RELATIONS AMONG QUALITATIVE DATA

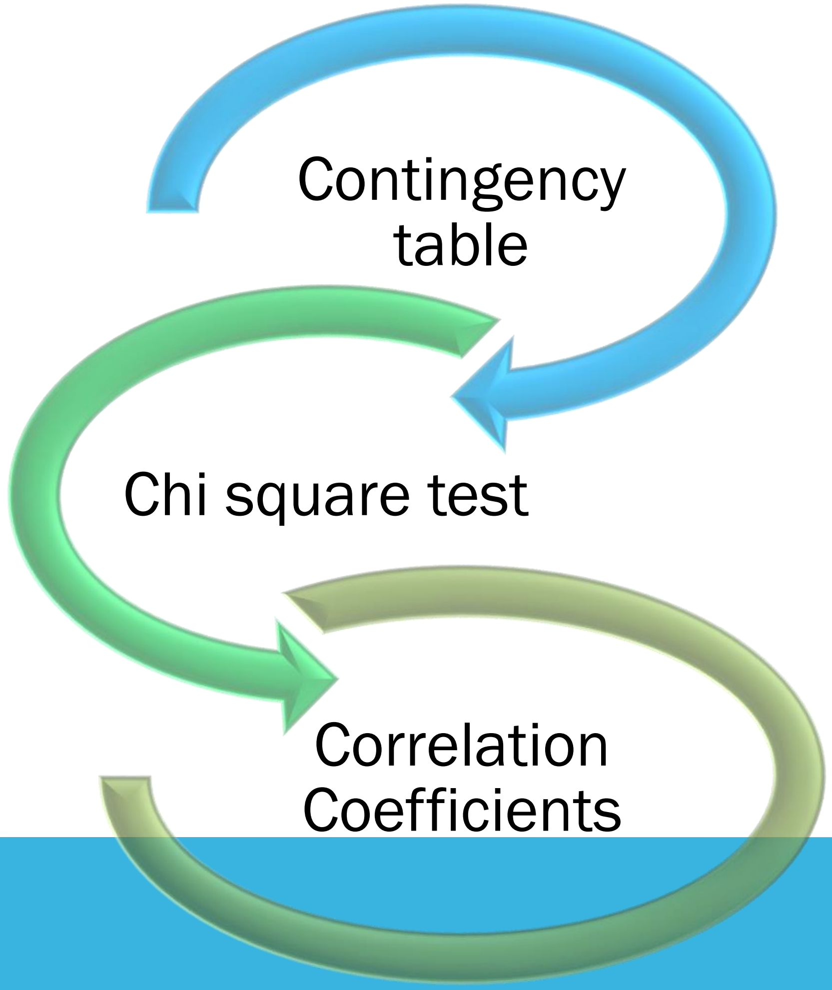
Karolina Tura-Gawron, PhD

AGENDA

I. Contingency tables
II. Chi square test
III. Correlation coefficients
IV. Statistica

ANALYSIS OF THE RELATIONS AMONG QUALITATIVE DATA

Steps



CONTINGENCY TABLE

A rectangular table such as this, in which items from a population are classified accrording to the two characteristics

$X \backslash Y$	Y_1	Y_2	...	...	Y_p	
X_1	n_{11}	n_{12}	...	...	n_{1p}	$\sum_{j=1}^{p} n_{1j}$
X_2	n_{21}	n_{22}	...	...	n_{2p}	$\sum_{j=1}^{p} n_{2j}$
...	...	...	...	...	...	
...	...	...	...	...	...	
X_k	n_{k1}	n_{k2}	...	...	n_{kp}	$\sum_{j=1}^{p} n_{kj}$
	$\sum_{i=1}^{k} n_{i1}$	$\sum_{i=1}^{k} n_{i2}$			$\sum_{i=1}^{k} n_{ip}$	$\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} = N$

CHI SQUARE TEST FOR INDEPENDENCE (1)

Conditions Required for a Valid Chi square Test: Contingency Table

I. The *n* observed counts are a random sample from the population of interest. We may consider this to be a multinomial experiment with *k x p* possible outcomes

II. The sample size , *n,* will be large enough so that, for every cell, the expected count, E_{ij}, will be equal to 5 or more.

