



Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

RNI1006 REGRESSION AND NONPARAMETRIC INFERENCE

Your pathway to Curtin. On campus. On track.

www.curtincollege.edu.au

COMMONWEALTH OF AUSTRALIA

Copyright Regulations 1969

WARNING

*This material has been reproduced and communicated to you or on behalf of
Curtin College pursuant to Part VB of the Copyright Act 1968 (the Act).*

*The material in this communication may be subject to copyright under the Act.
Any further reproduction or communication of this material by you may be the
subject of copyright protection under the ACT.*

Do not remove this notice.

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Aims of this lecture

1. The Bootstrap Idea

(Moore et al Ch 16)

1.1 First Steps in Using the Bootstrap

1.2 How Accurate Is a Bootstrap Distribution?

- Use the bootstrap method to estimate properties of an estimator.

1.3 Bootstrap Confidence Intervals

R package: *boot*

2. Significance Testing Using Permutation Tests

(Moore et al Ch 16)

2.1 Procedure

2.2 Two sample test

2.3 Paired/Matched test

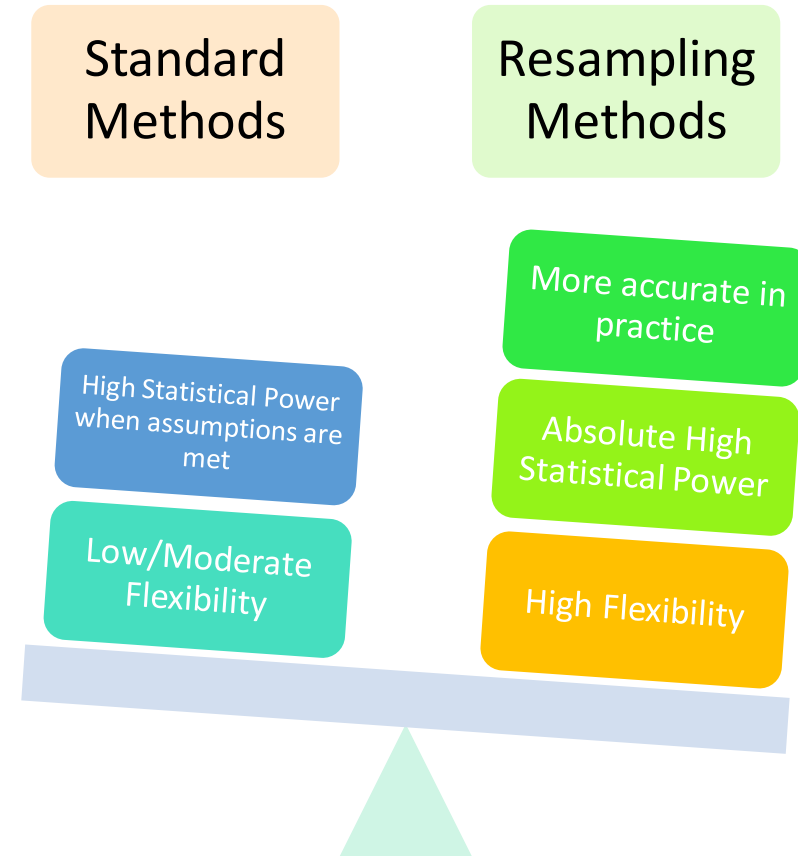
R package: *CarletonStats*

Resampling methods

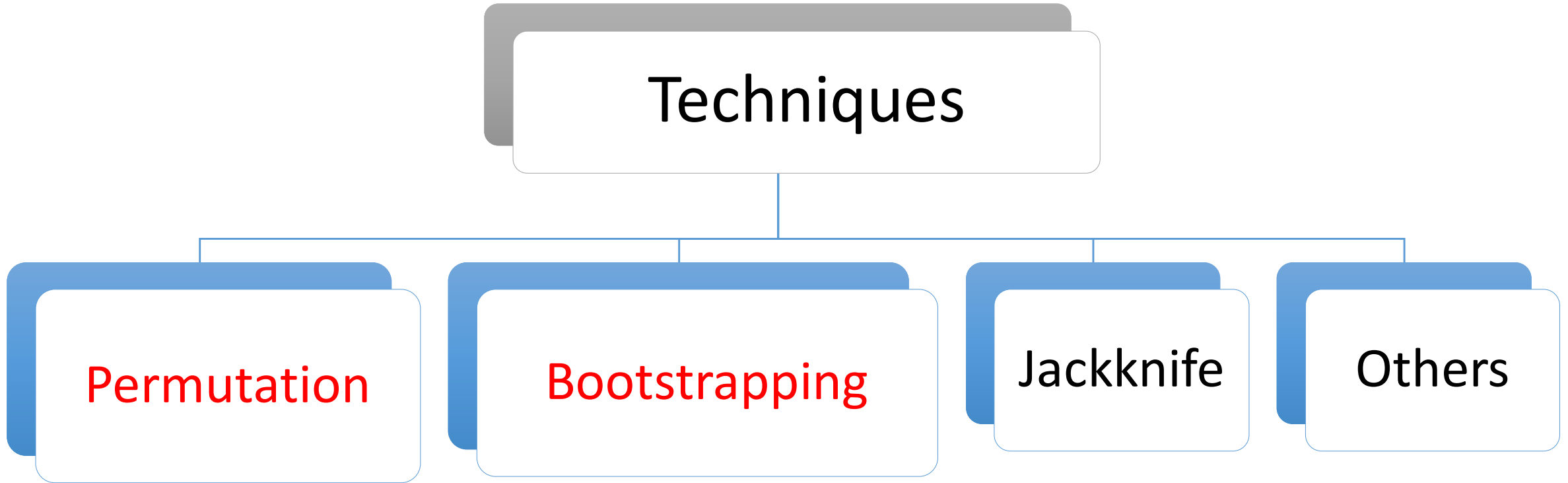
- Nonparametric methods we have seen so far allows us to **relax some of the assumptions of parametric methods**
- Another class, that has become more widely used because of the widespread availability of computing, includes **bootstrap methods** and **permutation tests**
- Still require the assumption of **random sampling** from a population
- Conceptually simple: “How likely is what we actually got relative to the sampling distribution?”
 - Difference is in how the sampling distribution is constructed

Why Resampling?

- In most cases, people accept assumptions for standard statistics “as if” they are satisfied.
- Some “awkward” but “interesting” statistics, that standard statistics fail to be applied to.
- Saves us from onerous formulas for different problems.
- More accurate in practice than standard methods.



Resampling Techniques



Bootstrap and Permutation

- The two methods of —bootstrap and permutation tests—apply the power of the computer to relax some of the conditions needed for traditional inference and to do inference in new settings.
- The big ideas of statistical inference remain the same. The fundamental reasoning is still based on asking, “What would happen if we applied this method many times?” Answers to this question are still given by confidence levels and P-values based on the sampling distributions of statistics.

Bootstrap and Permutation

- **Bootstrap intervals and permutation tests** are conceptually simple because they appeal directly to the basis of all inference: the sampling distribution that shows what would happen if we took very many samples under the same conditions.
- **The new methods do have limitations**, some of which we will illustrate. But their effectiveness and range of use are so great that they are now widely used in a variety of settings.
- There are **several packages of functions for resampling in R**. We will focus on the **boot** package, which offers the most capabilities.

Aim 1 Bootstrap Idea

- Statistical inference is based on **the sampling distributions of sample statistics**. A sampling distribution is based on many random samples from the population. **The bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample.**

- **The BOOTsTrAp IdeA**

The original sample is representative of the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population. The **bootstrap distribution** of a statistic, based on the resamples, represents the sampling distribution of the statistic.

Aim 1.1 Bootstrapping

- **Step 1: Resampling.** In place of many samples from the population, create many resamples by **repeatedly sampling with replacement** from this one random sample. **Each resample is the same size as the original random sample.**
- **Step 2: Bootstrap distribution.** The sampling distribution of a statistic describes the values taken by the statistic in all possible samples of the population of the same size. The **bootstrap distribution** of a statistic summarizes the values taken by the statistic in all possible resamples of the same size. The bootstrap distribution gives information (that is, **shape and spread**) about the sampling distribution.

Example 1: Average time looking at a Facebook profile

- In Example 12.17 (Moore et al 2017, page 670), we compared the amount of time a Facebook user spends reading different types of profiles.
- Here, let's focus on just the average time for the **fourth profile** (negative male).
- Figure 16.1(a) gives a histogram, and Figure 16.1(b) gives the Normal quantile plot of the 21 observations.
- The data are **skewed to the right**. Given the relatively small sample size, we have some **concerns about using the t procedures** for these data.

