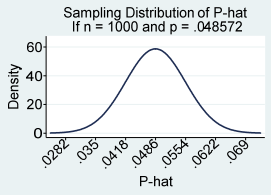
Hypothesis Testing Using a P-value Approach

Lecture 13

Reading: Sections 12.1 – 12.3

Review Sampling Distribution of $\hat{P}$

- 1,673,780 of Aboriginal identity of a total population of 34,460,065
- Population parameter $p = \frac{1,673,780}{34,460,065} = 0.048572$
- Consider $n = 1{,}000$
- What is sampling distribution of $\hat{P}$?



$E[\hat{P}] = p$ and $SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}}$

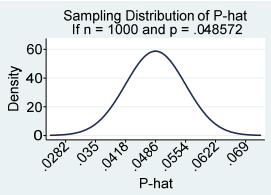
Rule of thumb, Normal: expect ≥ 10 Abor. ident. and ≥ 10 not

Stats Canada, "Aboriginal Peoples Highlight Tables, 2016 Census," https://www150.statcan.gc.ca/n1/en/pub/11-627-m/11-627-m2017027-eng.pdf

Sampling Distributions: Basis of Inference

- Suppose $\hat{P} = 0.054$
- Three ways to explain discrepancies between statistic and parameter:
  1. Sampling error
  2. Non-sampling errors
  3. Parameter differs from the claimed value

Is sampling error a *plausible* explanation for a sample proportion as large as 0.054?



But usually we don't know p: we need to make an *inference*

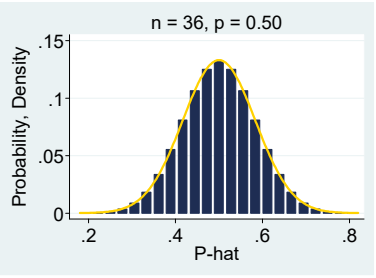
- Estimation starts with $\hat{P}$
- Hypothesis testing starts with a claim to test

Is this coin fair?

- For this coin, consider your initial presumption about the proportion of heads if you *imagine* tossing it an *infinite* number of times
  - What is your initial presumption about?
  - Is it based on any direct evidence about this coin?
- Sample to inform inference about the coin's fairness:
  - $n =$ _____; $\hat{P} =$ _____

4

Sampling Distribution for Fair Coin



Which is easier to prove: a coin is fair or a coin is unfair?

If $\hat{P} = \frac{20}{36}$, conclude?
- Coin is fair
- Coin is unfair
- Neither

If $\hat{P} = \frac{13}{36}$, conclude?
- Coin is fair
- Coin is unfair
- Neither

If $\hat{P} = \frac{29}{36}$, conclude?
- Coin is fair
- Coin is unfair
- Neither

5

Hypothesis Testing Overview

- Null hypothesis (H_0): Initial presumption about the unknown parameter, which is not based on any evidence (data)
  - E.g. H_0: $p = 0.50$ (coin is perfectly fair)
- Collect data and compute sample statistics
- If observed data are surprising (i.e. not very plausible) given the null hypothesis then reject H_0
  - Formal statistical methods calculate exactly how surprising the data are conditional on H_0 being true

6

Research (aka Alternative) Hypothesis

- If reject H_0 have sufficient evidence to infer H_1
- Research or Alternative Hypothesis (H_1 or H_A): Can be proven given evidence (data)
  - E.g. H_1: $p > 0.50$ (coin has a head bias)
  - E.g. H_1: $p < 0.50$ (coin has a tail bias)
  - E.g. H_1: $p \neq 0.50$ (coin is biased)
  - One-tailed test: H_1 has directional sign: > or <
  - Two-tailed test: H_1 has not equal sign: ≠

7

Analogy: Jury Makes an Inference

- H_0: Defendant is innocent
- H_1: Defendant is guilty
- Acquit: insufficient evidence to infer guilt
  - Fail to reject presumption of innocence (H_0)
    - Does this mean there is no evidence of guilt?
    - If you are acquitted does that prove innocence?
- Convict: enough evidence to infer guilt
  - Reject presumption of innocence (H_0) and infer guilty: prove H_1 beyond a reasonable doubt

