Lecture 10

Data transformation

χ^2 distribution

Normalization

For normal distributions, a large number of tests were developed which allowed for drawing conclusions about the truth of various hypotheses.

What to do when assumptions about normality are not met?

You can apply a nonparametric test or perform a data transformation.

Data transformation

Transforming non-normal data to normal distribution.

Frequently used transformations (transformations) of a variable *x*:

Arc sin X

Box-Cox transformation $\frac{x^{\lambda}-1}{\lambda}$

logarithm, exponentiation, and square root $\sqrt{x}$

itp.

Arc sin (the so-called Bliss transformation) - we usually use for data with a binomial distribution expressed as a percentage, most often taking values in the range (0-20% or 80-100%)

The Box-Cox transformation - this is a frequently used transformation in the case of asymmetric distributions (left or right oblique or "truncated" normal distributions)

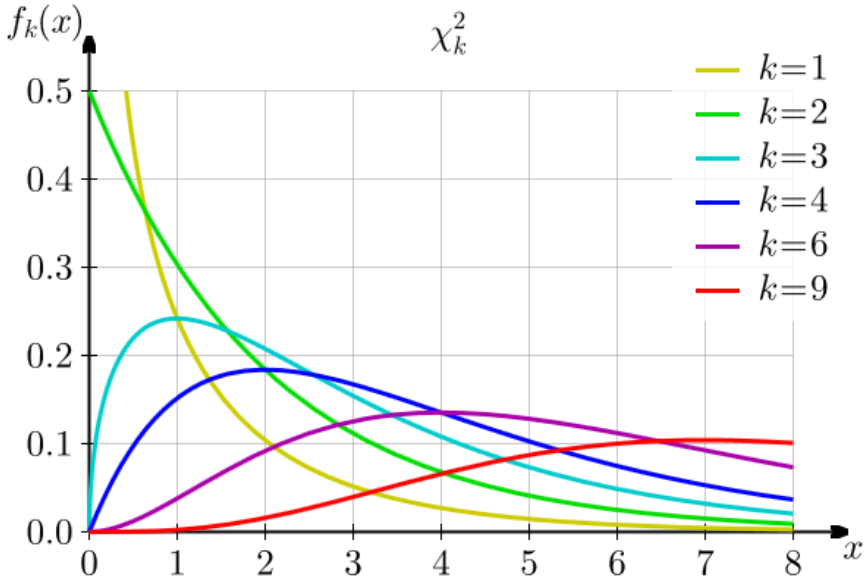
Logarithm - Usually used when the variance (and thus the standard deviation) increases with the increase of the mean value, i.e. there is a correlation between the mean and the variance. It may not be possible to apply a log (x) transformation, e.g. in this case if the variable takes the value 0, then a log (x + 1) transformation can be applied.

Rooting - we use for distributions close to the Poisson distribution, i.e. in right-skewed distributions where the mean value is close to the variance. As with the log (x) transformation, there can be a problem if the variable is set to 0 (or negative values).