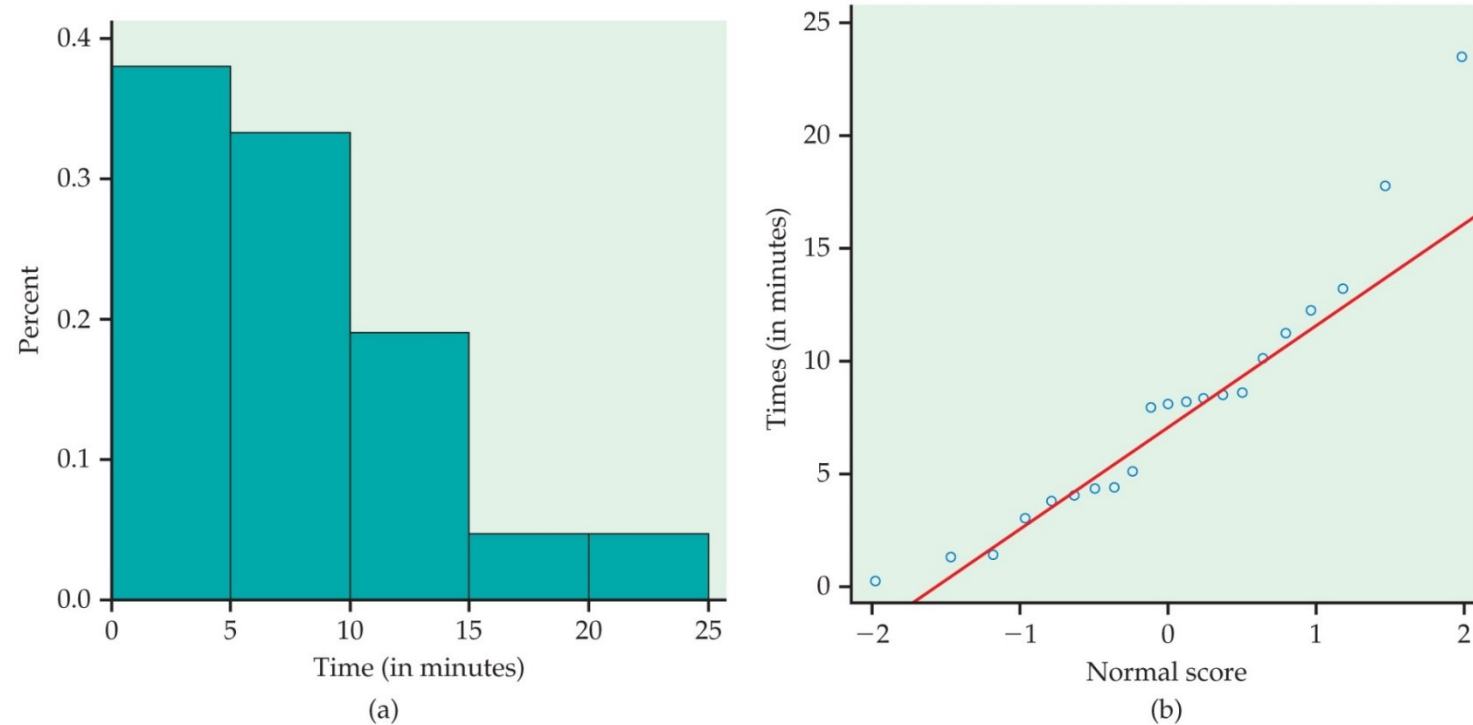


Figure 16.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Figure 16.1 (a) The distribution of times (minutes) looking at a negative male Facebook profile page. (b) Normal quantile plot of the times, Example 16.1. The distribution is right-skewed.

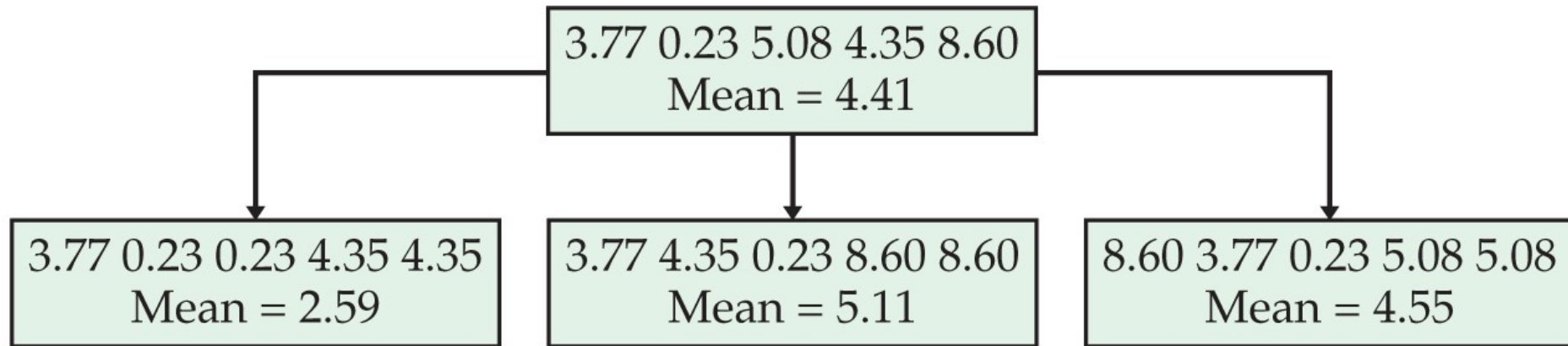
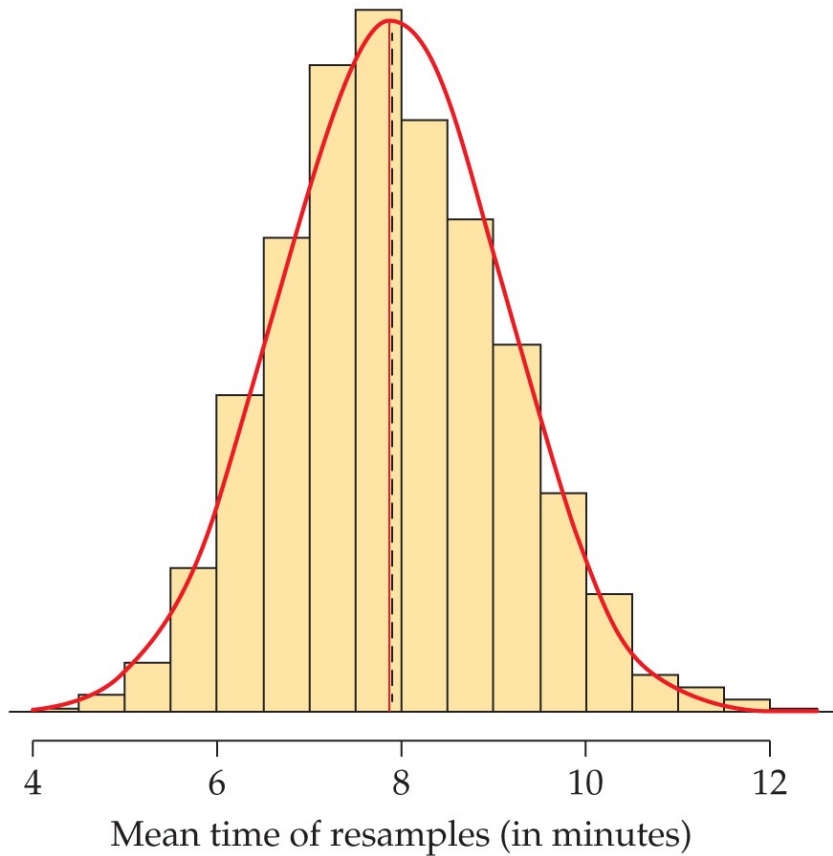


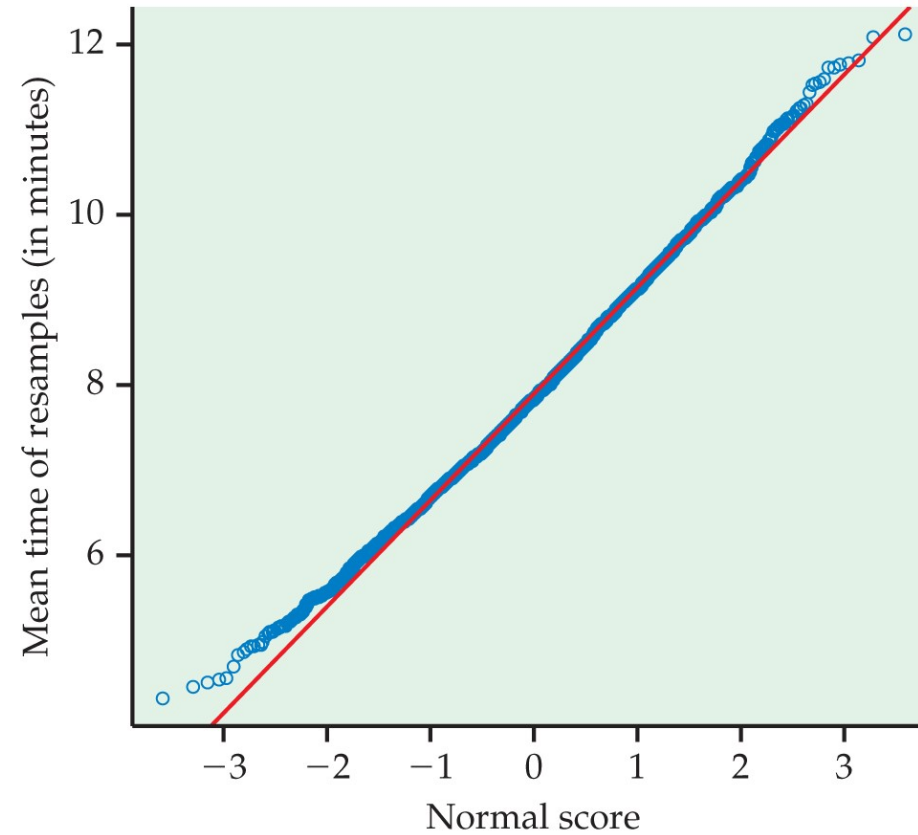
Figure 16.2

Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Figure 16.2 The resampling idea. The top box is a sample of size $n = 5$ from the Facebook profile viewing time data. The three lower boxes are three resamples from this original sample. **Some values from the original sample are repeated in the resamples because each resample is formed by sampling with replacement.** We calculate the statistic of interest, the sample mean in this example, for the original sample and each resample.



(a)



(b)

Figure 16.3

Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Figure 16.3 (a) The bootstrap distribution of 3000 resample means from the sample of Facebook profile viewing time data. The smooth curve is the Normal density function for the distribution which matches the mean and standard deviation of the distribution of the resample means. (b) The Normal quantile plot confirms that the bootstrap distribution is slightly skewed to the right but fits the Normal distribution quite well.

