Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 2
Investigating Relationships (part 1)

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Relationship between variables

If we analyze the relationship between a pair of variables one of them is called independent (or explanatory) the other – dependent (or explainable). The quality of the variables relationship model could be estimated by the relation coefficients.

Depending on the task of the research, the variables scales and the data availability we should select an appropriate method for analyzing the relationship.

Contingency table

- Relationship between categorical variables (nominal or ordinal) could be presented by a contingency table.

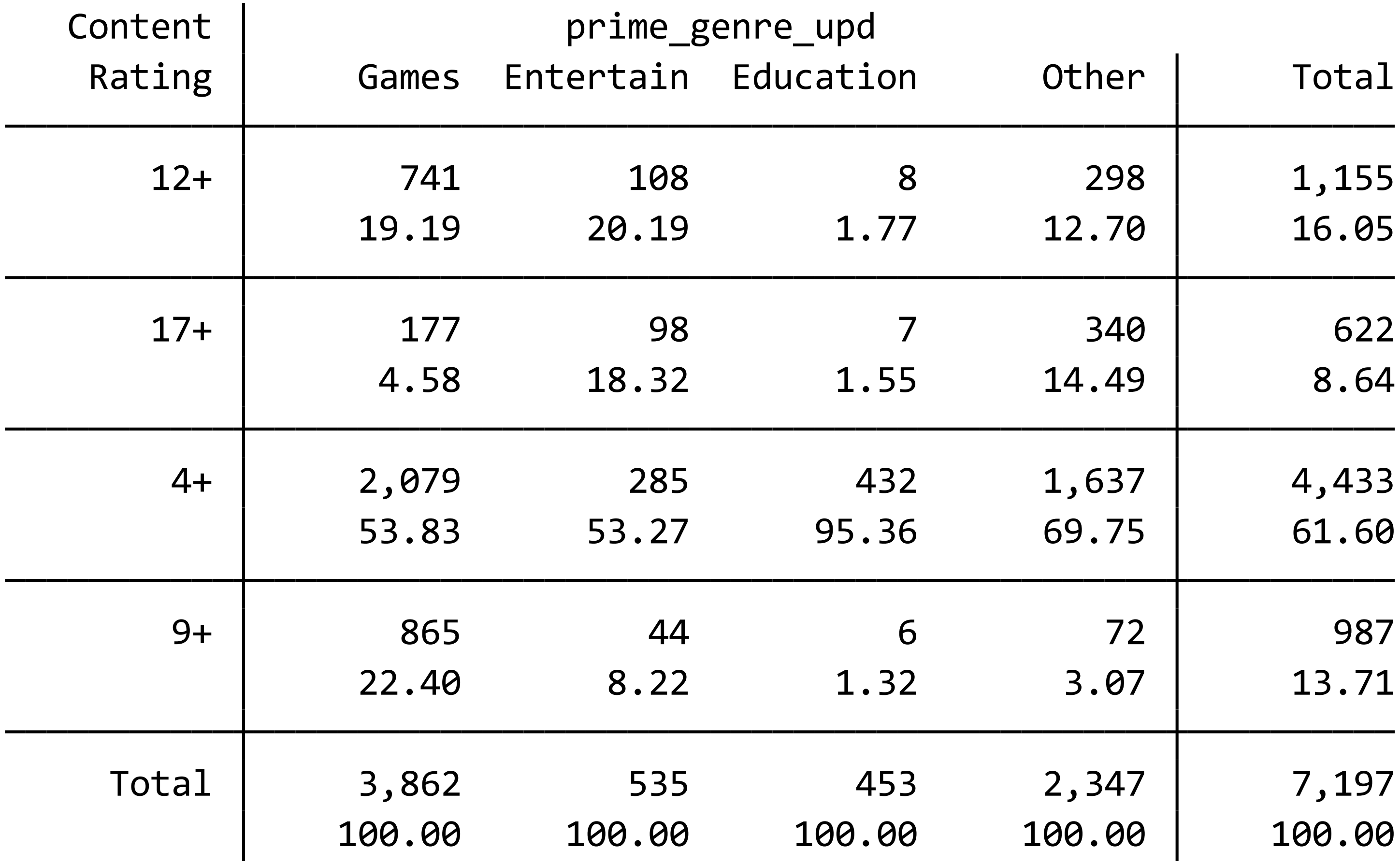
- The contingency tables could help to reveal the existence of statistical, but not causal relationship.

- Together with the contingency tables different coefficients that evaluate the force of the variable relationship could be calculated.

Contingency table with frequencies

```
   Content |              prime_genre_upd
    Rating |     Games  Entertain  Education      Other |     Total
-----------+--------------------------------------------+----------
       12+ |       741        108          8        298 |     1,155
       17+ |       177         98          7        340 |       622
        4+ |     2,079        285        432      1,637 |     4,433
        9+ |       865         44          6         72 |       987
-----------+--------------------------------------------+----------
     Total |     3,862        535        453      2,347 |     7,197
```

Contingency table with frequencies and column percentages

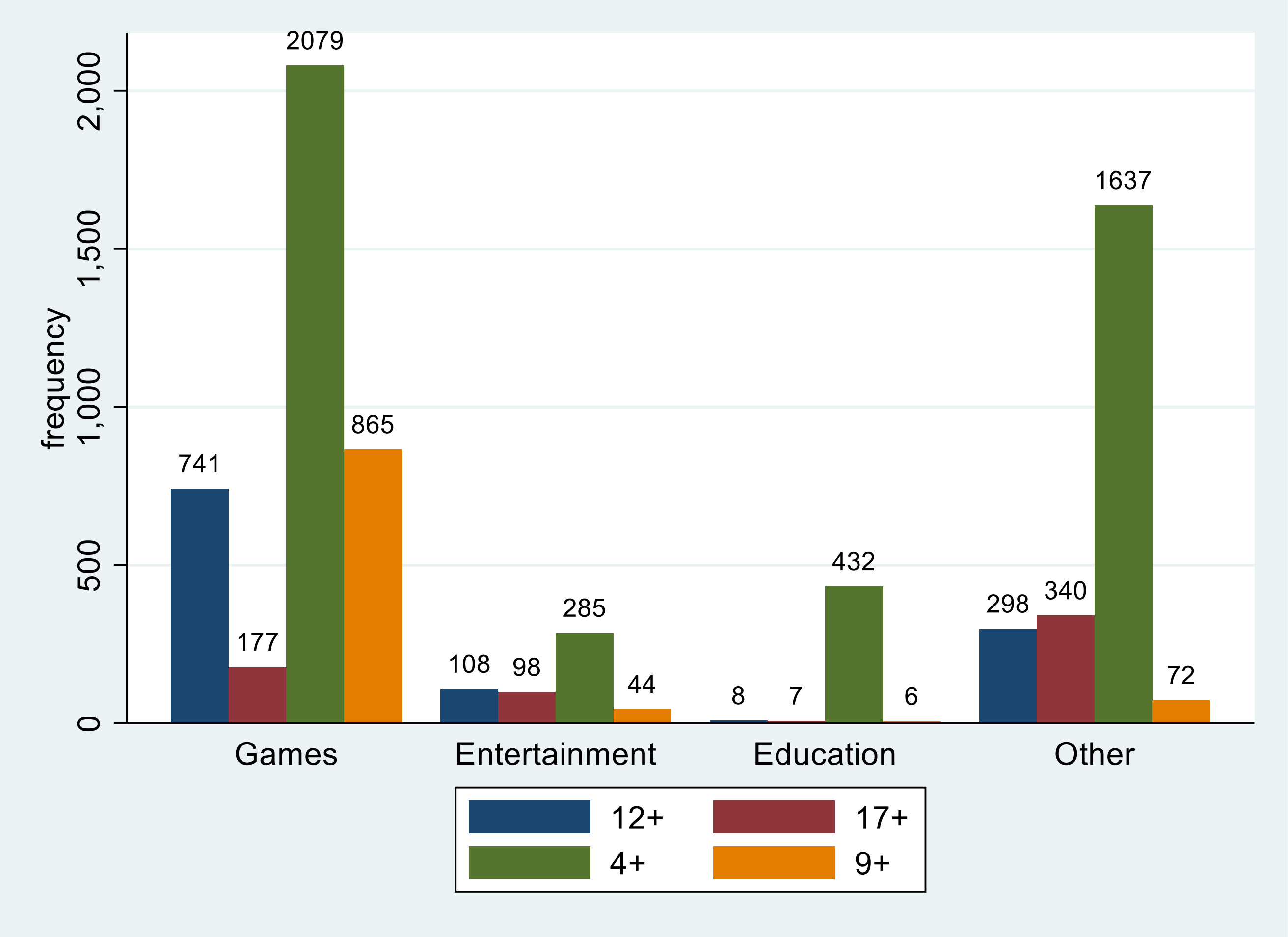


```
  Content |                 prime_genre_upd
   Rating |     Games  Entertain  Education      Other |     Total
----------+--------------------------------------------+----------
      12+ |       741        108          8        298 |     1,155 
          |     19.19      20.19       1.77      12.70 |     16.05 
----------+--------------------------------------------+----------
      17+ |       177         98          7        340 |       622 
          |      4.58      18.32       1.55      14.49 |      8.64 
----------+--------------------------------------------+----------
       4+ |     2,079        285        432      1,637 |     4,433 
          |     53.83      53.27      95.36      69.75 |     61.60 
----------+--------------------------------------------+----------
       9+ |       865         44          6         72 |       987 
          |     22.40       8.22       1.32       3.07 |     13.71 
----------+--------------------------------------------+----------
    Total |     3,862        535        453      2,347 |     7,197 
          |    100.00     100.00     100.00     100.00 |    100.00 
```

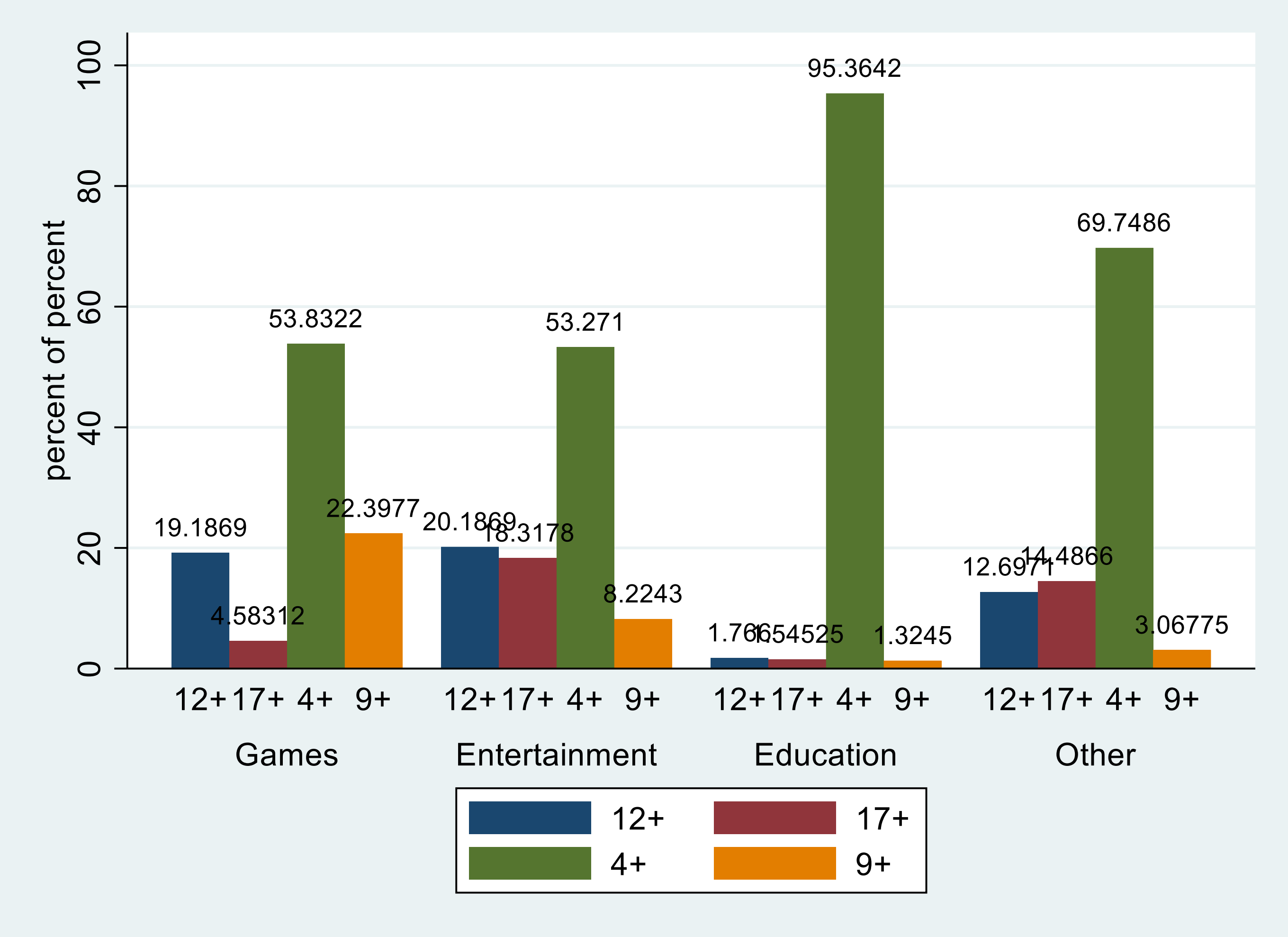
Contingency table with frequencies and row percentages

```
    Content |               prime_genre_upd
     Rating |     Games  Entertain  Education      Other |     Total
------------+--------------------------------------------+----------
        12+ |       741        108          8        298 |     1,155 
            |     64.16       9.35       0.69      25.80 |    100.00 
------------+--------------------------------------------+----------
        17+ |       177         98          7        340 |       622 
            |     28.46      15.76       1.13      54.66 |    100.00 
------------+--------------------------------------------+----------
         4+ |     2,079        285        432      1,637 |     4,433 
            |     46.90       6.43       9.75      36.93 |    100.00 
------------+--------------------------------------------+----------
         9+ |       865         44          6         72 |       987 
            |     87.64       4.46       0.61       7.29 |    100.00 
------------+--------------------------------------------+----------
      Total |     3,862        535        453      2,347 |     7,197 
            |     53.66       7.43       6.29      32.61 |    100.00 
```

Bar chart with frequencies in multiple groups



Bar chart with percentages in multiple groups



Inferential Statistics

Inferential statistics – consists of generalizing from sample to populations, performing hypothesis testing, determining relationships among variables and making predictions. We use probability to determine whether it is likely that a particular test outcome is representative of the population.

