Chapter 8: Comparing two independent populations

Duzhe Wang

Part 3
https://dzwang91.github.io/stat371/

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Review of two-sample T-test

- The data consists of separate simple random samples from two independent populations, label them 1 and 2. Let
  - μ_1 = true mean of population 1
  - μ_2 = true mean of population 2
  - n_1 = sample size taken from population 1
  - n_2 = sample size taken from population 2
  - σ_1^2 = true variance of population 1
  - σ_2^2 = true variance of population 2
- We wish to test: $H_0 : \mu_1 - \mu_2 = \delta$ vs. $H_A : \mu_1 - \mu_2 \neq \delta$.
- Good numerical and graphical summaries to explore the data might include means, medians, standard deviations, QQ plot, side-by-side boxplots, side-by-side dotplots, histograms.

Review of two-sample T-test continued

- After exploring the data, if we are willing to assume:
  - All of the data points are independent, both within and between populations
  - The two populations each follow normal distributions
  - The variances of the two populations are equal so that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

  then the test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  Where:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

- Compute the p-value using T distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom and then make a conclusion using given α.

Review of Welch T-test

- If $\sigma_1 \neq \sigma_2$, the test statistic: $t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, with the degrees of freedom:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

  and round down.

Part 3: comparing two independent population proportions

Example

- "Does handedness differ according to sex?"
- A sample of $n_M = 54$ males and $n_F = 21$ females was taken, and each person was asked to indicate which was their dominant hand.
- The data are as follows:

Female: 12 left, 9 right
Male: 23 left, 31 right

Example

- "Does handedness differ according to sex?"
- A sample of $n_M = 54$ males and $n_F = 21$ females was taken, and each person was asked to indicate which was their dominant hand.
- The data are as follows:

  Female: 12 left, 9 right
  Male: 23 left, 31 right

- If we let π_{FL} be the proportion of females that are left-handed, and π_{ML} be the proportion of males that are left-handed, then our hypotheses could be expressed as:

$$H_0 : \pi_{FL} - \pi_{ML} = 0$$
$$H_A : \pi_{FL} - \pi_{ML} \neq 0$$

Example continued

What test statistic do we choose?

Example continued

What test statistic do we choose?

- Let P_{FL} be the sample proportion of females that are left-handed, and P_{ML} be the sample proportion of males that are left-handed.
- When $\pi_{FL}n_F, (1-\pi_{FL})n_F, \pi_{ML}n_M, (1-\pi_{ML})n_M$ are all greater than 5,

$$P_{FL} \dot{\sim} N(\pi_{FL}, \frac{\pi_{FL}(1-\pi_{FL})}{n_F})$$

$$P_{ML} \dot{\sim} N(\pi_{ML}, \frac{\pi_{ML}(1-\pi_{ML})}{n_M})$$

Example continued

What test statistic do we choose?

- Let P_{FL} be the sample proportion of females that are left-handed, and P_{ML} be the sample proportion of males that are left-handed.
- When $\pi_{FL} n_F, (1 - \pi_{FL}) n_F, \pi_{ML} n_M, (1 - \pi_{ML}) n_M$ are all greater than 5,

$$P_{FL} \dot{\sim} N(\pi_{FL}, \frac{\pi_{FL}(1-\pi_{FL})}{n_F})$$

$$P_{ML} \dot{\sim} N(\pi_{ML}, \frac{\pi_{ML}(1-\pi_{ML})}{n_M})$$

- Therefore,

$$P_{FL} - P_{ML} \dot{\sim} N(\pi_{FL} - \pi_{ML}, \ \frac{\pi_{FL}(1-\pi_{FL})}{n_F} + \frac{\pi_{ML}(1-\pi_{ML})}{n_M}).$$

Example continued

- Under H_0, $\pi_{FL} = \pi_{ML} = \pi_L$. Then when $\pi_L n_F, (1-\pi_L)n_F, \pi_L n_M, (1-\pi_L)n_M$ are all greater than 5,
$$P_{FL} - P_{ML} \dot{\sim} N(0,\ \pi_L(1-\pi_L)(\tfrac{1}{n_F} + \tfrac{1}{n_M}).$$

Example continued

- Under H_0, $\pi_{FL} = \pi_{ML} = \pi_L$. Then when $\pi_L n_F, (1-\pi_L)n_F, \pi_L n_M, (1-\pi_L)n_M$ are all greater than 5,
$$P_{FL} - P_{ML} \dot{\sim} N(0,\ \pi_L(1-\pi_L)(\tfrac{1}{n_F} + \tfrac{1}{n_M})).$$
- π_L is unknown, but can be estimated using a weighted average of the two individual sample proportions:
$$P_L = \tfrac{P_{FL} n_F + P_{ML} n_M}{n_F + n_M}.$$

Example continued

- Under H_0, $\pi_{FL} = \pi_{ML} = \pi_L$. Then when $\pi_L n_F, (1-\pi_L)n_F, \pi_L n_M, (1-\pi_L)n_M$ are all greater than 5,
$$P_{FL} - P_{ML} \dot{\sim} N(0,\ \pi_L(1-\pi_L)(\frac{1}{n_F} + \frac{1}{n_M})).$$
- π_L is unknown, but can be estimated using a weighted average of the two individual sample proportions:
$$P_L = \frac{P_{FL}n_F + P_{ML}n_M}{n_F + n_M}.$$
- Then the test statistic is given in standardized form:
$$Z = \frac{P_{FL} - P_{ML}}{\sqrt{P_L(1-P_L)(\frac{1}{n_F} + \frac{1}{n_M})}} \dot{\sim} N(0,1).$$

Example continued

- $P_{FL} = 0.571$, $P_{ML} = 0.426$, and $P_L = \frac{12+23}{21+54} = 0.467$.
- The realization of test statistic:

$$Z_{obs} = \frac{0.571 - 0.426 - 0}{\sqrt{0.467(1-0.467)(\frac{1}{21}+\frac{1}{54})}} = 1.13.$$

- p-value $= 2 \times P(Z > 1.13) = 0.258$. If $\alpha = 0.05$, we conclude that there is not enough evidence to say that males and females have a different proportion of left-handed individuals.

Recap

- The data consists of separate samples from two populations, label them 1 and 2. Let:

  π_1 = true proportion in population 1

  π_2 = true proportion in population 2

  n_1 = sample size taken from population 1

  n_2 = sample size taken from population 2

- We wish to test: $H_0 : \pi_1 - \pi_2 = 0$ vs. $H_A : \pi_1 - \pi_2 \neq 0$.
- When the null is true, $\pi_1 = \pi_2 = \pi$. The unknown π can be estimated using a weighted average of the two individual sample proportions:

  $$P = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2}.$$

  where P_1 and P_2 are the sample proportions as computed from the two samples.

Recap continued

- If based on our prior knowledge we are willing to assume:
  - All of the data points are independent, both within and between populations
  - The sample sizes are large enough (πn_1, $(1-\pi)n_1$, πn_2, and $(1-\pi)n_2$ are all greater than 5).
- Then the test statistic is:

$$Z = \frac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1} + \frac{1}{n_2})}} \dot{\sim} N(0,1).$$

- Calculate the p-value and compare to the given significance level α.

What's the next?

We'll discuss how to compare two paired populations in Chapter 9.