STATISTICS

Non-parametric tests

Paweł Strzelecki

Lucas van der Velde

Warsaw School of Economics
Winter Semester 2018

Lecture is based on J.T. Mc Clave, P.G. Benson,T. Sincich: Statistics for Business and Economics, 11th Edition, 2010

Testing the normality assumption

- We often assumed a variable is normally distributed. Such assumption should be tested first!
- There are some simple descriptive methods for assessing the normality. They allow us to get some first insight into the data.
- Nevertheless, if we have sample data and we want to know whether the distribution in the population is normal we should conduct a statistical test (non-parametric test).

Outline

- Descriptive methods for assessing normality
- Non-parametric test

Descriptive methods for assessing normality

Descriptive methods for assessing normality

1. Construct a histogram and observe the shape of the graph
2. Compute the intervals $\bar{x} \pm S,\ \bar{x} \pm 2S,\ \bar{x} \pm 3S$ and determine the proportions of the data falling into each (these should be 68%, 95% and 100% respectively)
3. Construct a normal probability plot

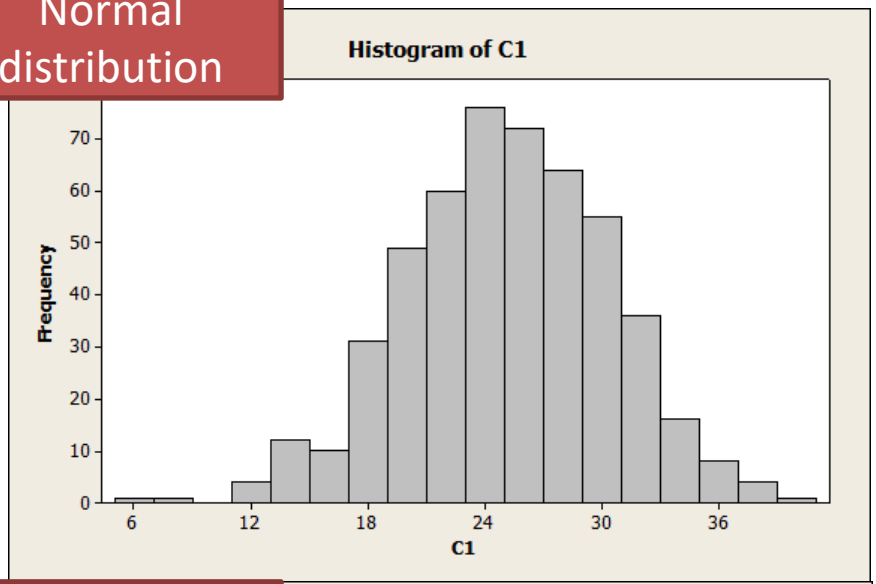
A normal probability plot

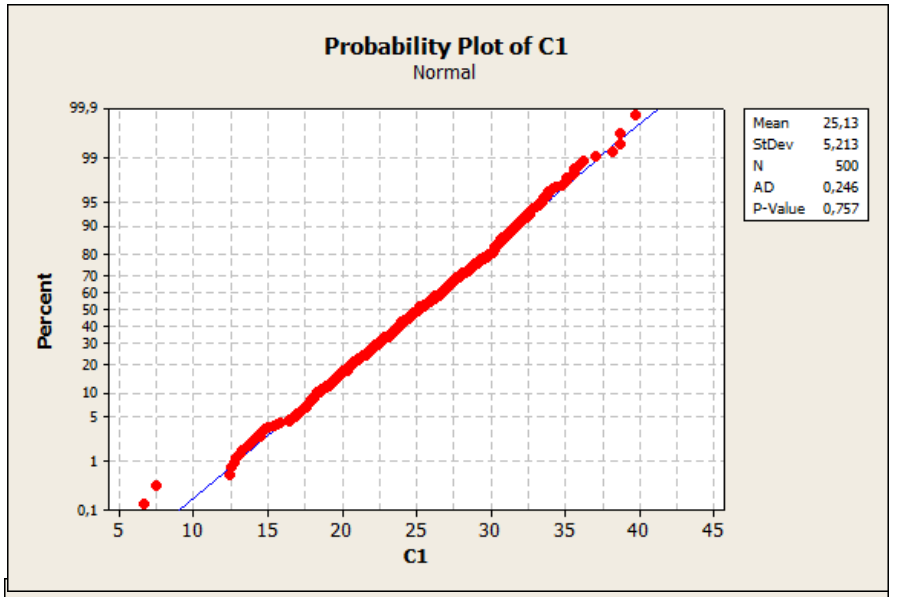
A normal probability plot is a scatterplot with the ranked data values on X axis and their corresponding expected z-scores from the normal distribution on the Y axis.

