

# Session 09: Hypothesis Testing

---

Stats 60/Psych 10  
Ismael Lemhadri  
Summer 2020

# This time (and next week)

---

- Hypothesis testing
- What p-values mean - and don't mean
- Connection to z-scores

# The three fundamental goals of statistics

---

- Describe
  - Decide
  - Predict
- 
- Hypothesis testing provides us with a tool to make decisions in the face of uncertainty using data

# Do checklists improve surgical outcomes?

## A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population

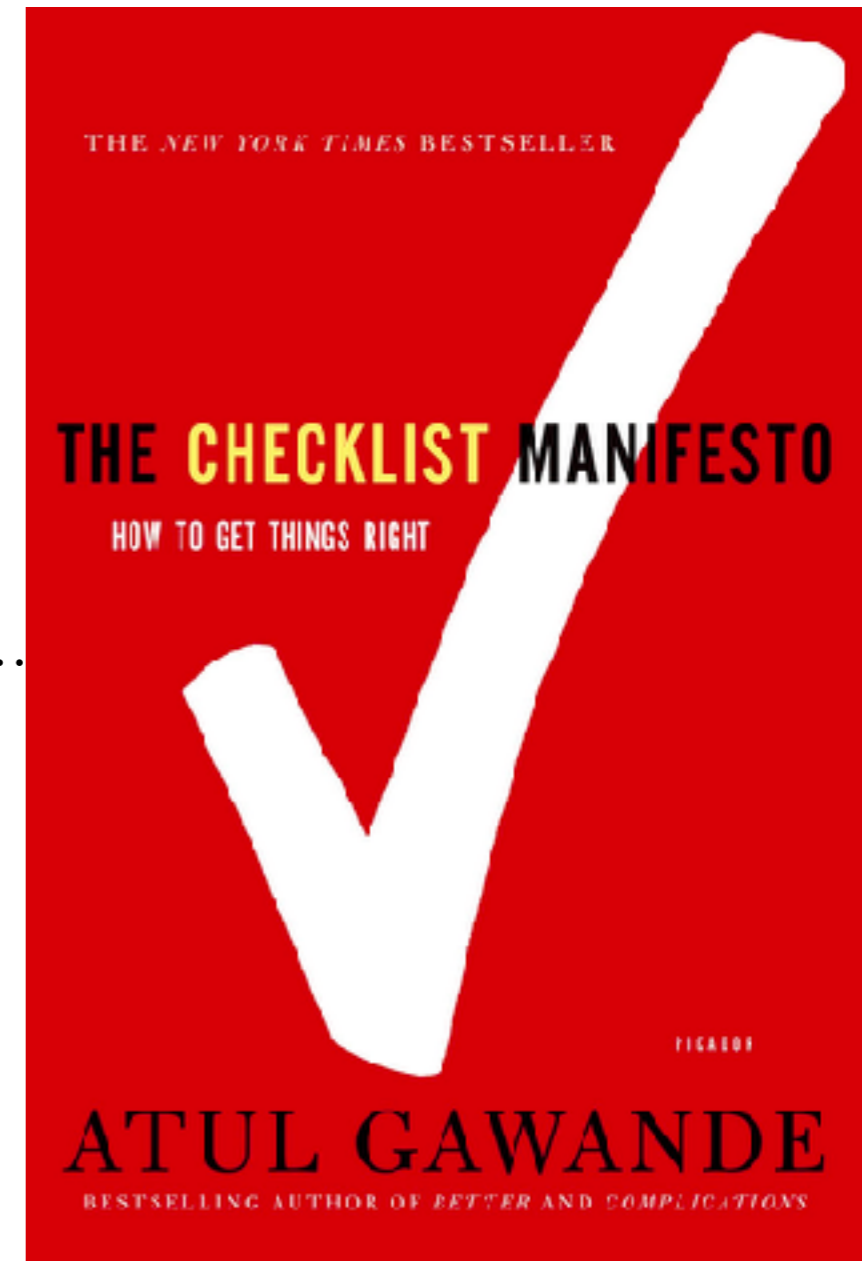
N ENGL J MED 360;5 NEJM.ORG JANUARY 29, 2009

We hypothesized that a program to implement a 19-item surgical safety checklist designed to improve team communication and consistency of care would reduce complications and deaths associated with surgery.

Between October 2007 and September 2008, eight hospitals in eight cities... participated in the World Health Organization's Safe Surgery Saves Lives program.

The rate of death was 1.5% before the checklist was introduced and declined to 0.8% afterward ( $P = 0.003$ ). Inpatient complications occurred in 11.0% of patients at baseline and in 7.0% after introduction of the checklist ( $P < 0.001$ ).

Huh?



# Do body-worn cameras improve policing?

- 2,224 DC Metro PD officers randomly assigned to wear BWC or not
- Compared use of force and number of complaints between groups

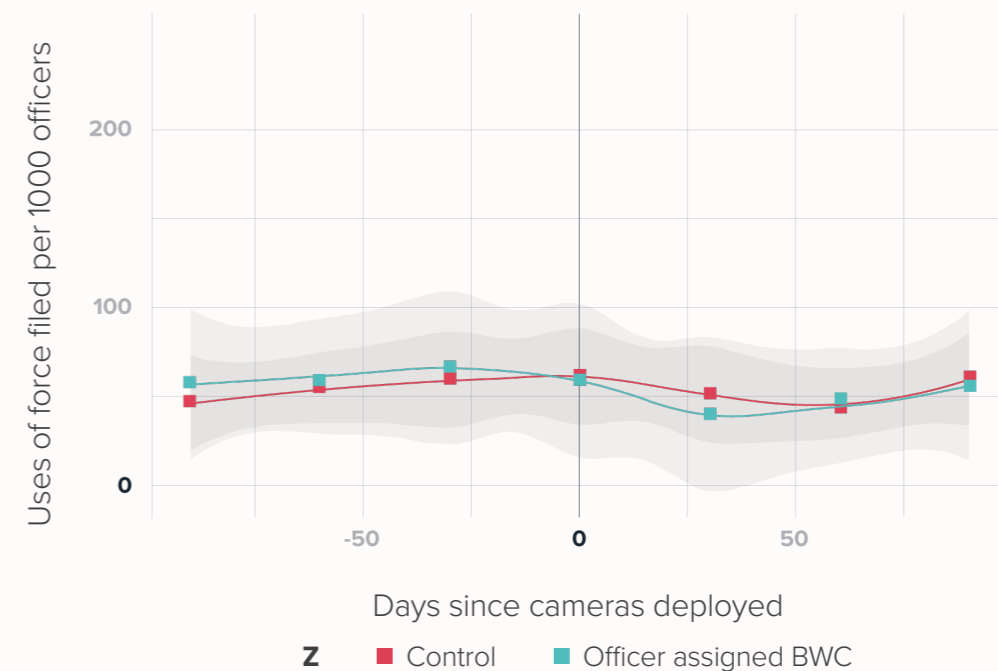


# Body worn cameras: no effect on policing outcomes

- “We are unable to reject the null hypotheses that BWCs have no effect on police use of force, citizen complaints, policing activity, or judicial outcomes.”
- Did they just use a triple negative?
  - “unable to reject the null hypotheses”

**FIG. 4. Uses of Force per 1,000 Officers, 90 days before and after BWC deployment.**

This figure plots pre- and post-treatment uses of force for both control and treatment group officers. As the chart indicates, there is no statistically significant difference between the two groups in either the 90-day period before or after the deployment of BWCs (which occurs on day 0).



