Teaching Mathematics and Statistics in Sciences, IPA HU-SRB/0901/221/088 - 2011

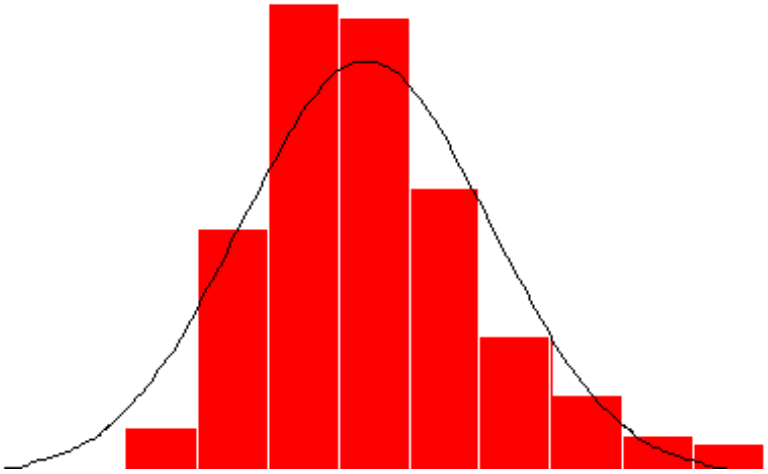
Biostatistics

Author: **Krisztina Boda** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

Nonparametric tests, rank-based tests, chi-square tests



Parametric tests

- **Parameter:** a parameter is a number characterizing an aspect of a population (such as the mean of some variable for the population), or that characterizes a theoretical distribution shape.
- Usually, population parameters cannot be known exactly; in many cases we make assumptions about them.
- Parameters of the normal distribution: μ, σ
- Parameter of the binomial distribution: n, p
- Parameter of the Poisson distribution: λ

Parametric tests

- The null hypothesis contains a parameter of a distribution. The assumptions of the tests are that the samples are drawn from a normally distributed population. So t-tests are parametric tests.
- One sample t-test: H0: μ=c,
- Two sample t-test: H_0: $\mu_1=\mu_2$, assumptions: $\sigma_1^2= \sigma_2^2$

Nonparametric tests

- We do not need to make specific assumptions about the distribution of data.
- They can be used when
  - The distribution is not normal
  - The shape of the distribution is not evident
  - Data are measured on an ordinal scale (low-normal-high, passed – acceptable – good – very good)

Ranking data

- Nonparametric tests can't use the estimations of population parameters. They use ranks instead. Instead of the original sample data we have to use its rank.
- To show the ranking procedure suppose we have the following sample of measurements:
  199, 126, 81, 68, 112, 112.
- Sort the data in ascending order: 68, 81,112,112,126,199
- Give ranks from 1 to *n*: 1, 2, 3, 4, 5, 6
- Two cases are equal, they are assigned a rank of 3.5, the average rank of 3 and 4. We say that case 5 and 6 are tied.
- Ranks corrected for ties: 1, 2, 3.5, 3.5, 5, 6

Result of ranking data

Case	Data	Rank	**Ranks corrected for ties**
4	68	1	**1**
3	81	2	**2**
5	112	3	**3.5**
6	112	4	**3.5**
2	126	5	**5**
1	199	6	**6**

- The sum of all ranks must be $\sum_{i=1}^{n} r_i = \frac{n(n+1)}{2}$
- Using this formula we can check our computations.
- Now the sum of ranks is 21, and 6(7)/2=21.

Nonparametric tests for paired data
(nonparametric alternatives of paired t-test)

- Sign test
- Wilcoxon's matched pairs test
- Null hypothesis: the paired samples are drawn from the same population

Nonparametric test for data in two independent groups
(nonparametric alternatives of two sample t-test)

- Mann-Whitney U test
- Null hypothesis: the samples are drawn from the same population
- Assumption: the distribution of variables is continuous and the density functions have the same shape