8

Two Options: Reject or Fail to Reject

- If sample gives conclusive evidence, then reject null (H_0) and infer the research hypothesis (H_1) is true (it has been proven)
- If inconclusive (weak) evidence in favor of H_1, then fail to reject H_0
- With a sample, cannot "prove" or "accept" H_0
  - Cannot prove coin is fair
- Asymmetry: cannot infer H_0 is true but can reject H_0 to prove H_1
  - Researcher hopes to prove H_1: the research hyp. is more important

The burden of proof is on the research hypotheses (H_1)
(e.g. prosecution must prove guilty beyond a reasonable doubt)

9

Writing Hypotheses in Formal Notation

- To show that more than 1 in 10 Canadians give to the United Way:
  - H_0: $p = 0.10$
  - H_1: $p > 0.10$
- To show that average debt of 23 year olds is less than $50K:
  - H_0: $\mu = 50{,}000$
  - H_1: $\mu < 50{,}000$
- To show that proportion of women who smoke differs from men:
  - H_0: $p_{women} = p_{men}$
  - H_1: $p_{women} \neq p_{men}$
- To show mean donation is higher for 2:1 match than a 1:1 match
  - H_0: $(\mu_2 - \mu_1) = 0$
  - H_1: $(\mu_2 - \mu_1) > 0$

10

Test Statistic: A Summary of Evidence

- Test statistic: Summary of evidence (one number) in a sample to compare with the initial presumption
  - For inference about p, test statistic based on $\hat{P}$
    - For example, test $H_0: p = p_0$ versus $H_1: p \neq p_0$
    - Standardized test statistic, $z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
  - For inference about μ, test statistic based on $\bar{X}$
    - For example, test $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$

11

Coupon Example (p. 396)

- A report claims that 15 percent of printable Internet coupons are redeemed
  - A firm wants to show that its Internet coupons have a higher redemption rate
    - H_0: $p = 0.15$
    - H_1: $p > 0.15$
  - For a random sample of 3,000 broadcast coupons, 483 are redeemed: $\hat{P} = 0.161$ (= 483/3,000)
    - *Any* evidence in favor of the research hypothesis?

12

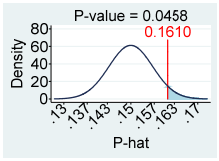
P-value Approach

- P-value approach: quantitatively measure how surprising the data are if H_0 were true
  - P-value: Probability of a test statistic at least as extreme as the one we got (in the direction of H_1) presuming H_0 is true
    - Small P-value: sampling error cannot plausibly explain why data differ so much from the initial presumption
    - P-value measures strength of evidence in favor of H_1: as it gets closer to 0 the strength of our evidence for H_1 increases (as data increasingly unlikely if H_0 were true)

13

$H_0: p = 0.15$ and $H_1: p > 0.15$

- Presuming H_0 were true $\hat{P} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$
  - For the values $p_0 = 0.15$ and $n = 3{,}000$, we have $\hat{P} \sim N(0.15, 0.0000425)$
- $\hat{P} = \frac{483}{3000} = 0.161$
  - P-value $= P(\hat{P} \geq 0.161) = P\left(Z \geq \frac{0.161-0.15}{\sqrt{0.0000425}}\right) = P(Z \geq 1.69) = 0.0458$



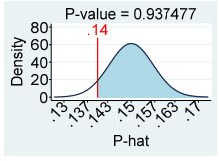
Observed sample is rather surprising if H_0 were true: we can reject the null and infer that the redemption rate is greater than 15%

14

When Evidence Contradicts H_1

- $H_0: p = 0.15$
- $H_1: p > 0.15$
- What if the evidence in the sample says the opposite of what the research hypothesis says? For example, $\hat{P} = \frac{420}{3{,}000} = 0.14$

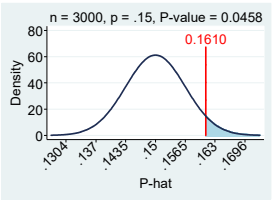
How about a P-value of 0.003?

