Two sample inference

ST551 Lecture 18

Charlotte Wickham

2017-11-01

Finish last time's slides

Two sample inference

Two sample setting

Setting: two **independent** samples

$Y_1, \ldots, Y_n$ i.i.d from population with c.d.f F_Y, and
$X_1, \ldots, X_m$ i.i.d from population with c.d.f F_X

Parameter: now focus on some comparison between the two populations F_Y and F_X

Alternative view

Setting: two **independent** samples

$$(Y_1, G_1), (Y_2, G_2), \ldots, (Y_n, G_n), (Y_{n+1}, G_{n+1}), \ldots, (Y_{n+m}, G_{n+m})$$

where G is a binary *grouping* variable which indicates which population the observation came from:

$$G_i = \begin{cases} 0, & \text{observation from } Y \\ 1, & \text{observation from } X \end{cases}$$

Two views are equivalent

Depending on sampling scheme one view may seem more natural:

- I sample 40 OSU graduate students and 20 OSU undergraduate students:
  - Y_i = graduate student time to complete 1 mile run, $i = 1, \ldots, 40$
  - X_i = undergraduate student time to complete 1 mile run, $i = 1, \ldots, 20$
- I sample 60 OSU students and record:
  - Y_i = time to complete 1 mile run, $i = 1, \ldots, 60$
  - G_i = student's level (0 = graduate, 1 = undergraduate), $i = 1, \ldots, 60$

In second view, if we condition on the counts in each group, inference is the same as first view.

Two sample inference for difference in population means

To compare population means: $\mu_Y = E(Y_i)$, $\mu_X = E(X_i)$, we might look at their difference:

$$\delta = \mu_Y - \mu_X$$

(In alternative view: equivalent to
$\delta = E(Y_i \mid G_i = 0) - E(Y_i \mid G_i = 1)$)

- Estimate for δ
- Test for $H_0 : \delta = \delta_0$
- Confidence interval for δ

Difference in sample means

It seems reasonable to use:

$$\hat{\delta} = \overline{Y} - \overline{X}$$

as a good starting point for inference on $\delta = \mu_X - \mu_Y$.

Complete worksheet (Charlotte will provide)

Leads to two sample Z-test and intervals

Assume known population variances: $Var(Y_i) = \sigma_Y^2$ $Var(X_i) = \sigma_X^2$.

$$Z(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\sigma_Y^2/n + \sigma_X^2/m}}$$

Reference Distribution: If null hypothesis $H_0 : \delta = \delta_0$ is true, then

$$Z(\delta_0) \dot{\sim} N(0,1)$$

Rejection Regions:

- $H_A : \delta > \delta_0$, reject H_0 for $Z(\delta_0) > z_{1-\alpha}$
- $H_A : \delta < \delta_0$, reject H_0 for $Z(\delta_0) < z_{\alpha}$
- $H_A : \delta \neq \delta_0$, reject H_0 for $|Z(\delta_0)| > z_{1-\alpha/2}$

Leads to two sample Z-test and intervals

$(1-\alpha)100\%$ Confidence interval for $\delta = \mu_Y - \mu_X$

$$(\overline{Y} - \overline{X}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}}$$

Next time...

What if population variances aren't known?