MAT20306 - Advanced Statistics

Lecture 4: Chi-square tests and correlation

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to Life

Situation 13: 1 sample, 2 nominal variables

Model : **one**(!) random sample of units (students), we measure:

2 nominal variables with r and c categories, respectively

X (continent, r=4) and Y (favorite color, c=3).

The data come in the form a cross-table

Result (e.g):

	Y			
nij	green	yellow	red	ni.
Africa	16	2	5	23
America	4	9	9	22
Asia	7	7	6	20
Europe	9	10	2	21
n.j	36	28	22	86

n_{ij}=number of units with i-th X-category, and j-th Y-category (random)

$n_{i.}$ and $n_{.j}$ are the marginal totals (random); n = n.. sample size (fixed)

Research question: association between X and Y? (→ H_a)

H_0: X and Y are independent

Before the party: Chi-squared test for independence

H_0: $\pi_{ij} = \pi_{i.} \times \pi_{.j}$ i=1,..r ; j=1,..,c
H_a: at least one equality above does not hold

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

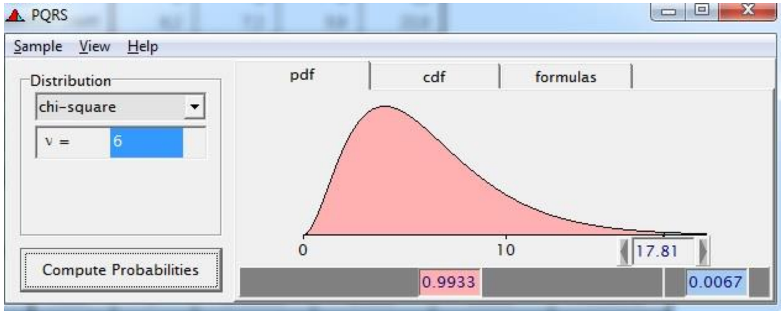
$\hat{E}_{ij}$ = estimated expected cell frequency under H_0,
$n_{ij} = O_{ij}$ = observed cell frequency

Under H_0, χ^2 approximately follows a $\boldsymbol{\chi^2_{(r-1)(c-1)}}$ distribution

Approximation OK if all $\hat{E}_{ij}$'s > 1, and 80% of $\hat{E}_{ij}$'s >5

Under H_a X^2 tends to larger values than under H_0

We use a right-sided R.R. / p value



WAGENINGEN UR
For quality of life

During the party: Descriptive (Sample) Statistics

		Y		
nij	green	yellow	red	ni.
Africa	16	2	5	23
America	4	9	9	22
Asia	7	7	6	20
Europe	9	10	2	21
n.j	36	28	22	86

The after party with PQRS

		Y		
nij	green	yellow	red	ni.
Africa	16	2	5	23
America	4	9	9	22
Asia	7	7	6	20
Europe	9	10	2	21
n.j	36	28	22	86

$E_{ij} = n * \pi_{ij}$; If X and Y are independent, $\pi_{ij} = \pi_{i.} * \pi_{.j}$, so

under H_0 $E_{ij} = n * \pi_{ij} = n * \pi_{i.} * \pi_{.j}$ →

$$\hat{E}_{ij} = n\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = n\,\frac{n_{i.}}{n}\,\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

Given $n_{i.}$ and $n_{.j}$ what do we expect for n_{ij}'s under H_0?

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$ =

= 17.81

WAGENINGEN UR
For quality of life

The after party with PQRS

	Y			
nij	green	yellow	red	ni.
Africa	16	2	5	23
America	4	9	9	22
Asia	7	7	6	20
Europe	9	10	2	21
n.j	36	28	22	86

$$\hat{E}_{ij} = n\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = n\,\frac{n_{i.}}{n}\,\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\mathrm{O}_{\mathrm{ij}}-\hat{E}_{\mathrm{ij}})^2}{\hat{E}_{\mathrm{ij}}} =$$

