ØAMET2200
Business Decision Making Using Data

Lecture 4

Instructor: Fenella Carpena

September 13, 2019

Announcements

- Problem Set 2 has been posted
  - Due Oct. 3 at 11:59 PM
  - After today's lecture, you should be able to do Exercise 3
  - Please **write your name** (or names of all group members) on the **first page** of your write-up
  - If your group members have changed since Problem Set 1, please contact me (**important for grades**)
- OH/Lecture Schedule
  - No OH on the following dates: Sept. 20 (Friday), Sept. 24 (Tuesday), Sept. 27 (Friday)
  - No lecture on Sept. 27

Parts 2-3 of this Course: Making decisions based on inference; Experiments and causality

- Given **limited data**, business decision makers need to choose how to act on it
  - You have two groups you want to compare
  - How do you know this difference isn't due to noise?
  - What is the likelihood of making a mistake in your decision?
- **Statistical tests** provide us with some clarity
- In many comparisons you want to make, you want to attribute differences to some **causal factor**
- **Experiments** and **A/B tests** are a uniquely powerful method for testing causal effects
  - Comparisons from observational (i.e., non-experimental) studies may suffer from **confounding** or **lurking** variables

Agenda for Today

- Chapter 17.1-17.4
- Experiments and A/B tests
- Confounding variables
- Two-sample tests for differences in means or proportions
- Confidence intervals for differences in means or proportions
- Hawthorne effects

What you should take away from this lecture

- How to **conduct** statistical tests comparing two groups

- How to **understand** the results of statistical tests comparing two groups

- How to **interpret** comparisons, keeping in mind the “ideal experiment” and how close you are to it

Case Study for Today: Cookie Cats

- A mobile game puzzle developed by Tactile Entertainment
- Players must connect at least 3 tiles of the same color to clear the board and win the level
- As the game progresses, players encounter gates that force them to wait for some time or to make in-app purchases before they can continue

Case Study for Today: Cookie Cats

- Currently, the first gate in the game is at **level 40**, but the company is deciding whether to move it to **level 30**
- Will moving the first gate to level 30 **cause** an **increase** player's use of the game?
- Two **measures of game use** that we will examine:
  - **Number of game rounds** that a user plays during the first 14 days after installation
  - **Player retention**, an **indicator** variable equal to 1 if the player comes back to the app 7 days after installing, and 0 otherwise

Cookie Cats: Comparison of Two Gate Levels

- What is the **population of interest?** All new and future users of the game
- What are the **population parameters** we are interested in?
  - We'll focus on the number of game rounds played for now
  - Call them μ_{30} and μ_{40}
- μ_{30}: average number of game rounds played during the first 14 days after installation **if the first gate is at level 30**
- μ_{40}: average number of game rounds played during the first 14 days after installation **if the first gate is at level 40**

Cookie Cats: Hypothesis Test

- The first gate is currently at gate 40. Will moving it to gate 30 **cause** an increase in the average # of game rounds played?
- State the **hypothesis test** as
  - $H_0 : \mu_{30} \leq \mu_{40}$
  - $H_1 : \mu_{30} > \mu_{40}$
  - This leads to a one-sided test
- Note that the above hypothesis test is equivalent to
  - $H_0 : \mu_{30} - \mu_{40} \leq 0$
  - $H_1 : \mu_{30} - \mu_{40} > 0$
- Why do we have $\mu_{30} \leq \mu_{40}$ in H_0 instead of $\mu_{30} \geq \mu_{40}$?
- Note that H_0 always contains the cut-off (technical point)

Data for Comparisons

- To test the hypothesis, we need sample data from 2 groups
  1. Players who encountered the first gate at level 40
  2. Players who encountered the first gate at level 30
- Data used to compare two groups typically arise in the following ways:
  1. Run an **experiment** that isolates a specific cause
  2. Obtain random samples from **two populations**
  3. Compare two sets of **observations**
- Experiments are the most reliable

