Lecture 10: Sep 24, 2018 - v2

Hypothesis Testing

- *Testing Frameworks*

James Balamuta
STAT 385 @ UIUC

Announcements

- **hw04** is due **Friday, Sep 21st, 2018** at **6:00 PM**
- **Quiz 05** covers Week 4 contents @ CBTF.
  - Window: Sep 25th - 27th
  - Sign up: https://cbtf.engr.illinois.edu/sched
- Got caught using GitHub's web interface in hw01 or hw02? Let's chat.

Last Time

- **Random Variables**
  - Variables that take on an unknown value
  - LCM is one way to generate random numbers
- **Distributions**
  - Probability Density/Mass Functions (PD/MF) is **d***
  - Cumulative Distribution Function (CDF) is **p***
  - Inverse CDF is **q***
  - Random Variables from Distribution is **r***
- **Sampling**
  - Sample **without** replacement **does not** add the picked object back.
  - Sample **with** replacement **does** add the picked object back.
- **Caches**
  - Speed up computationally intensive reports by storing results and re-using them.

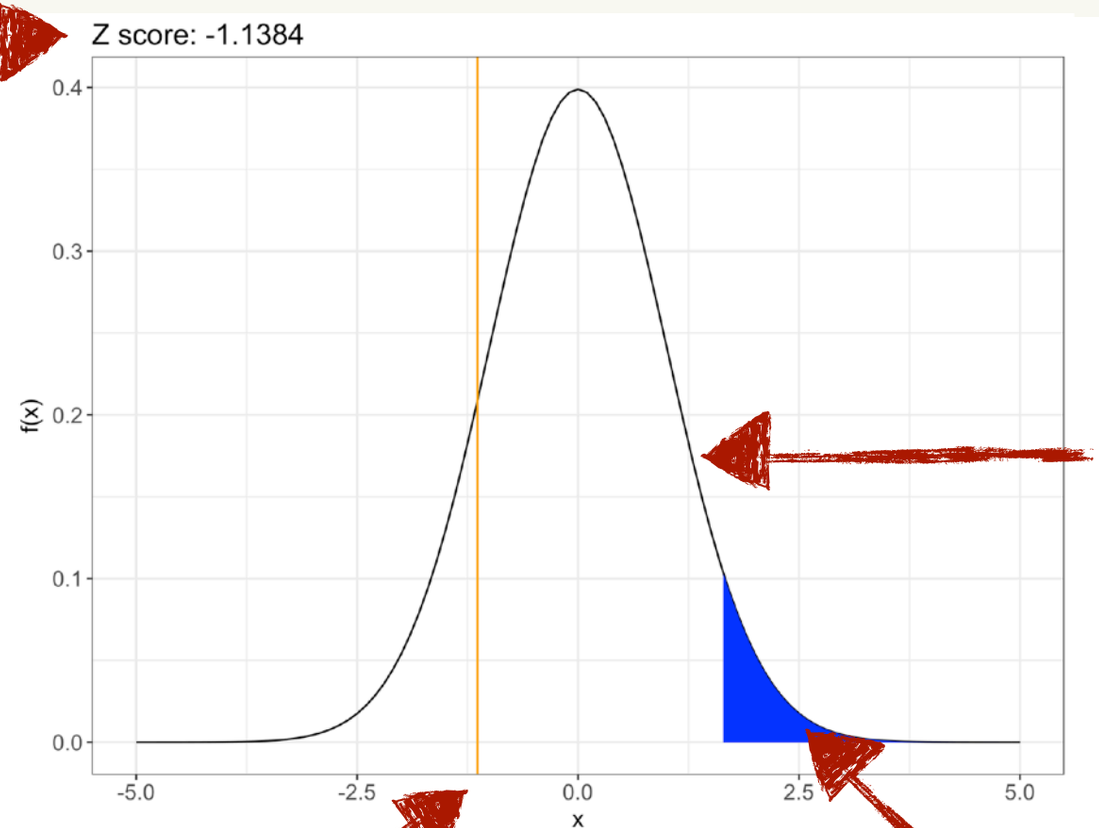
Lecture Objectives

- **Describe** how hypothesis testing is part of confirmatory data analysis and the importance of inference.
- **Differentiate** between different hypothesis testing frameworks.
- **Implement** hypothesis testing algorithms.

Testing Frameworks

Definition:
Hypothesis testing or *confirmatory data analysis* is the act of examining whether random variables (RVs) adhere to stated assumptions or differ.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