The after party with PQRS

$i \downarrow$ / $j \to$	nij	green (1)	yellow (2)	red (3)	ni.
1	Africa	16	2	5	23
2	America	4	9	9	22
3	Asia	7	7	6	20
4	Europe	9	10	2	21
	n.j	36	28	22	86

i=1

j=1

$E_{ij} = n * \pi_{ij}$; If X and Y are independent, $\pi_{ij} = \pi_{i.} * \pi_{.j}$, so

under H_0 $E_{ij} = n * \pi_{ij} = n * \pi_{i.} * \pi_{.j} \rightarrow$

$$\hat{E}_{ij} = n\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = n\,\frac{n_{i.}}{n}\,\frac{n_{.j}}{n} = \frac{n_{i.}\,n_{.j}}{n}$$

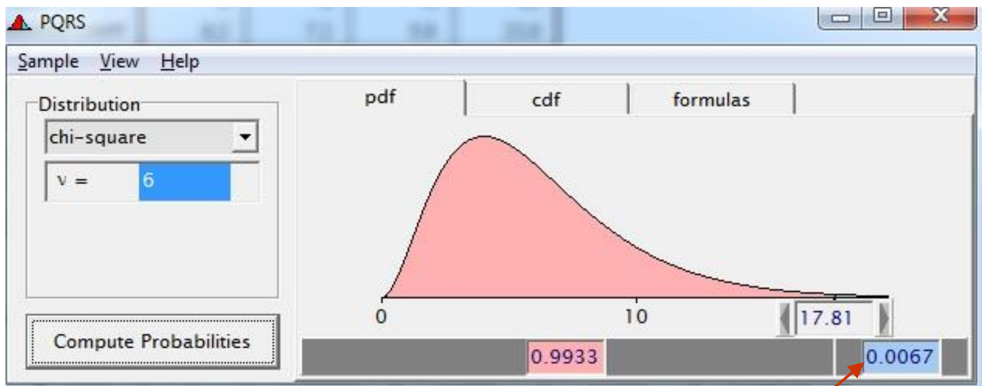
i=j=1

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} =$$

$$\chi^2 = \frac{(O_{11}-\hat{E}_{11})^2}{\hat{E}_{11}} + \frac{(O_{12}-\hat{E}_{12})^2}{\hat{E}_{12}} + \cdots$$

$$\chi^2 = \frac{\left(16 - \frac{23 \cdot 36}{86}\right)^2}{\frac{23\cdot 36}{86}} + \frac{\left(2 - \frac{23\cdot 28}{86}\right)^2}{23\cdot 28/86} + \cdots$$

The after party with PQRS



P-value < 0.05

H_0 rejected H_a shown

There is association between X and Y

The after party with SPSS output

Continent * Colour Crosstabulation

			Colour			Total
			1,00	2,00	3,00	
Continent	1,00	Count	16	2	5	23
		Expected Count	9,6	7,5	5,9	23,0
	2,00	Count	4	9	9	22
		Expected Count	9,2	7,2	5,6	22,0
	3,00	Count	7	7	6	20
		Expected Count	8,4	6,5	5,1	20,0
	4,00	Count	9	10	2	21
		Expected Count	8,8	6,8	5,4	21,0
Total		Count	36	28	22	86
		Expected Count	36,0	28,0	22,0	86,0

H_0 rejected H_a shown
There is association between X and Y.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	17,810[a]	6	,007	,006		
Likelihood Ratio	19,728	6	,003	,005		
Fisher's Exact Test	18,436			,004		
Linear-by-Linear Association	,077[b]	1	,781	,814	,414	,045
N of Valid Cases	86					

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,12.

b. The standardized statistic is ,277.

Situation 14: r samples, 1 nominal variable

Model 2: r random samples are taken (students from r continents),
1 nominal variable is measured, Y, with c possible outcomes

The data come again in a cross-table; the $n_{i.}$ are fixed sample sizes, chosen in advance in the design.

		Y		
nij	green	yellow	red	ni.
Africa	15	2	3	20
America	5	10	5	20
Asia	7	7	6	20
Europe	9	9	2	20
n.j	36	28	16	80

The nominal variable has c possible outcomes.

H_0 : the probabilities for these c outcomes are the same across the r samples.