Statistics

Descriptive

The purpose is to describe and summarize the data obtained from a sample. We can use numbers and graphs: bar and pie charts, histogram (skewness and kurtosis), frequency distribution table, central tendency statistics (mean, median, mode), measures of variability (range, variance, standard deviation).

Inferential

The purpose is to generalize to a larger population from the data obtained in a sample. We use probability to define how confident we can be that the conclusions we make are correct. It incudes hypotheses testing.

Hypotheses Testing

Hypothesis is a testable prediction about a real-world phenomenon.

Experimental hypothesis proposes a starting point that will be accepted or rejected by examining the evidence that supports or contradicts it. If the evidence supports the hypothesis, we **accept** it, if not we **reject** it.

Experimental Hypotheses

There are two Experimental Hypotheses:

- Null Hypothesis (H_0)
- Alternative Hypothesis (H_1 or H_A)

Null Hypothesis

Examples on Null Hypothesis:

H0: There is **no difference** in average values across groups.

H0: There is **no relationship** between the variables.

H0: There is **no difference between** the distributions.

Any differences or relationships that we observe are due to chance.

Alternative Hypothesis

Examples of Alternative Hypothesis:

H0: There is **a difference** in average values across groups.

H0: There is **a relationship** between the variables.

H0: There is **a difference between** the distributions.

Any differences or relationships are due to an effect.

Steps of statistical analysis

1. Generate hypothesis (or hypotheses).
2. Collect some useful data to check the hypothesis.
3. Fit a statistical model to the data – this model will test the original prediction.
4. Assess this model to see whether it supports the initial predictions.

Testing hypotheses

Research or Experimental hypothesis (H1) – prediction that your experimental manipulation will have some effect or that certain variables will relate to each other.

Null hypothesis (H0) – assumption that your prediction is wrong and that the predicted effect doesn’t exist.

Results of testing hypotheses:

H0 is rejected, H1 is accepted;

H0 is not rejected, H1 is rejected.

p-value

P-value is a number between 0 and 1. It shows how confident we should be that our experimental hypothesis (H1) is true. The closer p-value is to 0, the more confident are we that H1 is true. The question is how small does a p-value have to be so as we become sufficiently confident that H1 is true, what threshold can we use to make a good decision?

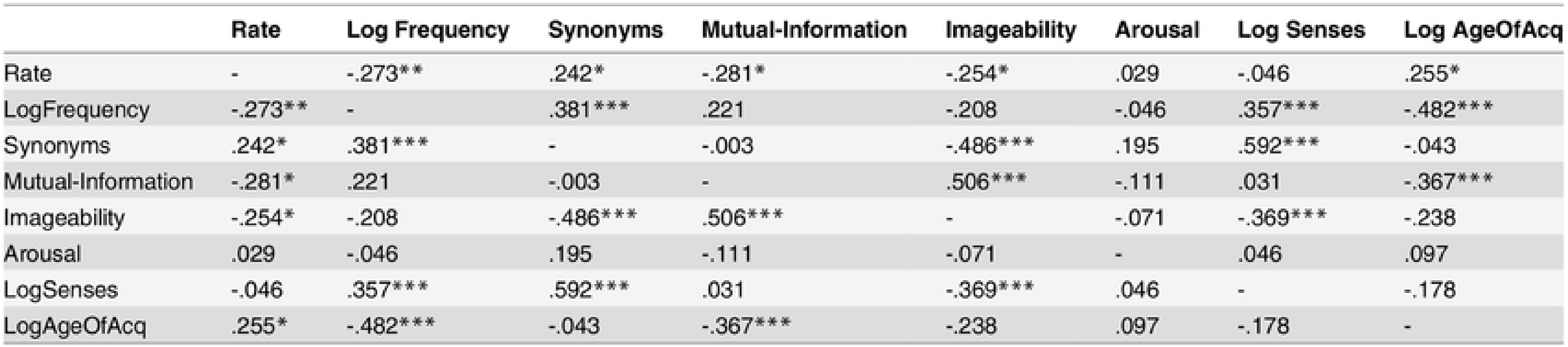
A commonly used threshold is **0.05**. It means that if we do the same analysis/experiments on different samples than only 5% of those experiments would give an opposite result.

Level of significance

If it’s extremely important that we are correct about the conclusions we make, then we can use a smaller threshold, like 0.00001. It means that we would be not correct only once every 100 000 experiments.

If it’s not that important to be correct, for example we are predicting if the bus will arrive on time, then we can use a larger threshold, like 0.2. It means that we would be not correct only 2 times out of 10.

The most common threshold is **0.05** because trying to reduce it often costs more than it worth.



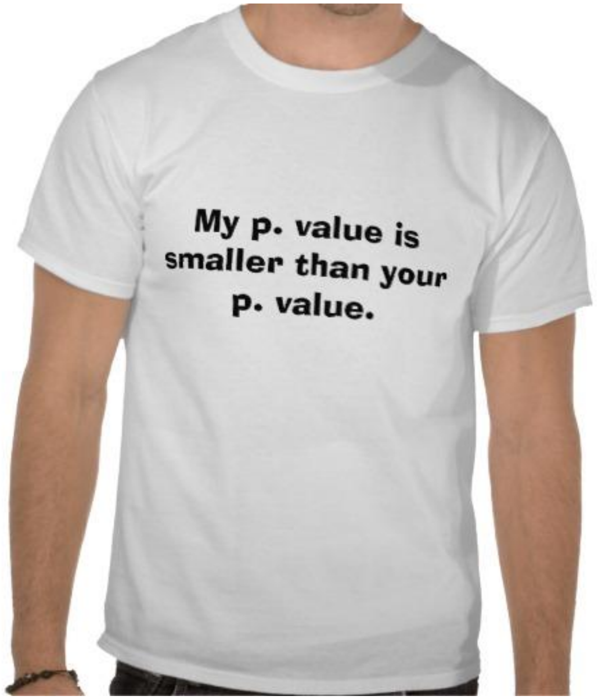
	Rate	Log Frequency	Synonyms	Mutual-Information	Imageability	Arousal	Log Senses	Log AgeOfAcq
Rate	-	-.273**	.242*	-.281*	-.254*	.029	-.046	.255*
LogFrequency	-.273**	-	.381***	.221	-.208	-.046	.357***	-.482***
Synonyms	.242*	.381***	-	-.003	-.486***	.195	.592***	-.043
Mutual-Information	-.281*	.221	-.003	-	.506***	-.111	.031	-.367***
Imageability	-.254*	-.208	-.486***	.506***	-	-.071	-.369***	-.238
Arousal	.029	-.046	.195	-.111	-.071	-	.046	.097
LogSenses	-.046	.357***	.592***	.031	-.369***	.046	-	-.178
LogAgeOfAcq	.255*	-.482***	-.043	-.367***	-.238	.097	-.178	-

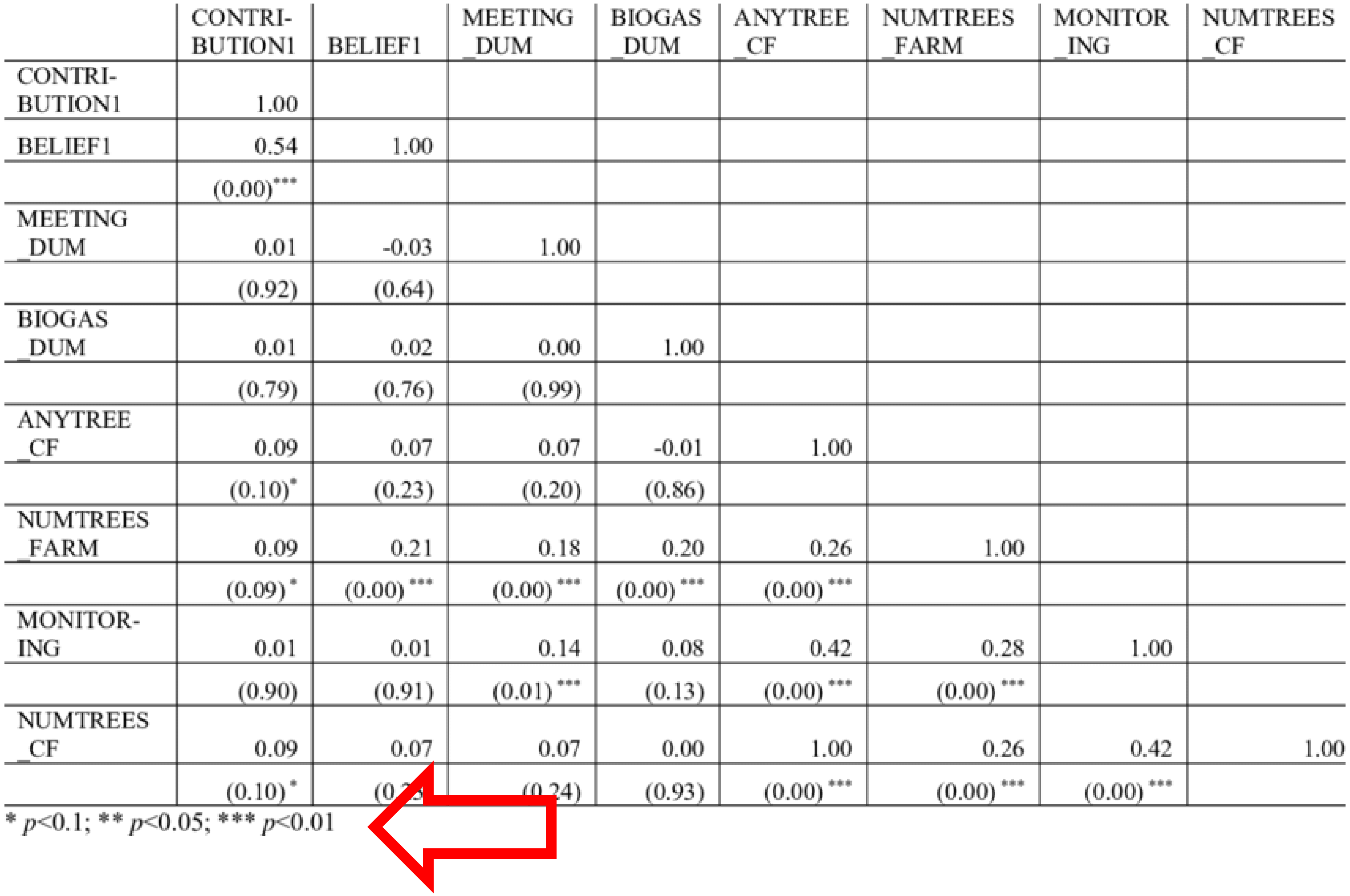
***: $p < .0001$

**: $p < .01$

* $p < .05$.