Patient	Change in body weight (kg)	Group	Rank	Rank corrected for ties
1.	-1	Diet	3	3
2.	5	Diet	16	16.5
3.	3	Diet	12	13
4.	10	Diet	21	21
5.	6	Diet	18	19
6.	4	Diet	15	15
7.	0	Diet	4	5.5
8.	1	Diet	8	9
9.	6	Diet	19	19
10.	6	Diet	20	19
Sum of ranks, R_1				140
11.	2	Control	11	11
12.	0	Control	5	5.5
13.	1	Control	9	9
14.	0	Control	6	5.5
15.	3	Control	13	13
16.	1	Control	10	9
17.	5	Control	17	16.5
18.	0	Control	7	5.5
19	-2	Control	1	1.5
20.	-2	Control	2	1.5
21.	3	Control	14	13
Sum of ranks R_2				91

SPSS output

Ranks

	group	N	Mean Rank	Sum of Ranks
VAR00001	1.00	10	14.00	140.00
	2.00	11	8.27	91.00
	Total	21		

Test Statistics[b]

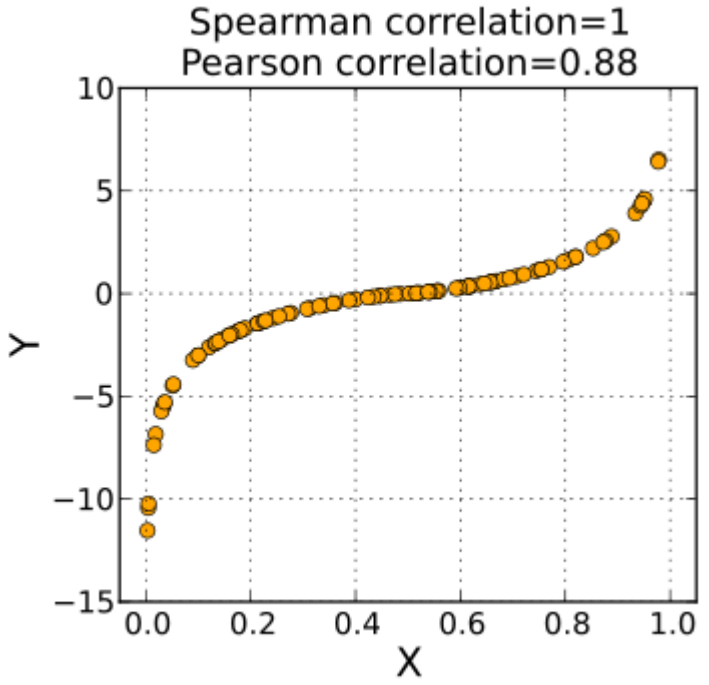
	VAR00001
Mann-Whitney U	25.000
Wilcoxon W	91.000
Z	-2.129
Asymp. Sig. (2-tailed)	.033
Exact Sig. [2*(1-tailed Sig.)]	.036[a]

a. Not corrected for ties.

