

A/B Testing

Treatment and control, randomization, test statistic, Null and Alternative Hypothesis, One-tailed and Two-Tailed Test, p-value, alpha, Bootstrap sampling, Statistical Simulation, independent and paired samples, comparing means (t-test), and comparing proportions (Z-test, chi-squared test, McNamir test)

Nagiza F. Samatova, samatova@csc.ncsu.edu

Professor, Department of Computer Science
North Carolina State University

Examples: A/B Tests

- Testing two prices to determine which yields more **net profit**
- Testing two **web headlines** to determine which produces more **clicks**
- Testing two **web ads** to determine which generates more **conversions**
- ...
- see 60 A/B tests chapter for more examples

A/B Tests: The Two Sample Comparison

Is treatment A different from treatment/control B?

- **Applications: Treatment, intervention**
 - a drug
 - a medical device
 - an advertising campaign
 - a price
 - a manufacturing procedure
- **Is the observed difference between two samples due to chance or due to a real difference in the treatments?**

A/B Test

- **A/B Test**

- An experiment with two groups to establish which of **two treatments**, products, procedures, or the like is **superior**
- Often one of the treatments is the standard existing treatment, or no treatment

- **Control vs. treatment**

- If a standard (or no treatment) is used, it is called the **control**

- **Typical Hypothesis**

- **Treatment is better than control**
- Example: *The product price A is more profitable than the price B*

- **Applications**

- Marketing
- Web design

A/B Testing: Key Ideas

- **Subjects are assigned (ideally, randomly) to two (A and B) groups**
 - that are treated exactly alike
 - except that the treatment under study differs from one another
- **A metric (**test statistic**) is used to compare group A to group B**

Test Statistic

- **Test statistic**

- A metric (e.g., difference in means, difference in proportions, difference in risks) used to compare group A to group B

- **Examples**

- **Binary variables**
 - click or no-click
 - buy or don't buy
 - fraud or no fraud
- **Categorical count variables:**
 - contracts signed
 - pages visited
- **Continuous variables**
 - purchase amount
 - profit
 - revenue per page-view (rather than conversion)

2 x 2 table for eCommerce Experiment

Outcome	Price A	Price B
Conversion	200	182
No conversion	2,353	2,240
Proportion	0.085	0.081

Mean Revenue per Page Conversion

Outcome	Price A	Price B
Revenue	\$3.87	\$4.11

A/B Testing: Comparing Two Means

INDEPENDENT SAMPLES OF **CONTINUOUS** **VARIABLE**

A/B Hypothesis Testing Procedures

Hypothesis to Test

R function

Statistical simulation

Mathematical Formula

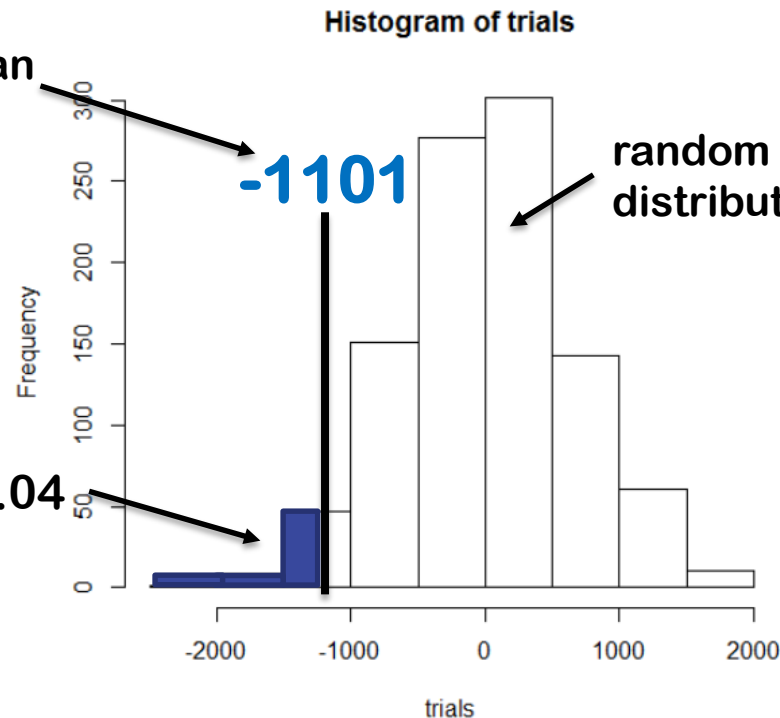
t.test()

sample mean difference

-1101

random distribution

p-value = 0.04



$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

t-statistic

A/B Hypothesis Testing Procedures

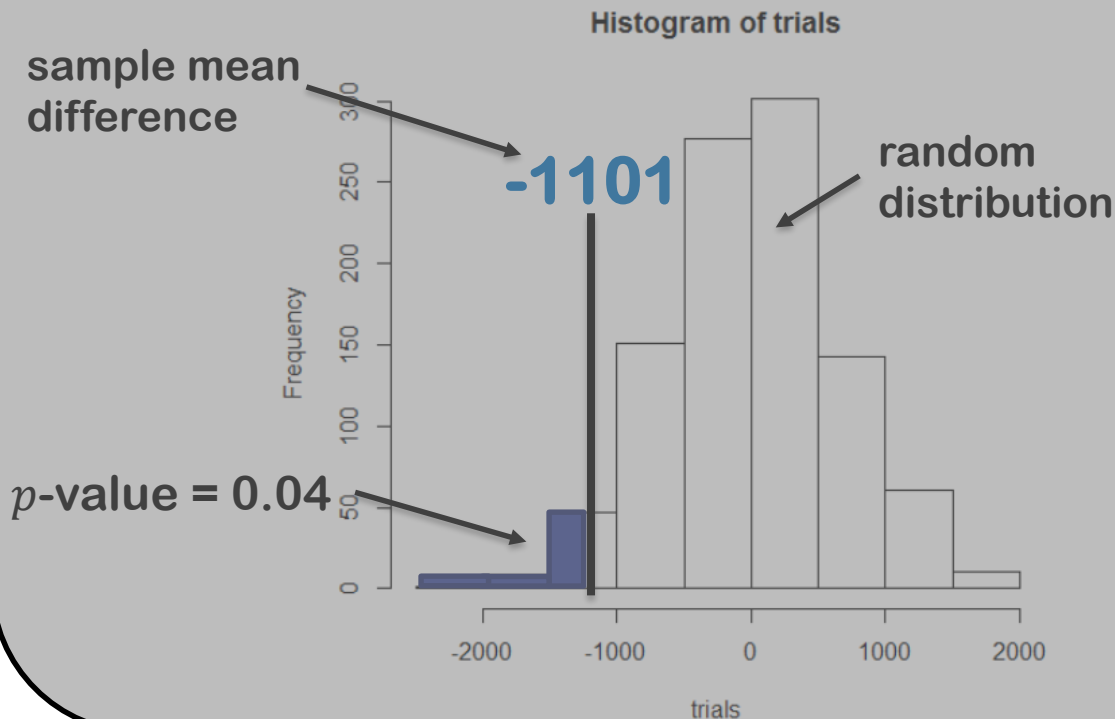
Hypothesis to Test

R function

`t.test()`

Statistical simulation

Mathematical Formula



$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

t-statistic

Example: Comparing Two Means

Is the difference in means between Treatment & Control statistically significant?

