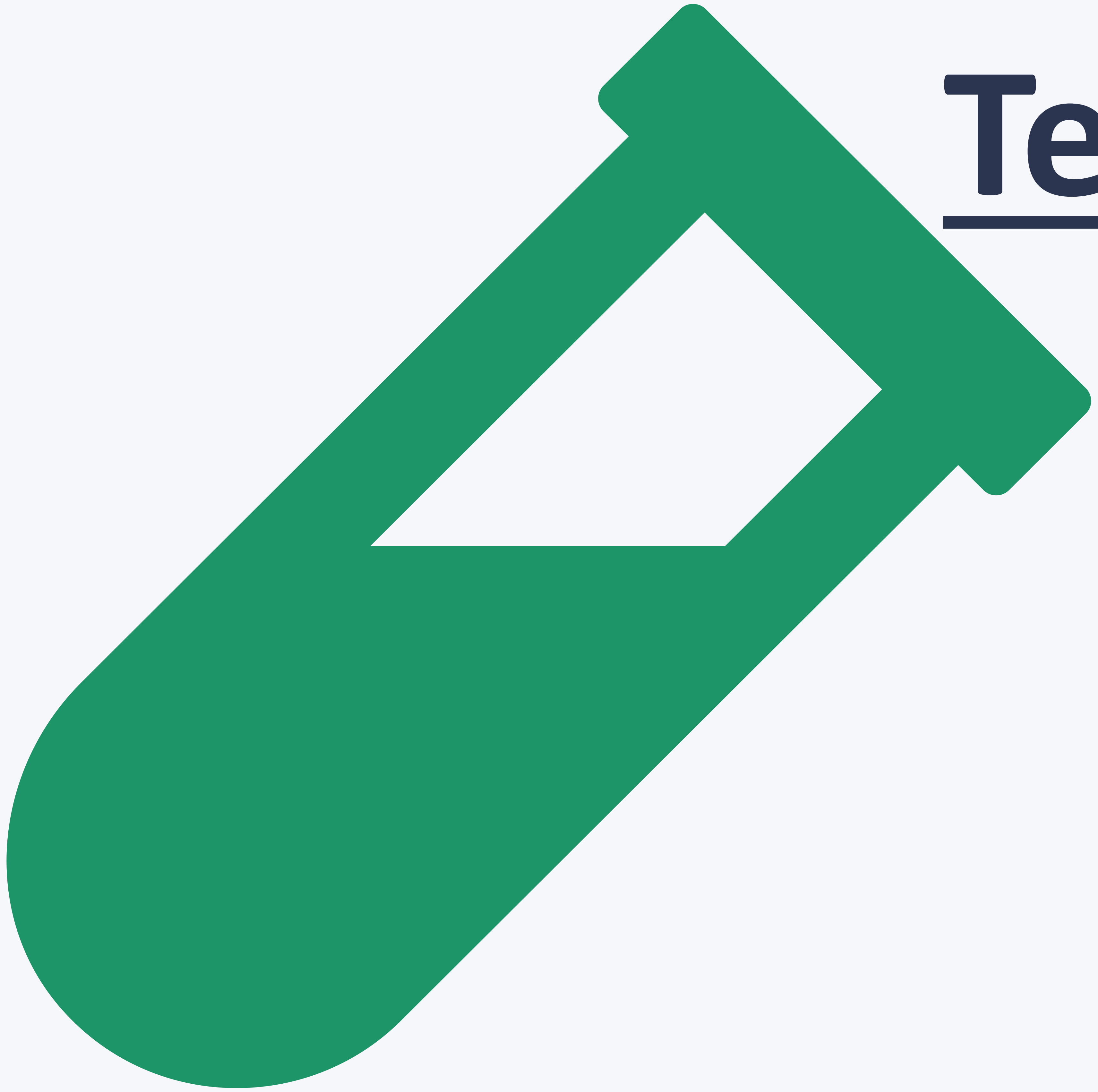


Test for Normality and Chi-Square Tests

Statistics Tutorial

Day 6

Prabesh Dhakal

2020 May 14

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

Test for Normality and Chi-square Tests

We talk about data distribution. We also talk about hypothesis testing.

