MONASH University

ETC2420

Statistical methods in Insurance

Week 1.
Review of hypothesis testing

29 July 2016

References

Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2014. **Introductory Statistics with Randomization and Simulation**. 1 edition. CreateSpace Independent Publishing Platform.

Hesterberg, Tim C. 2015. **"What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum."** The American Statistician 69 (4): 371–86.

Gender discrimination question

Are females unfairly discriminated against in promotion decisions compared to males?

- *48 male bank supervisors* were asked to assume the role of the personnel director of a bank
- They were given a personnel file to *judge whether the person should be promoted* to a branch manager position.
- The files given to the participants were identical, except that *half of them indicated the candidate was male and the other half indicated the candidate was female*
- These files were *randomly assigned* to the subjects.
- For each supervisor we recorded the *gender* associated with the assigned file and the *promotion decision*.

Gender discrimination data

		decision		
		promoted	not promoted	Total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

		decision		
		promoted	not promoted	Total
gender	male	0.875	0.125	1.000
	female	0.583	0.417	1.000
	Total	0.730	0.270	1.000

- 0.583 female promoted
- 0.875 male promoted
- Female are less likely to be promoted?
- *Statistical* evidence?

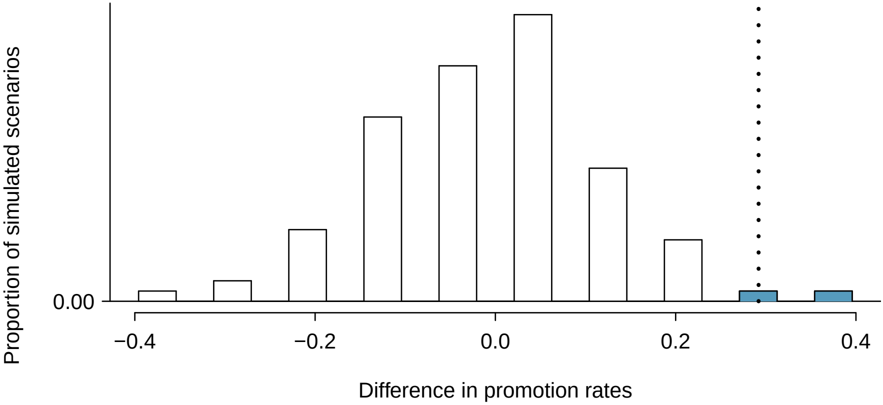
Permutation procedure

- Understand what sort of differences we might see *under the "no effect" scenario*, i.e. gender has no effect on promotion
- *If the bankers's decisions were independent of gender*, differences in promotion rates would be based only on random fluctuation

Permutation procedure

- We can simulate the “gender has no effect” scenario as follows:
  - Label 13 cards “not promoted”, and 35 other cards “promoted”
  - Shuffle these card throughly and divide them into two stacks of 24 people, representing the male and female groups
  - Tabulate the results and determine the fraction of male and female who were promoted
- If we do this multiple times, we can compute the *distribution of differences* from *chance alone*
- This procedure is formally called a *permutation test*

Permutation procedure



- Area in blue: 2%

CPR question

Blood thinners have an impact on survival, *either positive or negative*, but not zero

- Here we consider an experiment with patients who underwent Cardiopulmonary resuscitation (CPR) for a heart attack and were subsequently admitted to a hospital
- Each patient was randomly assigned to either receive a blood thinner (treatment group) or not receive a blood thinner (control group).
- The outcome variable of interest was whether the patient survived for at least 24 hours.

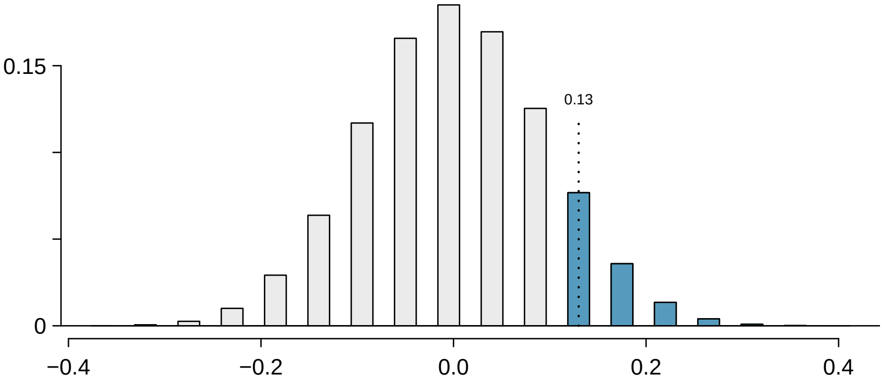
CPR data

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

	Survived	Died	Total
Control	0.22	0.78	1.00
Treatment	0.35	0.65	1.00
Total	0.27	0.73	1.00