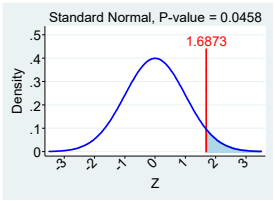


For one-tailed tests, P-values above 0.5 mean the evidence contradicts H_1: we have *no support* for H_1. In contrast, a P-value of 0.3 means there's some evidence but it's *insufficient support* for H_1.

15

Unstandardized vs. Standardized





$P(\hat{P} \geq 0.161 \mid H_0) = 0.0458$

$\hat{P}$ = 0.161 is the unstandardized test statistic

$P(Z \geq 1.6873) = 0.0458$

Z = 1.6873 is the standardized test statistic

16

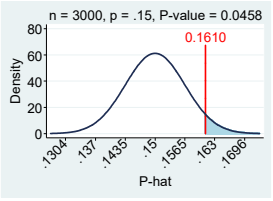
One Tailed v. Two Tailed Tests

- P-value: Probability of a test statistic at least as extreme *(in the direction of H_1)* as the one we got presuming that H_0 is true
- One-tailed test: P-value is area of one tail
  - If H_1 says > then upper tail area and if H_1 says < then lower tail area
- Two-tailed test: P-value is area of two tails
  - H_1 says ≠ so both tail areas (both directions)

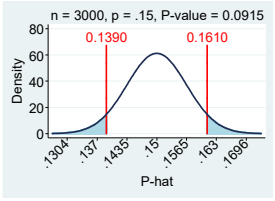
17

One Tailed vs. Two Tailed

$H_0: p = 0.15$
$H_1: p > 0.15$



$H_0: p = 0.15$
$H_1: p \neq 0.15$



In which case does the sample $\left(\hat{P} = \frac{483}{3000} = 0.161\right)$ provide stronger support for the research hypothesis?

18

CTV News: "Handgun ban supported by majority of Canadians: Nanos survey"

"A total ban on handgun ownership in Canada, exempting only police and security professionals, would enjoy significant support among Canadians, according to a new survey by Nanos Research.

The survey conducted for CTV News found that 48 per cent of Canadians would support such a ban, while another 19 per cent would somewhat support it.

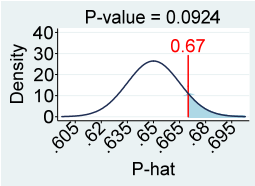
Twenty-one per cent of respondents said they would oppose a ban, and another 10 per cent said they would somewhat oppose it. Three per cent said they were unsure about their opinion.

[Results based] on a random survey of 1,000 Canadians, 18 years of age or older, between Aug. 25 and Aug. 27, 2018."

Published Sept., 2, 2018, https://www.ctvnews.ca/canada/handgun-ban-supported-by-majority-of-canadians-nanos-survey-1.4077763

Review P-value: Right-Tailed Test

- Can we prove over 65% support the handgun ban at least somewhat?
- $H_0: p = 0.65$
- $H_1: p > 0.65$
- $\hat{P} = 0.67, n = 1{,}000$

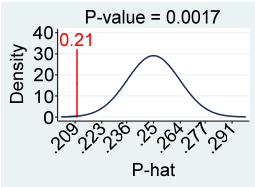


For a right-tailed test, P-value answers the question: "Presuming the null hypothesis were true, what is the chance sampling error could cause a sample proportion *as large as the one observed*?"

Do we have *any* evidence to support the research hypothesis?
How *strong* is the evidence to support the research hypothesis?

