PSY 503: Foundations of Statistics in Psychological Science

# NHST and *p*-values (Everything you ever wanted to know about *p*-values)

Jason Geller, Ph.D. (he/him/his)

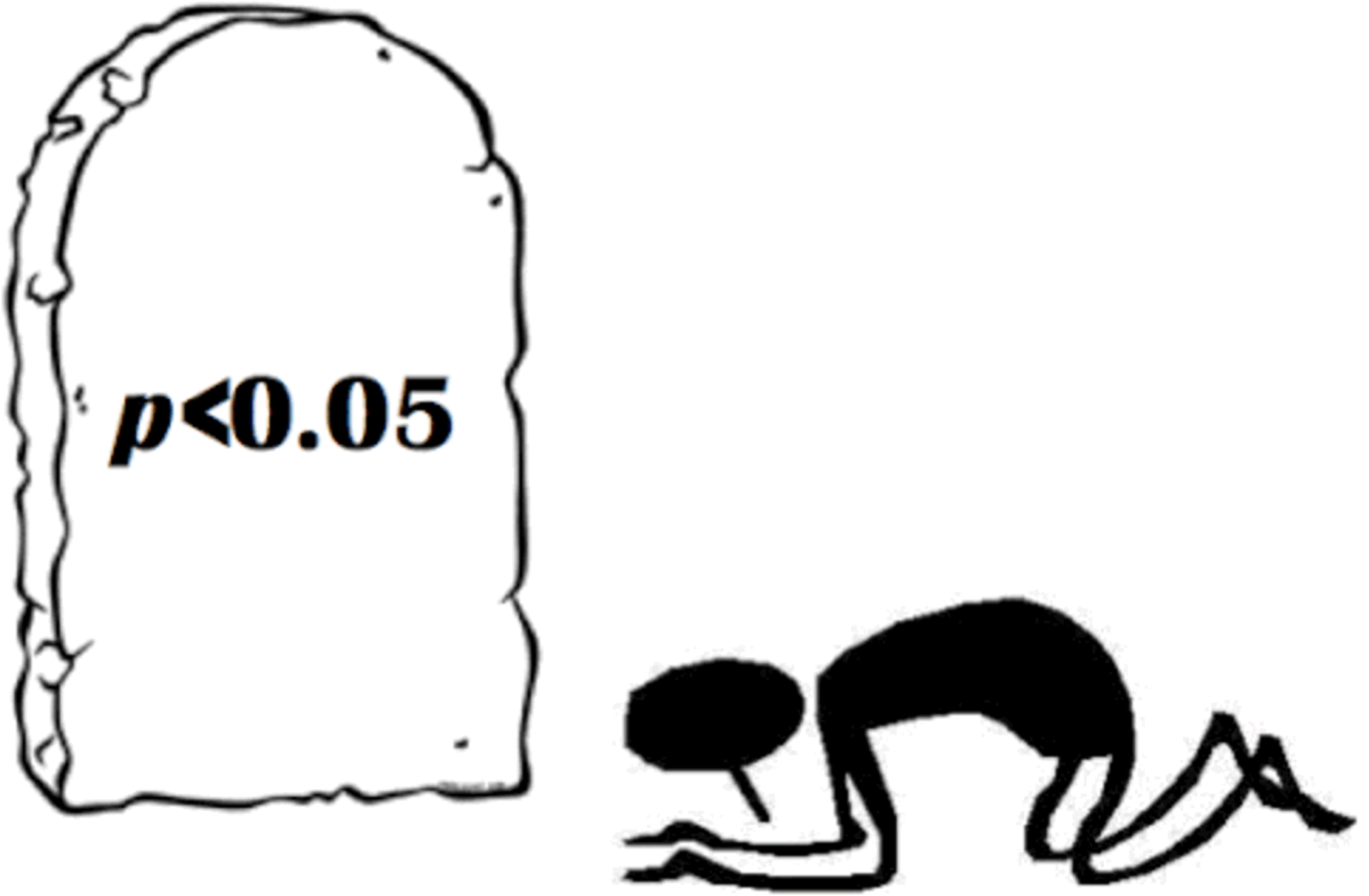
Princeton University

2022-10-05

Go to **www.menti.com/al2soqaxyin6**

# Name

| | | |
|---|---|---|
| sarah | Nicole | Jamie |
| Henna | Claire | Cody Dong |
| yes branson | | |


