Course Business

- Discuss midterm projects
  - Due today!
- Short-ish lecture on effect size & power
  - sleep.csv on CourseWeb
  - We'll also be finishing cuedrecall.csv from last week
- Next week = SPRING BREAK, WOO!
  - No class
  - Scheduled office hours will not be held, but I'm available over e-mail or by appointment

Week 9: Effect Size & Power

- Distributed Practice
- Finish `glmer()`
  - Interactions
  - Coding the Dependent Variable
  - Other Distributions
- Effect Size
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Distributed Practice

- Your colleague Arpad, who studies insomnia, ran a study examining whether (a) hours of exercise the day before and (b) amount of caffeine consumed predicted whether people successfully slept through the night:

```
InsomniaModel <- glmer(SleptThroughNight ~ 1 + HoursExercise + MgCaffeine + (1|Subject), data=sleep, family=binomial)
```

- Arpad would like help interpreting his R output.

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.102561   0.485201   4.333 1.47e-05 ***
HoursExercise 0.610568   0.225743   2.705  0.00684 **
MgCaffeine   -0.005148   0.003270  -1.575  0.11537
```

- Describe how hours of exercise affected sleeping through the night:

Distributed Practice

- Your colleague Arpad, who studies insomnia, ran a study examining whether (a) hours of exercise the day before and (b) amount of caffeine consumed predicted whether people successfully slept through the night:

```
InsomniaModel <- glmer(SleptThroughNight ~ 1 + HoursExercise + MgCaffeine + (1|Subject), data=sleep, family=binomial)
```

- Arpad would like help interpreting his R output.

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.102561   0.485201   4.333 1.47e-05 ***
HoursExercise 0.610568   0.225743   2.705  0.00684 **
MgCaffeine   -0.005148   0.003270  -1.575  0.11537
```

- Describe how hours of exercise affected sleeping through the night:
  - Every hour of exercise increased the odds of sleeping through the night by exp(0.61) = 1.84 times

Distributed Practice

- Sleep data from one subject wasn't properly recorded due to experimenter error

```
   Subject     Observation  HoursExercise      MgCaffeine      SleptThroughNight   HoursSleep
S001   :  6   S001-1 :  1   Min.   :0.0000   Min.   :  0.00   Min.   :0.000     Min.   : 0.600
S024   :  6   S001-2 :  1   1st Qu.:0.0000   1st Qu.: 35.79   1st Qu.:1.000     1st Qu.: 5.700
S028   :  6   S001-3 :  1   Median :0.5000   Median : 82.12   Median :1.000     Median : 7.300
S058   :  6   S001-4 :  1   Mean   :0.9015   Mean   : 87.47   Mean   :0.829     Mean   : 7.159
S063   :  6   S001-5 :  1   3rd Qu.:2.0000   3rd Qu.:129.20   3rd Qu.:1.000     3rd Qu.: 8.600
S068   :  6   S001-6 :  1   Max.   :3.0000   Max.   :297.61   Max.   :1.000     Max.   :12.600
(Other):238   (Other):268                                     NA's   :5         NA's   :5
```

- Since there is no reason to think this subject would be systematically different from the others, let's just remove those observations entirely. Which would **NOT** accomplish this?

(a) `sleep$HoursSleep <- ifelse(is.na(sleep$HoursSleep), 0, sleep$HoursSleep)`

(b) `sleep <- subset(sleep, is.na(sleep$HoursSleep) == FALSE)`

(c) `sleep <- sleep[is.na(sleep$HoursSleep) == FALSE, ]`

(d) `sleep <- na.omit(sleep)`

Distributed Practice

- Sleep data from one subject wasn't properly recorded due to experimenter error

```
   Subject     Observation  HoursExercise      MgCaffeine      SleptThroughNight   HoursSleep
S001   :  6   S001-1 :  1   Min.   :0.0000   Min.   :  0.00   Min.   :0.000     Min.   : 0.600
S024   :  6   S001-2 :  1   1st Qu.:0.0000   1st Qu.: 35.79   1st Qu.:1.000     1st Qu.: 5.700
S028   :  6   S001-3 :  1   Median :0.5000   Median : 82.12   Median :1.000     Median : 7.300
S058   :  6   S001-4 :  1   Mean   :0.9015   Mean   : 87.47   Mean   :0.829     Mean   : 7.159
S063   :  6   S001-5 :  1   3rd Qu.:2.0000   3rd Qu.:129.20   3rd Qu.:1.000     3rd Qu.: 8.600
S068   :  6   S001-6 :  1   Max.   :3.0000   Max.   :297.61   Max.   :1.000     Max.   :12.600
(Other):238   (Other):268                                     NA's   :5         NA's   :5
```

- Since there is no reason to think this subject would be systematically different from the others, let's just remove those observations entirely. Which would **NOT** accomplish this?

(a) `sleep$HoursSleep <- ifelse(is.na(sleep$HoursSleep), 0, sleep$HoursSleep)`

This would replace the missing values with 0s rather than remove them. That's not what we want here—failure to record the data doesn't mean that the person slept 0 hours

Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - Interactions
  - Coding the Dependent Variable
  - Other Distributions
- Effect Size
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

cuedrecall.csv

- Let's model our cued recall data with `glmer()`
  - 120 `Subjects`, all see the same 36 `WordPairs`
  - `AssocStrength` (property of **WordPairs**):
    - Two words have **Low** or **High** relation in meaning
      - VIKING—HELMET = high associative strength
      - VIKING—COLLEGE = low associative strength
  - Study `Strategy` (property of **Subjects**):
    - **Maintenance** rehearsal: Repeat it over & over
    - **Elaborative** rehearsal: Relate the two words

- Model with maximal random effects structure:

```
model1 <- glmer(Recalled ~
1 + AssocStrength * Strategy
+ (1 + AssocStrength|Subject)
+ (1 + Strategy|WordPair),
data=cuedrecall, family=binomial)
```

Interactions

```
Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.50485    0.04939  10.221  < 2e-16 ***
AssocStrength1               0.32199    0.09649   3.337 0.000847 ***
Strategy1                    0.72851    0.07735   9.418  < 2e-16 ***
AssocStrength1:Strategy1    -0.48515    0.14880  -3.261 0.001112 **
```

- Associative strength has a **+** effect on recall
- Study time has a **+** effect on recall
- But, their interaction has a **-** coefficient
- Interpretation?:
  - "With elaborative rehearsal, associative strength matters less"
  - "If pair has high associative strength, it matters less how you study it" (another way of saying the same thing)

Interactions

- We now understand the **sign** of the interaction
- What about the specific numeric **estimate**?
  - What does -.48515 mean in this context?

```
AssocStrength1:Strategy1 -0.48515    0.14880  -3.261 0.001112 **
```

- Descriptive stats: Log odds in each condition
  - Not something you have to do when running your own model—this is just to understand where the numbers come from

	Elaborative	Maintenance
High	0.9007865	0.4170527
Low	0.8079227	-0.1372521

- Low associative strength pair:
  - Elaborative rehearsal -> Increase of ≈ 0.97 logits
- High associative strength pair:
  - Elaborative rehearsal -> Increase of ≈ 0.49 logits

Interactions

```
AssocStrength1:Strategy1 -0.48515    0.14880  -3.261 0.001112 **
```

- Low associative strength pair:
  - Elaborative rehearsal -> Increase of 0.97 logits
- High associative strength pair:
  - Elaborative rehearsal -> Increase of 0.49 logits
- We can compute a difference in log odds:

0.49 – 0.97 = -0.48

- Or an **odds ratio** in terms of the odds:

$$\frac{\exp(.49)}{\exp(.97)} = \exp(-0.48) = 0.62$$

Interactions

```
AssocStrength1:Strategy1 -0.48515    0.14880  -3.261 0.001112 **
```

- Low associative strength pair:
  - Elaborative rehearsal -> Increase of 0.97 logits
- High associative strength pair:
  - Elaborative rehearsal -> Increase of 0.49 logits
- An **odds <u>ratio</u>** in terms of the odds:

$$\frac{\exp(.49)}{\exp(.97)} = \exp(-0.48) = 0.62$$

- "The ratio between the odds of recalling pairs with elaborative versus maintenance rehearsal was 0.62 times smaller for high associative strength items."

Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - Coding the Dependent Variable
  - Other Distributions
- Effect Size
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Coding the Dependent Variable

- So far, positive numbers in the results meant **better recall**

```
Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.50485    0.04939  10.221  < 2e-16 ***
AssocStrength1             0.32199    0.09649   3.337 0.000847 ***
Strategy1                  0.72851    0.07735   9.418  < 2e-16 ***
AssocStrength1:Strategy1  -0.48515    0.14880  -3.261 0.001112 **
```

- That's because we treat **correct recall** as a **1** ("hit") and an **error** as a **0** ("miss")
  - We're looking at things that **predict recall**

```
> contrasts(cuedrecall$Recalled)
           Remembered
Forgotten           0
Remembered          1
```

Coding the Dependent Variable

I don't trust these results. What if we'd coded it the other way, with "forgotten" as 1 and "remembered" as 0? Things might be totally different!

- This is also a totally plausible coding scheme
  - Variable that tracks whether you forgot something!
- Let's see if Evil Scott is right:
  - Step 1: Create a new variable that codes things the way Evil Scott wants
  - Step 2: Re-run the model
  - Step 3: ???
  - Step 4: PROFIT!

Coding the Dependent Variable

I don't trust these results. What if we'd coded it the other way, with "forgotten" as 1 and "remembered" as 0? Things might be totally different!

- This is also a totally plausible coding scheme
  - Variable that tracks whether you forgot something!
- Let's see if Evil Scott is right:
  - Step 1: Create a new variable that codes things the way Evil Scott wants
    - `cuedrecall$Forgotten <- ifelse(cuedrecall $Recalled == 'Forgotten', 1, 0)`
  - Step 2: Re-run the model
  - Step 3: ???
  - Step 4: PROFIT!

Coding the Dependent Variable

- Let's try running our model with the new coding:

Model of recall

```
Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.50485    0.04939  10.221  < 2e-16 ***
AssocStrength1               0.32199    0.09649   3.337 0.000847 ***
Strategy1                    0.72851    0.07735   9.418  < 2e-16 ***
AssocStrength1:Strategy1    -0.48515    0.14880  -3.261 0.001112 **
```

```
Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.50484    0.04939 -10.221  < 2e-16 ***
AssocStrength1              -0.32199    0.09649  -3.337 0.000847 ***
Strategy1                   -0.72850    0.07735  -9.418  < 2e-16 ***
AssocStrength1:Strategy1     0.48514    0.14880   3.260 0.001113 **
```

Model of for-getting

- All we've done is flip the signs
  - Anything that increases remembering decreases forgetting (and vice versa)
  - Remember how logits equally distant from even odds have the same absolute value?
  - Won't affect pattern of significance
- Conclusion: What we code as 1 vs 0 doesn't affect our conclusions (good!!)
  - Choose the coding that makes sense for your research question. Do you want to talk about "what predicts graduation" or "what predicts dropping out"?