Situation 14: Null-hypothesis of Homogeneity

		Y		
nij	green	yellow	red	ni.
Africa	15	2	3	20
America	5	10	5	20
Asia	7	7	6	20
Europe	9	9	2	20
n.j	36	28	16	80

Example: 4 Continents (r=4) and 3 colors, (c=3)

- $\pi_{1g} = \pi_{2g} = \pi_{3g} = \pi_{4g}$ (equal GREEN prob's)

  $\pi_{1y} = \pi_{2y} = \pi_{3y} = \pi_{4y}$ (equal YELLOW prob's)

  ($\pi_{1r} = \pi_{2r} = \pi_{3r} = \pi_{4r}$ then holds automatically)

Before the party: Chi-squared test for homogeneity

$H_0: \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj}$, $j=1,\ldots,c$

H_a: at least one equality above does not hold

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

$\hat{E}_{ij}$ = estimated expected cell frequency under H_0,

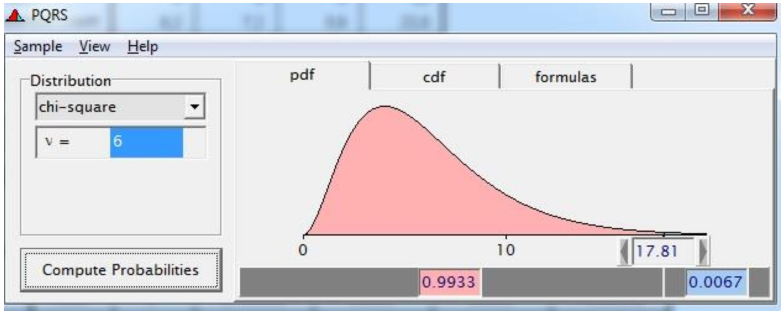
$n_{ij} = O_{ij}$ = observed cell frequency

Under H_0, χ^2 approximately follows a $\chi^2_{(r-1)(c-1)}$ distribution

Approximation OK if all $\hat{E}_{ij}$'s > 1, and 80% of $\hat{E}_{ij}$'s >5

Under H_a X^2 tends to larger values than under H_0

We use a right-sided R.R. / p value



During the party: Descriptive (Sample) Statistics

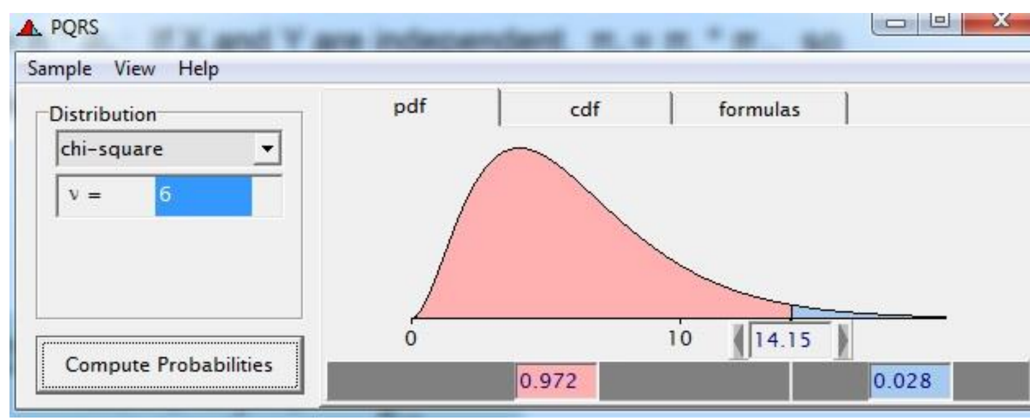
Y

nij	green	yellow	red	ni.
Africa	15	2	3	20
America	5	10	5	20
Asia	7	7	6	20
Europe	9	9	2	20
n.j	36	28	16	80

The after party with PQRS

$$\hat{E}_{ij} = n\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = n\,\frac{n_{i.}}{n}\,\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = 14.15$$



The after party with SPSS output

			Colour			Total
			1,00	2,00	3,00	
Continent	1,00	Count	15	2	3	20
		Expected Count	9,0	7,0	4,0	20,0
	2,00	Count	5	10	5	20
		Expected Count	9,0	7,0	4,0	20,0
	3,00	Count	7	7	6	20
		Expected Count	9,0	7,0	4,0	20,0
	4,00	Count	9	9	2	20
		Expected Count	9,0	7,0	4,0	20,0
Total		Count	36	28	16	80
		Expected Count	36,0	28,0	16,0	80,0

