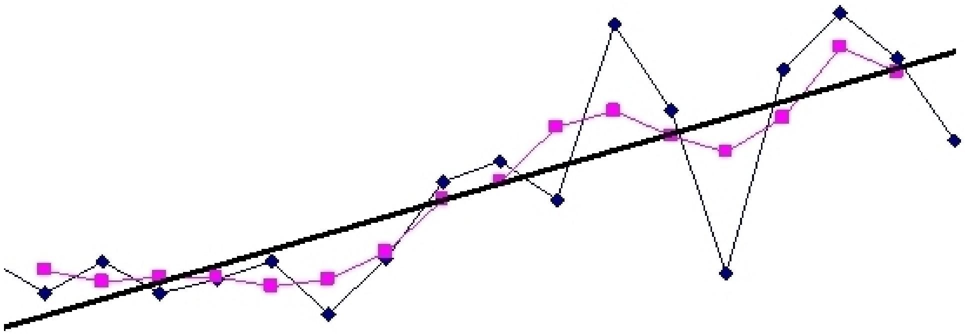
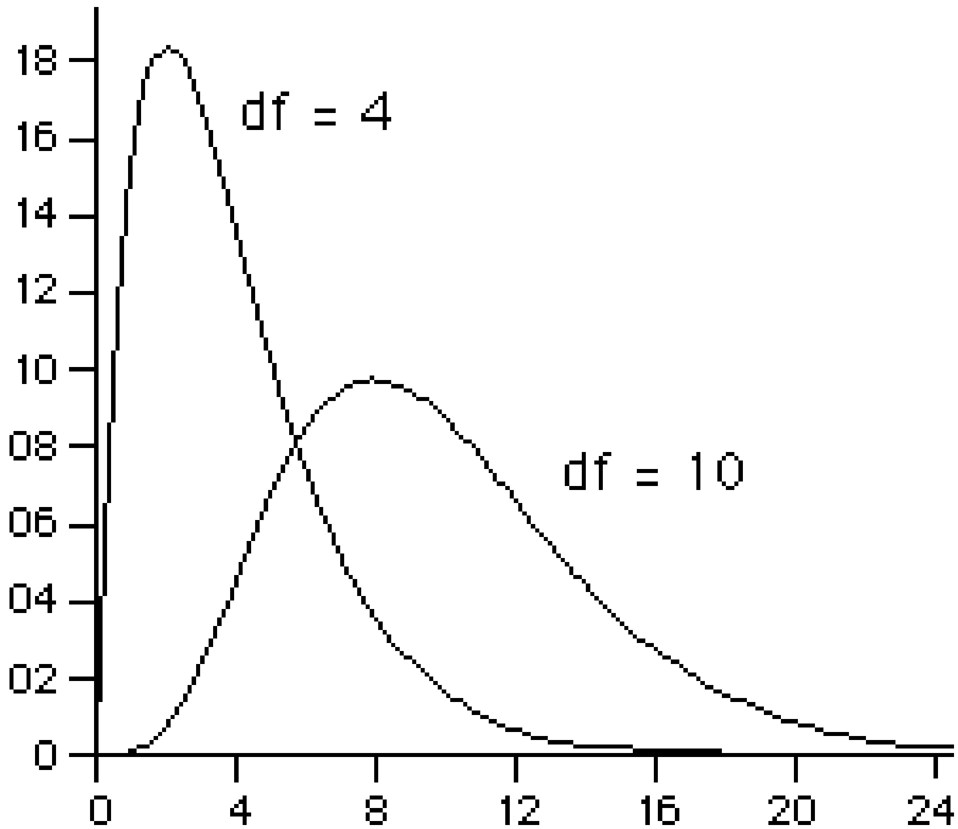
Author: Tibor Nyári PhD
University of Szeged
Department of Medical Physics and Informatics
www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

