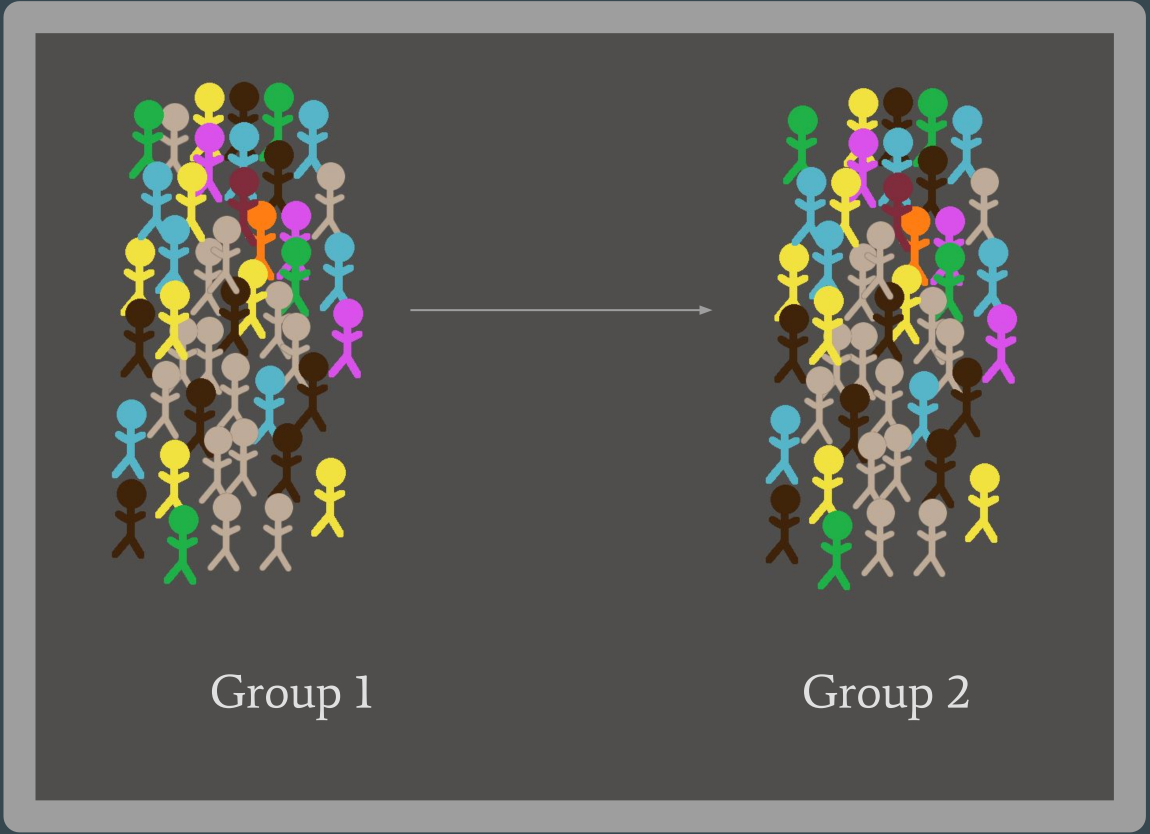
- A non-representative sample can be hard to detect but can jeopardize your ability to generalize to your population
- The best sampling is (usually) truly random (sometimes very difficult)
- The easiest is often to ask for volunteers (sometimes very biased, but called *haphazard sampling*)

The Actual Design - Group Assignment

- People come with history - history that might affect the DV in some unknown or unexpected way
- You want your IV groups to be as similar as possible otherwise you might lose the ability to say that your IV causes a change in your DV before you even collect any data
  - I.e., you might lose your internal validity
  - (They may act as alternative explanations to your findings)
- Random assignment is your friend