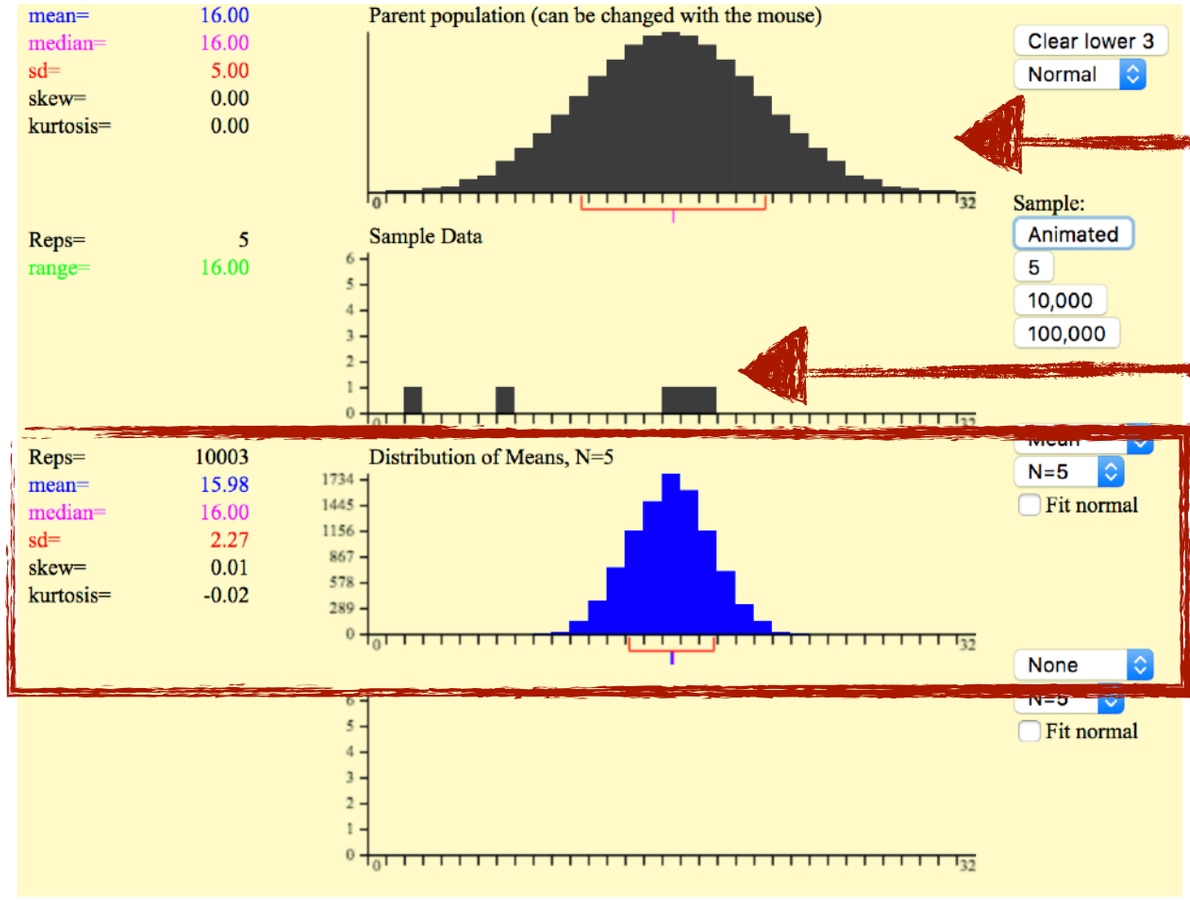


Assumed R.V Sampling Distribution

Test statistic z-score

Rejection Region

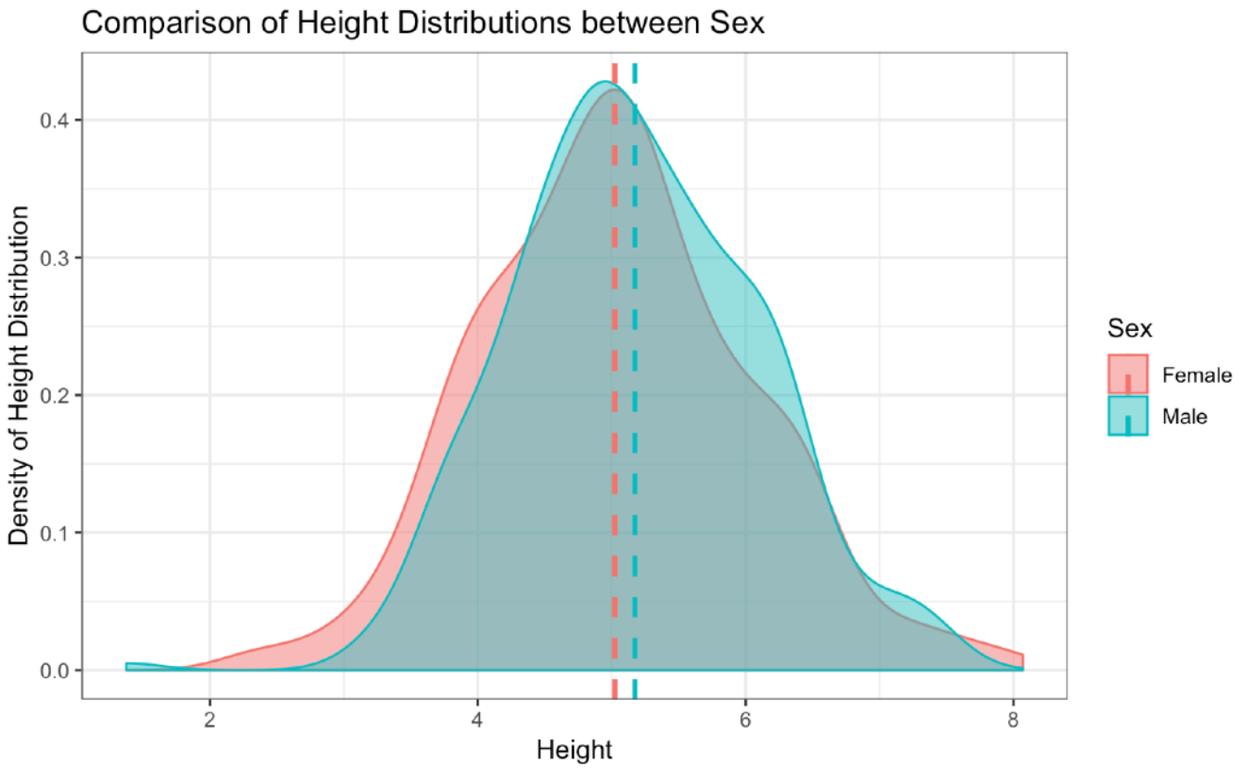
Definition:
Sampling Distribution is a probability distribution of statistics obtained from drawing many samples from a population of interest.

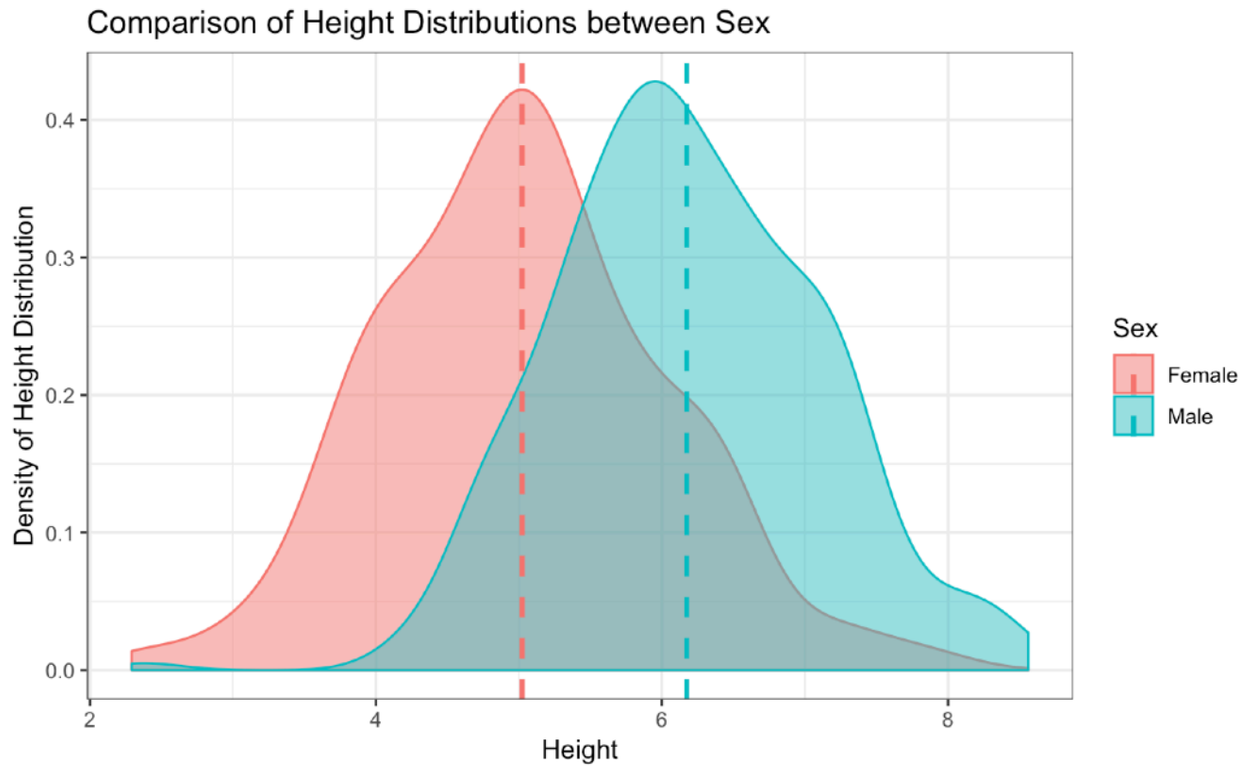


Source

Usefulness of Tests

… does one group differ from the next ???





VS.