b. Grouping Variable: group

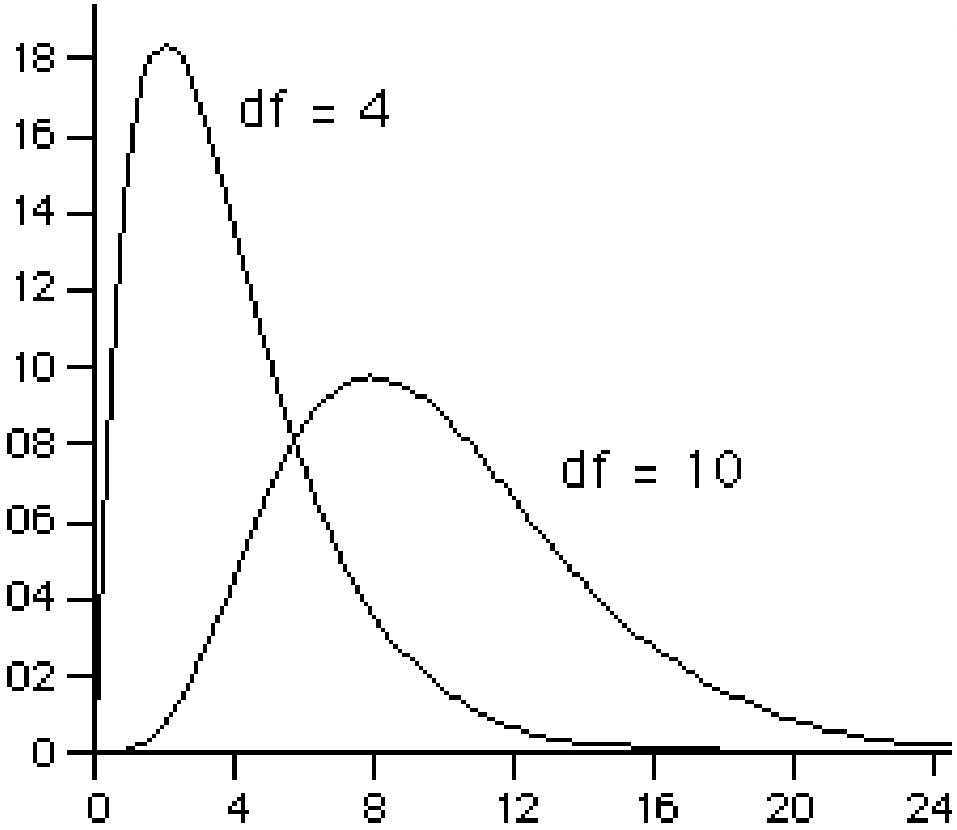
Nonparametric alternative of the correlation coefficient : Spearman's rank correlation coefficient.

- The rank correlation coefficient r_s one of the nonparametric measure of statistical dependence. It is the Pearson's correlation coefficient based on the ranks of the data if there are no ties (adjustments are made if some of the data are tied).
- $-1 \leq r_s \leq +1$
- Its significance can be tested using the same formula as in testing the Pearson's coefficient of correlation.



Nonparametric tests but they are not based on ranks

The chi-square distribution



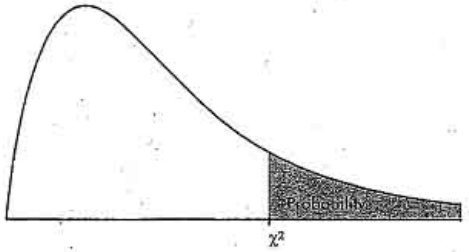


TABLE C: χ^2 CRITICAL VALUES

	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 $y_1, y_2,...,y_s$, respectively, which are the outcomes of the events $A_1,A_2,...,A_r$ and $B_1, B_2,...,B_s$. 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_s	Total
A_1	k_{11}	k_{12}	...	k_{1s}	k_{1+}
A_2	k_{21}	k_{22}	...	k_{2s}	k_{2+}
...	...	...	...	...	...
A_r	k_{r1}	k_{r2}	...	k_{rs}	k_{r+}
Total	k_{+1}	k_{+2}	...	k_{+s}	n

Frequency of A_i event

i=1,2,…r

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

Frequency of B_i event

j=1,2,…s

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

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 $(r-1)(s-1)$ degrees of freedom.
- 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_s	Total
A_1	k_{11}	k_{12}	...	k_{1j}	...	k_{1s}	k_{1+}
A_2	k_{21}	k_{22}	...	k_{2j}	...	k_{2s}	k_{2+}
...	...	...	...	...	...	...	...
A_i	k_{i1}	k_{i2}	...	k_{ij}	...	k_{is}	k_{i+}
...	...	...	...	...	...	...	...
A_r	k_{r1}	k_{r2}	...	k_{rj}	...	k_{rs}	k_{r+}
Total	k_{+1}	k_{+2}	...	k_{+j}	...	k_{+s}	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})=:
- Row total*column total/n

$$\frac{k_{i+}\cdot k_{+j}}{n}$$

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	Total
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	Total
Seasonal only	**42**	**238**	280
H1N1 only	**37.5**	**212.5**	250
Combined	**40.5**	**229.5**	270
Totals	120	680	800

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

$$\chi^2 = 0.024 + 0.004 + 5.607 + 0.975 + 5.932 + 1.047 = 13.5902$$

- χ^2=13.6
- Degrees of freedom: {(r–1)(c–1)=} (2-1)*(3-1)=2
- Here χ^2 =13.6> χ^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.

