Hypothesis testing with CLT (cont.)

Difference in proportions

10/30/24

Housekeeping

- No office hours Friday

Recap

- Hypothesis test for single proportion: $H_0 : p = p_0$ vs $H_A : p \neq p_0$ (or $>$, $<$)
  - Null distribution (assuming CLT holds under H_0):

$$\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

  - Obtain test statistic and calculate p-value

$$z = \frac{\hat{p}_{obs} - p_0}{\text{SE}_0} \sim N(0, 1)$$

Test of two proportions

Now suppose we have samples of binary outcomes from two populations.

Difference of two proportions

Suppose we have two populations 1 and 2, and want to conduct a hypothesis test for the difference in population proportions: $p_1 - p_2$

- We have samples of size n_1 and n_2
- Reasonable point estimate: $\hat{p}_{1,obs} - \hat{p}_{2,obs}$
- We will now work with the sampling distribution of the difference of two sample proportions

Sampling dist. of difference of two proportions

- In order to use CLT approximation, we have to ensure conditions are met:
  1. **Independence (extended)**: data are independent within *and* between groups
  2. **Success-failure (extended)**: success-failure conditions holds for ***both*** groups (must perform four total checks)
- If above hold, then:

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

where p_1 and p_2 are the population proportions

Hypothesis test for difference in proportions

1. Define hypotheses. Hypothesis tests for difference in proportions in this class will take the form:

$$
\begin{aligned}
H_0 &: \; p_1 = p_2 \\
H_A &: \; p_1 \neq p_2 \\
\text{or} \;\; & p_1 < p_2 \\
\text{or} \;\; & p_1 > p_2
\end{aligned}
$$

2. Set α and collect data/summarise (i.e. obtain $\hat{p}_{1,obs}$ and $\hat{p}_{2,obs}$)

Pooled proportion

- To verify success-failure conditions, need to know p_1 and p_2
  - We don't have these values, so maybe use $\hat{p}_{1,obs}$ and $\hat{p}_{2,obs}$?
- But wait! If $H_0 : p_1 = p_2$, then $\hat{p}_{1,obs}$ and $\hat{p}_{2,obs}$ come from the ***same*** population
- So under this null, we use a special proportion called the **pooled proportion**:

$$\hat{p}_{pooled} = \frac{\text{total \# of successes from both samples}}{\text{combined sample size}}$$

- This is the best estimate of both p_1 and p_2 if $H_0 : p_1 = p_2$ *is true*!
- For this reason, **use $\hat{p}_{pooled}$ to verify success-failure conditions**

Hypothesis test (cont.)

3. Obtain null distribution

- If conditions satisfied, then we know the sampling distribution of $\hat{p}_1 - \hat{p}_2$
- To obtain the **null distribution** we assume $H_0 : p_1 - p_2 = 0$ is true and we $\hat{p}_{pooled}$ to estimate p_1 and p_2 to approximate standard error:

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) \qquad \text{(CLT)}$$

$$\dot{\sim} N\Big(0, \underbrace{\sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_2}}}_{\widehat{\text{SE}}_0}\Big) \qquad (H_0)$$

Hypothesis test (cont.)

Obtain test-statistic:

$$z = \frac{\text{point estimate} - \text{null value}}{\text{SE}} \approx \frac{(\hat{p}_{1,obs} - \hat{p}_{2,obs}) - 0}{\widehat{\text{SE}}_0}$$

- To obtain p-value, we want $\Pr(Z \geq z)$ and/or $\Pr(Z \leq z)$ where $Z \sim N(0, 1)$
  - Obtain using `pnorm(z, 0, 1)`

Example: offshore drilling

A survey asked 827 randomly sampled registered voters in California: Do you support or oppose about drilling for oil and natural gas of the Coast of California? Or do you now know enough to say? We have the following distribution of responses separated by whether the respondent graduated from college:

position	no	yes	total
do_not_know	131	104	235
oppose	126	180	306
support	132	154	286
total	389	438	827

- Do the data provide strong evidence at the 0.05 level that the proportion of college graduates who support off-shore drilling in California is different than that of non-college graduates?

Example: offshore drilling (cont.)

Do the data provide strong evidence at the 0.05 level that the proportion of college graduates who support off-shore drilling in California is different than that of non-college graduates?

- Define parameters and hypotheses
  - Let p_c be the proportion of registered voters from California who are college-graduates who support off-shore drilling
  - Let p_{nc} be the proportion be of registered voters from California who are not college-graduates who support off-shore drilling
  - $H_0 : p_{nc} - p_c = 0$ and $H_A : p_{nc} - p_c \neq 0$

Example: offshore drilling (cont.)

Obtain observed proportions and pooled proportion.

position	no	yes	total
do_not_know	131	104	235
oppose	126	180	306
support	132	154	286
total	389	438	827

- $\hat{p}_{nc,obs} = \frac{132}{389} = 0.339$
- $\hat{p}_{c,obs} = \frac{154}{438} = 0.352$
- $\hat{p}_{pooled} = \frac{132+154}{389+438} = \frac{286}{827} = 0.346$

Check conditions for inference are satisfied.

Example: offshore drilling (cont.)