Previously

Understand the Algorithm

… weeks of programming saved hours of planning …

- *What* logic is being used?
- *How* does the logic apply in a procedural form?
- *Why* is this logic present?

Computation Types

… how can we test our assumptions ??

- **Parametric (Today)**
  - Assume a distribution and compute a test statistic based on asymptotic results.
- **Resampling (Next time)**
  - Draw samples **with** replacement from the sampling distribution.
- **Permutation (Next time)**
  - Shuffle the samples and sample **without** replacement.

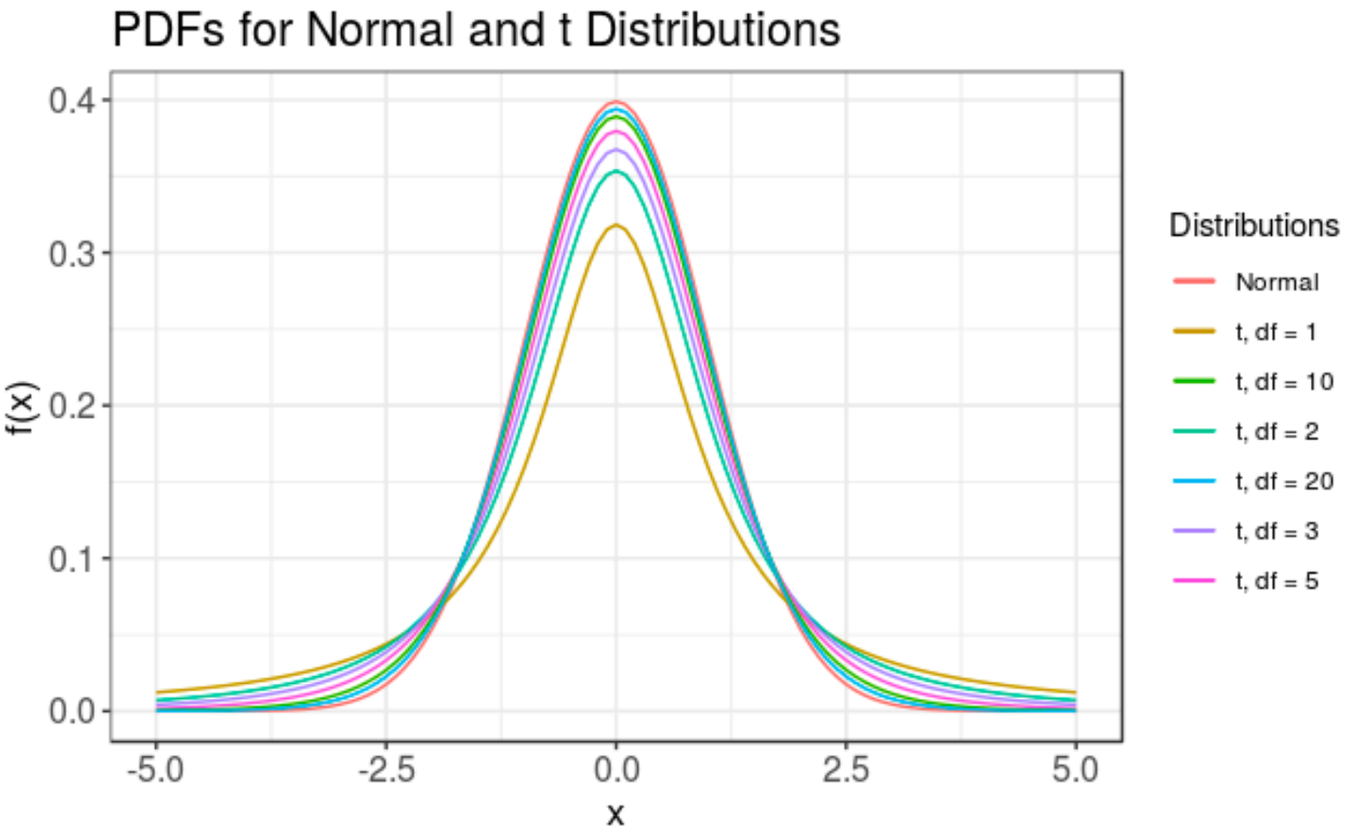
Breaking Down All Tests

… the logic of a hypothesis test …

1. **Make an assumption or null hypothesis…**
   … about the sampling distribution the data has …
2. **Compute a Test Statistic …**
   … Relevant to that sampling distribution …
3. **Compute a *P*-value or Critical Value …**
   … probability of an effect as big given the distribution …
4. **Make a decision …**
   … does evidence exist to suggest the assumption is not okay?

Unpaired (Two Sample) t-Test

Make an assumption...



Choose between:

- **z-test**
  - *Normal distribution*
  - Known Population Variance
  - Sample size: n > 30
- **t-test**
  - *Student's t*
  - Unknown Population Variance
  - Sample size: n < 30

Definition:
Null Hypothesis (H_0) states that the sample data gathered meets the sampling distributions assumptions.

	Means	Proportions
Single	$H_0 : \mu = \mu_0$	$H_0 : p = p_0$
Double	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : p_1 - p_2 = 0$

No difference

Definition:
Alternative Hypothesis (H_a or H_1) is the counter to the null hypothesis that emphasizes the underlying sample *may not* follow the assumptions laid out.

	Means	Proportions
One-sided	$H_a : \mu_1 - \mu_2 > 0$ $H_a : \mu > \mu_0$	$H_a : p > p_0$ $H_a : p < p_0$
Two-sided	$H_a : \mu_1 - \mu_2 \neq 0$ $H_a : \mu \neq \mu_0$	$H_a : p_1 - p_2 \neq 0$ $H_a : p \neq p_0$

Evidence suggests a difference

Example Hypothesis

… sample of forming a statement …

- Consider a coin flip.
- If the coin is fair, half the flips should be heads and the other half tails
- Otherwise, the number of flips between heads and tails will differ.

"Heads" "Tails"