EXERCISE

Applying

Q&A and Recap

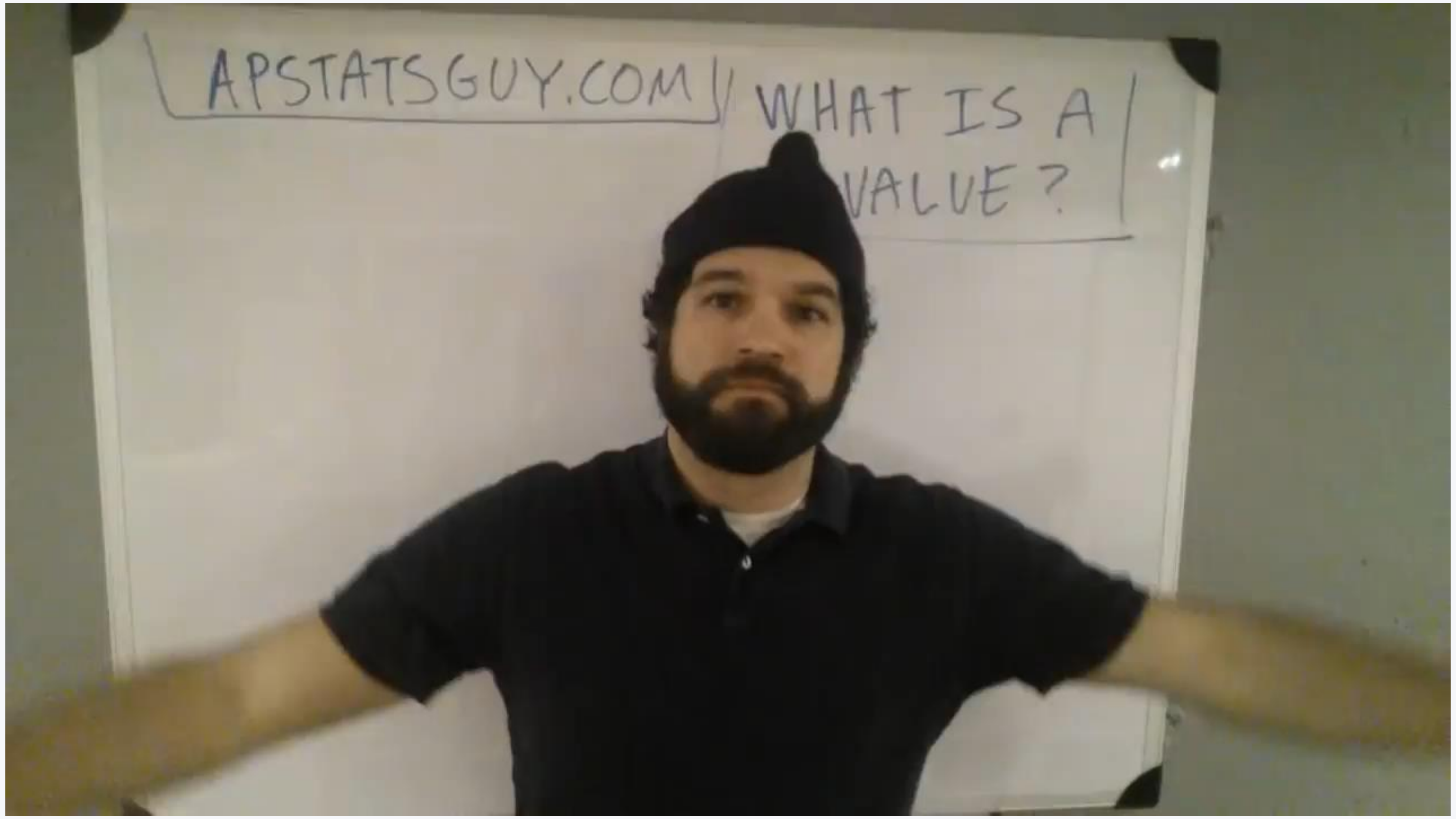
Please ask if you have any questions now.