H_0 rejected H_a accepted
There is no homogeneity of colours across continents.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	14,151[a]	6	,028	,026		
Likelihood Ratio	15,173	6	,019	,028		
Fisher's Exact Test	14,223			,023		
Linear-by-Linear Association	,824[b]	1	,364	,402	,201	,034
N of Valid Cases	80					

a. 4 cells (33,3%) have expected count less than 5. The minimum expected count is 4,00.

b. The standardized statistic is ,908.

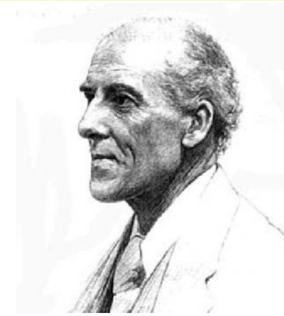
Final remarks about Chi-squared tests

- the test statistic of Pearson's chi-square

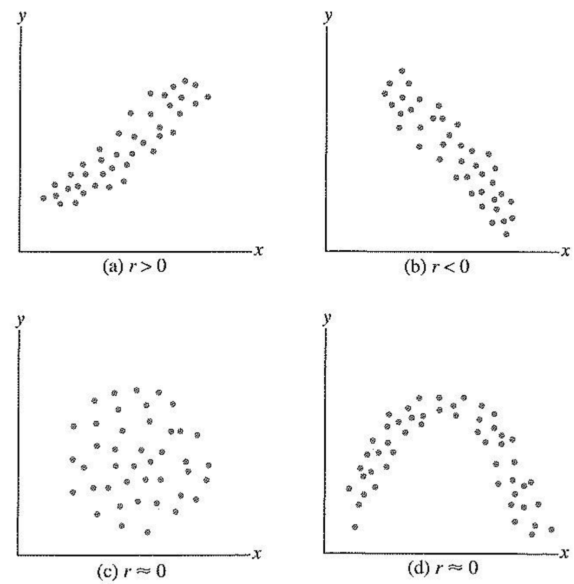
  can be used in combination with the chi-square approximation (rule of thumb about the expected *E* values),
- or it can be used in combination with an exact approach (without a need for the rule of thumb)
- In practice: when an exact P-value is available, we use that.
- In this course we use the approximate P-value obtained with the chi-square approximation, except for the 2x2 case in SPSS (Fisher exact test).

And when you think it’s done ... There is more !

Pearson correlation coefficient



- when people talk about a correlation or correlation coefficient they usually mean Pearson’s correlation coefficient
  - named after Karl Pearson (1857–1936), British statistician
- Pearson’s correlation coefficient ρ_{xy} measures the strength of the linear association between two quantitative variables *x* and *y*, see figure (O&L 11.20)
- ρ_{xy} is always between -1 and $+1$.
- values close to 1 or $-1 \Rightarrow$ strong (linear) association, values close to 0 $\Rightarrow$ little or no (linear) association
- when correlation ρ_{xy} =1 or ρ_{xy} -1,



Pearson correlation coefficient, continued

- There is no distinction between dependent and independent variables: $\rho_{xy} = \rho_{yx}$.

- The absolute value of ρ_{xy} is not affected by linear transformations of x or y, e.g. correlation between x and y is the same as between 2 x + 1 and 10 + 5 y.

  So, it does not matter whether measurements are in e.g. grams or kilograms.

- When x and y are independent, $\rho_{xy} = \rho_{yx} = 0$, but the reverse is not necessarily true.