$$H_0 : p = 0.5$$

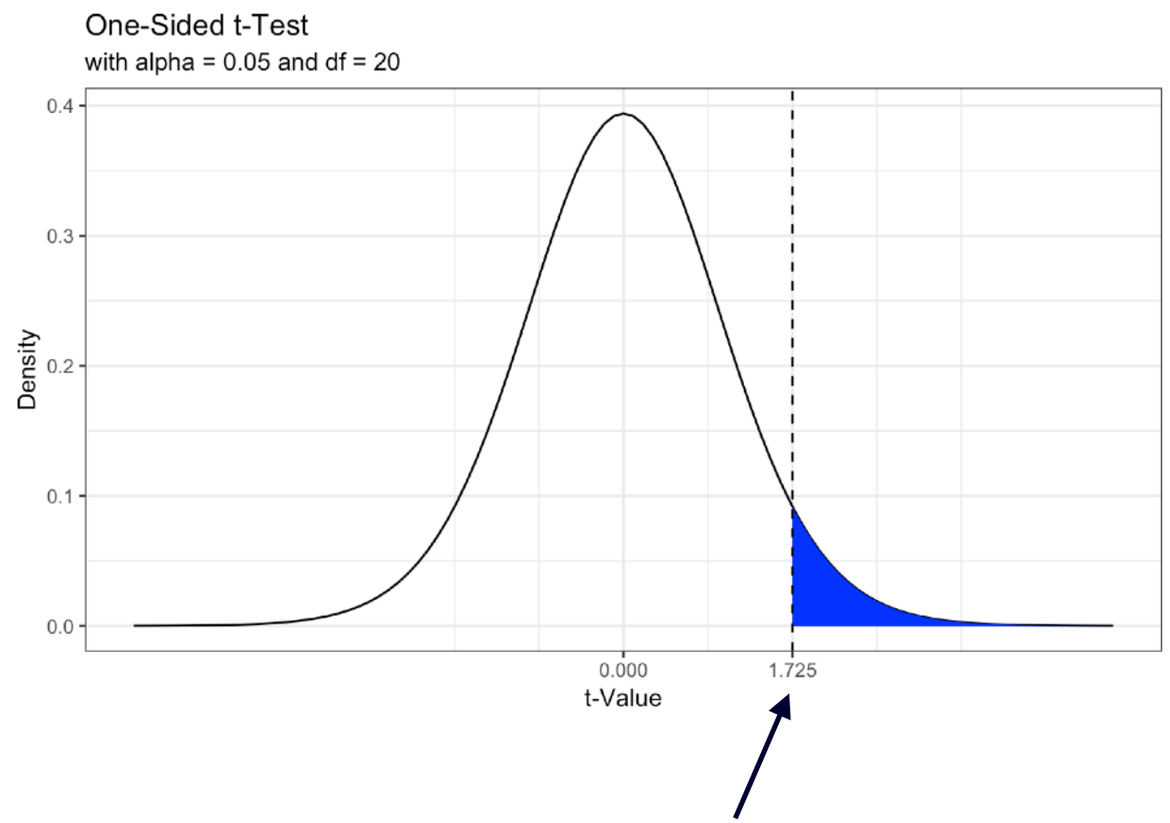
$$H_a : p \neq 0.5$$

One-Tailed Hypothesis

Research Question:
Is the height of males significantly *greater than* that of females?

H_0: The height of males is *not* significantly *greater than* females.

H_A: There is evidence to suggest the height of males *is* significantly greater than females.

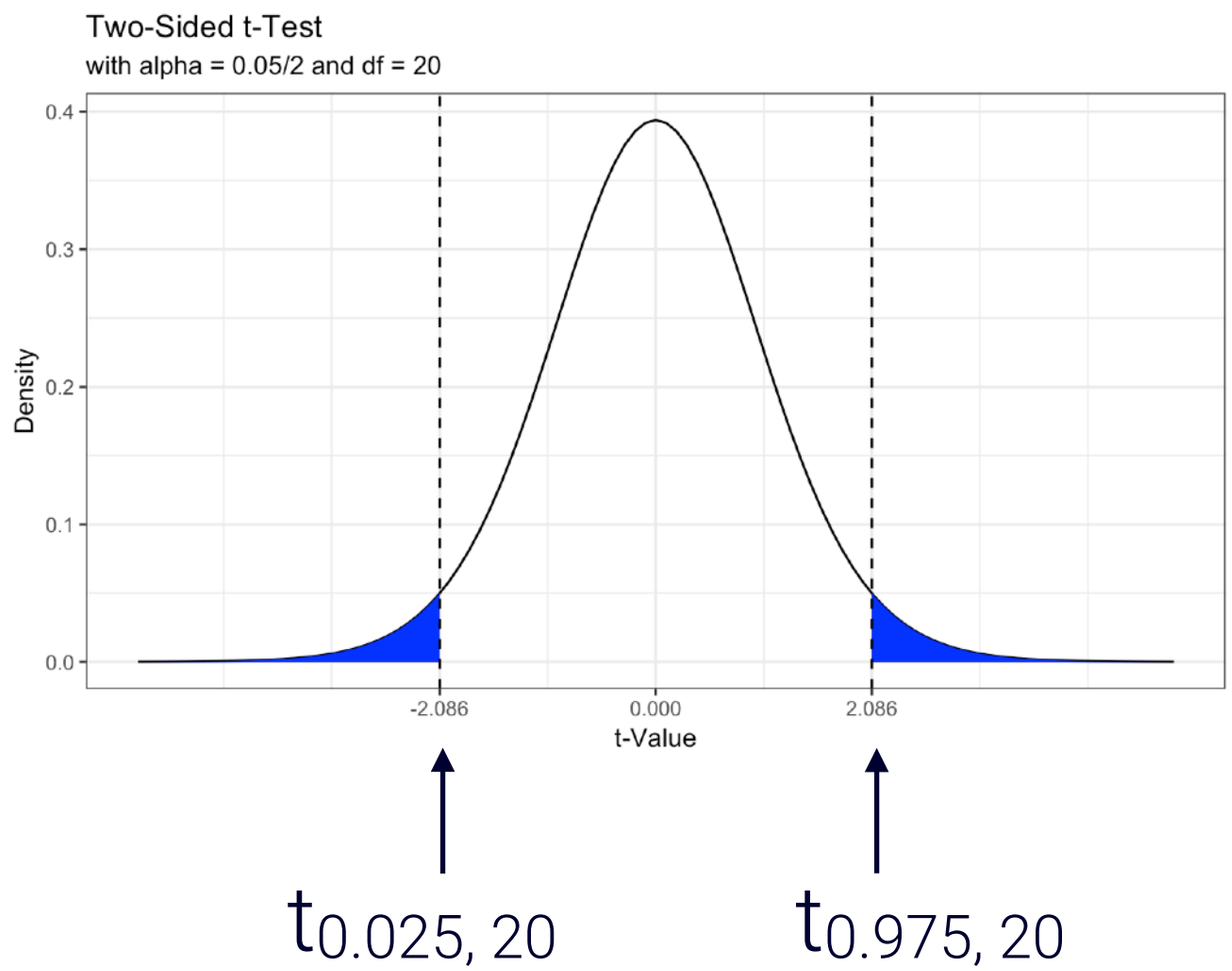


Two-Tailed Hypothesis

Research Question:
Is there a (statistically) significant difference between the height of males and females?

H_0: There is no (statistically) significant difference between the height of males and females.

H_A: There is evidence to suggest a (statistically) significant difference between the height of males and females.



Your Turn

Consider the research question…

Is the height of females significantly *less than* males?

What would be the Null and Alternative hypotheses?

How would the distribution look?

