Hypothesis testing with CLT

10/28/24

Housekeeping

- Homework 6 due tonight
- Modified office hours

Recap

- CLT -> sampling distribution for sample means -> confidence intervals for populations means
- Now we're returning to hypothesis testing!
  1. Two sets of hypotheses (competing claims)
  2. Collect data, calculate a statistic from the observed statistic, set significance level
  3. Obtain p-value from the null distribution: sampling distribution assuming if H_0 were true
     - p-value: probability of observing data as or more extreme as our own, assuming H_0 true
  4. Make a decision

Hypothesis testing using mathematical model

- We learned how to conduct hypothesis tests (HTs) using simulation to obtain null distribution
- But we can also use CLT to obtain null distribution!
- So the only step that will "look different" is #3: how we obtain our null distribution and p-value
  - Looks different depending on type of data
- Make a conclusion in terms of H_A

Hypothesis test for single proportion

1. Define hypotheses

Want to conduct a hypothesis test about a population proportion.

- $H_0 : p = p_0$
- $H_A : p \neq p_0$ or
  $H_A : p > p_0$ or
  $H_A : p < p_0$

- $H_0 : p \geq p_0$
- $H_A : p < p_0$

- $H_0 : p \leq p_0$
- $H_A : p > p_0$

- Remember, p_0 is our "null hypothesized value": the population proportion if H_0 were true

2. Collect data, set significance

- Obtain observed sample proportion $\hat{p}_{obs}$
- Set α significance level

3. Null distribution and p-value

Recall CLT for sample proportion: if we have n independent binary observations that satisfy the success-failure condition, then

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

- This is the *sampling distribution* of $\hat{p}$
- But we want the *null distribution* of $\hat{p}$: the sampling distribution under H_0
- We should operate in a world where H_0 is true, which means we operate assuming $p = p_0$
- So to use CLT, we must satisfy:
  - Independence
  - Success-failure condition under H_0: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

3. Null distribution and p-value (cont.)

If CLT holds and H_0 is true, then our **null distribution** is:

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

- We can standardize the null distribution by taking z-score:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1)$$

3. Test statistic

- p-value requires us to compare our observed data to the null distribution
- Calculate a **test statistic**: a quantity that assesses how consistent your sample data are with H_0
  - Our test statistic is of the form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE}}$$

  - For this specific test, our test statistic is:

$$z = \frac{\hat{p}_{\text{obs}} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  which is distributed $N(0, 1)$

Obtain p-value

- If $|z|$ large, then that usually means observed value is extremely unusual for H_0, which is convincing evidence against H_0
- p-value is then $\Pr(Z \geq z)$ or $\Pr(Z \leq z)$ (or both), depending on H_A
  - Easily obtained using `pnorm()`

Example: taste test

Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanted to test this claim using a hypothesis test at the 0.05 significance level.

- He randomly sampled 80 such people.
- He then filled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular.
- 53 participants correctly identified the soda.

Let p be the probability/rate of correctly identifying soda type.

Example: taste test (cont.)

1. **Define hypotheses**
    - $H_0: p = 0.5$ (random guessing)
    - $H_A: p > 0.5$ (better than random guessing)
    - Note: $p_0 = 0.5$ is our null hypothesized value!
2. **Collect data**
    - $\hat{p}_{\text{obs}} = \frac{53}{80} = 0.6625$

- Note: significance level already determined to be 0.05

Example: taste test (cont.)

3. **Obtain null distribution and p-value**

    i. Check conditions for inference satisfied

    - Independence: random sample
    - success-failure: $np_0 = 80(0.5) = 40 \geq 10$ and $n(1 - p_0) = 40 \geq 10$

    ii. Null distribution

$$\hat{p} \dot{\sim} N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{80}} = 0.056\right)$$

    iii. Test statistic:

$$z = \frac{\hat{p}_{obs} - p_0}{\text{SE}_0} = \frac{0.6625 - 0.5}{0.056} = 2.90$$

    - i.e. if H_0 true, our observed $\hat{p}_{obs}$ is 2.90 SDs above the mean

4. Example: taste test (cont.)

iv. Calculate p-value

- Remember $H_A: p > 0.5$

$$\text{p-value} = \Pr(Z \geq z) = \Pr(Z \geq 2.90) = \texttt{1 - pnorm(2.90, 0, 1) = 0.0019}$$

4. **Decision and conclusion**

- Since our p-value of 0.0019 is less than our significance level of 0.05, we reject H_0. The data provide strong evidence that the rate of correctly identifying a soda for these people is better than random guessing.