# “Null hypothesis statistical testing” (NHST)

---

- The most commonly used approach to perform statistical tests
  - Gerrig & Zimbardo (2002): NHST is the “backbone of psychological research”
- Almost all researchers continue to use it
- Many people think that it’s a bad way to do science
  - Bakan (1966): “The test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research”
  - Luce (1988): Hypothesis testing is “a wrongheaded view about what constitutes scientific progress”

# Prepare yourself for mental gymnastics

---

- Hypothesis testing is notoriously difficult to understand
- Because it's built in a way that violates our natural intuitions!





# How you might think hypothesis testing should work

---

- We start with a hypothesis
  - Body-worn cameras will reduce police misconduct
- We collect some data
  - Randomized controlled trial comparing BWC to no BWC
- We determine whether the data provide convincing evidence in favor of the hypothesis
  - What is the likelihood that the hypothesis is true, given the data along with everything else we know?

# How null hypothesis testing actually works

---

- We start with a hypothesis
  - Body-worn cameras will reduce police misconduct
- We flip it to generate a “null hypothesis”, which we assume is true
  - There is no effect of BWCs on police misconduct
- We collect some data
  - Randomized controlled trial comparing BWC to no BWC
- We determine how likely the data would have been, assuming that the hypothesis is wrong
  - If it is unlikely, then we we decide that we can “reject the null hypothesis “
  - If it is likely, then we “fail to reject the null hypothesis”
    - This doesn't mean that we decide that there is no effect!

# Why do you think we are spending two sessions talking about something that so many people think is a bad idea?

**Top**

# The steps of null hypothesis testing

---

1. Make predictions based on your hypothesis (*before seeing the data*)
2. Collect some data
3. Identify null and alternative hypotheses
4. Fit a model to the data that represents the alternative hypothesis and compute a test statistic
5. Compute the probability of the observed value of that statistic assuming that the null hypothesis is true
6. Assess the “statistical significance” of the result

# An example hypothesis: Is physical activity related to body mass index?

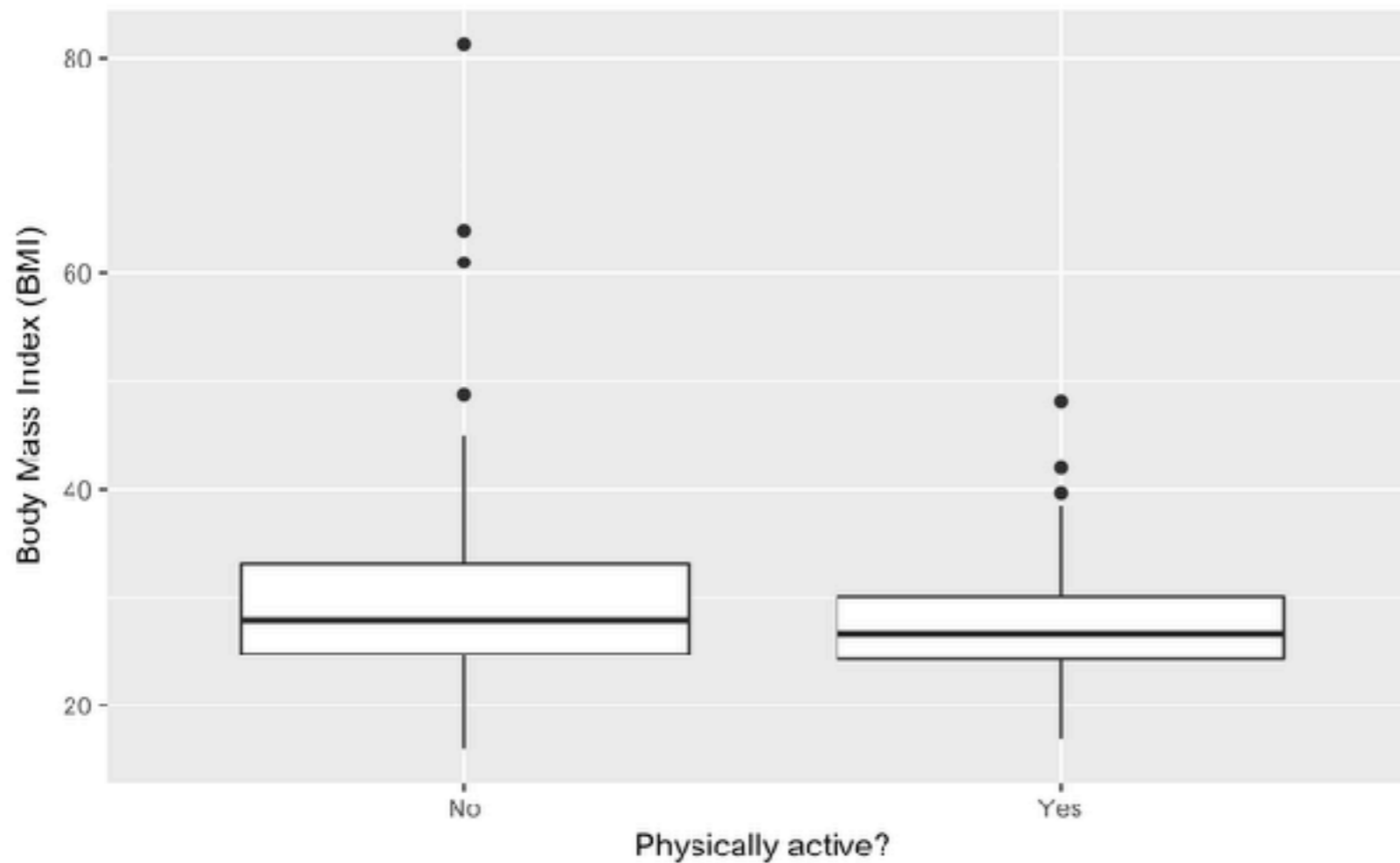
---

- In the NHANES dataset, participants were asked whether they engage regularly in moderate or vigorous-intensity sports, fitness or recreational activities
- Also measured height and weight and computed Body Mass Index

$$BMI = \frac{Weight(kg)}{Height(m)^2}$$

- Hypothesis of interest: BMI is related to physical activity
- Prediction: BMI should be greater for inactive vs. active individuals