Chi-square test

Testing for independeny
The r x c contingency tables
square test



The chi-square distribution



χ^2 CRITICAL VALUES

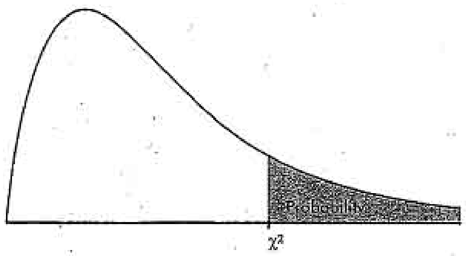


TABLE C: χ^2 CRITICAL VALUES

	Tail probability p										
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4

Example

- A study was carried out to investigate the proportion of persons getting influenza vary according to the type of vaccine. Given below is a 3 x 2 table of observed frequencies showing the number of persons who did or did not get influenza after inoculation with one of three vaccines.
- Does proportion of getting influenza depend on the type of vaccine?

Type of vaccine	Number getting influenza	Number not getting influenza	Total
Seasonal only	**43 (15.35%)**	**237**	280 (100%)
H1N1 only	**52 (20.8%)**	**198**	250 (100%)
Combined	**25 (9.2%)**	**245**	270 (100%)
Totals	120	680	800

Test of independence

- In biology the most common application for chi-squared is in comparing observed counts of particular cases to the expected counts.
- A total of n experiments may have been performed whose results are characterized by the values of two random variable *X* and *Y*.
- We assume that the variables are discrete and the values of *X* and *Y* are x_1, x_2,...,x_r and y_1, y_2,...,y_c, respectively, which are the outcomes of the events A_1,A_2,...,A_r and B_1, B_2,...,B_c . Let's denote by k_{ij} the number of the outcomes of the event (A_i, B_j). These numbers can be grouped into a matrix, called a contingency table. It has the following form:

Contingency table

	B_1	B_2	…	B_c	Total
A_1	k_{11}	k_{12}	…	k_{1c}	k_{1+}
A_2	k_{21}	k_{22}	…	k_{2c}	k_{2+}
…	…	…	…	…	…
A_r	k_{r1}	k_{r2}	…	k_{rc}	k_{r+}
Total	k_{+1}	k_{+2}	…	k_{+c}	n

Frequency of A_i event

i=1,2,…r

$$k_{i+} = \sum_{j=1}^{c} k_{ij}$$

Frequency of B_i event

j=1,2,…c

$$k_{+j} = \sum_{i=1}^{r} k_{ij}$$

Chi-square test (Pearson)

- H_0: The two variables are independent. Mathematically: $P(A_i\, B_j) = P\,(A_i)\, P(\, B_j\,)$
- Test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{k_{i+} \cdot k_{+j}}{n}\right)^2}{\frac{k + \cdot k_{+j}}{n}}$$

- If H_0 is true, then χ^2 has asymptotically χ^2 distribution with (r-1)(c-1) degrees of freedom., namely (number of rows -1)(number of columns -)
- Decision: if $\chi^2 > \chi^2_{table}$ then we reject the null hypothesis that the two variables are independent, in the opposite case we do not reject the null hypothesis.

Observed and expected frequencies

	B_1	B_2	…	B_j	…	B_c	Total
A_1	k_{11}	k_{12}	…	k_{1j}	…	k_{1c}	k_{1+}
A_2	k_{21}	k_{22}	…	k_{2j}	…	k_{2c}	k_{2+}
…	…	…	…	…	…	…	…
A_i	k_{i1}	k_{i2}	…	k_{ij}	…	k_{ic}	k_{i+}
…	…	…	…	…	…	…	…
A_r	k_{r1}	k_{r2}	…	k_{rj}	…	k_{rc}	k_{r+}
Total	k_{+1}	k_{+2}	…	k_{+j}	…	k_{+c}	n

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{k_{i+} \cdot k_{+j}}{n}\right)^2}{\frac{k_{i+} \cdot k_{+j}}{n}} =$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Observed (O_{ij})= k_{ij}
- Expected (E_{ij})=: $\frac{k_{i+} \cdot k_{+j}}{n}$
- Row total*column total/n

Example

- A study was carried out to investigate the proportion of persons getting influenza vary according to the type of vaccine. Given below is a 3 x 2 table of observed frequencies showing the number of persons who did or did not get influenza after inoculation with one of three vaccines.