Unpaired (Two Sample) t-Test

… formula overview …

Statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ where $s^2 = \dfrac{\sum_{i=1}^{n_1}\left(x_i - \bar{x}_1\right)^2 + \sum_{j=1}^{n_2}\left(x_j - \bar{x}_2\right)^2}{n_1 + n_2 - 2}$

n_1 and n_2 represent the sample size of data,

and the sample mean is $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

Generating Sample Data

… creating two groups …

```
# Set seed for reproducibility
set.seed(881)

# Generate data
n = 10
x1 = round(rnorm(n), 1)
# [1]  0.7  0.4 -1.3 -1.7  1.1 -1.9  1.9  0.3 -0.5 -0.2
x2 = round(rnorm(n) + 1, 1)
# [1]  0.9 -0.4  0.0  1.8  0.8  1.2  0.5  1.2  0.7  0.8
```

Computing a Test Statistic

$$s^2 = \frac{\sum_{i=1}^{n_1}\left(x_i - \bar{x}_1\right)^2 + \sum_{j=1}^{n_2}\left(x_j - \bar{x}_2\right)^2}{n_1 + n_2 - 2}$$

$$= \frac{13.896 + 3.485}{18}$$

$$= \boxed{0.9656111}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{-0.12 - 0.75}{\sqrt{(0.9656111)\left(\frac{1}{10} + \frac{1}{10}\right)}}$$

$$= \boxed{-1.979717}$$

```
# Compute means of each group
x1_mu = mean(x1)
# [1] -0.12
x2_mu = mean(x2)
# [1] 0.75

# Compute length
# and degrees of freedom
n1 = length(x1)
# [1] 10
n2 = length(x2)
# [1] 10
ndf = n1 + n2 - 2
# [1] 18

# Calculate pooled variance
s2 = ((n1 - 1) * var(x1) + (n2 - 1) *
var(x2)) / ndf
# [1] 0.9656111

# Compute the t-statistic
tstat = (mean(x1) - mean(x2)) /
              sqrt(s2 * (1 / n1 + 1 / n2))
# [1] -1.979717
```

Why are there a lot more
digits after the decimal place?

Does that mean the answer is
more reliable?

Aside

Floating Point Stability

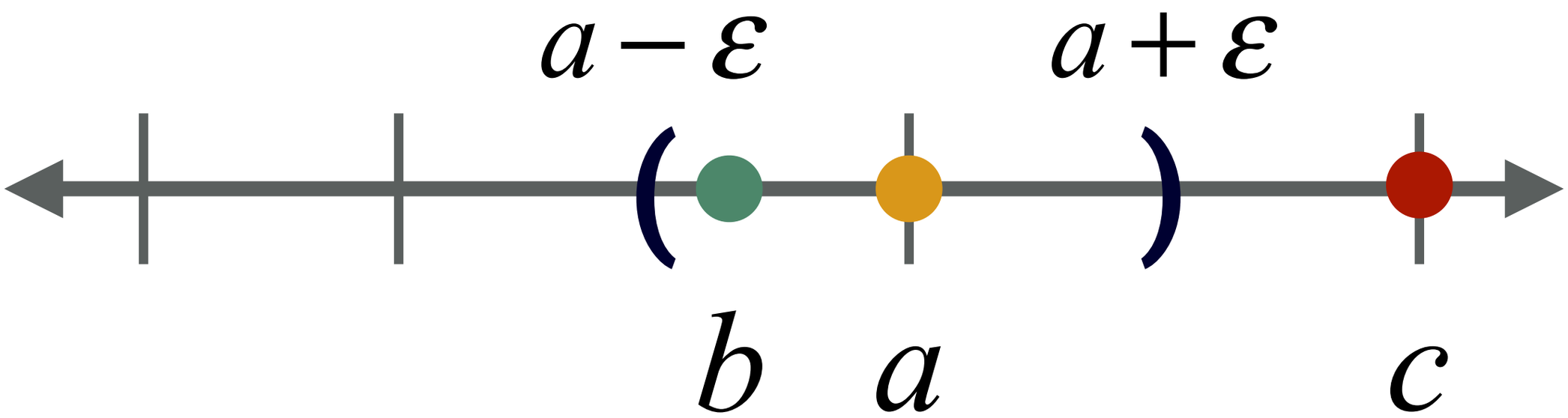
… false precision and tolerance …

```
# Numerics are problematic
0.10 + 0.05 == 0.15
# [1] FALSE

# Allow for tolerance with numerics via an epsilon neighborhood
all.equal(0.10 + 0.05, 0.15)
# [1] TRUE

# Lack of output stability though…
all.equal(0.12, 0.19)
# [1] "Mean relative difference: 0.5833333"

# Check for whether it is true
isTRUE(all.equal(0.12, 0.19))
# [1] FALSE
```