Otherwise, we can move on to the recap.

p-value

probability of you making the observations **if H_0 were true**

$$p-value \ = \ P(data \mid H_0 \ is \ true)$$



Video source: https://www.youtube.com/watch?v=-MKT3yLDkqk

SIGNIFICANCE LEVEL (α) AND $p - value$

When $p - value \leq \alpha$, we reject H_0

- The result is statistically significant
  - We are reasonably sure that there is something besides chance that gave us an observed sample

When $p - value > \alpha$, we fail to reject the H_0

- The result is not statistically significant.
  - We are reasonably sure that our observed data can be observed by chance alone

Hypothesis Testing

Tests for Normality

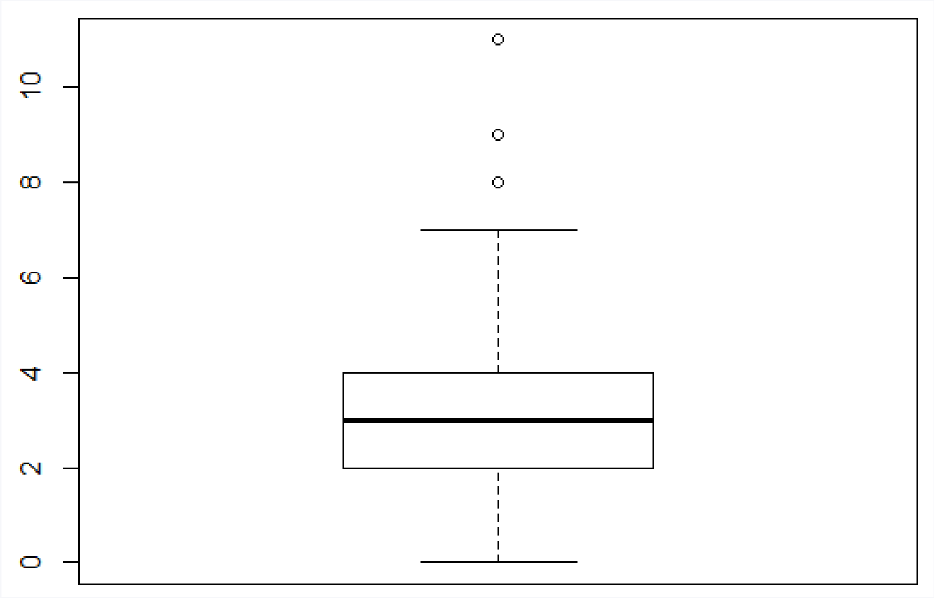
Chi-square Tests

TESTS OF NORMALITY

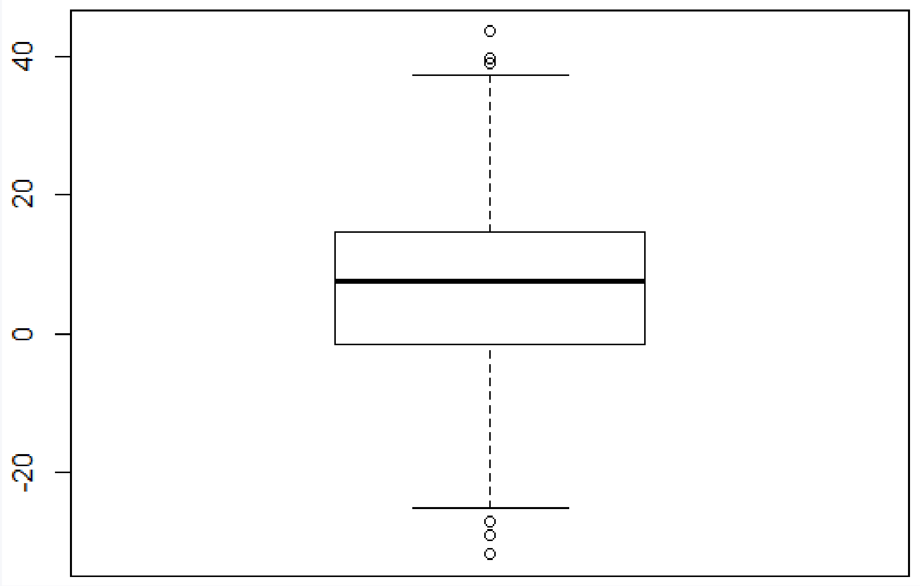
1. **Visual Approach**
   - Box Plot
   - Histogram
   - Density Plot
   - Q-Q Plot