Type of vaccine	Number getting influenza	Number not getting influenza	Total
Seasonal only	**43**	**237**	280
H1N1 only	**52**	**198**	250
Combined	**25**	**245**	270
Totals	120	680	800

- There are two categorical variables (type of vaccine, getting influenza)
- **H_0: The two variables are independent**
  - **proportions getting influenza are the same for each vaccine**

Calculation of the test statistic

Observed frequencies

Type of vaccine	Number getting influenza	Number not getting influenza	Tot al
Seasonal only	**43**	**237**	280
H1N1 only	**52**	**198**	250
Combined	**25**	**245**	270
Totals	120	680	800

Expected frequencies

Type of vaccine	Number getting influenza	Number not getting influenza	Tot al
Seasonal only	**42**	**238**	280
H1N1 only	**37.5**	**212.5**	250
Combined	**40.5**	**212.5**	270
Totals	120	680	800

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(k_{ij}-\frac{k_{i+}\cdot k_{+j}}{n}\right)^2}{\frac{k_{i+}\cdot k_{+j}}{n}} = \frac{(43-42)^2}{42}+\frac{(237-238)^2}{238}+\frac{(52-37.5)^2}{37.5}+\frac{(198-212.5)^2}{215.5}+\frac{(25-40.5)^2}{40.5}+\frac{(245-229.5)^2}{229.5}$$

$$\chi^2 = 0.024+0.004+5.607+0.989+5.932+1.047 = 13.60$$

- $\chi^2=13.603$
- Degrees of freedom: {(r–1)(c–1)=} (2-1)*(3-1)=2
- Here $\chi^2 = 13.603 > \chi^2_{table} = 5.991$; (df=2;α=0.05). We reject the null hypothesis
- We conclude that the proportions getting influenza are not the same for each type of vaccine

Assumption of the chi-square test

- Expected frequencies should be big enough
- The number of cells with expected frequencies < 5 can be maximum 20% of the cells.
- For example, in case of 6 cells, expected frequencies <5 can be in maximum 1 cell (20% of 6 is 1.2)

SPSS results

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13,603[a]	2	,001
Likelihood Ratio	13,941	2	,001
Linear-by-Linear Association	3,878	1	,049
N of Valid Cases	800		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 37,50.

- $\chi^2=13.036$ and $p=0.001$
- Here $p=0.001 < \alpha=0.05$ we reject the null hypothesis.
- We conclude that the proportions getting influenza are not the same for each type of vaccine

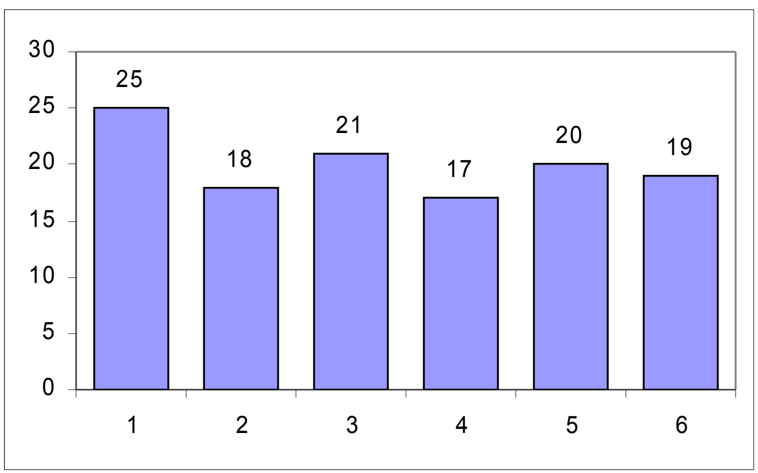
The chi-square test for goodness of fit

- Goodness of fit tests are used to determine whether sample observation fall into categories in the way they "should" according to some ideal model. When they come out as expected, we say that the data fit the model. The chi-square statistic helps us to decide whether the fit of the data to the model is good.
- H0: the distribution of the variable X is a given distribution