- Ten pigs were randomly assigned to a Treatment group with a new blood clotting agent
- Ten pigs were assigned to a Control group that did not receive the clotting agent
- Each pig's liver was injured in a controlled manner and blood loss was measured

- **Sample Statistic**

- Blood loss
- Continuous variable
- Compare means

```
8 df <- read.table(file = "../data_raw/pigblood.table.txt",
9                  header=TRUE)
10 control = df[,1]
11 treatment = df[,2]
12
13 cat("Mean Control: ", mean(control), "\n")
14 cat("Mean Treatment: ", mean(treatment), "\n")
15 cat("Difference in means: ",
16     mean(treatment) - mean(control), "\n")
```

File: AB_test_compare_means.R

Control	Treatment
786	543
375	666
4446	455
2886	823
478	1716
587	797
434	2828
4764	1251
3281	702
3837	1078
Mean Control: 2187	
Mean Treatment: 1086	
Difference: -1101	

Approach #1: Comparing Two Means via **t.test()**

Alternative Hypothesis: H1: **Control Mean is greater than Treatment Mean**

- **H₀**: no difference in the mean blood loss between two groups except for what chance might produce

```
22 # alternative = "greater":  
23 #   x has a larger mean than y  
24 t.test (x=control,  
25         y=treatment,  
26         alternative="greater")
```

File: AB_test_compare_means.R

Mean Control: 2187
Mean Treatment: 1086
Difference: **-1101**

```
      welch Two Sample t-test  
  
data: control and treatment  
t = 1.777, df = 11.717, p-value = 0.05075  
alternative hypothesis: true difference in means is  
greater than 0  
95 percent confidence interval:  
-5.494969      Inf  
sample estimates:  
mean of x mean of y  
  2187.4    1085.9
```

p-value is statistically significant (alpha = 0.1)

There is a ~5% chance of observing such a difference by chance

Examples: Null and Alternative Hypotheses

- **Null = H_0 = “no difference between the means of group A and group B”**
 - Alternative = H_1 = “A is different from B” (could be bigger or smaller)
- **Null = H_0 = “ $A \leq B$ ”**
 - Alternative = H_1 = “ $B > A$ ”
- **Null = H_0 = “ $A > B$ ”**
 - Alternative = H_1 = “ $B \leq A$ ”

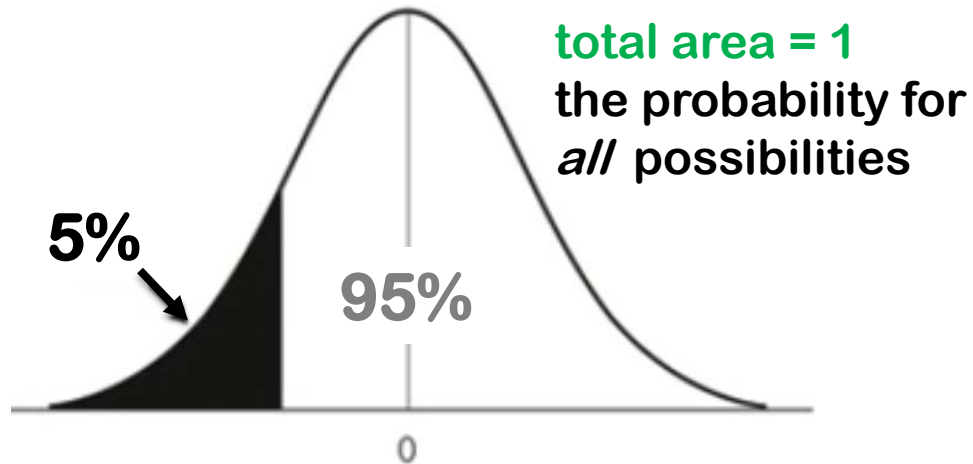
- The Null and Alternative Hypotheses must account for *all* possibilities.
- The nature of the null hypothesis determines the structure of the hypothesis test.

One-tailed or Two-tailed Hypothesis Tests

Statistical Significance Level, $\alpha = 0.05$

defined before the experiment starts

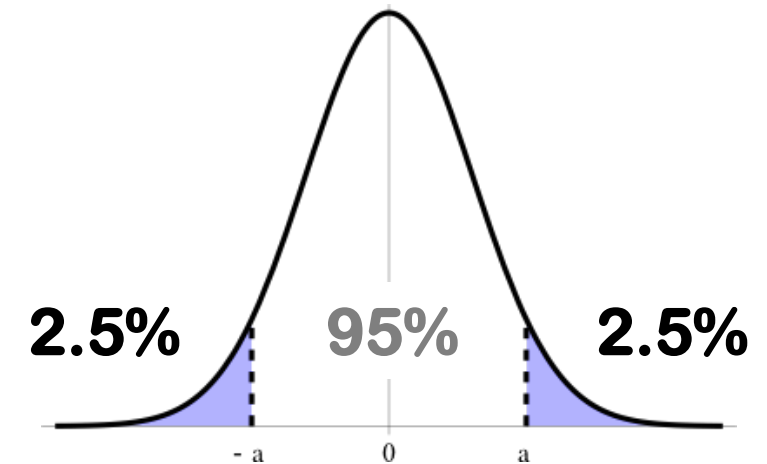
One-tailed test



p-value $< 0.05 = \alpha$
area under the curve < 0.05

alternative = c("less")