The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

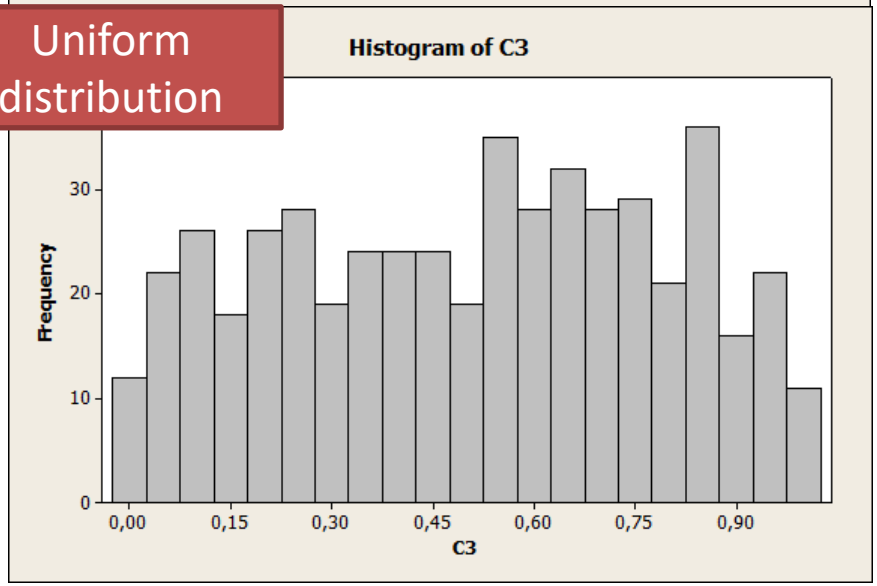
A normal probability plot

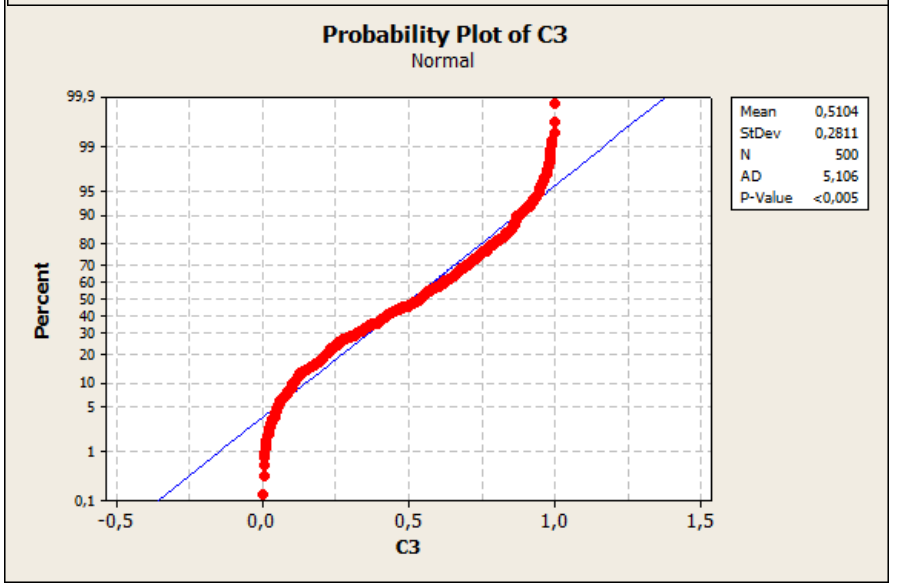
Normal distribution





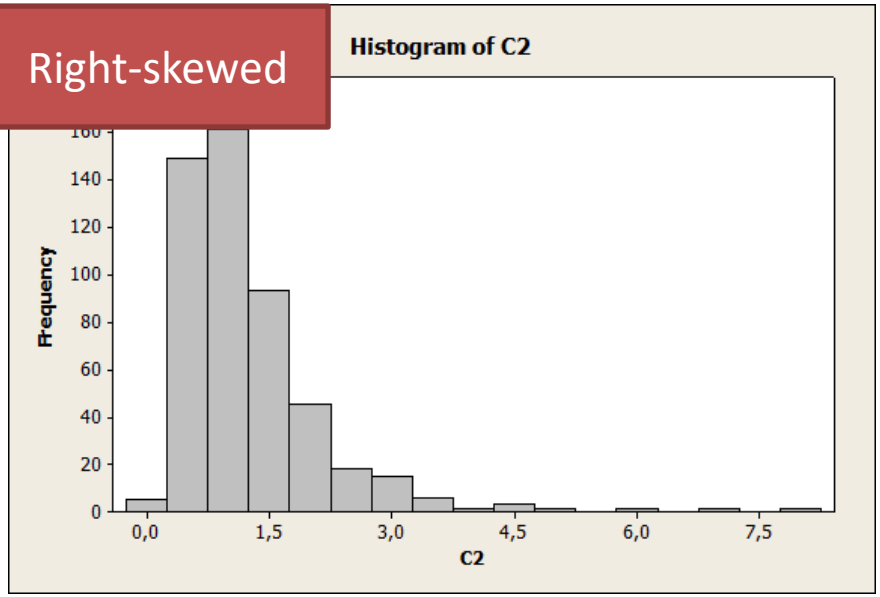
Uniform distribution