- The correlation ρ_{xy} is a population parameter that is estimated by the sample correlation r_{xy} :

$$r_{xy} = r_{yx} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad -1 \le r_{xy} \le 1$$

Correlation & inference

- Test on ρ_{xy}

1. H_0: $\rho_{xy} = 0$.

2. Test statistic is:

$$t = r_{xy} \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

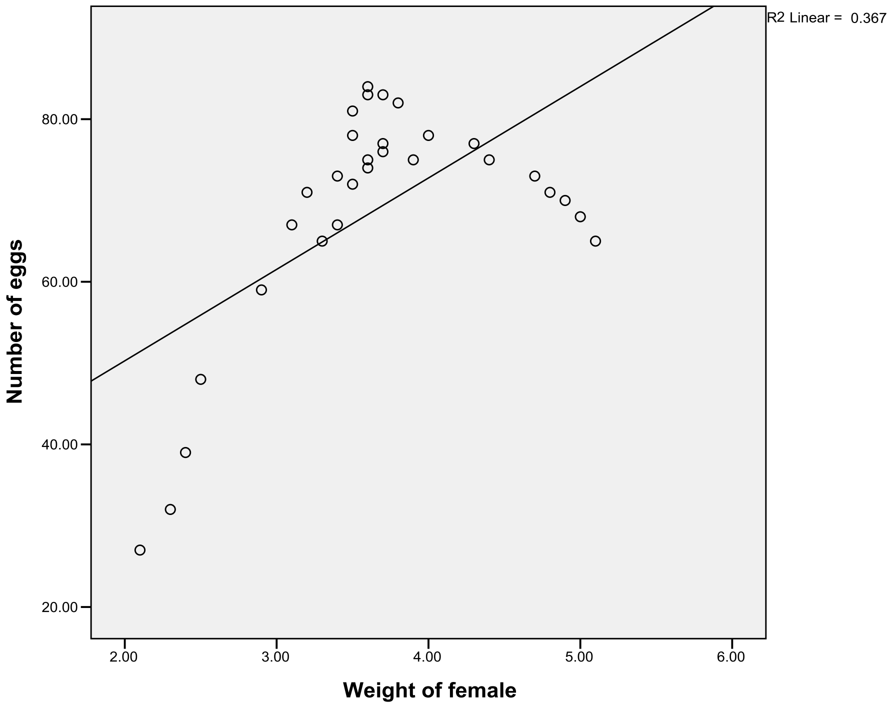
3. When H_0 is true, $t \sim t_{n-2}$

Grasshoppers (Example 11.13 in O&L)

Study of the reproductive success of grasshoppers.

An entomologist collected a sample of 30 female grasshoppers. She recorded the number of mature eggs produced and the body weight of each of the females (grams).

	Number	weight
1	27.00	2.10
2	32.00	2.30
3	39.00	2.40
4	48.00	2.50
5	59.00	2.90
6	67.00	3.10
7	71.00	3.20
8	65.00	3.30
9	73.00	3.40
10	67.00	3.40
11	78.00	3.50
12	72.00	3.50
13	81.00	3.50
14	74.00	3.60
15	83.00	3.60

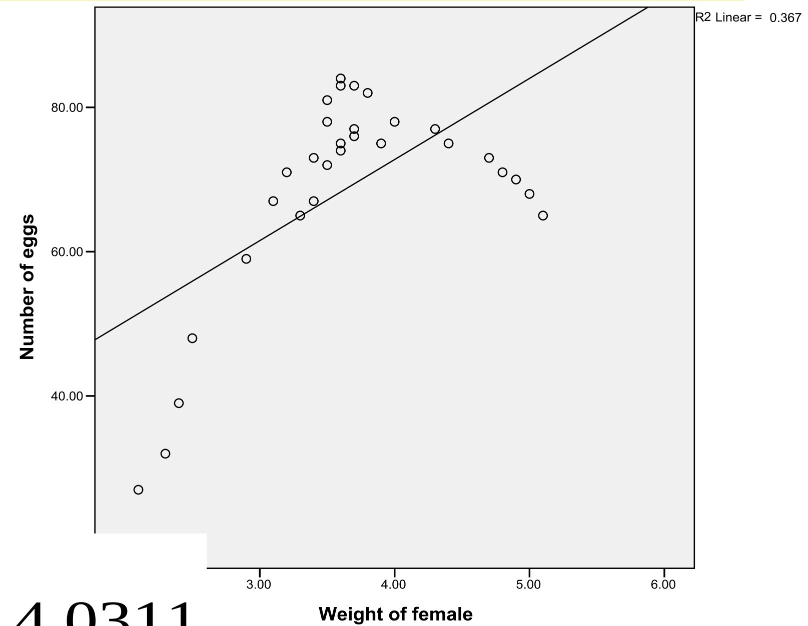


Grasshoppers (Example 11.13 in O&L)

Correlations

		Number of eggs	Weight of female
Number of eggs	Pearson Correlation	1	.606**
	Sig. (2-tailed)		.000
	N	30	30
Weight of female	Pearson Correlation	.606**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).



