Lecture 8.3

Hypothesis testing

(continued...)

Hypothesis testing (two populations)

Thus far, we only looked at hypothesis-testing procedures for a parameter of a single population.

We now extend these testing methods to situations involving the parameters of two different population distributions.

For this, we make the same independent assumptions:

1. $X_1, \ldots, X_{n_X}$ are iid,
2. $Y_1, \ldots, Y_{n_X}$ are iid, and
3. All X and Y samples are independent of one another.

test statistic

Again as in the case of single population, when testing H_0 about parameter(s) for most problems in this course, the test statistic will take the form

$$\text{Test Statistic} = \frac{\text{Estimator} - \text{Hypothesized Value}}{\text{SE of Estimator}}.$$

p-values

For hypothesis testing about parameters θ_1 and θ_2 of the two populations, we judge 'extreme' with respect to the alternative hypothesis. Suppose $H_0 : \theta_1 - \theta_2 = \theta_0$. Then,

- $H_1 : \theta_1 - \theta_2 > \theta_0$

$$\text{p-value} = \mathbb{P}_{H_0}(T \geq t).$$

- $H_1 : \theta_1 - \theta_2 < \theta_0$

$$\text{p-value} = \mathbb{P}_{H_0}(T \leq t).$$

- $H_1 : \theta_1 - \theta_2 \neq \theta_0$

$$\text{p-value} = 2 \min\{\mathbb{P}_{H_0}(T \leq t), \mathbb{P}_{H_0}(T \geq t)\}.$$

Example

A recent study investigated whether games like Minecraft can be used to improve a persons 3D spatial reasoning skills.

Spatial reasoning skill was assessed with the Mental Rotation Test (MRT). Thirty four participants were recruited to the study and randomly allocated to either Group A or Group B of equal size.

All participants completed the MRT at the beginning of the study.

Two weeks later participants in Group A completed a number of tasks in Minecraft. Participants in Group B did not.

All participants completed the MRT again at the end of the study and the change in their score recorded.

Example (continued...)

Question: Do MRT scores improve after the Minecraft training?

The 17 participants in Group A had an average change in MRT score of 6.41 points with a sample standard deviation of 5.56.

The 17 participants in Group B had an average change in MRT score of 3.94 points with a sample standard deviation of 4.72.

Let μ_A be the population mean change in MRT scores following Minecraft training, and let μ_B be the population mean change in MRT scores without Minecraft training. Suppose that both populations are normally distributed.

We would like to test:

$$H_0 : \mu_A - \mu_B = 0, \quad \text{against} \quad H_1 : \mu_A - \mu_B > 0.$$

Parametric vs. Non-parametric Hypothesis Testing

The randomization test in Lecture 1 was

$$\text{Data} \longrightarrow \overline{X}_1 - \overline{X}_2 \longrightarrow \text{Many Randomizations} \longrightarrow P,$$

which is an example of **non-parametric** hypothesis testing.

Now we take a different route and do

$$\text{Data} \longrightarrow \overline{X}_1 - \overline{X}_2 \longrightarrow T \longrightarrow P,$$

which is a **parametric** hypothesis testing.

Comparing Two Means

Suppose $X_1, \ldots, X_{n_X}$ is a simple random sample (i.e. independent and identically distributed) from $\mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_{n_Y}$ is a simple random sample from $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

We would like to perform a test of the hypotheses

$$H_0 : \mu_X - \mu_Y = 0, \quad \text{against} \quad H_1 : \mu_X - \mu_Y > 0.$$

Comparing Two Means

- If we can assume that $\sigma_X = \sigma_Y$, then we saw that

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X + n_Y - 2},$$

  where

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_X - 1)S_Y^2}{n_X + n_Y - 2},$$

  is the pooled estimator of the common variance.

- Otherwise, the statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \overset{\text{approx}}{\sim} t_\nu,$$

  where ν is given by the Welch degrees of freedom (or the conservative alternative).

Example (continued...)

Since $n_X = n_Y = 17$ and $\frac{\max\{5.56^2, 4.72^2\}}{\min\{5.56^2, 4.72^2\}} \approx 1.39$, we can use the pooled estimator of the common variance. Under H_0, our test statistic is then

$$T = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{S_p^2\,(1/17 + 1/17)}} \sim t_{32}.$$

We have $s_p^2 = ((17-1)5.56^2 + (17-1)4.72^2)/32 \approx 26.60$, so the observed test statistic is $t = (6.41 - 3.94)/\sqrt{26.60\,(2/17)} = 1.40$.

The p-value is then $\mathbb{P}(T_{32} \geq 1.40) = 0.086$.

What should we conclude? There is weak evidence against H_0, suggesting that the Minecraft training does improve 3D spatial reasoning skill.

Example (continued...)

