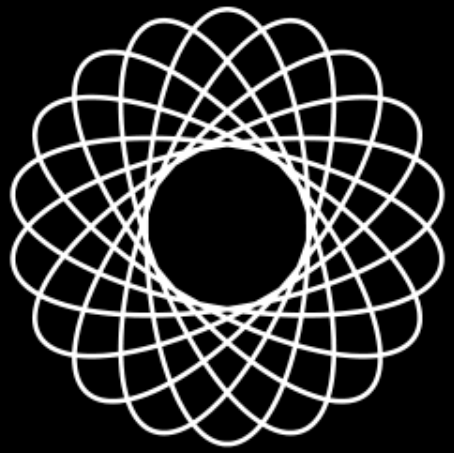
DATA SCIENCE



DATA SCIENCE

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Agenda

Anova
- One Way
- Two Way
- Post Hoc Tests

Chi Square
- Association Tests
- Goodness-of-Fit Tests

Chi Square Parametric
- Tests of Variance

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data

- **E.G –** is there a difference in the number of people responding Yes / No to a direct mail offer, based on income categories?

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data
- **E.G –** is there a difference in the number of people responding Yes / No to a direct mail offer, based on income categories?
- The DV or the target variable outcome is a frequency count

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data
- **E.G –** is there a difference in the number of people responding Yes / No to a direct mail offer, based on income categories?
- The DV or the target variable outcome is a frequency count
- There are two common applications of Chi Square tests:

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data
- **E.G –** is there a difference in the number of people responding Yes / No to a direct mail offer, based on income categories?
- The DV or the target variable outcome is a frequency count
- There are two common applications of Chi Square tests:
  - Tests of association

Chi-Square Tests

- Chi Square tests are multiple sample tests used when dealing with count or categorical data
- **E.G –** is there a difference in the number of people responding Yes / No to a direct mail offer, based on income categories?
- The DV or the target variable outcome is a frequency count
- There are two common applications of Chi Square tests:
  - Tests of association
  - Goodness-of-Fit

Chi-Square Tests

Example using categorical or tabular Data

As a retailer you look at brand ROI to assess shelf space effectiveness. Looking at a particular category, carbonated beverages, you know across all your stores the share of wallet for top Brands A, B and all other (C) is as listed in the first table.

Chi-Square Tests

Example using categorical or tabular Data

As a retailer you look at brand ROI to assess shelf space effectiveness. Looking at a particular category, carbonated beverages, you know across all your stores the share of wallet for top Brands A, B and all other (C) is as listed in the first table.

Brand	Transaction Share
A	52%
B	35%
C (All Other)	13%

Chi-Square Tests

Example using categorical or tabular Data

As a retailer you look at brand ROI to assess shelf space effectiveness. Looking at a particular category, carbonated beverages, you know across all your stores the share of wallet for top Brands A, B and all other (C) is as listed in the first table.

Brand	Transaction Share
A	52%
B	35%
C (All Other)	13%

You take a random sample of data from a particular store – 300 purchases of carbonated beverages.

Before you can start on any analysis, you first need to check if this difference implies this store is not like the population

Brand	# of Transactions	%
A	177	59%
B	78	26%
C (All Other)	45	15%

Random sample of 300 transactions from Store XXX

Chi-Square Tests

- The idea is to check the difference between what you see in your sample v/s what you expected in your sample, and then assess the chances of seeing that difference purely by chance

Column 1	Brand A	Brand B	Brand C
Observed	177	78	45
Expected	156	105	39

Chi-Square Tests

- The idea is to check the difference between what you see in your sample v/s what you expected in your sample, and then assess the chances of seeing that difference purely by chance

- If there was no difference between this store and all the other stores, what would be expect to see as the # of transactions for Brands A, B and all other (C)?

Column 1	Brand A	Brand B	Brand C
Observed	177	78	45
Expected	156	105	39

Chi-Square Tests

- A chi square test uses these “observed” and “expected” frequencies, to generate a conclusion about the statistical significance of the observed differences

Column 1	Brand A	Brand B	Brand C
Observed	177	78	45
Expected	156	105	39

Chi-Square Tests

- A chi square test uses these “observed” and “expected” frequencies, to generate a conclusion about the statistical significance of the observed differences

Mathematically, the quantity $$\sum \frac{(f_o - f_e)^2}{f_e}$$

Column 1	Brand A	Brand B	Brand C
Observed	177	78	45
Expected	156	105	39

Chi-Square Tests

- A chi square test uses these “observed” and “expected” frequencies, to generate a conclusion about the statistical significance of the observed differences