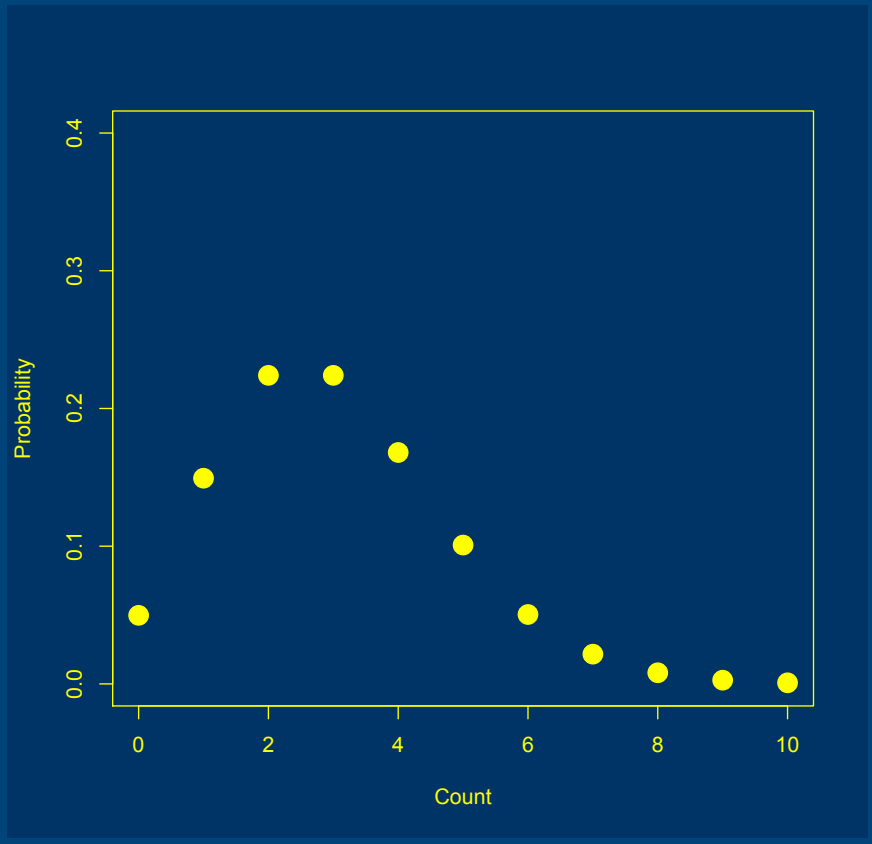
Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - Other Distributions
- Effect Size
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Other Distributions

- `glmer()` supports other non-normal distributions

- `family=poisson`
  - For count data
    - Examples:
      - Number of solutions you brainstormed for a problem
      - Number of gestures in a storytelling task
      - Number of doctor's visits
  - Counts range from 0 to positive infinity
  - Link is **log(count)**



Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- Effect Size
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Effect Size

- With `sleep.csv`, let's run a model predicting `HoursSleep` from fixed effects of `HoursExercise` and `MgCaffeine`, and a random intercept of `Subject`
  - Which fixed effects significantly influence the number of hours of sleep that people get?

Effect Size

- With `sleep.csv`, let's run a model predicting `HoursSleep` from fixed effects of `HoursExercise` and `MgCaffeine`, and a random intercept of `Subject`
  - Which fixed effects significantly influence the number of hours of sleep that people get?
- `SleepModel <- lmer(HoursSleep ~ 1 + HoursExercise + MgCaffeine + (1|Subject), data=sleep)`
  - We're back to lmer because this is a continuous DV

```
Fixed effects:
                Estimate Std. Erro  t value
(Intercept)     6.794429   0.25151   27.014
HoursExercise   0.720977   0.07249    9.945
MgCaffeine     -0.004190   0.00120   -3.475
```

They both do!

Effect Size

- *t* statistics and p-values tell us about **whether** there's an effect in the population
- A separate question is **how big** the effect is
  - **Effect size**

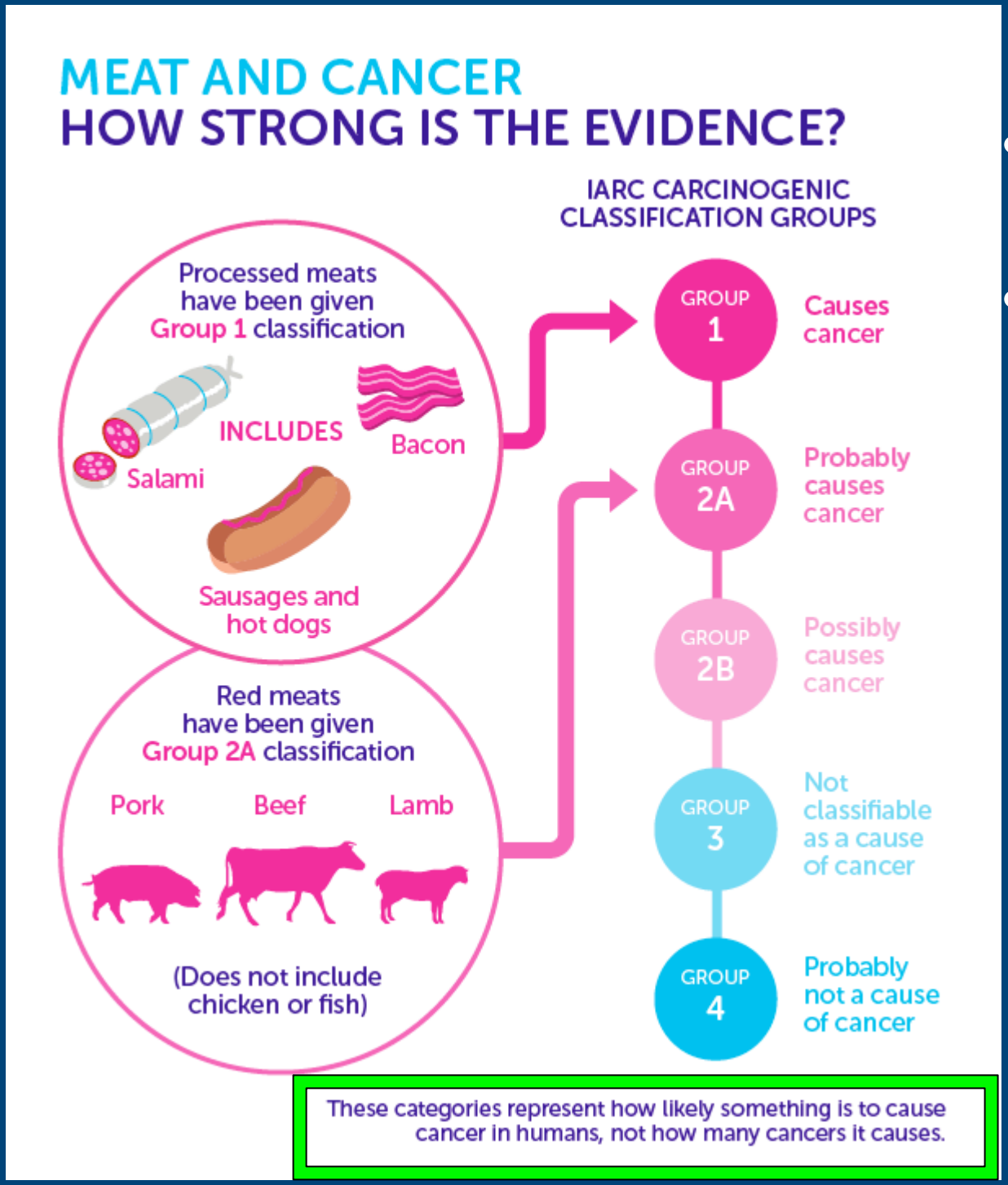
PBS NEWS HOUR **PBS NewsHour**
@NewsHour

Follow

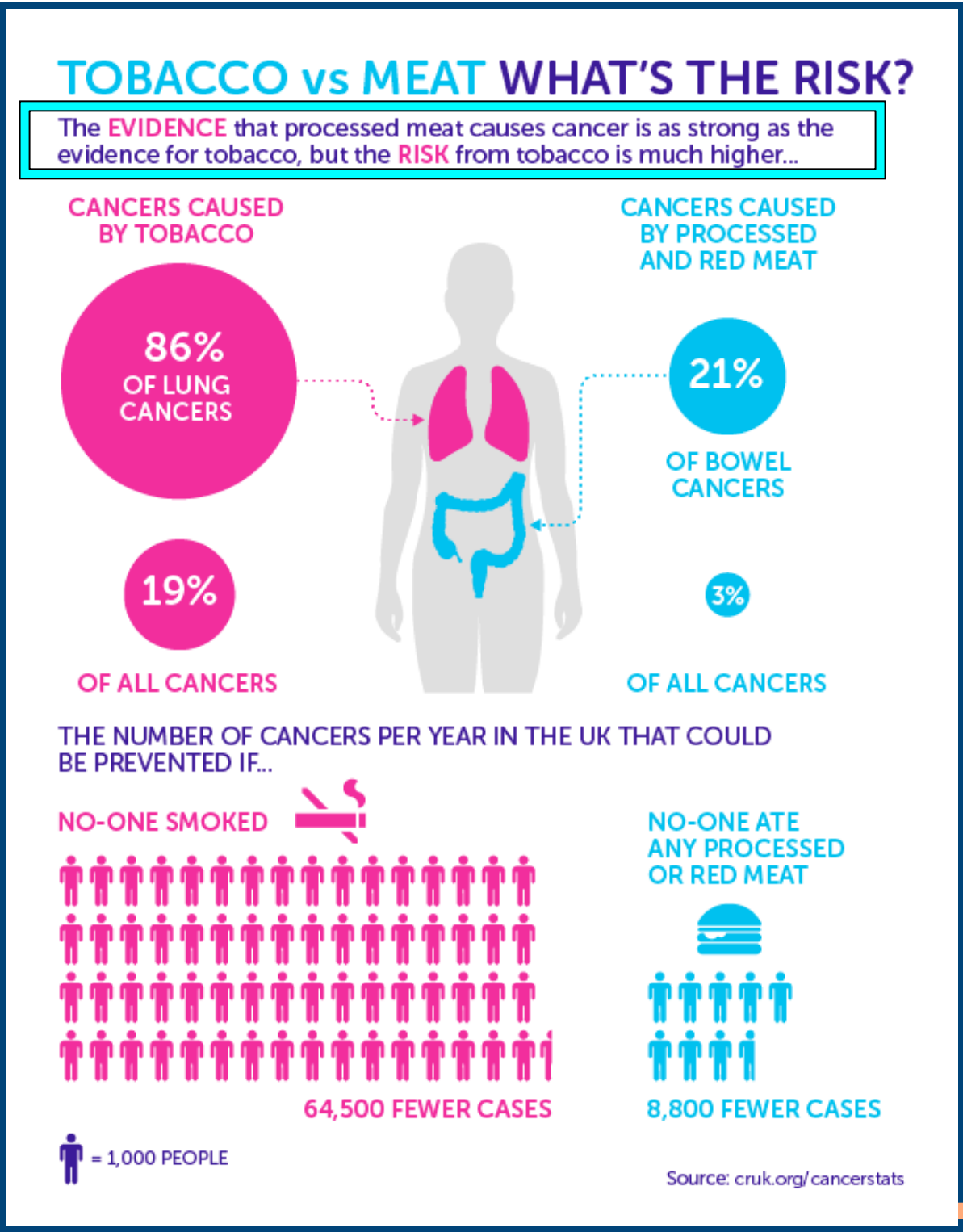
Bacon, hot dogs and processed meats cause cancer/are as dangerous as smoking, says @WHO. to.pbs.org/1WdDzBy

October 26, 2015

- Is bacon really this bad for you??