Aside

Epsilon (ε) Neighborhoods

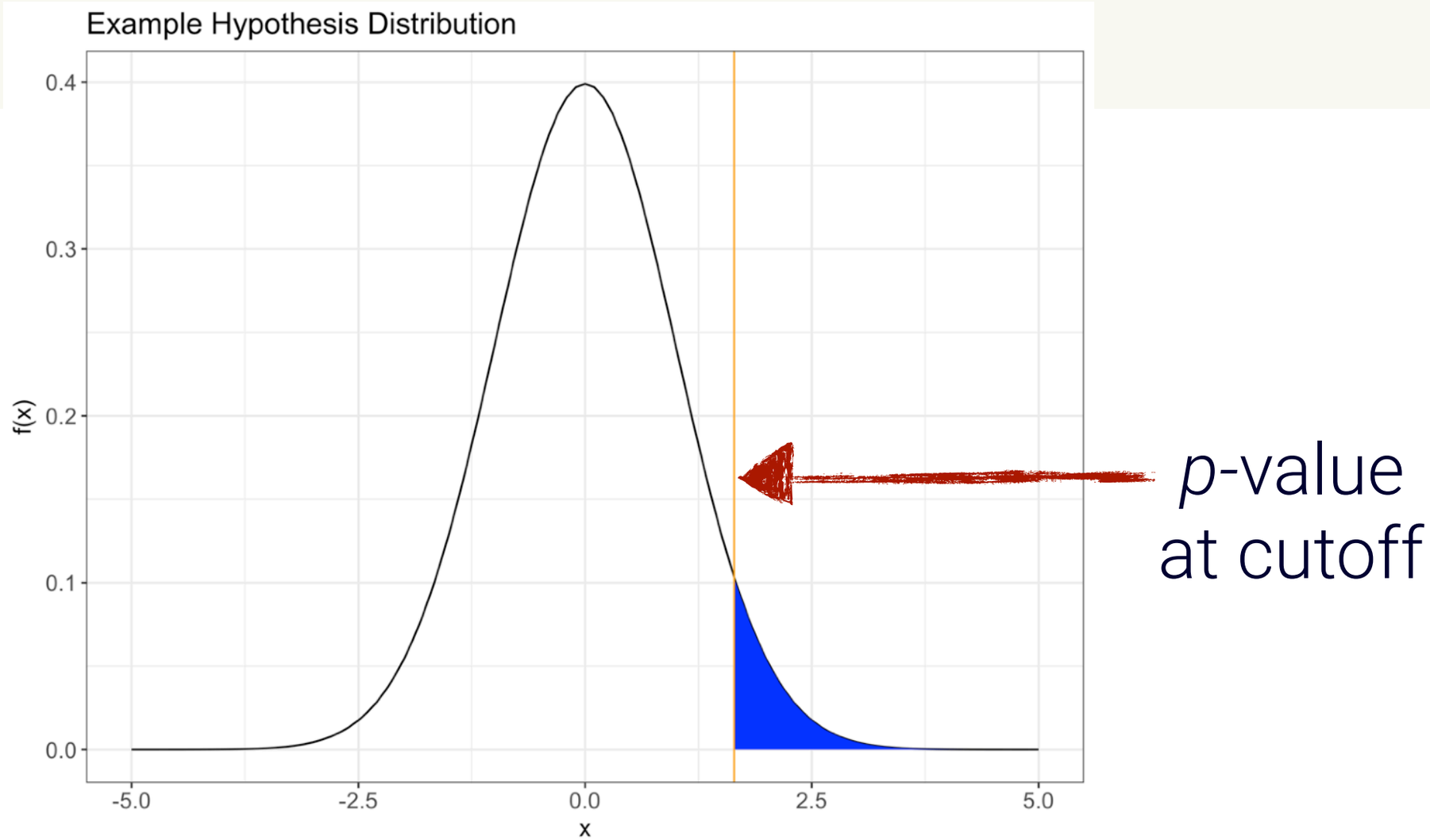
… where close enough is spot on …



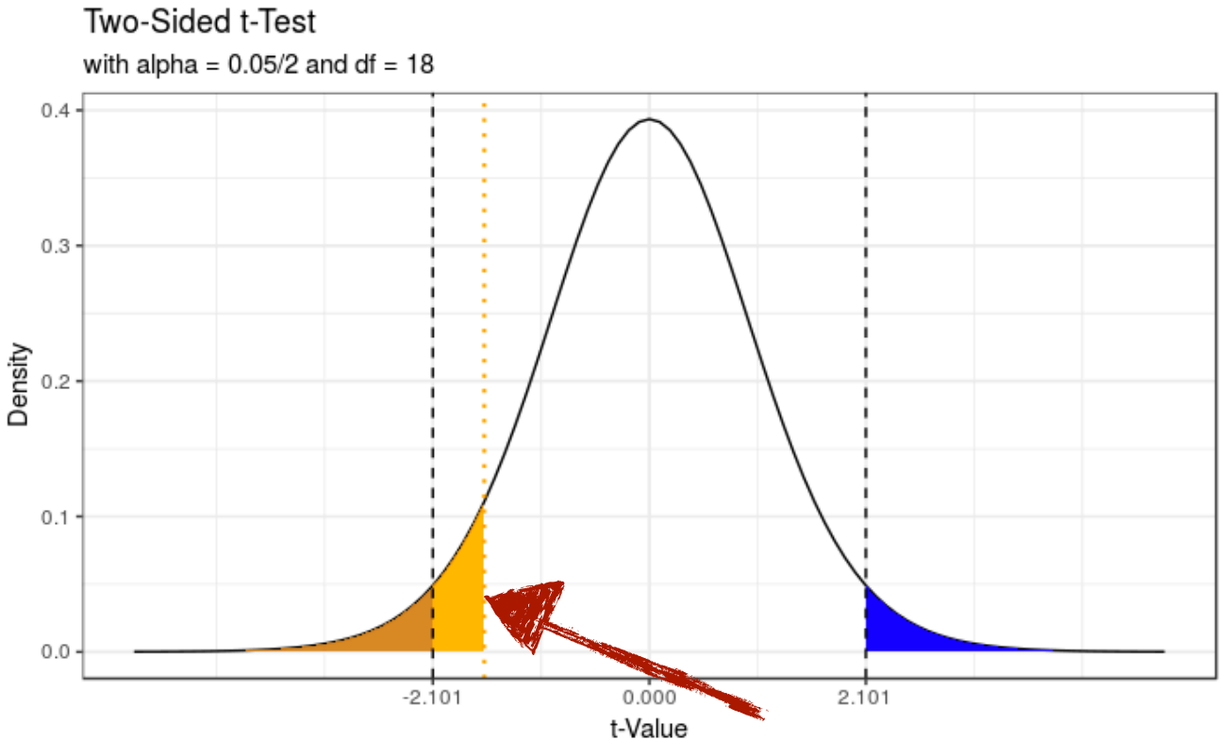
- Since ***b*** is inside the epsilon neighborhood of ***a*** they are **equal.**
- With ***c*** *outside* of ***a***'s epsilon neighborhood, the values are **not equal.**

Definition:

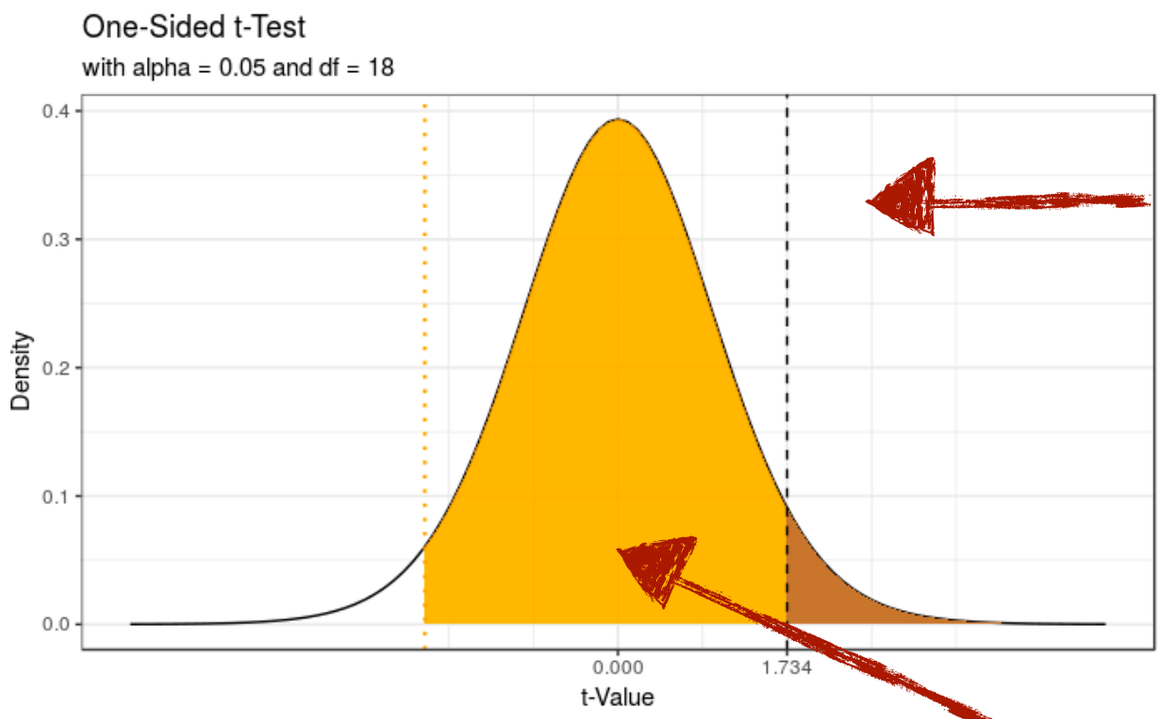
P-value is the probability associated with obtaining the observed value or more extreme values assuming the null hypothesis is true.



Computing a *p*-value



$$2 \cdot \left(1 - P\left(T \leq t\right)\right) = \boxed{0.06323542}$$



$$P\left(T > t\right) = 1 - P\left(T \leq t\right) = \boxed{0.9683823}$$

```
# Two-sided hypothesis
# p-value
2 * (1 - pt(abs(tstat), ndf))
# [1] 0.06323542

# Right tail hypothesis p-value
1 - pt(tstat, ndf)
# [1] 0.9683823

# Start from the right (upper)
# tail instead of the
# left (lower) tail
pt(tstat, ndf,
    lower.tail = FALSE)
# [1] 0.9683823
```

Toothpaste Kisses

… a story of p-values …

Dry

Wet

Wet + Toothpaste

Common Misconceptions

… error'd in our ways …

- *p*-values do **not** indicate the probability that either hypothesis is correct.
    - Probabilities computed relate to how likely the *result is extreme or more extreme* **given** *the null hypothesis is true*.