The distribution of the sample depending on the type of the variable

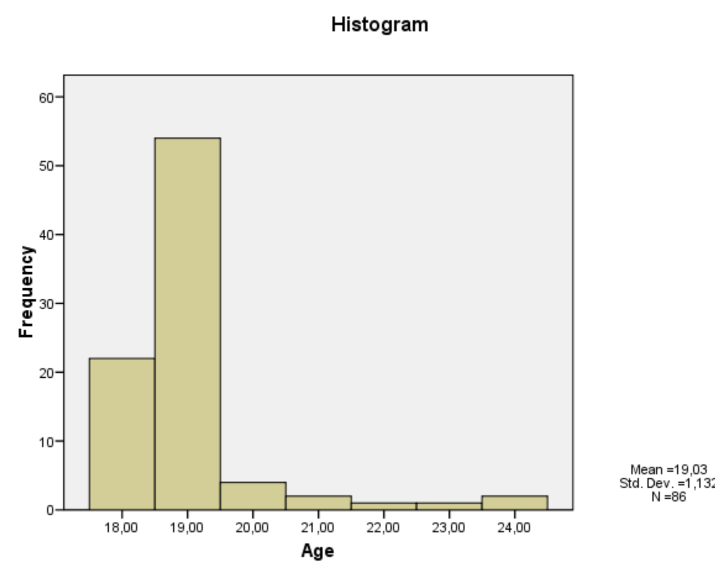
- **Categorical variable.**

Example. A dice is thrown 120 times. We would like to test whether the dice is fair or biased.
Observed frequencies



- **Continuous variable**

Example. We would like to test whether the sample is drawn from a normally distributed population.
Distribution of ages



- Suppose we have a sample of *n* observations. Let's prepare a bar chart or a histogram of the sample – depending on the type of the variable. In both cases, we have frequencies of categories or frequencies in the interval.
- Let's denote the frequency in the *i*-th category or interval by k_i, $i=1,2,\ldots,r$ (*r* is the number of categories).
- Let's denote p_i the probabilities of falling into a given category or interval in the case of the given distribution.
- If H0 is true and *n* is large, then the relative frequencies are approximations of p_i -s, $\frac{k_i}{n} \approx p_i$ or $k_i \approx np_i$.

Observed frequency

Expected frequency

- The formula of the test statistic has $\chi 2$ distribution with (r-1-s) degrees of freedom. Here s is the number of the parameters of the distribution (if there are).
- $$X^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{r} \frac{(k_i - n \cdot p_i)^2}{n \cdot p_i}$$

Test for uniform distribution

- **Example.** We would like to test whether a dice is fair or biased. The dice is thrown 120 times.
- H0: the dice is fair, the probability of each category, p_i=1/6.
- Calculation of expected frequencies: $n \cdot p_i$=120·1/6 = **20.**
- If it is fair, every throwing are equally probable so in ideal case we would expect 20 frequencies for each number.

	1	2	3	4	5	6
Observed frequencies	25	18	21	17	20	19
Expected frequencies	20	20	20	20	20	20

$$X^2 = \sum_{i=1}^{6} \frac{(k_i - 20)^2}{20} =$$

$$= \frac{1}{20}[(25-20)^2 + (18-20)^2 + (21-20)^2 + 17 - 20)^2 + (20-20)^2 + (19-20)^2 =$$

$$= \frac{1}{20}(25+4+1+9+0+1) = 2$$

The degrees of freedom is 5, the critical value in the table is =11.07.
As our test statistic, 2 < 11.07 we do not reject H0 and claim that the dice is fair.

Test for uniform distribution

- **Example 2.** We would like to test whether a dice is fair or biased. The dice is thrown 120 times.
- H0: the dice is fair, the probability of each category, p_i=1/6.
- Calculation of expected frequencies: $n \cdot p_i$=120·1/6 = **20.**
- If it is fair, every throwing are equally probable so in ideal case we would expect 20 frequencies for each number.