- Is bacon really this bad for you??
- **True** that we have as much evidence that bacon causes cancer as smoking causes cancer!
  - Same level of **statistical reliability**



- Is bacon really this bad for you??
- **True** that we have as much evidence that bacon causes cancer as smoking causes cancer!
  - Same level of **statistical reliability**
- But, **effect size** is much smaller for bacon

Effect Size: Parameter Estimate

- Simplest measure: Parameter estimates
  - Effect of 1-unit change in predictor on outcome variable
  - “Each hour of exercise the day before resulted in another 0.72 hours of sleep”
  - “On average, RT decreased by 18 ms for each additional trial of experience”
  - “Personalized math problems increased odds of passing exam by 1.2 times.”
  - Concrete! Good for “real-world” outcomes

```
Fixed effects:
                 Estimate  Std. Error t value
(Intercept)      6.794429    0.251513  27.014
HoursExercise    0.720977    0.072494   9.945
MgCaffeine      -0.004190    0.001206  -3.475
```

Effect Size: Standardization

- Which is the bigger effect?
  - 1 hour of exercise = 0.72 hours of sleep
  - 1 mg of caffeine = -0.004 hours of sleep
- Problem: These are measured in different units
  - Hours of exercise vs. mg of caffeine

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	6.794429	0.251513	27.014
HoursExercise	0.720977	0.072494	9.945
MgCaffeine	-0.004190	0.001206	-3.475

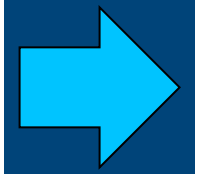
Effect Size: Standardization

- Which is the bigger effect?
  - 1 hour of exercise = 0.72 hours of sleep
  - 1 mg of caffeine = -0.004 hours of sleep
- Problem: These are measured in different units
  - Hours of exercise vs. mg of caffeine
- Convert to z-scores: # of standard deviations from the mean
  - This scale applies to anything!
  - **Standardized** scores

Effect Size: Standardization

- `scale()` puts things in terms of z-scores
- New z-scored version of HoursExercise:
  - `sleep$HoursExercise.z <- scale(sleep$HoursExercise)[,1]`
  - # of standard deviations above/below mean hours of exercise)

```
HoursExercise
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.8959
3rd Qu.:2.0000
Max.   :3.0000
```



```
HoursExercise.z
Min.   :-0.865233
1st Qu.:-0.865233
Median :-0.865233
Mean   :-0.005326
3rd Qu.: 1.054393
Max.   : 2.014206
```

- Then use these in a new model

Effect Size: Standardization

- `scale()` puts things in terms of z-scores
- New z-scored version of HoursExercise:
  - `sleep$HoursExercise.z <- scale(sleep$HoursExercise)[,1]`
  - # of standard deviations above/below mean hours of exercise)
- Then use these in a new model

- Try z-scoring MgCaffeine, too
- Then, run a model with the z-scored variables. Which has the largest effect?

Effect Size: Standardization

- `scale()` puts things in terms of z-scores
- New z-scored version of HoursExercise:
  - `sleep$HoursExercise.z <- scale(sleep$HoursExercise)[,1]`
  - # of standard deviations above/below mean hours of exercise)
- Then use these in a new model

```
Fixed effects:
                Estimate Std. Error t value
(Intercept)      7.07789    0.21879   32.35
HoursExercise.z  0.75116    0.07553    9.95
MgCaffeine.z    -0.26337    0.07579   -3.47
```

1 SD increase in exercise
=> 0.75 hours of sleep

1 SD increase in caffeine
=> -0.26 hours of sleep

Exercise effect is bigger

Interpreting Effect Size

- Consider **in context** of other effect sizes in this domain:

Our effect: .20

Other effect 1: .30

Other effect 2: .40

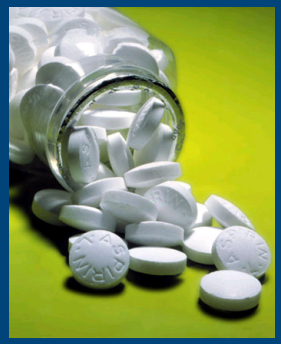
- vs:

Other effect 1: .10

Other effect 2: .15

Our effect: .20

- For interventions: Consider cost, difficulty of implementation, etc.
- Basic science: …predictions of competing theories

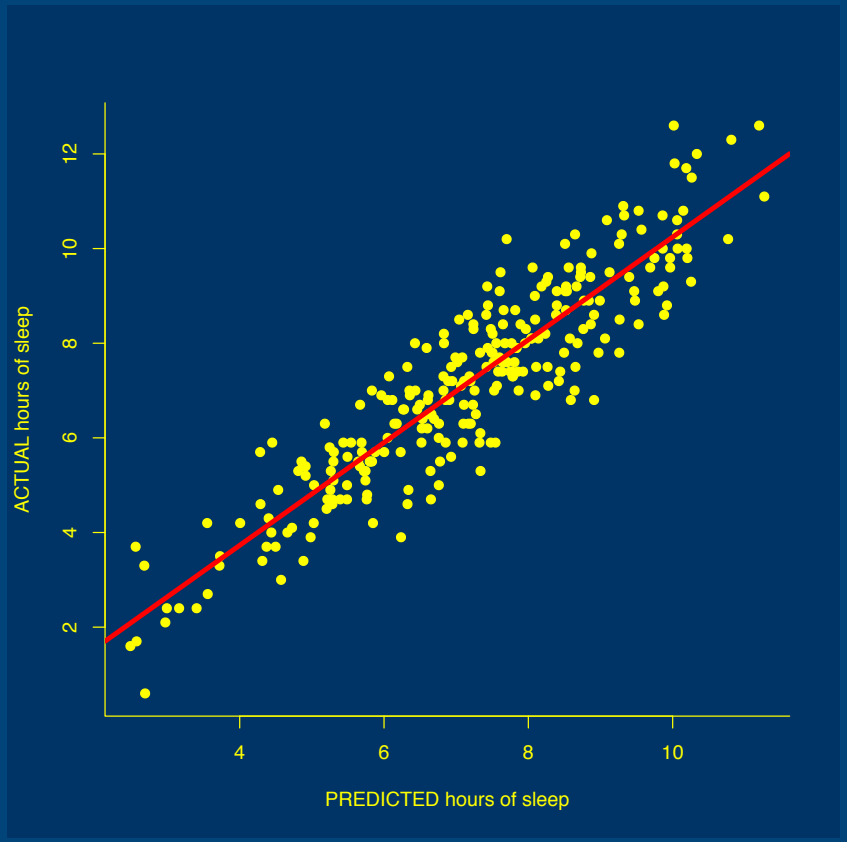


Overall Variance Explained

- How well can we explain this DV?
  - Test: Do *predicted* values match up well with the *actual* outcomes?
- R^2:

```
cor(fitted(SleepModel),
sleep$HoursSleep)^2
```

  - But, this includes what's predicted on basis of subjects (and other random effects)
  - Compare to the R^2 of a model with *just* the random effects & no fixed effects



Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - Type I and Type II Error
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Type I Error

- Does "brain training" affect general cognition?
  - H_0: There is no effect of brain training on cognition
    - $\gamma_1 = 0$ in the population
  - H_A: There is an effect of brain training on cognition
    - $\gamma_1 \neq 0$ in the population

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
2016, Vol. 17(3) 103–186

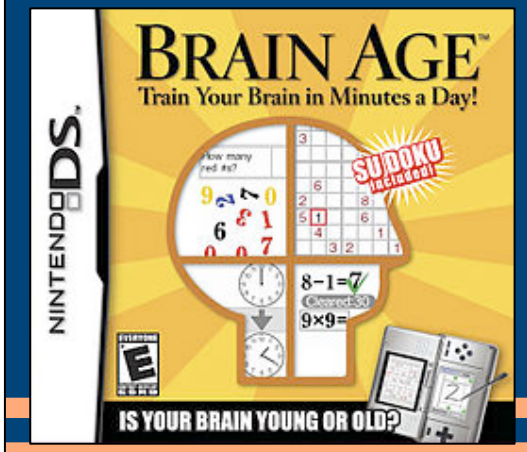
DOI: 10.1177/1529100616661983
pspi.sagepub.com
SAGE

Do "Brain-Training" Programs Work?