## Step 2: Collect some data

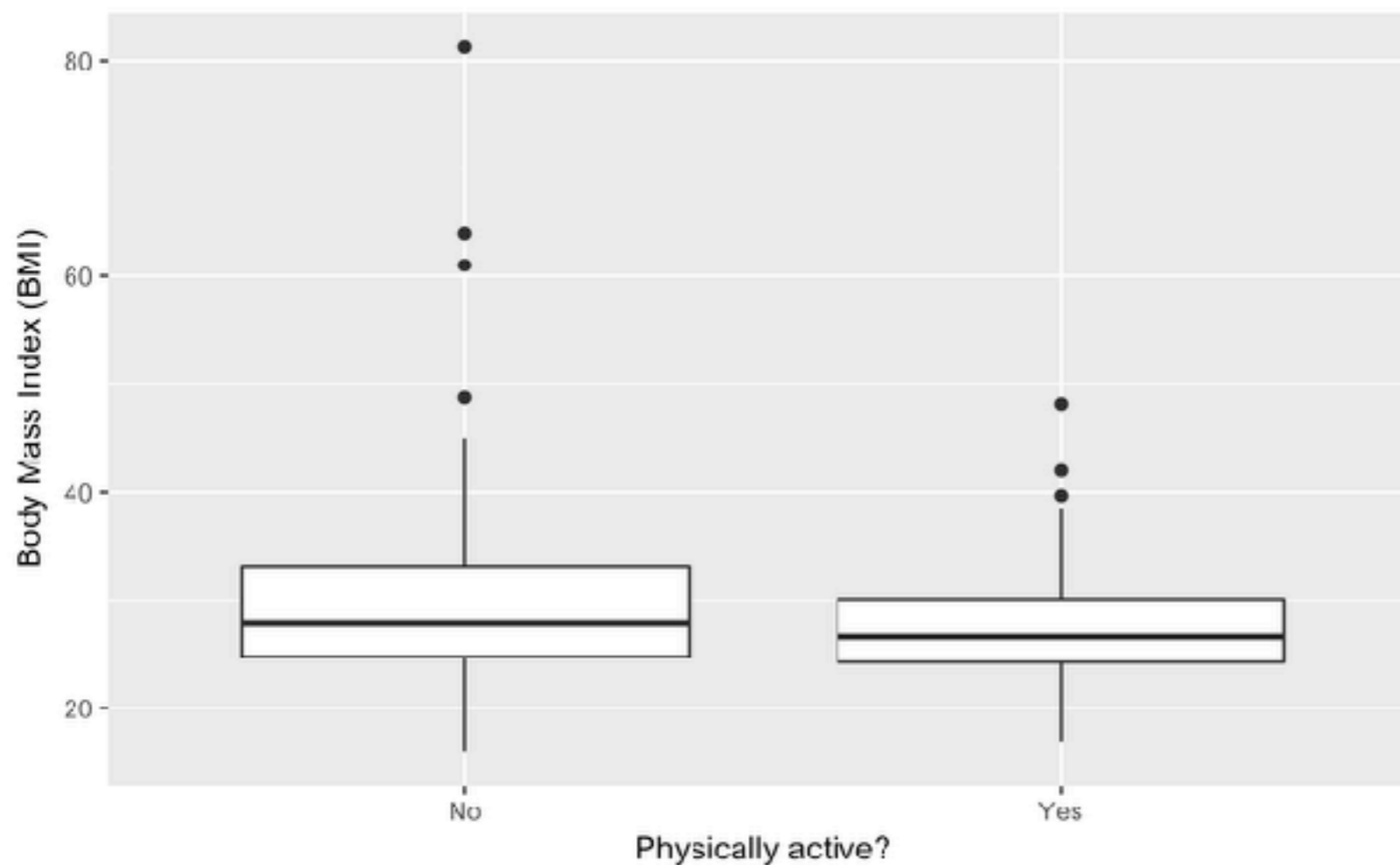


	N	mean BMI	SD
Active	125	27.41	5.07
Not Active	125	29.64	8.83

250 individuals sampled from NHANES

# Exercise: compute confidence intervals

- What are the confidence intervals for the mean for each group?



	N	mean BMI	SD
Active	125	27.41	5.07
Not Active	125	29.64	8.83

## Step 3: What are the “null hypothesis” ( $H_0$ ) and “alternative hypothesis” ( $H_A$ )?

---

- $H_0$ : The baseline against which we test our hypothesis of interest
  - What would the data look like if there was no effect?
  - Always involves some kind of equality ( $=$ ,  $\leq$ , or  $\geq$ )
- This is compared to an “alternative hypothesis” ( $H_A$ )
  - What we expect if there actually is an effect
  - Always involves some kind of inequality ( $\neq$ ,  $>$ , or  $<$ )
- *Null hypothesis testing operates under the assumption that the null hypothesis is true*



# BMI example: Null and alternative hypotheses

---

- $H_A$ :
  - BMI for active people is less than BMI for inactive people in the population
    - $\mu_{\text{active}} < \mu_{\text{inactive}}$
    - This is a “directional” hypothesis
  - Could also have a “non-directional” hypothesis
    - $\mu_{\text{active}} \neq \mu_{\text{inactive}}$
- $H_0$ :
  - BMI for active people is greater than or equal to BMI for inactive people in the population
    - $\mu_{\text{active}} \geq \mu_{\text{inactive}}$
    - $\mu_{\text{active}} = \mu_{\text{inactive}}$  (for non-directional  $H_A$ )

# Step 4: Fit a model to the sample data and compute a test statistic

---

$$\text{test statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{effect}}{\text{error}}$$

- The test statistic quantifies the amount of evidence against the null hypothesis, compared to the noise in the data
- It usually has a probability distribution associated with it
  - if not, then we can often compute one using simulation

# BMI: What is our test statistic of interest?

- “Student’s  $t$ ” statistic
  - Measures the difference of means between two groups
  - Distributed according to a  $t$  distribution when the sample size is small and the population SD is unknown

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

$\bar{X}_1$  : sample mean       $S_1^2$  : sample variance

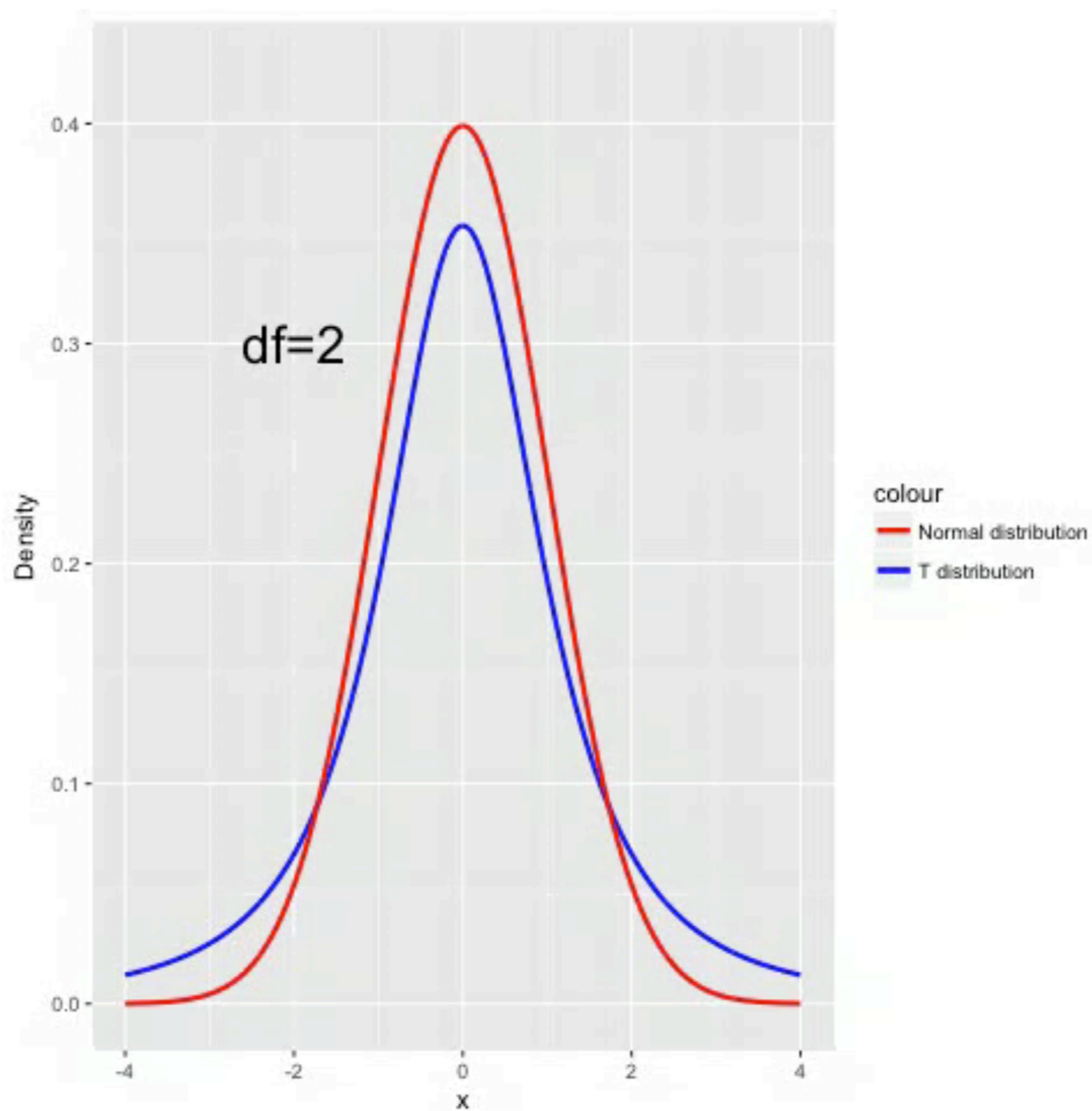
$N_1$  : sample size



Statistician William Sealy Gosset, AKA “Student”



