ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DIFFERENCE OF MEANS

HYPOTHESIS TESTING

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Hypothesis Testing for Difference of Means

- **Example (OpenIntro Ch.7) :**
  - A data set comprises a random sample of 150 cases of mothers and their newborns in North Carolina over a year. It has a *weight* variable representing the weights of the newborns and a *smoke* variable describing which mothers smoked while pregnant.
  - We would like to know: is there evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?
  - μ_n : mean weight for newborns of non-smoking mothers.
    μ_s : mean weight for newborns of smoking mothers.
  - H_0 : There is **no difference** in average birth weight, i.e., $\mu_n - \mu_s = 0.$
    H_1 : There is **some difference** in average birth weight, i.e., $\mu_n - \mu_s \neq 0.$

Hypothesis Testing for Difference of Means

- **Example (Cont'd) :** Suppose that the sample sizes are $n_n = 100,\ n_s = 50.$
- Can we model the sample difference in means using the t-distribution?
- As in the confidence interval case, we check two conditions:
  - **Independence:** data comes from a random sample so the observations are independent both within and between samples.
  - **Lack of outliers:** since $n > 30$, we only need to check for "extreme" outliers. Since this is data about weights of newborns, it is reasonable to assume none.
- Since both conditions are satisfied we can model the data using the t-distribution.

Hypothesis Testing for Difference of Means

- **Example (Cont'd) :** Here are the summary statistics associated with the data.

	n	Sample mean	Standard Deviation
Smoker	100	6.78	1.43
Non-smoker	50	7.18	1.6

- We would like to complete the hypothesis test at a significance level $\alpha = 0.05$.

- We start (as usual) by calculating the point estimate $\bar{x}_n - \bar{x}_s = 0.4$,

- And the standard error

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = 0.26$$

Hypothesis Testing for Difference of Means

- Since $\bar{x}_n - \bar{x}_s = 0.4$ and $SE = 0.26$, we can now calculate the test statistic

$$T = \frac{\text{(point estimate of difference of means)} - \text{(difference of means under the null)}}{SE}$$

$$T = \frac{0.4 - 0}{0.26} = 1.54$$

- To get a p-value, we now calculate the probability under the null that we get data this extreme, i.e., $\mathbb{P}(T > 1.54$ or $T < -1.54)$.
- To calculate this probability, we recognize that we have a t-distribution with $df = \min\{n_n, n_s\} - 1 = 49$.
- Using software, or tables, this gives $p - value = 0.1304$

Hypothesis Testing for Difference of Means

- So, we have that

$$p\text{-value} \geq \alpha$$

- We **do not reject** the null hypothesis.

- There is not enough evidence at the $\alpha = 0.05$ level to conclude that there is a difference in average weight of newborns from mothers who smoke and those who did not smoke during pregnancy.

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, the test statistic T, and the p-value.
  - **Conclude:** Provide a conclusion/interpretation in the context of the problem.