ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

ONE SAMPLE T-TESTS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

One Sample t-Tests

- **Example:** We'd like to determine whether UCSD students sleep less than 7 hrs a night on average. A random sample of 50 students is asked how much they sleep at night, and the point estimate suggests they sleep less than 7 hrs a night on average.

n	Sample mean	Sample SD (s)	Sample min	Sample max
50	6.74	0.71	5.81	8.98

- **Here, our hypotheses are**

  - H_0 : UCSD students sleep 7 hours a night, on average. $\iff$ $H_0 : \mu = 7$

  - H_1 : UCSD students sleep less than 7 hours a night, on average. $\iff$ $H_1 : \mu < 7$

One Sample t-Tests

n	Sample mean	s	Sample min	Sample max
50	6.74	0.71	5.81	8.98

- **Exercise:** Are the independence and normality conditions satisfied? (Yes.)
- **Question**: What is the standard error SE for our sample?
- **Answer:**

$$SE = \frac{s}{\sqrt{n}} = \frac{0.71}{\sqrt{50}} = 0.1004$$

- **Question:** What is df in this example?
- **Answer:**

$$df = n - 1 = 50 - 1 = 49$$

One Sample t-Tests

- We now know that $\bar{x} = 6.74$, $SE = 0.1004$, and $df = 49$. We'd like to now conduct our hypothesis test.

Categorical Data

Find Z-score using observed value, null value, and SE.

$$Z = \frac{\hat{p} - p_0}{SE}$$

Find p-value: the probability that data as extreme as the sample was generated, under the null.

Numerical Data

Find **T-score** using observed value, null value, and SE.

$$T = \frac{\bar{x} - \mu}{SE}$$

Find p-value: the probability that data as extreme as the sample was generated, under the null.

One Sample t-Tests

- **Question:** Having found $\bar{x} = 6.74$, $SE = 0.1004$, and $df = 49$, find the test statistic T, and the associated p-value for the given sample.
- **Answer:**
  - $T = \dfrac{\bar{x} - \mu}{SE} = \dfrac{6.74 - 7}{0.1004} \approx -2.59$
    (note that we used the mean μ under the null, i.e., $\mu = 7$)
  - Using software or a t-table we find that $\mathbb{P}(T < -2.59) = 0.0063$.
    (Note that we calculated the tail area only on one side of the distribution).
  - So the p-value is 0.0063 which means we reject the null hypothesis at the 95% confidence level (since 0.0063 < 0.05).

One Sample t-Tests

- **Question:** Interpret your result.

- **Answer:**

  - Because the p-value is smaller than 0.05, we reject the null hypothesis.

  - That is, the data provide strong evidence that the UCSD students sleep less than 7 hrs a night on average.

Summary

- There are 4 steps to conducting a one-mean hypothesis test
  - **Prepare:** Identify or calculate $\bar{x}$, s, n, and determine the significance level α to be used.
  - **Check:** Verify the conditions that $\bar{x}$ is nearly normal.
  - **Calculate:** If $\bar{x}$ is nearly normal, calculate SE, and calculate the T-score $T = \frac{x - \mu}{s}$. Calculate the p-value, which depends on df.
  - **Conclude:** compare the p-value to α and evaluate the hypothesis test. Provide a conclusion in the context of the problem.

Variation on a Theme: Paired Data

- Sometimes we work with paired data, e.g., item prices at two grocery stores.
- **Definition:** Two sets of observations are paired if each observation in one set has a special correspondence with exactly one observation in the other set.

	Whole Foods	Vons	Difference
Fuji Apples	1.89	1.49	**0.4**
Whole Milk	2.49	3.99	**-1.5**
…			
Yoghurt	5.89	5.99	**-0.1**

- In such cases, it often makes sense to examine *differences* in pairs of observations.
- We then analyze the differences using the t-distribution techniques we just saw.