doi:10.1371/journal.pone.0147924.t002



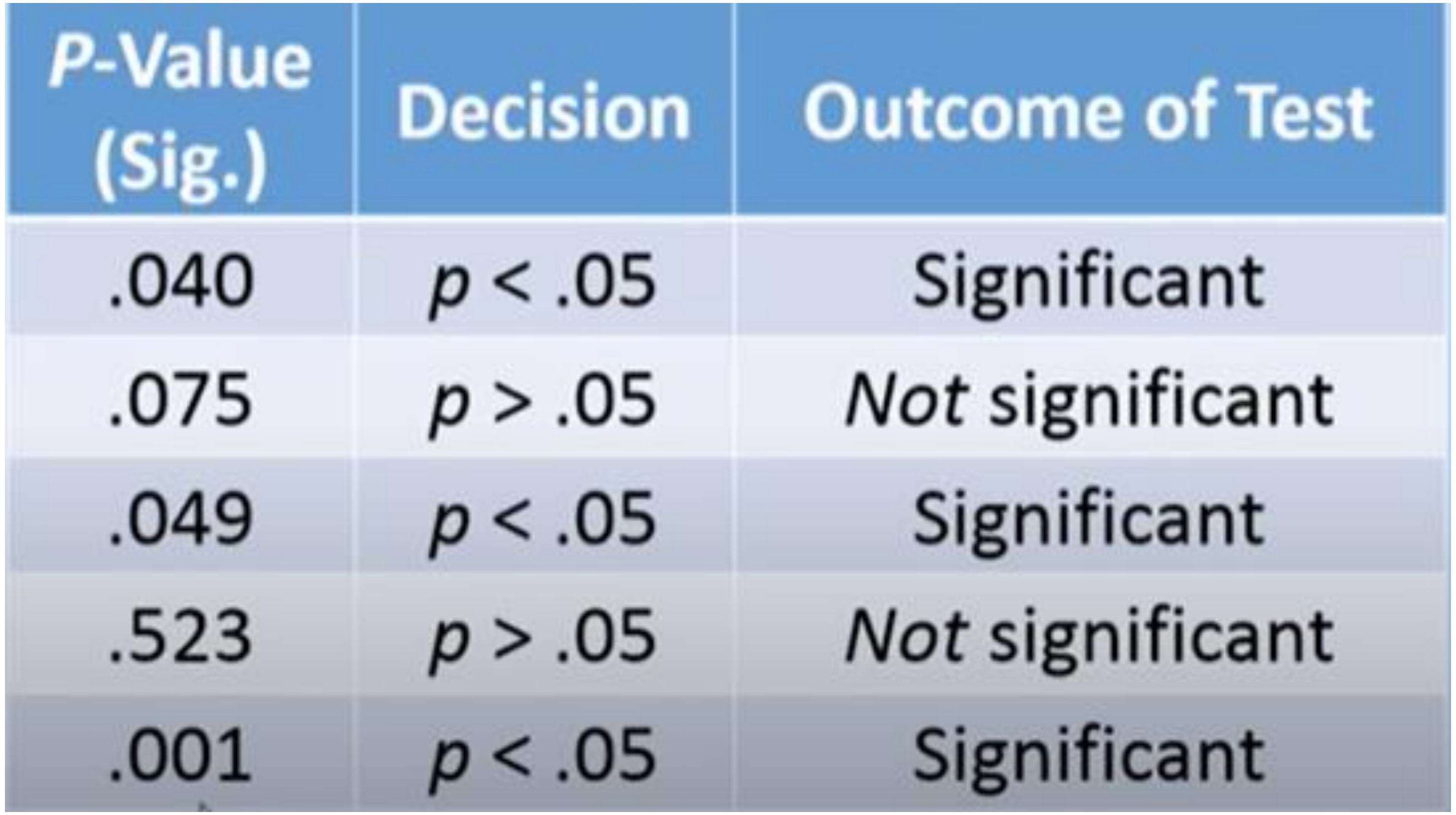


	CONTRIBUTION1	BELIEF1	MEETING_DUM	BIOGAS_DUM	ANYTREE_CF	NUMTREES_FARM	MONITOR_ING	NUMTREES_CF
CONTRIBUTION1	1.00							
BELIEF1	0.54	1.00						
	(0.00)***							
MEETING_DUM	0.01	-0.03	1.00					
	(0.92)	(0.64)						
BIOGAS_DUM	0.01	0.02	0.00	1.00				
	(0.79)	(0.76)	(0.99)					
ANYTREE_CF	0.09	0.07	0.07	-0.01	1.00			
	(0.10)*	(0.23)	(0.20)	(0.86)				
NUMTREES_FARM	0.09	0.21	0.18	0.20	0.26	1.00		
	(0.09)*	(0.00)***	(0.00)***	(0.00)***	(0.00)***			
MONITORING	0.01	0.01	0.14	0.08	0.42	0.28	1.00	
	(0.90)	(0.91)	(0.01)***	(0.13)	(0.00)***	(0.00)***		
NUMTREES_CF	0.09	0.07	0.07	0.00	1.00	0.26	0.42	1.00
	(0.10)*	(0.23)	(0.24)	(0.93)	(0.00)***	(0.00)***	(0.00)***	

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

P-value: examples

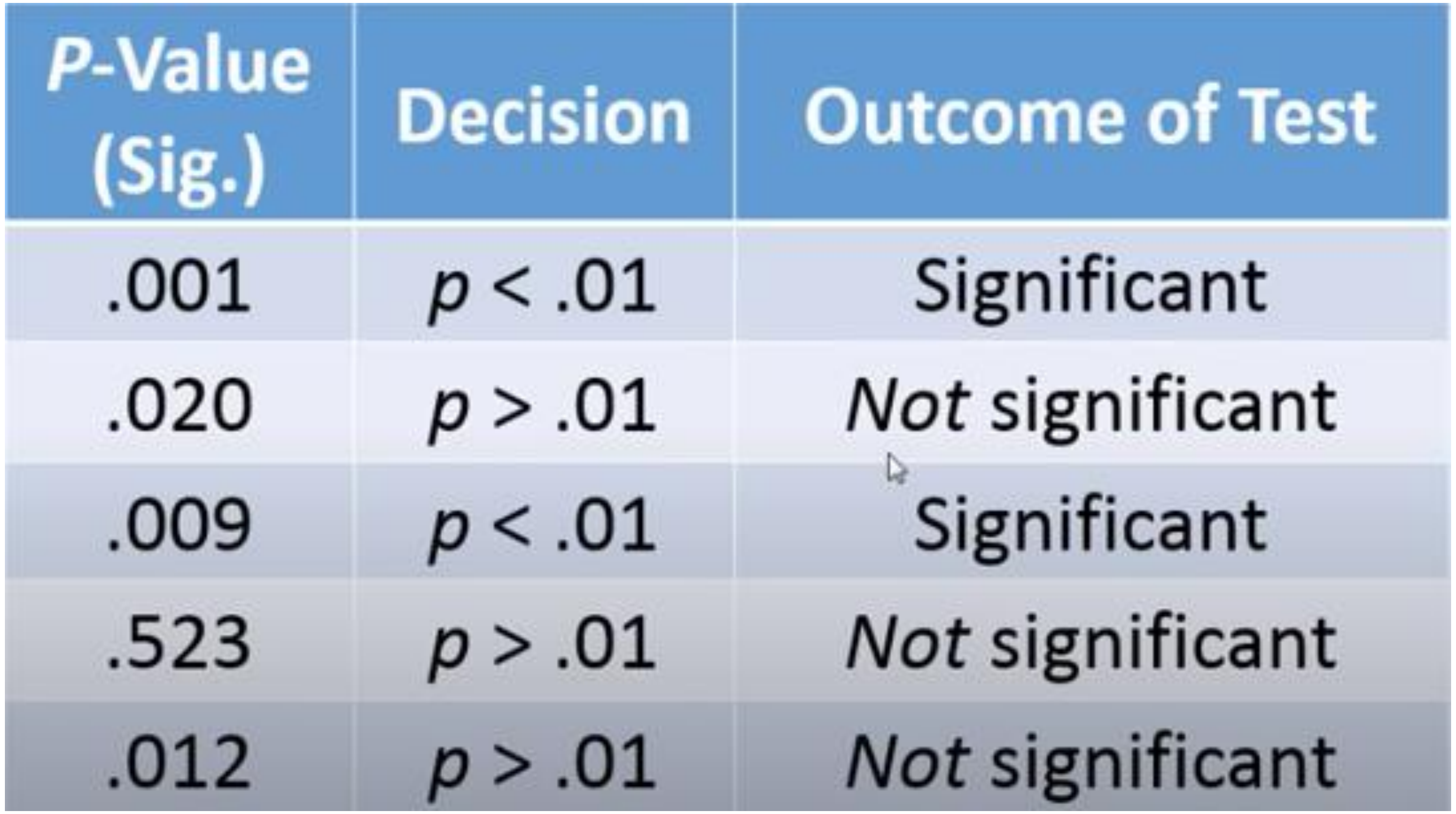
If the level of significance is 0.05



P-Value (Sig.)	Decision	Outcome of Test
.040	$p < .05$	Significant
.075	$p > .05$	*Not* significant
.049	$p < .05$	Significant
.523	$p > .05$	*Not* significant
.001	$p < .05$	Significant

P-value: examples

If the level of significance is 0.01



P-Value (Sig.)	Decision	Outcome of Test
.001	$p < .01$	Significant
.020	$p > .01$	*Not* significant
.009	$p < .01$	Significant
.523	$p > .01$	*Not* significant
.012	$p > .01$	*Not* significant

Hypothesis testing

H1: there is a relationship between students’ attendance and educational results

- p-value = 0.0000001 → H1 is accepted at 1% level
- p-value = 0.049 → H1 is accepted at 5% level
- p-value = 0.051 → H1 is rejected at 5% level, but can be accepted at 10% level
- p-value = 0.73 → H1 is rejected at 10% level

Note about p-value

H1: mean grades of boys and girls are not the same

H2: there is a relationship between students’ attendance and performance

H3: there is a difference between the distribution of students’ grades and normal distribution.

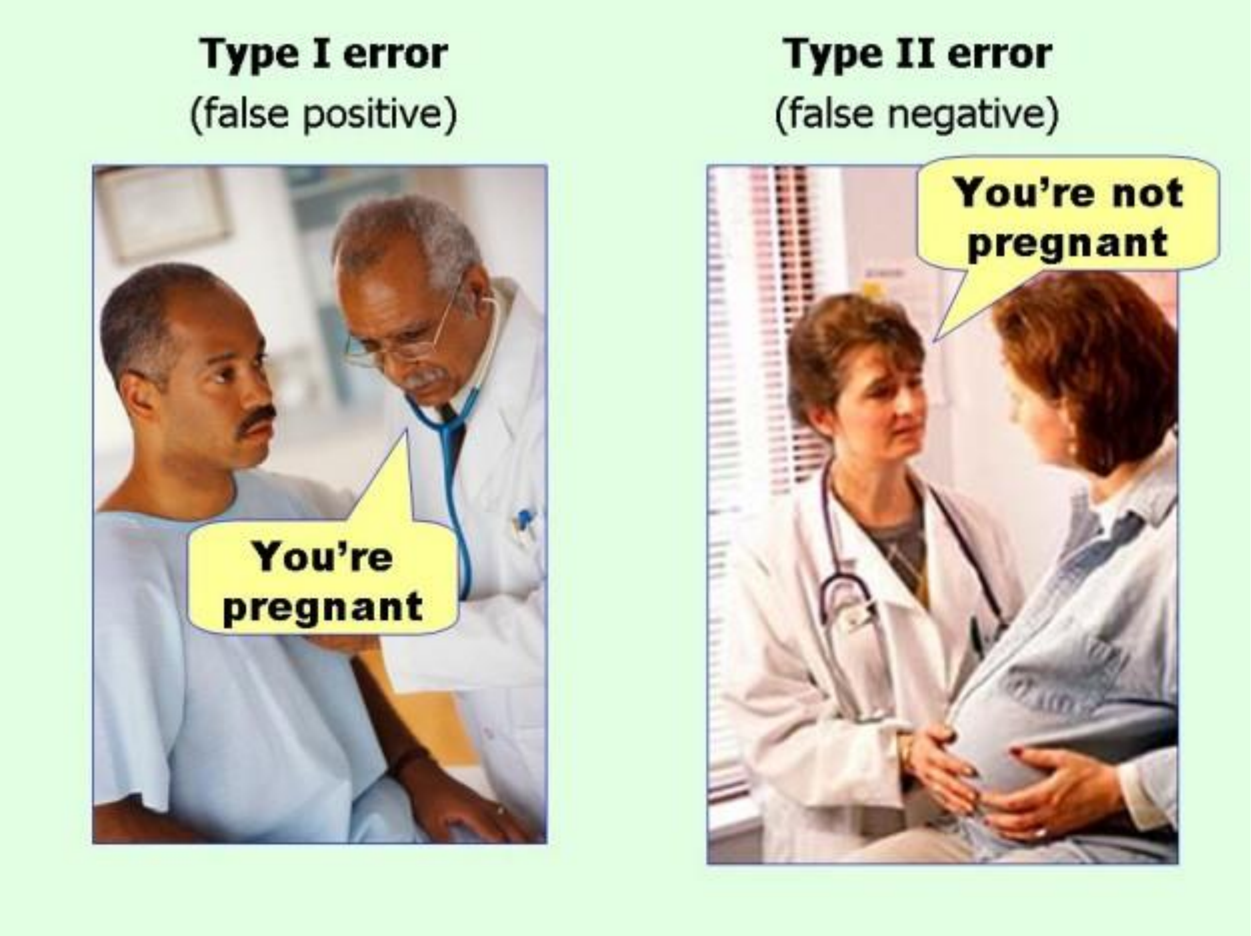
A small p-value is an indicator that we should accept these hypotheses, but it doesn’t tell us how different are the grades of boys and girls, how strong is the relationship between attendance and performance etc. So, you can have the same p-value in case of low and in case of high difference between grades of boys and girls.

