ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DIFFERENCE OF MEANS

CONFIDENCE INTERVALS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Confidence Intervals for Difference of Means

- We consider the difference in two population means $\mu_1 - \mu_2$ *when data is not paired.*

- **Idea:**
  - Identify conditions for using t-distribution with the point estimate $\bar{x}_1 - \bar{x}_2$.
  - Identify new formula for SE, and df, in this context.
  - Otherwise, proceed as in the one-sample (one-mean) case.

Confidence Intervals for Difference of Means

- **Example:** A small randomized control trial gives the following results for treating a particular condition. Here positive numbers indicate better outcomes.

	n	Sample mean	s
Treatment	9	3.5	5.17
Control	9	-4.33	2.76

- **Conditions for using t-distribution**
  - **Independence (extended):** data are independent within and between groups.
  - **Normality:** check the outliers rule of thumb for each group separately.

Confidence Intervals for Difference of Means

- **Expression for SE:** $$SE = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
- **Expression for df:** We approximate df using $df = min\{n_1, n_2\} - 1.$
- **Reason:**
  - Formal expression for df can be complicated.
  - Recall: In the one-sample (one-mean) case, t-distribution arose because we had to use an estimated standard deviation instead of the true one.
  - Here: In the two-sample (two-mean) case, our statistic does not exactly follow the t-distribution because we estimate two standard deviations instead of just one.
  - $df = min\{n_1, n_2\} - 1$ is a conservative estimate that allows us to circumvent this issue.

Back to Our Example

- **Example:** A small randomized control trial gives the following results for treating a particular condition. Here positive numbers indicate better outcomes.

	n	Sample mean	s
Treatment	9	3.5	5.17
Control	9	-4.33	2.76

- Since our data is independent, and assuming no clear outliers in our data (we can check this given the full data set), we can use the t-distribution to model the difference of the means.

- $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{5.17^2/9 + 2.76^2/9} = 1.95, \quad df = 9 - 1 = 8.$

Back to Our Example

- From our data and calculations, we now have
  - $\bar{x}_1 = 3.5, \ \bar{x}_2 = -4.33,$
  - $SE = 1.95, \quad df = 8.$
- **Exercise:** Calculate a $95\,\%$ Confidence Interval for the effect of the treatment on the change in outcomes.
- **Solution:** point-estimate = $\bar{x}_1 - \bar{x}_2 = 7.83$
  - $t_8^{\star} = 2.31$ (using software, or statistical tables).
  - Our interval is given by $(\bar{x}_1 - \bar{x}_2) \pm t_8^{\star} \times SE$
  - $\implies \ I = (3.32, \ 12.34)$

Summary

- As usual, there are 4 steps to conducting a two-mean hypothesis test
  - **Prepare:** Identify or calculate important parameters and determine the significance level α to be used.
  - **Check:** Verify the conditions for using t-distributions.
  - **Calculate:** calculate SE, and construct the confidence interval.
  - **Conclude:** Provide a conclusion/interpretation in the context of the problem.