	1	2	3	4	5	6
Observed frequencies	**5**	18	21	17	20	**36**
Expected frequencies	20	20	20	20	20	20

$$X^2 = \sum_{i=1}^{6} \frac{(k_i - 20)^2}{20} =$$

$$= \frac{1}{20}[(5-20)^2 + (18-20)^2 + (21-20)^2 + (17-20)^2 + (20-20)^2 + (39-20)^2 =$$

$$= \frac{1}{20}(225 + 4 + 1 + 9 + 0 + 361) = 30$$

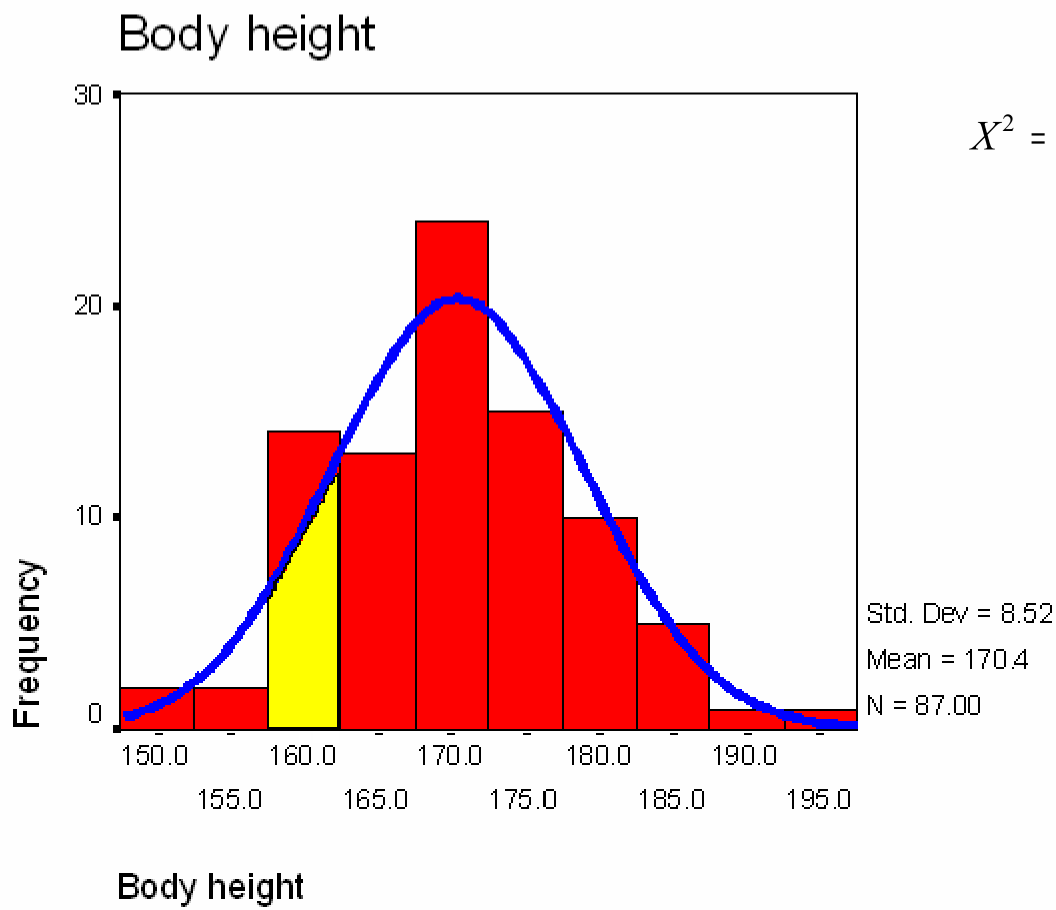
The degrees of freedom is 5, the critical value in the table is =11.07.
As our test statistic, 30 > 11.07 we reject H0 and claim that the dice is not fair.

