STA13: Elementary Statistics

Lecture 20
Book sections 7.4, 8.1-8.3

Dmitriy Izyumin

March 07 2018

Two Proportions

Just like with means, we sometimes want to compare the proportions of two different populations.

- Is there a difference between the proportions of successes in the populations?
- Are successes more likely in one population than in another?

Independent Samples

Examples:

- Is the proportion of Democrats the same for voters aged 18-25, as for voters older than 65?
- Is a certain vaccine more effective (proportion of successes) in children than in adults?
- Is the unemployment rate (proportion of unemployed members of the work force) the same for two different cities?

Notation

- Population 1 has proportion p_1.
- Sample 1 is taken from population 1, and has size n_1, and sample proportion $\hat{p}_1$.
- Population 2 has proportion p_2.
- Sample 2 is taken from population 2, and has size n_2, and sample proportion $\hat{p}_2$.

Parameter

$$p_1 - p_2$$

- p_1 is the proportion of successes in the first population.
- p_2 is the proportion of successes in the second population.
- We want to make inferences about $p_1 - p_2$, the difference in proportions between the two populations.

Statistic

$$\hat{p}_1 - \hat{p}_2$$

- $\hat{p}_1$ is the sample proportion of the first sample.
- $\hat{p}_2$ is the sample proportion of the second sample.
- $\hat{p}_1 - \hat{p}_2$ is a statistic and has a sampling distribution.
- We use the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to make inferences about $p_1 - p_2$.

Sampling Distribution

Properties of the Sampling Distribution of $(\hat{p}_1 - \hat{p}_2)$

1. The mean of the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is $(p_1 - p_2)$; that is,

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

   Thus, $(\hat{p}_1 - \hat{p}_2)$ is an unbiased estimator of $(p_1 - p_2)$.

2. The standard deviation of the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

3. If the sample sizes n_1 and n_2 are large (see Section 5.4 for a guideline), the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is approximately normal.

- $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$
- Need to check:
  $n_1\hat{p}_1 > 15$, $n_1(1-\hat{p}_1) > 15$, $n_2\hat{p}_2 > 15$, $n_2(1-\hat{p}_2) > 15$

Large Sample - CI

Large-Sample 100(1 − α)% Confidence Interval for $(p_1 - p_2)$: Normal (z) Statistic

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$
$$\approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

- $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$
- $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$

Large Sample - HT

Large-Sample Test of Hypothesis about $(p_1 - p_2)$: Normal (z) Statistic

One-Tailed Test

H_0: $(p_1 - p_2) = 0$*

H_a: $(p_1 - p_2) < 0$

[or H_a: $(p_1 - p_2) > 0$]

Two-Tailed Test

H_0: $(p_1 - p_2) = 0$

H_a: $(p_1 - p_2) \neq 0$

$$\textit{Test statistic: } z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

Rejection region: $z < -z_\alpha$

[or $z > z_\alpha$ when H_a: $(p_1 - p_2) > 0$]

Rejection region: $|z| > z_{\alpha/2}$

$$\textit{Note: } \sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \approx \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- We will use $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ as the estimate of the standard error.

Choosing the Sample Size

As before, we often want to plan ahead, and figure out the minimal sample sizes that would allow us to estimate the parameter to within a certain margin of error with a certain confidence.

- When two sample sizes are needed (n_1 and n_2), assume they are equal.
- Always round up to the nearest integer.

Difference of Means (Paired Differences)

Suppose you want to estimate μ_d to within some margin of error ME with $100(1-\alpha)\%$ confidence. To find the minimal sample sizes:

- Solve $n_d = \frac{(z_{\alpha/2})^2 \sigma_d^2}{ME^2}$
- Use prior information for values of σ_d^2 if possible.
- If no other info is available, use s_d^2 from a prior sample.

Difference of Means (Independent Samples)

Suppose you want to estimate $\mu_1 - \mu_2$ to within some margin of error ME with $100(1-\alpha)\%$ confidence. To find the minimal sample sizes:

- Use equal sample sizes $n_1 = n_2$
- Solve $n_1 = n_2 = \frac{(z_{\alpha/2})^2(\sigma_1^2+\sigma_2^2)}{ME^2}$
- Use prior information for values of σ_1^2 and σ_2^2 if possible.
- If no other info is available, use s_1^2 and s_2^2 from a prior sample.

Difference of Proportions

Suppose you want to estimate $p_1 - p_2$ to within some margin of error ME with $100(1-\alpha)\%$ confidence. To find the minimal sample sizes:

- Use equal sample sizes $n_1 = n_2$
- Solve $n_1 = n_2 = \frac{(z_{\alpha/2})^2(p_1(1-p_1)+p_2(1-p_2))}{ME^2}$
- Use prior information for values of p_1 and p_2 if possible.
- If no prior info is available, use $p_1 = p_2 = 0.5$.