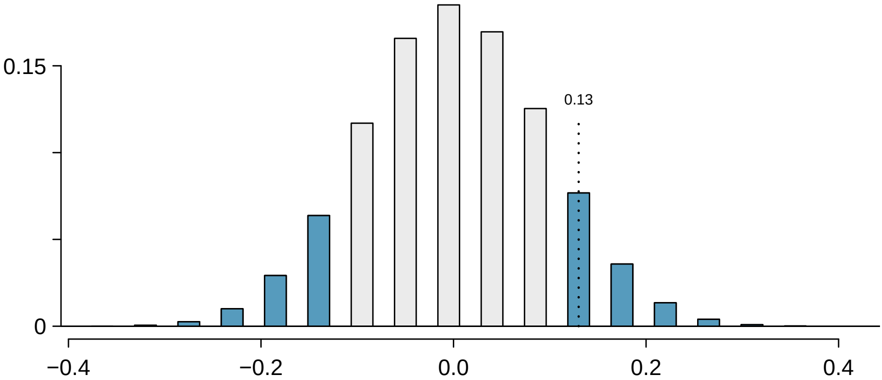
- 0.22 survival rate in control group
- 0.35 survival rate in treatment group
- Blood thinners have an impact?
- *Statistical* evidence?

Permutation procedure



- Area in blue: 12%

Example 3



- Area in blue: 24%

TV commercials question

They are more commercials in the “basic” TV channels than in the “extended” channels you pay extra for.

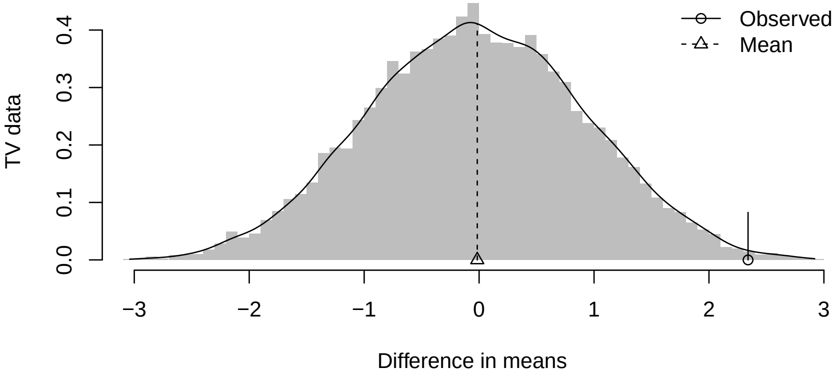
TV commercials data

Minutes of commercials per half-hour of TV:

Basic	6.9	10.0	10.1	8.5	7.6
	8.2	10.3	11.0	8.5	10.6
Extended	3.4	7.8	9.4	4.7	5.4
	7.6	5.0	8.0	7.8	9.6

- Average for Basic: 9.21
- Average for Extended: 6.87
- 9.21 - 6.87 = 2.34

Permutation procedure



- Area ≥ 2.34: 0.005

Permutation test vs Bootstrapping

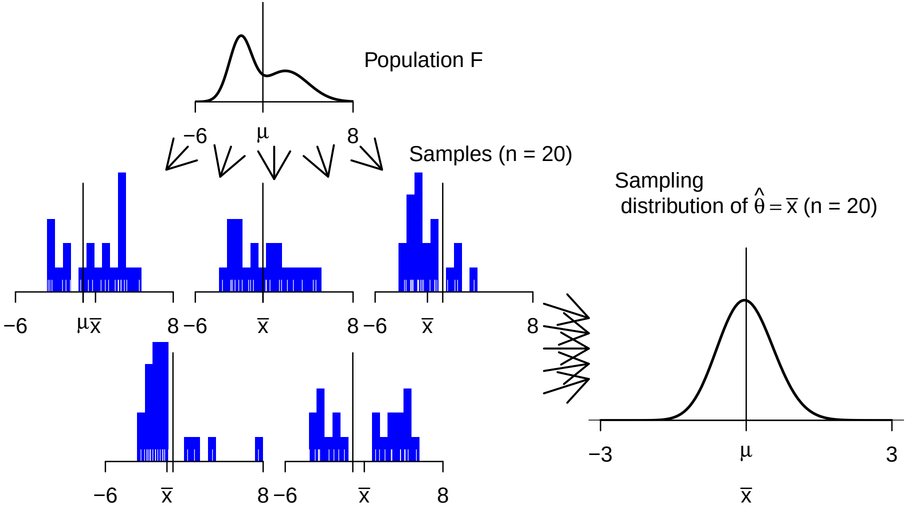
- See whether there is a significant difference