- Significant outcomes — where *p*-values are small — do **not** indicate a large effect.
    - Small probabilities indicate the unlikeliness of sample being representative of the null hypothesis.

- Non-significant outcomes — where *p*-values are large — do **not** conclusively indicate the

Computing a Critical Value

α

df	0.4	0.25	0.1	0.05	0.025	0.01	0.005	5e-04
1	0.3249197	1.0000000	3.077684	6.313752	2.706205	1.820516	63.656741	636.619249
2	0.2886751	0.8164966	1.885618	2.919986	4.302653	6.964557	9.924843	31.599055
3	0.2766707	0.7648923	1.637744	2.353363	3.182446	4.540703	5.840909	12.923979
4	0.2707223	0.7406971	1.533206	2.131847	2.776445	3.746947	4.604095	8.610302
5	0.2671809	0.7266868	1.475884	2.015048	2.570582	3.364930	4.032143	6.868827
6	0.2648345	0.7175582	1.439756	1.943180	2.446912	3.142668	3.707428	5.958816
7	0.2631669	0.7111418	1.414924	1.894579	2.364624	2.997952	3.499483	5.407883
8	0.2619211	0.7063866	1.396815	1.859548	2.306004	2.896459	3.355387	5.041305
9	0.2609553	0.7027221	1.383029	1.833113	2.262157	2.821438	3.249836	4.780913
10	0.2601848	0.6998121	1.372184	1.812461	2.228139	2.763769	3.169273	4.586894
11	0.2595559	0.6974453	1.363430	1.795885	2.200985	2.718079	3.105806	4.436979
12	0.2590327	0.6954829	1.356217	1.782288	2.178813	2.680998	3.054540	4.317791
13	0.2585909	0.6938293	1.350171	1.770933	2.160369	2.650309	3.012276	4.220832
14	0.2582127	0.6924171	1.345030	1.761310	2.144787	2.624494	2.976843	4.140454
15	0.2578853	0.6911969	1.340606	1.753050	2.131449	2.602480	2.946713	4.072765
16	0.2575992	0.6901323	1.336757	1.745884	2.119905	2.583487	2.920782	4.014996
17	0.2573470	0.6891951	1.333379	1.739607	2.109816	2.566934	2.898230	3.965126
18	0.2571230	0.6883638	1.330391	1.734064	2.100922	2.552380	2.878440	3.921646
19	0.2569228	0.6876215	1.327728	1.729133	2.093024	2.539483	2.860935	3.883406
20	0.2567428	0.6869545	1.325341	1.724718	2.085963	2.527977	2.845340	3.849516
21	0.2565799	0.6863520	1.323188	1.720743	2.079614	2.517648	2.831360	3.819277

$$t\left(1-\alpha/2, df\right) = t\left(1-0.05/2, 18\right)$$

$$= t\left(0.025, 18\right)$$

$$= \boxed{2.100922}$$

$$t\left(1-\alpha, df\right) = t\left(0.05, 18\right) = \boxed{1.734064}$$

```
# Significance Level
alpha = 0.05

# Critical value for
# two-sided test
qt(1 - alpha/2, ndf)
# [1] 2.100922

# Critical value for
# one-sided test
qt(1 - alpha, ndf)
# [1] 1.734064
```

my_ttest()

... implementation as a function...

```
my_ttest = function(x1, x2, test = c("two-sided", "lower", "upper"), alpha = 0.05) {
  # Force `test` to hold a pre-defined value
  test = match.arg(test)
  # Compute length and degrees of freedom
  n1 = length(x1); n2 = length(x2); ndf = n1 + n2 - 2
  # Calculate t-statistic
  s2 = ((n1 - 1) * var(x1) + (n2 - 1) * var(x2)) / ndf
  tstat = (mean(x1) - mean(x2)) / sqrt(s2 * (1 / n1 + 1 / n2))
  # Compute tail probability
  tail_prob = switch(test,
    "two-sided" = 2 * (1 - pt(abs(tstat), ndf)),
    "lower" = pt(tstat, ndf),
    "upper" = 1 - pt(tstat, ndf))
  # Format and return results
  results = list(tstat = tstat, df = ndf, reject = tail_prob < alpha, prob = tail_prob)
  return(results)
}
```

Testing my_ttest()

```r
# Set seed for reproducibility
set.seed(881)

# Generate data
n = 10
x1 = round(rnorm(n), 1)
x2 = round(rnorm(n) + 1, 1)
test_result = my_ttest(x1, x2)
test_result
# $tstat
# [1] -1.979717
# $df
# [1] 18
# $reject
# [1] FALSE
# $prob
# [1] 0.06323542

# Check against built in implementation
all.equal(test_result[-3],
          t.test(x1, x2, var.equal = TRUE)[1:3],
          check.attributes = FALSE)
# [1] TRUE
```

