ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

GOODNESS OF FIT TESTS

EVALUATING GOODNESS OF FIT FOR A DISTRIBUTION

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Evaluating Goodness of Fit for a Distribution

- **Advantage of chi-square goodness of fit tests:** can be applied to any single variable distribution for which we can calculate the cumulative distribution function.
- Suppose we suspect our data follows a specific distribution. We use our hypothesis testing framework to check this.

  H_0 : The data follows the distribution

  H_1 : The data does not follow the distribution

- Ok, but what is our test-statistic?

Evaluating Goodness of Fit for a Distribution

- **Test statistic:** For the chi-squared goodness of fit test, ***data is divided into k-bins*** *(we can do this ourselves if the data is not binned),* and the test statistic is

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Like before, O_i and E_i are the observed and expected counts for bin i.

- **Warning:** The test is sensitive to the choice of bins, but reasonable choices should produce similar results. A good rule of thumb is you need $E_i \geq 5$ for all bins.

Example: Evaluating Goodness of Fit for a Distribution

- **Example:** Suppose you are playing a dice-game where you roll two dice and your winnings depend on the number of sixes your roll. You play the game 200 times and observe the following counts.

	0 sixes	1 six	2 sixes
Number of times outcome appears	130	58	12

- **Question:** Conduct a chi-square goodness of fit test to determine if the dice are fair.
- **Solution:** We start with our Hypothesis.

H_0 : Dice are fair $\qquad$ H_1 : Dice are not fair

Evaluating Goodness of Fit for a Distribution

- **Solution (cont'd):** We start with our Hypothesis.

  - H_0 : Dice are fair $\qquad$ H_1 : Dice are not fair

- We interpret the dice being fair as meaning

  - The probability of rolling a 6 on any given roll to be $1/6$.
  - The dice are independent

- In other words, the number of sixes in 2 rolls is Binomial(2, 1/6)! So

  - $\mathbb{P}(\text{roll 0 sixes}) = 25/36$
  - $\mathbb{P}(\text{roll 1 sixes}) = 10/36$
  - $\mathbb{P}(\text{roll 2 sixes}) = 1/36$

Evaluating Goodness of Fit for a Distribution

- **Solution (cont'd):** From these probabilities we can calculate the expected numbers under the null hypothesis, and compare it to the results we observed

	0 sixes	1 six	2 sixes
Number of times outcome appears	130	58	12
Expected number	138.889	55.556	5.556

- So we can now calculate

- $$X^2 = \frac{(130 - 138.889)^2}{138.889} + \frac{(58 - 55.556)^2}{55.556} + \frac{(12 - 5.556)^2}{5.556} \approx 8.15$$

Evaluating Goodness of Fit for a Distribution

- **Solution (cont'd):** Remains to calculate the p-value by finding $\mathbb{P}(\chi^2 \geq X^2)$, the probability that a chi-square r.v. with $k-1$ degrees of freedom, is at least as extreme as X^2.

- We have 2 degrees of freedom (3 bins, have to sum to 200, so 1 degree of freedom is lost).

- Using tables or software, we determine
$$\mathbb{P}(\chi^2 \geq 8.15) = 0.017$$

- At the $\alpha = 0.05$ level, we can reject(!) the null-hypothesis that the dice are fair.

- Can you guess in which direction the dice are not fair?