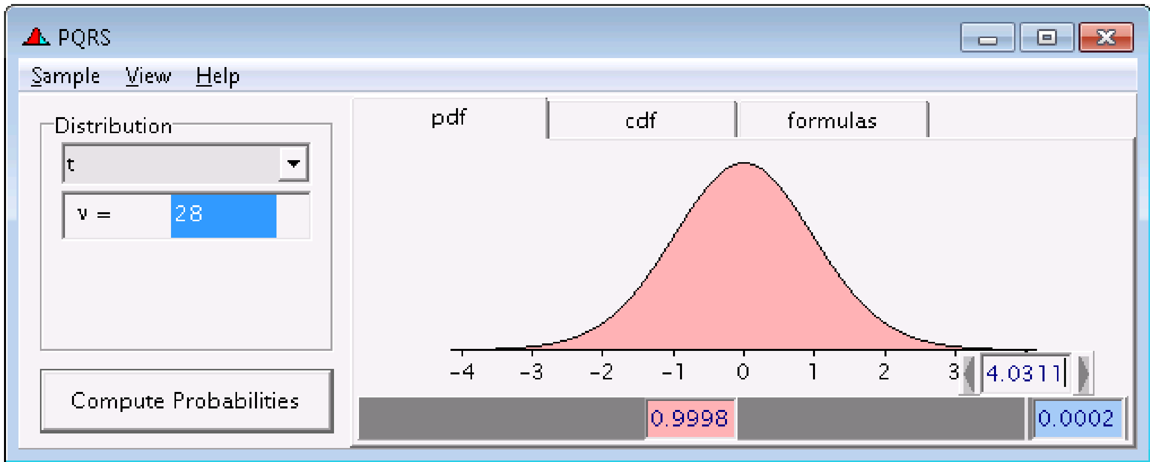
H_0: $\rho_{xy} = 0$ vs H_A: $\rho_{xy} > 0$

$$t = r_{xy}\frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = 0.606 \cdot \frac{\sqrt{30-2}}{\sqrt{1-0.606^2}} = 4.0311$$

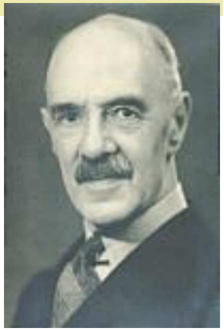
Under H_0: $t_{n-2}=t_{28}$ distribution

RSP=0.000 <0.05, so reject H_0

We have shown there is a positive correlation between weight and number of eggs

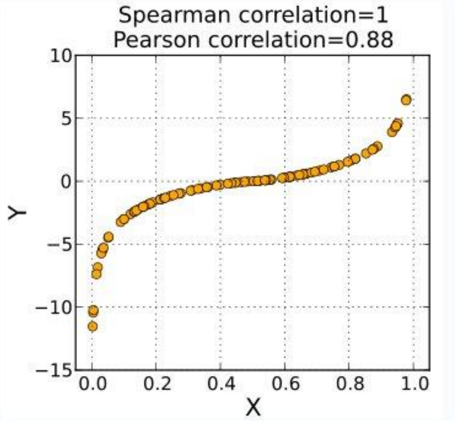


Spearman rank correlation



- r_{xy} is highly sensitive to outlying observations (outliers)
- an alternative is Spearman's rank correlation r_S (not mentioned in O&L), named after Charles Spearman (1863 – 1945), English psychologist

- observations are replaced by rank numbers
  ranking x and y separately, with mid ranks in case of ties
- Spearman's r_S is the ordinary correlation, but derived from these rank numbers

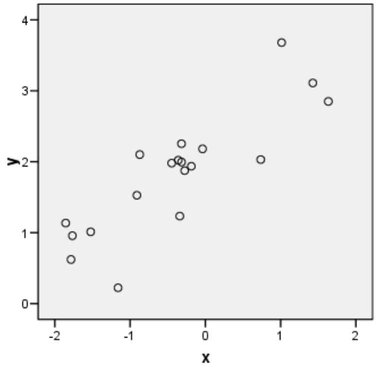


- r_s measures the strength of a monotonic relationship between two quantitative variables x and y.
  The relationship need not be linear, see figure from Wikipedia.