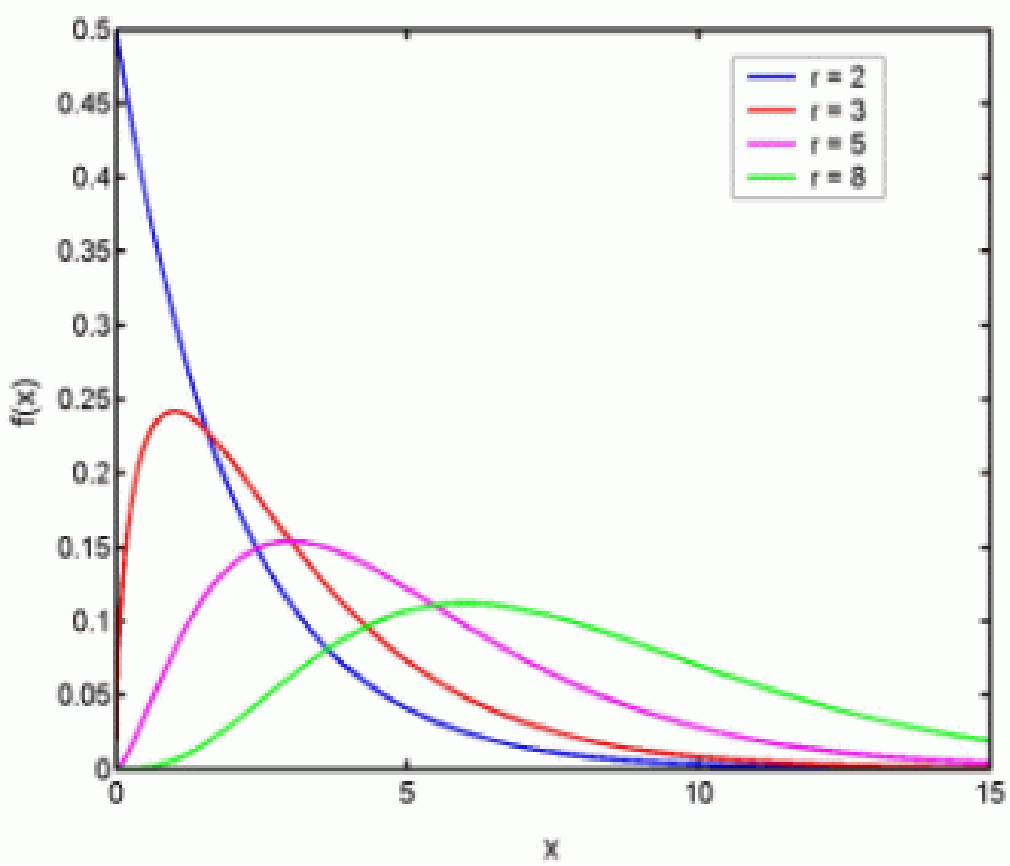
Mathematically, the quantity $$\sum \frac{(f_o - f_e)^2}{f_e}$$