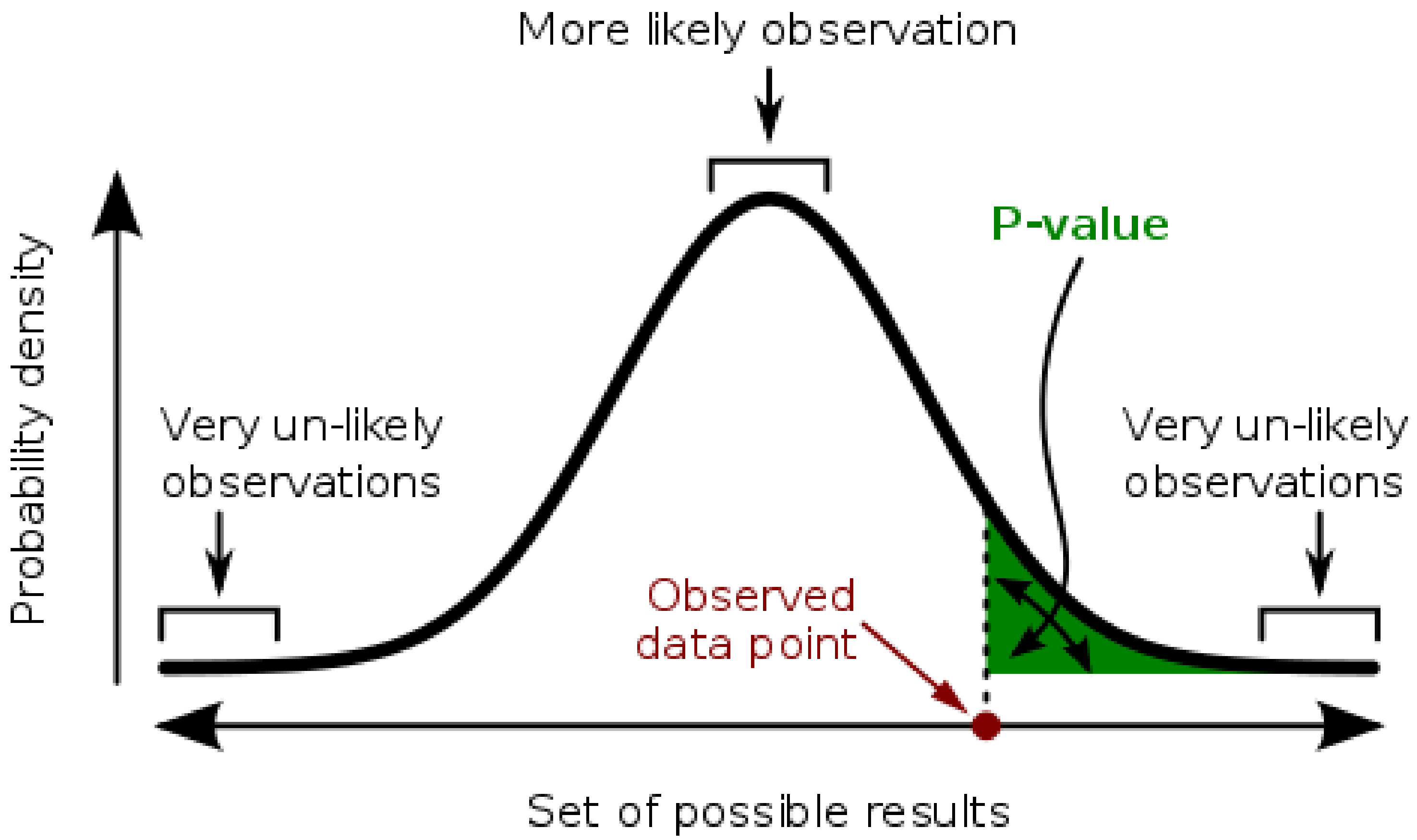
Type I and Type II errors

	H_0 is true	**H_0 is false**
H_0 is not rejected	H_0 is not rejected correctly	**Type 2 error**
H_0 is rejected	**Type 1 error**	H_0 is rejected correctly

- Type 1 error occurs when we believe that there is a genuine effect in our population, when in fact there isn’t. The probability of this error is the level of significance (0,05).
- Type 2 error occurs when we believe that there is no effect in the population when, in reality, there is.

Type I and Type II errors





A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Select appropriate test to analyze the relationship between variables

Step 1: purpose of the research

To select an appropriate test, you should define the purpose of your research. There are two possible major purposes: **comparison** or **relationship**.

Comparison tries to understand whether there is a difference between the groups. For example, grades of boys VS grades of girls, results of control group VS results of treatment group, the distribution of salary VS normal distribution. We can compare the variables' values in 2 or more groups.

Relationship tries to find a connection, for example, relation between age and salary, advertising budget and income.

Select appropriate test to analyze the relationship between variables

Step 2: type of data

We should consider the type of the data we are looking at. We could have **categorical** (qualitative characteristics) or **continuous** (quantitative or numerical) data.

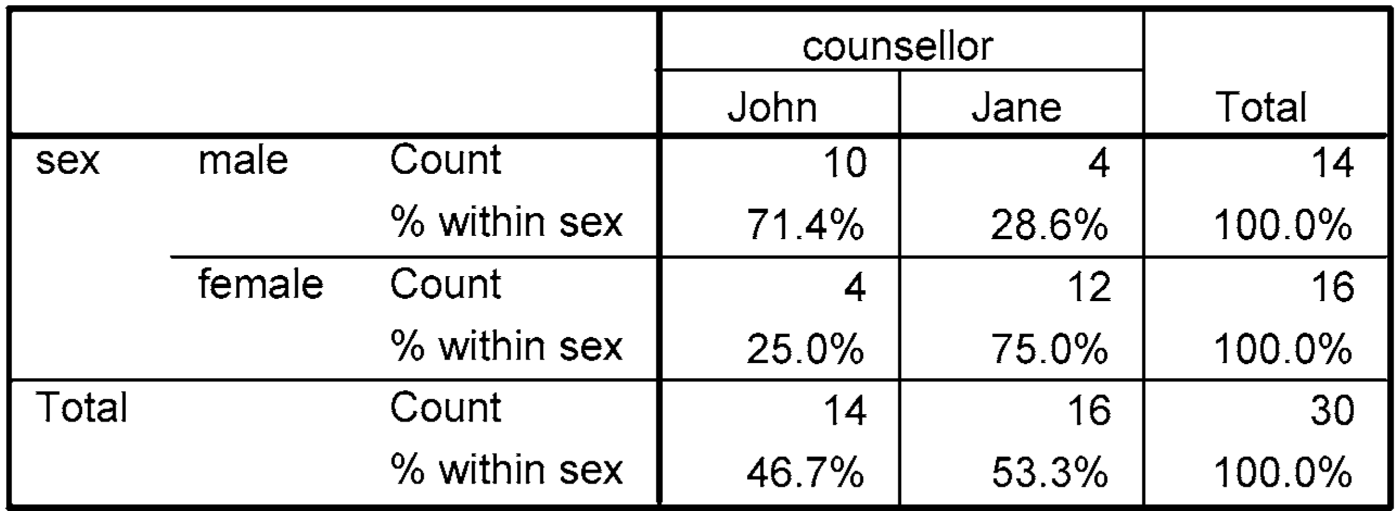
Once you have defined the purpose of the research and the type of the data you have you can select from 3 main families of statistical tests:

- **Chi-Squared**
- **Correlation**
- **t-test/ANOVA**

Chi-squared statistical test

Chi-squared statistical test is used to examine associations between variables. Most frequently it’s used with categorical variables. Let’s look at the contingency table between sex of the patient and the counsellors (Jane and John).

sex * counsellor Crosstabulation



			counsellor		Total
			John	Jane	
sex	male	Count	10	4	14
		% within sex	71.4%	28.6%	100.0%
	female	Count	4	12	16
		% within sex	25.0%	75.0%	100.0%
Total		Count	14	16	30
		% within sex	46.7%	53.3%	100.0%

Chi-squared statistical test

We can see that the sex of patients is not equally distributed across the counsellors. We might ask was this a chance result? If we started the service again and the patients were *randomly* assigned to the counsellors, might we have found a similar result? Perhaps even the opposite – with most male patients seeing the female counsellor. More likely of course, by chance, we would expect them to be distributed quite evenly across the sexes.

The question is: how much of an unequal distribution across the sexes does there have to be for us to conclude that there was a significant bias for male patients to see a male counsellor and vice versa? Or, to put it another way: how much of an unequal distribution across the sexes does there have to be for us to *reject the possibility that this has occurred by chance*? This is where a statistical test comes in handy.

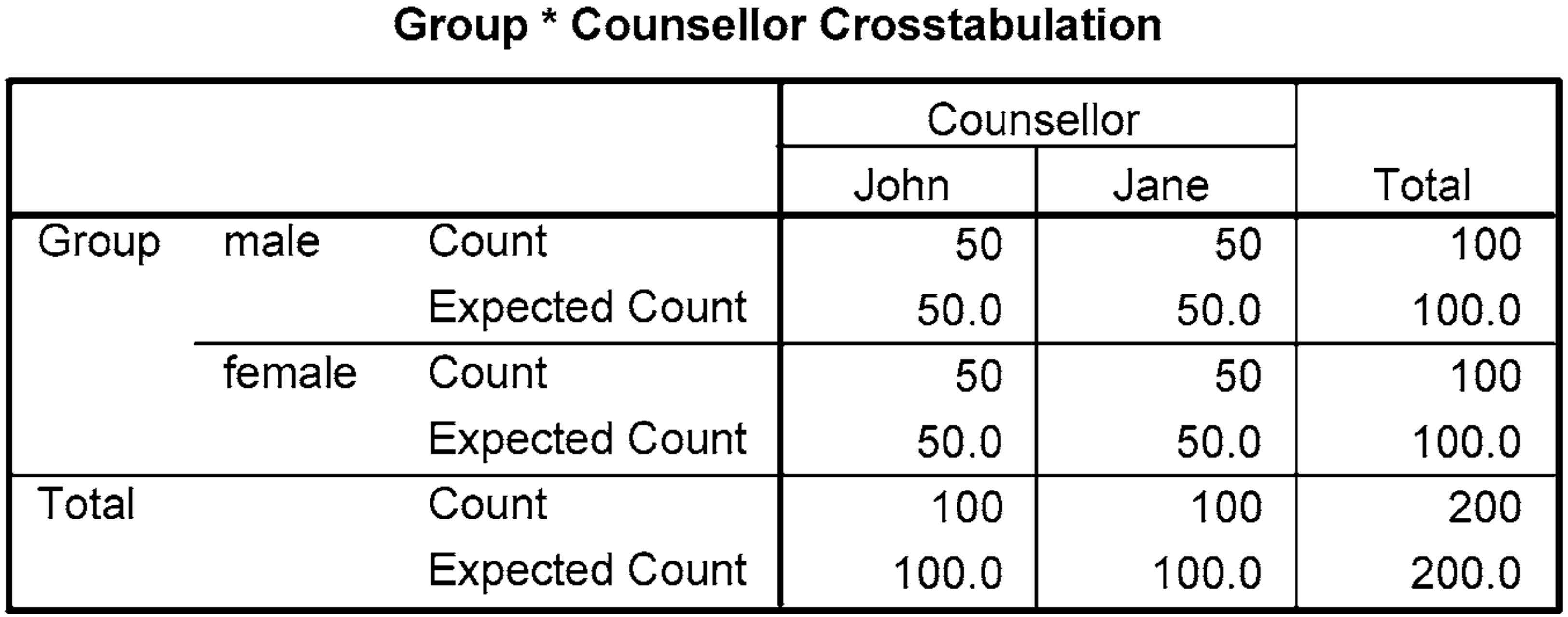
Chi-squared statistical test

Chi-squared applies a statistical test to cross-tabulation by comparing the actual *observed* frequencies in each cell of table with *expected* frequencies. Expected frequencies are those we would expect if data is ‘randomly distributed’.

Two variables are mutually independent, if an observed frequency in each cell is equal to an expected frequency.

Scenario 1

There is no difference across male/female patients and counsellor. The actual observed counts match the expected counts.

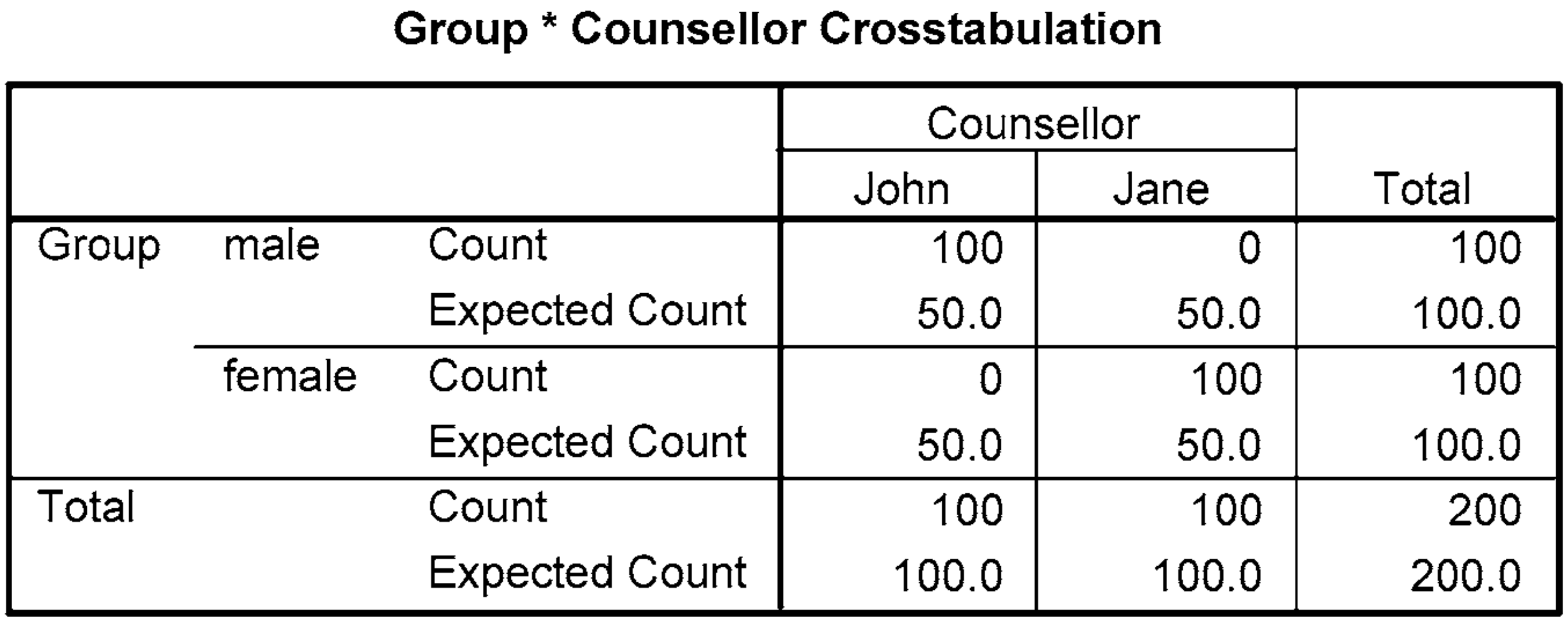


Group * Counsellor Crosstabulation

			Counsellor		Total
			John	Jane	
Group	male	Count	50	50	100
		Expected Count	50.0	50.0	100.0
	female	Count	50	50	100
		Expected Count	50.0	50.0	100.0
Total		Count	100	100	200
		Expected Count	100.0	100.0	200.0

