ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

# GOODNESS OF FIT TESTS

## PART 2

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

# The Chi-Square Distribution

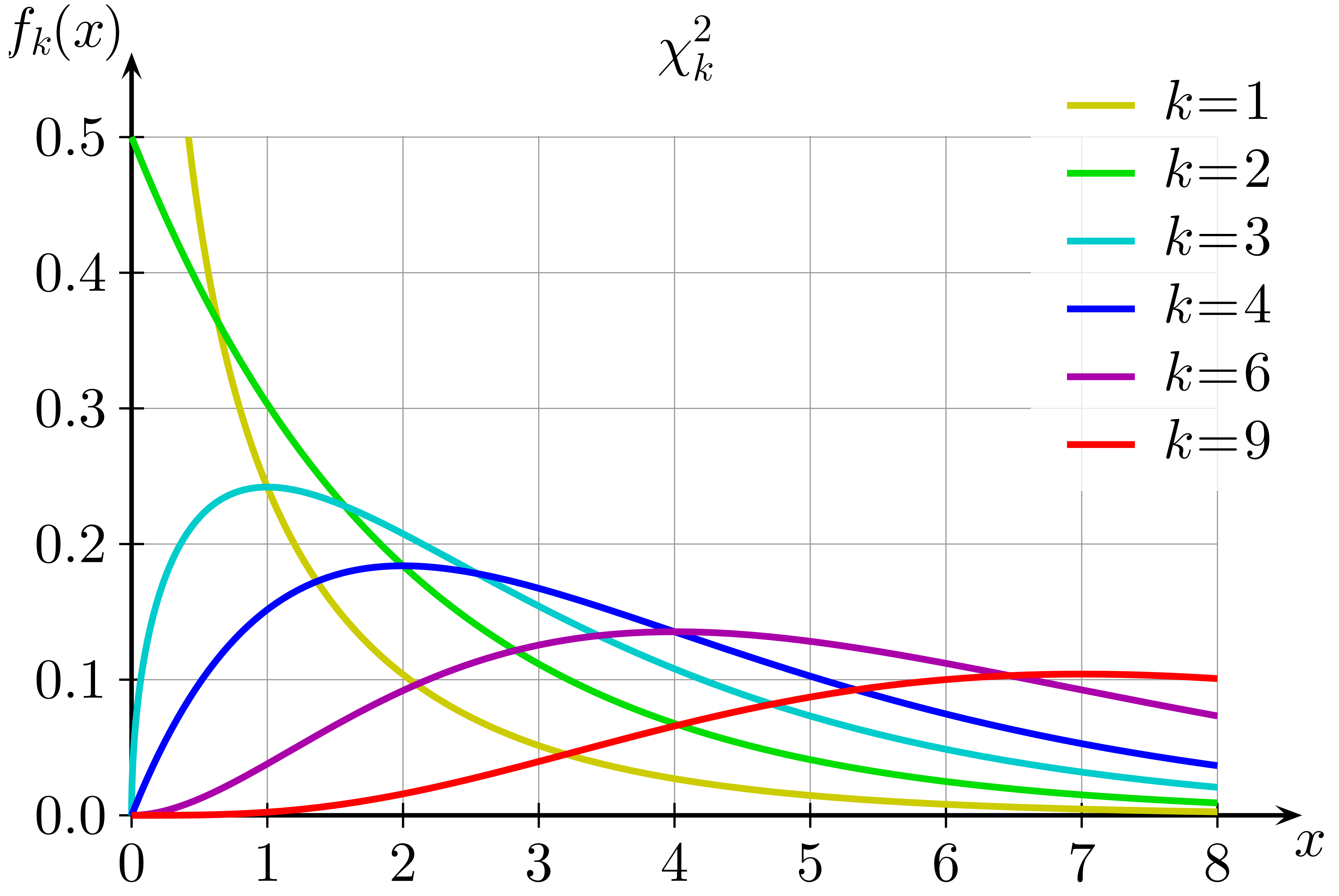
- **Recall:** We proposed a test statistic $X^2 = \sum_{i=1}^{k} Z_i^2$ where

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

- To see why this makes sense, we will need to take a small detour.
- **Definition:** The ***chi-squared distribution*** ($\chi^2$-distribution) with $k$ degrees of freedom is the distribution of a sum of the squares of $k$ independent standard normal random variables.
- So a $\chi^2$ random variable is always non-negative!

# The $\chi^2$ Distribution

- The $\chi^2$-squared distribution, has a single parameter: $k$
- In stats this parameter is often called the degrees of freedom, or **df.**
- **df** determines the shape, center, and spread of the distributions.
- Using the figure, can you see how? How about using the definition?




The probability density function of chi-squared distributions with k degrees of freedom.

## Example: the Chi-Square Distribution

- **Example:** Given a $\chi^2$ random variable, $x$, with 3 degrees of freedom, what is the probability that $x$ exceeds 6.25.
- **Solution:** We need to calculate $\mathbb{P}(x \geq 6.25) = \mathbb{P}(\xi_1^2 + \xi_2^2 + \xi_3^2 \geq 6.25)$
- Can use software, or tables to do this. Turns out

$$\mathbb{P}(x \geq 6.25) \approx 0.1001$$

- **Question:** Do you expect this probability to increase or decrease with more degrees of freedom?
- **Answer:** It should increase, because increasing df means we are adding more positive terms, which increases the probability that the sum is large.

# Finding a P-Value for the Chi-Square Distribution

- **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.

- **Recall:** In our juror example we calculated $X^2 = \sum_i^4 Z_i^2 = 5.89$

- **If our null hypothesis (of no racial bias) were true, then our test-statistic $X^2$ would follow the $\chi^2$ distribution with 3 degrees of freedom.**

- Why 3, not 4, degrees of freedom? Roughly speaking, the sum of all the proportions has to add up to 1, taking away one degree of freedom.

## The Chi-Square Test Statistic

**Example (recall):** In our jury example, we calculated

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

- $Z_1 = \dfrac{205 - 198}{\sqrt{198}} = 0.5$
- $Z_2 = \dfrac{26 - 19.25}{\sqrt{19.25}} = 1.54, \quad Z_3 = \dfrac{25 - 33}{\sqrt{33}} = -1.39, \quad Z_4 = \dfrac{19 - 24.75}{\sqrt{24.75}} = -1.16$
- $\implies X^2 = (0.5)^2 + (1.54)^2 + (-1.39)^2 + (-1.16)^2 = 5.8993$

## Example: the Chi-Square Test Statistic

**Example (recall):** Now, we need to calculate the probability, under the null (**$\chi^2$-distribution with 3 degrees of freedom**), of obtaining a statistic at least as extreme as 5.8993.

- So we calculate $$\mathbb{P}(X^2 \geq 5.8993) \approx 0.1116$$
- This is our p-value
- Since our p-value is large (for example it is larger than $\alpha = 0.05$), we do not reject the null hypothesis (of no racial bias).

# Summary: Chi-Square Test for One-Way Table

- To evaluate whether observed counts $O_1, O_2, \ldots, O_k$ in k categories are different from what we'd expect from a null-hypothesis where the expected counts are $E_1, E_2, \ldots E_k$, we

- Calculate the test statistic $X^2 = \dfrac{(O_1 - E_1)^2}{E_1} + \dfrac{(O_2 - E_2)^2}{E_2} + \ldots + \dfrac{(O_k - E_k)^2}{E_k}$

- 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 $X^2$.

## Summary: Conditions for a Chi-Square Test

As always, we should check two conditions before performing a chi-square test

- **Independence**
- **Sample size:** each particular scenario (i.e., number in the table) must have at least 5 expected cases.