p<0.05

## Today

- Null hypothesis significance testing (NHST)
  - 1 vs 2 tailed tests
  - *p*-values
  - Steps in NHST
  - Type 1 and Type 2 error
  - *p*-value misconceptions

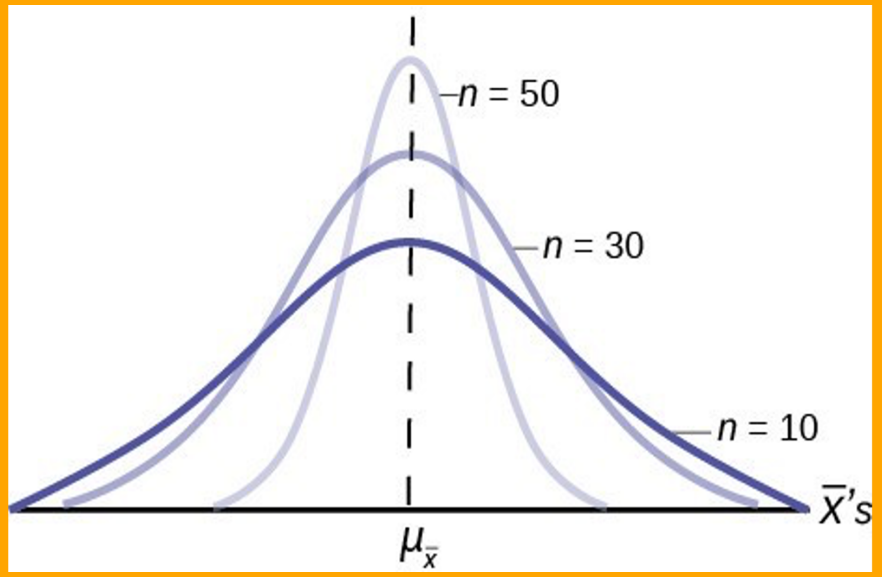
## Next Week

- Parametric
  - 1 sample, 2 sample (Welch’s), & paired t-tests
- Non-parametric
  - Mann-Witney U test (between)
  - Wilcoxon signed-rank (paired)
- Correct for Multiple Comparisons

# Recap

- Sampling Distribution: The probability distribution of a given statistic (e.g., mean) taken from a random sample
- Constructing a Sampling Distribution
  - Randomly draw n sample points from a finite population with size N
  - Compute statistic of interest
  - List different observed values of the statistic with their corresponding frequencies

# Recap




- As n increases we become more confident (less spread) of our estimate of the population mean

# Recap: CIs

- CIs: Interval or range that encompasses true parameter value
    - Level of confidence
        - 95% is most common

Calculation

- Lower: estimate - MoE
- Upper: estimate + MoE

# Hypothesis Testing: The General Framework

# Statistical Inference

1. Estimation

2. Test relationships (hypothesis testing)

## Proof by contradiction

> To prove a mathematical statement, A, you assume temporarily that A is false. If that assumption leads to a contradiction, you conclude that A must actually be true

# NHST

- Negate the conclusion: Begin by assuming the opposite – that there is no relationship between X and Y.
- Analyze the consequences of this premise: If there is no relationship between X and Y in the population, what would the sampling distribution of the estimate of the relationship between X and Y look like?
- Look for a contradiction: Compare the relationship between X and Y observed in your sample to this sampling distribution. How (un)likely is this observed relationship?
    - If small, then there is evidence there is a relationship

# NHST

- Null Hypothesis $H_0$: There is no significant difference
  - 0 in population
  - No difference between two or more groups
- Alternative Hypothesis $H_1$: There is a statistically significant difference

# An Example

Toftness, Carpenter, Geller, Lauber, Johnson, and Armstrong (2017)

- Does fluency of the instructor lead to better learning of the material?