  $\rightarrow$ **Permutation test**

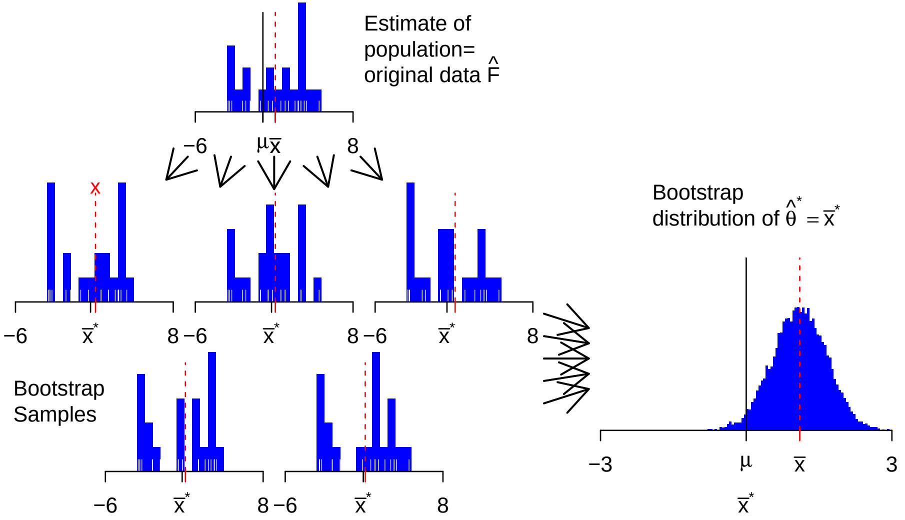
- Quantifying the random variability in the estimates and in the estimated difference

  $\rightarrow$ **Bootstrapping**

Bootstrapping: Ideal world



Bootstrapping: Bootstrap world



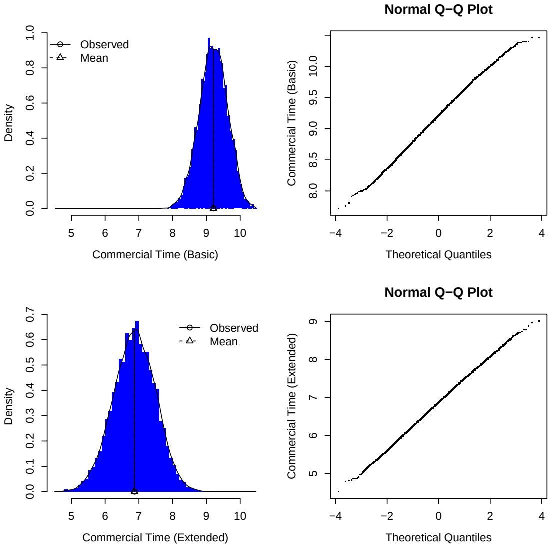
Bootstrapping

Minutes of commercials per half-hour of TV:

Basic	6.9	10.0	10.1	8.5	7.6
	8.2	10.3	11.0	8.5	10.6
Extended	3.4	7.8	9.4	4.7	5.4
	7.6	5.0	8.0	7.8	9.6