# The t distribution vs. the normal (Z) distribution



## Step 5: Determine the probability of the test statistic under the null hypothesis

---

- How likely is it that we would see an effect of this size if there really is no effect?
- To do this, we need to know the distribution of the statistic under the null hypothesis
- We can then ask how likely our observed value is within that distribution
- Two ways to determine this:
  - Theoretical distribution
  - Null distribution obtained using simulation

# A simple example: Is this coin fair?



- Do an experiment: 100 flips
- Statistic of interest: 70 heads
- $H_0$ :  $p(\text{heads})=0.5$
- $H_A$ :  $p(\text{heads}) \neq 0.5$
- How likely are we to observe 70 heads on 100 flips if  $H_0$  is true?

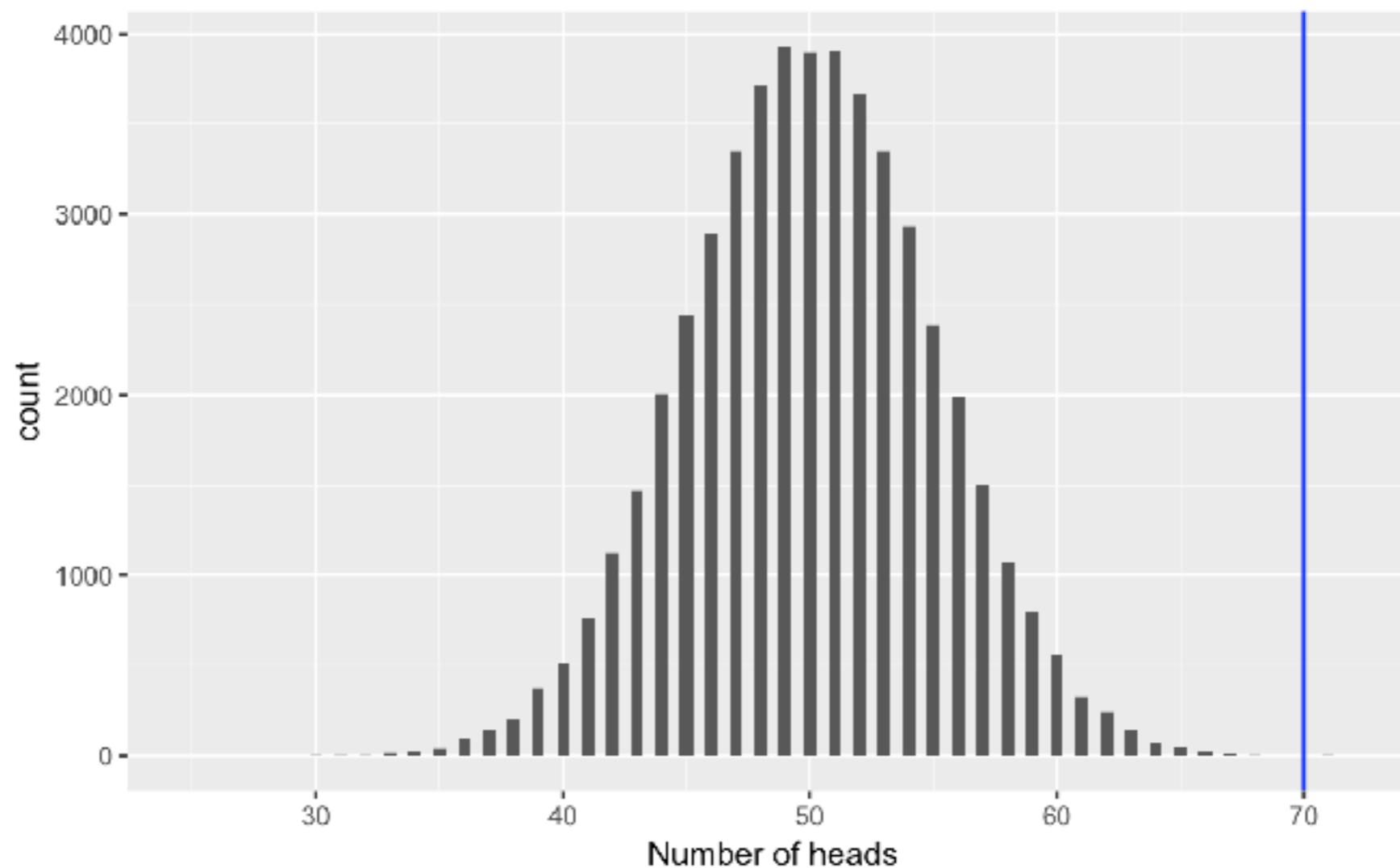
binomial distribution 
$$P(X \leq k) = \sum_{i=0}^k \binom{N}{k} p^i (1-p)^{n-i}$$

$$P(X \leq 69|p=0.5) = 0.99996$$

$$P(X \geq 70|p=0.5) = 1 - 0.99996 = 0.00004$$

# Using random sampling to generate an empirical null distribution

- Draw random samples from a binomial distribution (using `rbinom()`)
- Compare them to the observed data



$$P(X \geq 70 | p=0.5) = 3/500000 = 0.000006$$

# BMI example

---

- What would the t statistic look like if there was really no difference in BMI between active and inactive people?



# Randomization

---

- We can make the null hypothesis true (on average) by randomly reordering group membership

Team	Squat
Football	325
Football	290
Football	290
Football	305
Football	370
XC	165
XC	180
XC	215
XC	175
XC	125

$$t = 6.92$$

$$df = 8,$$

$$p(t_8 \geq 6.92) = 0.0001$$

# Randomization

---

- We can make the null hypothesis true (on average) by randomly reordering group membership

Team	Squat
Football	325
Football	290
XC	290
XC	305
Football	370
Football	165
Football	180
XC	215
XC	175
XC	125

$$t = 0.83$$

$$df = 8$$

$$p(t_8 \geq 0.83) = 0.43$$

# Randomization

---

- We can make the null hypothesis true (on average) by randomly reordering group membership

Team	Squat
XC	325
XC	290
Football	290
Football	305
Football	370
XC	165
Football	180
Football	215
XC	175
XC	125