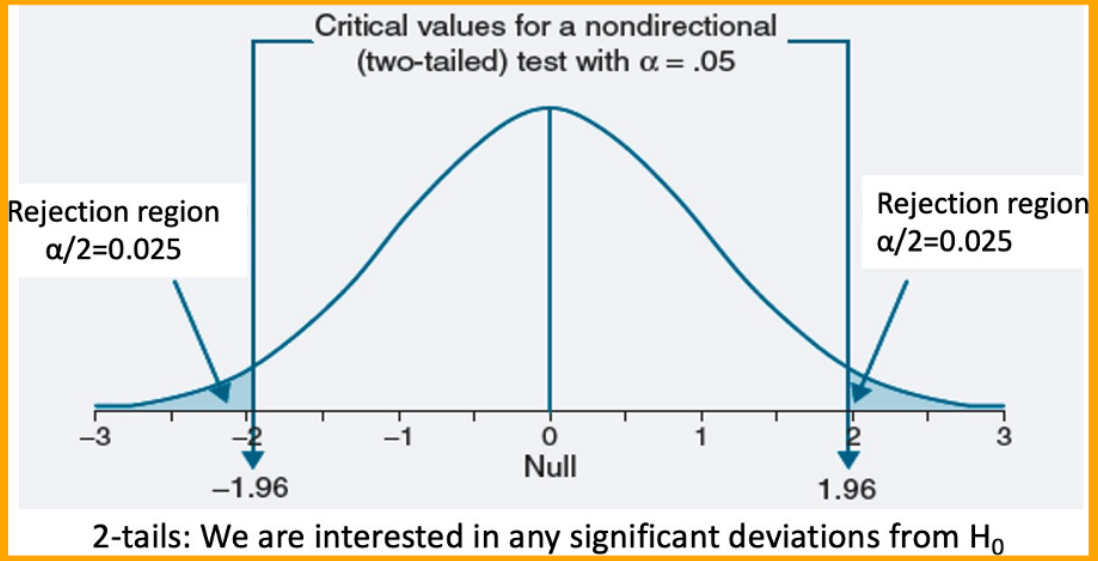
# Null and Alternative Example

- Does fluency of the instructor lead to better learning of the material?
  - Null Hypothesis: H0 : $\mu_f = \mu_d$
  - Alternative Hypothesis: HA: $\mu_f \neq \mu_d$

## Two-sided and One-sided Alternative Hypotheses

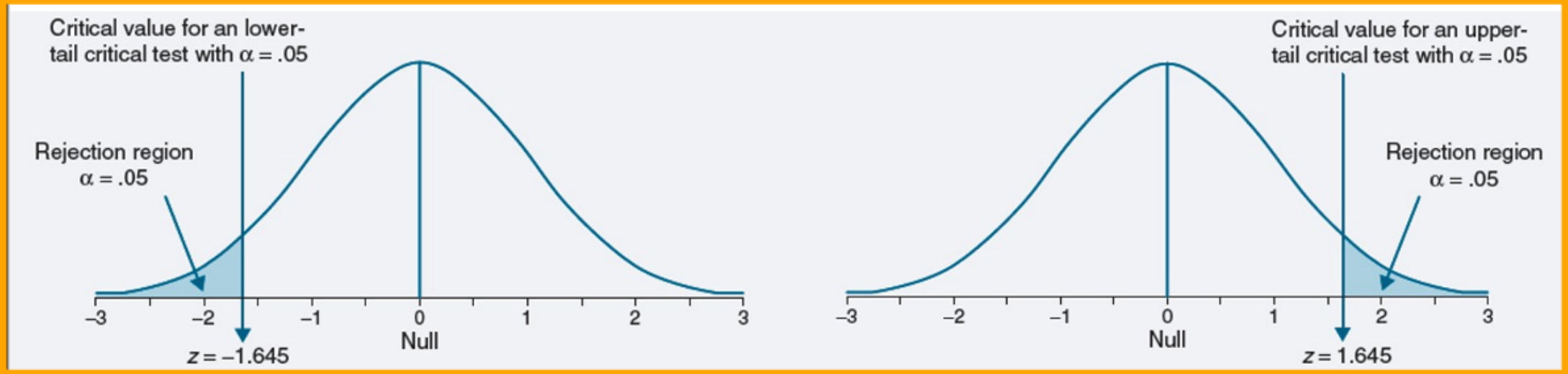
- Two-tailed: H0 : $\mu_f = \mu_d$ ; HA = $\mu_f \neq \mu_d$
- One-tailed: H0 = $\mu_f = \mu_d$ ; HA = $\mu_f > \mu_d$
- One-tailed: H0 = $\mu_f = \mu_d$ ; HA = $\mu_f < \mu_d$
- Only use a one-tailed / directional hypothesis if you have a strong theoretical prediction (for example, from a model) or you preregister it
  - Can gain statistical power
  - But sometimes findings are meaningful and interesting if they go in an unexpected way
- We can accommodate both two-tailed and one-tailed tests statistically

