# Chapter 3: Hypothesis Testing

In **Hypothesis Testing**

- A claim is made about the population, and researchers want to determine which of the two competing claims is more likely.
- Researchers collect data and compare statistics.
- Based on data collected, researchers assess which of the claims seems more likely.

Some examples of questions that could be answered using a hypothesis test are:

- What is a better method to help smokers quit: give them money for successfully quitting, or penalize them if they don't stop?
- What is the better mailer to send in order to nudge voters into actually voting?
- What is a more effective government assistance program, giving people food stamps or cash?

We refer to the two competing hypotheses as the **null hypothesis**, denoted by $H_0$, and the **alternative hypothesis**, denoted by $H_a$.

- $H_0$ is the boring claim that nothing interesting is happening.
- $H_a$ is new or different result a researcher is trying to establish or find evidence for.

**Collect sample data. Then assess the competing claims.**

1. Telepathy is the ability of an individual to communicate thoughts and ideas by means other than the known senses. I claim that I do have telepathy. There are two possibilities: either I do or I do not. Which claim is the null hypothesis and which is the alternative?

$H_0$: I don't have telepathy (boring).
$H_a$: I do have telepathy (interesting).

2. There are many experiments we could try to run to test these competing claims. For example, I will think of a letter A, B, C, or D and communicate this letter to each of you. If I could collect data from everyone in the population, and let $p$ denote the proportion of all people that say the letter I was thinking of. If $H_0$ is true, what would you expect the value of $p$ to be? If $H_a$ is true, what would you expect the value of $p$ to be?

$p = 1/4$

3. There are (about) ~~25~~ 10 people in this room. If $\hat{p}$ denotes the proportion of the people in this class that say the letter I was thinking of. What would be enough evidence to convince you that I do have telepathy? What would need to be true about $\hat{p}$?

Larger $\hat{p}$ is more convincing.

## Section 3.2: Hypotheses and Significance

The general steps for a hypothesis test are summarized below.

- Set the hypotheses in terms of population parameters. Use an equal sign in the null hypothesis. Depending on what researchers are hoping to prove use $\neq$, $<$, or $>$ in the alternative hypothesis.
- Collect data and define a statistic that can be used to assess the hypotheses. Compute the **test statistic** usingn the collected data.
- Assume $H_0$ is true. Under this assmuption, is the observed test statistic likely? Unlikely? If the **test statistic is very unlikely** (under the assumption in $H_0$):
  - The test is **statistically significant**.
  - We have convincing evidence the null hypothesis is wrong.
  - **We reject $\mathbf{H_0}$ and accept the alternative hypothesis**.
- If the **test statistic seems plausiible** (under the assumption in $H_0$):
  - The test is **not statistically significant**.
  - We cannot be sure whether the claim in $H_0$ is true or not.
  - The **test is inconclusive**. We neither reject nor accept $H_0$.

4. A 2004 article[11] from the Economic Journal studied the so called unscrupulous diner's dilemma.

> The unscrupulous diner's dilemma is a problem faced frequently in social settings. When a group of diners jointly enjoys a meal at a restaurant, often an unspoken agreement exists to divide the check equally. A selfish diner could thereby enjoy exceptional dinners at bargain prices. . . This dilemma typifies a class of serious social problems[12] from environmental protection and resource conservation to eliciting charity donations and slowing arms races.

Researchers wanted to test whether people order more food and beverages when they know the bill is going to split evenly, or do they order the same amount regardless of whether they are splitting the bill or paying individually.

$\mu_{even}$ = mean paid by even split

$\mu_{control}$

(a) State the null and alternative hypotheses in words.

$H_0$: Boring Claim. People order same amount of food either way. How bill is split does not make a difference.

$\mu_{control} - \mu_{even} = 0$

$H_a$: what they hope to prove.

★ When bill is split evenly, people order more food. ★

$\mu_{control} - \mu_{even} < 0$

[11] http://rady.ucsd.edu/faculty/directory/gneezy/pub/docs/splitting-bill.pdf

[12] http://www.uvm.edu/ pdodds/files/papers/others/1994/glance1994a.pdf

(b) To test the claims, participants were randomly assigned into two tables, each with four people. One table (even-split group) was randomly picked and told they were going to evenly-split the bill. The other table (control) was told each person was going to pay for what they ordered. The mean amount ordered by the control group was $8.67. Which of following samples for the even-split group is the most statistically significant? Support your answer with an explanation.

(i) $4.67 (ii) $8.50 (iii) $8.80 [✓ (iv) $11.23] ☆

If $H_0$ is true, we'd expect even split to also have mean around $8.67

control orders more

unlikely if $H_0$ were true and provides evidence for $H_a$.

(c) Restate the hypotheses in terms of the parameters $\mu_{\text{even}}$ and $\mu_{\text{control}}$, the true mean amount ordered by people that evenly-split and invidiually pay for the bill, respectively.

$$H_0: \mu_{even} - \mu_{control} = 0$$

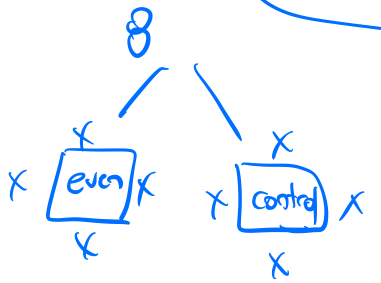
$$H_a: \mu_{even} - \mu_{control} > 0$$

☆

(d) If the table below gives the amounts ordered by each of the four people in each group, what is the value of the test statistic?

| Even-Split | | | |
|---|---|---|---|
| $15.00 | $8.00 | $8.75 | $13.17 |

| Control | | | |
|---|---|---|---|
| $8.50 | $7.90 | $10.85 | $7.43 |



$$\bar{x}_{even} = \$11.23$$

$$\bar{x}_{control} = \$8.67$$

$$\boxed{t = \bar{x}_{even} - \bar{x}_{control} = \$2.56}$$

(e) Based on the test statistic, what do you think we can conclude about the two competing claims?