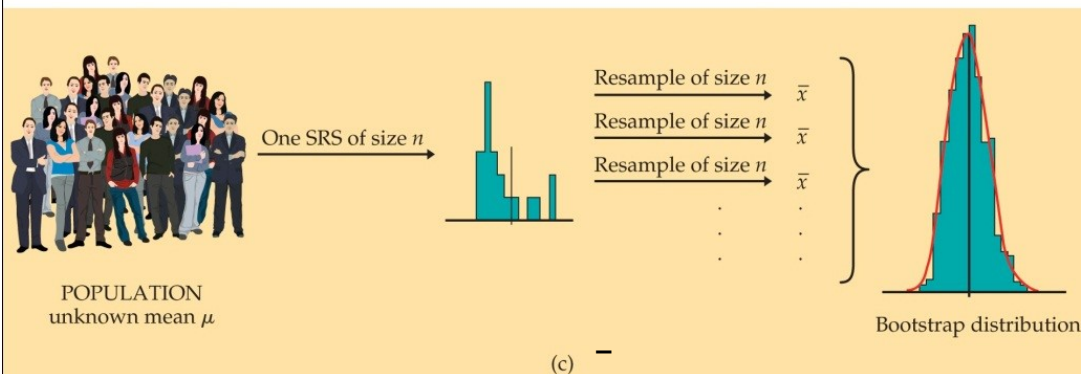
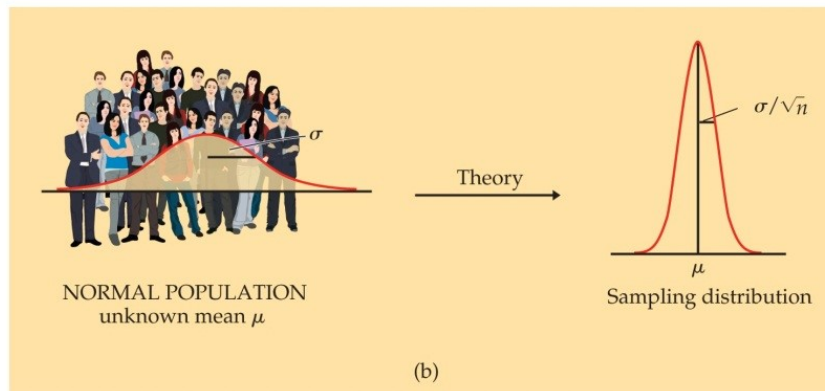
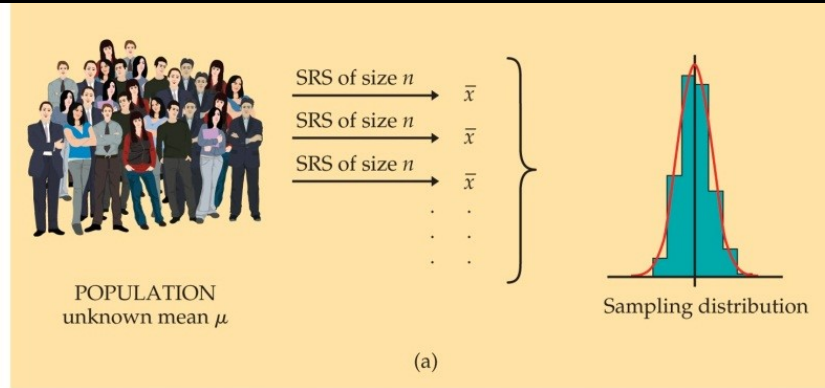


Figure 16.4 (a) The idea of the sampling distribution of the sample mean \bar{x} : take very many samples, collect the \bar{x} values from each, and look at the distribution of these values.

(b) The theory shortcut: if we know that the population values follow a Normal distribution, theory tells us that the sampling distribution of \bar{x} is also Normal.

(c) The bootstrap idea: when theory fails and we can afford only one sample, that sample stands in for the population, and the distribution of \bar{x} in many resamples stands in for the sampling distribution.

ORDINARY NONPARAMETRIC BOOTSTRAP

Bias=bootstrap mean – original mean

Call:

boot(data = Time, statistic = theta, R = 3000)

Bootstrap Statistics :

	original	bias	std. error
t1*	7.870476	0.02295317	1.216978

Figure 16.5

Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e,
© 2017 W. H. Freeman and Company

Figure 16.5 R output for the Facebook profile
viewing time bootstrap, Example 16.3.

An example of R code in Comp
Lab 4

```
``{r}
# Generate bootstrap distribution for
n=1000
set.seed(9854)
gosset.boot1000 <- boot(gosset.data,
theta, 1000)
gosset.boot0 <- gosset.boot1000$t0 ##
the observed value statistic (mean)
gosset.boot1000
gosset.boot10001 <- gosset.boot1000$t
## a matrix consists of the statistics from n
resampling
summary(gosset.boot10001)
head(gosset.boot10001)
```

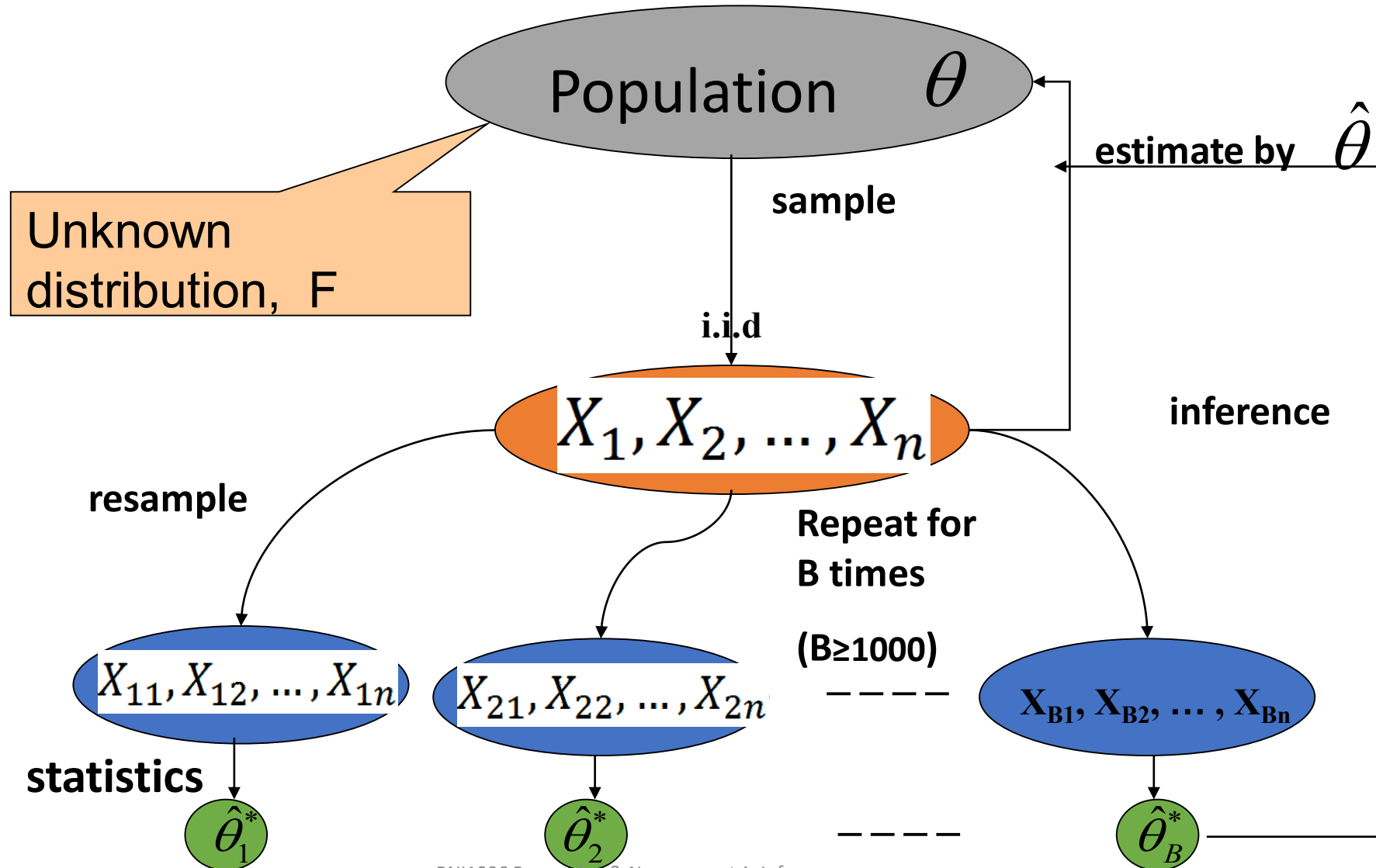
What is new?

- The bootstrap distribution of the resample means is used only to estimate how the sample mean of one actual sample of size n would vary because of random sampling.
- Using the same data for two purposes
 - to estimate a parameter and also to estimate the variability of the estimate—is perfectly legitimate. We do exactly this when we calculate \bar{x} to estimate μ and then calculate s/\sqrt{n} from the same data to estimate the variability of x .
- **What is new?**
 - First of all, **we don't rely on the formula to estimate the standard deviation of x** . Instead, we use the ordinary standard deviation of the many x -values from our many resamples. Suppose that we take B resamples and call the means of these resamples x^* to distinguish them from the mean \bar{x} of the original sample. We would then find the mean and standard deviation of the x^* 's in the usual way.