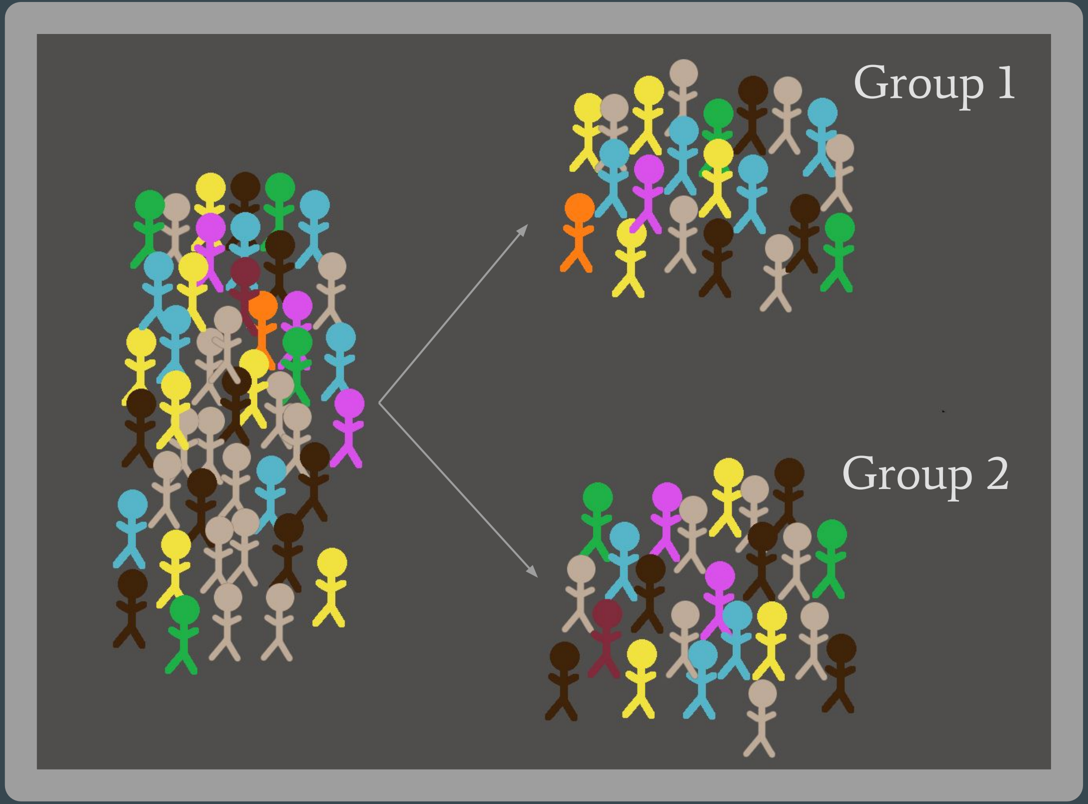
Option 1

Within-Participants (aka, Repeated Measures): Measure 'em once, change the IV level, then measure 'em again



Option 2

Between-Participants: Collect ‘em then (randomly) split ‘em and give each group a different level of the IV



When To Use What?

	Within Participants	**Across Participants**
Number of participants	One group	One group per condition
Learning effects	Can’t cope very well	Not a problem
Time constraints	Can be very long	Shorter participation time
Group differences	Identical	Participant Differences

The Actual Design - Validity

- The degree to which a study answers the question it says it does
    - *Construct:*
        - How well did you operationalize your variables?
    - *Internal*: The degree to which the relationship between the IV(s) and DV(s) can be established
        - Basically, how well the experiment was run
        - Are there any *confounds* (i.e., other variables differing between your groups that can explain between-group differences)
    - *External*: The degree to which your study can be generalized to the real world
        - How artificial is your experiment's environment?
        - How artificial is your task?

The Trade Off

Internal Validity
External Validity

Validity - Remember People Are Smart

- *Reactivity* is when people change their behavior because they’re being studied
- *Demand Characteristics* are things that tell your participants what you’re studying
    - If people realize they’re in a study they will try to guess what it’s about
    - They will also try to guess what your expectations are
    - They will often try to live up to those expectations
- *Learning effects* occur when people get better at your task

!YOU NEED TO CONTROL FOR THESE THINGS!

Validity - Remember People are Sensitive

- *Order effects* occur when the order in which things are presented in the experiment changes the outcome
  - E.g., *Primacy effects*
- *Stimulus Saliency* can affect what gets noticed

Yes

No

!YOU NEED TO CONTROL FOR THESE THINGS!

Useful Controls

- For demand characteristics:
    - Don’t tell people what your experiment is about
        - If there’s risk involved tell them what they need to know to make an informed decision
    - Use “filler trials”
- For order effects & learning effects:
    - Randomize when possible
    - Important to *counterbalance* when using repeated measures
        - Include all treatment orders

Validity - Remember People Are Sensitive

- Things about people can interact with things about the experiment
  - The day
  - The environment
  - The task (especially the questions)
- These need to be the same across conditions
  - If they aren’t then they can be *confounding variables*
- When doing a between participants design make sure to randomize group assignment

Reliability

- Based on the idea that good results should be repeatable
  - If you did your study again, would you get approximately the same findings?
- You can think of a given DV score as

