MA22004

Seminar 6

Dr Eric Hall

12/11/2020

Announcements

Reminders

- It is week 6! You should have read the remainder of §4 of the notes on **Perusall**.
- Feedback for Class Test 1 is posted.

Upcoming

- Lab 4 due **Thursday 12 Nov** at **17:00**: upload to **Gradescope**; late submissions will be accepted until Sat 14 Nov at 13:00.
- Worksheet #6 (Two sample inferences) will be posted to **Blackboard** tonight: start before next workshop (should last two weeks).
- Investigation #5 on **Perusall**: do before next workshop.
- Reading assignment #7 (§5) on **Perusall**: do before next seminar.

Two sample inferences

- Compare and contrast CI and hypothesis tests for one and two sample inferences for population proportions.
- Two sample inferences for population variances (F distribution).

Population proportions

Consider a population containing a proportion p_X of individuals satisfying a given property. For a sample of size m from this population, we denote the sample proportion by $\widehat{p}_X$. Likewise, p_Y, n, $\widehat{p}_Y$.

We assume the samples from the X and Y populations are independent.

The natural estimator for the difference in population proportions

$$p_X - p_Y$$

is the difference in the sample proportions

$$\widehat{p}_X - \widehat{p}_Y .$$

Population proportions and CLT

$$\mu_{(\widehat{p}_X - \widehat{p}_Y)} = \mathbf{E}[\widehat{p}_X - \widehat{p}_Y] = p_X - p_Y \ ,$$

and

$$\sigma^2_{(\widehat{p}_X - \widehat{p}_Y)} = \mathrm{Var}[\widehat{p}_X - \widehat{p}_Y] = \frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n} \ .$$

If m and n are large, difference between to proportions:

$$\widehat{p}_X - \widehat{p}_Y \sim \mathsf{N}\left(p_X - p_Y, \frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}\right)$$

$100(1-\alpha)\%$ CI for $p_X - p_Y$

$$\widehat{p}_X - \widehat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{m} + \frac{\widehat{p}_Y(1-\widehat{p}_Y)}{n}},$$

As a rule of thumb, if $m\widehat{p}_X$, $m(1-\widehat{p}_X)$, $n\widehat{p}_Y$, and $n(1-\widehat{p}_Y)$ are greater than or equal to 10.

Hypothesis test on equality

If we are considering a hypothesis test concerning the equality of the population proportions,

$$H_0 : p_X - p_Y = 0\,,$$

then we assume $p_X = p_Y$ as our default position.

We must replace the standard error with a pooled estimator for the standard error of the population proportion,

$$\widehat{p} = \frac{m}{m+n}\widehat{p}_X + \frac{n}{m+n}\widehat{p}_Y$$

Pooled estimator

Consider $H_0 : p_X - p_Y = 0$.

The test statistic is

$$Z = \frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\widehat{p}(1 - \widehat{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} .$$

Hypothesis test for difference of proportions

For a test at level α:

If $H_a : p_X - p_Y > 0$, then $P = 1 - \Phi(z)$, i.e., upper-tail $R = \{z > z_\alpha\}$.

If $H_a : p_X - p_Y < 0$, then $P = \Phi(z)$, i.e., lower-tail $R = \{z < -z_\alpha\}$.

If $H_a : p_X - p_Y \neq 0$, then $P = 2(1 - \Phi(|z|))$, i.e., two-tailed $R = \{|z| > z_{\alpha/2}\}$.

> As a rule of thumb: if $m\widehat{p}_X, m(1 - \widehat{p}_X), n\widehat{p}_Y, n(1 - \widehat{p}_Y)$ are all greater than 10.

Comparing variances

For a random sample

$$X_1, \dots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and an independent random sample

$$Y_1, \dots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)\,,$$

the rv

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \qquad \sim \mathsf{F}(m-1, n-1)\,,$$

that is, F has an F distribution with df $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

Making comparisons with a ratio

The statistic F comprises the *ratio* of variances σ_X^2/σ_Y^2 and not the difference; therefore, the plausibility of $\sigma_X^2 = \sigma_Y^2$ will be based on how much the ratio differs from 1.

For $H_0 : \sigma_X^2 = \sigma_Y^2$, and the P-values are determined by the $\mathsf{F}(m-1, n-1)$ curve where m and n are the respective sample sizes.

Hypothesis test for comparing variances

For a hypothesis test at level α, we use the following procedure:

If $H_a : \sigma_X^2 > \sigma_Y^2$, then P-value is $A_R =$ area under the F$(m-1, n-1)$ curve to the right of f.

If $H_a : \sigma_X^2 < \sigma_Y^2$, then P-value is $A_L =$ area under the F$(m-1, n-1)$ curve to the left of f.

If $H_a : \sigma_X^2 \neq \sigma_Y^2$, then P-value is $2 \cdot \min(A_R, A_L)$.

> Assume the population distributions are normal and the random samples are both independent of one another.

$100(1-\alpha)\%$ CI for comparing variances

For the *ratio* of population variances σ_X^2/σ_Y^2, is given by

$$\left(F^{-1}_{\alpha/2,m-1,n-1} s_X^2/s_Y^2 \, , F^{-1}_{1-\alpha/2,m-1,n-1} s_X^2/s_Y^2\right) \ .$$

Summary

Today we discussed CI and hypothesis tests for comparing:

- population proportions
- population variances

from two different groups.

The latter required us to consider the F-distribution.