MA22004

Seminar 5

Dr Eric Hall

05/11/2020

Announcements

Reminders

- It is week 4! You should have read the remainder of §3 of the notes on **Perusall**.
- Feedback for Lab 2 is posted.

Upcoming

- Lab 3 due **Thursday 5 Nov** at **17:00**: upload to **Gradescope**; late submissions will be accepted until Sat 7 Nov at 13:00.
- Worksheet #5 (Two sample inferences) is on **Blackboard**: start before next workshop (should last two weeks).
- Investigation #4 on **Perusall**: do before next workshop.
- Reading assignment #6 (rest of §4) on **Perusall**: do before next seminar.

Inferences for means based on two samples

Today we compare **means** based on two samples from different groups.

E.g.,

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i\,, \qquad X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \qquad Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$$

Two types of sampling

Comparisons for means fall into two types:

- sets of observations are dependent (i.e. **paired** between the groups)
- sets of observations are independent (i.e. between the groups)

The samples must still be independent *within* each set of observations.

Paired data

When two sets of observations have a special correspondence (i.e. are dependent) the sets of observations are said to be paired.

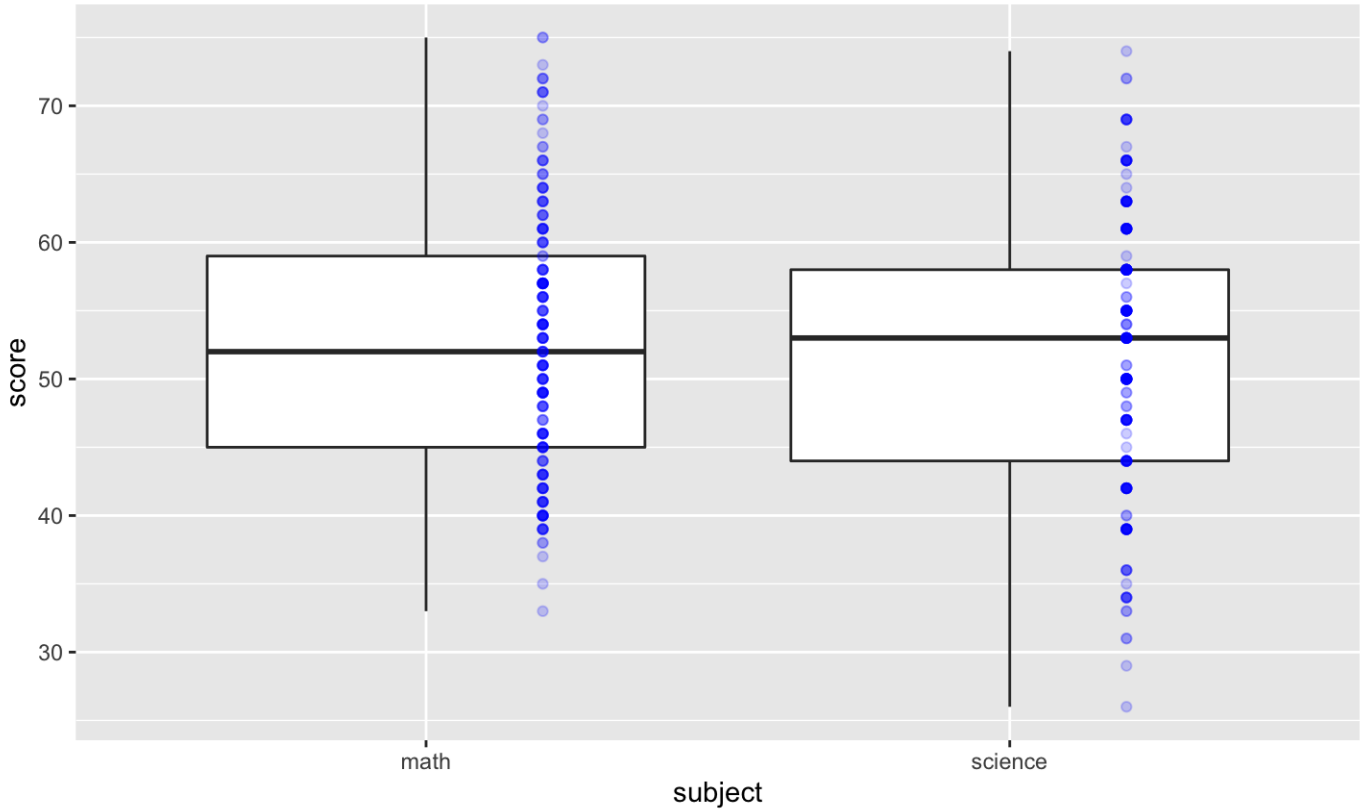
What is the approach?