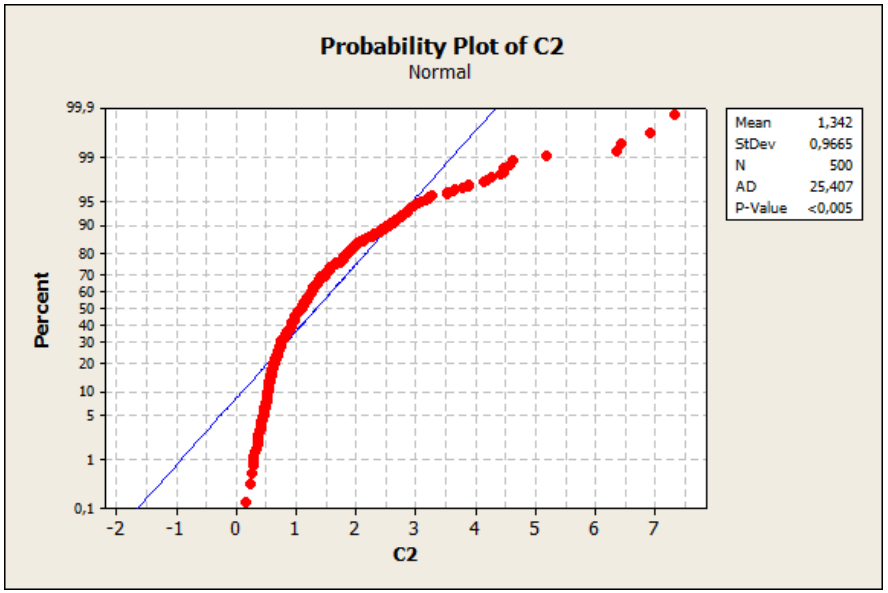




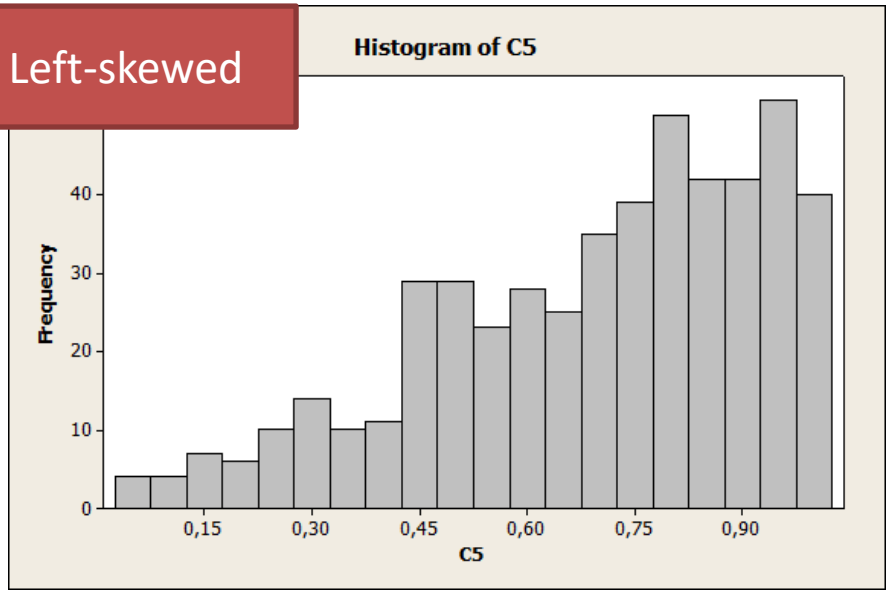
A normal probability plot

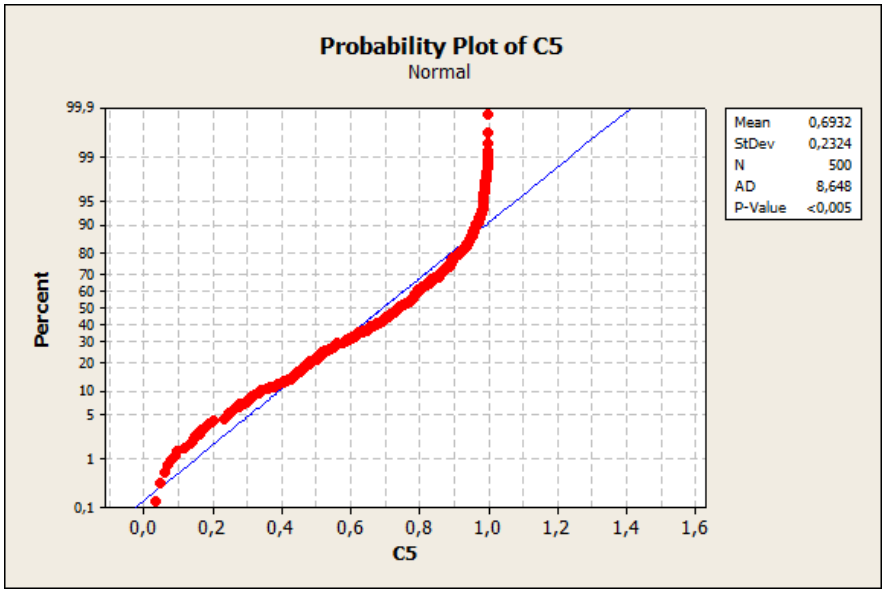
Right-skewed





Left-skewed





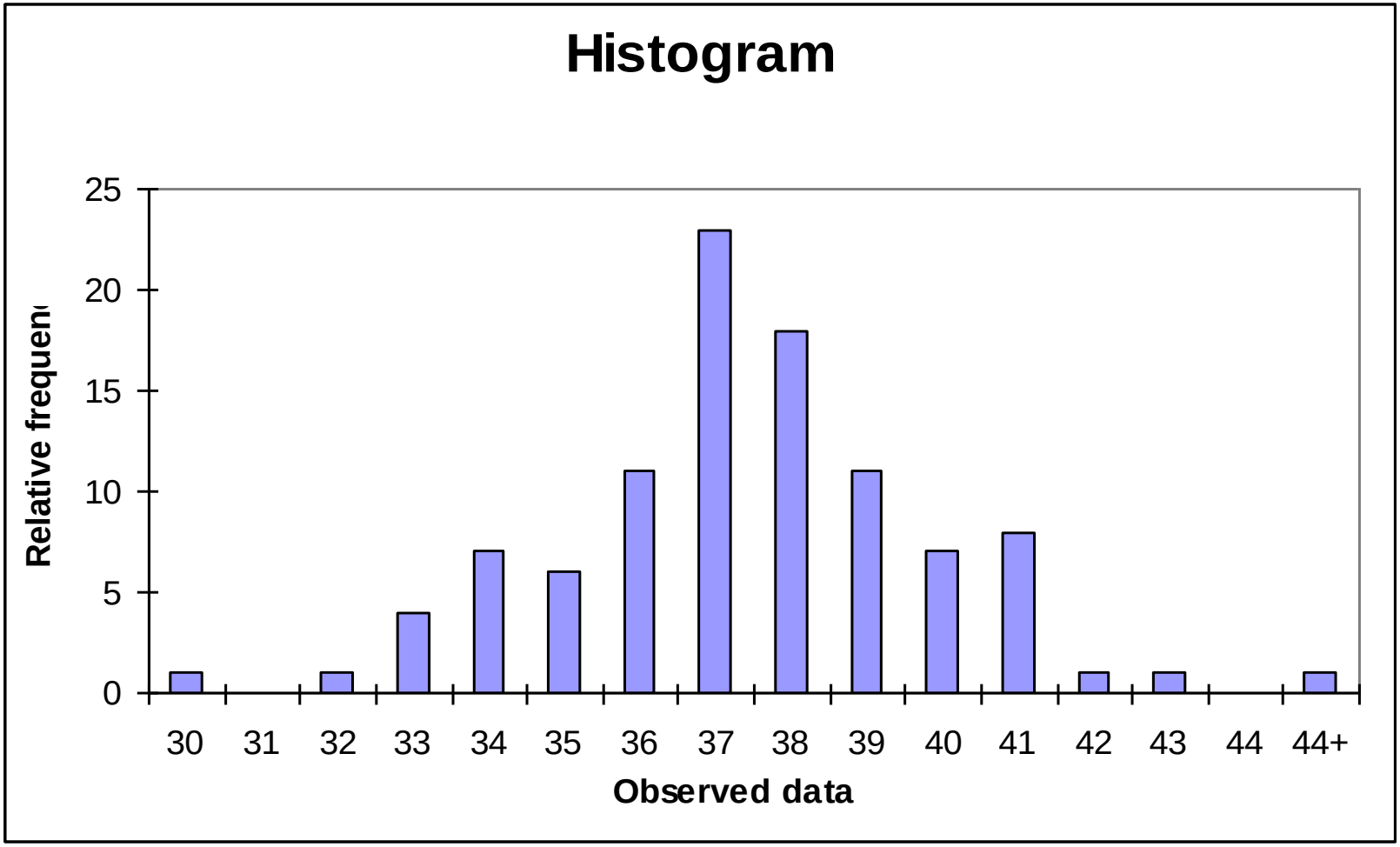
Example

The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. The results of 100 EPA tests on a certain new car model are displayed in table below. Using descriptive methods for assessing normality determine whether the mileage ratings in the sample of 100 EPA tests were normally distributed.