Daniel J. Simons[1], Walter R. Boot[2], Neil Charness[2,3], Susan E. Gathercole[4,5], Christopher F. Chabris[6,7], David Z. Hambrick[8], and Elizabeth A. L. Stine-Morrow[9,10]

[1]Department of Psychology, University of Illinois at Urbana-Champaign; [2]Department of Psychology, Florida State University; [3]Institute for Successful Longevity, Florida State University; [4]Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK; [5]School of Clinical Medicine, University of Cambridge; [6]Department of Psychology, Union College; [7]Geisinger Health System, Danville, PA; [8]Department of Psychology, Michigan State University; [9]Department of Educational Psychology, University of Illinois at Urbana-Champaign; and [10]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign

lumosity



Type I Error

- Does “brain training” affect general cognition?
  - H_0: There is no effect of brain training on cognition
    - $\gamma_1 = 0$ in the population
  - H_A: There is an effect of brain training on cognition
    - $\gamma_1 \neq 0$ in the population

Estimate	Std. Error	z value	Pr(>|z|)	
0.50485	0.04939	10.221	< 2e-16	***
0.32199	0.09649	3.337	0.000847	***
0.72851	0.07735	9.418	< 2e-16	***
-0.48515	0.14880	-3.261	0.001112	**

Type I Error

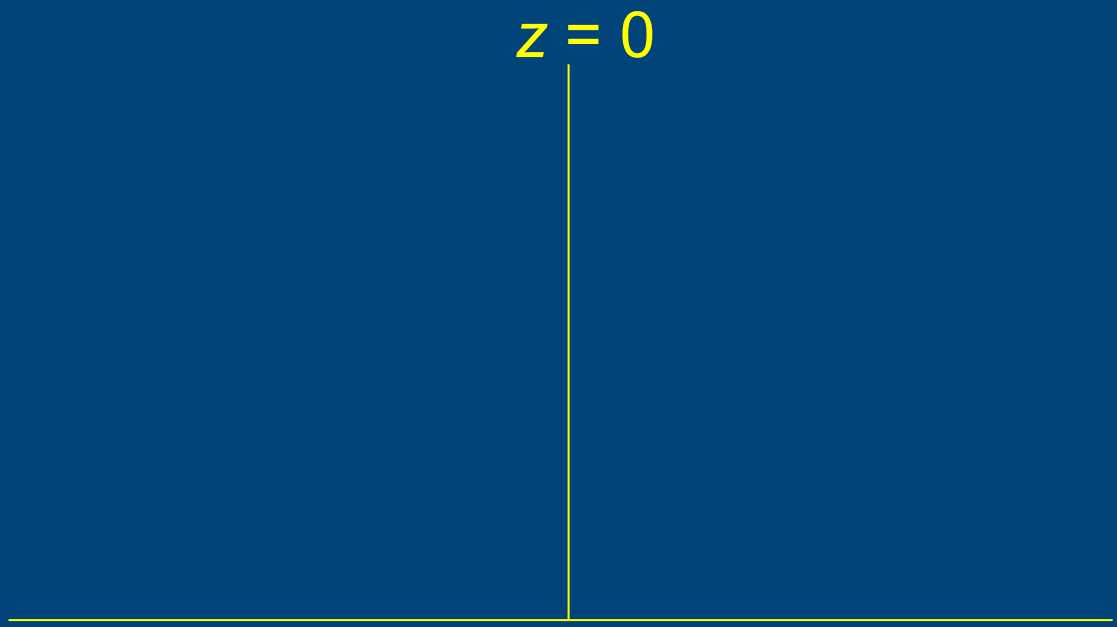
- Is a *z* score of 3.3 good evidence against H_0?
- In a world where brain training has **no** effect on cognition (H_0), the most probable z score would have been 0

13 Movie Trailers That Actually Use the Phrase 'In a World...'

Matt Singer | *February 19, 2015 @ 9:19 AM*

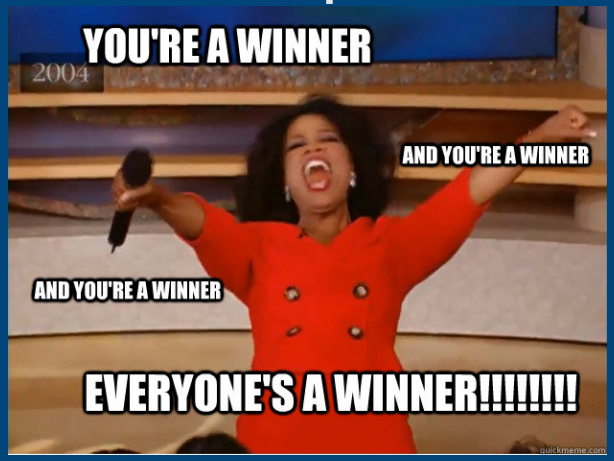
Type I Error

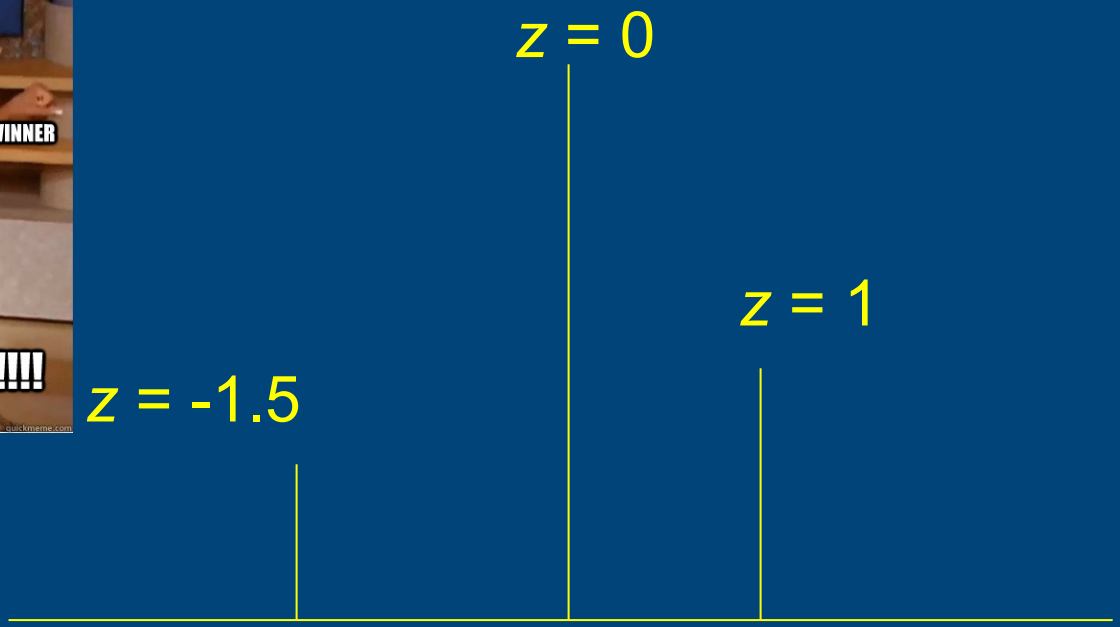
- Is a *z* score of 3.3 good evidence against **H_0**?
- In a world where brain training has **no** effect on cognition (**H_0**), the most probable *z* score would have been 0



Type I Error

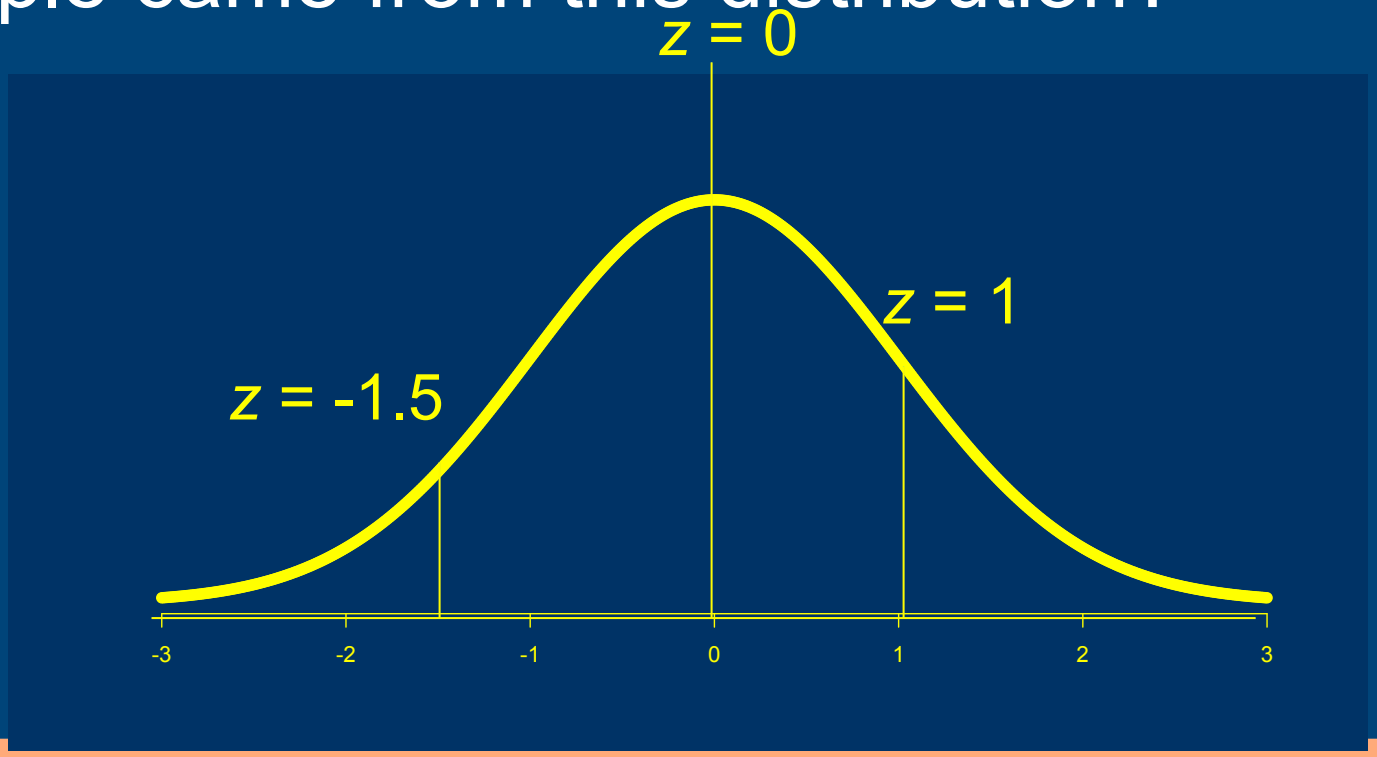
- But even under H_0, we wouldn't always expect to get *exactly* a z-score of 0 in our sample
  - Observed effect will sometimes be higher or lower just by chance (but these values have lower probability) – **sampling error**