2. **Inferential Approach**
   - Shapiro-Wilk's Test

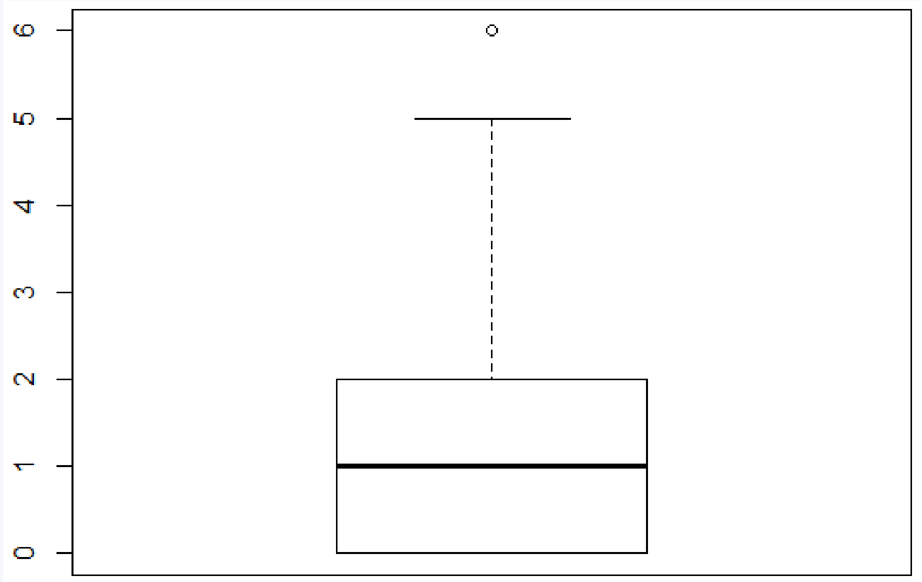
WHICH OF THESE ARE NORMALLY DISTRIBUTED?