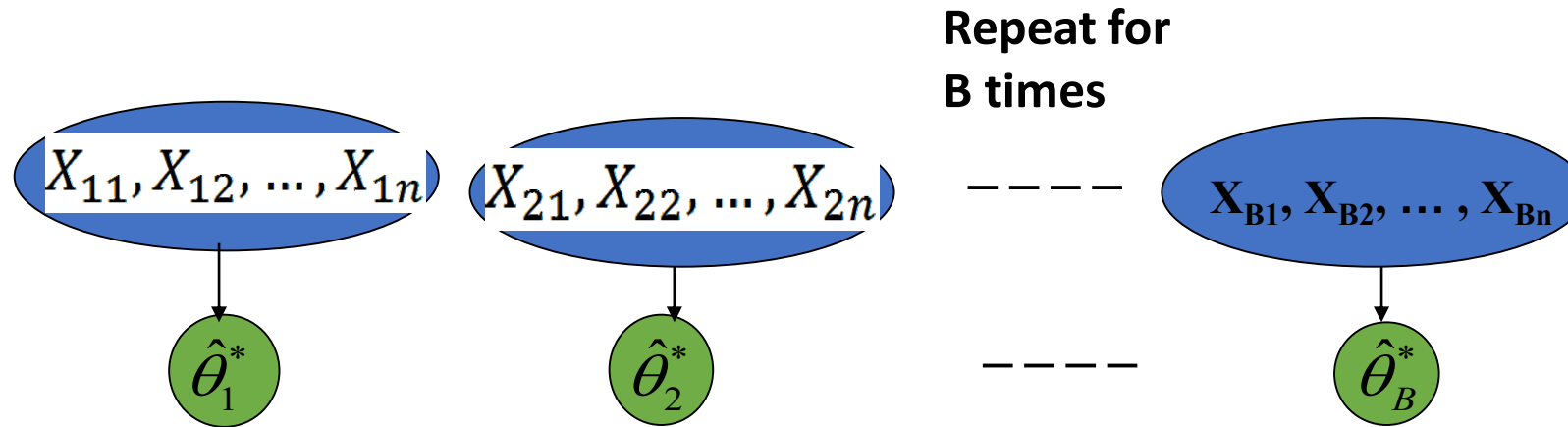
To make clear that these are the mean and standard deviation of the means of the B resamples rather than the mean \bar{x} and standard deviation s of the original sample, we use a distinct notation:

- Second, we don't appeal to the central limit theorem or other theory to tell us that a sampling distribution is roughly Normal. We look at the bootstrap distribution to see if it is roughly Normal (or not). In most cases, the bootstrap distribution has approximately the same shape and spread as the sampling distribution, but **it is centered at the original sample statistic value rather than the parameter value**.

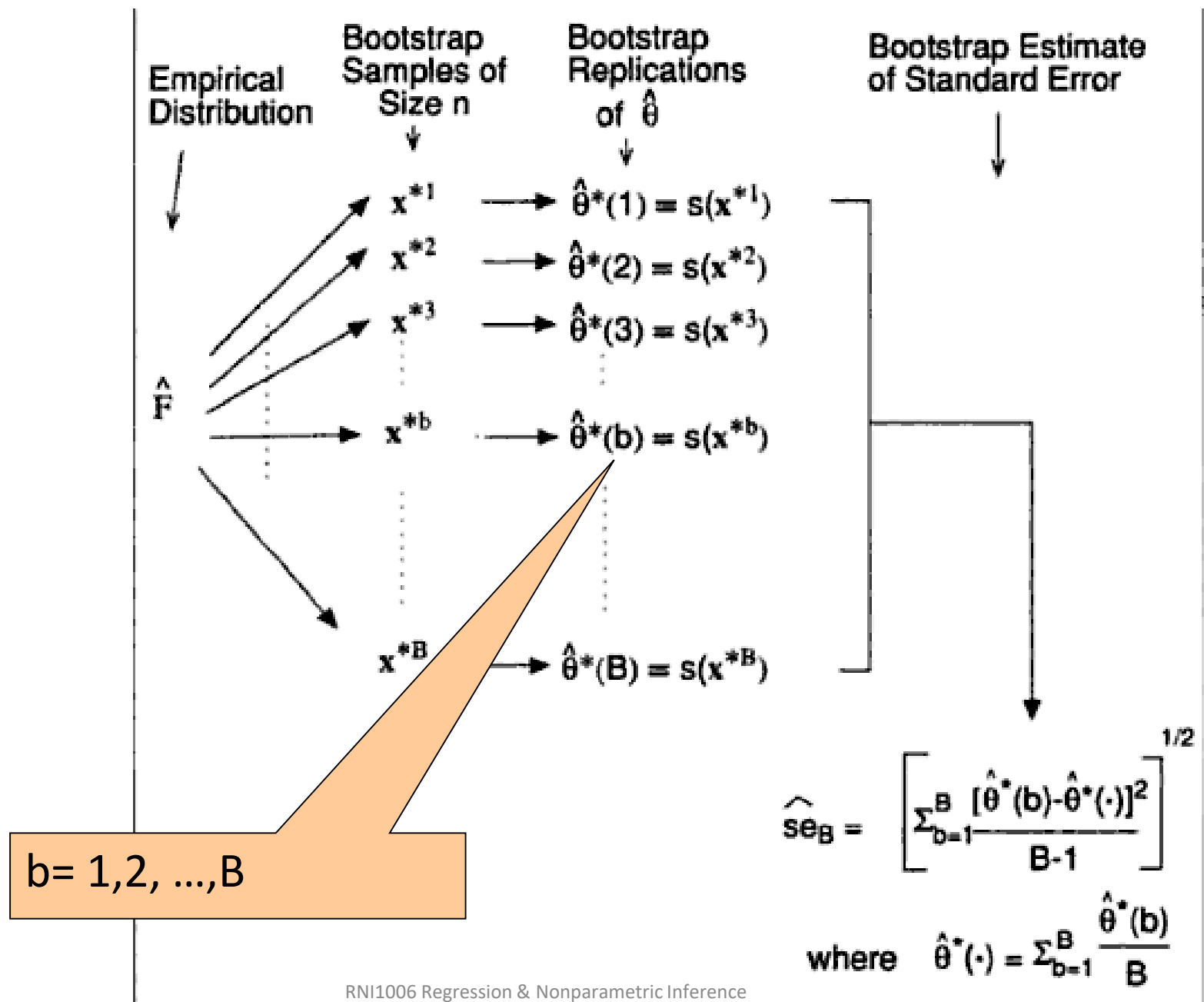
Aim 1.2 The nonparametric of Bootstrap - estimator



Step 3: statistics



STEP 3: Regard X_1, X_2, \dots, X_n as the new population and resample it B times with replacement, $X_{b1}, X_{b2}, \dots, X_{bn}$ where $b=1, 2, \dots, B$



Bootstrap Standar Error calculation

(Continued)

- Bootstrap replicates $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \dots, s(\mathbf{x}^{*B})$ are obtained by calculating the value of the statistic $s(\mathbf{x})$ on each bootstrap sample.
- The standard deviation of the values $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \dots, s(\mathbf{x}^{*B})$ is the estimate of the standard error of $s(\mathbf{x})$.

Aim 1.3 The percentile method for confidence interval

- 1) Construct \hat{F} , the empirical distribution function of the observed data. \hat{F} places probability $1/n$ on each observed data point X_1, X_2, \dots, X_n .
- 2) Draw a bootstrap sample $X_1^*, X_2^*, \dots, X_n^*$ of size n with replacement from \hat{F}

Then calculate

$$\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \dots, X_n^*).$$

- 3) Repeat Step (2) a large number of times, say 1000, and then rank the values $\hat{\theta}^*$.

The percentile method (Continued)

For a 95% confidence interval, after ranking the bootstrapped theta coefficients, simply take the 2.5 % as the lower confidence limit and the 97.5% as the upper confidence limit.

The percentile $(1-\alpha)$ 100% confidence interval for a population mean is:

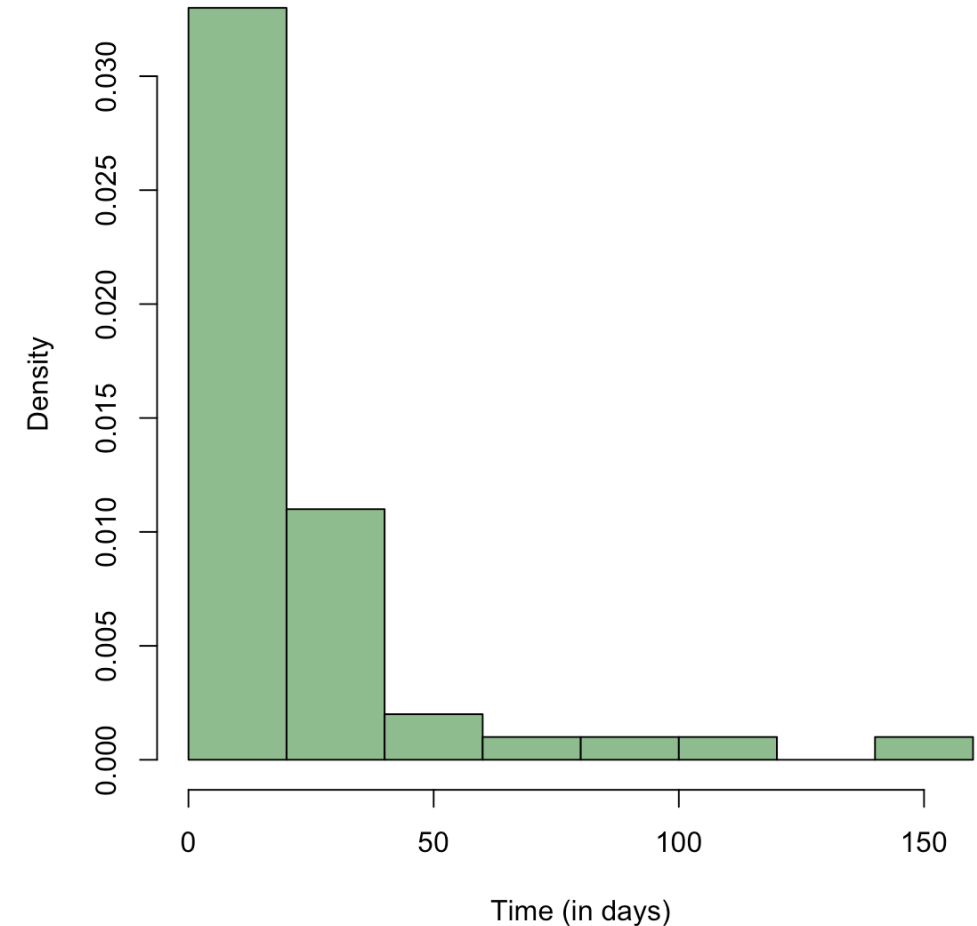
$$(\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)})$$