Scenario 2

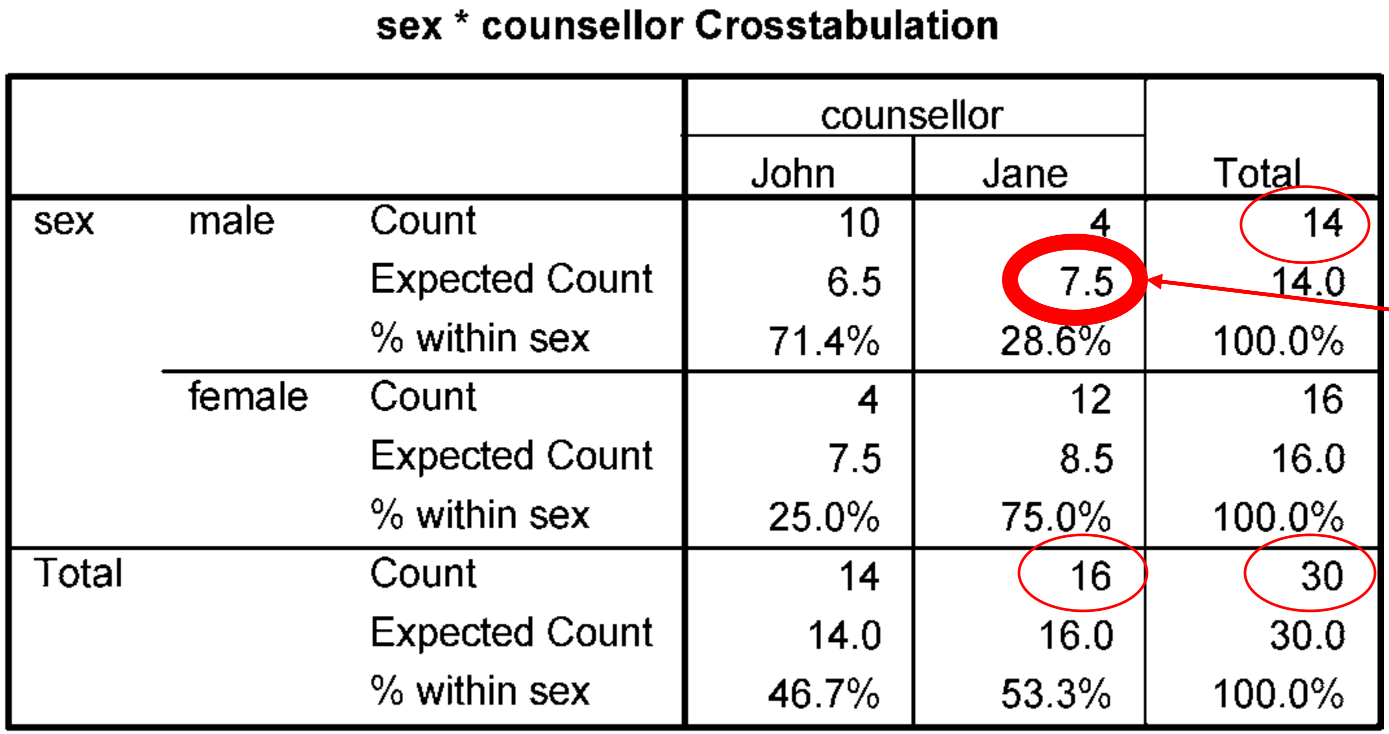
All the males saw John, whereas all the females saw Jane. Notice that the *expected* frequencies remain the same – 50 in each cell. The divergence from expected frequencies would strongly suggest that there is a relationship between sex of the patient and the counsellor they saw.



Group * Counsellor Crosstabulation

			Counsellor		
			John	Jane	Total
Group	male	Count	100	0	100
		Expected Count	50.0	50.0	100.0
	female	Count	0	100	100
		Expected Count	50.0	50.0	100.0
Total		Count	100	100	200
		Expected Count	100.0	100.0	200.0

Scenario 3



sex * counsellor Crosstabulation

			counsellor		Total
			John	Jane	
sex	male	Count	10	4	14
		Expected Count	6.5	7.5	14.0
		% within sex	71.4%	28.6%	100.0%
	female	Count	4	12	16
		Expected Count	7.5	8.5	16.0
		% within sex	25.0%	75.0%	100.0%
Total		Count	14	16	30
		Expected Count	14.0	16.0	30.0
		% within sex	46.7%	53.3%	100.0%

$$7.5 = \frac{14 * 16}{30}$$

$$E_i = \frac{(\text{row total} \times \text{column total})}{\text{Table total}}$$

Calculating Chi-squared

The formula for Chi-Square statistic is defined as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

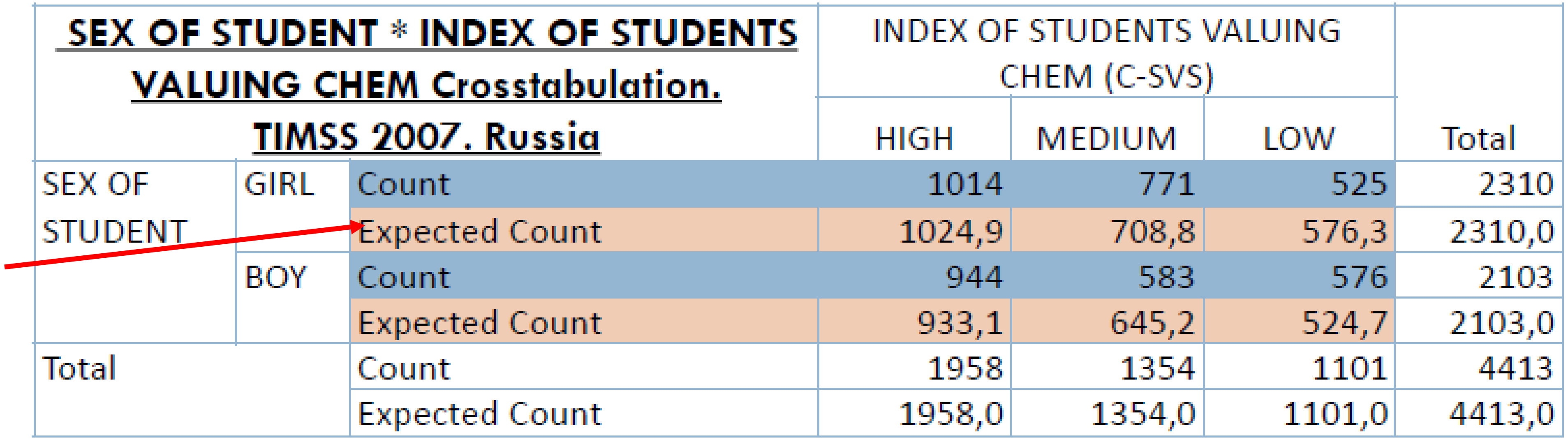
Here, O_i = observed frequency and E_i = expected frequency.

Example

*(2310*1958)/4413 = 1024,9*

*(2310*1354)/4413 = 708,8*

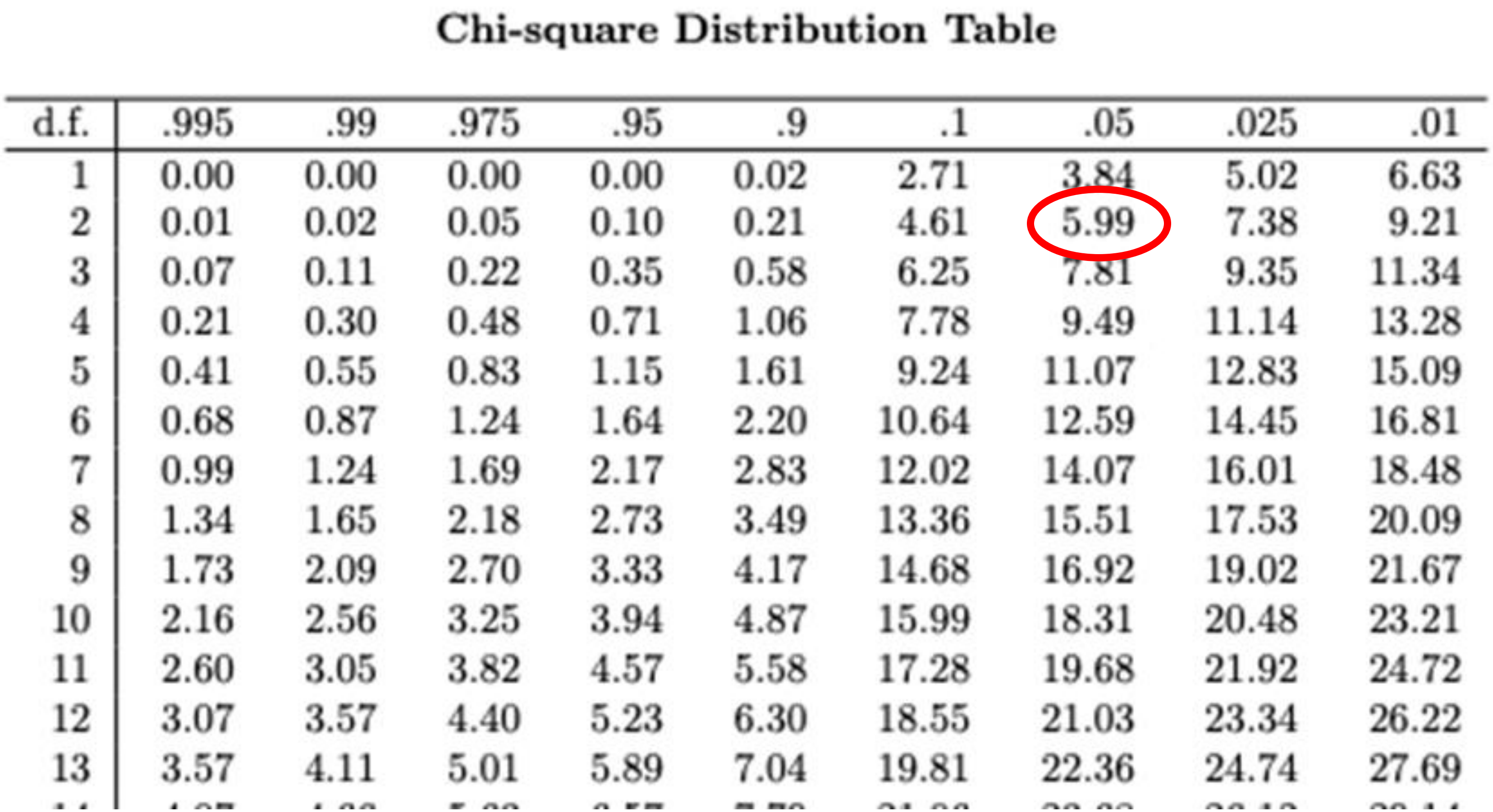
*(2103*1958)/4413 = 933,1*



SEX OF STUDENT * INDEX OF STUDENTS VALUING CHEM Crosstabulation. TIMSS 2007. Russia			INDEX OF STUDENTS VALUING CHEM (C-SVS)			Total
			HIGH	MEDIUM	LOW	
SEX OF STUDENT	GIRL	Count	1014	771	525	2310
		Expected Count	1024,9	708,8	576,3	2310,0
	BOY	Count	944	583	576	2103
		Expected Count	933,1	645,2	524,7	2103,0
Total		Count	1958	1354	1101	4413
		Expected Count	1958,0	1354,0	1101,0	4413,0

$$\chi^2 = \frac{(1014-1024{,}9)^2}{1024{,}9} + \frac{(771-708{,}8)^2}{708{,}8} + \frac{(525-576{,}3)^2}{576{,}3} + \frac{(944-933{,}1)^2}{933{,}1} + \frac{(583-645{,}2)^2}{645{,}2} + \frac{(576-524{,}7)^2}{524{,}7} = 21{,}306$$

df = (rows -1) * (columns – 1) = (2-1)*(3-1) = 2



Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69

Chi-squared calculator: https://www.di-mgt.com.au/chisquare-calculator.html

Conclusions

Ho: There is no relationship between the sex and the grades in chemistry.

H1: There is a relationship between the sex and the grades in chemistry.

In the table of Chi-squared statistical distribution we can find that the critical value for df = (2-1)*(3-1) = 2 and significance level p=0,05 is 5,99. So if there is no relation between the variables the Chi-squared statistics should be not grater than 5,99. In our case 21,3>5,99 so we should reject Ho and accept H1 – **there is a relationship** between variables.

Properties of Chi-squared

- Interval of values from 0 to + ∞.
- The value doesn't permit to know which variable is dependent and which is independent.
- Doesn’t evaluate the force of the relationship. The bigger the sample – the bigger the value of chi-squared.
- The conventional rule of thumb is that if all (more than 95%) of the expected counts in the contingency table are greater than 5, it's acceptable to use the chi-squared test.