Experiments: Basic Terminology

- **Experiment**: procedure that uses random assignment to produce data that reveal causation
- **Treatment**: something done to an experiment participant in order to see its effect, for example:
  - `treatment = 1` if participant received a drug for diabetes
  - `treatment = 0` if participant received a placebo
- **Response:** a variable we measure to see if the treatment had an effect (e.g., person's blood sugar level)
- **A/B Test:** a type of experiment with two variants, A and B; often used in website, app, and product design
- **Observational Study**: comparing two groups without an experiment

The Ideal Experiment or A/B Test

- We have a random sample from the population
  - Crucial for **external validity**: can you generalize the results from your sample?
- Subjects are **randomly assigned** to either a treatment (T) group or a control (C) group
  - Crucial for **internal validity**: have you actually identified the causal effect of a specific treatment?
  - Must also have **full compliance**: no one in C group got treatment, and everyone in T group got treatment
- We compare the response between T and C groups
  - The difference in response is **caused by** the treatment

Why is an experiment or A/B test useful?

1. Experiments are **transparent**
   - Simple to explain to non-experts
   - When implemented properly, results are highly credible
2. Random assignment to T and C group through experiments eliminates **confounding**

Confounding

- For a comparison to be valid, the **only** difference between the T and C group should be that T group received treatment

- If there are other differences, we say there are **potential confounding factors** or **lurking variables**

- Confounding means we are **mixing effects** of two or more causes when making a comparison

- Confounding matters because **incorrect beliefs about cause and effect** lead to poor decisions

Examples of Confounding in Observational Studies

- People who receive heart transplants are likely to have shorter life expectancies than people who do not
  - Are heart transplants a bad thing?
- Adults who drink a glass of wine everyday have lower risk of cardiovascular disease than adults who do not
  - Does this mean drinking wine leads to better health?
- Students who hire a tutor do better on exams than students who do not
  - Is it all coming from the tutoring?

Espresso House: Exercise

Espresso House is testing 2 types of ads to promote its new høstlatte. Version A includes the price of the latte, and Version B does not. Which is more effective for increasing sales of høstlatte?

Consider the following approaches for comparisons. Explain whether the approach is contaminated by confounding. If so, identify a possible confounding or lurking variable.

(a) Version A is used at a store in Bergen and Version B in Oslo.

(b) Version A is used on Mondays at a store in Oslo, and Version B is used on Fridays at the same store.

(c) The høstlatte advertisement is sent by mail to people living in Oslo. Version A was sent to randomly selected set of people, and the same is done for Version B.

Using data from an experiment or A/B test to compare sample averages

- To assess whether the difference between **sample average** between the T and C groups is statistically significant, we can use a **two-sample *t*-test**

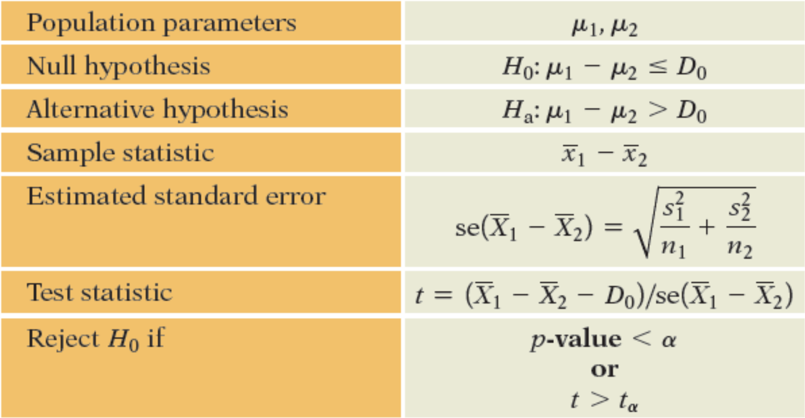
- The **two-sample *t*-statistic** is

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{se(\overline{X}_1 - \overline{X}_2)}$$

- Sample size condition must be satisfied **in both groups**
  - $n > 10 \cdot |K_4|$ in Group 1 and also in Group 2