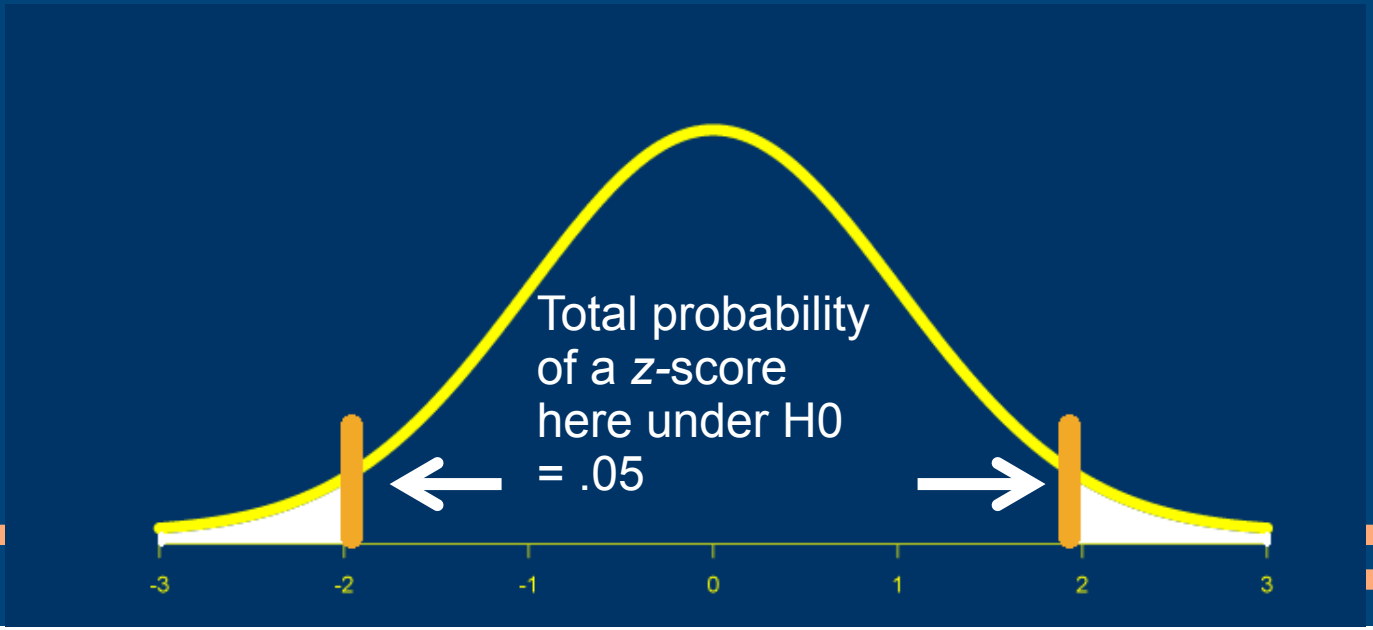
Type I Error

- In a world where H_0 is true, the distribution of *z*-scores should look like this
  - The normal distribution of *z*-scores has mean 0 and std. dev. 1—the **standard normal**
- How plausible is it that the z-score for *our* sample came from this distribution?



Type I Error

- *p*-value: Probability of obtaining a result this extreme under the null hypothesis of no effect
- We reject H_0 when the observed *t* or *z* has < .05 probability of arising under H_0
- But, still *possible* to get this *z* when H_0 is true



Type I Error

- *p*-value: Probability of obtaining a result this extreme under the null hypothesis of no effect
- We reject H_0 when the observed *t* or *z* has < .05 probability of arising under H_0
- But, still *possible* to get this *z* when H_0 is true

A Consensus on the Brain Training Industry from the Scientific Community

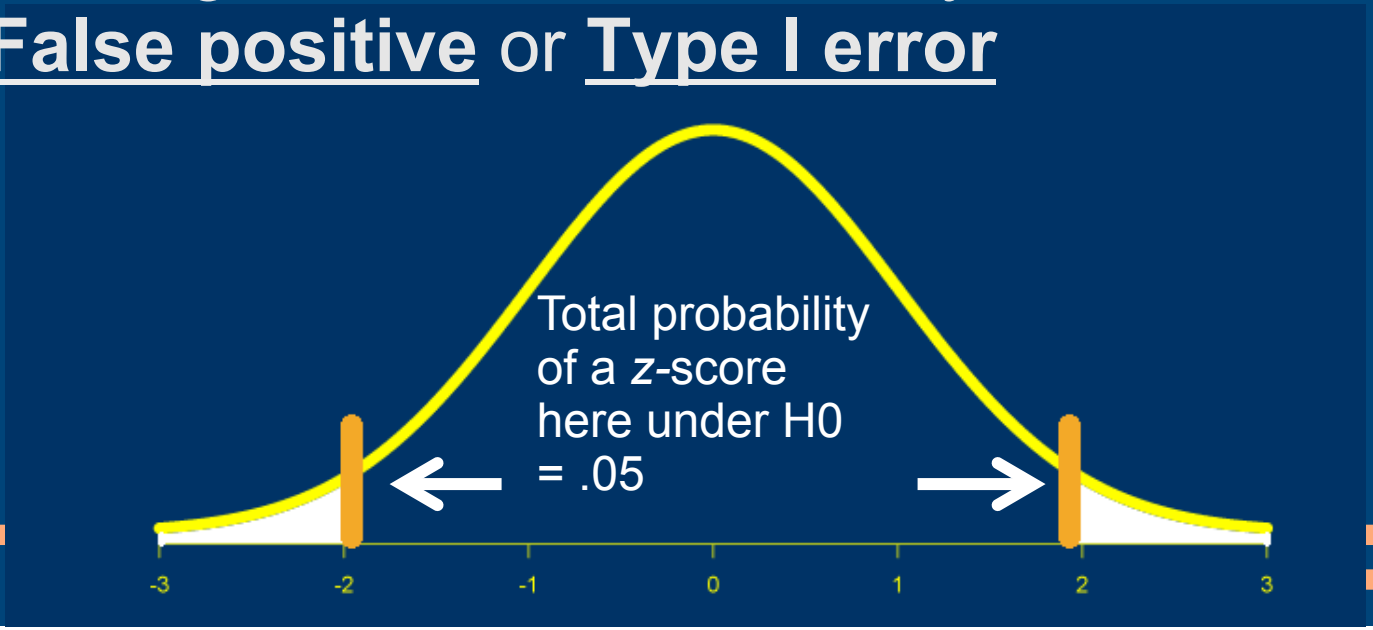
STANFORD CENTER ON LONGEVITY

Max-Planck-Institut für Bildungsforschung
Max Planck Institute for Human Development

To date, there is little evidence that playing brain games improves underlying broad cognitive abilities, or that it enables one to better navigate a complex realm of everyday life. Some intriguing

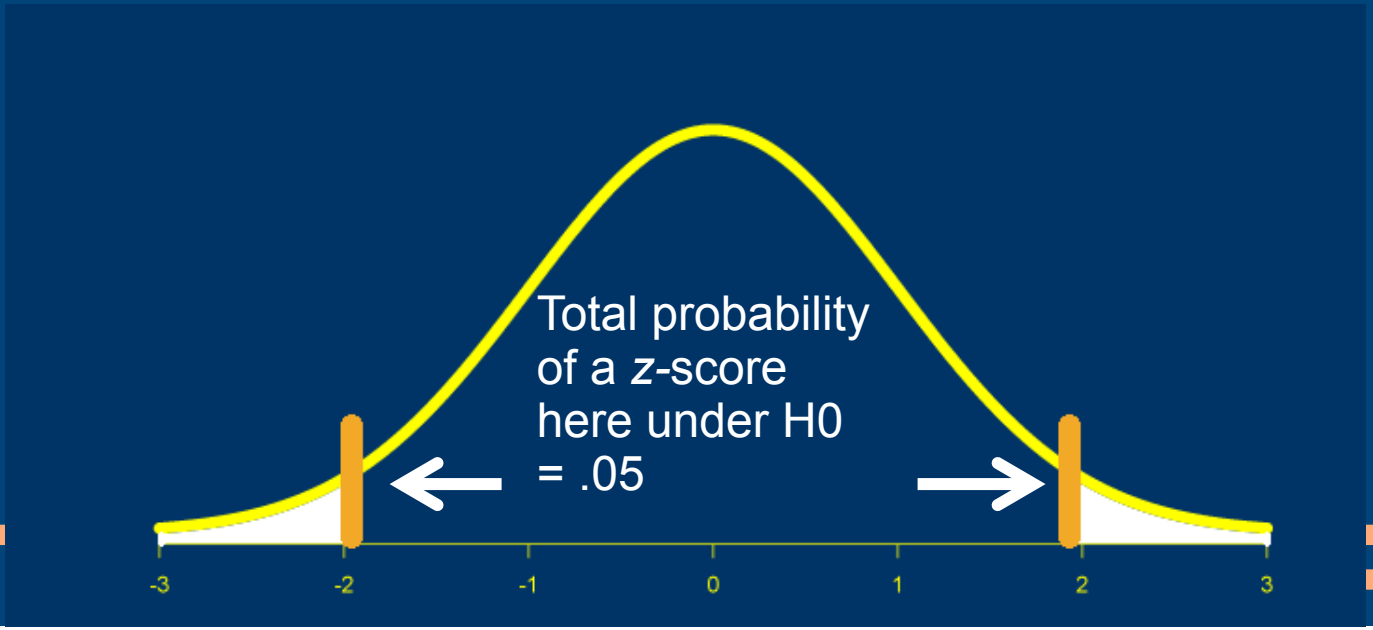
Type I Error

- *p*-value: Probability of obtaining a result this extreme under the null hypothesis of no effect
- We reject H_0 when the observed *t* or *z* has < .05 probability of arising under H_0
- But, still *possible* to get this *z* when H_0 is true
  - In that case, we'd incorrectly conclude that brain training works when it actually doesn't
  - **False positive** or **Type I error**