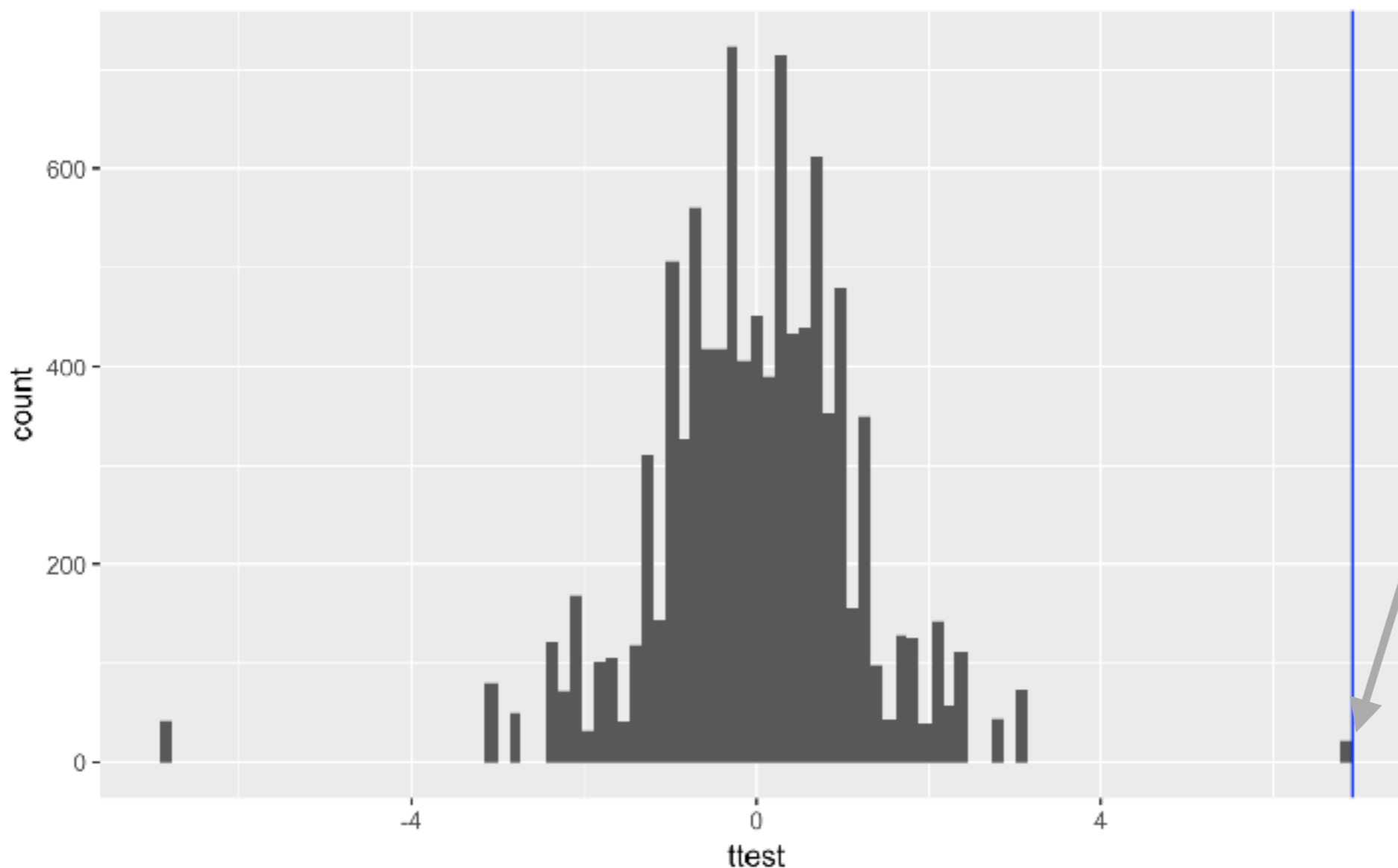
$$t = 1.09$$

$$df = 8$$

$$p(t_8 \geq 1.09) = 0.30$$

- Scramble 10,000 times to get distribution of t values under null hypothesis

$$P(t_{\text{random}} \geq t_{\text{observed}}) = .0021$$

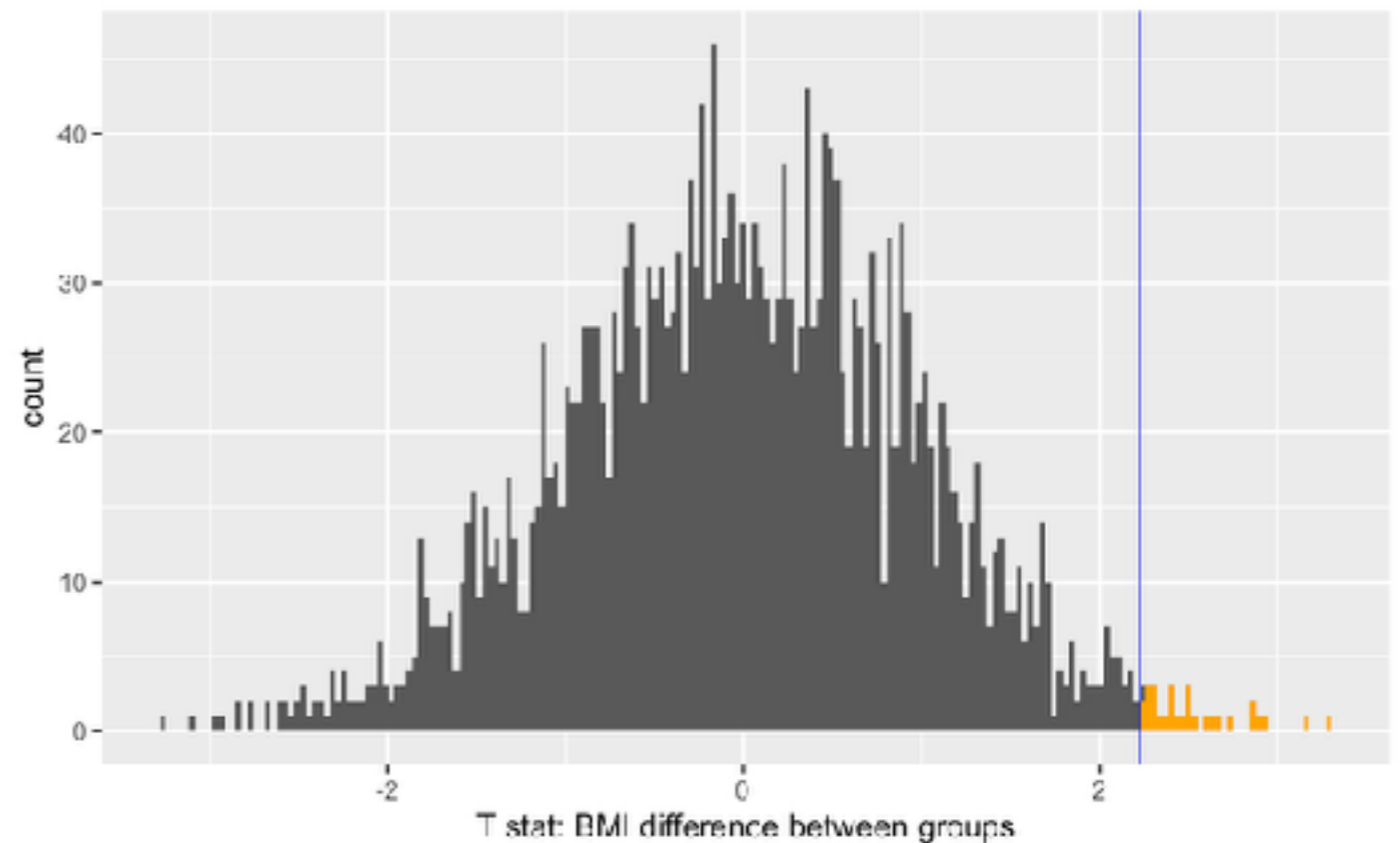


What  
happened  
here?

there are  
3,628,800  
possible  
permutations of  
10 items

# BMI example: randomization

- If there is no difference between groups, then the result should be no different from what we see if activity levels are randomly shuffled between people