- Degrees of freedom: complicated!
  - Can use the rule of thumb $df = n_1 + n_2 - 2$ (if the variances in the two groups are similar)
  - But use Stata for more accuracy

Comparing Differences in Sample Means: Details



Population parameters	μ_1, μ_2
Null hypothesis	$H_0\colon \mu_1 - \mu_2 \leq D_0$
Alternative hypothesis	$H_a\colon \mu_1 - \mu_2 > D_0$
Sample statistic	$\bar{x}_1 - \bar{x}_2$
Estimated standard error	$\mathrm{se}(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Test statistic	$t = (\overline{X}_1 - \overline{X}_2 - D_0)/\mathrm{se}(\overline{X}_1 - \overline{X}_2)$
Reject H_0 if	p-**value** $< \alpha$ **or** $t > t_\alpha$

Cookie Cats: A/B Test

- To decide whether they should move the first gate to level 30, the Cookie Cats game developers conducted an A/B test

- Sample of 90,189 new players who downloaded the app; players randomly assigned to Version A or B of the game

- 44,700 players got Version A of the game (1st gate at level 30)

- 45,589 players got Version B of the game (1st gate at level 40)

Cookie Cats: Descriptive Statistics

		# Game Rounds Played		
Game Version	Sample Size	Mean	SD	Kurtosis
Level 30	44,700	52.46	256.72	31687.84
Level 40	45,589	51.30	103.29	66.10

- Use the following command in Stata to get the above info:

  ```
  summarize gamerounds if version == "gate_30", detail
  summarize gamerounds if version == "gate_40", detail
  ```

- The **sample average** # of games rounds played (during first 2 weeks after installation) is higher in the level 30 than level 40 version

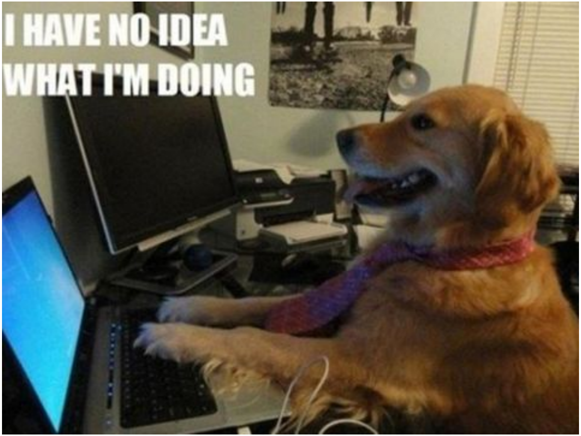
- Is this difference statistically significant?

Cookie Cats: Exercise

(a) Does the A/B test show that moving the first gate to level 30 increases the average number of game rounds played? Use 1% significance level for the hypothesis test.

(b) What is the *p*-value of the test from part (a)?

What are we doing in a hypothesis test?



Cookie Cats: Exercise

(c) Interpret the results from part (b).

(d) The hypothesis test in part (a) requires that sample size conditions hold. Is this the case?

Comparing Differences: Summary of Steps

Step 0 Check that SRS and sample size conditions hold

Step 1 Choose a level of significance α (or it is given to you)

Step 2 Formulate the null and alternative hypothesis.

- H_0 always includes the “cutoff” value
- H_0, H_1 involve population parameters, NOT sample statistics

Step 3 Calculate the test statistic (e.g., t-statistic)

$$\frac{\text{sample statistic} - \text{“cutoff” value from hypothesis}}{\text{se(sample statistic)}}$$

Step 4 Find the critical value, which depends on α

Step 5 Compare test statistic (from Step 3) with the critical value (from Step 4). Or, calculate the p-value.