III. N>40

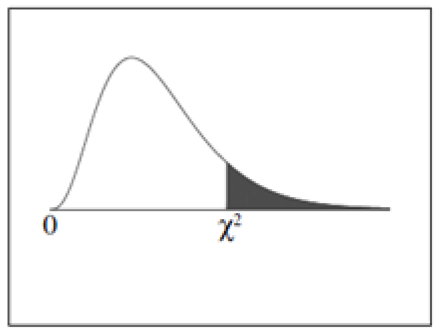
CHI SQUARE TEST FOR INDEPENDENCE (2)

H_0: X and Y are independent
H_1: X a nd Y are dependent

Rejection region: $\chi^2 > \chi^2_\alpha$

Where χ^2_α has *(k-1)(p-1)* df

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

CHI SQUARE TEST FOR INDEPENDENCE (3)

Observed frequencies and expected frequencies comparison

$$E_{ij} = \frac{(row\ i\ total) * (column\ j\ total)}{total\ sum} = \frac{\sum_{j=1}^{p} n_{ij} \sum_{i=1}^{k} n_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}}$$

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

O – observed cell's frequency

E – expected cell's frequency

TASK 1. (1)

A certain Cola company sells four types of cola throughout North America. To help determine if the same marketing approach used in the United States can be used in Canada and Mexico, one of the firm's marketing analysts wishes to ascertain if there is an association between the type of cola preferred: regular (A), caffeine free (B), both caffeine- and sugar free (C), sugar free only (D). The second classification might consist of three nationalities:American (N1), Canadian (N2) and Mexican (N3). The marketing analyst then interviews a arndom sample of 250 cola drinkers from the three countries, classifies each according to the two criteria, and records the observed frequency of drinkers falling into each of the twelve possibles cells. Take α=0.01.

	Cola Preference				
Nationality	A	B	C	D	Total
N1	72	8	12	23	115
N2	26	10	16	33	85
N3	7	10	14	19	50
Total	105	28	42	75	250

TASK. 1.(2) CONTINGENCY TABLE CLASSIFYING COLA DRINKERS

$$E_{11} = \frac{(row\ 1\ total) * (column\ 1\ total)}{total\ sum} = \frac{115 * 105}{250} = 48.3$$

...

$$E_{34} = \frac{50 * 75}{250} = 15$$

	Cola Preference				
Nationality	A	B	C	D	Total
N1	72 (48.3)	8 (12.88)	12 (19.32)	23 (34.5)	115
N2	26 (35.7)	10 (9.52)	16 (14.28)	33 (25.5)	85
N3	7 (21)	10 (5.6)	14 (8.4)	19 (15)	50
Total	105	28	42	75	250

TASK. 1.(3) COMPUTATION OF CHI SQUARE FOR COLA DRINKERS

Cell i	O_i	E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
1	72	48.30	561.69	11.63
2	26	35.70	94.09	2.64
3	7	21.00	196.00	9.33
4	8	12.88	23.81	1.85
5	10	9.52	0.23	0.02
6	10	5.60	19.36	3.46
7	12	19.32	53.58	2.77
8	16	14.28	2.96	0.21
9	14	8.40	31.36	3.73
10	23	34.50	132.25	3.83
11	33	25.50	56.25	2.21
12	19	15.00	16.00	1.07
Total	250	250.00	x	42.75

χ^2

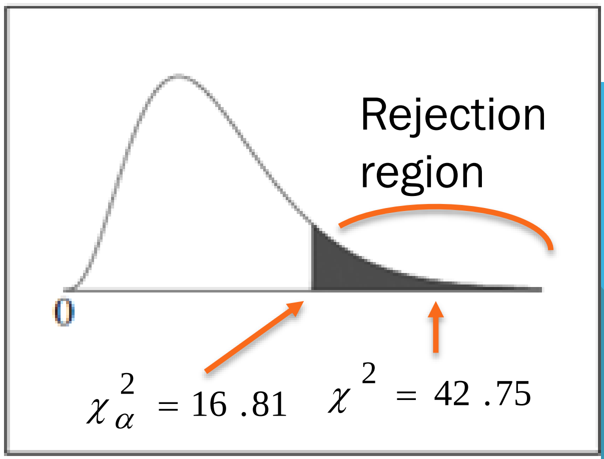
TASK. 1.(4) CHI SQUARE TEST FOR INDEPENDENCE

$$\chi^2 = 42.75$$

$$d.f. = (n.rows - 1)(n.columns - 1) = (3-1)(4-1) = 6$$

$$\chi^2_{\alpha} = \chi^2_{0.01,6} = 16.81$$

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955



We reject the null hypothesis that the two classifications are independent.

