Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Hypothesis Testing I: χ^2 distribution

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

OCTOBER 4, 2016

Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

1	2	3	4	5	6
8	5	9	2	7	5

- H_0: fair die
- How far does it deviate from uniform distribution?

Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

1	2	3	4	5	6
8	5	9	2	7	5

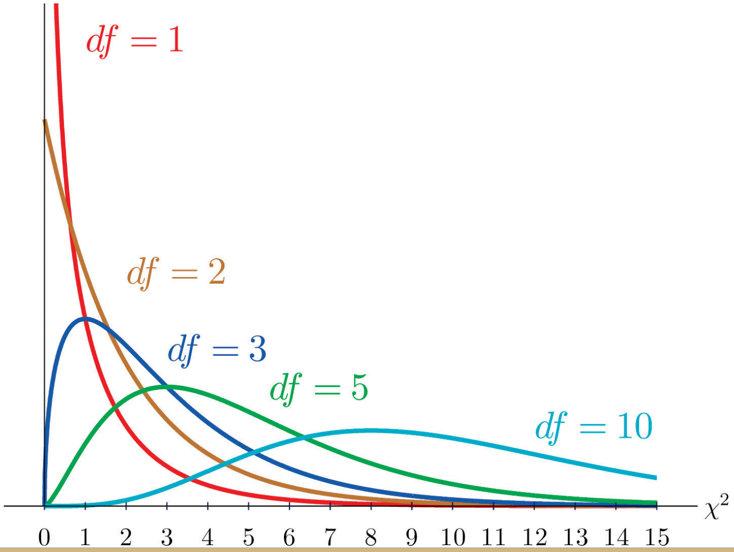
- H_0: fair die
- How far does it deviate from uniform distribution?
- χ^2 distribution

Chi-Square Definition

Let $Z_1, \ldots Z_n$ be independent random variables distributed $N(0,1)$. The χ^2 distribution with n degrees of freedom can be defined by

$$\chi_n^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{1}$$

Chi-Square Definition



Chi-Square Distributions

PDF

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\{-x/2\}$$

CDF

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \gamma\left(\frac{n}{2}, \frac{x}{2}\right)$$

- $\gamma(s,x) \equiv \int_0^x t^{s-1} \exp\{-t\}\, dt$
- $\Gamma(x) \equiv \int_0^\infty t^{x-1} \exp\{-t\}\, dt$, $\Gamma(n) = (n-1)!$

Goodness of Fit

	1	2	3	4	5	6
Observed	8	5	9	2	7	5
Expected	6	6	6	6	6	6

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

Goodness of Fit

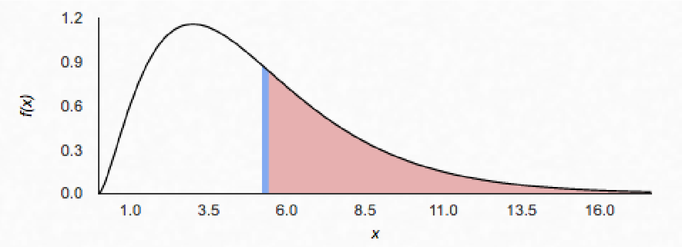
	1	2	3	4	5	6
Observed	8	5	9	2	7	5
Expected	6	6	6	6	6	6

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

- In our example, 5.33
- Approximately distributed as χ^2 with $k-1$ degrees of freedom

Test Statistic and p-value



- Expected value of χ^2 with df=5 is 5
- 5.33 is not that far away
- 0.38 probability of rejecting the null

Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of $k-1$ degrees of freedom

Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of $k-1$ degrees of freedom
- Important because it specifies which χ^2 distribution to use