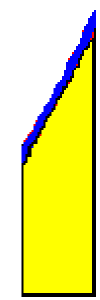
Test for normality

- Let's suppose we have a sample and would like to know whether it comes from a normally distributed population.
- H0: the sample is drawn from a normally distributed population .
- Let's make a histogram from the sample, so we get the "observed" frequencies . To test the null hypothesis we need the expected frequencies.
- **We have to estimate the parameters** of the normal density functions. We use the sample mean and sample standard deviation. The expected frequencies can be computed using the tables of the normal distribution

$$X^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i}$$

k_i

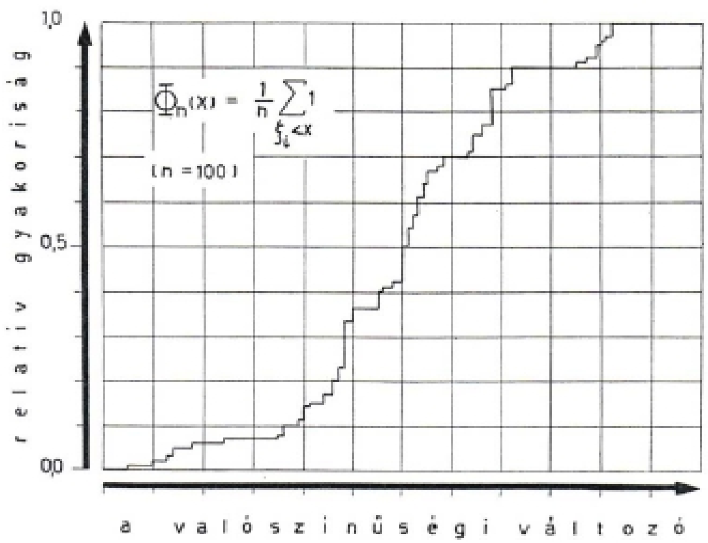




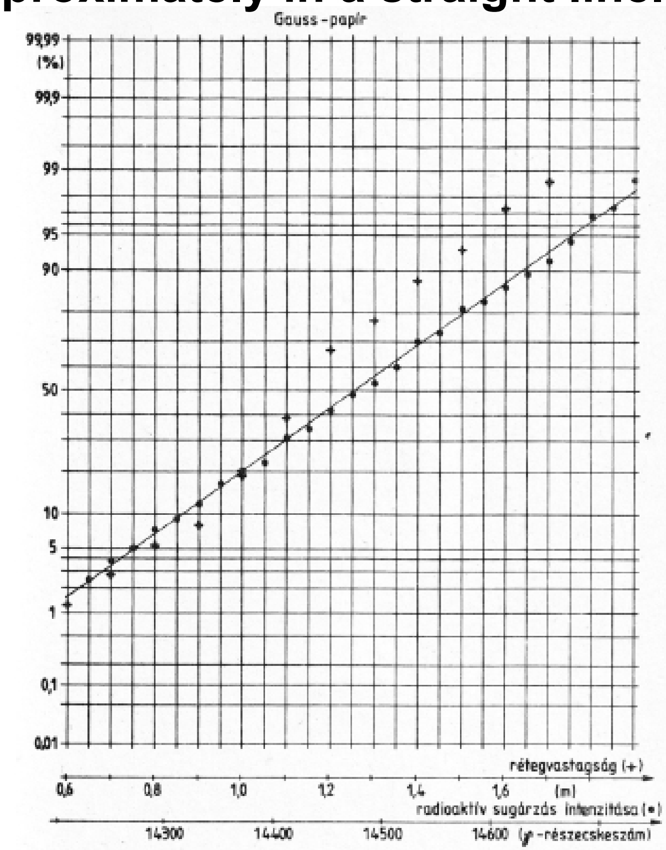
np_i

Using Gauss-paper

There is a graphical method to check normality . The "Gauss-paper" is a special coordinate system, the tick marks of the y axis are the inverse of the normal distribution and are given in percentages. We simply have to draw the distribution function of the sample into this paper. In the case of normality the points are arranged approximately in a straight line.



1. ábra. Egy mért valószínűségi változó (n adat) empirikus eloszlásfüggvénye.



http://www.hidrotanszek.hu/hallgato/Adatfeldolgozas.pdf

SPSS: Q-Q plot (quantile-quantile plot)