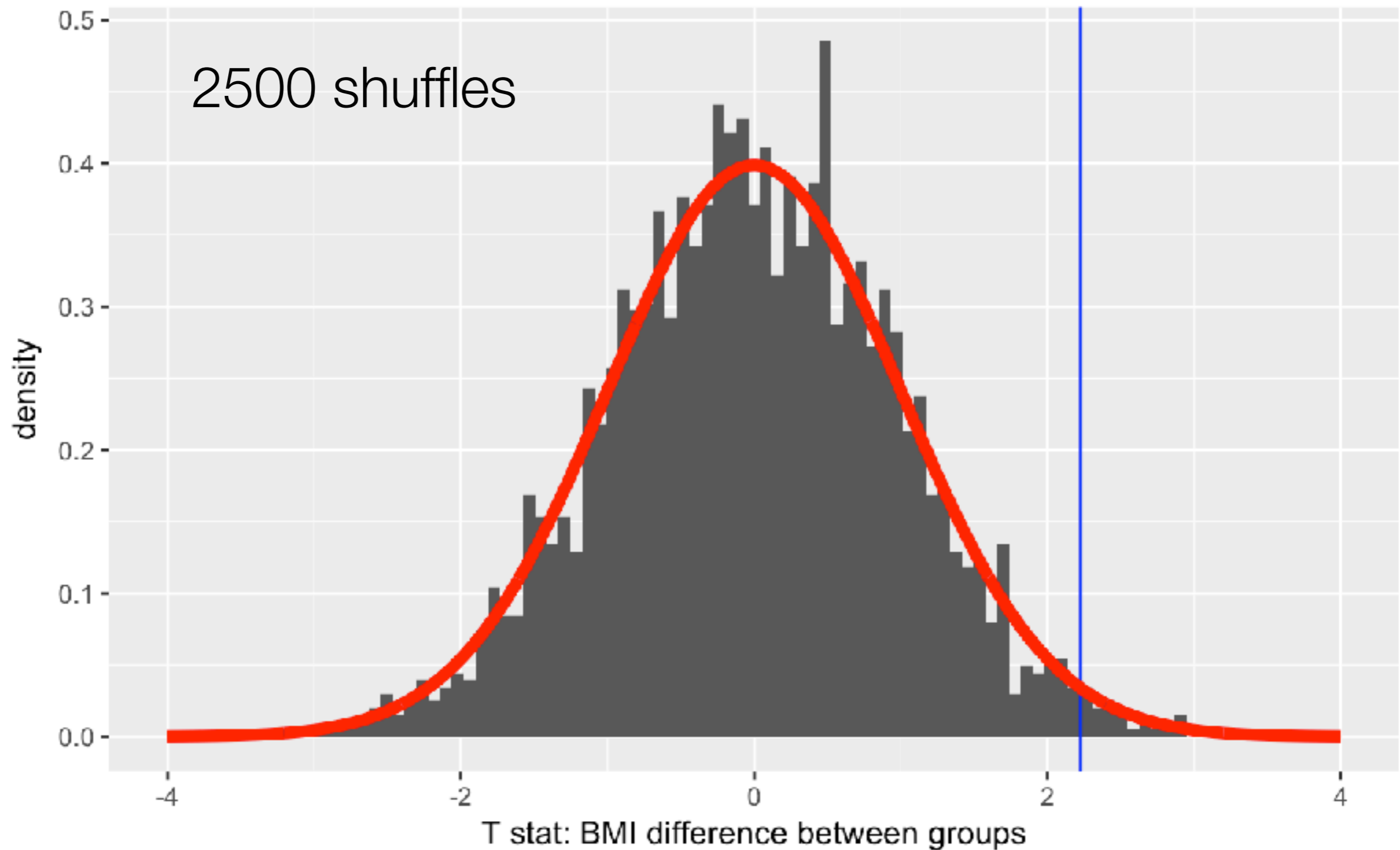
Largest difference in 2500 random shuffles: 3.21

Observed difference in actual data: 2.22

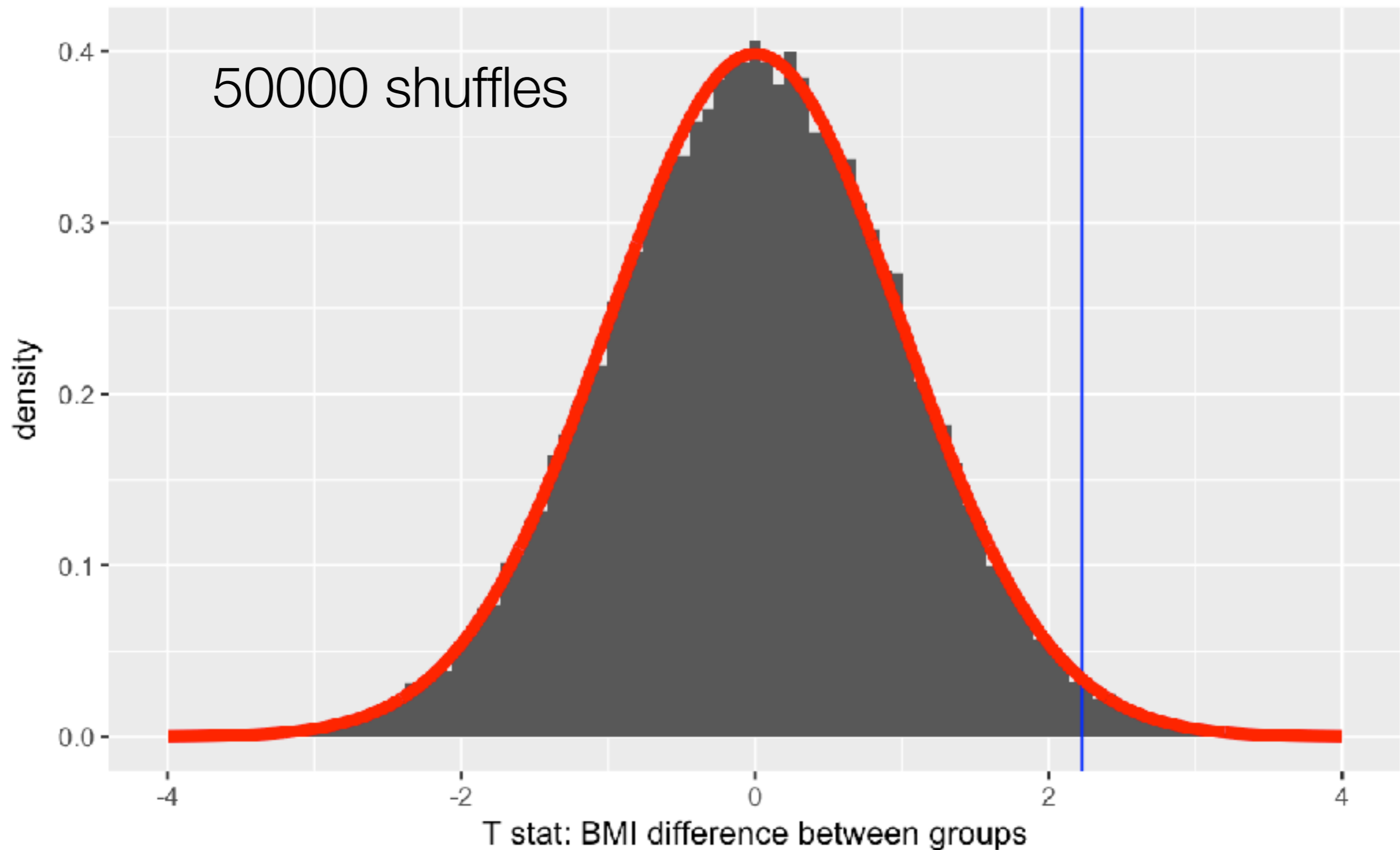
Number of shuffles with  $t \geq 2.22$ : 16

$$p(t \geq 2.22 | H_0) = 16/2500 = 0.0064$$

# The $t$ distribution vs permutation distribution



With enough random shuffles, the nonparametric and theoretical distributions can become very similar



Performing a  $t$  test in R“formula notation:  $y \sim x$ ”

---

```
ttestResult = t.test(BMI~PhysActive,  
data=NHANES_sample,var.equal=TRUE,  
alternative='greater')
```

BMI ~ PhysActive →

“Does BMI differ as a function of the different values of PhysActive?”