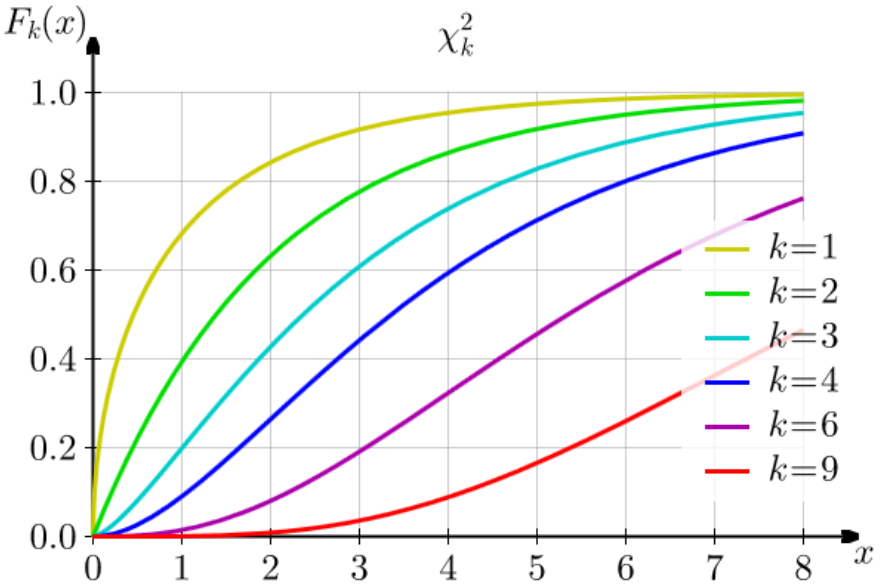
Data transformation - problems

1) The inability to transform some distributions to a normal distribution, e.g. you cannot transform a step variable to a continuous variable, so if the variable is a discrete (discrete) variable that takes a small number of values (e.g. 1, 2, 3, 4 and 5) it is impossible to apply the transformation so that the distribution of this variable is normal
2) Difficulty interpreting the results. Due to the fact that after the transformation the parameter values (eg mean value) change, it is impossible to infer, for example, the percentage difference between the means on the basis of parameters calculated on the transformed variable.

χ^2

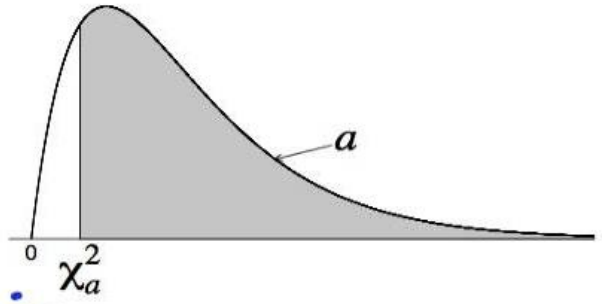


The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.



http://en.wikipedia.org/wiki/Chi-squared_distribution

χ^2



df	$\chi^2_{0.9995}$	$\chi^2_{0.999}$	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.85}$	$\chi^2_{0.80}$
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.036	0.064
2	0.001	0.002	0.010	0.020	0.051	0.103	0.211	0.325	0.446
3	0.015	0.024	0.072	0.115	0.216	0.352	0.584	0.798	1.005
4	0.064	0.091	0.207	0.297	0.484	0.711	1.064	1.366	1.649
5	0.158	0.210	0.412	0.554	0.831	1.145	1.610	1.994	2.343
6	0.299	0.381	0.676	0.872	1.237	1.635	2.204	2.661	3.070
7	0.485	0.598	0.989	1.239	1.690	2.167	2.833	3.358	3.822
8	0.710	0.857	1.344	1.646	2.180	2.733	3.490	4.078	4.594
9	0.972	1.152	1.735	2.088	2.700	3.325	4.168	4.817	5.380
10	1.265	1.479	2.156	2.558	3.247	3.940	4.865	5.570	6.179

https://www.youtube.com/watch?v=HwD7ekD5l0g

Pearson's chi-squared test - χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Pearson's chi-squared test χ^2 - example

The petrograph examines microscopically a thin sample of igneous rock. Its task is to give the tested scale the correct name. Using a special apparatus coupled with a microscope, it counts 100 crystals present in the sample. It is known from the literature that in granite the ratio of the 4 main minerals is 4: 1: 2: 3. Can the tested sample, in which the ratio of the respective crystals is 35: 12: 22: 31, be called granite?

Pearson's chi-squared test χ^2 - example

H_0 – the distributions are consistent

	100		
	E	O	$(O-E)^2/E$
4	40	35	0.625
1	10	12	0.4
2	20	22	0.2
3	30	31	0.033333
			$\sum$
			1.258333

Test statistic χ^2 is 1.258

p=ROZKŁAD.CHI(1.258;3)=0,739

$p>\alpha$

H_0 we assume, we can say that the distributions are consistent

Pearson's chi-squared test χ^2 - example

https://www.statskingdom.com/310GoodnessChi.html

Pearson's chi-squared test χ^2 - example

The structure of the forest was studied at five sites. The number of species in all three forest floors was determined. Check whether the species richness of these forests depends on the floor.

	position				
	A	B	C	D	E
Layer I	24	33	52	7	64
Layer II	22	11	6	16	23
Layer III	5	7	6	11	4

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

Pearson's chi-squared test χ^2 - example

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

H_0 the variables are independent or the species richness of these forests does not depend on the layer

The chi-square statistic is 52.0032. The p-value is < 0.00001. The result is significant at $p < .05$.

we reject H_0

the variables are dependent or the species richness of these forests depends on the layer