follows a Chi Square Distribution, with k -1 degrees of freedom

Column 1	Brand A	Brand B	Brand C
Observed	177	78	45
Expected	156	105	39

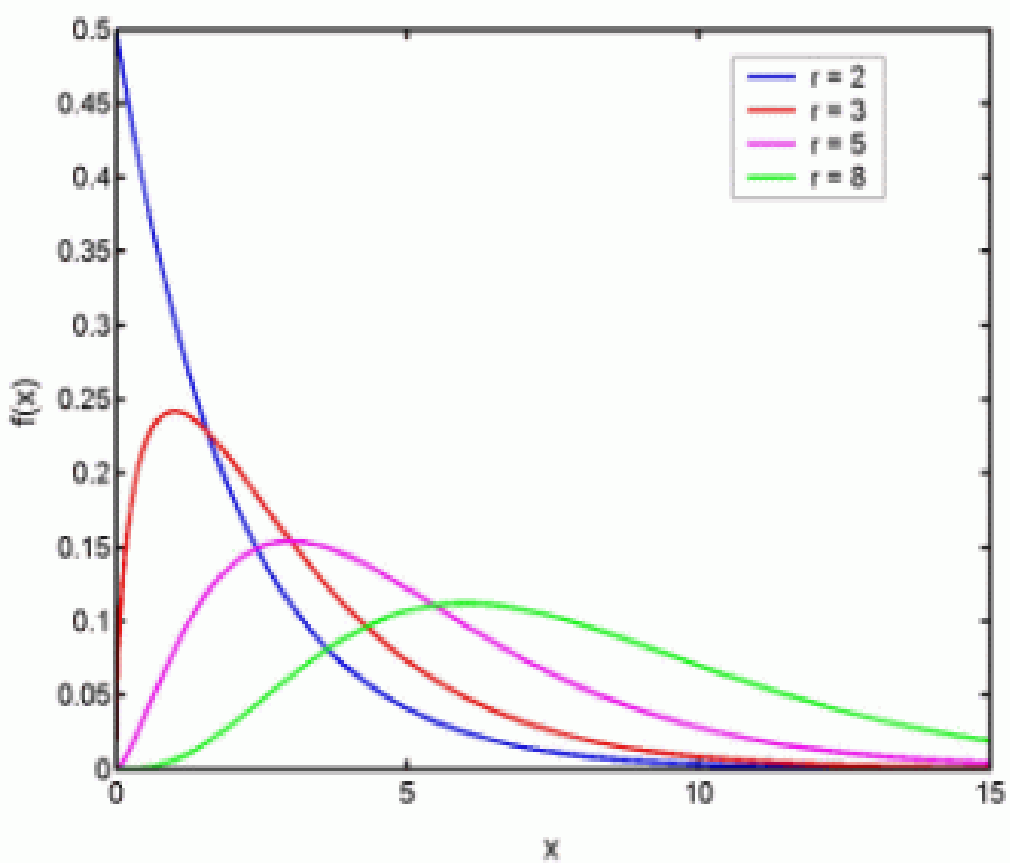
Chi-Square Tests

- A Chi Square distribution is an asymmetric distribution that depends only on sample size
- It is generated as the square of std scores (Z) from a normal distribution
- As sample size increases, Chi Square tends to normal



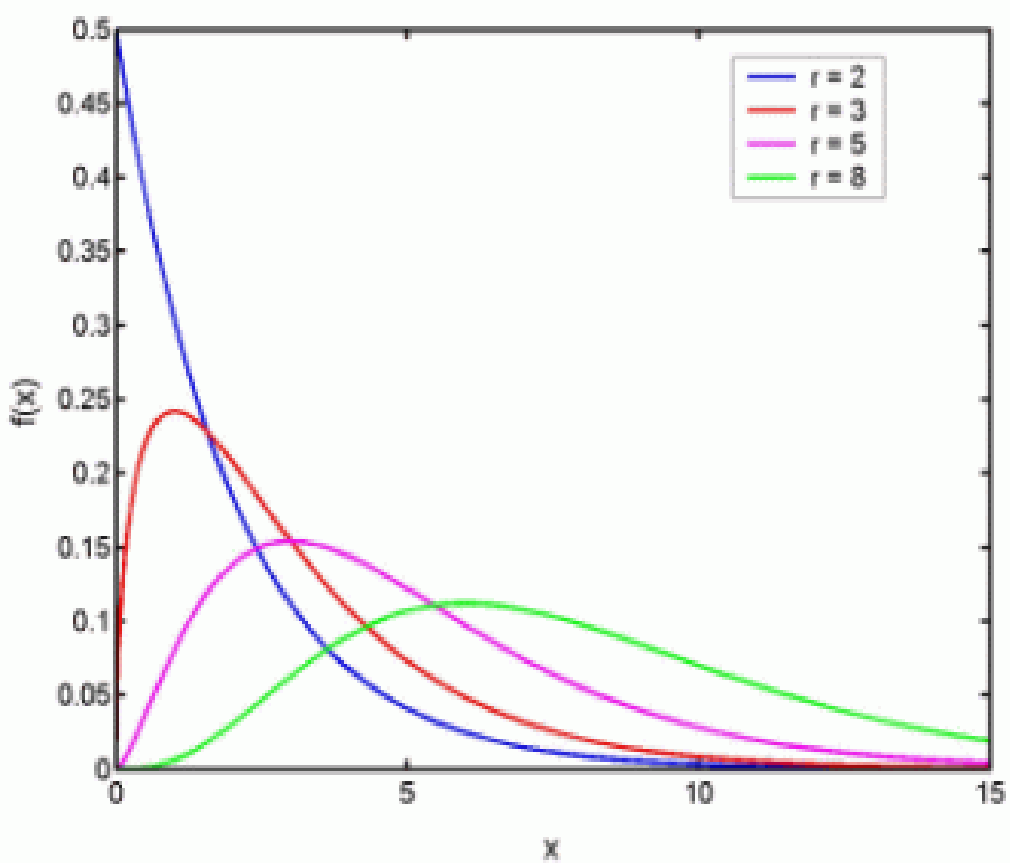
Chi-Square Tests

- A Chi Square distribution is an asymmetric distribution that depends only on sample size