# Two Tailed Test




- The sum of the tails sums to α (0.025 in each tail for a two-tailed when α = 0.05)
- See where the statistic lies relative to a ‘critical score’ that depends on defined alpha (same procedure used to calculate confidence intervals)

# One Tailed Test




- 0.05 in each tail
- If statistic within rejection region = reject null & accept alternative
- Do you see why you get power with a one-sided / directional hypothesis?

## Should I use a one-tailed or two-tailed test?

- Always use two-tailed when there is no directional expectation
    - There are two competing predictions
- Can use one-tailed when strong justification for directional prediction
- Never follow up with one-tailed if two-tailed is not statistically significant

## Define your level of significance (α)

- Level of significance (α): Probability of rejecting the NULL hypothesis due simply to chance
  - α = 0.05
  - Some use others (e.g. 0.01, 0.0000003 particle discovery in high energy physics)
  - How do we determine what the probability of rejecting the null hypothesis given our data?

## What is a *p*-value?

The probability of observing the sample data, or more extreme data, assuming the null hypothesis is true

How surprising something is

# *p*-value

## *p*-value: Schools of Thought

- Ronald Fisher:
  - Quantifying evidence
    - Smaller *p*-value provides stronger evidence against the null hypothesis
- Neyman and Pearson:
  - Null and Alternative hypotheses
  - *p*-value is only used to check if it is smaller than the chosen $\alpha$ level, but it does not matter how much smaller it is

# *p*-value: Schools of Thought

| | FISHER | NEYMAN/PEARSON | HYBRID NHST* |
|---|---|---|---|
| explicit & serious alternative $H_a$ | X | ✓ | X |
| when to set-up statistical model | after data collection | before data collection | after data collection |
| goal of statistical analysis | quantify evidence against $H_0$ | decide action: adopt $H_0$ or $H_a$ | decide action: adopt $H_0$ or $\neg H_0$ |
| power calculation | X | ✓ | X |

* this is a worst-case portrait of modern NHST ; this is *not* how it *should* be done

## Steps for Hypothesis Testing

1. State null hypothesis and alternative hypothesis
2. Calculate the corresponding test statistic and compare the result against the “critical value”
3. State your conclusion

## Steps for Hypothesis Testing

- Step 1
  - Convert the research question to null and alternative hypotheses
    - The null hypothesis ($H_0$) is a claim of “no difference in the population”
      - No difference between a population parameter and hypothesized value: $H_0 : \mu = 7.56$
      - No difference between one population parameter and another: $H_0 : \mu_f = \mu_d$
      - *We usually want to reject this hypothesis*
- The alternative hypothesis ($H_1$) claims $H_0$ is false: There is some difference

## Example

- The problem: In the 1970s, 20–29 year old men in the U.S. had a mean μ body weight of 170 pounds. Standard deviation σ was 40 pounds. We test whether mean body weight in the population now differs.
    - What is null here?
    - What is alternative?