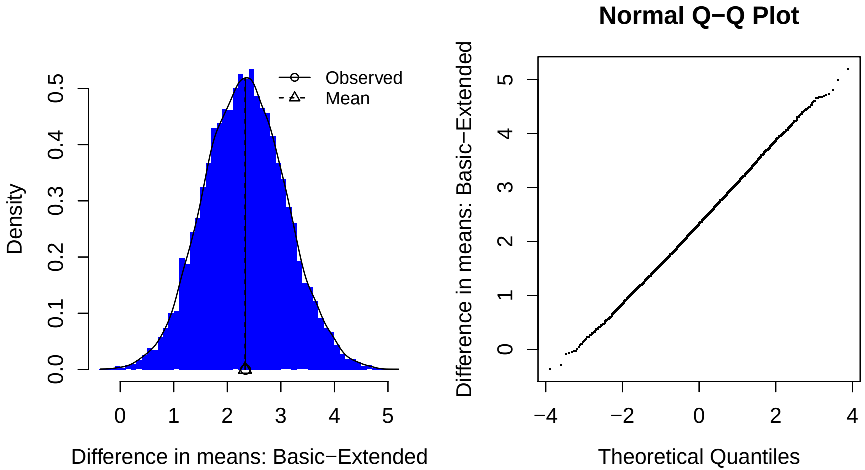
- **Bootstrap sample**: draw *n* observations with replacement from the original data
- **Bootstrap estimate**: estimate computed on a bootstrap sample
- **Bootstrap distribution**: *sampling distribution* for the bootstrap estimate

Bootstrap distribution



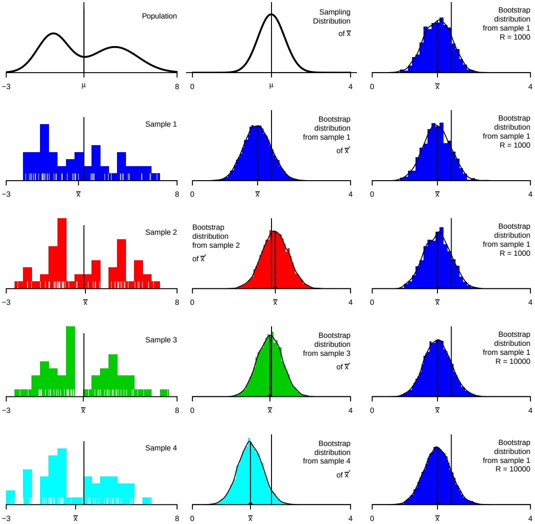
Center? spread? shape?

Bootstrap distribution



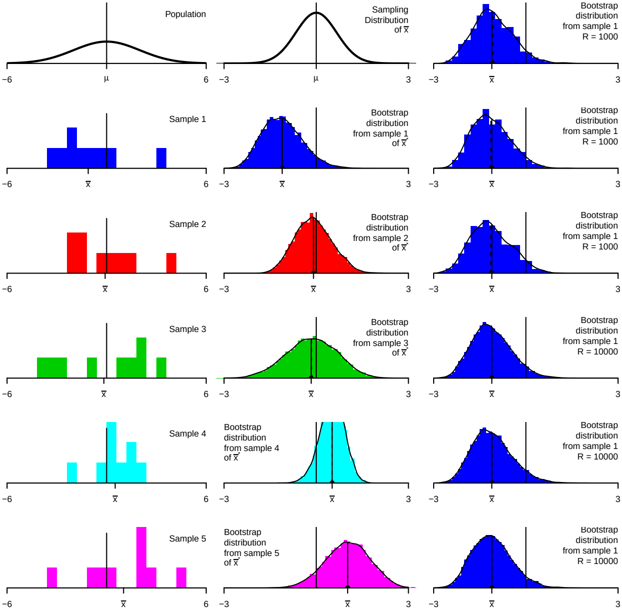
Difference with permutation distribution?

Sample mean: $n = 50$



- Two types of random variation

Sample mean: $n = 9$



Bootstrapping for confidence intervals

- We can use bootstrapping to **assess the uncertainty** surrounding **any sample statistic**
- Assumption: the variability of the bootstrap statistics is “similar” to the variability of the sample statistics
- The **standard error** of a statistic is the standard deviation of the sample statistic, which can be estimated from a bootstrap distribution
- **Confidence intervals** can be created using the **standard error** or the **percentiles** of a bootstrap distribution

Hypothesis testing

"A **hypothesis test** is a statistical technique used to evaluate **competing claims** using **data**. Often, the **null hypothesis** takes a stance of *no difference* or *no effect*. If the null hypothesis and the data notably disagree, then we will reject the null hypothesis in favor of the **alternative hypothesis**."

