Statistics 305: Introduction to Biostatistical Methods for Health Sciences

Chapter 14 : Inference for Proportions

Jinko Graham

2018-08-23

Inference for Proportions (Chapter 14)

- Instead of quantitative measurements, we classify each sampled individual into one of two categories.
  - success, failure (canonical)
  - breast cancer, no breast cancer
  - Etc.
- Want to make inference about the proportion p of successes in a population, or about the difference between the proportions p_1 and p_2 of successes in two populations.

Example: Women's Health Initiative (WHI)

- A randomized controlled trial, called the Women's Health Initiative, randomized 16,608 post-menopausal women aged 50-79 years to receive either hormone replacement therapy in the form of estrogen plus progestin (EP; $n_1 = 8506$), or a placebo ($n_2 = 8102$).
- After five years, 166 of those in the EP group had developed invasive breast cancer, compared to 122 in the placebo group.
- The populations to compare are postmenopausal women aged 50-79 years at the start of the trial, who are taking EP (population 1) or placebo (population 2).
- The estimated populaton proportions are $\hat{p}_1 = 166/8506 = 0.0195$ and $\hat{p}_2 = 122/8102 = 0.0151$
- It looks like the EP group has a higher risk of breast cancer, but could this difference be due to chance?

Outline of Approach

- Similar approach to inference as inference of population means (quantitative data), with some minor differences.
- Inference is based on the sampling distribution of $\hat{p}_1 - \hat{p}_2$
- Two-stage reasoning:
  - 1. Transform $\hat{p}_1 - \hat{p}_2$ into an initial pivotal quantity, Z_1, whose denominator depends on the unknown p_i's
  - 2. Get a final pivotal quantity, Z, by replacing the unknown p_i's in the denominator of Z_1 with estimates.
- Confidence intervals and hypothesis tests follow from the approximate sampling distribution of the final pivotal quantity, Z.
- Note: We will not cover *Inference for a single proportion* in the text (Sections 14.2 – 14.5).

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

- Assume that we have independent simple random samples (SRSs) of size n_1 and n_2, from the two parent populations.
- Then the distribution of $\hat{p}_1 - \hat{p}_2$ has
  - mean $p_1 - p_2$ and
  - **approximate** SD

$$\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}. \tag{1}$$

- If the sample sizes are “large”, the CLT tells us that the shape of this distribution is approximately normal.
  - (The CLT applies because it turns out that a proportion is an average ... of 0’s and 1’s.)

Initial Transformation

- For large samples, $\hat{p}_1 - \hat{p}_2$ is **approximately** normally distributed with
  - mean $p_1 - p_2$, and
  - SD **approximated** by

    $$\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

    in equation (1).
- So the distribution of

  $$Z_1 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

  is **approximately** $N(0,1)$.

Z_1 with Estimated SDs

- Inserting the estimates $\hat{p}_1$ and $\hat{p}_2$ for the unknown parameters p_1 and p_2 into Z_1 above gives

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$$

- What is the distribution of Z?
    - Under certain conditions (see below) it is reasonable to assume that Z is approximately normal.

Rules of Thumb: SD Approximation and Normal Approximation

- The formula for approximating the SD of $\hat{p}_1 - \hat{p}_2$ in equation (1) on pg 5 is considered accurate when the population is at least 10 times as large as the sample.
- The normal approximation to the distribution of Z is considered reliable when the underlying p_j's are moderate (i.e. not too close to 0 or 1)
  - Rule: Use the normal approximation only when there are at least 5 successes *and* 5 failures in both sample 1 and sample 2 (some texts recommend ten, rather than five).
  - Rationale: p_1 close to 0 or 1 will lead to almost no successes or failures, respectively, in sample 1. Same story for p_2.
- In a nutshell: Should have
  1. Population at least 10 times as large as the sample, and
  2. At least 5 successes *and* 5 failures in both sample 1 and sample 2 (for Z to be approximately normal and the statistical inference to be valid).

Checking rules of thumb for WHI Data

1. The population is at least 10 times larger than the sample of 16,608 used in the study. True.
2. At least 5 successes and failures in both samples.
   - True: 166 cancer, 8340 cancer-free in the EP group; 122 cancer, 7980 cancer-free in the placebo group.