Two-tailed test

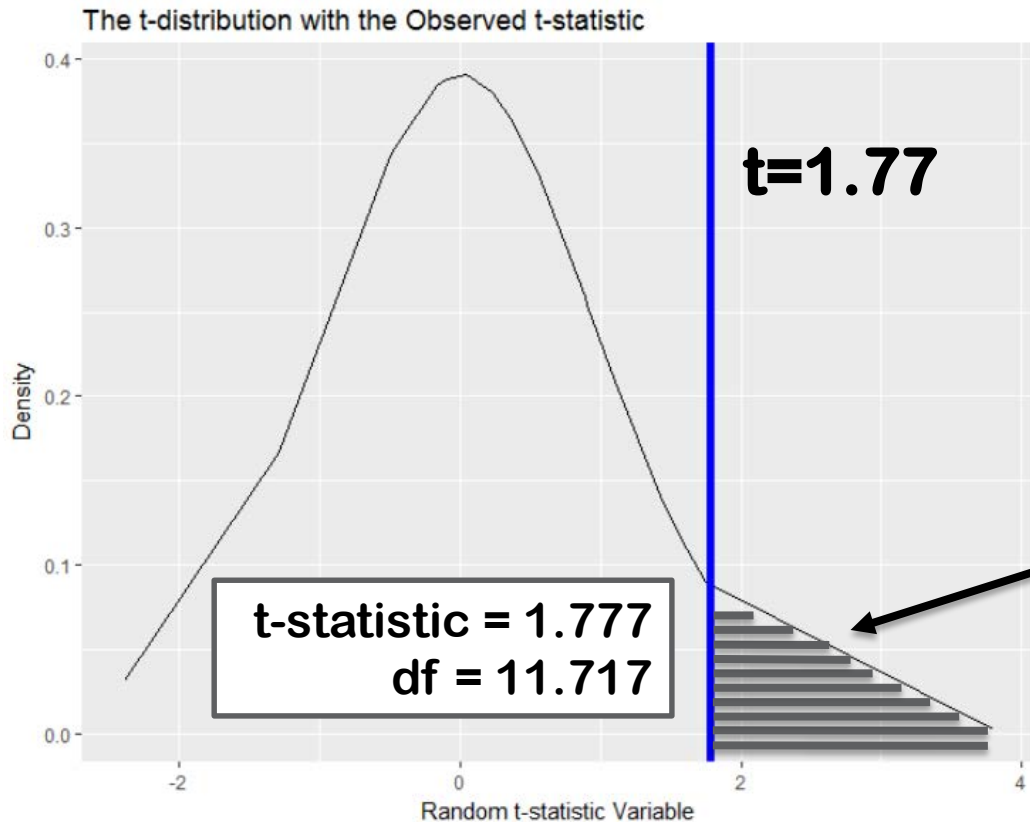


p-value $< 0.025 = \frac{\alpha}{2}$

alternative = c("two.sided")

p-value is the probability of observing such an effect by random chance

Approach #1: **t-statistic** and **df** reported by t.test()



```
31 # access individual values
32 res <- t.test (x=control,
33               y=treatment,
34               alternative="greater")
35 res$p.value
36 res$statistic ← t-statistic
37 res$parameter ← degrees of freedom
```

File: AB_test_compare_means.R

```
18 # P[X > t.statistic]
19 pt (t.statistic,
20     df=degrees.of.freedom,
21     lower.tail = FALSE)
```

File: AB_test_t_distribution.R

```
14 # Generate n random numbers from a t-distribution
15 # with the specified degrees of freedom
16
17 tRandom <- rt (n, df=degrees.of.freedom)
```

Statistical Distributions & Functions in R




Distribution	Random Number Generator	Density	Distribution	Quantile
Normal	r norm	d norm	p norm	q norm
<i>t</i>	rt	dt	pt	qt
<i>F</i>	rf	df	pf	qf
χ^2	rchisq	dchisq	pchisq	qchisq

{d p q r} *distribution_abbreviation*()

- **d** = density
- **p** = distribution function
- **q** = quantile function
- **r** = random generation

- **pnorm(a)** $\equiv P(X \leq a)$: probability that *a* or smaller number occurs
- **pnorm(b) - pnorm(a)** $\equiv P(a \leq X \leq b)$: probability that the variable falls between two points
- **qnorm()**: given the cumulative probability distribution, it returns the quantile

Statistical Distributions: Mean and Variance



Distribution	Degrees of freedom	Mean	Variance
Normal		μ	σ^2
t	n	0	$n/(n-2)$
F	n_1 and n_2	$n_2/(n_2-2)$	a/b
χ^2	r	r	$2r$

$$a = 2n_2^2(n_1 + n_2 - 2)$$

$$b = n_1(n_2 - 2)^2(n_2 - 4)$$

A/B Hypothesis Testing Procedures

Hypothesis to Test

R function

Statistical simulation

Mathematical Formula

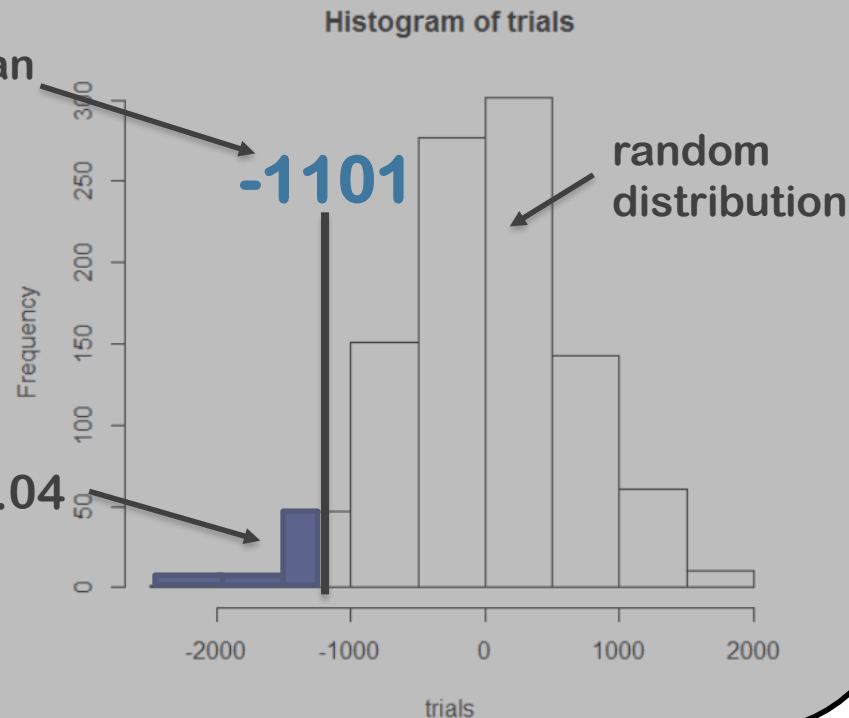
`t.test()`

sample mean difference

-1101

random distribution

p-value = 0.04



$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

t-statistic

Test Statistic: Typical General Form

$$\text{Test Statistic} = \frac{\text{Observed Differences}}{\text{Standard Error}}$$

- **Observed difference**

- Difference between the observed mean of group A and observed mean of group B
- Difference between the observed proportions for group A and observed proportions for group B (e.g., proportions: ad clicks vs. total web page visits)

- **Standard error:**

- The variability (standard deviation) of a sample statistic over many samples

- **Standard deviation:**

- refers to variability of individual data values within a single sample

Approach #2: **Formula** for t -statistic to compare two means

The difference between the mean of the treatment and the mean of the control comes from the **t-distribution**. The statistical significance of the observed difference is the probability **P [X > t.statistic]** (one-tailed).