## Steps for Hypothesis Testing

- Step 2
  - Calculate the corresponding test statistic and compare the result against the “critical value”
    - *t*, *z*, *F*
      - Is the test statistic > or < than critical value?
  - A value of the test statistic is interesting if it has only a small chance of occurring when the null hypothesis is true

## Steps for Hypothesis Testing

- Step 2
  - Define your level of significance (α)
- α Interpretation: If we were do this experiment many many times, we would only expect 5% (or another level of significance) to be Type 1 Error

## Steps for Hypothesis Testing

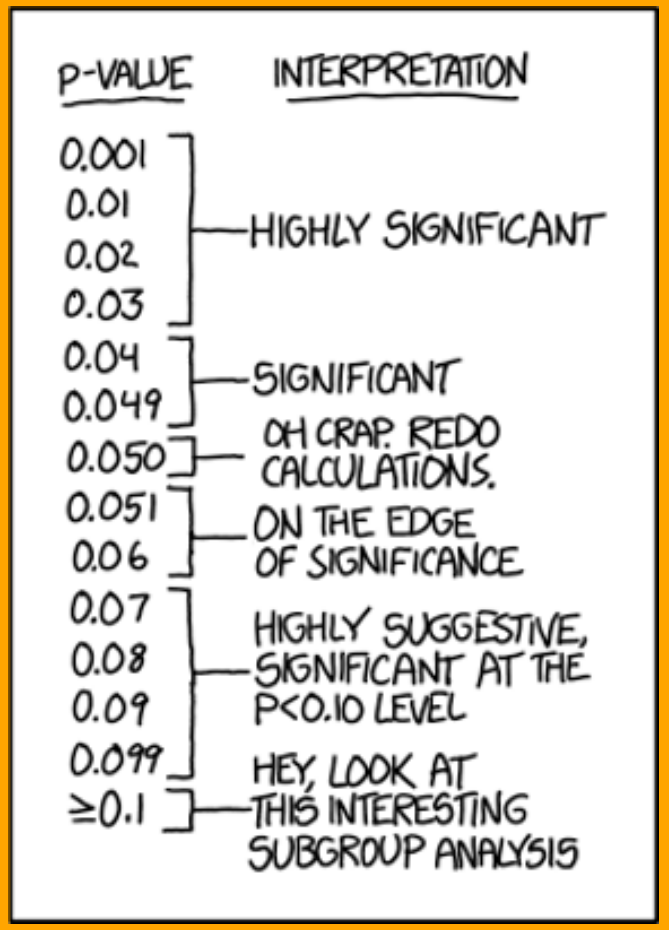
- Step 3
  - State your conclusion
    - Reject the null $p < \alpha$
    - Fail to reject the null $p > \alpha$

Never say we found no difference. Instead say you found no statistically significant difference. Can never affirm the null

# *p*-value Conventions

- Conventions:
  - $p > 0.10$: non-significant evidence against H0
  - $0.05 < p < 0.10$: marginally significant evidence
  - $p < 0.05$: significant evidence against H0

.pull-right [

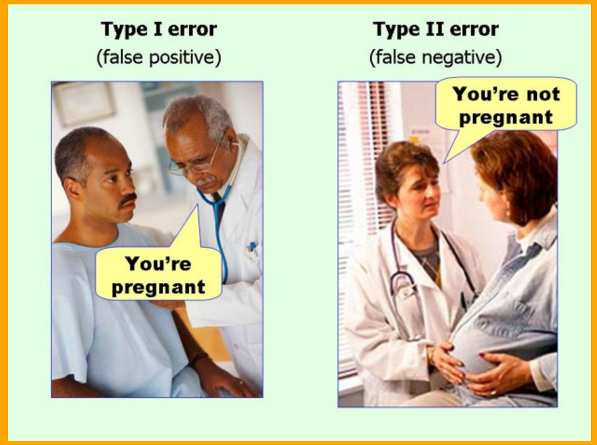



]