Based on the sample data, we conclude that at the 1% level of significance that there is a relationship between the preferences of cola drinkers and their nationality.

CHI SQUARE TEST- CONTINGENCY TABLE 2 X 2

a	b
c	d

$$\chi^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

YATES CORRECTION FOR A 2 X 2 CONTINGENCY TABLE

20<N<40 and at least one of E<5

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

a	b
c	d

$$\chi^2 = \frac{\left(|ad - bc| - \frac{N}{2}\right)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

HOW TO MEASURE THE STRENGHT OF THE RELATION?

I. Yula's coefficient $\phi = \sqrt{\frac{\chi^2}{N}}$

I. V-Cramer's coefficient $V = \sqrt{\frac{\chi^2}{n(\min(k;p)-1)}}$

II. Pearson's contingency coefficient $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

III. Czuprow's coefficient $t = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(p-1)}}}$

Strenght <0,1)
No direction!

TASK 1. (1)

I. Yula's coefficient

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{42.75}{250}} = 0.41$$

II. V-Cramer's coefficient

$$V = \sqrt{\frac{\chi^2}{n(\min(k;p)-1)}} = \sqrt{\frac{42.75}{250(\min(4,3)-1)}} = 0.29$$

TASK 1. (2)

III. Pearson's contingency coefficient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{42.75}{42.75 + 250}} = 0.38$$

IV. Czuprow's coefficient

$$t = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(p-1)}}} = \sqrt{\frac{42.75}{250\sqrt{(4-1)(3-1)}}} = 0.26$$

STATISTICA- TASK 2. (1)

The data are available in the file „Characteristics.sta". Analyse the relation among the eye colour and hair colour.

1. Contingency table
2. Chi square test for independence
3. Strength of the relation

TASK 2. (2)

The data are available in the file „Characteristics.sta". Analyse the relation among the eye colour and hair colour.

1. Contingency table

Summary Frequency Table (Characteristics)
Marked cells have counts > 5
(Marginal summaries are not marked)

Eye Color	Hair Color brown	Hair Color red	Hair Color black	Hair Color blonde	Row Totals
blue	22	11	8	0	41
green	9	7	6	0	22
brown	15	4	13	5	37
All Grps	46	22	27	5	100

Summary Table: Expected Frequencies (Characteristics)
Marked cells have counts > 5
Pearson Chi-square: 14,6631, df=6, p=,023045