# BMI example: dissecting the parametric test in R

---

```
>ttestResult <- t.test(BMI~PhysActive,data=NHANES_sample,  
  var.equal=TRUE,alternative='greater')
```

# BMI example: parametric test in R

---

```
>ttestResult=t.test(BMI~PhysActive,data=NHANES_sample,  
var.equal=TRUE,alternative='greater')
```

Two Sample t-test

data: BMI by PhysActive

t = 2.4452, df = 248, p-value = 0.007587

alternative hypothesis: true difference in means is  
greater than 0

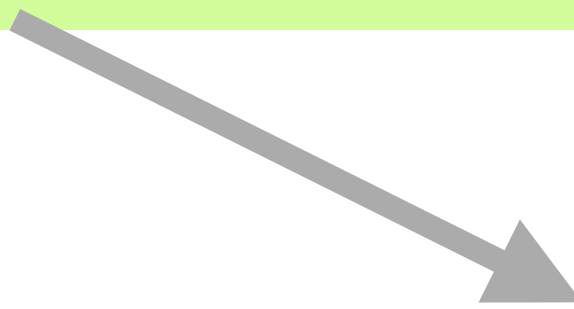
95 percent confidence interval:

0.7230215            Inf

sample estimates:

mean of x mean of y

29.63752    27.41136



Directional  
alternative  
hypothesis

# BMI example: parametric test in R

---

```
>ttestResult=t.test(BMI~PhysActive,data=NHANES_sample,  
var.equal=TRUE,alternative='greater')
```

Two Sample t-test

data: BMI by PhysActive

**t = 2.4452**, df = 248, p-value = 0.007587

alternative hypothesis: true difference in means is  
greater than 0

95 percent confidence interval:

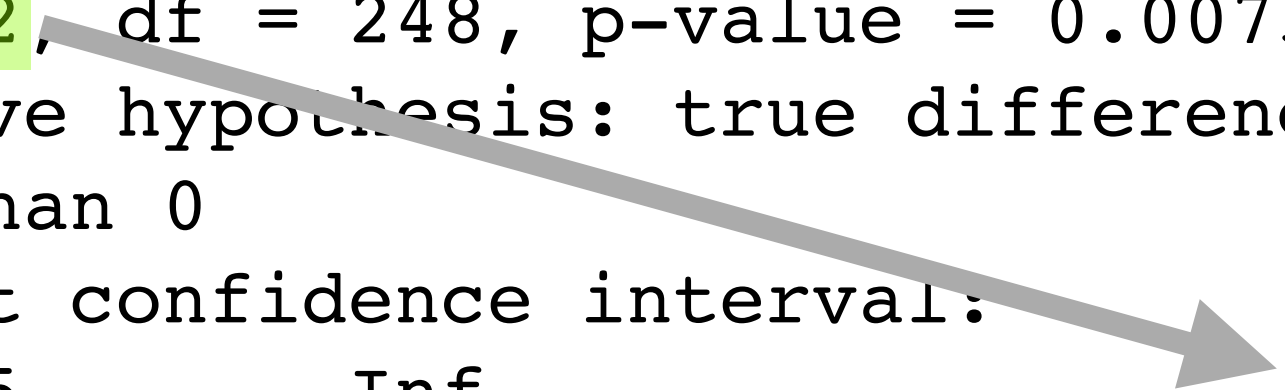
0.7230215                      Inf

sample estimates:

mean of x mean of y

29.63752    27.41136

t statistic  
computed on  
observed sample



# BMI example: parametric test in R

---

```
>ttestResult=t.test(BMI~PhysActive,data=NHANES_sample,  
var.equal=TRUE,alternative='greater')
```

## Two Sample t-test

```
data: BMI by PhysActive
```

```
t = 2.4452, df = 248, p-value = 0.007587
```

```
alternative hypothesis: true difference in means is  
greater than 0
```

```
95 percent confidence interval:
```

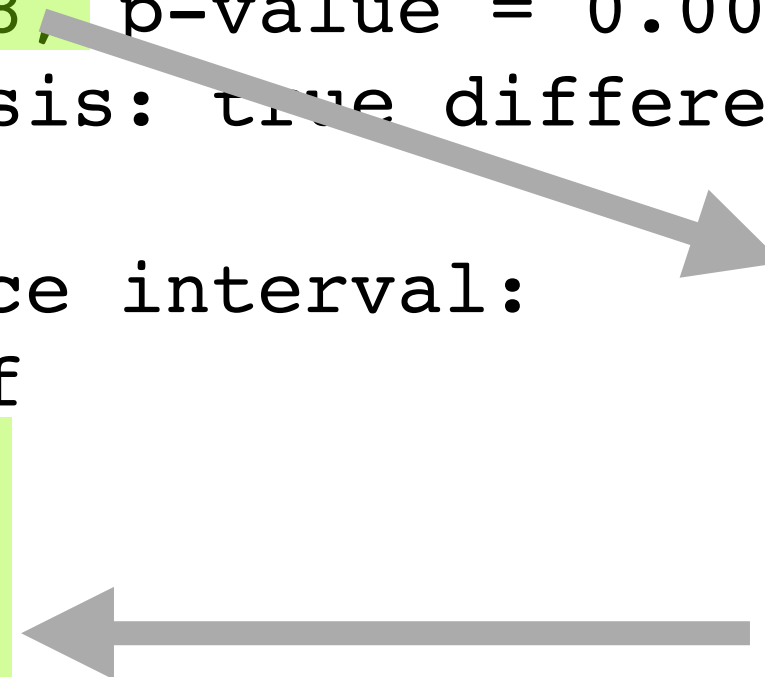
```
0.7230215      Inf
```

```
sample estimates:
```

```
mean of x mean of y
```

```
29.63752  27.41136
```

N - 2 degrees  
of freedom  
(because we are  
estimating two  
parameters)



# BMI example: parametric test in R

---

```
>ttestResult=t.test(BMI~PhysActive,data=NHANES_sample,  
var.equal=TRUE,alternative='greater')
```

## Two Sample t-test

```
data: BMI by PhysActive
```

```
t = 2.4452, df = 248, p-value = 0.007587
```

```
alternative hypothesis: true difference in means is  
greater than 0
```

```
95 percent confidence interval:
```

```
0.7230215      Inf
```

```
sample estimates:
```

```
mean of x mean of y
```

```
29.63752  27.41136
```



probability of

$t \geq 2.44$  for

$t(248)$

in R:

```
1 - pt(2.4452, 248)
```

**if we reran the test as a two-tailed (non-directional) test, the p-value would be:**

the same  
(0.0075)

