Chapter 1: Introduction to data

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Statistics

The general process of investigation has 4 stages:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

We hope to make stages 2-4 objective, rigorous, and efficient.

How best can we collect data? How should it be analyzed? And what can we infer from the analysis?

Our intuitions are often wrong

“Common sense is nothing more than a deposit of prejudices laid down by the mind before you reach age 18.” - Albert Einstein

Our intuitions are often wrong

“Common sense is nothing more than a deposit of prejudices laid down by the mind before you reach age 18.” - Albert Einstein

“Common sense is what tells us the earth is flat.” - Stuart Chase

Our intuitions are often wrong

"Common sense is nothing more than a deposit of prejudices laid down by the mind before you reach age 18." - Albert Einstein

"Common sense is what tells us the earth is flat." - Stuart Chase

Together, a bat and ball cost $\$1.10$. The bat costs $\$1.00$ **more** than the ball. How much does the ball cost?

Our intuitions are often wrong

"Common sense is nothing more than a deposit of prejudices laid down by the mind before you reach age 18." - Albert Einstein

"Common sense is what tells us the earth is flat." - Stuart Chase

Together, a bat and ball cost $\$1.10$. The bat costs $\$1.00$ **more** than the ball. How much does the ball cost?

The ball costs 5 cents.

Incorrect intuitions - survivorship bias

In WWII, the Center for Naval Analyses considered adding armor to planes.

Incorrect intuitions - survivorship bias

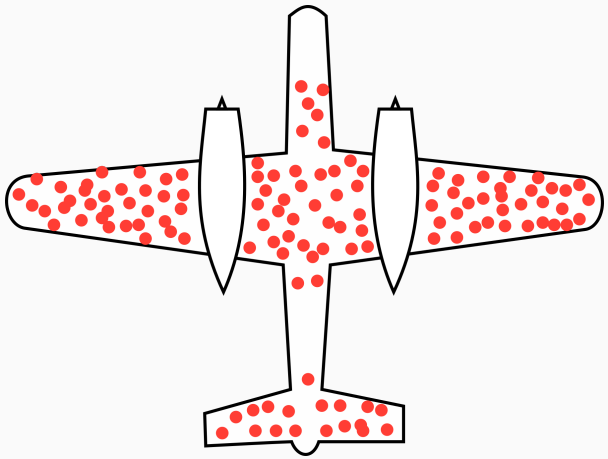
In WWII, the Center for Naval Analyses considered adding armor to planes.

- When planes returned after getting hit by bullets, the locations were documented.

Incorrect intuitions - survivorship bias

In WWII, the Center for Naval Analyses considered adding armor to planes.

- When planes returned after getting hit by bullets, the locations were documented.
- The figure below characterizes the locations of bullet holes in the planes that returned.

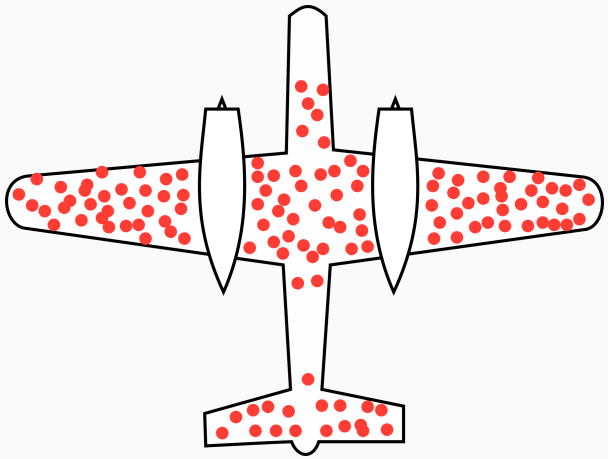


Where should armor be added?

Incorrect intuitions - survivorship bias

In WWII, the Center for Naval Analyses considered adding armor to planes.

- When planes returned after getting hit by bullets, the locations were documented.
- The figure below characterizes the locations of bullet holes in the planes that returned.



Where should armor be added?

Add armor where holes are **not** *located (engines, cockpit, and tail).*

This class will not be intuitive.

- This class will **not** be intuitive.

This class will not be intuitive.

- This class will **not** be intuitive.
- You will need to practice.

This class will not be intuitive.

- This class will **not** be intuitive.
- You will need to practice.
- A lot.

Case study

Using Stents to Prevent Strokes

- Objective: Evaluate the effectiveness of stents in preventing strokes.
- Treatment group: Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- Control group: Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis.

New England Journal of Medicine 365:993-1003.

Results of Using Stents to Prevent Strokes

	stroke	no event
treatment	45	179
control	28	199
Total	73	378

Table 1: Descriptive statistics for the stent study (after 1 year).

Results of Using Stents to Prevent Strokes

	stroke	no event
treatment	45	179
control	28	199
Total	73	378

Table 1: Descriptive statistics for the stent study (after 1 year).

What proportion of treatment group had a stroke?

Results of Using Stents to Prevent Strokes

	stroke	no event
treatment	45	179
control	28	199
Total	73	378

Table 1: Descriptive statistics for the stent study (after 1 year).

What proportion of treatment group had a stroke?

$\frac{45}{45+179} = 20\%$

Results of Using Stents to Prevent Strokes

	stroke	no event
treatment	45	179
control	28	199
Total	73	378

Table 1: Descriptive statistics for the stent study (after 1 year).

What proportion of treatment group had a stroke?

$\frac{45}{45+179} = 20\%$

What proportion of control group had a stroke?

Results of Using Stents to Prevent Strokes

	stroke	no event
treatment	45	179
control	28	199
Total	73	378

Table 1: Descriptive statistics for the stent study (after 1 year).

What proportion of treatment group had a stroke?

$\frac{45}{45+179} = 20\%$

What proportion of control group had a stroke?

$\frac{28}{28+199} = 12\%$

Understanding the results

Do the data show a “real” difference between the groups?

Understanding the results

Do the data show a “real” difference between the groups?

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won’t observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- The observed difference between the two groups (20 - 12 = 8%) may be real, or may be due to natural variation.
- We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

Treating Chronic Fatigue Syndrome

- Objective: Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
- Participant pool: 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- Actual participants: Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Deale et. al. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial*. The American Journal of Psychiatry 154.3 (1997).

Study design

- Patients randomly assigned to treatment and control groups, 30 patients in each group:
  - *Treatment*: Cognitive behavior therapy – collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
  - *Control:* Relaxation – No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

Understanding the results

Do the data show a “real” difference between the groups?

Understanding the results

Do the data show a “real” difference between the groups?

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won’t observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- The observed difference between the two groups (70 - 19 = 51%) may be real, or may be due to natural variation.
- Since the difference is quite large, it is more believable that the difference is real.
- We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

These patients had specific characteristics and volunteered to be a part of this study, therefore they may not be representative of all patients with chronic fatigue syndrome. While we cannot immediately generalize the results to all patients, this first study is encouraging. The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.