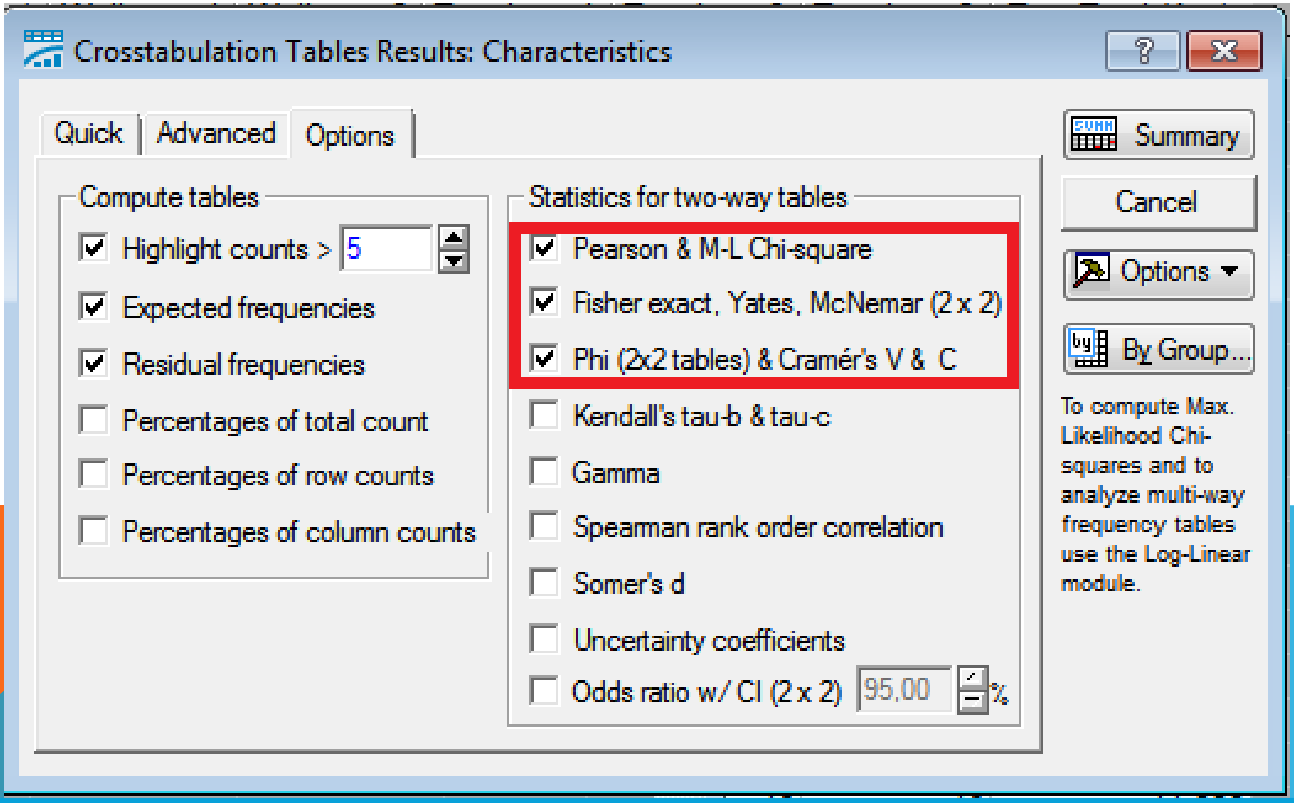
Eye Color	Hair Color brown	Hair Color red	Hair Color black	Hair Color blonde	Row Totals
blue	18,86000	9,02000	11,07000	2,050000	41,0000
green	10,12000	4,84000	5,94000	1,100000	22,0000
brown	17,02000	8,14000	9,99000	1,850000	37,0000
All Grps	46,00000	22,00000	27,00000	5,000000	100,0000

Basic statistics> Tables and banners>Select variables

STATISTICA- TASK 2. (3)

The data are available in the file „Characteristics.sta". Analyse the relation among the eye colour and hair colour.

1. Contingency table
2. Chi square test for independence



Basic statistics> Tables and banners>Select variables>Advanced

STATISTICA- TASK 2. (4)

The data are available in the file „Characteristics.sta". Analyse the relation among the eye colour and hair colour.

1. Contingency table
2. Chi square test for independence
3. Strength of the relation

	Statistics: Eye Color(3) x Hair Color(4) (Characteristics)		
Statistic	Chi-square	df	p
Pearson Chi-square	14,66310	df=6	p=,02305
M-L Chi-square	16,43634	df=6	p=,01159
Phi	,3829243		
Contingency coefficient	,3576030		
Cramér's V	,2707684		