# Type 1 and Type 2 Error Rates

- You think the manipulation worked, but it really doesn't
  - Type 1 error
- You don't think the manipulation worked, but it really does
  - Type 2 error

```{r echo=FALSE,out.width="100%",fig.cap="",fig.show='hold',fig.align='center'}
```



# Type 1 and Type 2 Error Rates

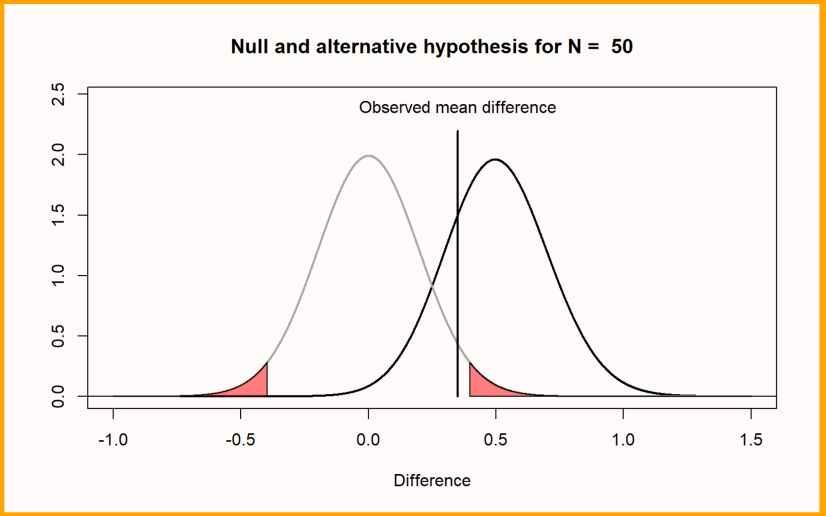
| Type I and Type II Error | | |
|---|---|---|
| **Null hypothesis is ...** | **True** | **False** |
| **Rejected** | **Type I error**<br>False positive<br>Probability = α | **Correct decision**<br>True positive<br>Probability = 1 - β |
| **Not rejected** | **Correct decision**<br>True negative<br>Probability = 1 - α | **Type II error**<br>False negative<br>Probability = β |

## Correctly reporting and interpreting p-values

- Exact *p*-values!
- *p*-values reference the observed data and not a theory
- Report $\alpha$
- Do not use *p*-values as a measure of evidence