- a. Assumption for expected frequencies are OK
- χ^2=13.06 and p=0.001
- Here p=0.001 < α=0.05 we reject the null hypothesis.
- We conclude that the proportions getting influenza are not the same for each type of vaccine

Testing for independence

Chi-squre test: 2 by 2 tables

	Risk factor		Total
	Yes	No	
Group 1	k_{11}	k_{12}	k_{1+}
Group 2	k_{21}	k_{22}	k_{2+}
Total	k_{+1}	k_{+2}	n

Chi-square test for 2x2 tables

- Formula of the test statistic

$$\chi_p^{\,2} = \frac{n\left(k_{11}k_{22} - k_{12}k_{21}\right)^2}{k_{+1}k_{+2}k_{1+}k_{2+}}, \mathrm{df} = 1;$$

- Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-square test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 × 2 contingency table. This reduces the chi-square value obtained and thus increases its *p*-value.

$$\chi^2 = \frac{n\left(\left|k_{11}k_{22} - k_{12}k_{21}\right| - n/2\right)^2}{k_{+1}k_{+2}k_{1+}k_{2+}}, \mathrm{df} = 1; \mathrm{Yates}$$

Example

- We are going to compare the proportions of two different treatments' output. Our data are tabulated below.
- H0: the outcome is independent of treatment in the population.

	outcome		
Treatment	Death	Survival	Total
A	5	45	50
B	8	42	50
Total	13	87	100

$$\chi_p^2 = \frac{n(k_{11}k_{22} - k_{12}k_{21})^2}{k_{+1}k_{+2}k_{1+}k_{2+}} = \frac{100(5*42 - 8*45)^2}{50*50*13*87} = 0.79, \text{ df} = 1;$$

Solution based on obseved and expected frequencies

Observed

Treatment	outcome: Death	Survival	Total
A	5	45	50
B	8	42	50
Total	13	87	100

Expected

$$\frac{50 \cdot 13}{100} = \frac{13}{2}$$

Treatment	outcom: Death	Survival	Total
A	6.5	43.5	50
B	6.5	43.5	50
Total	13	87	100

$$\chi_p^2 = \frac{(5-6.5)^2}{6.5} + \frac{(8-6.5)^2}{6.5} + \frac{(45-43.5)^2}{43.5} + \frac{(42-43.5)^2}{43.5} =$$

$$= \frac{2.25}{6.5} + \frac{2.25}{6.5} + \frac{2.25}{43.5} + \frac{2.25}{43.5} = 0.79, \quad \text{df} = 1;$$

Decision

- Here Pearson χ^2 =0.796 <χ^2_{table} =3.841 thus we accept the null hypothesis that the two variables are independent
- SPSS p-value (=0.372) is greater than α=0.05 so thus we accept also the null hypothesis that the two variables are independent

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,796[b]	1	,372		
Continuity Correction[a]	,354	1	,552		
Likelihood Ratio	,802	1	,370		
Fisher's Exact Test				,554	,277
Linear-by-Linear Association	,788	1	,375		
N of Valid Cases	100				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,50.

Notes

- Both variables are dichotomous
- The Chi-squares give only an estimate of the true Chi-square and associated probability value, an estimate which might not be very good in the case of the marginals being very uneven or with a small value (~less than five) in one of the cells
- In that case the Fisher Exact is a good alternative for the Chi-square. However, with a large number of cases the Chi-square is preferred as the Fisher is difficult to calculate.

Fisher’s-exact test

Calculation of the p-value is based on the permutational distribution of the test Statistic (without using chi-square formula).