Step 6 Apply rejection rule, e.g., if p-value $< \alpha$, reject H_0

Comparing Differences in Sample Proportions

- Previously, we compared differences in the **sample average** (game rounds played) between the level 30 vs. level 40
- We can also compare differences in **sample proportions**
- Do the data show that the level 30 version produces a higher **proportion of users** who return to the app 7 days after installation?
- State the **hypothesis test** as
  - $H_0 : p_{30} - p_{40} \leq 0$
  - $H_1 : p_{30} - p_{40} > 0$
- p_{30}: proportion of users in gate level 30 version who return 7 days after installation
- p_{40}: proportion of users in gate level 40 version who return 7 days after installation

Using data from an experiment or A/B test to compare sample proportions

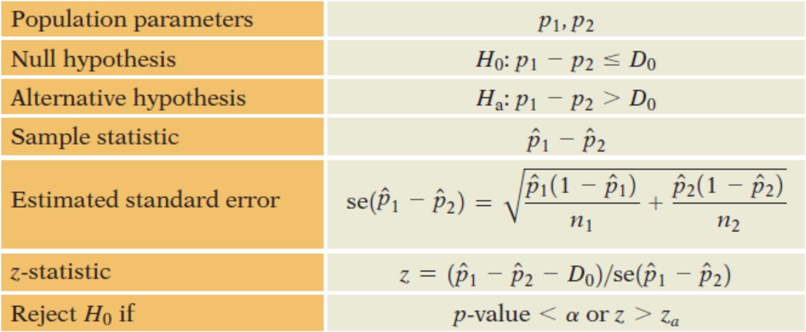
- To assess whether the difference between the **sample proportions** is statistically significant, we can use a **two-sample *z*-test**

- The **two-sample *z*-statistic** is

$$z = \frac{(\widehat{p}_1 - \widehat{p}_2) - D_0}{se(\widehat{p}_1 - \widehat{p}_2)}$$

- Sample size condition must be satisfied **in both groups**
  - $n \cdot \widehat{p} \geq 10$, $n \cdot (1 - \widehat{p}) > 10$ in Group 1 and also in Group 2

Comparing Sample Proportions: Details



Population parameters	p_1, p_2
Null hypothesis	$H_0\colon p_1 - p_2 \leq D_0$
Alternative hypothesis	$H_a\colon p_1 - p_2 > D_0$
Sample statistic	$\hat{p}_1 - \hat{p}_2$
Estimated standard error	$\text{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$
z-statistic	$z = (\hat{p}_1 - \hat{p}_2 - D_0)/\text{se}(\hat{p}_1 - \hat{p}_2)$
Reject H_0 if	p-value $< \alpha$ or $z > z_a$

Cookie Cats: Descriptive Statistics for Proportion

		Retention Rate
Game Version	Sample Size	Proportion who return after 7 days
Level 30	44,700	0.1902
Level 40	45,589	0.1820

- The proportion of users who return to the game 7 days after installation is higher with level 30 than level 40.

- Is this difference statistically significant?

Cookie Cats: Exercise

(e) Does the A/B test show that moving the first gate to level 30 increases the proportion of users who come back to the game 7 days after installation? Use 1% significance level for the hypothesis test.

(f) What is the *p*-value of the test from part (e)?

Cookie Cats: Exercise

(g) Interpret the results from part (f).

(h) The hypothesis test in part (e) requires that sample size conditions hold. Is this the case?

Extension: Using values other than zero in hypothesis tests

- So far, we have been comparing the difference in means and the difference in proportions to the value **zero**

$$H_0 : \mu_1 - \mu_2 \leq 0,\ H_1 : \mu_1 - \mu_2 > 0$$
$$H_0 : p_1 - p_2 \leq 0,\ H_1 : p_1 - p_2 > 0$$

- It's also possible to compare the difference to a **different** value
- Replace "D_0" in the formula for the t-stat or the z-stat

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{se(\overline{X}_1 - \overline{X}_2)} \qquad z = \frac{(\widehat{p}_1 - \widehat{p}_2) - D_0}{se(\widehat{p}_1 - \widehat{p}_2)}$$

- Here's an example:

$$H_0 : p_1 - p_2 \leq 0.04$$
$$H_1 : p_1 - p_2 > 0.04 \qquad z = \frac{(\widehat{p}_1 - \widehat{p}_2) - 0.04}{se(\widehat{p}_1 - \widehat{p}_2)}$$

Cookie Cats: Exercise

(i) Suppose that moving the first gate from level 40 to level 30 is profitable for Cookie Cats only if the proportion of users who come back to the game after 7 days increase by more than 0.005 (half a percentage point).

