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

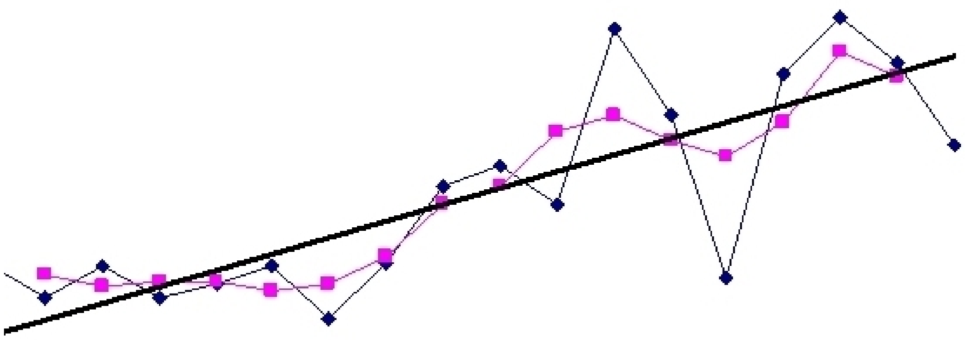
Mathematical and Statistical Modelling in Medicine

Author: **Tibor Nyári** PhD

University of Szeged
Department of Medical Physics and Informatics

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

Testing for proportions



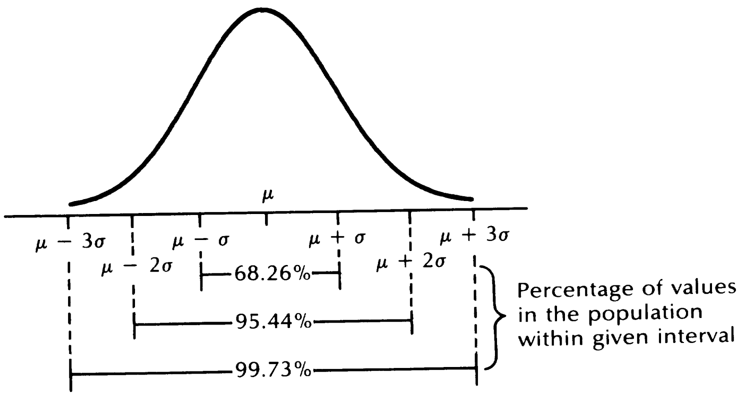
Testing forsingle proportion

The square contingency table

Pearson -, Yates – and Fisher exact tests

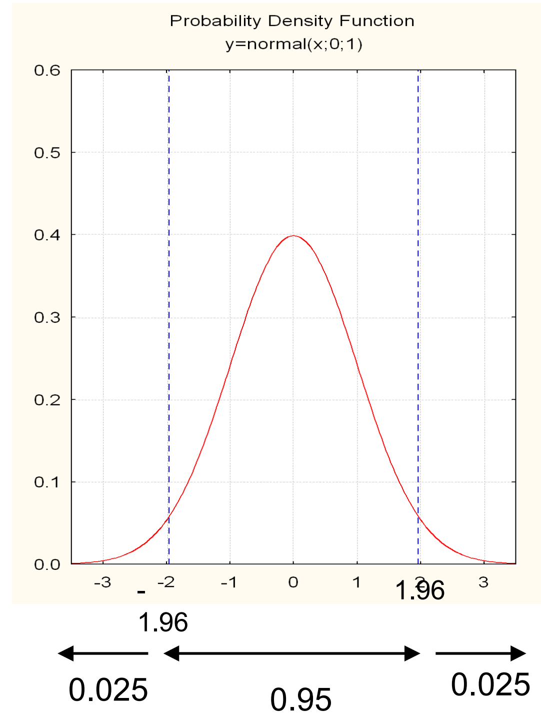
Special distribution of a continuous variable: the normal distribution N(μ,σ)

- The density curves are symmetric, single-peaked and bell-shaped
- The 68-95-99.7 rule .
  In the normal distribution with mean μ and standard deviation σ:
  - 68% of the observations fall within σ of the mean μ
  - 95% of the observations fall within 2σ of the mean μ
  - 99.7% of the observations fall within 3σ of the mean μ