Confidence Intervals

- The level-C CI for $p_1 - p_2$ is of the form:

$$\text{estimate} \pm \text{margin of error}$$

- The estimate is $\hat{p}_1 - \hat{p}_2$
- The margin of error is $z^* \times SE$ where
  - z^* is the upper $(1 - C)/2$ critical value of the standard normal distribution.
  - SE is the estimated SD of $\hat{p}_1 - \hat{p}_2$ in the denominator of Z; namely, $SE = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$

Example (WHI)

- Recall: 16,608 women aged 50-79 years randomized to receive either estrogen plus progestin (EP; $n_1 = 8506$), or a placebo ($n_2 = 8102$). After five years, 166 in the EP group developed invasive breast cancer, compared to 122 in placebo group.
- For EP, $\hat{p}_1 = 166/8506$ and, for placebo, $\hat{p}_2 = 122/8102$.

- 95% CI is estimate $\pm$ margin of error, where
  - estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = 0.0044$
  - margin of error is a critical value times standard error of difference.
- The critical value is 1.96 (see R demo).
- The standard error is $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = 0.002$
- The margin of error is therefore $1.96 * .002 = 0.00392$.
- Putting it all together, the CI is 0.0044 ± 0.00392 or approximately $(0.0005, 0.008)$.

Test Statistic

- The null hypothesis is $H_0 : p_1 - p_2 = 0$.
- Numerator of the test statistic is therefore the estimated difference $(\hat{p}_1 - \hat{p}_2)$ minus 0.
- Denominator of the test statistic is

$$SE = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$$

- BUT, under H_0, we have $p_1 = p_2$. Call this common value p.
- Assuming a common proportion p in the two populations, we pool the 2 samples to obtain an estimate $\hat{p}$; i.e.,

$$\hat{p} = (\text{number of cancers in both samples})/(n_1 + n_2).$$

- The formula for the SE of $\hat{p}_1 - \hat{p}_2$ simplifies to $\sqrt{\hat{p}(1 - \hat{p}) \times (1/n_1 + 1/n_2)}$
- So the statistic for testing $H_0 : p_1 - p_2 = 0$ is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})\,(1/n_1 + 1/n_2)}}$$

Example (WHI, continued)

- $H_0 : p_1 - p_2 = 0$.
- Test statistic numerator is $\hat{p}_1 - \hat{p}_2 = 0.0044$.
- Test statistic denominator is the SE based on the pooled estimate of $p = p_1 = p_2$.
  - Pooled estimate of the common population proportion is $\hat{p} = (166 + 122)/(8506 + 8102) = 0.0173$.
  - So the SE is

$$\begin{aligned} & \sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)} \\ = \quad & \sqrt{0.0173(1-0.0173)\,(1/8506 + 1/8102)} \\ = \quad & 0.002. \end{aligned}$$

- The test statistic value is $z = 0.0044/0.002 = 2.2$.

p-value

- The p-value is the chance of a value of the test statistic that is as or more extreme than what we did observe in our data, when the null hypothesis is true.
- Same logic as we saw earlier for inference of population means (Chapter 11).
- Let Z be a standard normal random variable and z be the observed value of the test statistic. Then:
  - For $H_a : p_1 - p_2 \neq 0$, p-value $= 2P(Z \geq |z|)$.
  - For $H_a : p_1 - p_2 > 0$, p-value $= P(Z \geq z)$.
  - For $H_a : p_1 - p_2 < 0$, p-value $= P(Z \leq z)$.

Example (WHI, continued)

- Suppose we wish to test
  H_0: $p_1 - p_2 = 0$ vs. H_a: $p_1 - p_2 \neq 0$ at level $\alpha = 0.05$.
- For an observed value of the test statistic $z = 2.2$, computer software calculates a pvalue of about 0.03 (see R demo)
- We therefore reject H_0 at the 5% level: There is statistical evidence that women taking EP have a higher risk of invasive breast cancer than those taking the placebo.
  - Note: We say that the EP group has higher risk than the placebo group because $\hat{p}_1$ (for EP) is greater than $\hat{p}_2$ (for placebo), as evidenced by $z = 2.2$ being greater than zero.

Summary

- Inference for the difference $p_1 - p_2$ between two population proportions is based on a pivotal quantity.
- Confidence intervals are of the form **estimate** $\pm$ **margin of error**, where
  - estimate is difference between sample means, and
  - margin of error is a critical value times the standard error of the difference in sample proportions
- To test the null hypothesis $H_0 : p_1 - p_2 = 0$ against an alternative H_a we calculate a test statistic and the p-value
  - The p-value is the chance of seeing a value of the test statistic as or more extreme than the value that was observed in our data, under the null hypothesis.
  - Compare the p-value to a significance level α to obtain a statistical hypothesis test
- The statistical inference is considered reliable when:
  - the population is at least $10\times$ larger than the size of the sample
  - there are sufficient numbers of successes and failures in each sample (at least 5 successes and failures in each sample).