ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

GOODNESS OF FIT TESTS

PART 1

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Why Goodness of Fit Tests?

- **When might you be interested in goodness of fit tests?**

  - You have a sample that can be classified into several groups, and you'd like to know if the sample is representative of the population.

  - You have data, and you would like to know if the data can reasonably be modeled as being drawn from a particular distribution (e.g., the normal distribution).

- **We can answer such questions with hypothesis tests!**

Example: Goodness of Fit Tests

- **Example:** A random sample of 275 jurors from a small county had jurors identify their racial group. We'd like to know if the sample is representative of the population.

	White	Black	Hispanic	Other	Total
On Juries (number)	205	26	25	19	275
Registered Voters (Proportion)	0.72	0.07	0.12	0.09	1

Example is from OpenIntro Statistics (Chapter 6)

- **Issue:** We'd like to examine all groups simultaneously (so not just pairs). This means our previous test statistics (for means and differences of means) don't apply.

The Chi-Square Test Statistic

- **Example:** Replacing the second row by expected valued, we get.

	White	Black	Hispanic	Other	Total
On Juries	205	26	25	19	275
Expected Counts	198	19.25	33	24.75	275

- Do the differences provide evidence that the jurors are not a random sample?
  - H_0 : The jurors are a random sample (no racial bias).
  - H_1 : The jurors are *not* a random sample (racial bias in juror selection).

The Chi-Square Test Statistic

- **Recall:** In hypothesis testing, we construct a ***test statistic,*** and we calculate how likely it is to have a statistic at least this extreme, under the null hypothesis.
- **Our test statistic here:** Suppose we have k groups.

For each group, calculate the Z-score

$$Z_i = \frac{\text{(observed count from group i)} - \text{(expected count under null of group i)}}{\text{SE of group i}}$$

Combine them to get the test statistic

$$X^2 = \sum_{i=1}^{k} Z_i^2$$

Example: the Chi-Square Test Statistic

- **Example (cont'd):** So in our jury example, we calculate

$$Z_1 = \frac{\text{(observed count of white jurors)} - \text{(expected white juror count under null)}}{\text{SE of observed white count}}$$

- $Z_1 = \dfrac{205 - 198}{\sqrt{198}} = 0.5$

- $Z_2 = \dfrac{26 - 19.25}{\sqrt{19.25}} = 1.54, \quad Z_3 = \dfrac{25 - 33}{\sqrt{33}} = -1.39, \quad Z_4 = \dfrac{19 - 24.75}{\sqrt{24.75}} = -1.16$

- $\Longrightarrow X^2 = (0.5)^2 + (1.54)^2 + (-1.39)^2 + (-1.16)^2 = 5.8993$

The Chi-Square Test Statistic

- X^2 gives us a sense of how much the observed counts deviate from the null counts.
- So higher values for X^2 indicate more deviation, but how do we figure out whether the null-hypothesis is true or not based on this test statistic?
- **Next, we will show that *if the null-hypothesis is true*, X^2 follows a distribution known as the chi-square distribution.**
- By using the properties of this distribution, we will obtain a p-value to evaluate our hypotheses.