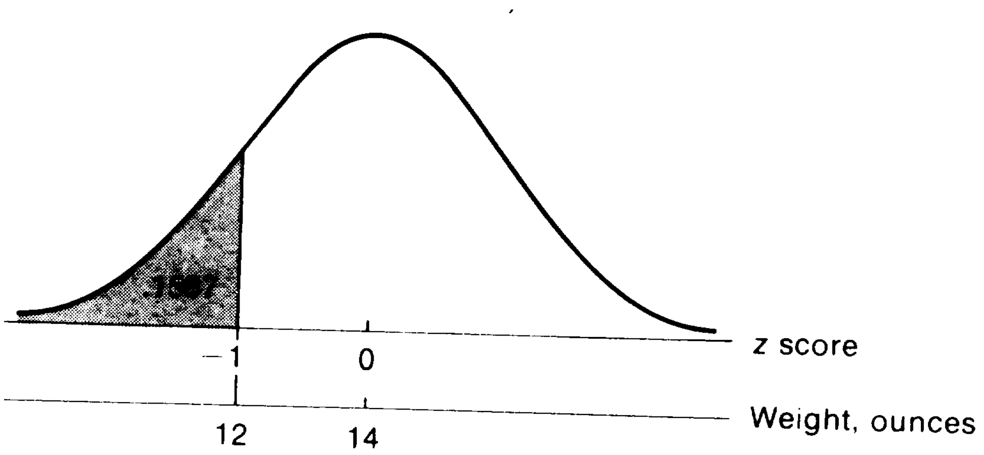


Standard normal probabilities

x	(x): proportion of area to the left of x
-4	0.0003
-3	0.0013
-2.58	0.0049
-2.33	0.0099
-2	0.0228
-1.96	0.0250
-1.65	0.0495
-1	0.1587
0	0.5
1	0.8413
1.65	0.9505
1.96	0.975
2	0.9772
2.33	0.9901
2.58	0.9951
3	0.9987
4	0.99997



The use of the normal curve table



Example

- Find the area under a standard normal curve between $x=-1.96$ and $x=1.96$.
- Solution.
  - $P(-1.96<X<1.96)=P(X<1.96)-P(X<-1.96)=0.95$
  - $\Phi(-1.65)=0.025$, $\Phi(1.96)=0.975$. We find the area by subtracting. Thus, the area between is $0.975-0.025=0.95$

PROPORTIONS

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%

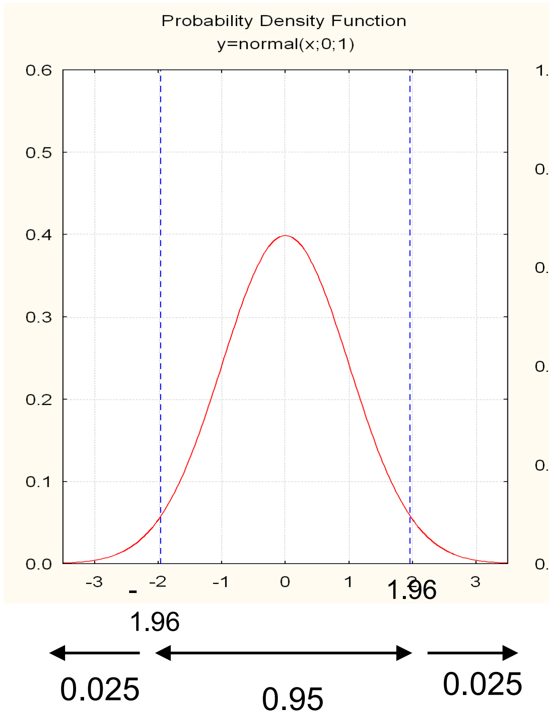
$$z = \frac{\hat{p}_1 - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{(515/2146) - 0.22}{\sqrt{\frac{0.22 \cdot 0.78}{2146}}} = \frac{0.24 - 0.22}{0.0089} = 2.234$$

Decision

- As the calculaterd |z| score was greater than 1.96, thus Nullhypothesis is rejected , and the alternative hypothesis is accepted, namely the diffence is significant

Standard normal probabilities

x	(x): proportion of area to the left of x
-4	0.0003
-3	0.0013
-2.58	0.0049
-2.33	0.0099
-2	0.0228
-1.96	0.0250
-1.65	0.0495
-1	0.1587
0	0.5
1	0.8413
1.65	0.9505
1.96	0.975
2	0.9772
2.33	0.9901
2.58	0.9951
3	0.9987
4	0.99997



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 as below.
- H_0: the outcome is independent of treatment in the population.

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

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

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 :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.