TABLE 5.2 EPA Gas Mileage Ratings for 100 Cars (miles per gallon)

36.3	41.0	36.9	37.1	44.9	36.8	30.0	37.2	42.1	36.7
32.7	37.3	41.2	36.6	32.9	36.5	33.2	37.4	37.5	33.6
40.5	36.5	37.6	33.9	40.2	36.4	37.7	37.7	40.0	34.2
36.2	37.9	36.0	37.9	35.9	38.2	38.3	35.7	35.6	35.1
38.5	39.0	35.5	34.8	38.6	39.4	35.3	34.4	38.8	39.7
36.3	36.8	32.5	36.4	40.5	36.6	36.1	38.2	38.4	39.3
41.0	31.8	37.3	33.1	37.0	37.6	37.0	38.7	39.0	35.8
37.0	37.2	40.7	37.4	37.1	37.8	35.9	35.6	36.7	34.5
37.1	40.3	36.7	37.0	33.9	40.1	38.0	35.2	34.8	39.5
39.9	36.9	32.9	33.8	39.8	34.0	36.8	35.0	38.1	36.9

Example



The histogram shows the distribution is symetric.

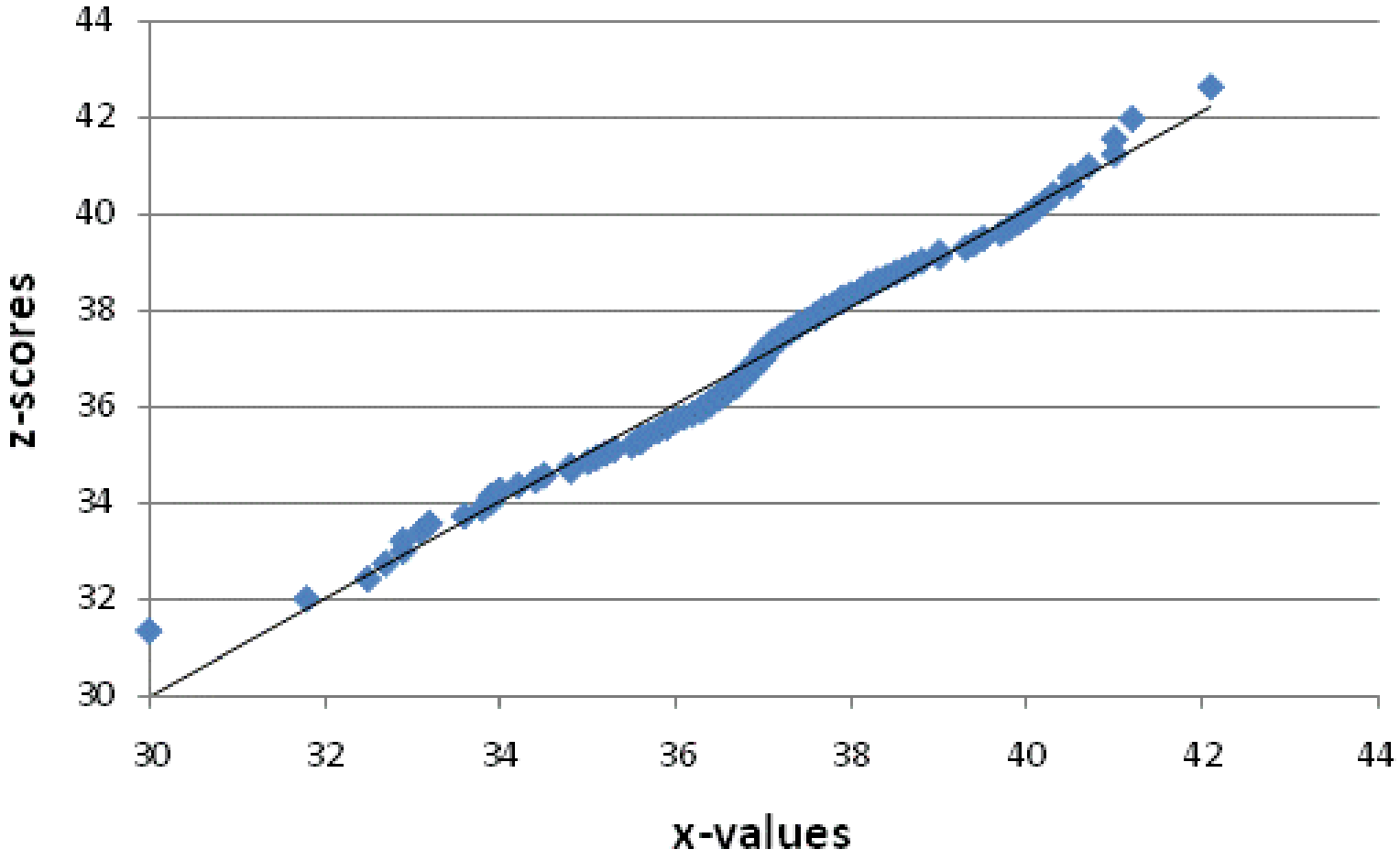
Example

The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. The results of 100 EPA tests on a certain new car model are displayed in table below. Using descriptive methods for assessing normality determine whether the mileage ratings in the sample of 100 EPA tests were normally distributed.