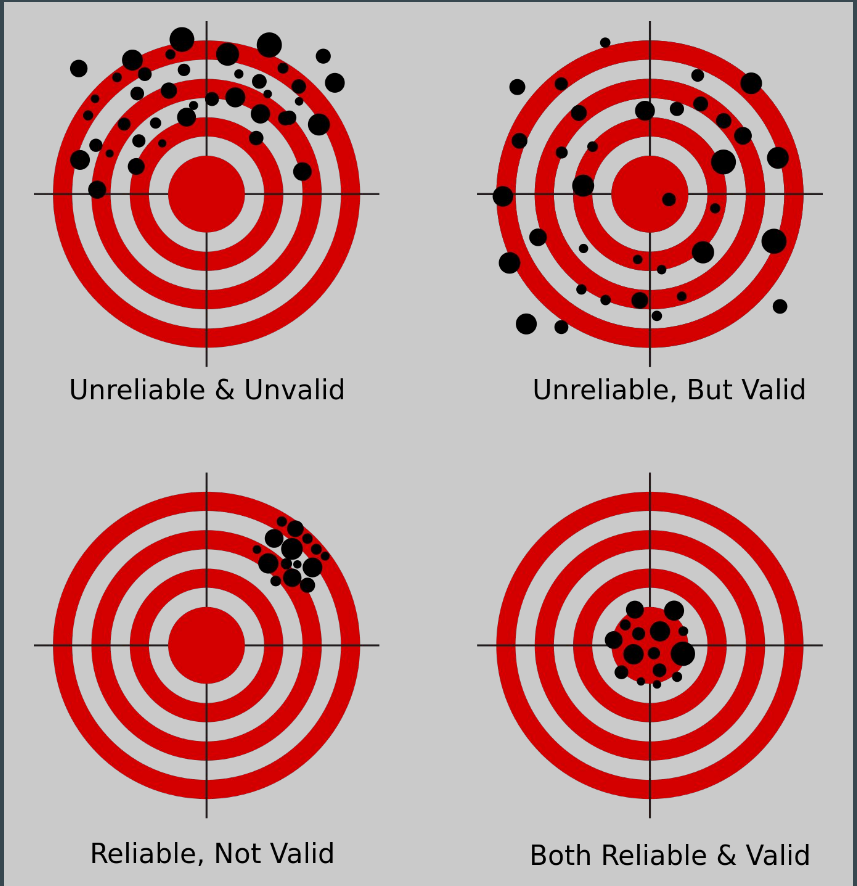
$$\boxed{\text{true score + error}}$$

Measurement Error
which can come from

- equipment
- training
- the items used

Reliability

- Many types
    - Some are trivial (e.g., measurement tools can be evaluated by testing and re-testing them)
    - Some are more difficult
        - Inter-rater reliability must be evaluated when coding is involved
            - Percent agreement with a cut-off
            - Cohen’s Kappa
                - Looks at the the agreement between raters versus chance agreement



**Shamelessly taken from Wikipedia*

A Note About Ethics

- Your studies can have *real consequences* for *real people*
  - Ask yourself what could happen to them for being in your study
  - Ask yourself how your company could look if the public found out about the study
- E.g., Facebook manipulated the emotional content in the newsfeeds of > 1 million users to see how positivity and negativity could spread around the Internet
  - Are the outcomes worth the findings?
  - Does consent really cover things like this?
  - Would *you* want to be in a study like this?
- If possible, try to debrief after the study

Collecting Data

- Determine how many participants you'll need prior to running your study
    - Convention is that you run until you find a significant result
        - This is **terrible** practice and a form of *p-hacking*
        - http://multithreaded.stitchfix.com/blog/2015/10/15/multiple-hypothesis-testing/
    - A better way is to run a *power analysis*
        - Still not perfect
        - Involves estimating
            - your expected group means (i.e., performance under each level of the IV)
            - your variance
            - your alpha

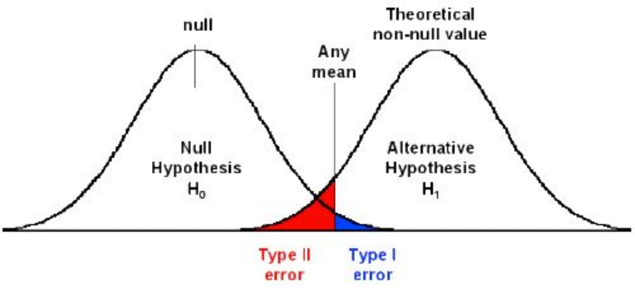
Once You Have Your Data

1. Get an idea of what your data looks like
2. Pick a level of risk (an alpha level)
3. Select the appropriate statistical test
4. Run that test
5. Interpret your results with reference to your hypotheses
6. Draw a conclusion about your population

Every experiment will use this basic procedure!