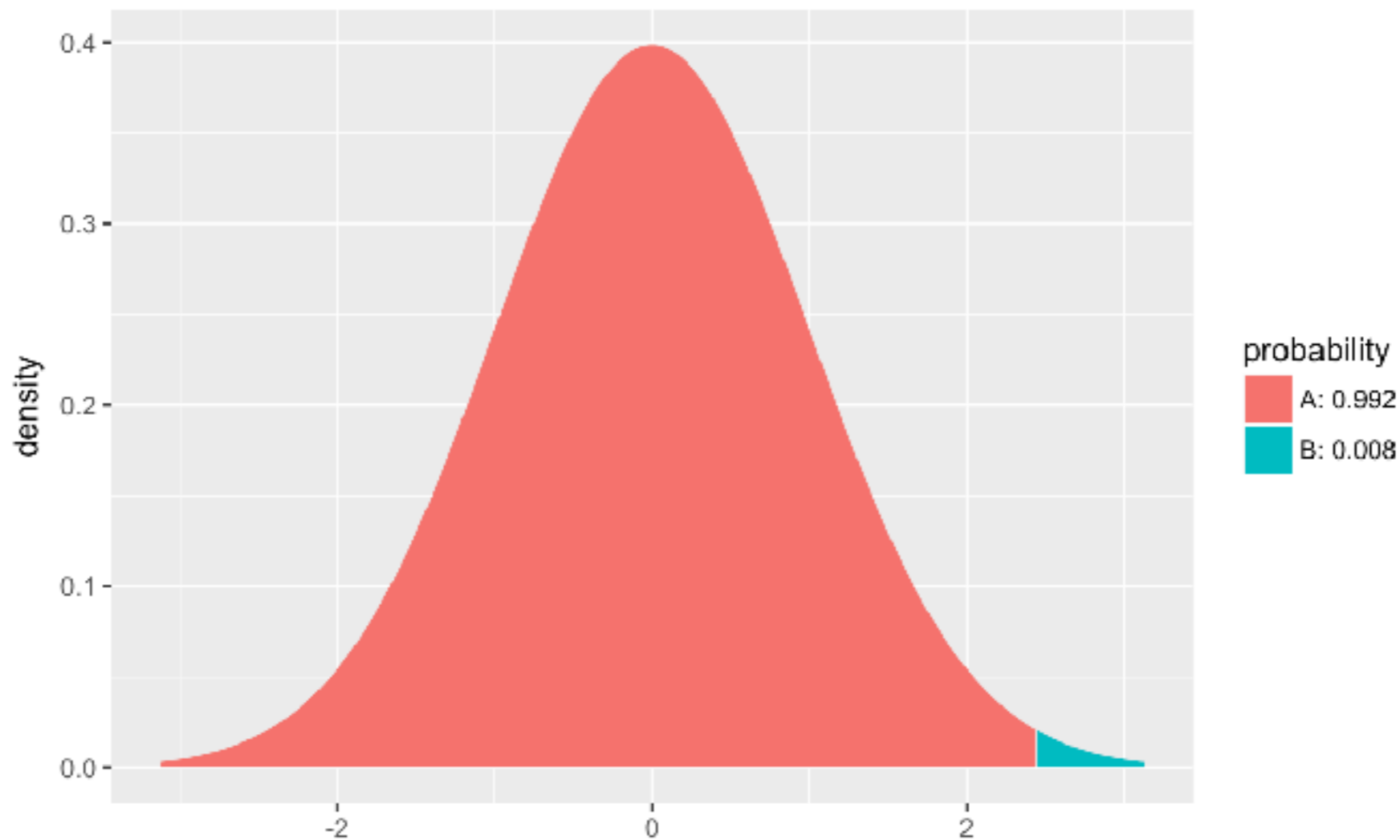
twice as large  
(0.015)

half as large  
(0.00375)

# One-tailed vs two-tailed tests

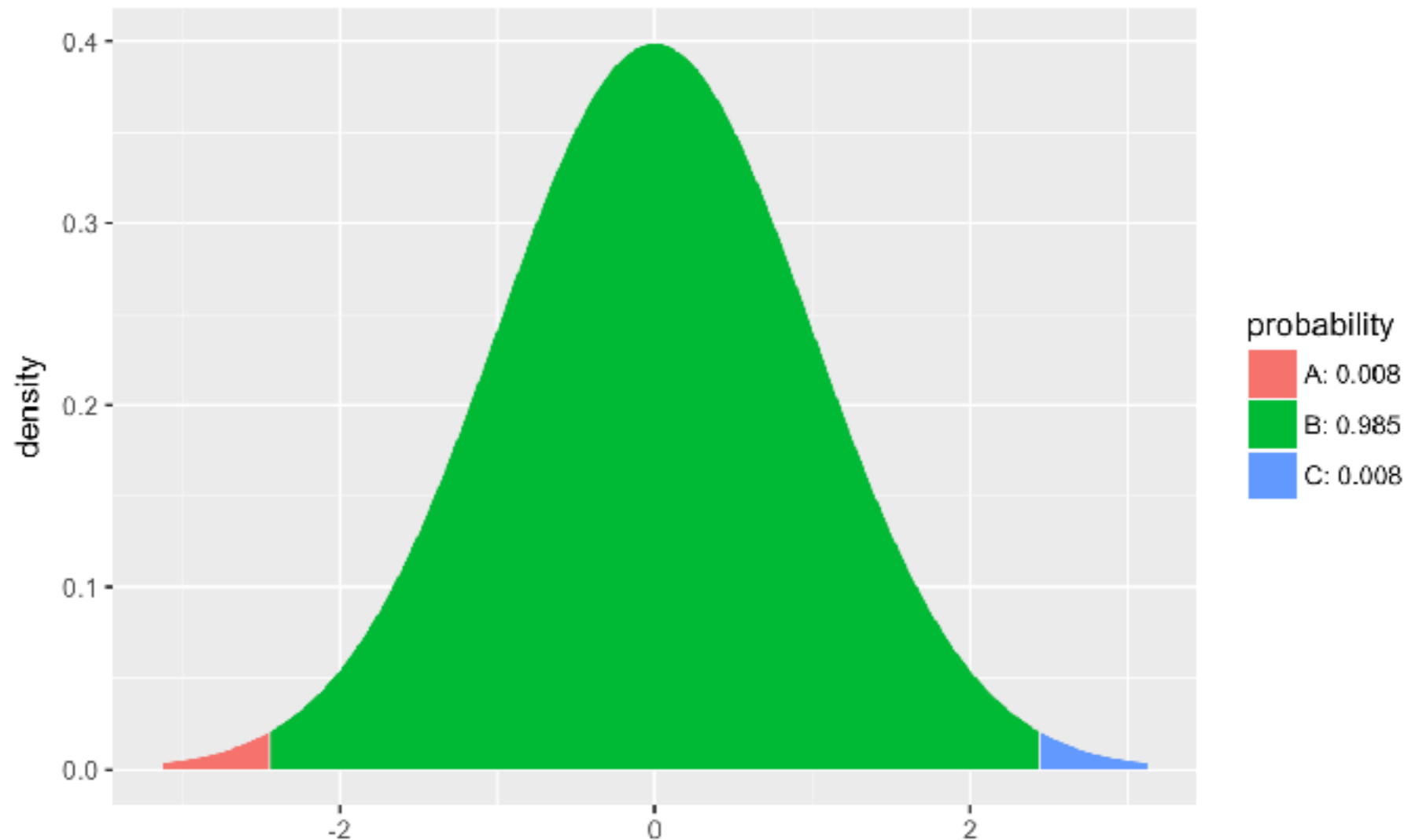
---

- Directional test:
  - $p\text{-value} = 1 - p(t_{\text{observed}} \geq t_{248})$



# One-tailed vs two-tailed tests

- Two-tailed (non-directional test)
  - $p\text{-value} = 1 - p(t_{\text{observed}} \geq t_{248}) + p(t_{\text{observed}} \leq t_{248})$





# Two-tailed results

---

```
ttestResult = t.test(BMI~PhysActive, data=NHANES_sample, var.equal=TRUE,  
                    alternative='two.sided')
```

Two Sample t-test

```
data: BMI by PhysActive
```

```
t = 2.4452, df = 248, p-value = 0.01517
```

```
alternative hypothesis: true difference in means is not equal  
to 0
```


```
95 percent confidence interval:
```

```
0.4329999 4.0193201
```

```
sample estimates:
```

```
mean of x mean of y
```

```
29.63752 27.41136
```



p-value is twice  
as large for two-  
tailed test versus  
one-tailed test:  
data are less  
surprising!

# Step 6: Assess the “statistical significance” of the result

---

- What does “statistical significance” mean?
- How much evidence against the null hypothesis do we require before rejecting it?