Gov 50: 20. Hypothesis testing: One-sample Tests

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. Statistical thought experiments

3. Hypothesis tests

4. p-values

5. Small samples

1/ Today's agenda

Where are we? Where are we going?

- Up to now: learned the properties of estimators.
- Formed confidence intervals around our estimates.
- Now: test hypotheses about the population from our sample.

- Logistics:
  - Last DataCamp due Thursday.
  - Draft analyses for final project due next.
  - Midterm 2: definitely harder, but remember the curve!
  - You all are learning a ton in this class and it shows.

2/ Statistical thought experiments

The lady tasting tea

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea poured before the milk. You stop by Tealuxe and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You are skeptical that she can really tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to friend in a **random** order
  - Ask friend to pick which 4 of the 8 were milk-first.

Assuming we know the truth

- Friend picks out all 4 milk-first cups correctly!
- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly**?
  - Only one way to choose all 4 correct cups.
  - But 70 ways of choosing 4 cups among 8.
  - Choosing at random ≈ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- $\rightsquigarrow$ the guessing hypothesis might be implausible.

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**.
- What would the world look like **if we knew the truth**?
- Example 1:
  - An analyst claims that the proportion of poor households in Boston is 20%.
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line.
  - Should you conclude that the analyst is wrong?
- Example 2:
  - Trump won 46% of the vote in the 2016 election.
  - A recent poll of 1600 registered voters found that 43% of the public approves of Donald Trump.
  - Has support for Trump changed or could this poll be just due to random chance?

3/ Hypothesis tests

Hypothesis testing procedure

Conducted with several steps:

1. Generate your **null** and **alternative hypotheses**
2. Collect sample of data
3. Calculate appropriate **test statistic**
4. Use that value to calculate a probability called a **p-value** (more later)
5. Use p-value to decide whether to reject the null hypothesis or not

- This procedure is general, but we'll focus on tests of a single population mean today.

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters.
  - The "devil's advocate" hypothesis $\rightsquigarrow$ assumes what you seek to prove wrong.
  - Ex: your friend is guessing randomly about the tea.
  - Ex: Trump's approval is the same as election result.
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0.
  - It is the opposite of the null hypothesis.
  - Ex: your friend can tell milk-first vs. tea-first.
  - Ex: Trump's support has changed since the election.
  - Denoted H_1 or H_a
- **Probabilistic** proof by contradiction: try to disprove the null.

Hypothesis testing example

- Has Trump's approval changed since the election?
- What is the parameter we want to learn about?
  - True population proportion supporting Trump, p.
- **Step 1**: How to state the hypotheses
  - Null hypothesis: $H_0 : p = 0.46$
  - Alternative hypothesis: $H_1 : p \neq 0.46$
- **Step 2**: gather data (done for us)
  - Poll has $\overline{X} = 0.43$ with $n = 1600$.

Test statistic

- **Step 3**: calculate appropriate test statistic
  - A **test statistic** will help us adjudicate between the null and alternative.
  - Bigger values of the test statistic $\rightsquigarrow$ null less plausible
  - How to calculate?

Calculating the test statistic

- By the CLT, we know that the sample proportion is roughly normal:

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- We can use the transformations of normals to standardize this:

$$Z = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

- Put differently:

$$Z = \frac{\text{observed - null guess}}{SE}$$

  - How many SEs away from the null guess is the sample mean?

Calculating the test statistic

- Test statistic:

$$Z = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}}$$

- This poll: $\overline{X} = 0.43$ with sample size $n = 1600$
- Key step: statistical thought experiment is to **assume the null is true** $p = 0.46$
- Estimate the test statistic:

$$Z_{\text{obs}} = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} = \frac{0.43 - 0.46}{\sqrt{0.46(1-0.46)/1600}}$$
$$\approx -2.4$$