Bootstrap: key ideas

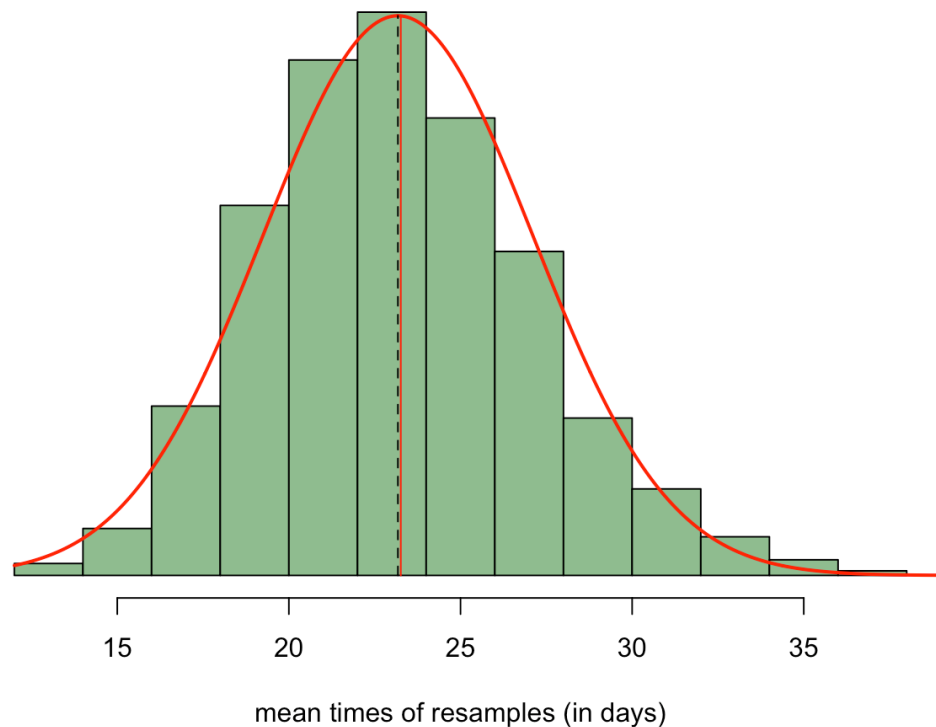
- Sampling distribution based on many random samples from population
 - Bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample
- 1. Create many resamples by repeatedly sampling **with replacement** from this **one sample**. Each resample is of the **same size** as the original random sample
- 2. For each resample, calculate the statistic of interest. This yields the **bootstrap distribution** of that statistic, which gives information about the **sampling distribution**.
- 3. This bootstrap distribution can be used to calculate, e.g., mean, SE, confidence intervals

Example 2 World Bank data: Bootstrap

- World Bank collects information about the time required to start business in a random sample of 50 countries
- Data are strongly right-skewed
- Median is 12 days, and mean is twice as large, about 23.3 days
- Clearly, using t -procedures may not be appropriate if we want to estimate some measure of the central tendency



Example 2 World Bank data: Bootstrap



- The bootstrap distribution using 3000 resamples is nearly Normal – but not quite
- It is centered close to the mean of the original sample ($\bar{x}_b = 23.18$ days versus $\bar{x} = 23.26$ days)
- The standard error of the bootstrap distribution is its standard deviation is $SE_b = 3.94$ versus $s/\sqrt{n} = 3.99$ for the original sample

Example 2 World Bank data: Bootstrap

Confidence intervals for mean time to set up business

- t - interval:
 - (15.25, 31.27)
- Bootstrap t -confidence interval using $\bar{x}_b \pm t_{\alpha/2, n-1} \times SE_b$
 - (15.27, 31.09) (we will not cover this one)
- **Bootstrap percentile confidence interval**
 - (16.22 31.66)

Bootstrap: remarks

- Bootstrap appears to create new data out of nothing – like Baron Munchausen pulling himself out of a hole using his bootstraps
 - The bootstrap distribution of the resample means is used only to estimate how the sample mean of one actual sample of size (50 in the previous example) would vary because of random sampling
 - Can be extended to different statistics – **medians, trimmed means, correlations, ratios of means, ...**
 - Can be extended to very complex models beyond comparing means – regression, errors-in-variables models, spatial models, ...
- This has been a very simple exposition - much more to be said!

Aim 2 What is Permutation Tests?

- Permutation tests are significance tests based on **permutation resamples** drawn at random from **the original data**. Permutation resamples are drawn **without replacement**.
- Also called **randomization tests, re-randomization tests, exact tests**.
- Introduced by R.A. Fisher and E.J.G. Pitman in the 1930s.



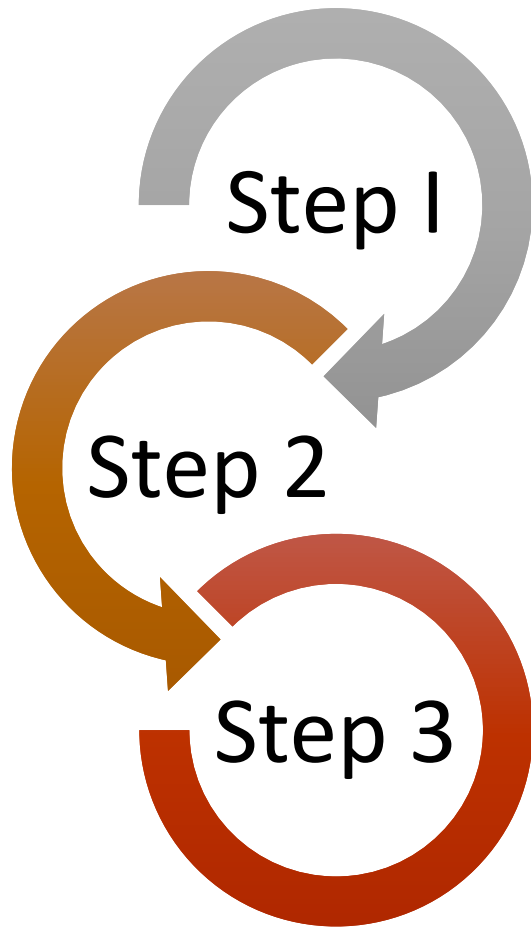
When Can We Use Permutation Tests?

- Only when we can see how to **resample** in a way that is **consistent** with the **study design** and with the **null hypothesis**.
- If we cannot do a permutation test, we can often calculate a **bootstrap confidence interval** instead.

Permutation tests

- Fisher's basic idea:
 - Let's say that we've measured the heights of 100 **Englishmen** and 100 **Frenchmen**
 - If there is no difference between their average heights (**the null hypothesis**), then they can be “shuffled without regard to nationality, and divided at random into two new groups of a hundred each. This division could be done in an enormous number of ways [F]or each of these ways the difference between the two average statures is calculated.”
 - This distribution of “average statures” is the reference or permutation distribution against which we compare **the difference we actually observed**
 - If a difference as large as or larger occurs rarely in the permutation distribution, we would reject the null hypothesis

Aim 2.1 Procedure of Permutation Tests



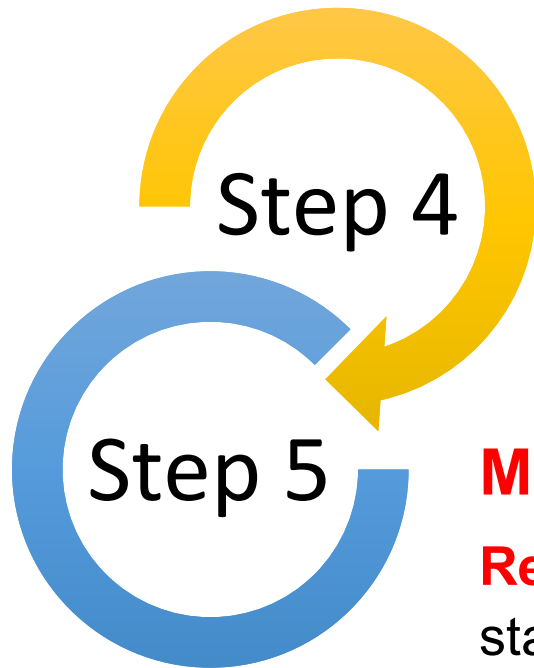
Analyze the problem.

- What is the hypothesis and alternative?
- What distribution is the data drawn from?
- What losses are associated with bad decisions?

Choose a test statistic which will distinguish the hypothesis from the alternative.

Compute the test statistic for the original data of the observations.

Procedure of Permutation Tests



Step 4

Rearrange the Observations

Compute the test statistic for **all possible permutations** (rearrangements) of the data of the observations

Step 5

Make a decision

Reject the Hypothesis: if the value of the test statistic for the original data is an **extreme value** in the permutation distribution of the statistic.

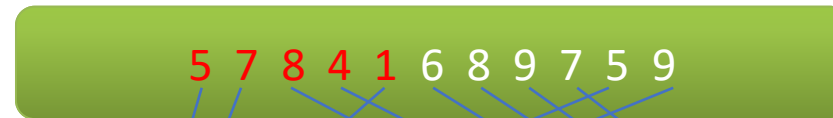
Otherwise, accept the hypothesis and reject the alternative.