t-statistic formula

$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

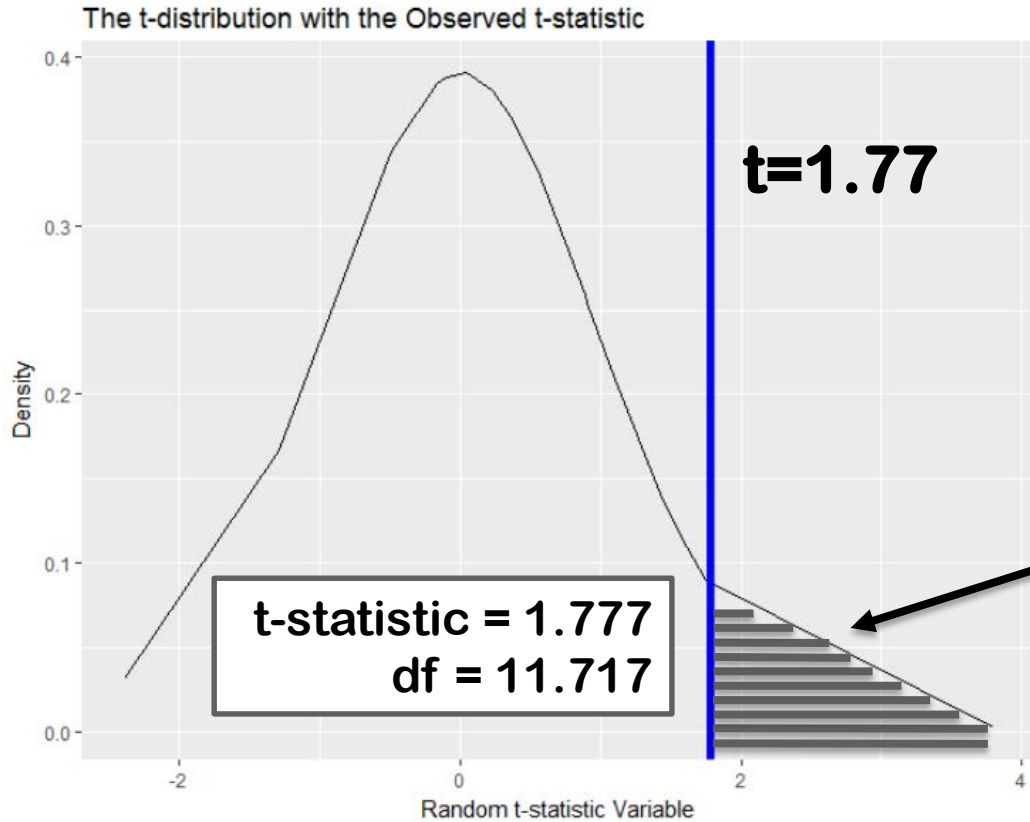
\bar{x}_B = The sample average for treatment B

s_B^2 = The variance for treatment B

n_B = The sample size for treatment B

```
8 df <- read.table(file = "../data_raw/pigblood.table.txt",
9                   header=TRUE)
10 control <- df[,1]
11 treatment <- df[,2]
12 n.control <- length(control)
13 n.treatment <- length(treatment)
14
15 observed.difference <- mean (control) -
16                        mean (treatment)
17
18 standard.error <- sqrt (var(treatment)/n.treatment +
19                        var(control)/n.control)
20
21 t.statistic <- observed.difference / standard.error
22 t.statistic
```

Approach #2: **t-statistic** and **df** from math formulas



$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{b + a}}$$

$$df = \frac{b + a}{\frac{b^2}{n_B - 1} + \frac{a^2}{n_A - 1}}$$

$$a = \frac{s_A^2}{n_A}$$

$$b = \frac{s_B^2}{n_B}$$

P [X > t.statistic] (one-tailed)

```
19 pt (t.statistic,  
20   df=degrees.of.freedom,  
21   lower.tail = FALSE)
```

File: AB_test_t_distribution.R

```
14 # Generate n random numbers from a t-distribution  
15 # with the specified degrees of freedom  
16  
17 tRandom <- rt (n, df=degrees.of.freedom)
```

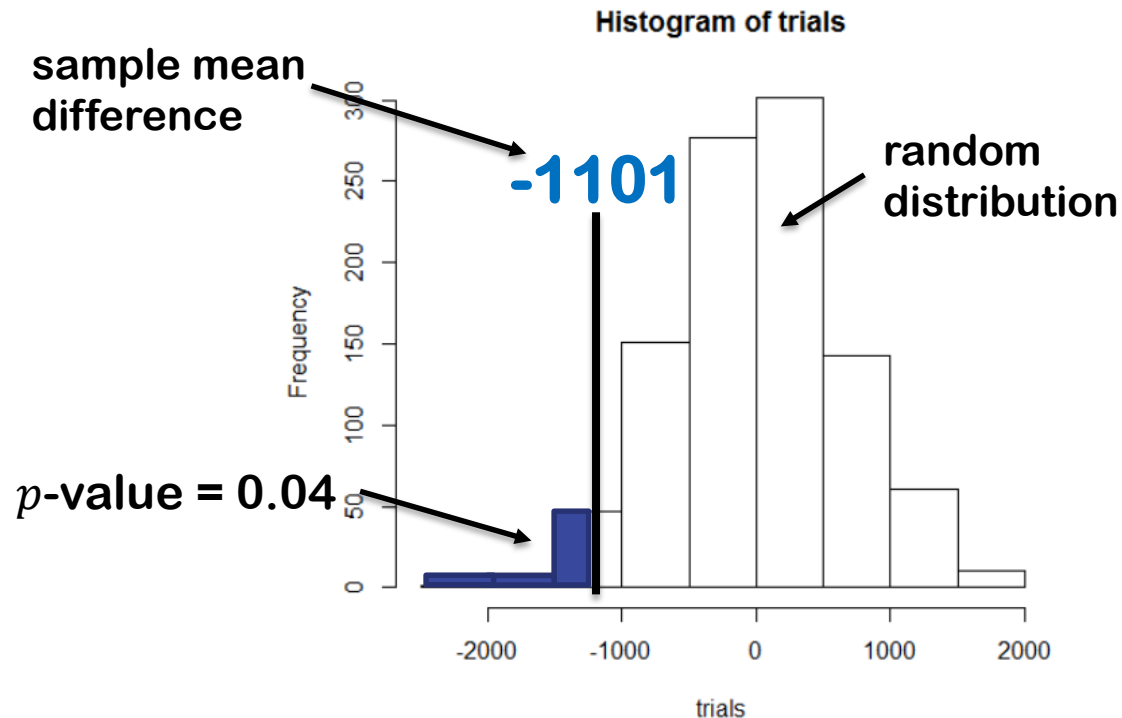
A/B Hypothesis Testing Procedures

Hypothesis to Test

R function

`t.test()`

Statistical simulation



Mathematical Formula

$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

t-statistic

Approach #3: **Statistical Simulation** to Compare two Means

Null Hypothesis: Ho: Differences between Control & Treatment Means are the same

- Ho: no difference in the mean blood loss between two groups except for what chance might produce

Original Sample →

Control Treatment

786	543
375	666
4446	455
2886	823
478	1716
587	797
434	2828
4764	1251
3281	702
3837	1078

Difference: -1101

1. Shuffle Combined Control and Treatment (A+B) data

```
> shuf = sample (vector.data, replace = FALSE)
> shuf
[1] 543 823 4446 3281 786 3837 1716 587 666 1251
[11] 455 2828 375 1078 797 2886 702 478 4764 434
```

2. Split into New Control and New Treatment data

```
> control = shuf[1:10]
> treatment = shuf[11:20]
> control
[1] 543 823 4446 3281 786 3837 1716 587 666 1251
> treatment
[1] 455 2828 375 1078 797 2886 702 478 4764 434
```