Testing the normality of the distribution

- We can look at the histogram, but to receive an objective information weather our distribution is different from “normal” or not we have to run a statistical test.
- The Kolmogorov-Smirnov and Shapiro-Wilks tests allow to test if the distribution deviates from a comparable normal distribution. These tests compare the scores in the sample to a normally distributed set of scores with same mean and standard deviation. Shapiro-Wilks is calculated if the sample size is less than 50.

Normality test

Hypothesis 0: The values are sampled from normally-distributed population.

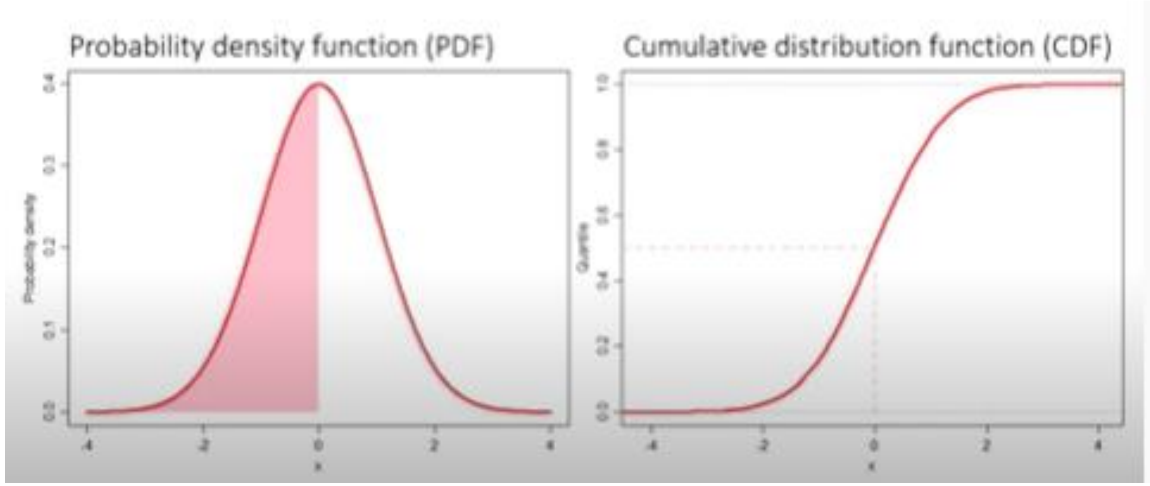
Hypothesis 1: The values are not sampled from normally-distributed population.

If the significance of the test is > 0,05 we reject H1 and conclude that our distribution **is not significantly different from the normal** distribution.

If the significance of the test is <= 0,05 we reject H0, accept H1 and conclude that our distribution is **significantly different from the normal** distribution.

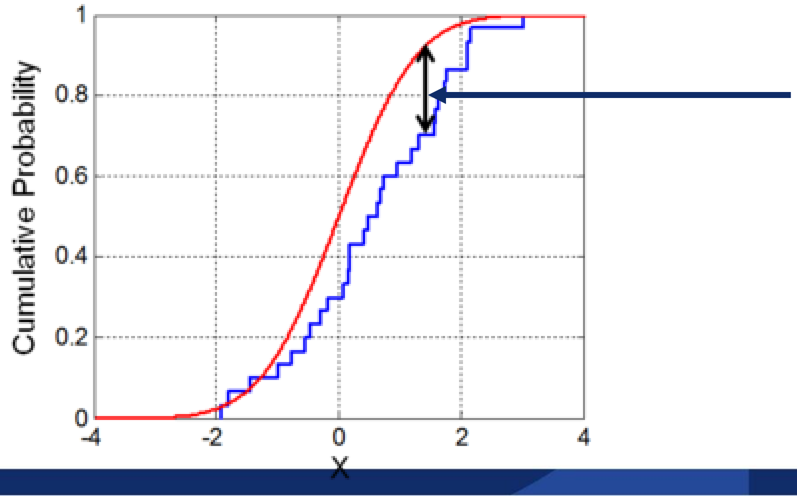
Kolmogorov-Smirnov normality test

The idea behind the Kolmogorov-Smirnov test is that the maximum difference between the assumed cumulative normal distribution function and the sample to be investigated is used to decide whether the random sample belongs to the distribution or not.



Probability density function (PDF)

Cumulative distribution function (CDF)



D statistic

The bigger the D statistic, the higher is the difference between variable’s distribution and normal distribution.

Shapiro-Wilk normality test

The basic idea is to estimate the variance of the sample. The values should approximately equal in the case of a normal distribution and thus should result in a quotient of close to 1.0. If the quotient is significantly lower than 1.0 then the null hypothesis (of having a normal distribution) should be rejected.

Details here:
http://www.statistics4u.info/fundstat_eng/ee_shapiro_wilk_test.html

Quantile-Quantile (Q-Q) plot

Q-Q plot allows to visually inspect similarity between two distributions by plotting the quantiles of both distributions against each other. The extent of the deviation from the straight line indicates the discrepancy between the distributions. The Q-Q plot is often used to check the normality of the distribution of a particular variable.

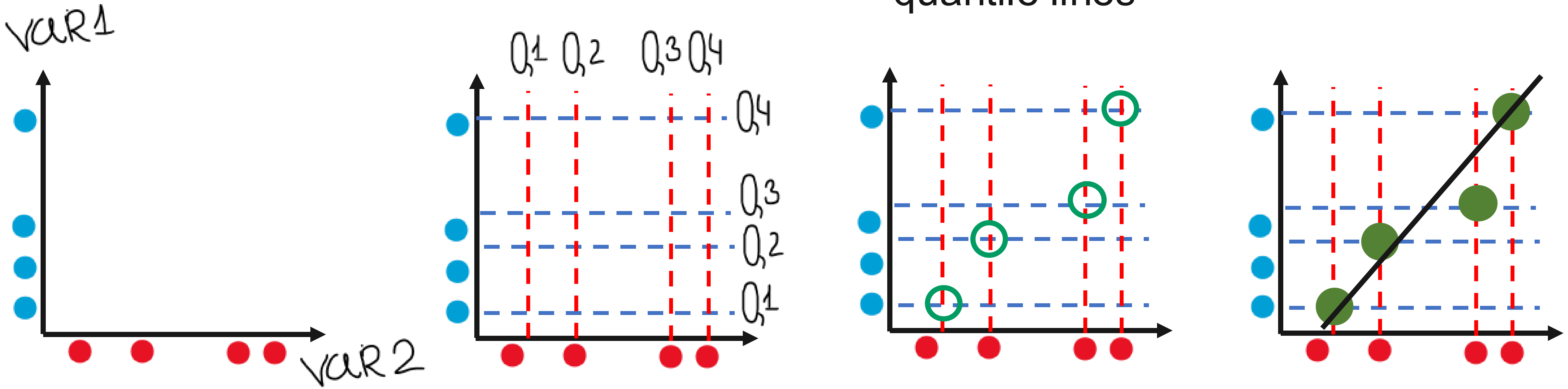
Quantile-Quantile (Q-Q) plot step-by-step

1) Arrange points along axes

2) Draw quantile lines

3) Mark the interception of corresponding quantile lines

4) Add reference line *(i.e., a line connecting the first and the last quantiles or fit a regression line)*



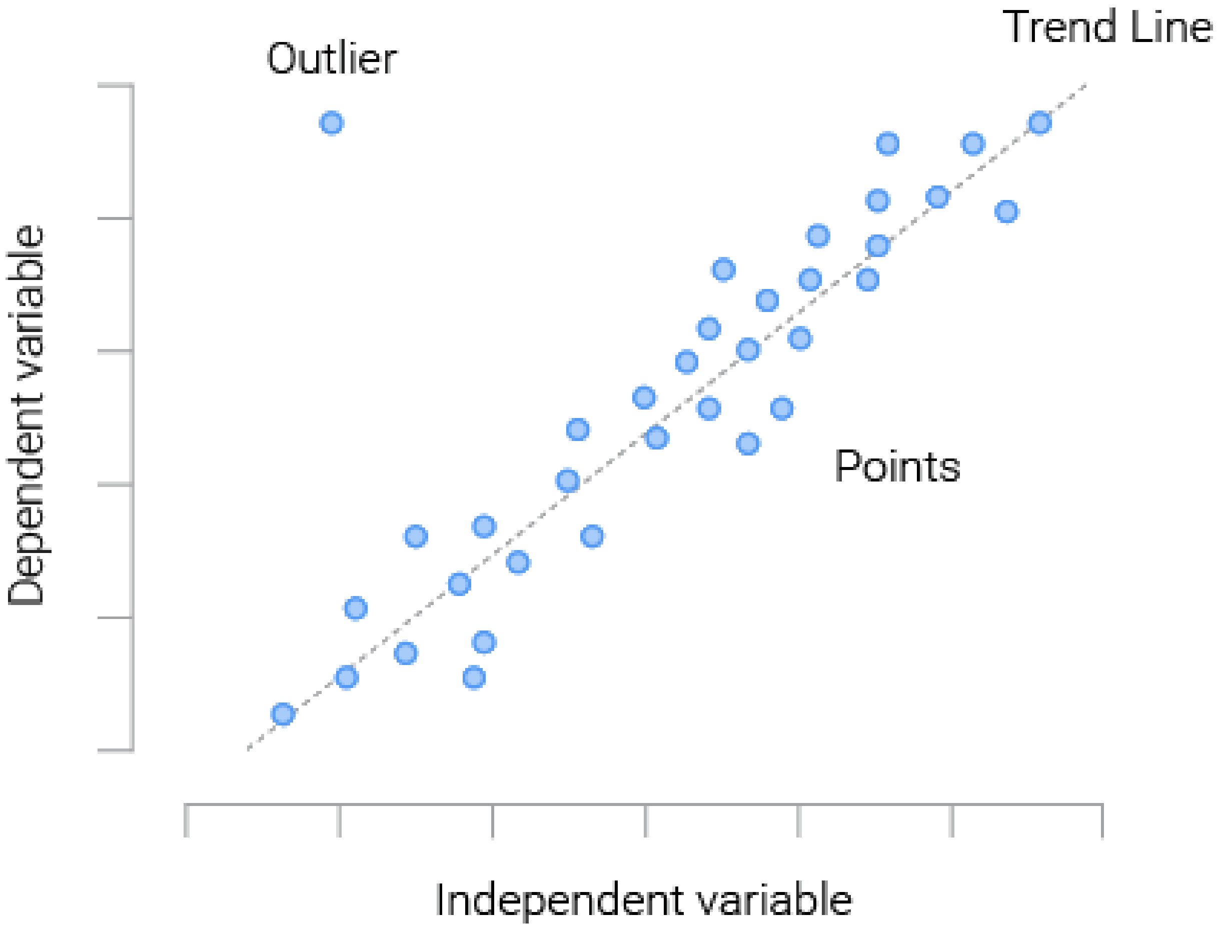
Correlation Analysis: stages

1. Create a scatterplot to investigate the type of relationship between the variables and find out possible outliers;
2. Select appropriate correlation coefficient depending on the specific characteristics of the variables;
3. Calculate correlation coefficient and the corresponding p-value;
4. Interpret the significance, strength and direction of the relationship.

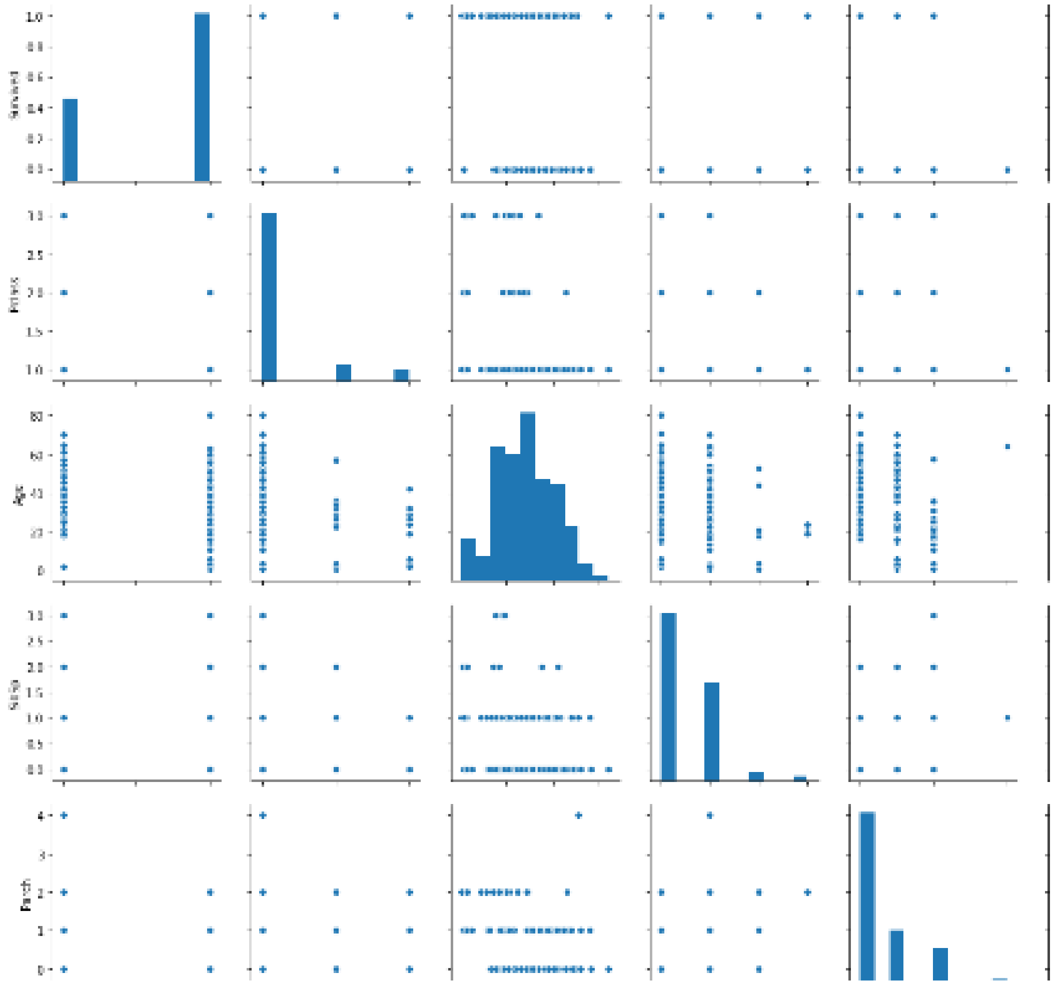
Scatterplot

Before conducting any correlational analysis it’s essential to plot a scatterplot to look at the **general trend** of the data. A scatterplot is simply a graph that plots each objects score on one variable against their score on another and tells whether there seems to be a **relationship between variables**, what kind of relationship it is and whether any cases are markedly different from the others (**outliers**). The outliers could severely bias the correlation coefficient.