Display of data

	Disease status		
	Disease	No	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	n

Fisher-exact test

- The procedure, ascribed to Sir Ronald Fisher, works by first using probability theory to calculate the probability of observed table, given fixed marginal totals.
  - Note: *n* factorial: n!=1·2·3·…·n, 0!=1

$$\frac{(a+c)!(b+d)!(a+b)!(c+d)!}{n!a!b!c!d!}$$

Example

	Disease status		
	Yes	No	Total
Exposed	2	3	5
Non-exposed	4	0	4
Total	6	3	9

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3,600[b]	1	,058		
Continuity Correction[a]	1,406	1	,236		
Likelihood Ratio	4,727	1	,030		
Fisher's Exact Test				,167	,119
Linear-by-Linear Association	3,200	1	,074		
N of Valid Cases	9				

a. Computed only for a 2x2 table

b. 4 cells (100,0%) have expected count less than 5. The minimum expected count is 1,33.

Observed probabilities

Original table

	Disease	Status
	Yes	No
Exposed	2	3
Non-exposed	4	0

$$p_{obs} = \frac{5!4!6!3!}{9!2!3!4!0!} = \frac{12441600}{104509440} = 0{,}1190$$

Possible re-arrangements

	Disease	Status
	Yes	No
Exposed	3	2
Non-exposed	3	1
		p=0,4762
Exposed	4	1
Non-exposed	2	2
		p=0,3571
Exposed	5	0
Non-exposed	1	3
		P=0,0476

Fisher's p-value=0,119+0,0476=0,167

Fisher showed that to generate a significance level, we need consider only the cases where the marginal totals are the same as in the observed table, and among those, only the cases where the arrangement is as extreme as the observed arrangement, or more so.

The chi-square test for goodness of fit

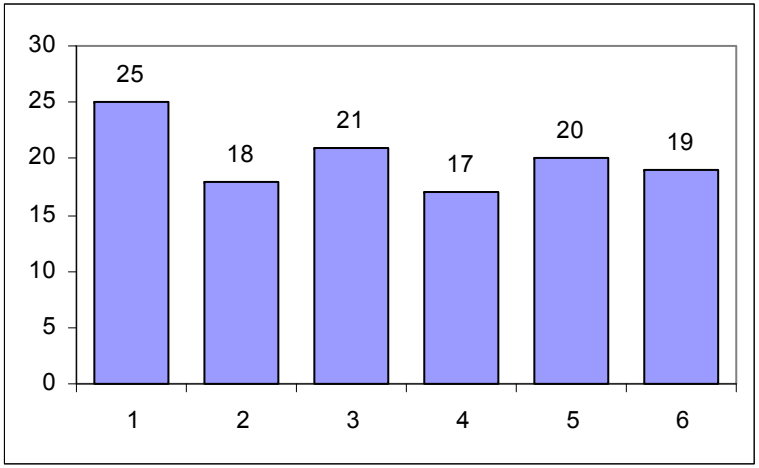
The chi-square test for goodness of fit

- 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

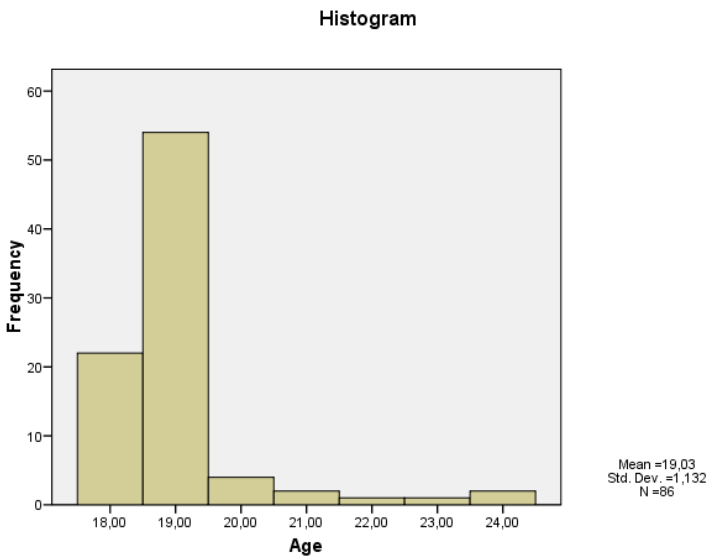
- **Categorical variable**.