Permutation Resampling Process

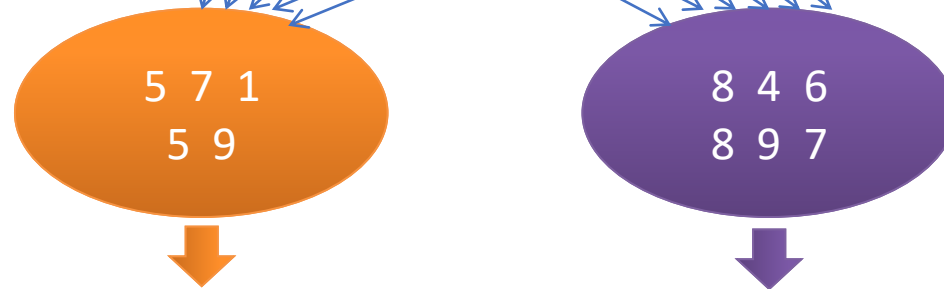
Collect Data from Control & Treatment Groups



Merge Samples to form a pseudo population



Sample without replacement from pseudo population to simulate Control and Treatment Groups



Compute **target statistic** for each resample

Median (5 7 1 5 9)

Median (8 4 6 8 9 7)

Compute “difference statistic”, save result in table and repeat resampling process 1000+ iterations

Aim 2.2 Example 3 t-test vs Permutation test

Here is a simple sample:

- data

X			Y		
A	B	C	D	E	F
121	118	110	34	12	22
$\bar{x}_n = 116.33$			$\bar{y}_n = 22.67$		

- random samples $X = \{x_1, \dots, x_{n_x}\}$ and $Y = \{y_1, \dots, y_{n_y}\}$
- test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{(n_x - 1) + (n_y - 1)}}} \cdot \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}$$

(\bar{X} , \bar{Y} and S_x^2 , S_y^2 are sample means and sample variances, n_x , n_y sample sizes)

Using the original T-test to Find P-value

- under H_0 holds $T \sim t_{n_x+n_y-2}$ (t-distribution with $n_x + n_y - 2$ d.f.)

- compute the p -value for the observed value t of test statistic T

$$p = 1 - P(|T| \leq |t| | H_0) = 2[1 - P(T \leq |t| | H_0)] = 2[1 - F_{t, n_x+n_y-2}(|t|)]$$

- Decision rule: reject H_0 if $p\text{-value} \leq \alpha$

**T-test statistic: T=13.0875,
Two-sided P-value=0.0002**

Example 3: Permutation Resampling Process

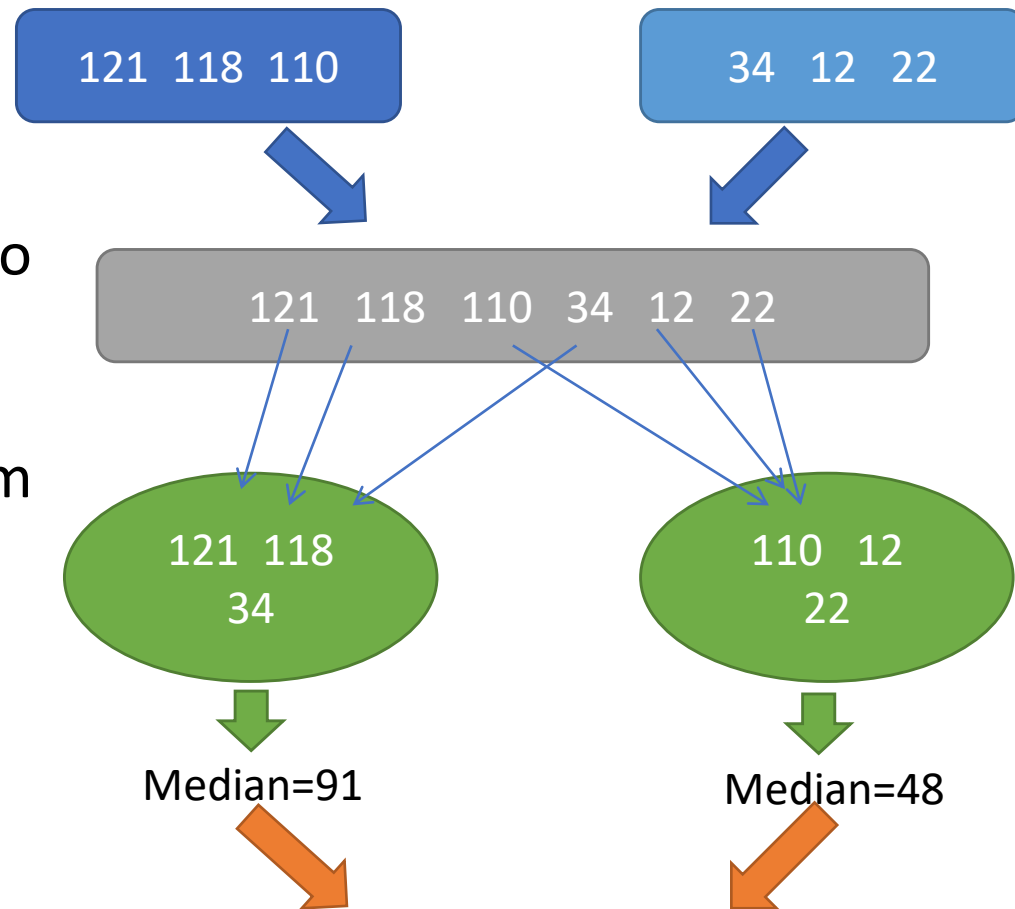
Collect Data from Control & Treatment Groups

Merge Samples to form a pseudo population

Sample without replacement from pseudo population to simulate Control and Treatment Groups

Compute **target statistic** for each resample

Compute “difference statistic”, save result in table and **repeat resampling process 1000+ iterations**



Example 3: After one permutation

- after one permutation:

X			Y		
A	B	D	C	E	F
121	118	34	110	12	22
$\bar{x}_n = 91$			$\bar{y}_n = 48$		

- test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{(n_x-1) + (n_y-1)}}} \cdot \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}$$

T-test statistic: T=1.019

Formula in Permutation needed

- Note: The number of permutations for comparing two groups of size m and $n - m$ is

$$C_m^n = \binom{n}{m} = \frac{n!}{m! \cdot (n - m)!}$$

- how many permutations exist?

$$C_3^6 = \binom{6}{3} = \frac{6!}{3! \cdot 3!} = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = 20$$

Example 3: All 20 Permutation data

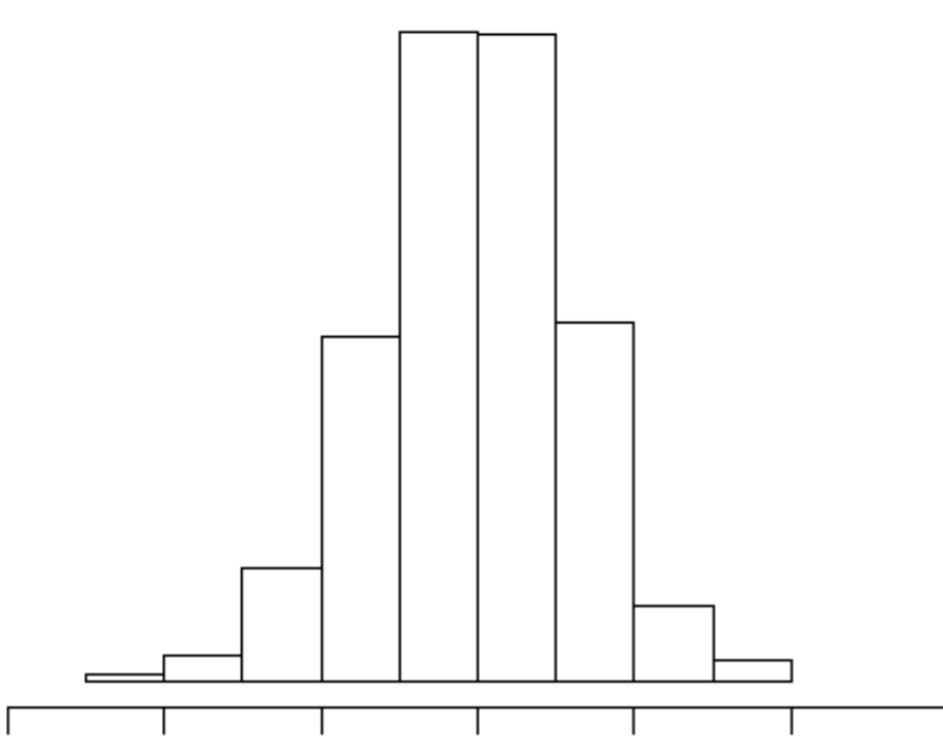
permutation	X	Y	\bar{x}_n	\bar{y}_n	$\bar{x}_n - \bar{y}_n$	t
1	ABC	DEF	116.33	22.67	93.67	13.087
2	ABD	CEF	91.00	48.00	43.00	1.019
3	ABE	CDF	87.00	52.00	35.00	0.795
4	ABF	CDE	83.67	55.33	28.33	0.627
5	ACD	BEF	88.33	50.67	37.67	0.866
6	ACE	BDF	84.33	54.67	29.67	0.659
7	ACF	BDE	81.00	58.00	23.00	0.500
8	ADE	BCF	59.00	80.00	-21.00	-0.455
9	ADF	BCE	55.67	83.33	-27.67	-0.611
10	AEF	BCD	51.67	87.33	-35.67	-0.813
11	BCD	AEF	87.33	51.67	35.67	0.813
12	BCE	ADF	83.33	55.67	27.67	0.611
13	BCF	ADE	80.00	59.00	21.00	0.455
14	BDE	ACF	58.00	81.00	-23.00	-0.500
15	BDF	ACE	54.67	84.33	-29.67	-0.659
16	BEF	ACD	50.67	88.33	-37.67	-0.866
17	CDE	ABF	55.33	83.67	-28.33	-0.627
18	CDF	ABE	52.00	87.00	-35.00	-0.795
19	CEF	ABD	48.00	91.00	-43.00	-1.019
20	DEF	ABC	22.67	116.33	-93.67	-13.087

How is the conclusion

- Test decision. The absolute value of the test statistic $t \geq |t| = 13.0875$ we obtained for the original labeling.
- We obtain the exact p value. $p = 2/20 = 0.1$.
- Note: If both groups have equal size, only half of permutations is really needed (symmetry)