- Conditions for inference:
  - Independence: random sample
  - Success-failure:
    - $n_{nc}\hat{p}_{pooled} = 389 \times 0.346 = 134.59 > 10$
    - $n_{nc}(1 - \hat{p}_{pooled}) = 389 \times (1 - 0.346) = 254.41 > 10$
    - $n_c\hat{p}_{pooled} = 438 \times 0.346 = 151.55 > 10$
    - $n_c(1 - \hat{p}_{pooled}) = 438 \times (1 - 0.346) = 286.45 > 10$
- Since conditions are met, we can proceed

Example: offshore drilling (cont.)

- Find the null distribution for $\hat{p}_{nc} - \hat{p}_c$

$$\hat{p}_{nc} - \hat{p}_c \dot{\sim} N\left(0, \sqrt{\frac{0.346(1-0.346)}{389} + \frac{0.346(1-0.346)}{438}} = 0.033\right)$$

- Set up calculation for test statistic

$$z = \frac{(\hat{p}_{nc,obs} - \hat{p}_{c,obs}) - 0}{\text{SE}_0} = \frac{(0.339 - 0.352) - 0}{0.033} = -0.394$$

- Draw picture and write code for p-value

Example: offshore drilling (cont.)

p-value calculation:

- $\Pr(Z \leq z) + \Pr(Z \geq -z) = 2 \times \Pr(Z \geq 0.394)$
- `2 * (1 - pnorm(0.394))` = 0.694

Make a decision and conclusion in context.

- Since our p-value is greater the 0.05, we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling among California voters.

Test for a single mean

Hypotheses and null distribution

Want to conduct a hypothesis test for the mean μ of a population.

- Hypotheses: $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ (or $\mu > \mu_0$ or $\mu < \mu_0$)
- Verify conditions for CLT
    1. Independence
    2. Approximate normality or large sample size
- If conditions satisfied, the CLT under H_0 gives us **null distribution** for $\bar{X}$:

$$\bar{X} \dot{\sim} N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

z-test and *t*-test

- If σ known, we perform a **z-test** where our test-statistic is:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

and we obtain our p-value using `pnorm()`

- If σ unknown, we perform a ***t*-test** by estimating σ with s. Our test statistic is:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{df} \qquad df = n - 1$$

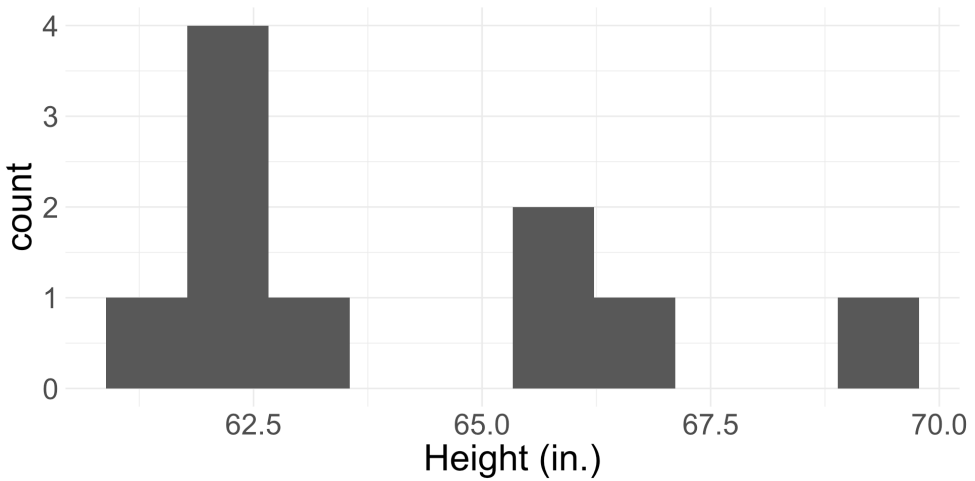
and we obtain our p-value using `pt()`

Example: height

- In the US, the average height for women is 5'3.5" or 63.5 inches
- Let's conduct a hypothesis test to see if the average height of female-identifying students in STAT 201 is equal to national average.
- Define parameters and hypotheses
- I took a random sample of 10 female-identifying students across both sections. Set $\alpha = 0.10$

Example: height (cont.)

n	mean	sd
10	64	2.748737



- Conditions:
  - Independence: random sample
  - Approximate normality: $n = 10 < 30$, but no clear outliers
- Since we don't know σ, we perform a t-test:
  - $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{64 - 63.5}{2.749/\sqrt{10}} = 0.575$

- Are conditions for inference met?
- If so, what test (z-test or t-test) should we perform? What is our test-statistic?

Example: height (cont.)

- Draw a picture and write code to find our p-value
  - $df = n - 1 = 9$
  - p-value is $\Pr(T \geq 0.575) + \Pr(T \leq -0.575) = 2 \times \Pr(T \geq 0.575)$ where $T \sim t_9$
  - $2 \times$ (1-`pt(0.575, 9)`) = 0.5793797
- Make a decision and conclusion in context
  - Since our p-value is greater than 0.01, we fail to reject H_0. The data do not provide sufficient evidence to suggest that the average height of female-identifying students in STAT 201 is different from national average.