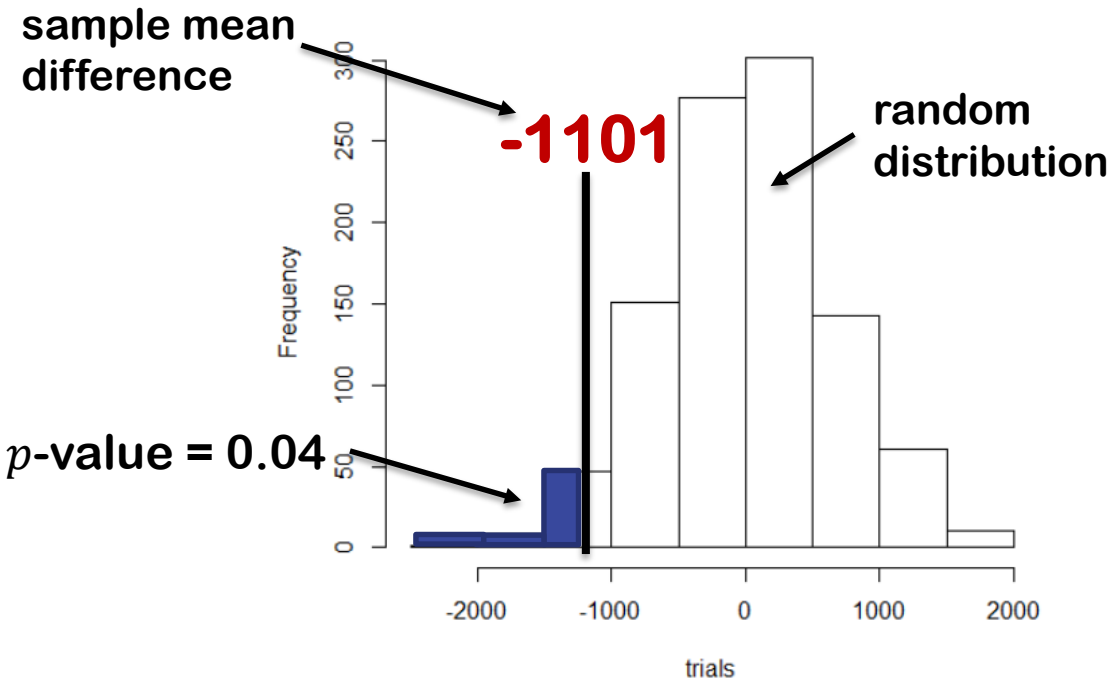
3. Compare means for New Control and New Treatment data

```
> trial = mean (treatment) - mean (control)
> trial
[1] -313.9
```

4. Repeat steps 1-3 for 1,000 trials, plot histogram, compute p-value

A/B Test: Pig Blood

Histogram of trials



$$p\text{-value} = 0.044 < 0.05$$

A difference in blood loss ≤ -1101 ml occurred only 44 times out of 1,000 under the chance model → chance is not responsible and treatment is effective.

Result is statistically significant →
Reject the Null Hypothesis

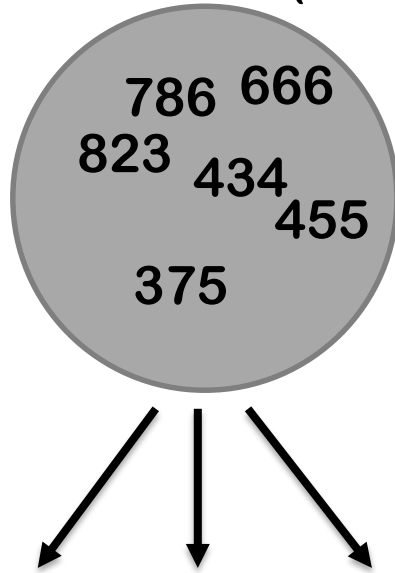
```
39 n.repeats <- 1000
40 with.replacement <- TRUE
41 trials <- replicate(n.repeats,
42                     simulation.trial(vector.data,
43                                     with.replacement,
44                                     mean))
45 hist(trials)
46 # The mean.diff is negative, so use <=
47 pval <- ifelse(trials <= mean.diff, 1, 0)
48 cat("p-value: ", sum(pval)/n.repeats, "\n")
49
```

```
9 simulation.trial <- function(data.vector,
10                              replace=TRUE,
11                              fun.name) {
12   shuffle <- sample(data.vector,
13                     length(data.vector),
14                     replace)
15   shuffle.matrix <- matrix(shuffle, ncol=2,
16                             nrow=length(data.vector),
17                             byrow=TRUE)
18   statistic <- fun.name(shuffle.matrix[,2]) -
19                 fun.name(shuffle.matrix[,1])
20   return(statistic)
21 }
```

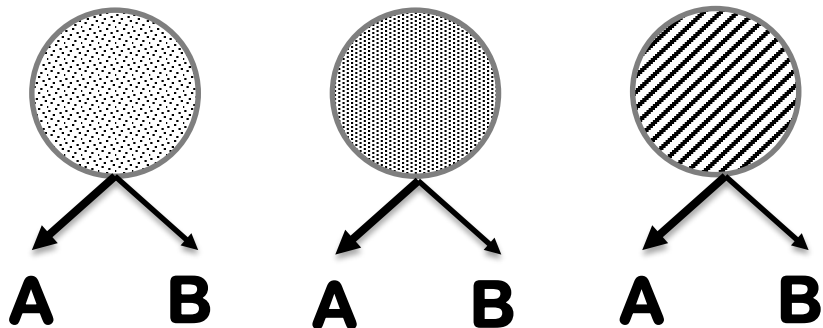
File: AB_test_compare_means_simulation.R

Bootstrap Sampling: With or Without Replacement

Original Combined (A+B) Sample



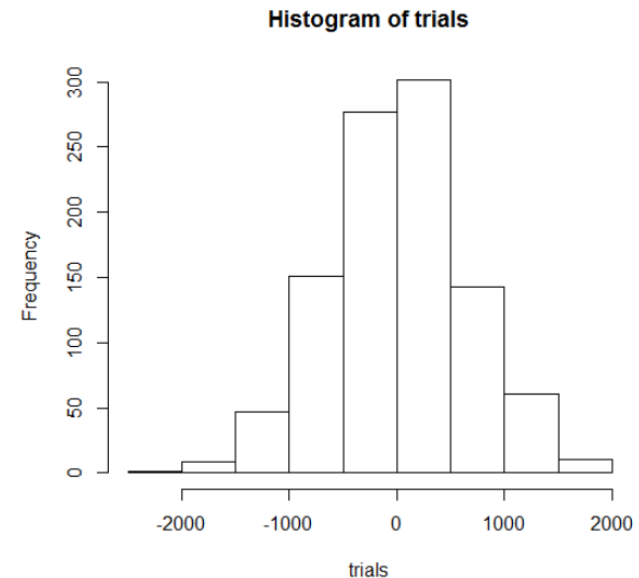
Draw 1,000 **resamples**, **with** or **without replacement**: the **same size** as A+B



histogram of:
 $\text{mean}(B) - \text{mean}(A)$

```
9 simulation.trial <- function(data.vector,  
10                             replace=TRUE,  
11                             fun.name) {  
12   shuffle <- sample(data.vector,  
13                     length(data.vector),  
14                     replace)  
15   shuffle.matrix <- matrix(shuffle, ncol=2,  
16                           nrow=length(data.vector),  
17                           byrow=TRUE)  
18   statistic <- fun.name(shuffle.matrix[,2]) -  
19                 fun.name(shuffle.matrix[,1])  
20   return(statistic)  
21 }
```