Scatterplot



Scatterplot Matrix



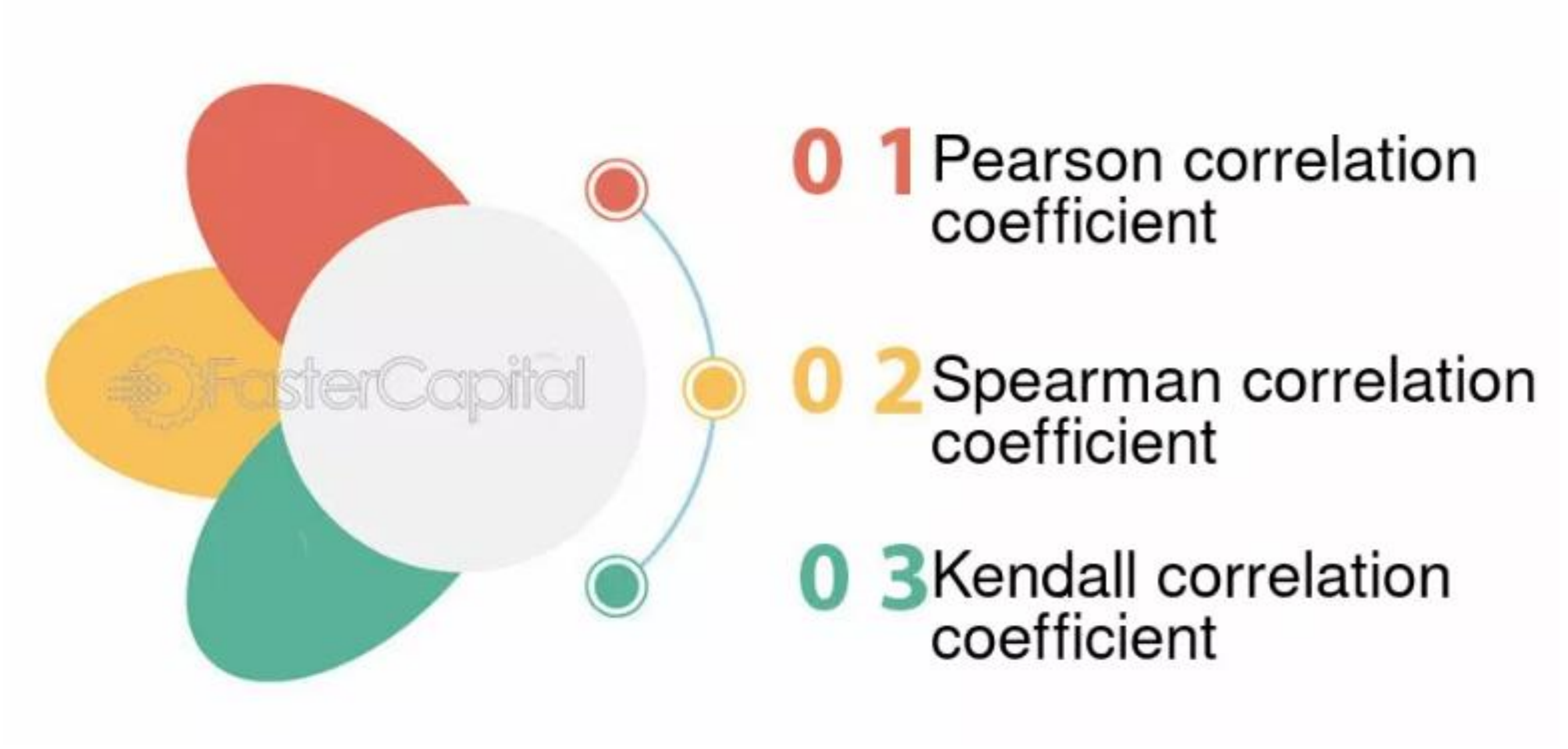
Correlation

Correlation describes the **direction** and **strength** of a relationship between two variables (e.g., height and shoe size). The direction can be **positive** or **negative**.

Positive correlation: an **increase** in values for one variable is associated with an **increase** in values for the other variable, for example, as height increases so does shoe size.

Negative correlation: an **increase** in values for one variable is associated with a **decrease** in values on another variable, for example, as temperature *reduces* the use of electricity for heating *increases*.

Correlation coefficients



01 Pearson correlation coefficient

02 Spearman correlation coefficient

03 Kendall correlation coefficient

Pearson correlation coefficient

Evaluates the force of linear relation between two variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}\left[\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)\right]}{\left(n-1\right)\cdot\sigma_x \cdot\sigma_y}$$

- ➢ The value of the coefficient lies between -1 and +1.
- ➢ Could be calculated only for interval normally distributed variables
- ➢ Measures the direction (positive vs. negative) and strength (value) of the relationship.

Interpretation of Pearson correlation coefficient's values

A coefficient of +1 indicates that the two variables are perfectly positively correlated, so as one variable increases the other increases by a proportionate amount.

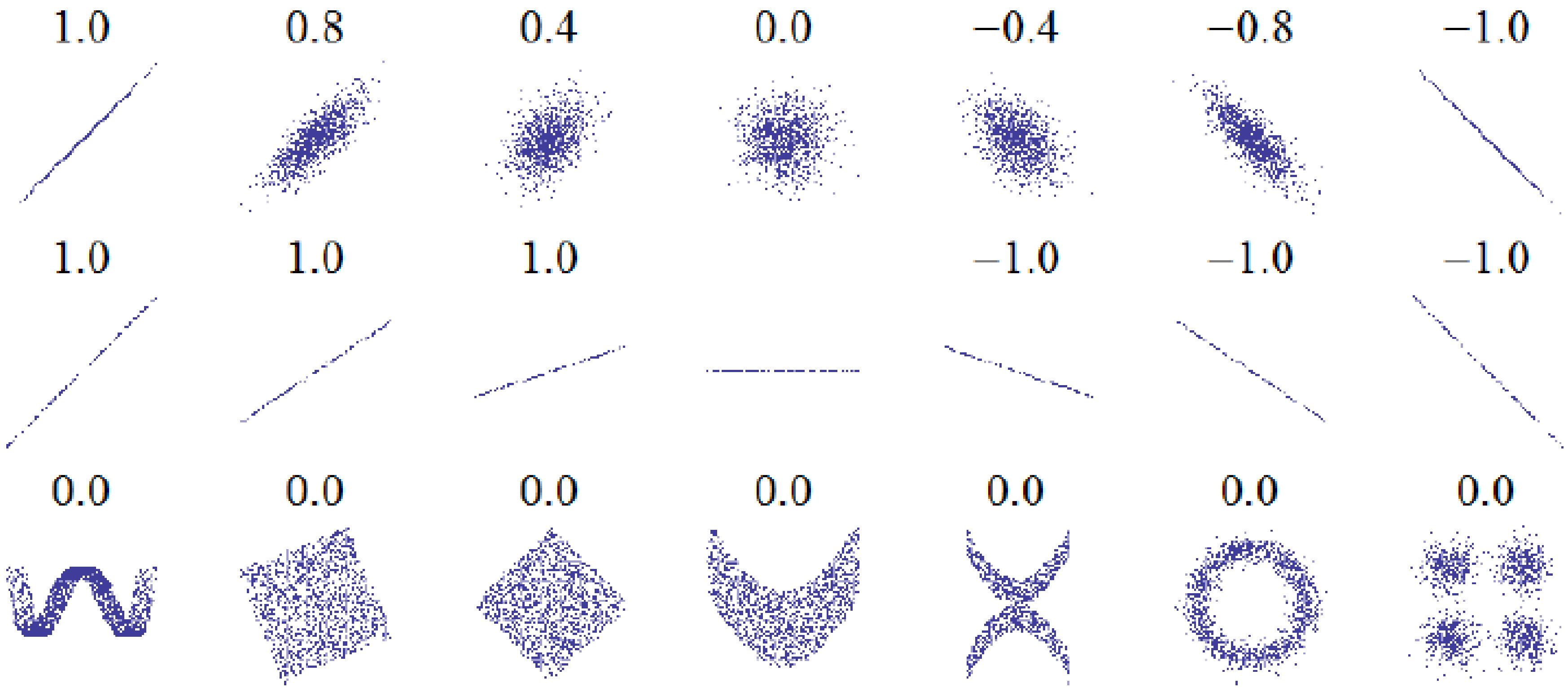
Conversely, a coefficient of -1 indicates a perfect negative relationship: if one variable increases the other decreases by a proportionate amount.

A coefficient of 0 indicates no linear relationship at all.

Interpretation of Pearson correlation coefficient’s values

Value of Pearson’s correlation coefficient	Interpretation
0 < r <= 0,2	Very weak correlation
0,2 < r <= 0,5	Weak correlation
0,5 < r <= 0,7	Medium correlation
0,7 < r <= 0,9	Strong correlation
0,9 < r <= 1	Very strong correlation

Scatterplot for different values of r



Wikipedia

Significance of Pearson correlation coefficient

1. H_0: r = 0, H_1: r != 0
2. Calculate t-criterion

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

3. According to the table of Student's distribution determine the critical t-value for a given level of significance and degrees of freedom.
4. If the actual value (calculated in step 2) exceeds the critical value (calculated in step 3), then the hypothesis H_0 is rejected and H_1 is accepted.
5. Python calculates p-value which should be compared with the significance level selected by a researcher (0,05 or 0,01).

Interpretation: causality

The correlation coefficients give no indication of the direction of *causality*. So, for example, even if we can conclude that exam performance goes down as anxiety about that exam goes up, we cannot say the high exam anxiety *causes* bad exam performance. This caution is for two reasons:

- the third-variable problem;
- direction of causality.

Using R2 for interpretation

R^2 – coefficient of determination is calculated as squared correlation coefficient. It’s a measure of the amount of variability in one variable that is explained by the other.

Rank correlation coefficients

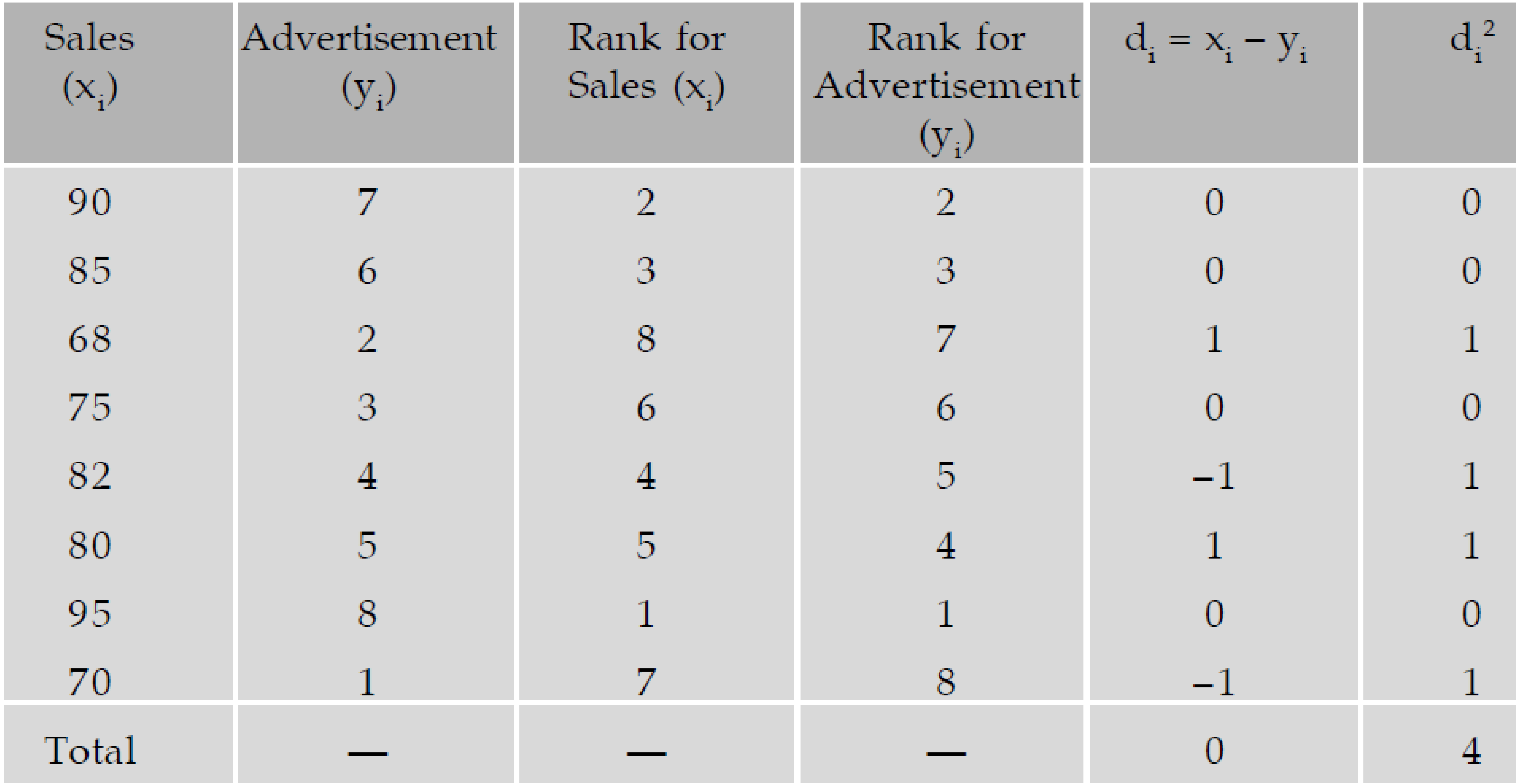
- Are used to evaluate the relationship between ordinal variables or internal variables that are not normally distributed.
- To calculate a coefficient the ranks of the values are used.
- There are two most popular rank correlation coefficients: Spearmen’s and Kendall’s.
- Kendall’s coefficient should be selected if we have many tied ranks. If most values are unique, we select Spearmen’s coefficient.