4/ p-values

Finding the p-value

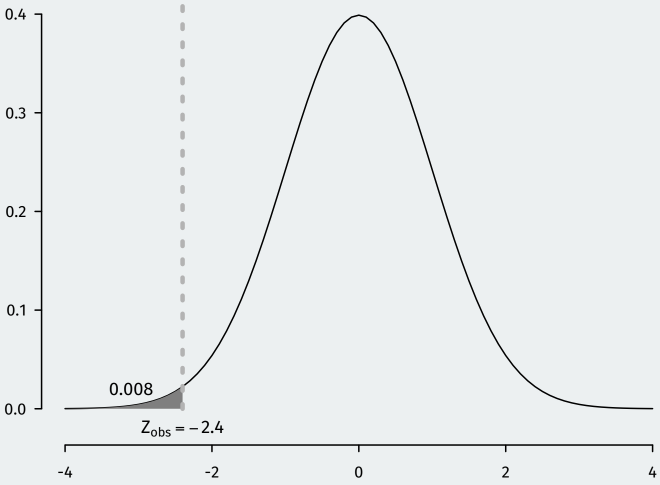
- **Step 4**: determine the p-value.
  - The **p-value** is the probability of observing a test statistic as extreme as Z_{obs}, if the null hypothesis is true.
  - smaller p-values $\rightsquigarrow$ more unlikely test statistic under the null $\rightsquigarrow$ null less plausible
- How to calculate?
  - We know Z is distributed standard normal $\rightsquigarrow$ use R!

Standard normal probabilities in R

- The `pnorm(x)` function will give the probability of being less x in a standard normal:

```
pnorm(-2.4)
```

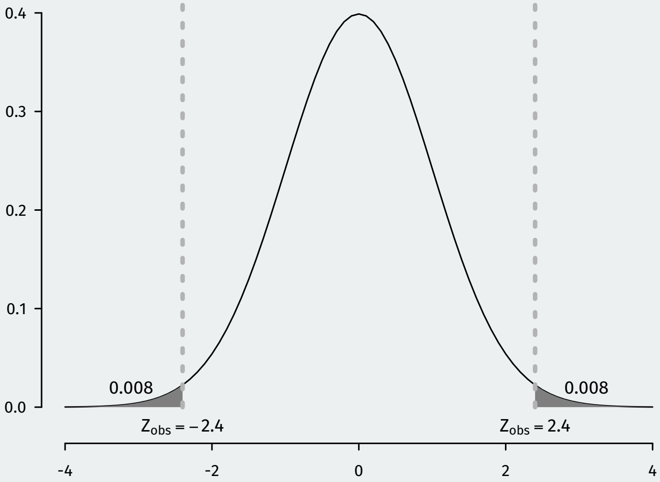
```
## [1] 0.0082
```



0.4
0.3
0.2
0.1
0.0
0.008
$Z_{obs} = -2.4$
-4
-2
0
2
4

One-sided vs. two-sided tests

- p-value depends on the form of the alternative hypothesis
- Can either use a **one-sided** or **two-sided** alternative:
  - $H_1 : p < 0.46$ (one-sided, Trump has lost support)
  - $H_1 : p > 0.46$ (one-sided, Trump has gained support)
  - $H_1 : p \neq 0.46$ (two-sided)
- One sided p-value: $\mathbb{P}(Z > Z_{\text{obs}})$ or $\mathbb{P}(Z < Z_{\text{obs}})$
- Two sided p-value: $\mathbb{P}(Z > |Z_{\text{obs}}|) + \mathbb{P}(Z < -|Z_{\text{obs}}|)$
  - Because of the symmetry of the normal: $2 \times \mathbb{P}(Z > |Z_{\text{obs}}|)$



- two-sided p-value: 0.016

p-value interpretation

- **p-value**: probability of the data assuming the null is true.
  - **statistical thought experiment**: how likely is our data in a world where the null hypothesis is true.
- **NOT** the probability that the null is true!
- We can interpret it as the strength of evidence against the null.
  - Smaller p-values $\rightsquigarrow$ more evidence against the null.

Rejecting the null

- **Step 5**: use p-value to decide whether to reject null or not.
- Choose a threshold below which you'll reject the null.
  - **Test level** α: the threshold for a test.
  - Decision rule: "reject the null if the p-value is below α"
  - Otherwise "fail to reject" or "retain", not "accept the null"
- Common (arbitrary) thresholds:
  - $p \geq 0.1$ "not statistically significant"
  - $p < 0.05$ "statistically significant"
  - $p < 0.01$ "highly significant"

Testing errors

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true.
  - $\rightsquigarrow$ 5% of the time we'll reject the null when it is actually true.
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- Type I error because it's the worst (like convicting an innocent null hypothesis)
- Type II error less serious (missed out on an awesome finding)

5/ Small samples

Problem of small samples

- Up to now, we've relied on the CLT to justify confidence intervals and hypothesis tests.
  - "Sums and means of random variables tend to be normally distributed as sample sizes get big."
- What if our sample sizes are low?
  - Distribution of $\overline{X}$ will be unknown
  - $\rightsquigarrow$ can't determine p-values
  - $\rightsquigarrow$ can't get z values for confidence intervals
- Very difficult to get around this problem without more information.