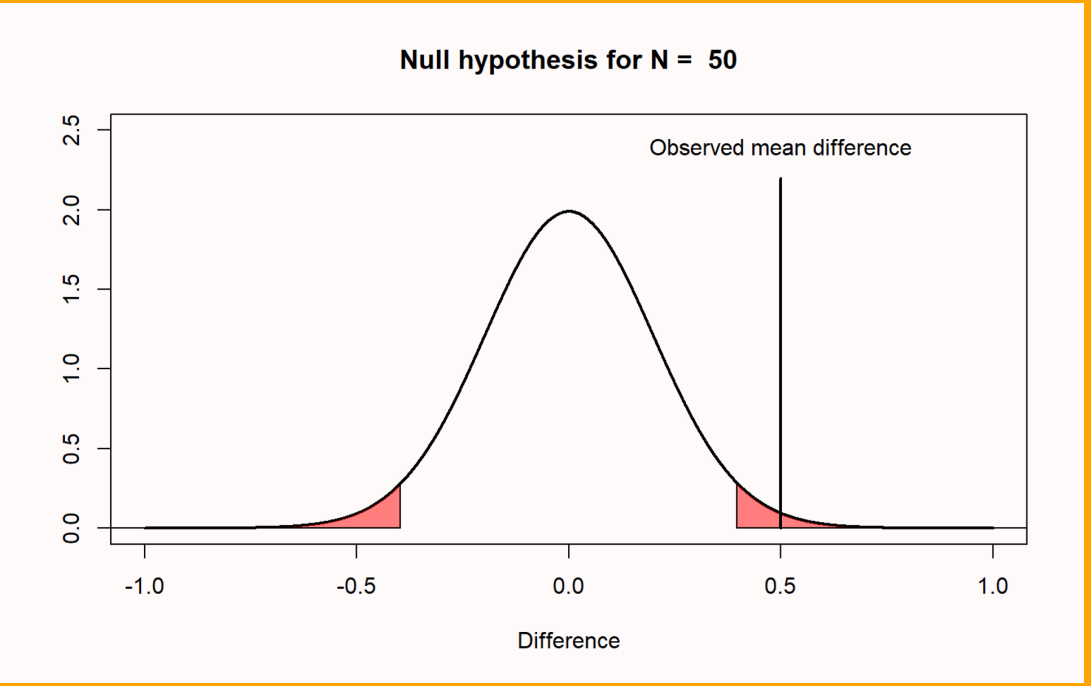
# *p*-value Misconceptions

- A non-significant *p*-value does not mean that the null hypothesis is true
  - "There was no difference"
  - "The null is true"



# *p*-value Misconceptions

- A significant p-value means that the null hypothesis is false
  - p < .05, therefore there is an effect',
  - or 'there is a difference between the two groups, p < .05'

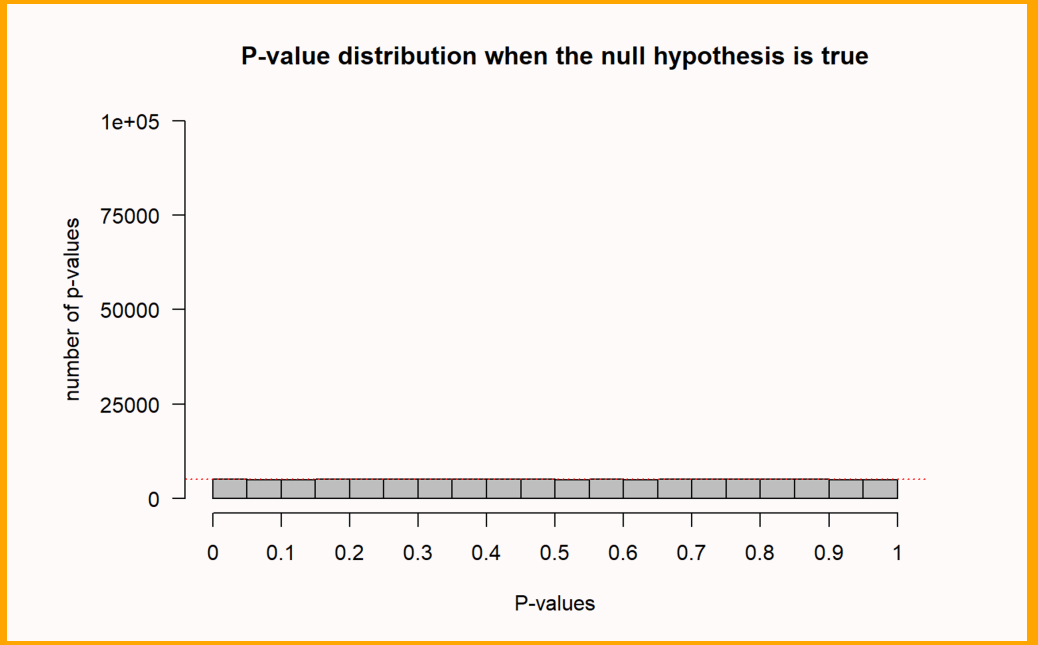


# *p*-value Misconceptions

- A significant p-value does not mean that a practically important effect has been discovered

# *p*-value Misconceptions

- If you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%
  - Type 1 error rate references all studies we will perform in the future where the null hypothesis is true
  - Not more than 5% of our observed mean differences will fall in the red tail areas.