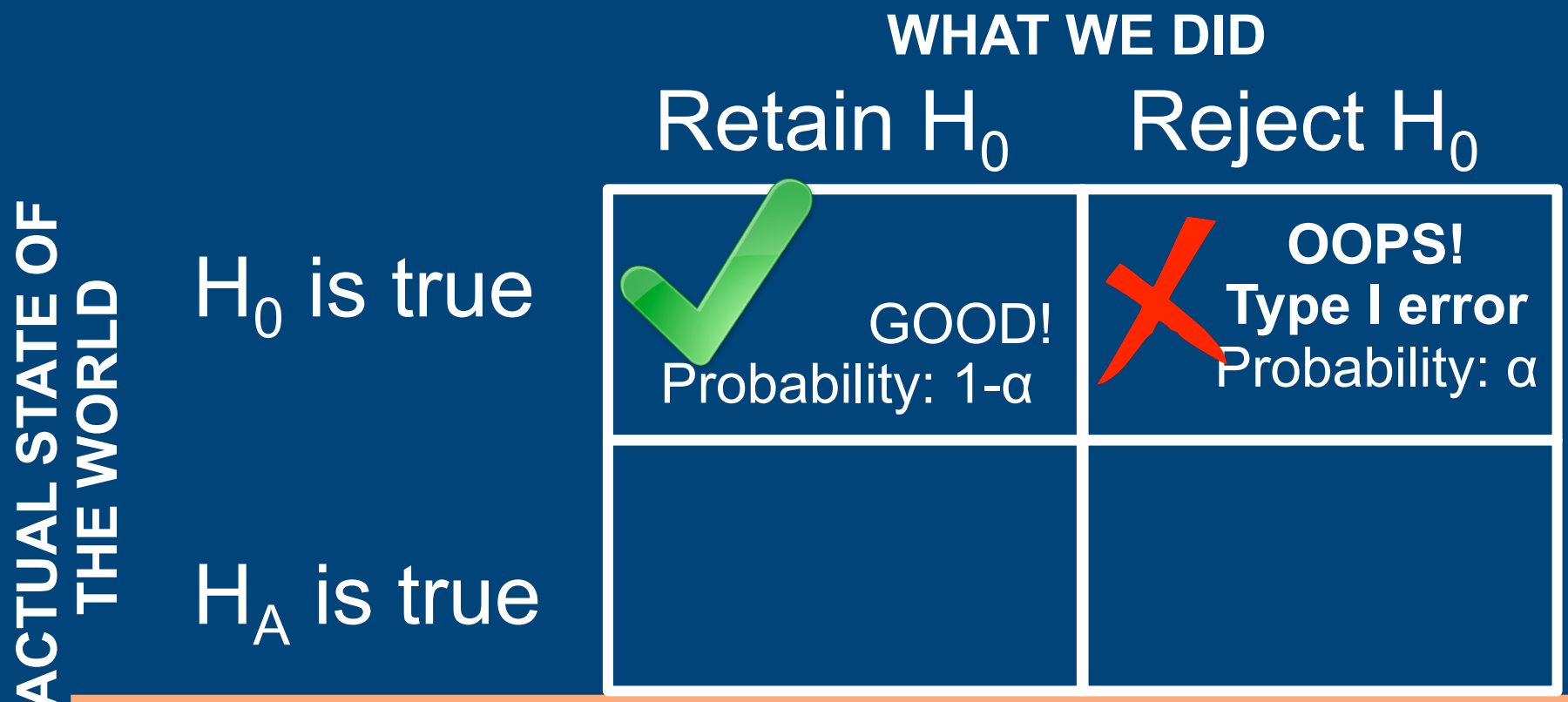
Type I Error

- What is our rate of Type I error?
  - Even in a world where H_0 is true, 5% of *z* values fall in white area
  - Thus, a 5% probability
  - α = rate of Type I error = .05



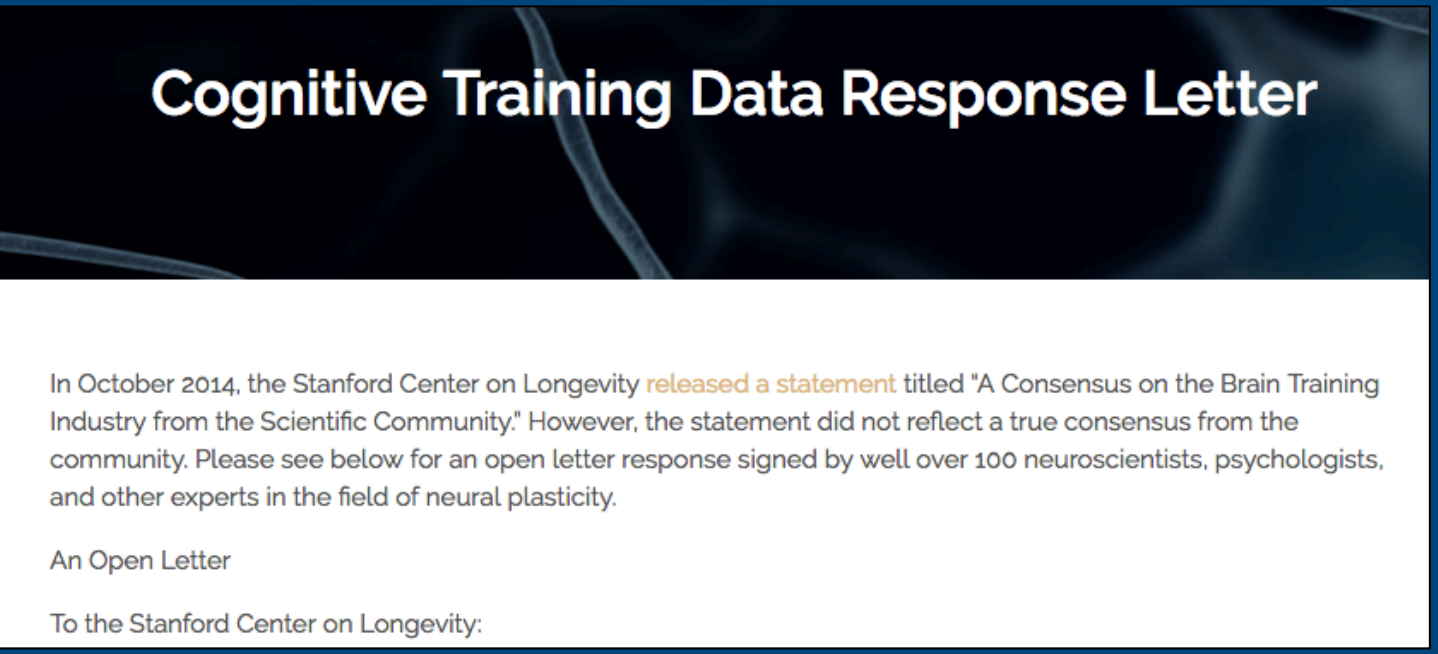
Type I Error and Type II Error

- So, in a world where **H_0** is true, two outcomes possible



Type I Error and Type II Error

- What about a world where H_A is true?



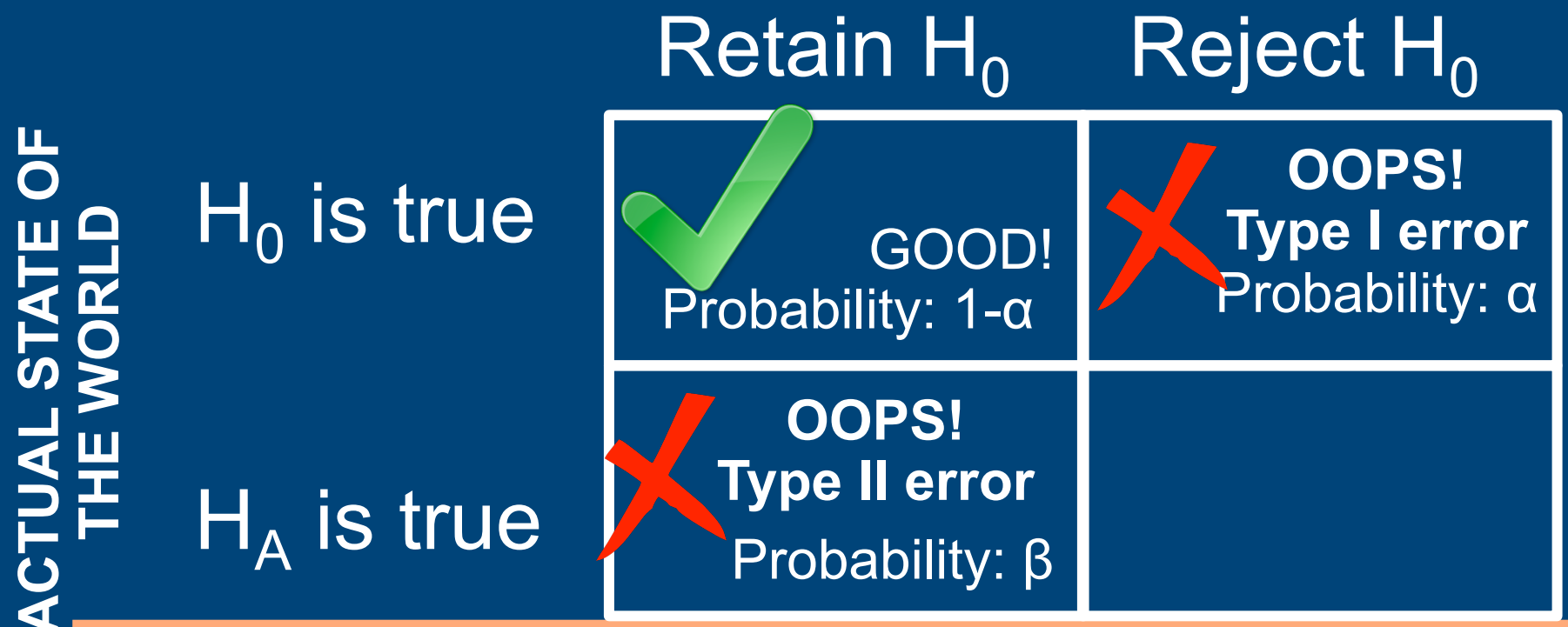
The Controversy

Does scientific evidence show brain training works?

Yes, although not all scientists agree. That may be because the question itself is flawed: it assumes all brain training is essentially the same.