What is the null and alternative hypothesis?

(j) Carry out the test in part (i) using the 5% significance level. What do you conclude?

CI for Differences in Means and Proportions

- The two-sample *t*-test (or *z*-test) determine **if** the sample mean (or proportion) is statistically significantly larger in one group vs. another

- We get a **yes** or **no** answer from the test

- In some applications, we might want to know the **range** for the difference instead: a **CI** is more appropriate here

CI for Differences in Means and Proportions

- The $100(1-\alpha)\%$ two-sample CI for the difference in **means** is

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} \cdot se(\overline{X}_1 - \overline{X}_2)$$

- The $100(1-\alpha)\%$ two-sample CI for the diff. in **proportions** is

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2} \cdot se(\widehat{p}_1 - \widehat{p}_2)$$

- If the 95% CI does not include zero, the difference (in means or proportion) is statistically significant at the 5% level

Cookie Cats: Exercise

(k) What is the 95% CI for the difference in the 7-day retention rate (i.e., the proportion of users who come back to the game 7 days after installation) between the two game versions?

(l) Interpret the 95% CI from part (k).

Cookie Cats: Summary of our analysis
4M Analytics decision making strategy

Motivation

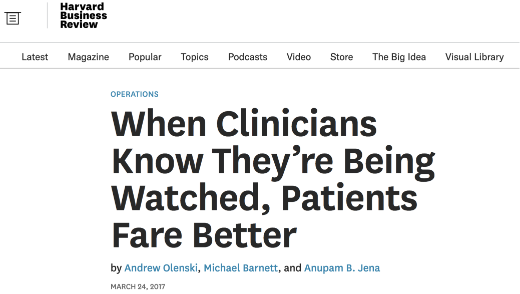
Method

Mechanics

Message

Hawthorne Effects

- Experiment or A/B test participants may change their behavior if they know they are being studied
- This phenomenon is called the "Hawthorne Effect"
  - Named after a series of studies conducted at a factory called Hawthorne Works
- Threatens the validity of an experiment or A/B test



OPERATIONS

When Clinicians Know They're Being Watched, Patients Fare Better

by Andrew Olenski, Michael Barnett, and Anupam B. Jena

MARCH 24, 2017

Colgate vs. Solidox: Exercise

Colgate and Solidox are two brands of toothpaste in Norway. Both brands have a line of teeth whitening toothpaste: Colgate Total White and Solidox White Activator.

Colgate has developed a new toothpaste formula. They would like to test if it is more effective at teeth whitening than Solidox, so they conducted an experiment. 150 people in the treatment group used Colgate for 3 days. 220 people in the control group used Solidox, also for 3 days. After 3 days,72% of treatment group had whiter teeth, in comparison to 60% in the control group.

As a data scientist at the company, you have been tasked with analyzing the data from the experiment.

(a) Do the results of experiment show that Colgate is better at teeth whitening than Solidox? Use the 5% significance level.

(b) The marketing head at Colgate wants to change the packaging to say “You’re 30 percentage points more likely to get whiter teeth with Colgate than Solidox.” How do you respond?

Colgate vs. Solidox: Exercise

Takeaways

- Observational studies may suffer from **confounding** variables
- After observing a pattern in data, you need to ask:
  - What causal effect is **suggested**?
  - What **alternative stories** are there for the observed patterns?
- Use experiments or A/B tests to discover **causal** relationships
- Use two-sample t-tests or z-tests to compare sample means and proportions to a particular **value**
- Use two-sample CI for the difference in means/proportions to get a **range of values** of the population difference