File: AB_test_compare_means_simulation.R



Summary: What we covered so far

- **A/B Test**

- Control (A) and Treatment (B) groups or
- Treatment (A) and Treatment (B)

- **Null Hypothesis and Alternative Hypothesis**

- One-tailed vs. Two-tailed Test
- p-value ($p\text{-value} < \alpha$): to have a statistically significant conclusion
- statistical significance, α : probability of having the observed effect due to chance

- **Hypothesis Testing Procedures:**

- R function call
- Mathematical formula
- Statistical simulation
 - Bootstrap sampling (with or without replacement)

- **Test Statistics Metric**

- t-statistic: comparing the differences in group means (continuous variable)

- **t-distribution**

- `rt()` and `pt()`

A/B Testing: Comparing Two Means

PAIRED SAMPLES OF **CONTINUOUS** VARIABLE

Control (A) and Treatment (B) Groups

Two (2) Samples: A and B

Independent

- Customers randomly assigned to one of the two groups

Dependent, or **Paired**

- **Follow-up studies:**
 - Group A: At the beginning of the study
 - Group B: At the end of the study
- The **same** customers exposed to both:
 - Marketing strategy A
 - Marketing strategy B

Paired Comparison: Dependent Samples

- **Example: Contribution of music to cognitive learning**
 - Reading comprehension scores for 11 subjects:
 - Sample-A: initially after reading a passage without background music
 - Sample-B: a week later after reading a similar passage with background music
- **Sample-A and Sample-B are NOT independent:**
 - the **same subjects** participated in the intervention
 - the observed **mean difference is 1.45**

Without Music	With Music	Difference
24	27	+3
79	80	+1
17	18	+1
50	50	0
98	99	+1
45	47	+2
97	97	0
67	70	+3
78	79	+1
85	87	+2
76	78	+2

Question:

Is this observed difference statistically significant?

A/B Hypothesis Testing Procedures

Hypothesis to Test

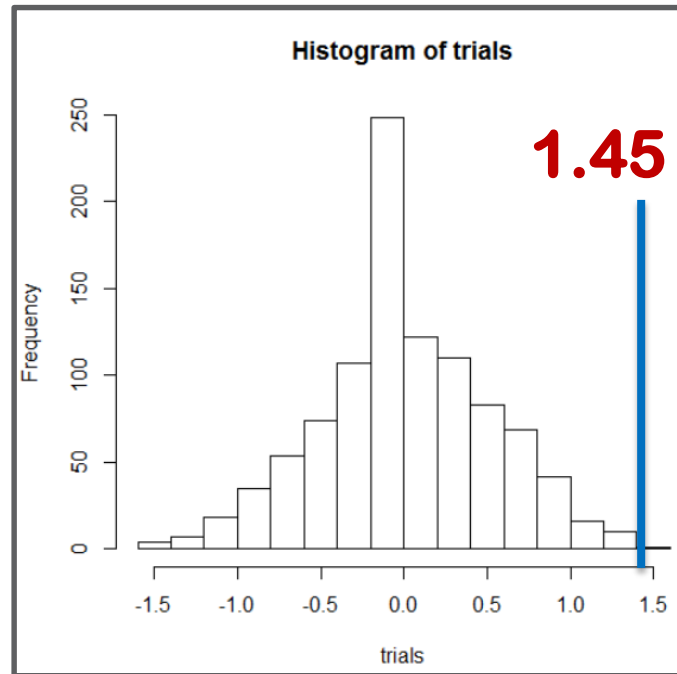
R function

Statistical simulation

Mathematical Formula

t.test()

paired = TRUE



$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

t-statistic

(BONUS paper on paired t-test)

Approach #1: Paired t.test()

Alternative Hypothesis: H1: Treatment Mean is greater than Control Mean
Ho: no difference in the mean reading scores between two groups except for what chance might produce

Paired t.test()

```
24 # alternative = "greater":  
25 #     x has a larger mean than y  
26 # paired = TRUE  
27 t.test (x=treatment,  
28         y=control,  
29         alternative="greater",  
30         paired = TRUE)
```