One proportion z-test

… another kind of hypothesis test …

Null → $H_0 : p = 0.5$

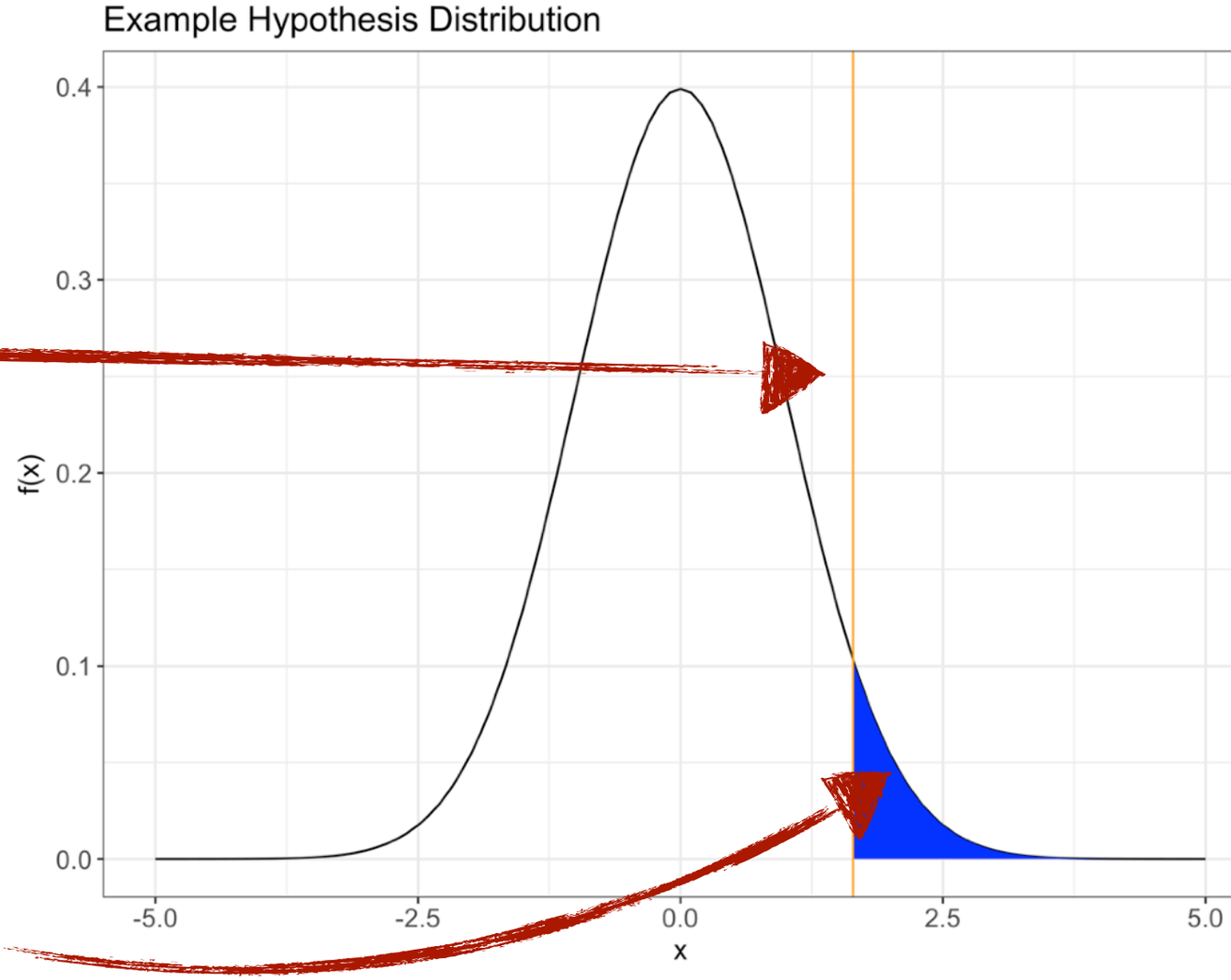
$H_a : p > 0.5$ ← Alternative

$$\hat{p} = \frac{observed}{total}$$

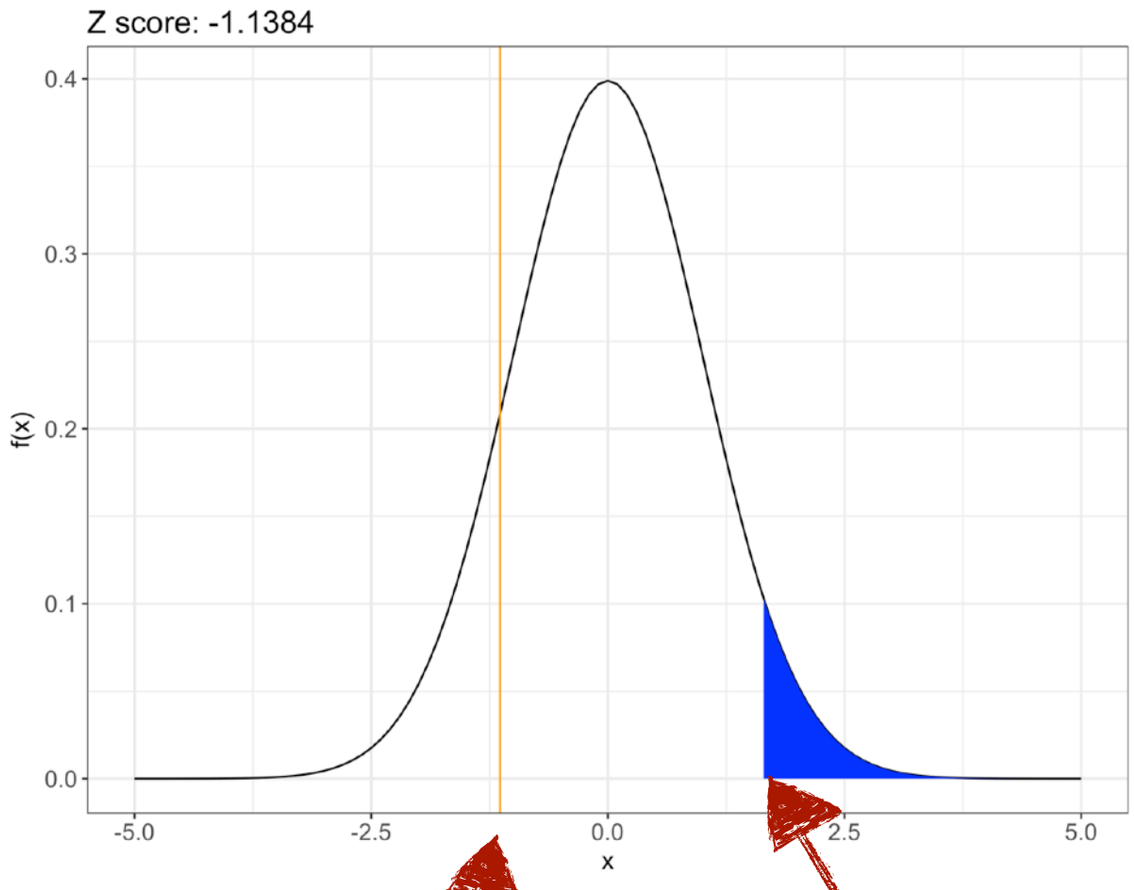
Z score →

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_{1-\alpha}$$

Critical Value →



Coin Flip Bias Example



Test statistic z-score

Critical Region

```
# Initial values
n = 1000; p0 = 0.5
alpha = 0.05

# Simulate Data
set.seed(1337)
x = rbinom(n, 1, 0.5)

# Calculate test values
p_hat = sum(x) / n
z_score = (p_hat - p0) /
                  sqrt(p0 * (1 - p0) / n )
z_crit = qnorm(1 - alpha)
z_score > z_crit
# [1] FALSE
```

And more!

… many different hypothesis tests exist …

Hypotheses	Assumptions	Critical Region
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ or n large, σ^2 known	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ σ^2 unknown	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$z = \dfrac{\bar{x} - \bar{y} - 0}{\sqrt{(\sigma_X^2/n) + (\sigma_Y^2/m)}} \geq z_\alpha$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	Variances unknown, large samples	$z = \dfrac{\bar{x} - \bar{y} - 0}{\sqrt{(s_x^2/n) + (s_y^2/m)}} \geq z_\alpha$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$t = \dfrac{\bar{x} - \bar{y} - 0}{s_p\sqrt{(1/n) + (1/m)}} \geq t_\alpha(n+m-2)$ $s_p = \sqrt{\dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$H_0: \mu_D = \mu_X - \mu_Y = 0$ $H_1: \mu_D = \mu_X - \mu_Y > 0$	X and Y normal, but dependent	$t = \dfrac{\bar{d} - 0}{s_d/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0: p = p_0$ $H_1: p > p_0$	$b(n, p)$ n is large	$z = \dfrac{(y/n) - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$
$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 > 0$	$b(n_1, p_1)$ $b(n_2, p_2)$	$z = \dfrac{(y_1/n_1) - (y_2/n_2) - 0}{\sqrt{\left(\dfrac{y_1+y_2}{n_1+n_2}\right)\left(1 - \dfrac{y_1+y_2}{n_1+n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \geq z_\alpha$

Back cover of: Probability and Statistical Inference (9th Edition)

Recap

- **Testing Frameworks**
  - Steps of a Hypothesis Test
  - Sampling distributions play a large role.
  - Case study in implementing a method for assessing group difference.

Acknowledgements

Acknowledgements

- Chapter 4: Classes of S Programming by W.N. Venable and B.D. Ripley

Changelog

- v2
  - Fixed probability area of one graph, included book cover jacket containing hypothesis test