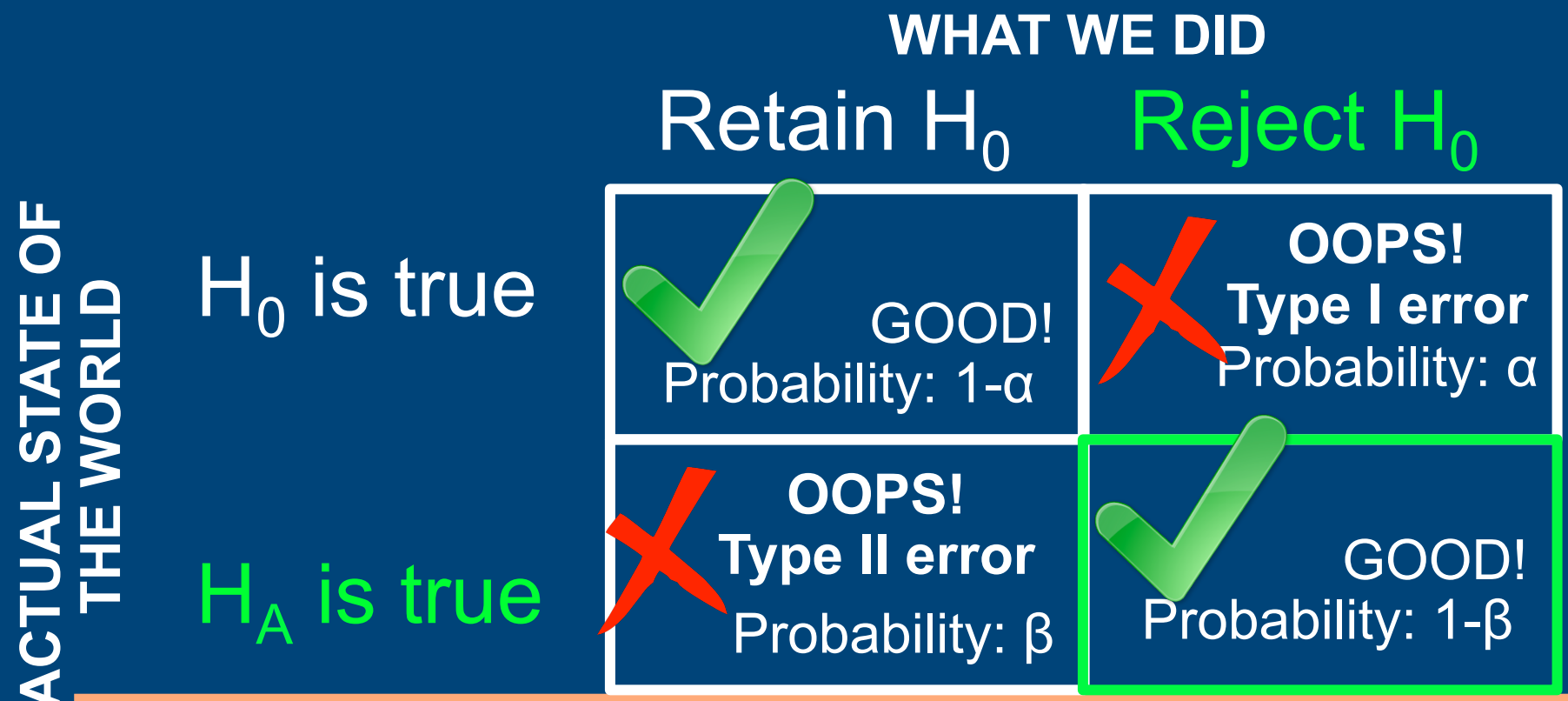
Type I Error and Type II Error

- Another mistake we could make: There really is an effect, but we retained H_0
  - **False negative / Type II error**
  - Traditionally, not considered as "bad" as Type I
  - Probability: β

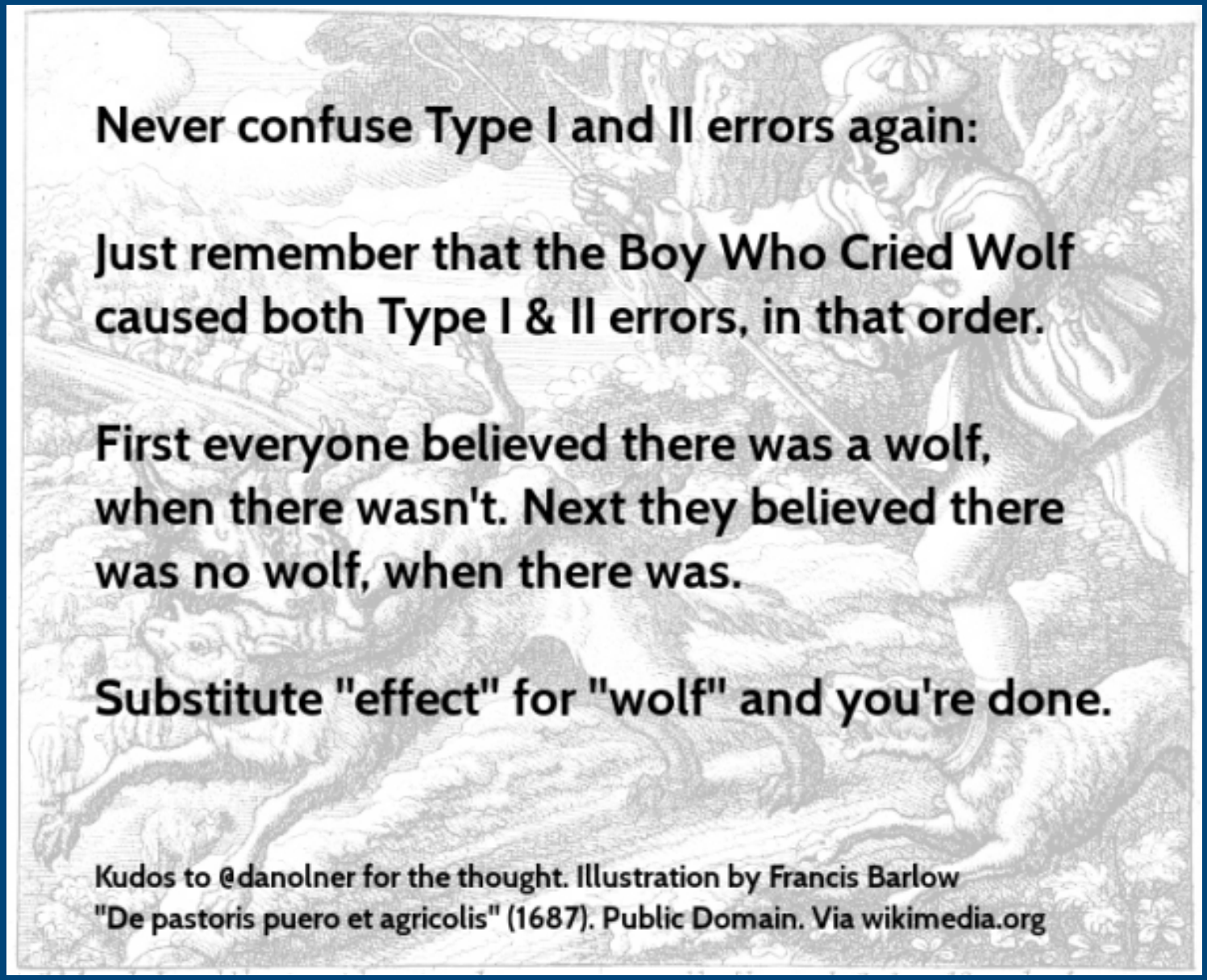


Type I Error and Type II Error

- **POWER (1-β):** Probability of correct rejection of H_0: detecting the effect when it really exists
  - If our hypothesis (H_A) is right, what probability is there of obtaining significant evidence for it?
  - Can we find the thing we're looking for?



Type I Error and Type II Error



Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - ~~Type I and Type II Error~~
  - Why Should We Care?
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Why Do We Care About Power?

1. **Grant agencies** now want to see it
   - Don't want to fund a study with low probability of showing anything
   - e.g., Our theory *predicts* greater activity in Broca's area in condition A than condition B. But our experiment has only a 16% probability of *detecting* that difference. Not good!

Why Do We Care About Power?

1. **Grant agencies** now want to see it
   - Don't want to fund a study with low probability of showing anything

2. **Efficiency**: Don't spend resources on studies with low power to find anything interesting
   - Societal resources: Money, participant hours
   - **Your** resources: Time!!

3. Interpreting **null effects**
   - Null effect of WM training on intelligence, 20% power
     - Maybe effect exists & we just didn't detect it
   - Null effect of WM training on intelligence, 80% power
     - Informative!!

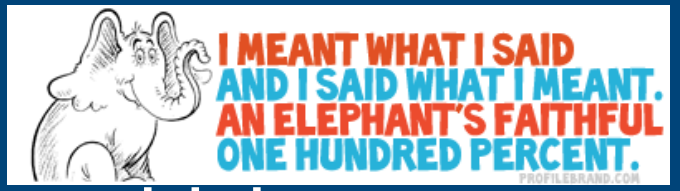
Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - ~~Type I and Type II Error~~
  - ~~Why Should We Care?~~
  - Assessing Power
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Data Simulations

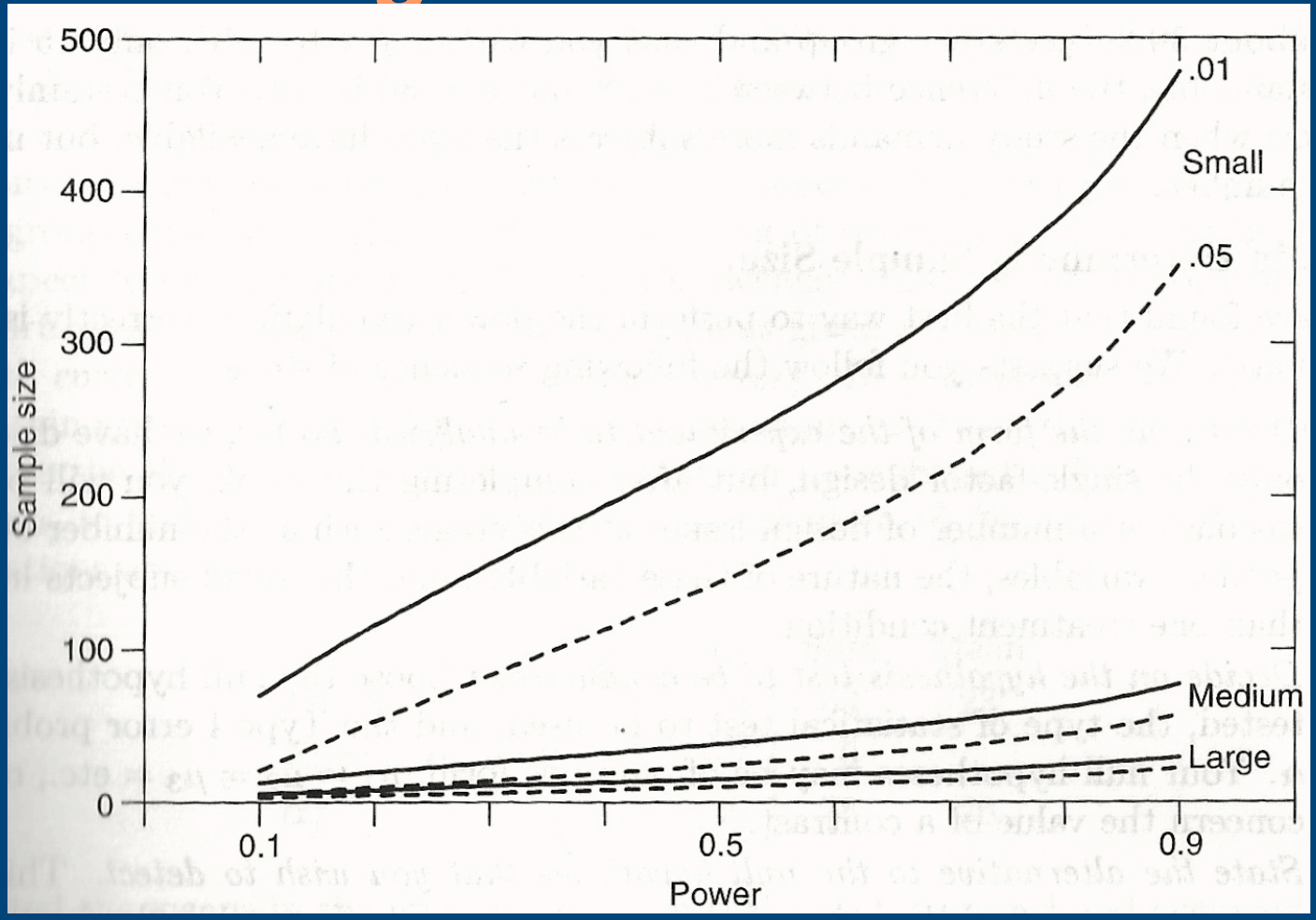
- If we say "α = .05"…
  - Significant differences should be false positives 5% of the time
  - BAD if test yields more false positives than claimed

- Is this true for a given test?
  - i.e., what proportion of our significant differences are false positives?
  - Achieved **nominal** false positive rate if the rate is indeed what we said our α is



- Problem: We usually don't *know* which differences truly exist in the population
  - That's what we're doing the study to find out!