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

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

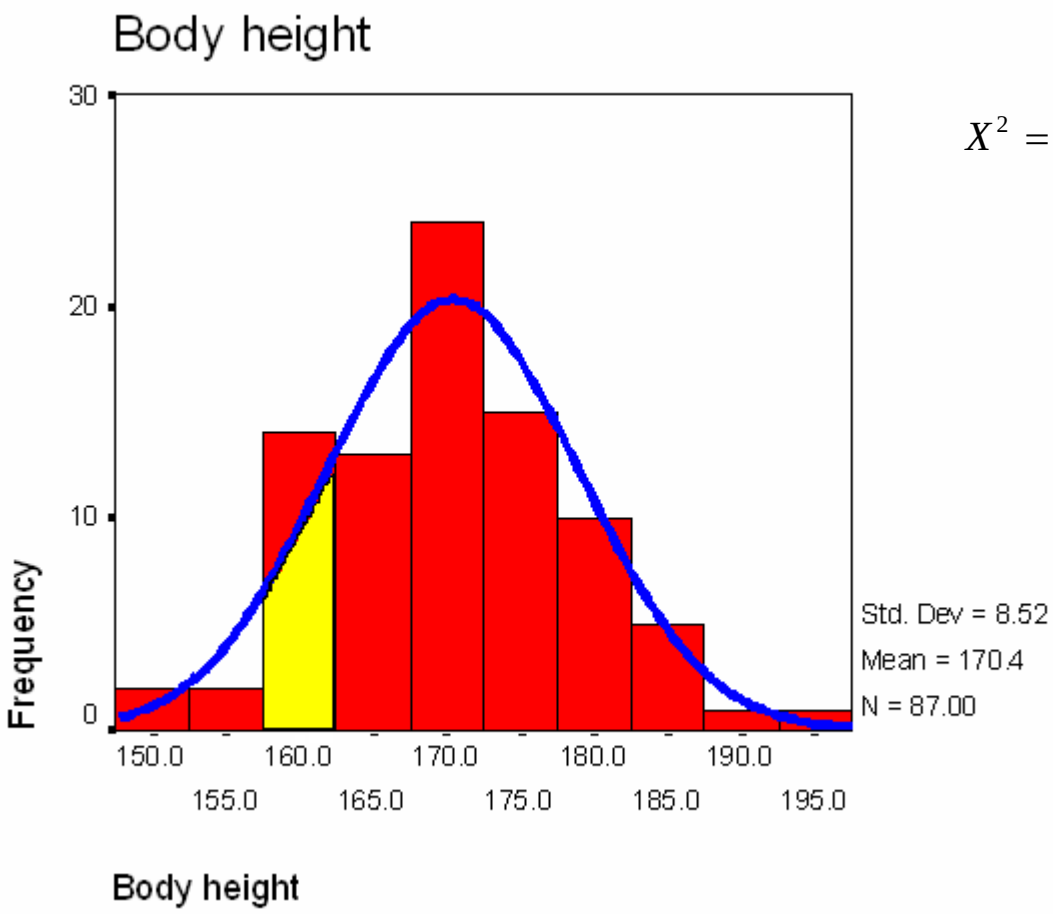
$$= \frac{1}{20}[(5 - 20)^2 + (18 - 20)^2 + (21 - 20)^2 + (17 - 20)^2 + (20 - 20)^2 + (39 - 20)^2 =$$

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

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



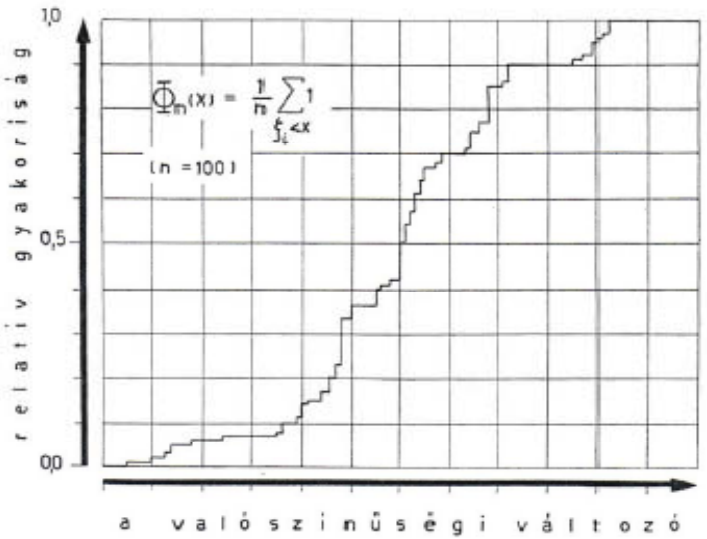
$$X^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i}$$