Review P-value: Left-Tailed Test

- Can we prove less than 25% oppose the handgun ban?
- $H_0: p = 0.25$
- $H_1: p < 0.25$
- $\hat{P} = 0.21, n = 1{,}000$

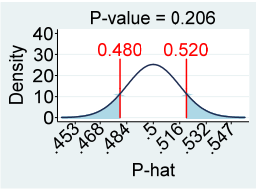


For a left-tailed test, P-value answers the question: "Presuming the null hypothesis were true, what is the chance sampling error could cause a sample proportion *as small as the one observed*?"

Do we have *any* evidence to support the research hypothesis?
How *strong* is the evidence to support the research hypothesis?

Review P-value: Two-Tailed Test

- Can we prove this coin is unfair?
- $H_0: p = 0.50$
- $H_1: p \neq 0.50$
- $\hat{P} = 0.480, n = 1{,}000$



For a two-tailed test, P-value answers the question: "Presuming the null hypothesis were true, what is the chance sampling error could cause a sample proportion *as far from null as the one observed*?"

How *strong* is the evidence to support the research hypothesis?

Missing Girls in Canada?

- "India's skewed sex ratio: An aversion to having daughters is leading to millions of missing girls" *The Economist*, April 2011
- "There has been much discussion about whether female feticide occurs in certain immigrant groups in Canada. We examined data on live births in Ontario and compared sex ratios in different groups according to the mother's country of birth and parity." Ray et al. (2012)

Links to sources: https://www.economist.com/asia/2011/04/07/seven-brothers and http://www.cmaj.ca/content/184/9/E492

Natural proportion boys born: 51.2%

- Indian-born mothers subset of complete data for 2002 – 07 for all births in ON:
  - 1st baby: 14,789 babies; 7,546 males; $\hat{P}$ = 0.510
  - 2nd baby: 13,076 babies; 6,873 males; $\hat{P}$ = 0.526
  - 3rd baby: 3,268 babies; 1,883 males; $\hat{P}$ = 0.576
- Chinese-born mothers subset of complete data for 2002 – 07 for all births in ON:
  - 1st baby: 12,339 babies; 6,429 males; $\hat{P}$ = 0.521
  - 2nd baby: 9,852 babies; 5,080 males; $\hat{P}$ = 0.516
  - 3rd baby: 1,403 babies; 715 males; $\hat{P}$ = 0.510

1st baby: parity = 0; 2nd baby: parity = 1; 3rd baby: parity = 2

H_0: $p = 0.512$; H_1: $p > 0.512$

Why choose a one-tailed hypothesis test?

- Parity = 0, Indian
  - n = 14,789; $\hat{P} = 0.510$
  - P-value = 0.665 (see *)
- Parity = 1, Indian
  - n = 13,076; $\hat{P} = 0.526$
  - P-value = 0.001
- Parity = 2, Indian
  - n = 3,268; $\hat{P} = 0.576$
  - P-value = 1.1×10^{-13}
- Parity = 0, Chinese
  - n = 12,339; $\hat{P} = 0.521$
  - P-value = 0.022
- Parity = 1, Chinese
  - n = 9,852; $\hat{P} = 0.516$
  - P-value = 0.235
- Parity = 2, Chinese
  - n = 1,403; $\hat{P} = 0.510$
  - P-value = 0.571

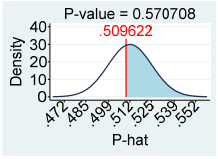
** For P-value accurate to 3rd decimal place, use $\hat{P} = \frac{7,546}{14,789}$ (not $\hat{P}$ =0.510).*

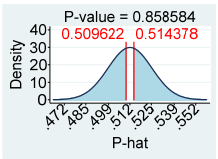
How to interpret these results?

25

Careful: P-value not always double

- Parity = 2, Chinese
  - n = 1,403; $\hat{P} = 0.509622$
  - For H_0: $p = 0.512$ versus H_1: $p > 0.512$ the P-value is 0.571
  - What would be the P-value for H_0: $p = 0.512$ versus H_1: $p \neq 0.512$?
  - It *cannot* be 2*0.571 because a P-value is a probability (never > 1)





26