If we do not assume equal variances, under H_0, our test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{17} + \frac{S_Y^2}{17}}} \overset{\text{approx}}{\sim} t_{31.19},$$

using the Welch degrees of freedom.

The observed test statistic is $t = \dfrac{6.41 - 3.94}{\sqrt{5.56^2/17 + 4.72^2/17}} \approx 1.40.$

So, the p-value is $\mathbb{P}(T_{31.19} \geq 1.40) \approx 0.086.$

Do we change our conclusion now?

Comparing Two Proportions

Suppose $X_1, \ldots, X_{n_X}$ is a simple random sample (iid) from $\text{Ber}(p_X)$ and $Y_1, \ldots, Y_{n_Y}$ is a simple random sample (iid) from $\text{Ber}(p_Y)$.

Consider the respective sample proportions as

$$\hat{P}_X = \frac{\sum_{i=1}^{n_X} X_i}{n_X}, \quad \text{and} \quad \hat{P}_Y = \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y}$$

Recall

$$\frac{\left(\hat{P}_X - \hat{P}_Y\right) - (p_X - p_Y)}{\sqrt{\dfrac{p_X(1-p_X)}{n_X} + \dfrac{p_Y(1-p_Y)}{n_Y}}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1)$$

Depending on H_0, we can generally encounter three cases.

Comparing Two Proportions

Case 1: Suppose $H_0 : p_X - p_Y = p_0$, then under H_0

$$\frac{(\hat{P}_X - \hat{P}_Y) - p_0}{\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

As we don't know p_X and p_Y, we can use their estimates and by CLT and LLN, we get

$$\frac{(\hat{P}_X - \hat{P}_Y) - p_0}{\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

Comparing Two Proportions

Case 2: However, if $H_0 : p_X = p_Y = p$, then under H_0

$$\frac{\hat{P}_X - \hat{P}_Y}{\sqrt{p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

Case 3: If $H_0 : p_X = p_Y$, i.e., H_0 does not specify p, we estimate it by pooling the proportions

$$\hat{P} = \frac{1}{n_X + n_Y}\left(\sum_{i=1}^{n_X} X_i + \sum_{i=1}^{n_Y} Y_i\right).$$

So, by CLT and LLN, we finally have

$$\frac{\hat{P}_X - \hat{P}_Y}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

Example

A comparative experiment[1] looked at the effectiveness of the sweetener xylitol in preventing ear infections in preschool children.

Syrup	Infection	No Infection	Total
Placebo	68	97	165
Xylitol	46	113	159

Is there any evidence that Xylitol reduces the risk of middle ear infection in children?

[1] Uhari, M., Kenteerkar, T. and Niemala, M. (1998) "A novel use of Xylitol sugar in preventing acute otitis media". *Pediatrics*, **102**, 879–884. Appearing in Utts, J.M. and Heckard, R.F. (2007) *Mind on Statistics* (3rd edition).

Example (continued)

The parameter of interest is the difference $p_1 - p_2$, where p_1 is the true proportion of ear infection for those who used xylitol and p_2 is the true proportion of ear infection for those who did not.

If xylitol is useful, then we expect to have $p_1 < p_2$. The relevant hypotheses are therefore

$$H_0 : p_1 = p_2, \quad \text{against} \quad H_1 : p_1 - p_2 < 0.$$

Parameter estimates are

$\hat{p}_1 = 46/159,\ \hat{p}_2 = 68/165,$ and $\hat{p} = (68 + 46)/(165 + 159) = 0.352.$

Example (continued)

Under H_0, the observed test statistic is

$$z = \frac{(46/159 - 68/165)}{\sqrt{(0.352 \times (1 - 0.352)) \times (1/159 + 1/165)}} \approx -2.314.$$

So, the p-value is $\mathbb{P}(Z \leq -2.314) \approx 0.0103$.

What do we conclude?

```
> prop.test(c(46,68),c(159,165),alternative = "less",
+ correct = FALSE)
```

Comparing two populations: large samples

Comparing means: If the populations are non-normal, then we can approximate the distribution of the test statistic as we did, if the sample sizes are large enough; see slide 17 of Lecture 7.3 and apply to both n_X and n_Y.

- **Populations have the same variance:** We can use the pooled variance estimator and we approximately have a t-distribution with $n_X + n_y - 2$ degree of freedom.

- **Populations have possibly unequal variances:** We can use the Smith-Satterthwaite approximation (or the conservative variant) and have a t-distribution with ν degree of freedom.

Comparing proportions: In all three cases, we can use normal approximation when

$$n_X \cdot \min\{\hat{p}_X, 1 - \hat{p}_X\} \geq 8, \quad \text{and} \quad n_Y \cdot \min\{\hat{p}_Y, 1 - \hat{p}_Y\} \geq 8.$$