k_i

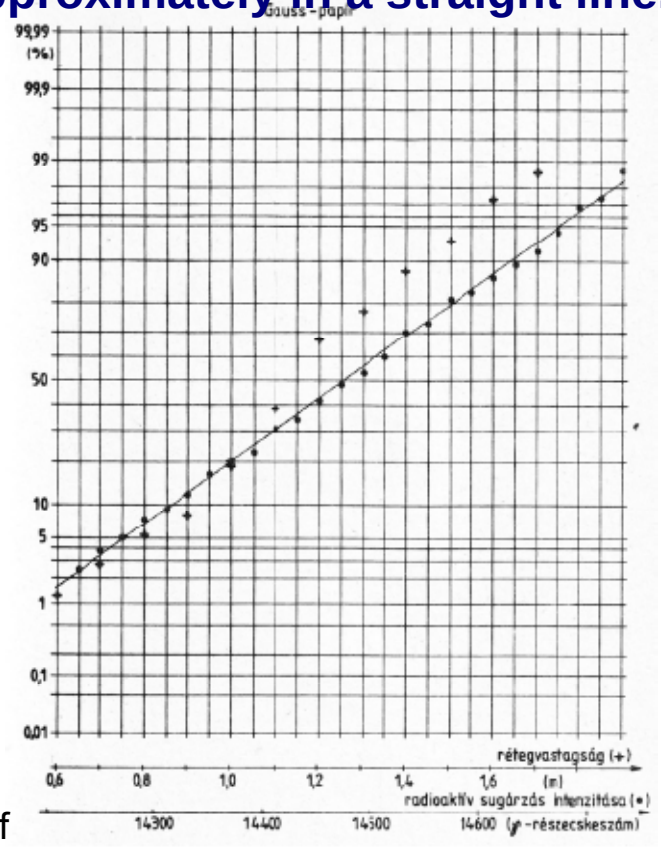
np_i

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.

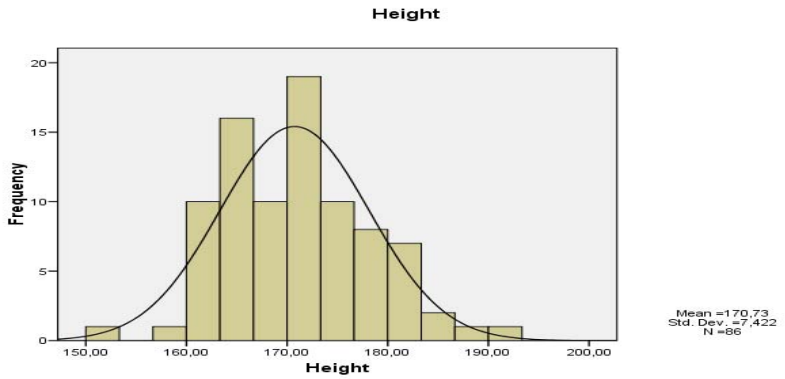


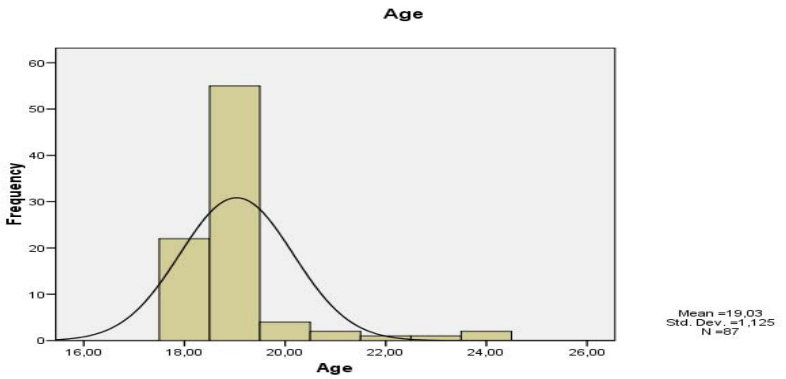
1. ábra. Egy mért valószínűségi változó (n adat) empirikus eloszlásfüggvénye.