Interval	% in interval
$\overline{x} \pm S = (34.6, 39.4)$	68
$\overline{x} \pm 2S = (32.2, 41.8)$	96
$\overline{x} \pm 3S = (29.8, 44.2)$	99

The table shows that the proportion of observatiosn contained in each interval are close to what they should be according to 3-sigma rule.

Example



The z-scores fall close to a straight line when plotted against the expected z-scores from normal distribution → we can expect the distribution of X to be normal.

Non-parametric test for assessing normality

Chi-square test for assessing normality

H0: F(x) = F0(x)

Ha: F(x) ≠ F0(x)

where F0(x) – cumulative distribution function of a random variable $X \sim N(\mu, \sigma)$

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

where:

n_i observed absolute frequencies in the interval *i*

$\hat{n}_i$ absolute frequencies in the interval *i* to be observed if $X \sim N(\mu, \sigma)$

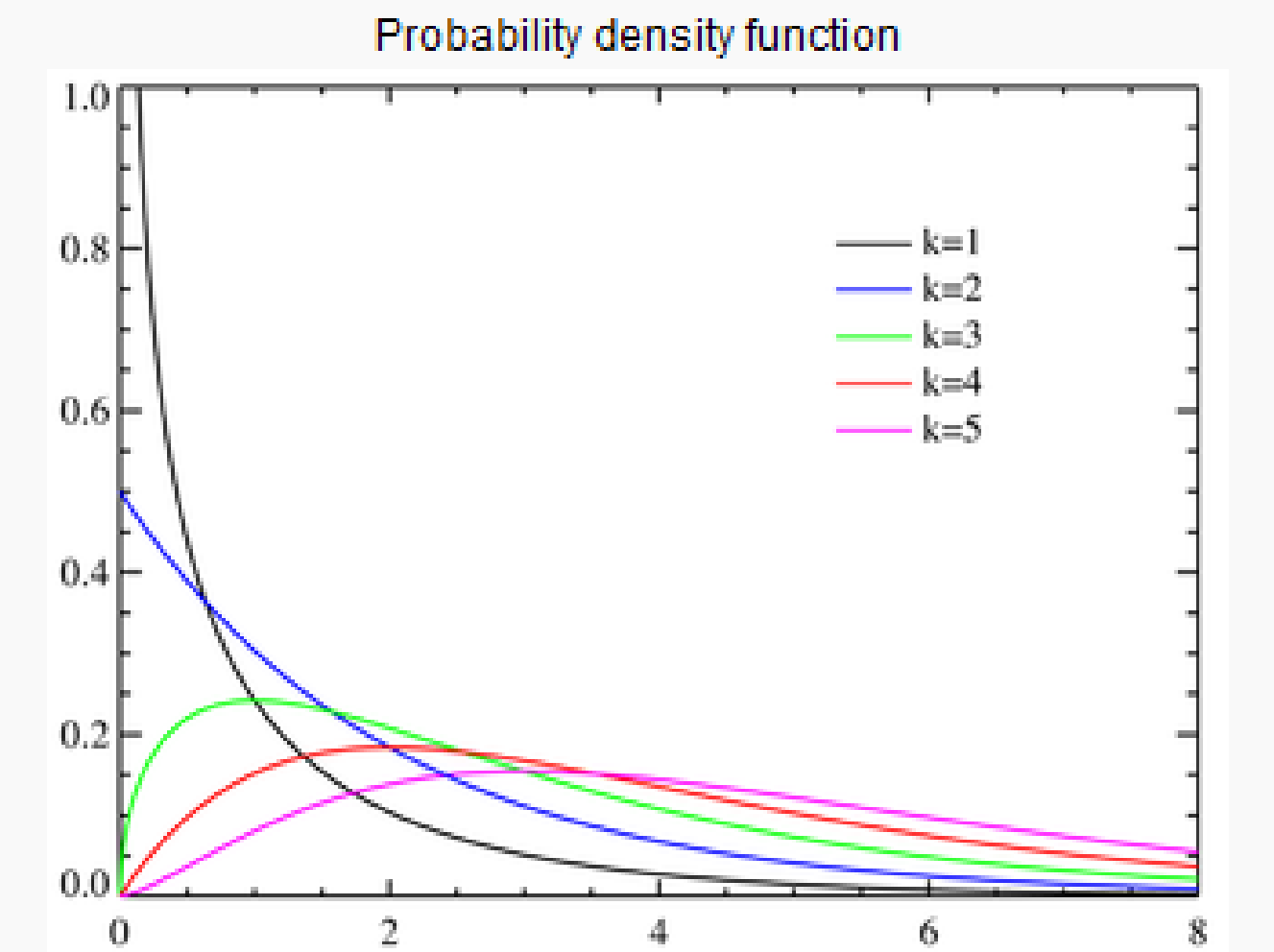