Hypothesis testing components

- Null hypothesis
- Alternative hypothesis
- Assumptions
- Significance level
- Sampling distribution under the null hypothesis
- Test statistic
- P-value
- Decision

Null and alternative hypothesis

- The **null hypothesis** (H_0) is often a statement that no effect or no difference is present. For example, the value of a population parameter (e.g. mean) is equal to some claimed value
- The **alternative hypothesis** (H_a) represents an alternative claim under consideration and is often represented by a range of possible values for the value of interest.
- Failing to find strong evidence for the alternative hypothesis is not equivalent to providing evidence that the null hypothesis is true.

Null and alternative hypothesis

- The hypotheses are stated in terms of **population parameters**
- We observe a **sample** from the population

Population parameter	**Sample statistic**
Mean μ	Sample mean $\bar{X}$
Proportion π	Sample Proportion p
Difference between means $\mu_2 - \mu_1$	Difference between sample means $\bar{X}_2 - \bar{X}_1$
Difference in proportions $\pi_2 - \pi_1$	Difference in sample proportions $p_2 - p_1$
Correlation ρ	Sample correlation r
...	...

Null and alternative hypothesis

- Directional tests (one-sided)
  - one sample:
    $H_0 : \mu = \mu_0$ and $H_a : \mu > (or <)\ \mu_0$
  - two samples:
    $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 > (or <)\ \mu_2$
  - Keywords: reduce, improve, higher, lower, greater than, etc
- Non-directional tests (two-sided)
  - one sample:
    $H_0 : \mu = \mu_0$ and $H_a : \mu \neq \mu_0$
  - two samples:
    $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$
  - Keywords: change, different from, etc

Significance level and decision errors

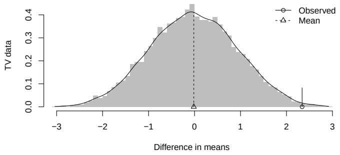
		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error (α)
	H_A true	Type 2 Error (β)	okay

- α: Probability of Type I Error
  The defendant is innocent (H_0 true) but wrongly convicted
- β: Probability of Type II Error
  The court failed to reject H_0 (i.e. failed to convict the person) when he was in fact guilty (H_a true).
- $1 - \beta$: Power of the test

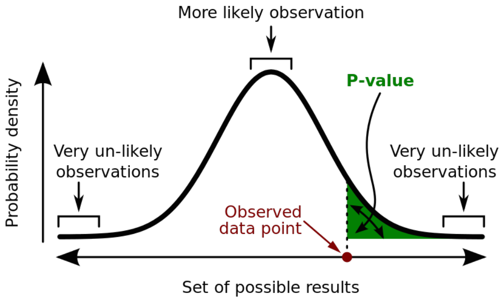
The signifcance level selected for a test should reflect **the real-world consequences** associated with making a Type 1 or Type 2 Error.

Test statistic

- A **test statistic** is a numerical summary of the data that measures compatibility between the null hypothesis and the data.
- The **null distribution** is the probability distribution of the test statistic *when the null hypothesis is true*



The p-value



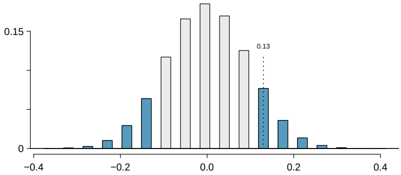
The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, *if the null hypothesis were true*. We typically use a test statistic (a summary statistic of the data) to compute the p-value.

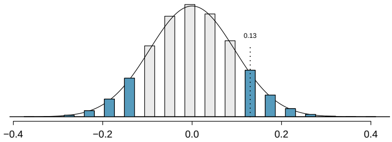
The p-value and decision

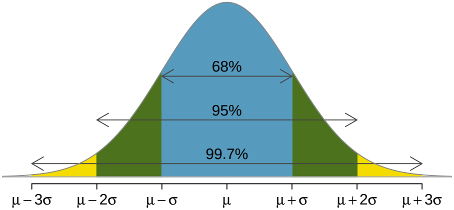
We say that the data provide **statistically significant** evidence against the null hypothesis if the p-value is less than the significance level α (this result would rarely occur just by chance).

Nuzzo, Regina. 2014. **“Scientific Method: Statistical Errors.”** Nature 506 (7487): 150–52.

Standard hypothesis tests







Read Chapter 2, 3 and 4 of “Introductory Statistics with Randomization and Simulation”

Quiz 1

QUIZ 1