To analyze paired data we will consider the difference of each paired observation:

$$\mu_D = \mu_X - \mu_Y$$

Paired data: math and science scores

Consider 200 observations of students that took a standardized science and math test. How are the distributions similar? How are they different?



Paired or not?

Can the math and science scores for a given student be assumed to be independent of each other?

id	math	science	diff
70	41	47	-6
121	53	63	-10
86	54	58	-4
141	47	53	-6

Means of paired data

Parameter of interest

$$\mu_{\text{diff}}$$

Average difference between math and science scores of **all** students.

Point estimator

$$\bar{x}_{\text{diff}}$$

Average difference between math and science scores of 200 **sampled** students.

Hypothesis test for paired data

$$H_0 : \mu_{\text{diff}} = 0\,, \quad \text{(there is no difference between scores)}$$

vs

$$H_a : \mu_{\text{diff}} \neq 0\,, \quad \text{(there is a difference between scores)}$$

Calculate an appropriate test statistic for the *new* parameter μ_{diff}.

$$\bar{x}_{\text{diff}} = 0.795\,, \quad s_{\text{diff}} = 8.2938\,, \quad n_{\text{diff}} = 200\,.$$

Nothing new: carry out inference on a single sample population mean.

Calculate test statistic

Let $\alpha = 0.10$

$H_0 : \mu_{\text{diff}} = 0$

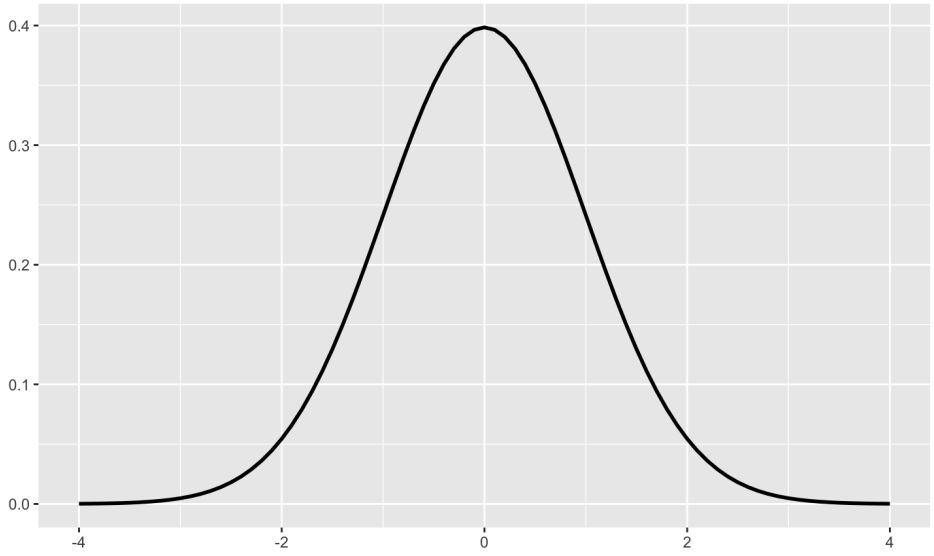
$H_a : \mu_{\text{diff}} \neq 0$

$\bar{x}_{\text{diff}} = 0.7950$

$s_{\text{diff}} = 8.2938$

$n_{\text{diff}} = 200$

Recap: P-values



What is the correct interpretation of the P-value?

Difference of two independent means

General interval estimate:

$$\text{point estimate} \pm \text{margin of error}$$

Now for the parameter of interest $\mu_D = \mu_X - \mu_Y$:

$$(\bar{x} - \bar{y}) \pm t_{\alpha/2,\nu} \cdot \hat{\sigma}_{(\bar{x}-\bar{y})}$$

Only new concept:

$$\hat{\sigma}_{(\bar{x}-\bar{y})} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

Tricky parts…

Complicated to compute true df!

A conservative estimate for the degrees of freedom ν is

$$\nu = \min(m - 1, n - 1) \, .$$

Check conditions

1. Independence of samples both within and between groups.
2. Sample size and skew (more skewed distributions need larger number of samples).

Summary

Today we discussed CI and hypothesis tests for:

- differences of means of paired observations
- differences of means of independent observations

from two different groups.