AB_paired_test_music_affect_on_reading.R

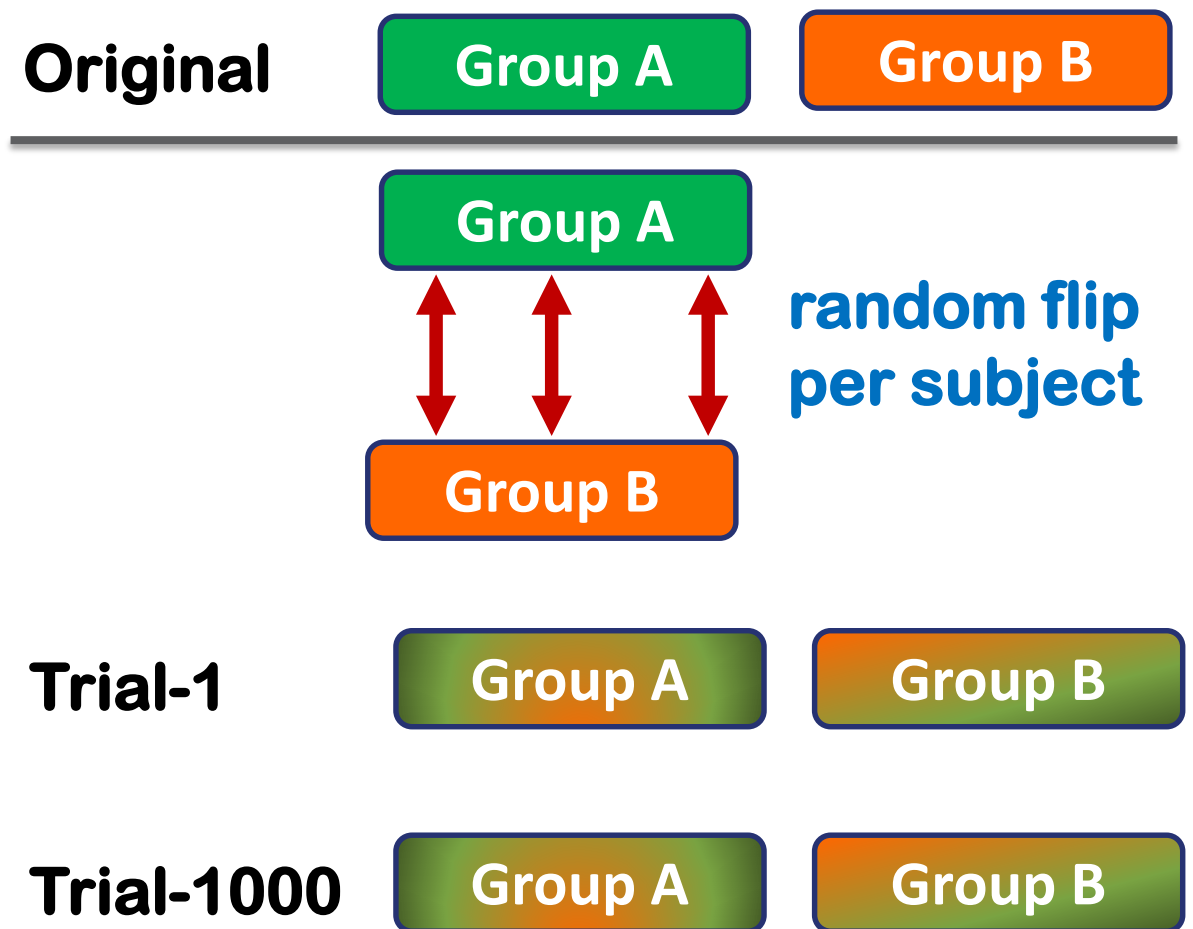
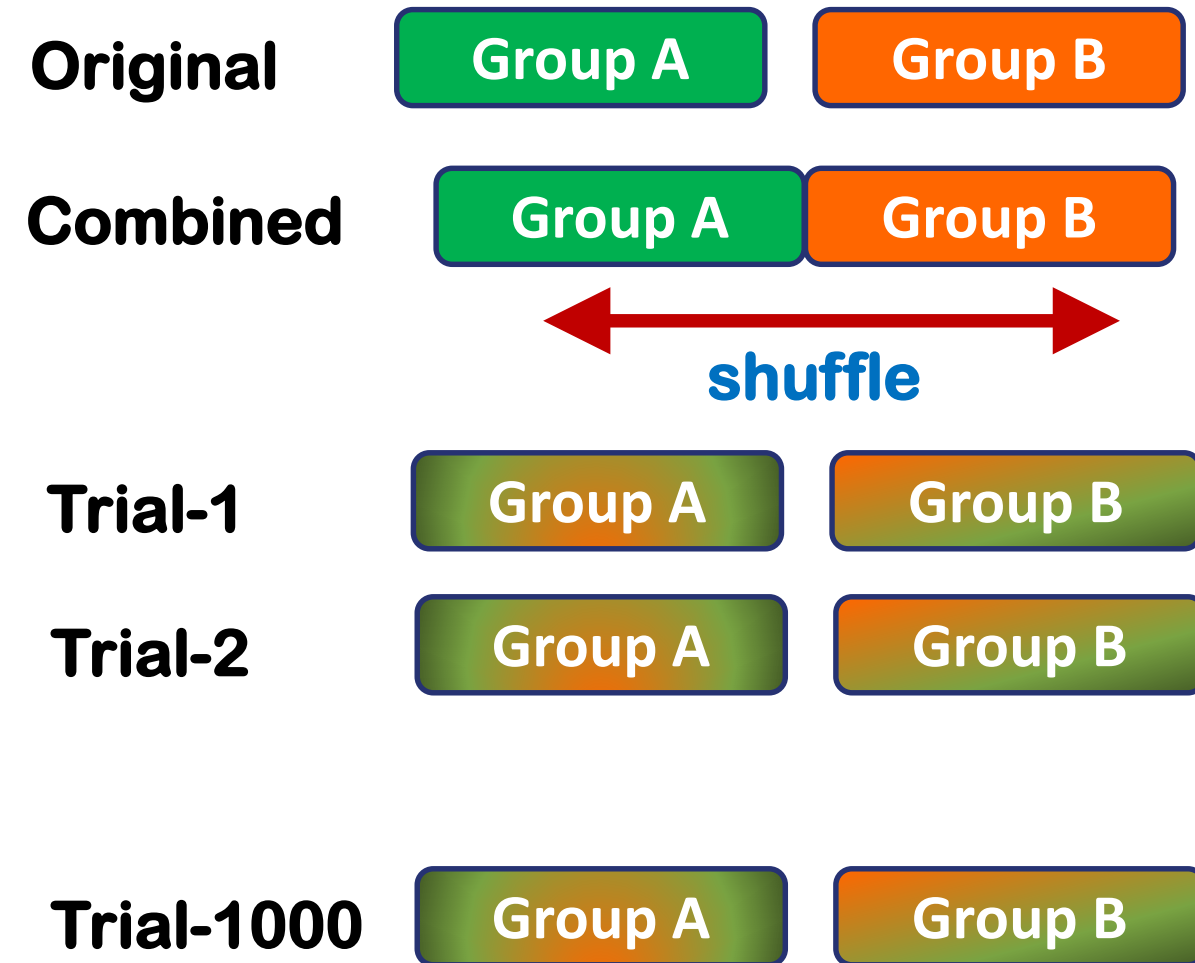
Output

```
      Paired t-test  
data:  treatment and control  
t = 4.6578, df = 10, p-value = 0.0004487  
alternative hypothesis: true difference in means is  
greater than 0  
95 percent confidence interval:  
 0.8885447      Inf  
sample estimates:  
mean of the differences  
          1.454545
```