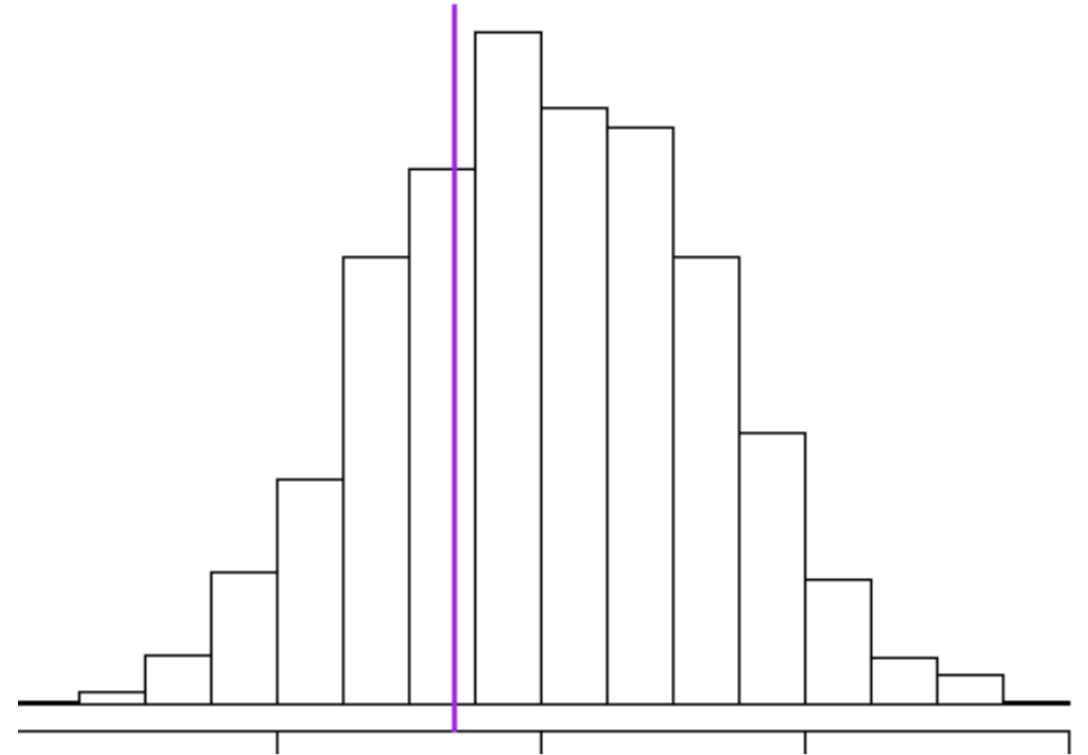
Example: null is false

Histogram of many.falsenull



Example: null is true

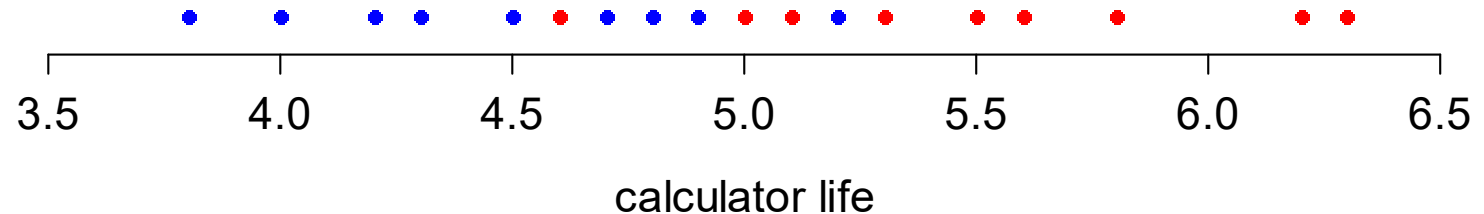
Histogram of many.truennull



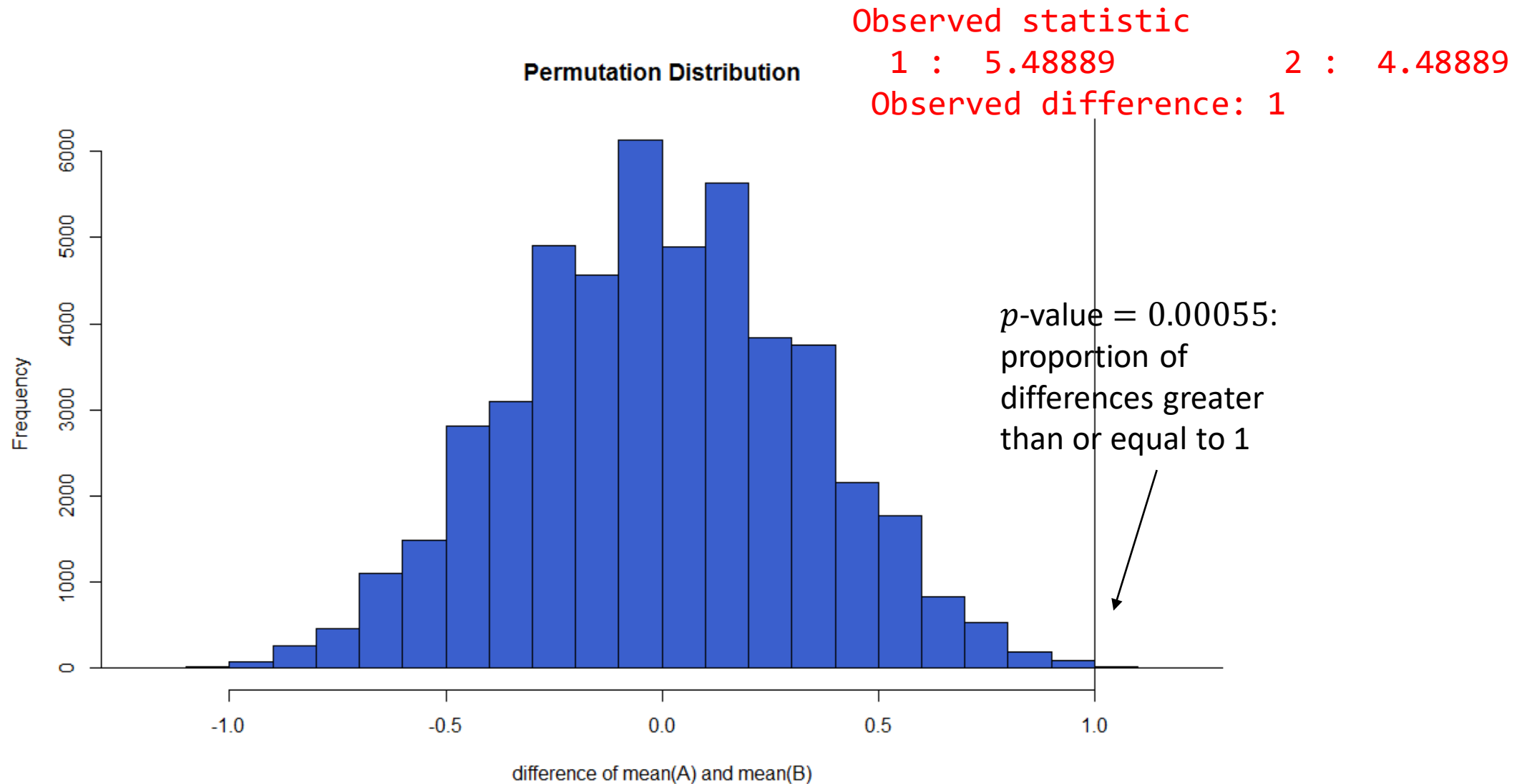
Example 4 (WMMY9e, Ex. 16.17) – revisited

- The following data represent the number of hours that two different brands of calculators operate without requiring a recharge. Does brand A operate **longer than** brand B?

A	B
5.5	3.8
5.6	4.8
6.3	4.3
4.6	4.2
5.3	4.0
5.0	4.9
6.2	4.7
5.8	5.2
5.1	4.5



Example 4 (WMMY9e, Ex. 16.17) – revisited



Permutation tests – a summary of general procedure

- Calculate the **statistic** for the original data (e.g., difference in means)
- Choose **permutation resamples** from the data without replacement in a way that is **consistent with the null hypothesis of the test and with the study design**.
- Construct the permutation distribution of the statistic from its values in a large number of resamples
- Find the p -value by **locating the original statistic on the permutation distribution**
- **How would a matched pairs permutation test be conducted?**

Example 4 (WMMY9e, Ex. 16.17) – revisited

```
## install package "CarletonStats"
```

```
library(CarletonStats)
```

```
brandA<-c(5.5, 5.6, 6.3, 4.6, 5.3, 5.0, 6.2, 5.8, 5.1)
```

```
brandB<-c(3.8, 4.8, 4.3, 4.2, 4.0, 4.9, 4.7, 5.2, 4.5)
```

```
obs <- c(brandA, brandB)
```

```
groups <- c(rep(1,9), rep(2,9))
```

```
obs
```

```
groups
```

```
set.seed(123)
```

```
Perm.Test(obs, groups, alternative="greater")
```

```
** Permutation test **
```

```
Permutation test with alternative: greater
```

```
Observed statistic
```