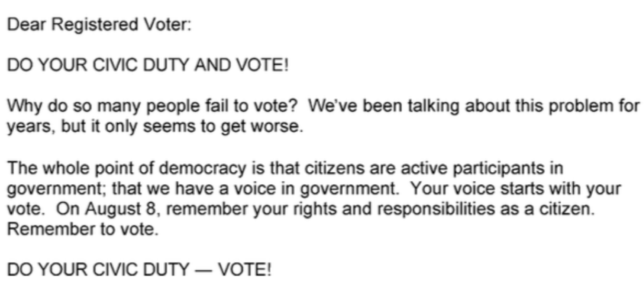
?? It depends on a number of factors.

## Example: Social Pressure and Voter Turnout

5. A 2008 experiment[13] at Yale aimed to determine whether positive or negative pressure is more effective at improving voter turnout.

Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty. We distinguish between two aspects of this type of utility,

One group received a mailing emphasizing the intrinsic (internal) statisfaction for voting:



Dear Registered Voter:

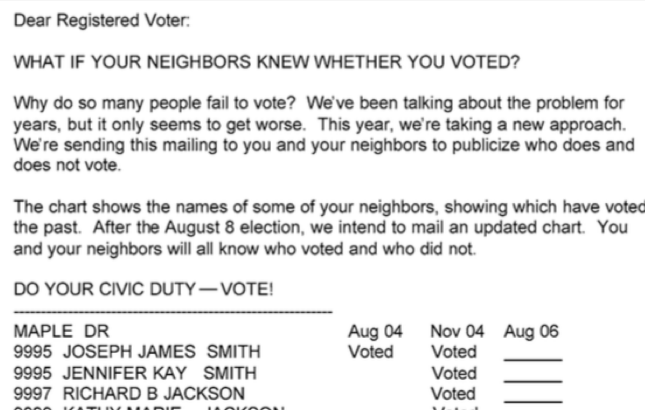
DO YOUR CIVIC DUTY AND VOTE!

Why do so many people fail to vote? We've been talking about this problem for years, but it only seems to get worse.

The whole point of democracy is that citizens are active participants in government; that we have a voice in government. Your voice starts with your vote. On August 8, remember your rights and responsibilities as a citizen. Remember to vote.

DO YOUR CIVIC DUTY — VOTE!

Another group received a mailing placing extrinsic (outside) pressure on people to vote:



Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

| MAPLE DR | Aug 04 | Nov 04 | Aug 06 |
|---|---|---|---|
| 9995 JOSEPH JAMES SMITH | Voted | Voted | ______ |
| 9995 JENNIFER KAY SMITH | | Voted | ______ |
| 9997 RICHARD B JACKSON | | Voted | ______ |

(a) State the null and alternative hypotheses in words the researches can use to test whether positive or negative pressure is more effective at improving voter turnout.

$H_0$: There is no difference. $p_{int} - p_{ext} = 0$

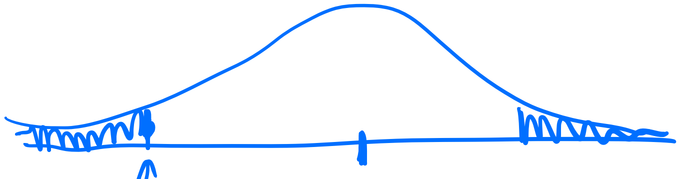
$H_a$: There is a difference. $p_{int} - p_{ext} \neq 0$

(b) What is a possible test statistic the researchers could use to assess the competing claims in your previous answer?

$$t = \hat{p}_{int} - \hat{p}_{ext}$$

(c) Using the test statistic in your previous answer, restate your null and alternative hypotheses using appropriate notation.

Area in tails = P-value



extreme? $p_1 - p_2 = 0$

[13] http://isps.yale.edu/sites/default/files/publication/2012/12/ISPS08-001.pdf

## Calculating $P$-Values

- ☆ The **P-value** is the probability that you would get a random sample with a test statistic as or more extreme as the observed test statistic if the null hypothesis were true.
- The smaller the $P$-value is, the less likely the sample is, and there is evidence that contradicts $H_0$ and supports $H_a$.
- **Thus, the smaller the $P$-value, the more statistically significant the result is.**

6. In the telepathy example, let $T$ be the number of people out of ~~25~~ 10 that say the letter I was thinking. If we observed that ~~20~~ out of ~~25~~ people say the letter I was thinking.

4 out of 10

(a) Calculate the $P$-value of the observed test statistic. In other words, in ~~25~~ 10 identical and independent trials each with likelihood of success $p = 0.25$, compute $P(T \geq \cancel{20})$ 4.

$H_0: p = 1/4$

$H_a: p > \frac{1}{4}$

✱ Let's assume $p = 1/4$

Our sample has $n = 10$

$T = $ # of people who correctly guess

$$\boxed{P\text{-value} = P(T \geq 4) = 1 - \text{pbinom}(3, 10, \tfrac{1}{4})} \boxed{= 0.2241}$$

(b) Based on the value of the $P$-value, what can we conclude about my telepathy ability?

22.41% chance of 4 or more people out of 10 getting correct letter if $p = 1/4$.

The **null distribution** is the distribution of the test statistic if the null hypothesis is true.

7. What is the null distribution for the previous telephathy example?

$$T \sim \text{Binom}(10, 1/4)$$