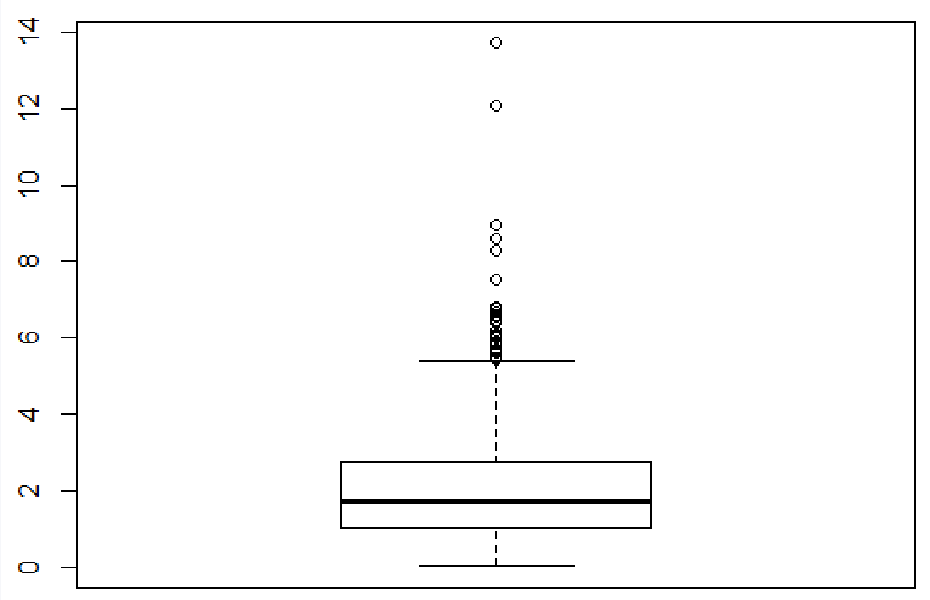
A



B



C

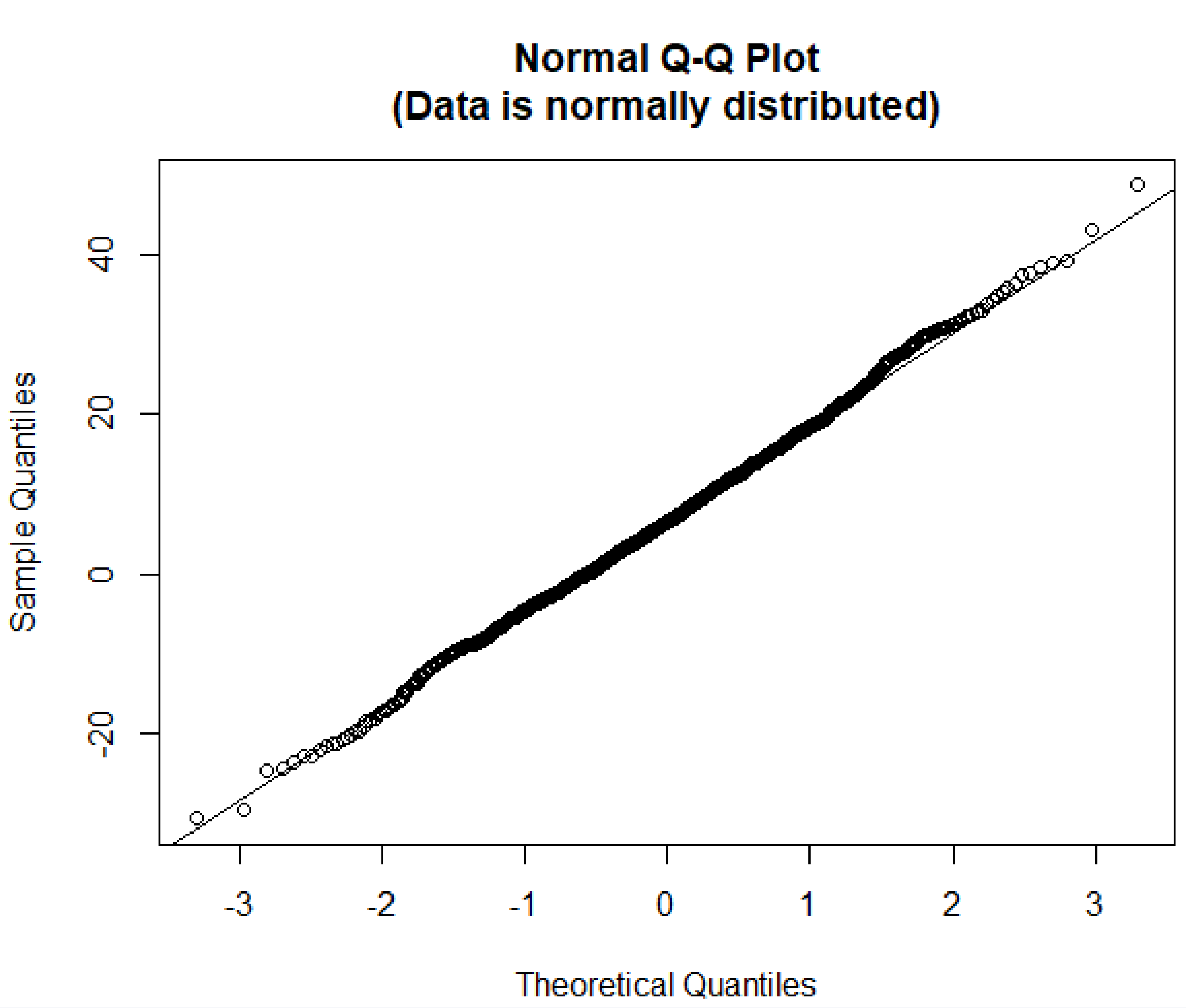


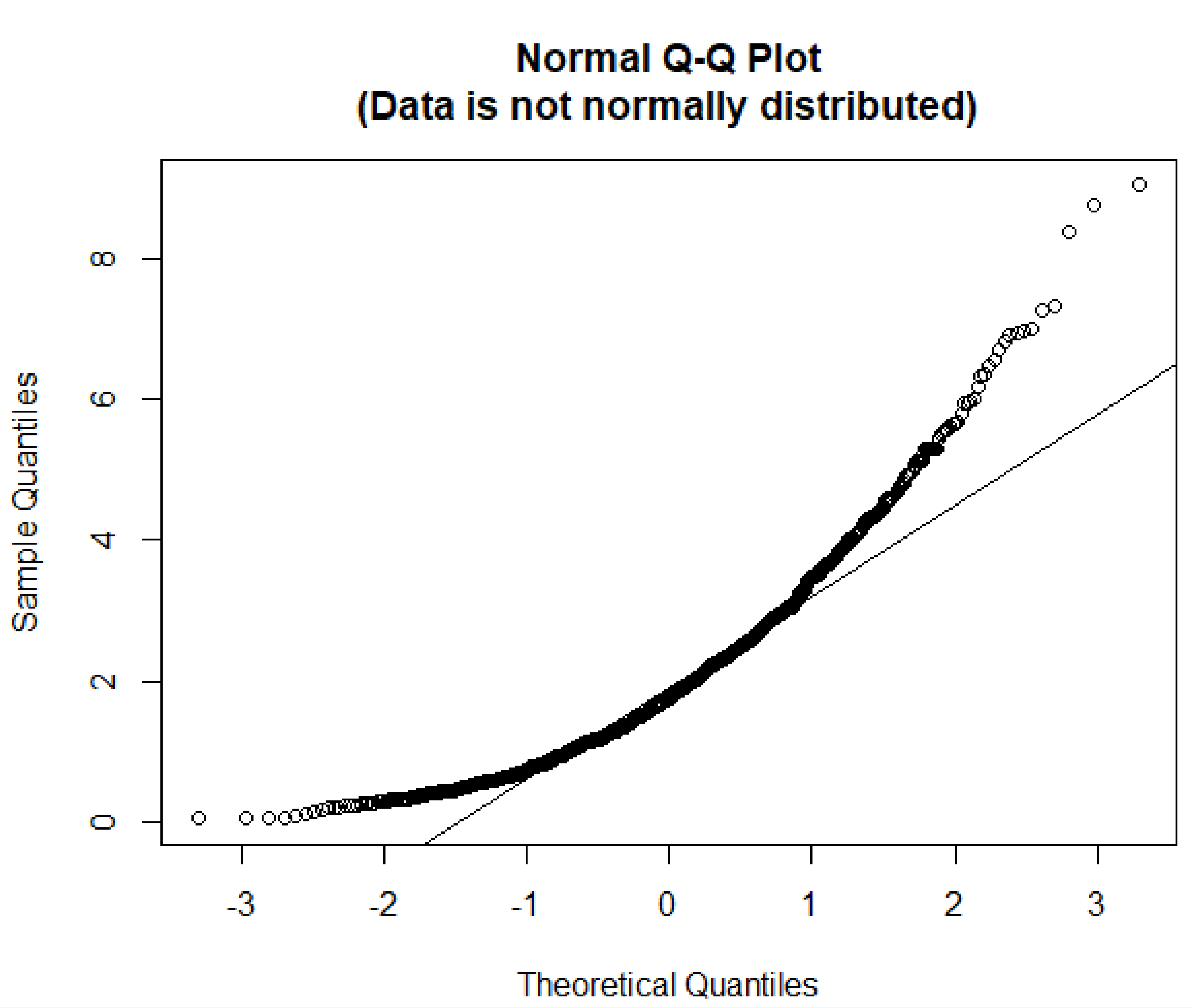
D

Q-Q PLOTS

An alternative graphical method that works well for small sample size.

It compares the data to a perfect normal distribution.





SHAPIRO-WILK'S TEST FOR NORMALITY

Works well for smaller data sets. However, you get more false negatives with larger datasets.

- **H0 : The data comes from a normally distributed population**
- **H1 : The data does not come from a normally distributed population**

Fail to reject H0 if $p - value \ <= \ 0.05$

CHI-SQUARE TESTS

Allows you to look at differences between categorical variables

E.g. Gender, political differences, etc.