# Practice NHST

# 1 Sample t-test (1 tailed)

- 1 sample t-test (1 tailed)

$$t(x) = \frac{\bar{X} - \mu}{s/\sqrt{(n)}}$$

  - where μ is some value we are comparing our sample to

- Find *t*

- Find *t* critical
- Find *p*-value

# 1 Sample t-test (1 tailed)

- Sample of 12 people lost 0.61 kg with standard deviation of s= 1.62 kg. Can we conclude their weight is less than than their original weight?
  - What do we know?
  - What are the hypotheses?

$$t(x) = \frac{\bar{X} - \mu}{s/\sqrt{(n)}}$$

# 1 Sample *t* (1 tailed)

```
t0 =( -0.61 - 0)/(1.62/(12)^(1/2))
alpha =0.05
tcrit0 = qt(alpha,12    - 1, lower.tail = TRUE) # (alpha, df) - NOTICE alpha!
pval0 = pt(t0,12 - 1, lower.tail = TRUE) # looking at LEFT tail
t0
```

```
## [1] -1.304384
```

```
tcrit0
```

```
## [1] -1.795885
```

```
pval0
```

```
## [1] 0.1093673
```

# What do we conclude?

- Decision, conclusion, and *p* value

# 1 Sample *t* (Example 2)

- Among 157 African-American men seen in the emergency department at the hospital, the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. Can we conclude based on this data that the mean systolic blood pressure for a population of African-American men is greater than 140 (i.e., μ) at the 95% confidence level?
  - What do we know?
  - What are the hypotheses?

$$t(x) = \frac{\bar{X} - \mu}{s/\sqrt{(n)}}$$

# 1 Sample *t* (Example 2)

```
t0 =(146-140) /(27/(157)^(1/2))
alpha =0.05
tcrit0 = qt(alpha,157    - 1, lower.tail = FALSE) # (1-alpha, n - 1) - NOTICE 1-alpha!
pval0 =pt (t0,157 - 1, lower.tail = FALSE) # looking at RIGHT tail (1 - p)!
t0
```

```
## [1] 2.784436
```

```
tcrit0
```

```
## [1] 1.65468
```

```
pval0
```

```
## [1] 0.003013078
```

# 1 Sample *t* (Example 2)

- Decision, Conclusion, and p-value?

# 1 Sample *t* (Example 2)

- Use `t.test` function in R

```
#####simulate data#######
n1 =157
alpha1 =0.05
data1 <- rnorm(n1,mean =146,sd=27)
############ Run ttest#########

res1 <- t.test(data1,mu =140,alternative ="greater")
#alternative options: "less", "greater", "two.sided"

res1
```

```
##
##     One Sample t-test
##
## data:  data1
## t = 4.7079, df = 156, p-value = 2.743e-06
## alternative hypothesis: true mean is greater than 140
## 95 percent confidence interval:
##  146.6637      Inf
## sample estimates:
## mean of x
##   150.275
```

# 1 Sample *t* (Example 3)

- A cookie company claims that there are 15 chocolate chips per cookie, but you aren't convinced. You take 10 cookies and count the number of chocolate chips in each cookie. Here is what the data looks like: [13,14,15,17,18,19,21,20,19,20]. Are the number of chocolate chips significantly different from 15?
- What do we know?
- What are the hypotheses?

$$t(x) = \frac{\bar{X} - \mu}{s/\sqrt{(n)}}$$

```r
data2 = c(13,14,15,17,18,19,21,20,19,20)

res2 <- t.test(data2,mu =15,alternative ="two.sided")
res2
```

```
##
##     One Sample t-test
##
## data:  data2
## t = 2.9824, df = 9, p-value = 0.01539
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  15.6279 19.5721
## sample estimates:
## mean of x
##      17.6
```

- Decision, Conclusion, and p value

# 2 Sample Welch's t-test

- 2 sample tests are interested in whether there are differences between 2 groups
- There are several traditional 2 sample t-tests, where the appropriate version depending on equal or unequal sample size or variance (giant flow charts!)
- Welch's t-test gives gives equivalent answer to traditional t-test when there is an equal sample size or variances, BUT can also handle unequal sample size and variance.
    - http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html