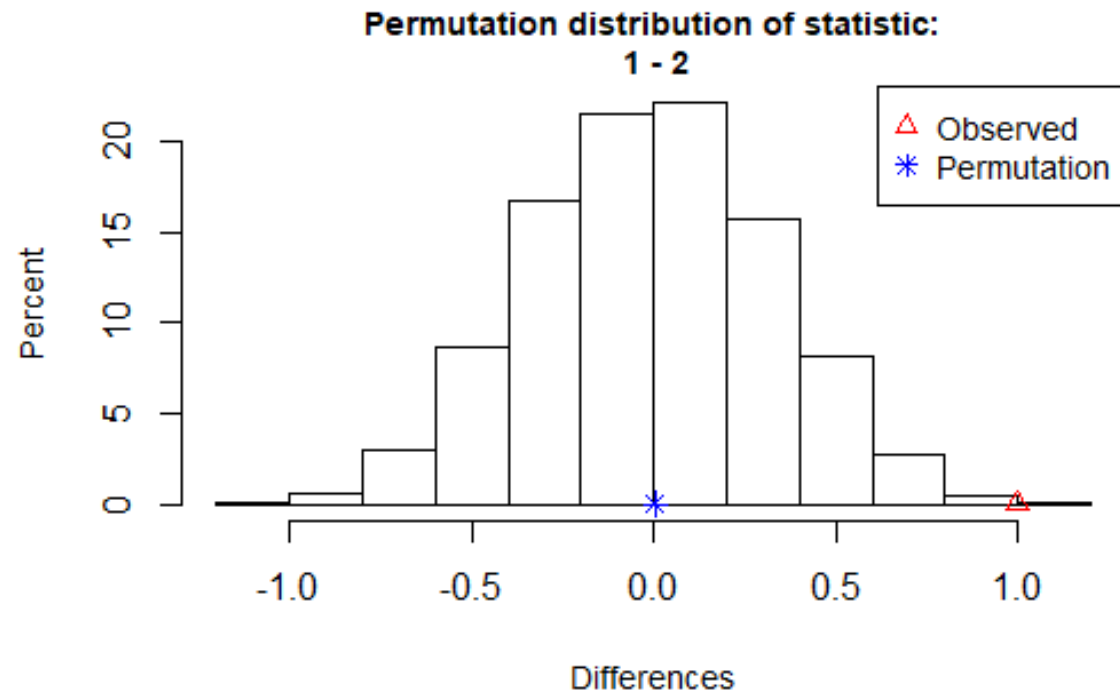
```
1 : 5.48889      2 : 4.48889
```

```
Observed difference: 1
```

```
Mean of permutation distribution: 0.00303
```

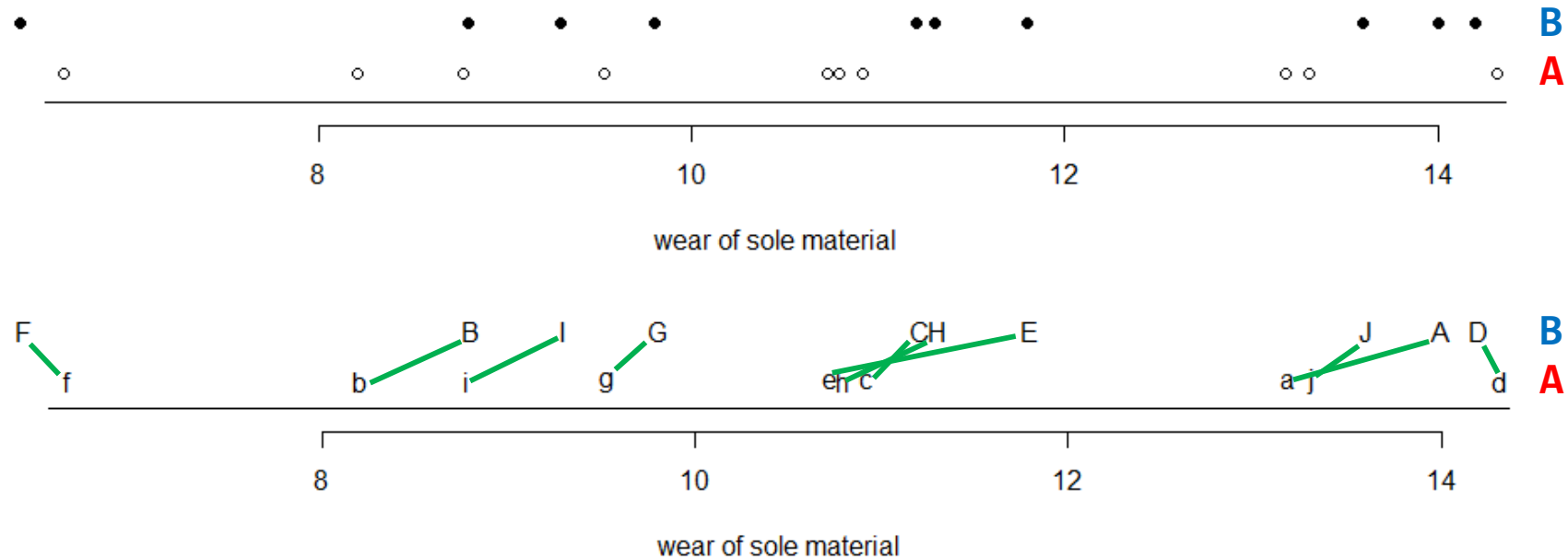
```
Standard error of permutation distribution: 0.33252
```

```
P-value: 5e-04
```



Aim 2.3 Example 5- matched pairs

We want to compare the amount of wear of two different materials of shoe sole material. The experiment is designed so that we randomly assign to the left or right shoe of an individual, either material A or B.



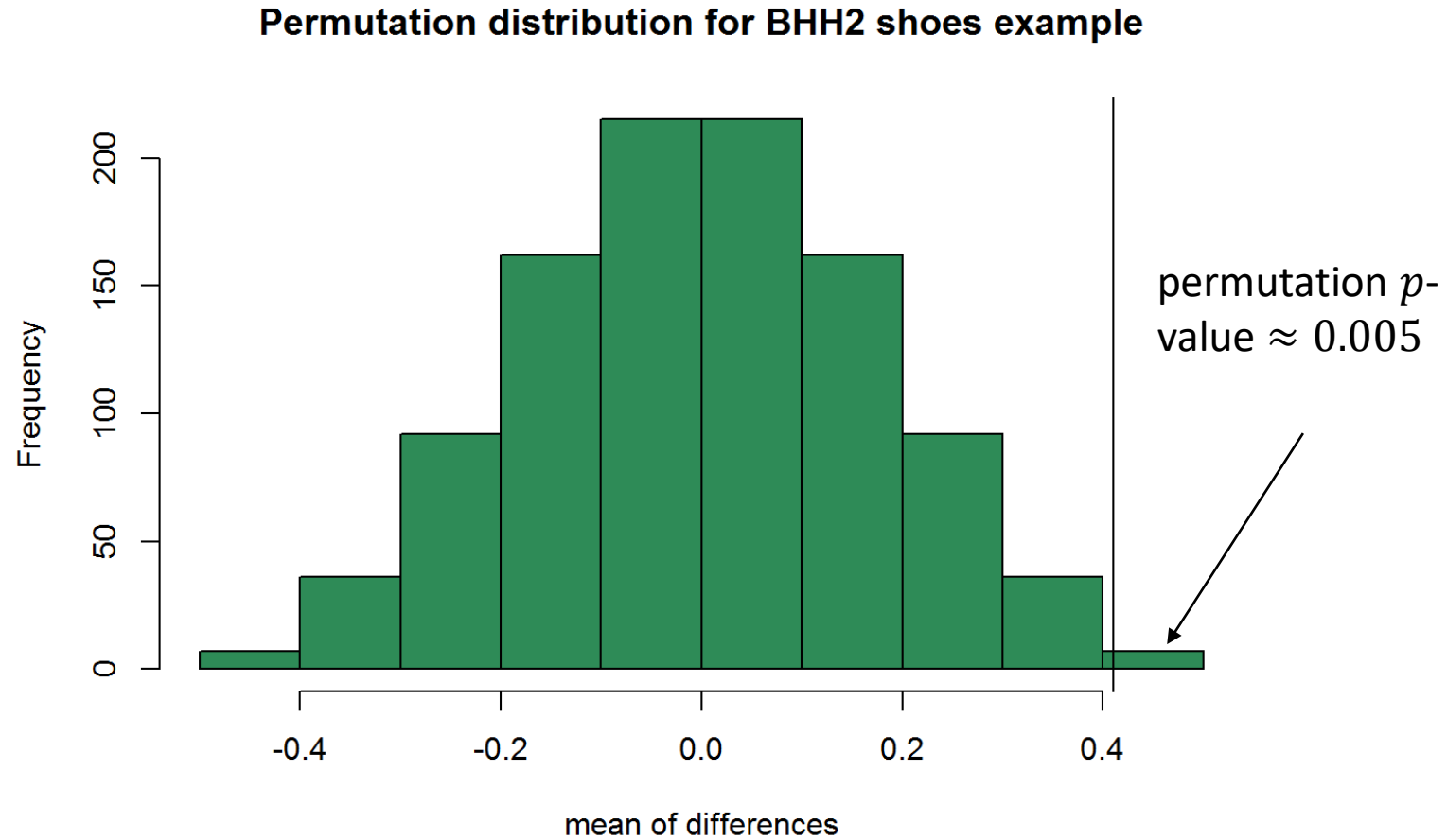
Example 5- matched pairs

A	B	Z=B-A
13.2	14	0.8
8.2	8.8	0.6
10.9	11.2	0.3
14.3	14.2	-0.1
10.7	11.8	1.1
6.6	6.4	-0.2
9.5	9.8	0.3
10.8	11.3	0.5
8.8	9.3	0.5
13.3	13.6	0.3

$$\bar{Z} = 0.41$$

- Test the null hypothesis of the same wear for both materials against the alternative that the wear of material B is higher
- How would we construct a permutation distribution for these data?
- (Parametric t -test gives a p -value of 0.004)

Example 5- matched pairs



Advantages



Exist for **any test statistic, regardless** of whether or not its **distribution is known**



Free to choose the statistic which best discriminates between hypothesis and alternative and which minimizes losses

Can be used for:

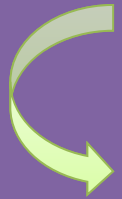
- Analyzing **unbalanced designs**;
- Combining **dependent tests** on mixtures of **categorical, ordinal, and metric data**.

Limitations



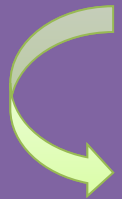
An Important Assumption

The **observations are exchangeable**
under the null hypothesis



Consequence

Tests of difference in location (like a permutation t-test) **require equal variance**



In this respect

The **permutation t-test** shares the **same weakness** as the **classical Student's t-test**.

Comparison and Contrast

	Application	Sampling Procedure
Permutation	Hypothesis testing	Samples drawn at random, without replacement.
Bootstrap	Standard deviation, confidence interval, hypothesis testing, bias	Samples drawn at random, with replacement