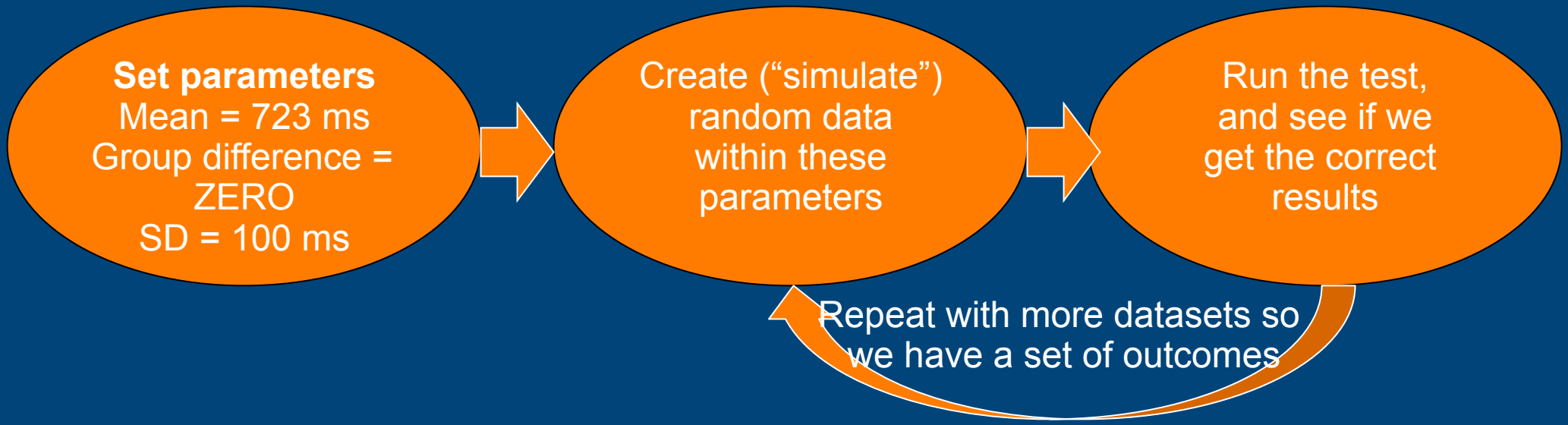
Determining Power



- Power for ANOVAs can be easily found from tables
  - Simpler design. Only 1 random effect (at most)
- More complicated for mixed effect models

Data Simulations

- Solution: Simulate data where we know what the results should be



- A way of evaluating statistical procedures
  - When there **is no** actual group difference, how often do we get false positives (Type I errors)?
  - When there **is** an actual group difference, what is our **power** to detect it?

Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - ~~Type I and Type II Error~~
  - ~~Why Should We Care?~~
  - ~~Assessing Power~~
  - Power of Mixed Effect Models
  - Doing Your Own Power Analysis

Mixed Effect Model Simulations: Results

	DESIGN / RANDOM EFFECTS	COMPARSION METHOD	CONTROL OF TYPE I ERROR	POWER
Barr et al. (2013) maximal model	2 CROSSED (BETWEEN OR WITHIN ITEMS)	ANOVA	=	+
Barr et al. (2013) intercepts only	2 CROSSED (BETWEEN OR WITHIN ITEMS)	ANOVA	-	n.a.
Quene & van den Bergh (2004)	1 (WITHIN ITEMS)	1 RM-ANOVA	n.a.	+
Quene & van den Bergh (2004)	2 (WITHIN ITEMS)	2 RM-ANOVAs	=	=
Baayen, Davidson, & Bates (2008) - 1	2 CROSSED (BETWEEN ITEMS)	2 RM-ANOVAs	=/- N=40 N=20	+ "especially with missing data"

Mixed Effect Model Simulations: Results

	DESIGN / RANDOM EFFECTS	COMPARSION METHOD	CONTROL OF TYPE I ERROR	POWER
Barr et al. (2013) maximal model	2 CROSSED (BETWEEN OR WITHIN ITEMS)	ANOVA	=	+
Barr et al. (2013) intercepts only	2 CROSSED (BETWEEN OR WITHIN ITEMS)	ANOVA	-	n.a.
Quene & van den Bergh (2004)	1 (WITHIN ITEMS)	1 RM-ANOVA	n.a.	+
Quene & van den Bergh (2004)	2 (WITHIN ITEMS)	2 RM-ANOVAs	=	=
Baayen, Davidson, & Bates (2008) - 1	2 CROSSED (BETWEEN ITEMS)	2 RM-ANOVAs	=/-	+
Baayen, Davidson, & Bates (2008) - 2	2 CROSSED (WITHIN ITEMS)	1 RM-ANOVA	=	+
Baayen, Davidson, & Bates (2008) - 3	2 CROSSED (BETWEEN ITEMS)	REGRESSION	+	n.a.

Data Simulations: Conclusions

- Type I error rates roughly **equal**
  - Assuming you do mixed effects models correctly
- Mixed effects models are **more powerful**
  - By-subjects ANOVA doesn't remove noise from item variability
  - By-items ANOVA doesn't remove noise from subject variability
  - Mixed effects models account for *both r*andom effects—data less noisy

Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - ~~Type I and Type II Error~~
  - ~~Why Should We Care?~~
  - ~~Assessing Power~~
  - ~~Power of Mixed Effect Models~~
  - Doing Your Own Power Analysis

Your Own Power Analysis

- Rationale behind power analyses:
  - Can we detect the kind & size of effect we're interested in?
  - What sample size would we need?

- In practice:
  - We can't control effect size; it's a property of nature
  - α is usually fixed (e.g., at .05) by convention
  - But, we **can** control our sample size *n*!
- So:
  - Determine desired power (often .80)
  - Estimate the effect size(s)
  - Calculate the necessary sample size *n*

Your Own Power Analysis

- Rationale behind power analyses:
  - Can we detect the kind & size of effect we're interested in?
  - What sample size would we need?

- Two ways to do this:
  - Use tables/software for ANOVA (e.g. G*Power)
    - Mixed effect models, if anything, will have *at least this much power* or *more*
  - Apply the simulation procedure to *your* design
    - Your fixed effect sizes
    - Your random effects structure & variance

Estimating Effect Size

- One reason we haven't always calculated power is it requires the effect size
- But, several ways to **estimate** effect size:
  1. **Prior literature**
     - What is the effect size in other studies in this domain or with a similar manipulation?

Estimating Effect Size

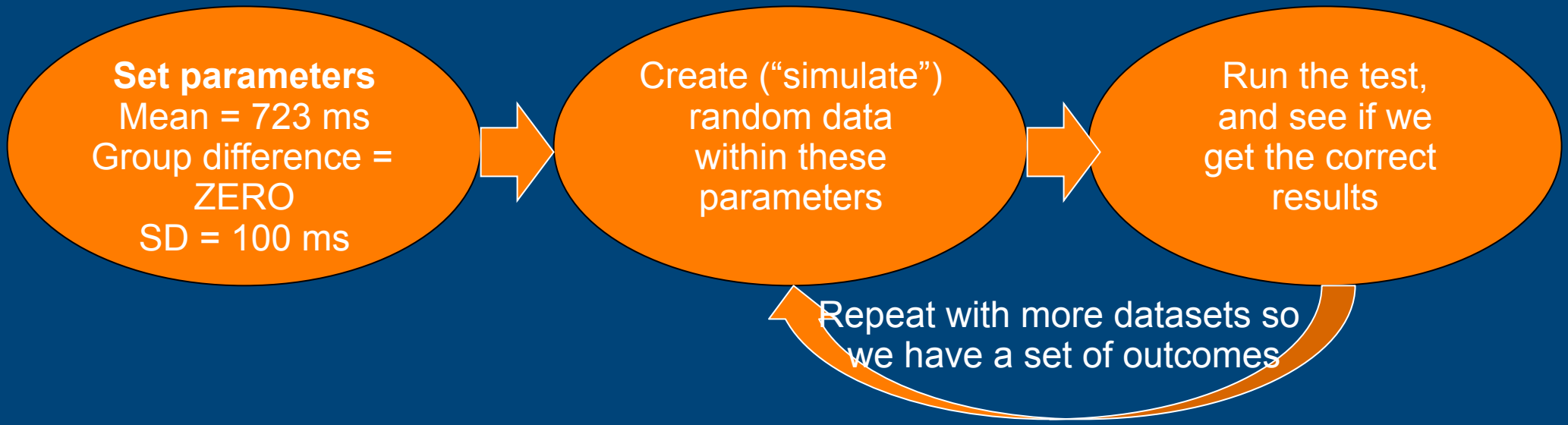
- One reason we haven't always calculated power is it requires the effect size
- But, several ways to **estimate** effect size:
  1. **Prior literature**
  2. **Pilot study**
     - Run a version of the study with a smaller *n*
     - Don't worry about whether effect is significant, just use data to estimate ω^2

Estimating Effect Size

- One reason we haven't always calculated power is it requires the effect size
- But, several ways to **estimate** effect size:
  1. **Prior literature**
  2. **Pilot study**
  3. Smallest **interesting** effect
     - Decide smallest effect size we'd care about
       - e.g., we want our educational intervention to have an effect size of at least .05 GPA
     - Calculate power based on that effect size
       - True that if actual effect is smaller than .05 GPA, our power would be lower, but the idea is *we no longer care* about the intervention if its effect is that small

Data Simulations

- Simulate data using your fixed effect sizes & random effects variances



- What sample size(s) do you need in order to detect the effect 80% of the time?
  - Will 40 subjects in each of 5 schools suffice?
  - What about 40 subjects in 10 schools?

Week 9: Effect Size & Power

- ~~Distributed Practice~~
- Finish `glmer()`
  - ~~Interactions~~
  - ~~Coding the Dependent Variable~~
  - ~~Other Distributions~~
- ~~Effect Size~~
- Power
  - ~~Type I and Type II Error~~
  - ~~Why Should We Care?~~
  - ~~Assessing Power~~
  - ~~Power of Mixed Effect Models~~
  - ~~Doing Your Own Power Analysis~~