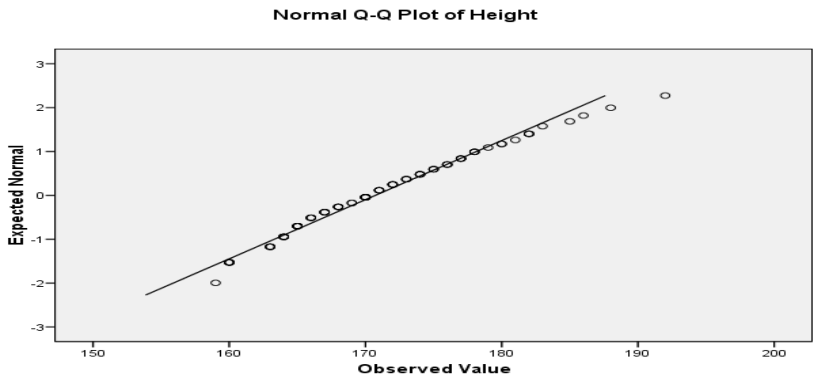


http://www.hidrotanszek.hu/hallgato/Adatfeldolgozas.pdf

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







Comparing a single proportion

- In a country hospital were 515 Cesarean section (CS) in 2146 live birth in 2001. Compare this proportion to the national proportion 22%. Does proportion of CS in this hospital differ from the national one?

- H_0: p=22%
- H_A: p≠22%

$$z = \frac{\hat{p}_1 - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{(515/2146) - 0.22}{\sqrt{\frac{0.22 \cdot 0.78}{2146}}} = \frac{0.24 - 0.22}{0.0089} = 2.234$$

Review questions and problems

- When to use nonparametric tests
- Ranking data
- The aim and null hypothesis of the test for independence
- Contingency table
- Observed and expected frequencies
- The assumption of the chi-square test
- Calculation of the degrees of freedom of chi-square test
- Calculation of the test statistic of chi-square test and decision based on table
- Evaluation possibilities of a 2x2 contingency table
- Fisher's exact test

Problems

1. The following table shows the results of placebo and aspirin in an experiment, with the number of people in each treatment group who did and did not develop thromboses. Decide whether the aspirin had or had not effect on thrombus formation.

	Developed thrombi	Free of thrombi
Placebo	10	5
Aspirin	10	20

Find the value of the test statistic, and give your conclusion. (alfa=0.05, $\chi 2_{table}$=3.84)
This conclusion was based on the value .....because.....

2. Two medicines are being compared regarding a particular side effect, 60 similar patients are split randomly into two groups, one on each drug. The results are presented in the following table:

	Side effects	no side effects
Drug A	10	20
Drug B	5	25

Are drug type and side effects independent?
Find the value of the test statistic, and give your conclusion. (alfa=0.05, $\chi 2_{table}$=3.84)
This conclusion was based on the value .....because.....

Problems

- Boys and girls were asked whether they find biostatistics necessary or not. The answers were evaluated by a chi-square test. Interpret the SPSS result

Sex * Is biostatistics necessary? Crosstabulation

			Is biostatistics necessary?		
			yes	no	Total
Sex	Male	Count	58	11	69
		% within Sex	84.1%	15.9%	100.0%
	Female	Count	34	11	45
		% within Sex	75.6%	24.4%	100.0%
Total		Count	92	22	114
		% within Sex	80.7%	19.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.264[b]	1	.261		
Continuity Correction[a]	.777	1	.378		
Likelihood Ratio	1.243	1	.265		
Fisher's Exact Test				.333	.188
Linear-by-Linear Association	1.253	1	.263		
N of Valid Cases	114				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.68.