$$\hat{n}_i = n \cdot p_i$$

Rejection region: $\chi^2 > \chi^2_{\alpha, r-k-1}$

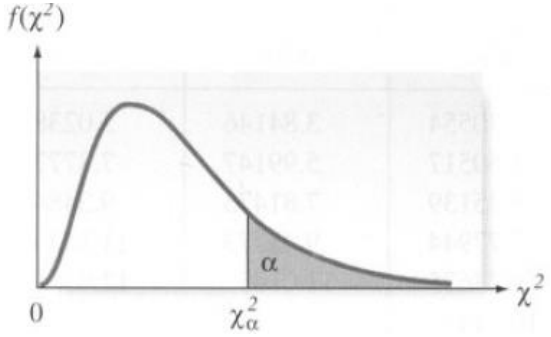
where r-number of intervals X is grouped into, k-number of parameters estimated based on the sample

Chi-square distribution

with k degrees of freedom



Chi-square table



Degrees of Freedom	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672

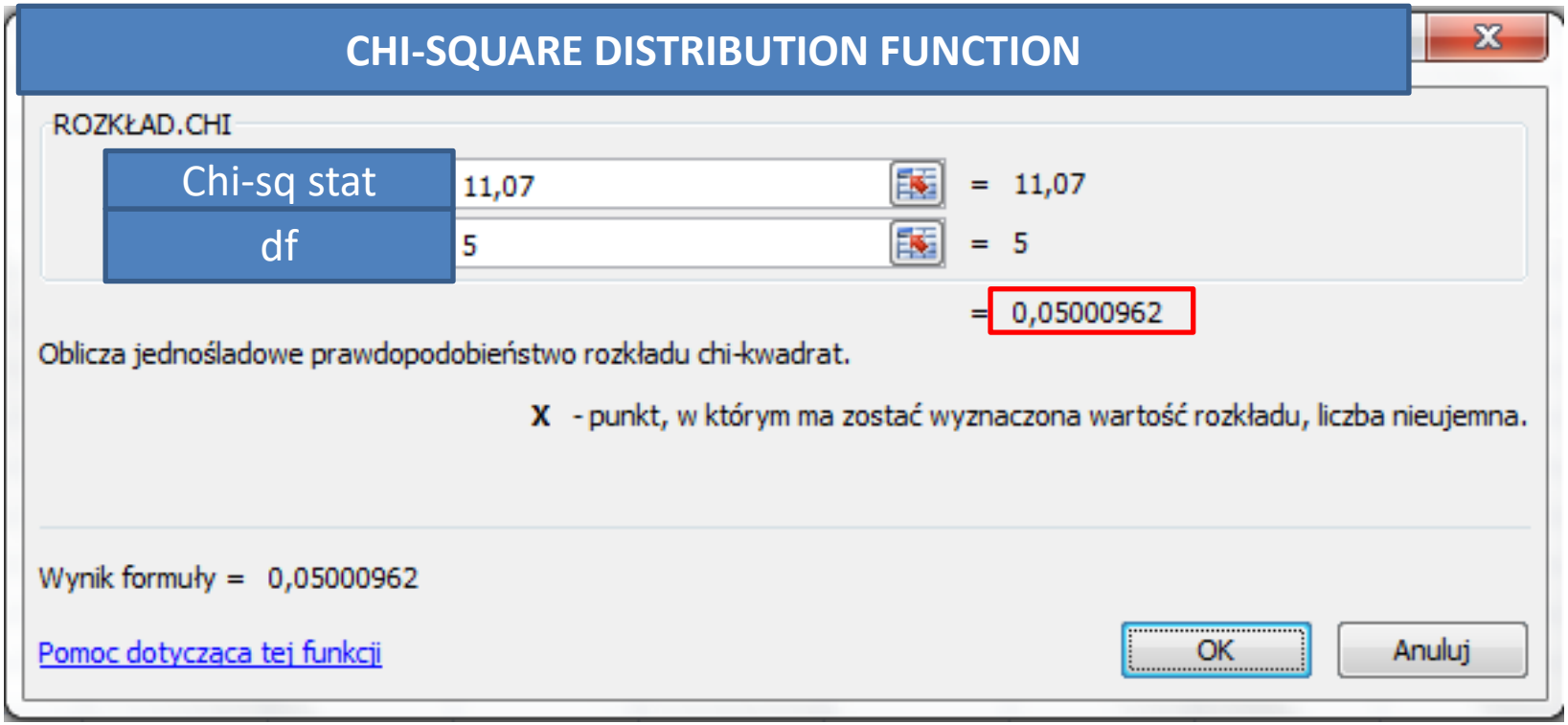
$$P(\chi^2 > \chi^2_{df,\alpha}) = \alpha$$

For instance for df=5:

$$P(\chi^2 > 11.07) = 0.05$$

Calculating probabilitities from Chi-square distribution in Excel

$$P(\chi^2 > 11.07) = 0.05$$



Example

The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. The results of 100 EPA tests on a certain new car model are displayed in table below. Using descriptive methods for assessing normality determine whether the mileage ratings in the sample of 100 EPA tests were normally distributed.