Solution to small samples?

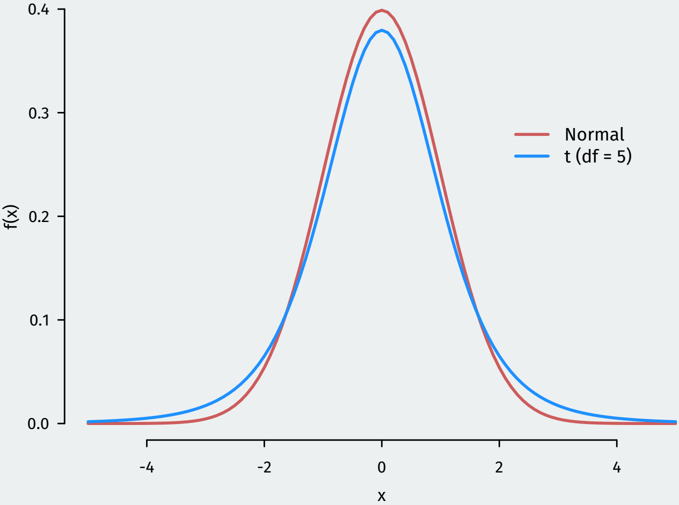
- Common approach: assume data X_i are **normally distributed**
  - THIS IS AN ASSUMPTION, PROBABLY IS WRONG.
  - For instance, if X_i is binary, then it is very wrong.
- If true, then we can determine the distribution of the following test statistic:

$$T = \frac{\overline{X} - \mu}{\widehat{\text{SE}}} \sim t_{n-1}$$

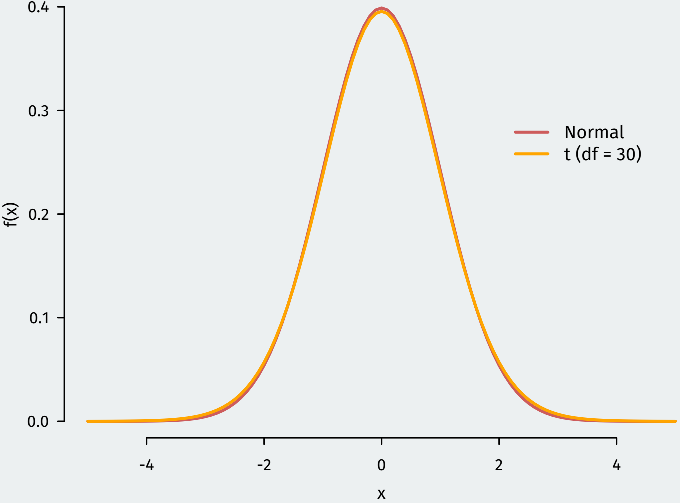
- In words: T follows a "t" distribution with $n-1$ degrees of freedom.
  - Called Student's t distribution and only has 1 parameter, "degrees of freedom."
  - Centered around 0
  - Similar to normal with fatter tails $\rightsquigarrow$ higher likelihood of extreme events.

Who was Student?

Student's t distribution



Student's t distribution



z-test vs t-test

- z-tests are what we have seen: relies on the normal distribution.
  - Justified in large samples (roughly $n > 30$) by CLT
- t-tests rely on the the t-distribution for calculating p-values.
  - Justified in small samples if data is normally distributed.
- Common practice is to use t-tests all the time because t is “conservative”
  - $\rightsquigarrow$ p-values will always be larger under t-test
  - $\rightsquigarrow$ always less likely to reject null under t
  - t-distribution converges to standard normal as $n \to \infty$
- R will almost always calculate p-values for you, so details of t-distribution aren’t massively important.

Next time

- Another round of testing.
- Two-sample tests: hypothesis tests for comparison of means.
- Next week: inference for regression.