Example: M&M's

M&M's reported that 14% of its candies are yellow. We are interested in testing this claim. In a random sample of 100 M&M's, 9 were found to be yellow. Conduct a hypothesis test at the 0.10 level.

$p =$ true proportion of yellow M&M's

1. Write out null and alternative hypotheses
2. Collect data (i.e. obtain our observed statistics)
3. i) Verify conditions for CLT are met

1. $H_0 : p = 0.14$ versus $H_A : p \neq 0.14$
2. $\hat{p}_{obs} = \frac{9}{100} = 0.09$
3. i) Independence: random sample

   Success-failure:
   $np_0 = 100(0.14) = 14 \geq 10$ and
   $n(1 - p_0) = 86 \geq 10$

Example: M&M's (cont.)

3. ii) Obtain null distribution

   iii) Obtain test statistic z

3. ii) By CLT, our null distribution is

$$\hat{p} \dot{\sim} N\left(0.14, \sqrt{\frac{0.14(1-0.14)}{100}}\right)$$

$$= N(0.14, 0.035)$$

iii) $z = \frac{\hat{p}_{obs} - p_0}{\text{SE}_0} = \frac{0.09 - 0.14}{0.035} = -1.43$

3. iv) Obtain p-value. Write out in Pr() notation or in code what we want to find. Drawing a picture may help!

3. iv) Since H_A is two-sided, we want

$$\begin{aligned} \text{p-value} &= \Pr(Z \leq -1.43 \cup Z \geq 1.43) \\ &= \Pr(Z \leq -1.43) + \Pr(Z \geq 1.43) \\ &= 2 \times \Pr(Z \geq 1.43) \\ &= 2 \texttt{ * (1 - pnorm(1.43))} \\ &= 0.153 \end{aligned}$$

Example: M&M's (cont.)

4. Make a decision and conclusion in context.
    - Since our p-value of 0.153 is greater than our significance level of 0.10, we fail to reject H_0. The data are not strong enough to suggest that the true proportion of yellow M&Ms is different from 14%.

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 wand 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: $\hat{p}_{1,obs} - \hat{p}_{2,obs}$
- Apply same process as in previous test for a single proportion, this time working 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:

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

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:

$$
\begin{aligned}
H_0 &: \ p_1 = p_2 && \Rightarrow && H_0 : p_1 - p_2 = 0 \\
H_A &: \ p_1 \neq p_2 && \Rightarrow && H_A : p_1 - p_2 \neq 0 \\
&\text{or } \ p_1 < p_2 && \Rightarrow && p_1 - p_2 < 0 \\
&\text{or } \ p_1 > p_2 && \Rightarrow && p_1 - p_2 > 0
\end{aligned}
$$

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

Pooled proportion

- To obtain null distribution, we would need to know p_1 and p_2 to verify success-failure conditions
- We obviously 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 $\hat{p}_{2,obs}$ in theory come from the *same* population
  - So under this null hypothesis, we use a special proportion called the **pooled proportion** to check the success-failure conditions:

$$\hat{p}_{pooled} = \frac{\text{total \# of successes from both samples}}{\text{combined sample size}} = \frac{n_1\hat{p}_{1,obs} + n_2\hat{p}_{2,obs}}{n_1 + n_2}$$

- This is the best estimate of both p_1 and p_2 *if null hypothesis of $p_1 = p_2$ is true*!

Hypothesis test (cont.)

3. Obtain null distribution (first verify conditions for inference using $\hat{p}_{pooled}$)
    - If conditions satisfied, then we have the "general" sampling distribution of $\hat{p}_1 - \hat{p}_2$ from previous slide.
    - But to obtain the **null distribution**, we assume $H_0: p_1 - p_2 = 0$ is true, and will estimate p_1 and p_2 using $\hat{p}_{pooled}$ to approximate standard error:

$$
\begin{aligned}
\hat{p}_1 - \hat{p}_2 &\dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) && \text{(CLT)} \\
&\dot{\sim} N\big(0, \underbrace{\sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_2}}}_{\widehat{\text{SE}}_0}\big) && (H_0)
\end{aligned}
$$

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)`