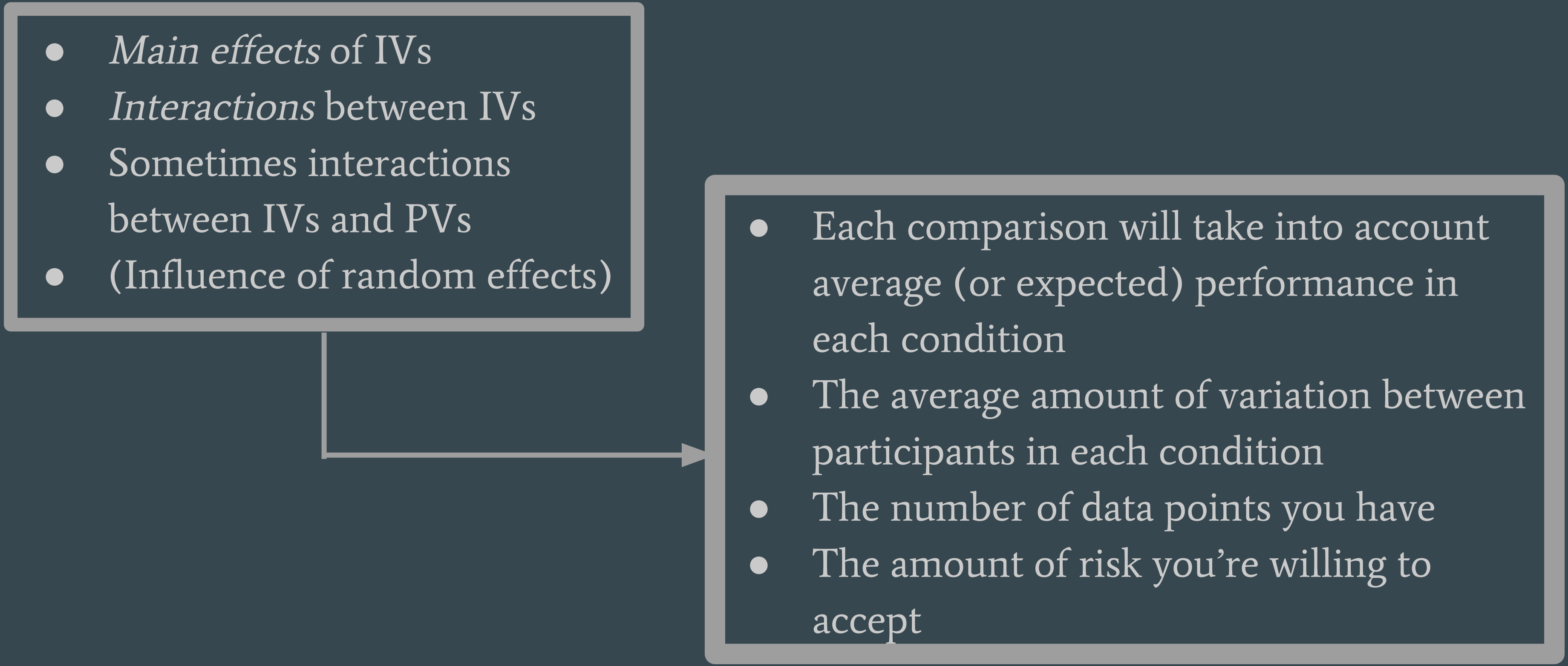
Experimental Risk

- Basically the probability that you're wrong about your findings
- Alpha level
  - The probability of making a Type I error
  - The probability of incorrectly rejecting a true null (i.e., a false positive)
  - Often set to .05
- Power
  - The probability of making a Type II error
  - The probability of failing to reject a false null (i.e., a false negative)



**Shamelessly taken from a random online page*

What You’re Usually Looking For…



- *Main effects* of IVs
- *Interactions* between IVs
- Sometimes interactions between IVs and PVs
- (Influence of random effects)

- Each comparison will take into account average (or expected) performance in each condition
- The average amount of variation between participants in each condition
- The number of data points you have
- The amount of risk you’re willing to accept

Research
Superstar

QUESTIONS?