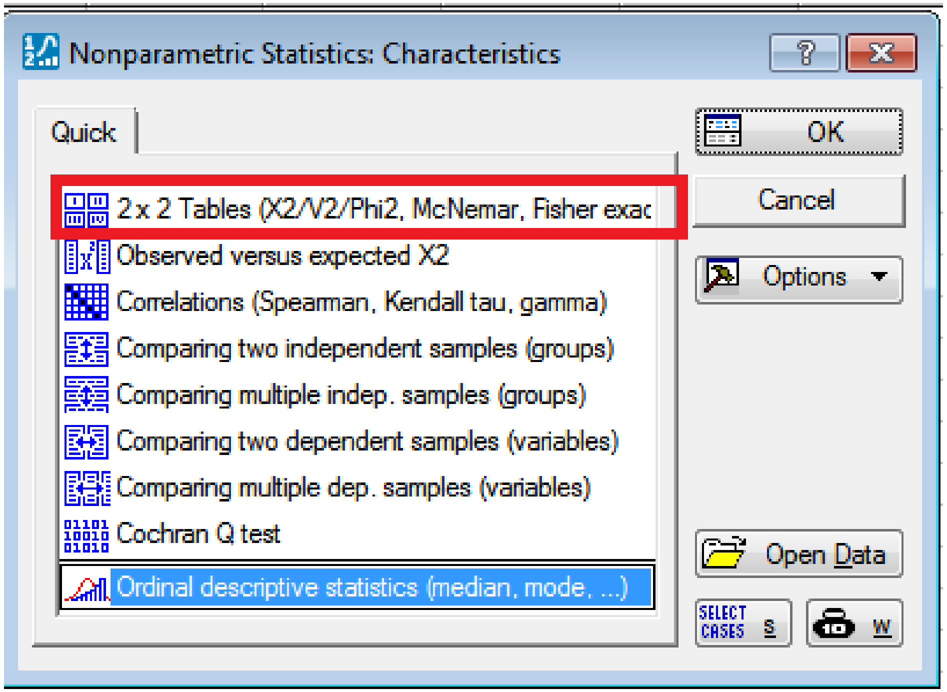
Basic statistics> tables and banners>Select variables>Advanced>Detailed Two-way Tables

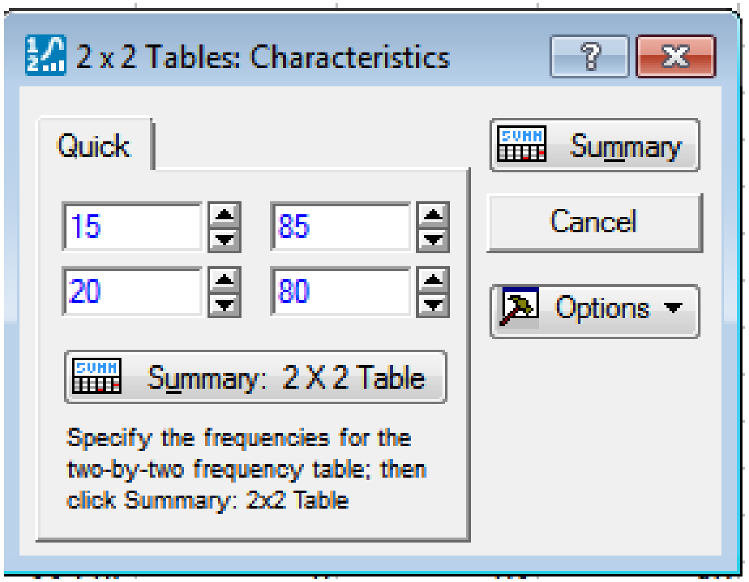
STATISTICA- TASK 3. (1)

In an investigation into eye colour and left- or right-handedness the following results were obtained. Is there eveidence, at the significance 5% level, of an association between eye colour and left- or right-handedness?

		Handedness	
		Left	Right
Eye colour	Blue	15	85
	Brown	20	80

STATISTICA- TASK 3. (2)





Nonparametric statistics> 2 x 2 Tables

STATISTICA- TASK 3. (3)

	2 x 2 Table Column 1	Column 2	Row Totals
Frequencies, row 1	15	85	100
Percent of total	7,500%	42,500%	50,000%
Frequencies, row 2	20	80	100
Percent of total	10,000%	40,000%	50,000%
Column totals	35	165	200
Percent of total	17,500%	82,500%	
Chi-square (df=1)	,87	p= ,3521	
V-square (df=1)	,86	p= ,3533	
Yates corrected Chi-square	,55	p= ,4566	
Phi-square	,00433		
Fisher exact p, one-tailed		p= ,2285	
two-tailed		p= ,4570	
McNemar Chi-square (A/D)	43,12	p= ,0000	
Chi-square (B/C)	39,01	p= ,0000	

LITERATURE

McClave, J., Benson, G. & Sincich, T. (2008). Statistics for Business & Economics. Pearson International Edition, p. 553-594

THANK YOU FOR YOUR ATTENTION