Here we group the data into 11 intervals:

interval	*frequency*
<32	2
32-33	4
33-34	7
34-35	6
35-36	11
36-37	23
37-38	18
38-39	11
39-40	7
40-41	8
>41	3

Example

The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. The results of 100 EPA tests on a certain new car model are displayed in table below. Using descriptive methods for assessing normality determine whether the mileage ratings in the sample of 100 EPA tests were normally distributed.

Here we group the data into 11 intervals:

interval	frequency
<32	2
32-33	4
33-34	7
34-35	6
35-36	11
36-37	23
37-38	18
38-39	11
39-40	7
40-41	8
>41	3

Solution:

H0: F(x) = F0(x)
Ha: F(x) ≠ F0(x)

Test statistic: $\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$

We need to compute $\hat{n}_i$:

$\bar{x} = 37$

$S = 2.4$

$$\hat{n}_1 = n \cdot P(X < 32) = 100 \cdot P(Z < \frac{32-37}{2.4}) = 100 \cdot 0.019 = 1.9$$

$$\hat{n}_2 = n \cdot P(32 < X < 33) = 100 \cdot \left(P(Z < \frac{33-37}{2.4}) - P(Z < \frac{32-37}{2.4}) \right) = 100 \cdot 0.0298 = 2.98$$

etc.

Example

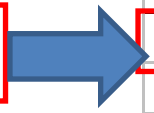
interval	*frequency*	*probabilities*	*freq observed if X normal*
<32	2	0.0194	2
32-33	4	0.0298	3
33-34	7	0.0585	6
34-35	6	0.0970	10
35-36	11	0.1357	14
36-37	23	0.1605	16
37-38	18	0.1603	16
38-39	11	0.1353	14
39-40	7	0.0965	10
40-41	8	0.0581	6
>41	3	0.0488	5

Attention!

Chi-square tests may give wrong results if hypothethical frequencies <5. In such a situation we merge the neighbouring intervals

Example

interval	*frequency*	*probabilities*	*freq observed if X normal*
<32	2	0.0194	2
32-33	4	0.0298	3
33-34	7	0.0585	6
34-35	6	0.0970	10
35-36	11	0.1357	14
36-37	23	0.1605	16
37-38	18	0.1603	16
38-39	11	0.1353	14
39-40	7	0.0965	10
40-41	8	0.0581	6
>41	3	0.0488	5



interval	*frequency*	*freq observed if X normal*
<33	6	5
33-34	7	6
34-35	6	10
35-36	11	14
36-37	23	16
37-38	18	16
38-39	11	14
39-40	7	10
40-41	8	6
>41	3	5

Example

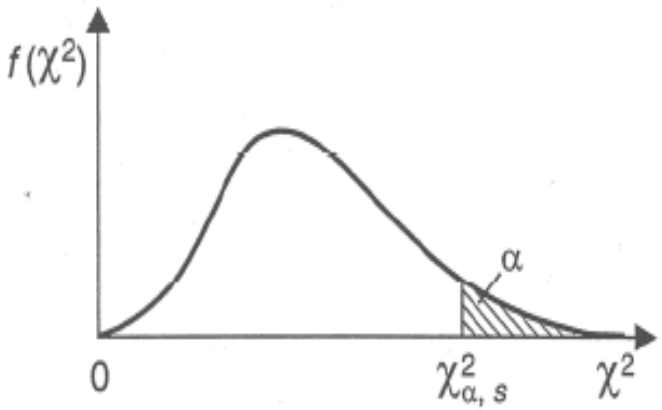
interval	frequency	freq observed if X normal
<33	6	5
33-34	7	6
34-35	6	10
35-36	11	14
36-37	23	16
37-38	18	16
38-39	11	14
39-40	7	10
40-41	8	6
>41	3	5

Solution:

H0: F(x) = F0(x)
Ha: F(x) ≠ F0(x)

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 8.32$$



Calculating p-value:

$$p - value = P(\chi^2 > 8.32) = 0.3052 > 0.1$$

The sample gives us no evidence to reject H0 about the normality of the distribution. This conclusion is consistent to what we observed on the basis of descriptive analysis.

Calculating probabilitities from Chi-square distribution in Excel

$$P(\chi^2 > 8.32) = 0.3052$$