Statistics

Week 5: Inference for Two Samples (Chapter 8)

ESD, SUTD

Term 5, 2017

Outline

Independent samples design

Suppose we wish to check if two treatments are significantly different, e. g., compare the effectiveness between two teaching methods, or investigate the salary gap between men and women.

In the 1st case, we can divide a class randomly into two groups, and use different methods to teach them. In the 2nd case, we can randomly sample some men and women from the same company.

Such situations can be modeled as follows. Take random samples from two populations:

Sample 1: $x_1, x_2, \ldots, x_n$

Sample 2: $y_1, y_2, \ldots, y_m$

The two samples are statistically *independent*, and n, m do *not* necessarily equal.

Graphical methods

Before doing any statistical analysis, one should investigate the two samples graphically, for example using

- Side-by-side box plots,
- A Q-Q plot. This is particularly easy if $n = m$, since it involves plotting $x_{(i)}$ vs $y_{(i)}$.

Why is a scatter plot not a good idea?

Compare the means

Suppose the two populations have means μ_1 and μ_2, and standard deviations σ_1 and σ_2, all of which are unknown. We are interested in the difference $\mu_1 - \mu_2$.

Consider the sample means $\bar{X}$ and $\bar{Y}$. We have:

$$\mathsf{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2,$$

$$\mathsf{Var}(\bar{X} - \bar{Y}) = \mathsf{Var}(\bar{X}) + \mathsf{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

(Make sure that you understand why these formulas are true.)

Large samples, CI

When both n and m are *large*, the random variable

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

is approximately standard normal, as a consequence of the CLT.

Since n and m are large, we can approximate σ_i by s_i.

Therefore, the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is given by:

$$\bar{x} - \bar{y} - z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}.$$

Hypothesis testing

We are often interested in testing
$H_0 : \mu_1 - \mu_2 = \delta_0$ vs $H_1 : \mu_1 - \mu_2 \neq \delta_0$,
where δ_0 is some specified value (often, $\delta_0 = 0$ is used).

We may use the CI to perform the hypothesis test. Alternatively, we can compute the statistic

$$z = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}},$$

and find the p-value, given by $P(|Z| \geq |z|)$.

Reject H_0 if the p-value is less than the predetermined value of α.

One-sided confidence intervals and hypothesis testing can be done similarly (try writing down the formulas yourself).

Further reading

What if n and m are not large? Then the formulas become more involved, and depend on whether the populations variances equal.

You can read about them in Section 8.3 of the textbook, under the heading 'Inferences for small samples'. This section is not a part of the course, but you may find it useful for your project.

Outline

Introduction

Independent samples design has a disadvantage: randomization may not ensure that the two groups are equal on all attributes (except for the treatment).

To overcome this, we can use **matched pairs design**. Form n matched pairs, (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Each pair is chosen to be *as similar as possible*, thus improving precision of the study.

Examples: to investigate the salary gap, for each pair, choose 1 man and 1 woman with the same job, qualification and experience.

To compare customers' preference between two types of coffee, ask each customer to drink both types.

To compare the rate of respiratory problems between smokers and non-smokers, find identical twins, only one in each pair smokes.

Drawback: it may not be easy or possible to form matched pairs.

Mean and variance

Assume that $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$, and that n is not necessarily large.

Now that the samples are paired up, it makes sense the consider the normal random variables $D_i = X_i - Y_i$.

$$\mathsf{E}(D_i) = \mu_1 - \mu_2,$$

$$\mathsf{Var}(D_i) = \mathsf{Var}(X_i) + \mathsf{Var}(Y_i) - 2\,\mathsf{Cov}(X_i, Y_i),$$

which is usually *smaller* than the sum of the variances, if the matching is done successfully. (Why?)

Consider the observed values x_i and y_i. Let

$$d_i = x_i - y_i, \quad \text{and} \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

Confidence interval

Also, compute the sample standard deviation of the d_i,

$$s_d = \left[\frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2 \right]^{1/2}.$$

Then, the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{d} - t_{n-1,\,1-\alpha/2} \frac{s_d}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \bar{d} + t_{n-1,\,1-\alpha/2} \frac{s_d}{\sqrt{n}}.$$

As the standard error here is expected to be less than the one in independent samples design, we expect matched pairs to give smaller CI.

Hypothesis testing

The CI can be use for hypothesis testing. Alternatively, compute

$$t = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}},$$

and then find the p-value.

Exercise:

1. Open the *Excel* file. We want to check if the body temperatures of males and females are different, using $\alpha = 0.05$. Which type of experimental design is this?
2. Set up H_0 and H_1 and perform the hypothesis test.