Test Statistic:

$$\chi^2 \; test\ statistic = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

Types of Test:

- Goodness of Fit Test
- Test of Independence

GOODNESS OF FIT TEST - I

- Compared observed distribution of the data against the expected distribution

H0 : The data follows the "expected"/specified distribution

H1 : The data does not come follow the specified distribution

$$\chi^2 \; test\ statistic = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

$\boldsymbol{E_i} = \boldsymbol{N} * \boldsymbol{p_i}$ for bin $\boldsymbol{i}$
$\boldsymbol{N}$ = the sample size
$\boldsymbol{p_i}$ = hypothesized distribution of bin $\boldsymbol{i}$

Degrees of freedom (df) = no. of categories – 1 = k - 1

GOODNESS OF FIT TEST - II

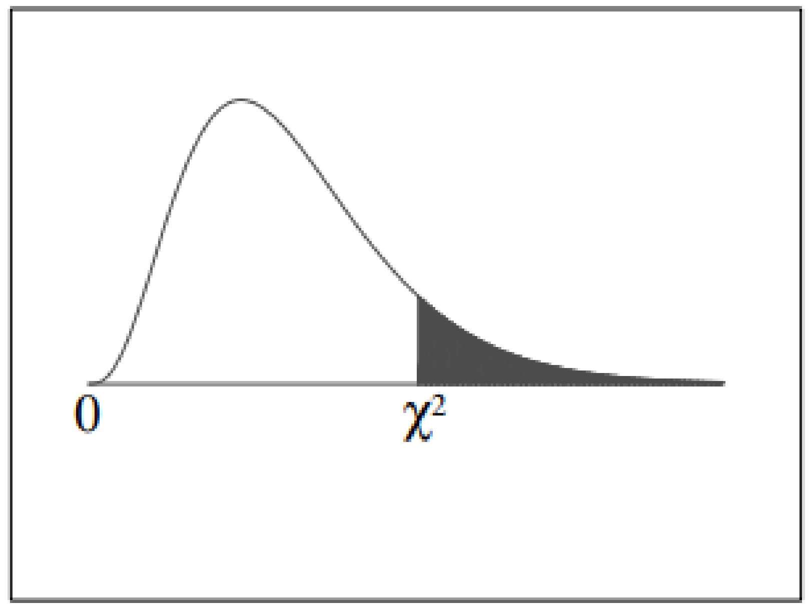
H0 : The data follows the “expected”/specified distribution

H1 : The data does not come follow the specified distribution

$$\chi^2\ test\ statistic = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E}$$

$\boldsymbol{E_i} = \boldsymbol{N} * \boldsymbol{p_i}$ for bin $\boldsymbol{i}$
$\boldsymbol{N}$ = the sample size
$\boldsymbol{p_i}$ = distribution of bin $\boldsymbol{i}$

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\boxed{\chi^2_{.050}}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
$\boxed{2}$	0.010	0.020	0.051	0.103	0.211	4.605	$\boxed{5.991}$	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

http://kisi.deu.edu.tr/joshua.cowley/Chi-square-table.pdf

Rejection Criteria:

- $p - value\ <=\ significance\ level\ (\alpha)$
- $test\ statistic > critical\ value$

GOODNESS OF FIT TEST - EXAMPLE

…

$$\chi^2 \; test\ statistic = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

$\boldsymbol{E_i} = \boldsymbol{N} * \boldsymbol{p_i}$ for bin $\boldsymbol{i}$
$\boldsymbol{N}$ = the sample size
$\boldsymbol{p_i}$ = distribution of bin $\boldsymbol{i}$

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

TEST OF INDEPENDENCE

- Check if two variables are independent

H0 : The two categorical variables are independent

H1 : The two variables are not independent

$$\chi^2 \; test \; statistic = \sum_{i=1}^{N} \frac{\left(O_i - E_{ij}\right)^2}{E_{ij}}$$

$$\boldsymbol{E_{ij}} = \frac{R_i C_j}{N}$$

$\boldsymbol{R}$ = row

$\boldsymbol{C}$ = *column*

Degrees of freedom (df) = (no. of rows - 1)*(no. of columns - 1)

Click here if you want to know how to do this by hand.

Exercise

Perform Shapiro-Wilk's Test and Chi-square Tests in R

ASSIGNMENT

Download the PDF file named “Problem Set 1” from MyStudy.

You can work either individually or in group.

PLAN FOR NEXT WEEK

That's it for today! :-)

Next week, we are going to discuss:

- t-Test
- F-test

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`