Approach #2: Paired Comparison Test: **Simulation**

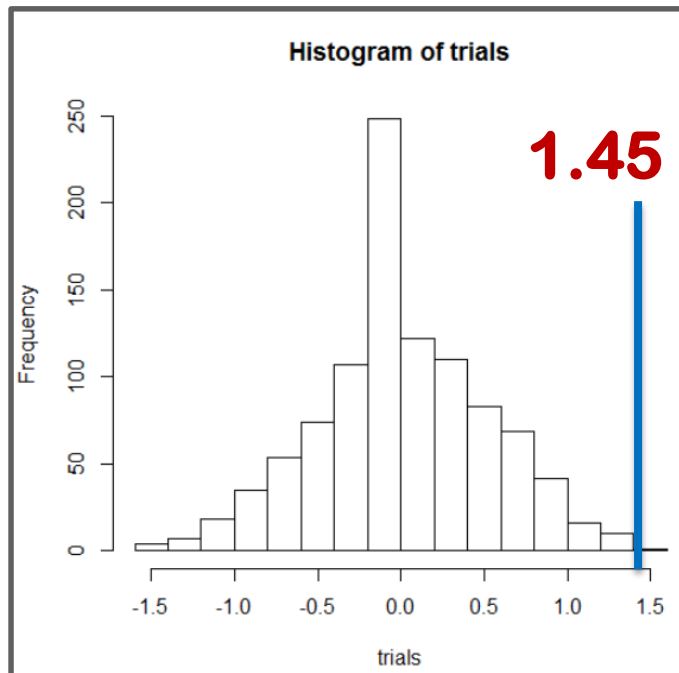
Previous approach will **NOT** work

CORRECT PAIRED Approach



Approach #2: Paired Comparison Test: **Simulation**

1. Randomly shuffle the two scores for the first subject into columns 1 and 2
2. Repeat the shuffle for the remaining 10 subjects
3. Calculate the mean score for columns 1 and 2 and record the difference: col 2 – col 1
4. Repeat steps one through three 1,000 times
5. Draw a histogram of the resampled differences and find out how often the resampled difference exceeds the observed value of 1.45



```
10 reading.scores <- scan(file="../data_raw/reading.txt")
11 score.matrix <- matrix(reading.scores,ncol=2,nrow=11,byrow=TRUE)
12 # verify the matrix is correct
13 write.table(score.matrix, sep=" ", row.names=FALSE,
14             col.names=c("Without Music", "With Music"))
15
16 set.seed (2018)
17
18 res <- function() {
19   # apply applies the sample function to each *row* in mat
20   shuffle.within.rows <- apply(score.matrix, 1,
21                               function(mat) sample(mat,replace=F))
22   # must transpose the shuffled matrix to get 2 column format
23   mat2 <- t(shuffle.within.rows)
24   diffmeans <- mean(mat2[,2]) - mean(mat2[,1])
25   return(diffmeans)
26 }
27
28 n <- 1000
29 trials <- replicate(n, res())
30 hist(trials)
31 pval <- sum(ifelse(trials>=1.45,1,0))
32 cat("Paired comparison p-value: ", pval/n,"\n")
```

AB_paired_test_music_affect_on_reading_simulation.R

$p\text{-value} = 0.001 < 0.05$

A/B Test: Comparing Two Proportions

INDEPENDENT SAMPLES OF **BINARY /**
CATEGORICAL VARIABLE

Test Statistic

- **Test statistic**

- A metric (e.g., difference in means, difference in proportions, difference in risks) used to compare group A to group B

- **Examples**

- **Binary variables**

- click or no-click
- buy or don't buy
- fraud or no fraud

- **Categorical count variables:**

- contracts signed
- pages visited

- **Continuous variables**

- purchase amount
- profit
- revenue per page-view

2 x 2 table for eCommerce Experiment

Outcome	Price A	Price B
Conversion	200	182
No conversion	2,353	2,240
Proportion	0.085	0.081

Mean Revenue per Page Conversion

Outcome	Price A	Price B
Revenue	\$3.87	\$4.11

Comparing Two Proportions

2 x 2 table for eCommerce Experiment

Outcome	Price A	Price B
Conversion	200	182
No conversion	2,353	2,240
Proportion	0.085	0.081

Test

Parametric: Z-test

Non-parametric: Chi-squared Test: `prop.test()`

$$Z = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{p(1-p) \left(\frac{1}{n_B} + \frac{1}{n_A} \right)}}$$

Z-statistic


Assumptions: the probability of common success is approximate 0.5, and the number of games is very high: i.e., a binomial distribution approximates a gaussian distribution).

```
6 converters <- c(200, 182)
7 group.sizes <- c(2353, 2240)
8
9 prop.test (x=converters, n=group.sizes,
10           alternative = "greater",
11           correct = FALSE)
12 AB_test_compare_proportions.R
```

2-sample test for equality of proportions
without continuity correction

```
data: converters out of group.sizes
X-squared = 0.21139, df = 1, p-value = 0.3228
alternative hypothesis: greater
95 percent confidence interval:
 -0.00965319  1.00000000
sample estimates:
 prop 1      prop 2
0.08499788 0.08125000
```

Statistical Distributions & Functions in R

Distribution	Random Number Generator	Density	Distribution	Quantile
Normal	r norm	d norm	p norm	q norm
<i>t</i>	rt	dt	pt	qt
<i>F</i>	rf	df	pf	qf
 χ^2	rchisq	dchisq	pchisq	qchisq

{d p q r} *distribution_abbreviation*()

- **d** = density
- **p** = distribution function
- **q** = quantile function
- **r** = random generation

- **pnorm(a)** $\equiv P(X \leq a)$: probability that *a* or smaller number occurs
- **pnorm(b) - pnorm(a)** $\equiv P(a \leq X \leq b)$: probability that the variable falls between two points
- **qnorm()**: given the cumulative probability distribution, it returns the quantile

Example: Z-statistic for **two proportions**

- **Metric:**
 - a percentage
 - a proportion

$$Z = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{p(1-p) \left(\frac{1}{n_B} + \frac{1}{n_A} \right)}}$$

Outcome	Price A	Price B
Conversions, X	42,480	42,551
Sample size, n	50,332	49,981

n_B = The sample size for treatment B

X_B = The number of conversions for treatment B

$\hat{p}_B = \frac{X_B}{n_B}$ = The point estimate for the proportion of converted for treatment B

$p = \frac{X_A + X_B}{n_A + n_B}$ = The combined conversion rate

Standard error for two proportions:

- obtained by combining all values (irrespective of original group), computing the combined proportion, and then normalizing by the two sample sizes

A/B Testing: Summary

Two Samples	Compare means	Compare proportions
Independent	Student's t-test: <code>t.test()</code>	Non-parametric Chi-squared test: <code>prop.test()</code> or <code>chisq.test()</code> Parametric: Z-test
Paired	<code>t.test(paired=TRUE)</code>	McNemar's test: <code>mcnemar.test()</code>

See more examples: <http://yatani.jp/teaching/doku.php?id=hcistats:chisquare>

Randomization in A/B Testing

- **Randomization**

- The process of **randomly assigning** subjects to treatment groups

- **Why is randomization necessary?**

- To make sure that any difference between treatment groups is due to one of two things:
 - The effect of the different treatments
 - Random assignment may have resulted in the naturally better-performing subjects being concentrated in A or B:
 - e.g. most African-American athletes from the sample ended up being on the same basketball team
 - e.g., most kids from economically advantageous families ended up being assigned to the same group for new computerized test design A vs. paper-based test design B

Why Have a Control Group?

- **Why not skip the control group and just run an experiment applying the treatment of interest to only group, and compare the outcome to prior experience?**
- **Rationale:**
 - Without a control group, there is no assurance that “other things are equal” and that any differences is really due to the treatment (or to chance)
 - Control groups is subject to the SAME conditions (EXCEPT for the treatment of interest) as the treatment group
 - If comparison is made simply to “baseline” or prior experience, then other factors, besides the treatment, might differ

Requesting Permission

- **In studies involving human or animal subjects, it is typically necessary to get their permissions (special process)**
- **Facebook Experiment with Emotional tone in users' newsfeeds (2014)**
 - Used sentiment analysis to classify newsfeed posts as positive or negative
 - Then altered the pos/neg balance in what it showed to users:
 - some randomly selected users experienced more positive posts
 - while others more negative ones
 - Finding: users exposed to more positive newsfeed were more likely to post positively themselves and vice versa
- **Issue with Facebook Experiment**
 - Users were subjects without their knowledge
 - What if Facebook pushed some extremely depressed users over the edge if they got negative version of their feed?