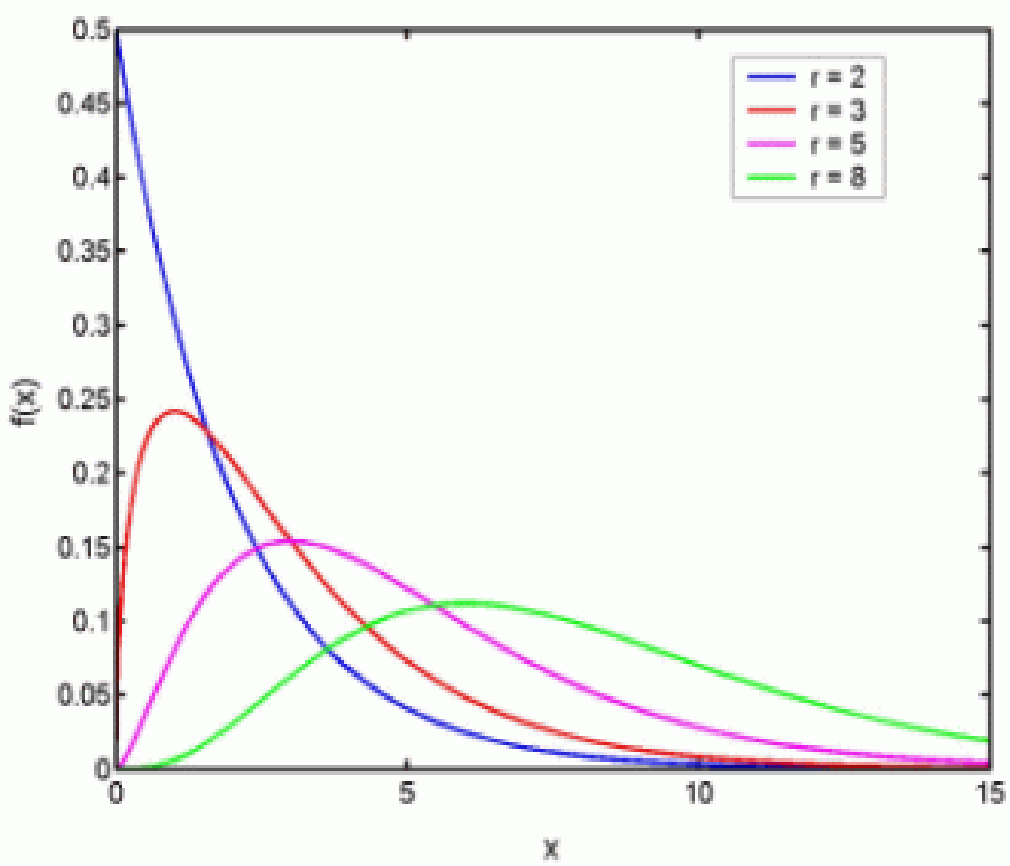
Chi-Square Tests

- A Chi Square distribution is an asymmetric distribution that depends only on sample size
- It is generated as the square of std scores (Z) from a normal distribution



Chi-Square Tests

- A Chi Square distribution is an asymmetric distribution that depends only on sample size
- It is generated as the square of std scores (Z) from a normal distribution
- As sample size increases, Chi Square tends to normal



Chi-Square Tests

The Chi-Square statistic is built as:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

For our example, Chi-Square Test Statistic:

Chi-Square Tests

The Chi-Square statistic is built as:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

For our example, Chi-Square Test Statistic:

(177-156)^2/156 + (78-105)^2/105 + (45-39)^2/39 = 10.69

Chi-Square Tests

The Chi-Square statistic is built as:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

For our example, Chi-Square Test Statistic:

(177-156)^2/156 + (78-105)^2/105 + (45-39)^2/39 = 10.69

To use a table: also need df

Degrees of Freedom = Number of cells - 1 = 3 - 1 = 2

Chi-Square Tests

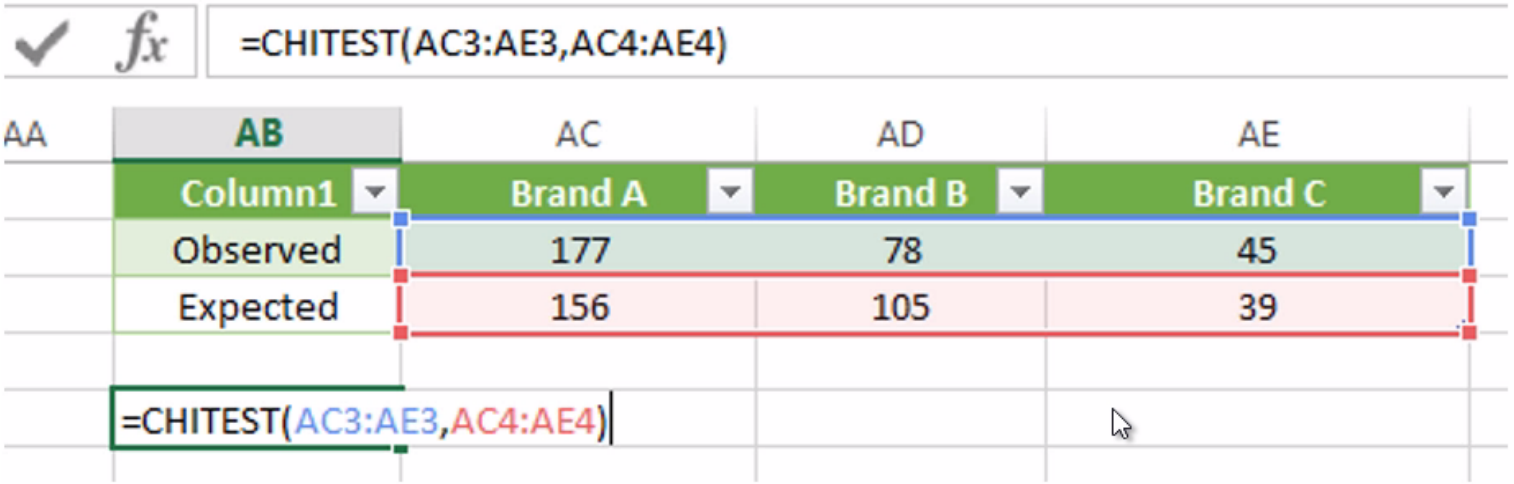
Chi Square Distribution Table

d.f.	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3

1. What was our null hypothesis?
2. What is the critical value here at the 5% significance level?
3. What is the conclusion based on your test statistic?

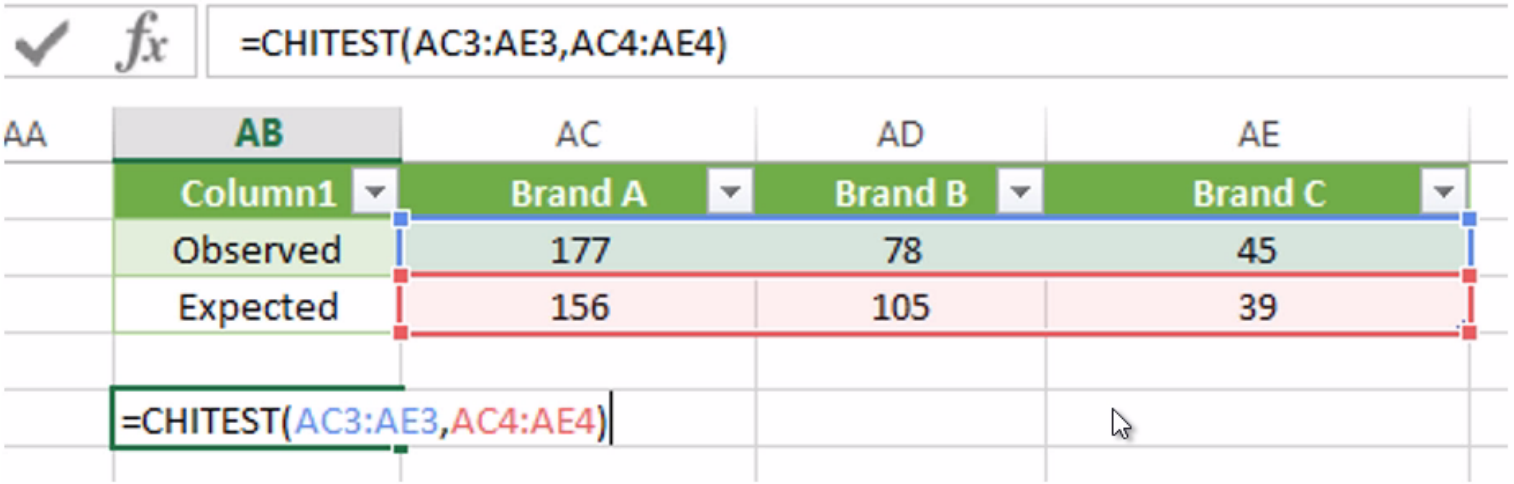
Chi-Square Tests

Using Excel:



Chi-Square Tests

Using Excel:



This will generate a p-value directly. In this example: 0.004765139

Coming Up

Chi Square:

- Association Tests
- Goodness-of-Fit Tests

THANK YOU