- when data are approximately normally distributed (without outliers), r_S and r_{xy} tend to be similar.
- but r_s is not estimating a population parameter, in contrast to r_{xy} ,

An example of Spearman’s rank correlation

x	y
-.91	1.53
-.04	2.18
-.28	1.88
-.36	2.02
-1.86	1.14
-1.77	.96
-.32	2.25
1.63	2.85
-.19	1.94
-.32	2.00
-.34	1.23
-1.16	.22
1.43	3.11
-.87	2.10
-.45	1.98
-1.79	.62
-1.52	1.01
1.01	3.68
.74	2.03



Correlations

		x	y
x	Pearson Correlation	1	.852**
	Sig. (2-tailed)		.000
	N	19	19
y	Pearson Correlation	.852**	1
	Sig. (2-tailed)	.000	
	N	19	19

**. Correlation is significant at the 0.01 level

Note that here Pearson correlation and Spearman rank correlation are similar.

x	y	Rx	Ry
-.91	1.53	6	7
-.04	2.18	15	15
-.28	1.88	13	8
-.36	2.02	9	12
-1.86	1.14	1	5
-1.77	.96	3	3
-.32	2.25	11	16
1.63	2.85	19	17
-.19	1.94	14	9
-.32	2.00	12	11
-.34	1.23	10	6
-1.16	.22	5	1
1.43	3.11	18	18
-.87	2.10	7	14
-.45	1.98	8	10
-1.79	.62	2	2
-1.52	1.01	4	4
1.01	3.68	17	19
.74	2.03	16	13

Spearman’ correlation of 0.821 can be Obtained by calculating Pearson’s correlation on rank numbers

Correlations

		Rank of x	Rank of y
Rank of x	Pearson Correlation	1	.821**
	Sig. (2-tailed)		.000
	N	19	19
Rank of y	Pearson Correlation	.821**	1
	Sig. (2-tailed)	.000	
	N	19	19

**. Correlation is significant at the 0.01 level (2-tailed).

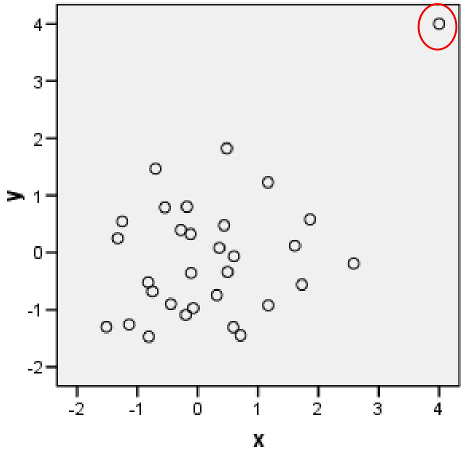
Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.821**
		Sig. (2-tailed)	.	.000
		N	19	19
	y	Correlation Coefficient	.821**	1.000
		Sig. (2-tailed)	.000	.
		N	19	19

**. Correlation is significant at the 0.01 level (2-tailed).

Another example of Spearman's rank correlation

Unrelated x and y, with one added outlying observation with both high x and y value.



Correlations

		x	y
x	Pearson Correlation	1	.445*
	Sig. (2-tailed)		.012
	N	31	31
y	Pearson Correlation	.445*	1
	Sig. (2-tailed)	.012	
	N	31	31

*. Correlation is significant at the 0.05 level (2-tailed).

Pearson correlation is sensitive to the outlier: relatively high correlation (and significantly different from 0)

Spearman correlation is not really sensitive to the outlier and consequently lower

Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.185
		Sig. (2-tailed)	.	.319
		N	31	31
	y	Correlation Coefficient	.185	1.000
		Sig. (2-tailed)	.319	.
		N	31	31

HOW ARE YOU DOING??

or