Spearman’s rho correlation coefficient

Assesses the monotonic relationship between variables. Unlike the Pearson coefficient, the Spearman coefficient does not assume a linear relationship. Instead, it measures the strength and direction of the monotonic association between variables. Could be calculated for non-normally distributed variables and ordinal variables. First the scores should be ranked. Is applicable if we have small number of tied ranks.

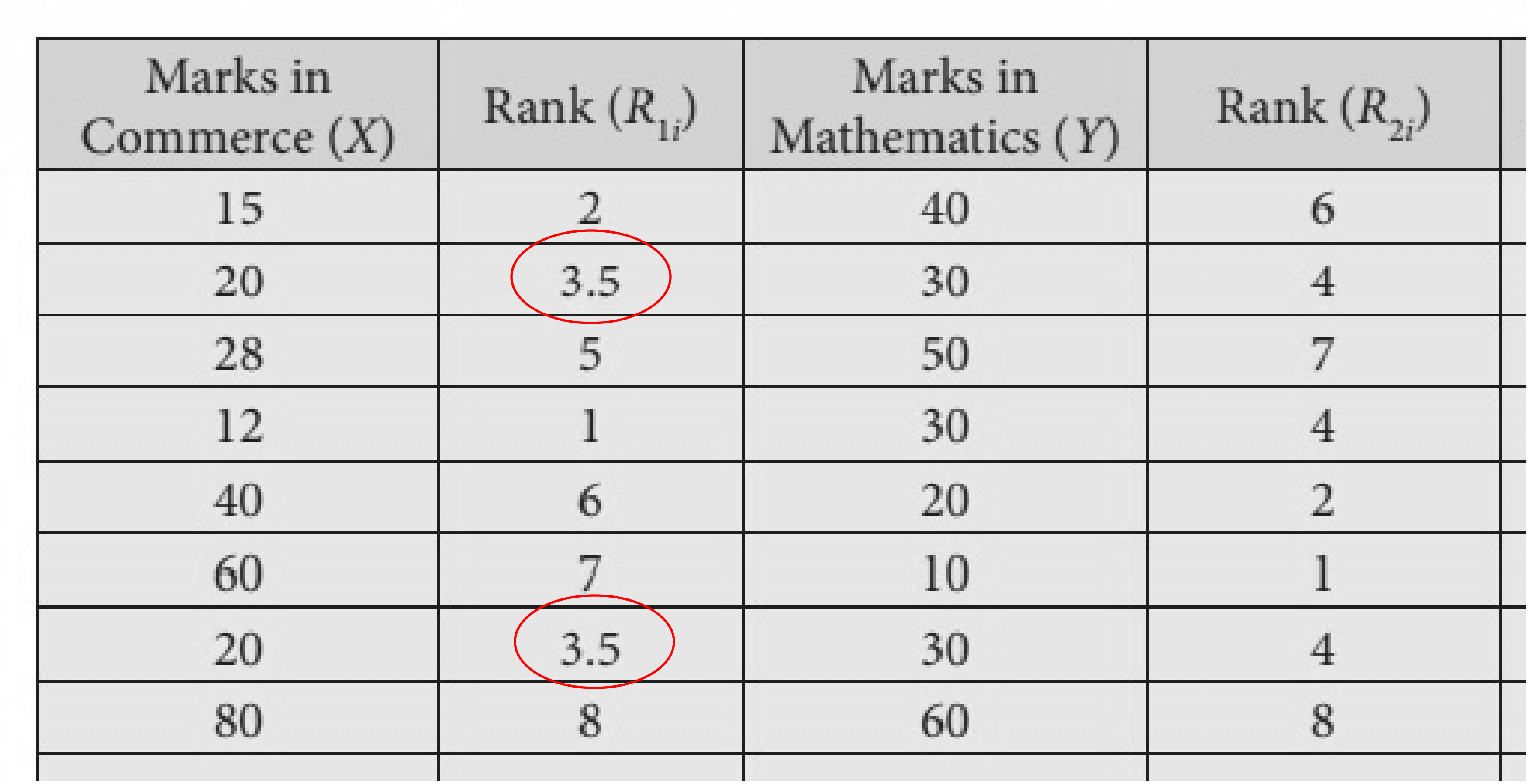
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Spearman’s correlation coefficient



Sales (x_i)	Advertisement (y_i)	Rank for Sales (x_i)	Rank for Advertisement (y_i)	$d_i = x_i - y_i$	d_i^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	–1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	–1	1
Total	—	—	—	0	4

Tied ranks



Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})
15	2	40	6
20	3.5	30	4
28	5	50	7
12	1	30	4
40	6	20	2
60	7	10	1
20	3.5	30	4
80	8	60	8

Kendall’s tau correlation coefficient

Kendall’s tau is another non-parametric correlation, and it should be used when you have a small data set with many tied ranks. This means that if you rank all the scores and many scores have the same rank, Kendall’s tau should be used. It represents the degree of concordance between two columns of ranked data. Like the Spearman coefficient, Kendall's tau does not assume a linear relationship.

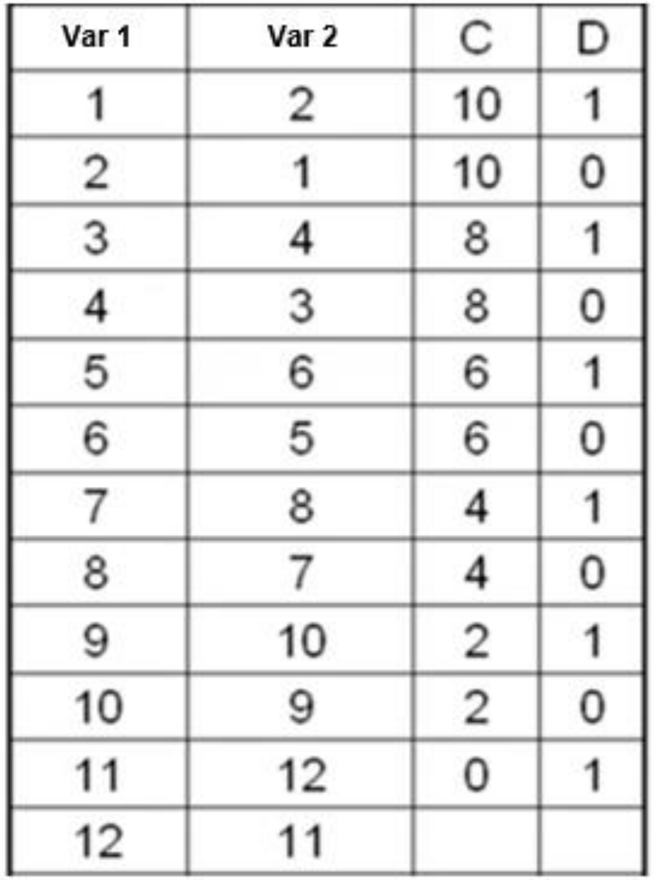
Kendall’s tau

$$Kendall's \cdot tau = \frac{C-D}{C+D}$$

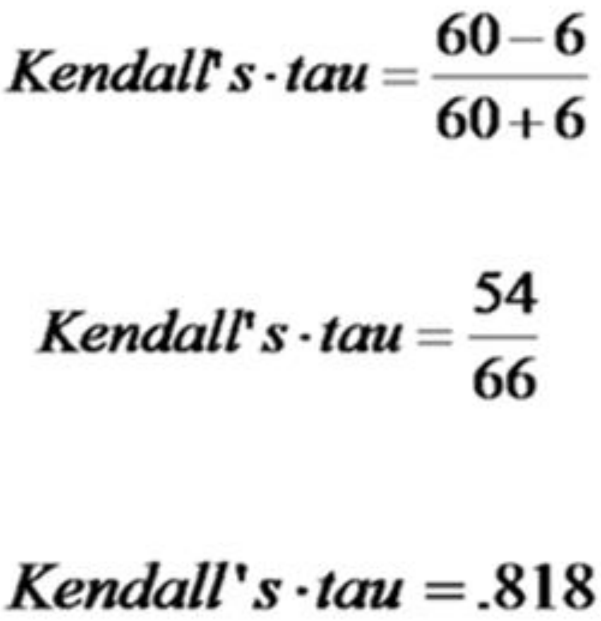
C – number of concordant pairs

D – number of discordant pairs

Kendall’s tau



Var 1	Var 2	C	D
1	2	10	1
2	1	10	0
3	4	8	1
4	3	8	0
5	6	6	1
6	5	6	0
7	8	4	1
8	7	4	0
9	10	2	1
10	9	2	0
11	12	0	1
12	11		

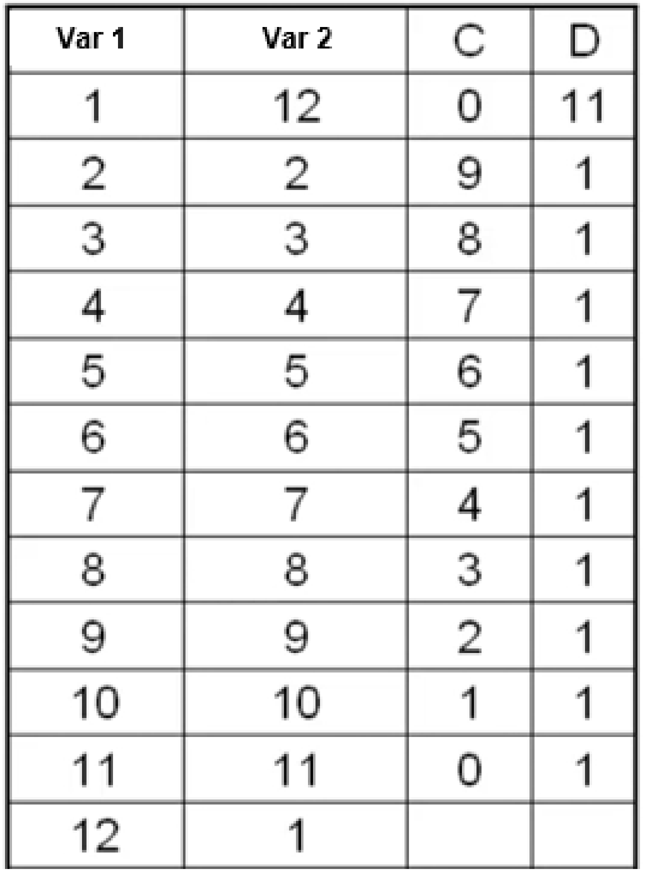


$$Kendall's \cdot tau = \frac{60-6}{60+6}$$

$$Kendall's \cdot tau = \frac{54}{66}$$

$$Kendall's \cdot tau = .818$$

Kendall’s tau



Var 1	Var 2	C	D
1	12	0	11
2	2	9	1
3	3	8	1
4	4	7	1
5	5	6	1
6	6	5	1
7	7	4	1
8	8	3	1
9	9	2	1
10	10	1	1
11	11	0	1
12	1		

$$Kendall's \cdot tau = \frac{45-21}{45+21}$$

$$Kendall's \cdot tau = \frac{24}{66}$$

$$Kendall's \cdot tau = .364$$

Comparing correlation coefficients

Correlation coefficient	Variables
Pearson’s correlation coefficient	Both variables are interval and normally distributed
Spearman's rank correlation coefficient	Both variables are interval, most of their values are unique, but they are not normally distributed
Kendall’s rank correlation coefficient	Variables are interval or ordinal, values are not unique (there are many tied ranks)

Partial correlation

Partial correlation looks at the relationship between two variables while “controlling” the effect of one or more additional variables.

Partial correlations are used to find out the size of the unique portion of variance.

H0: there is no relationship between two variables after controlling for effects of confounding variable

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Example of a partial correlation 1

Paul Broca in 1873 have found a strong relationship between gender and brain size: women’s brains are, on the whole, smaller than men’s. This was at the time used as evidence to argue that women were intellectually inferior to men. The obvious problem is that this relationship takes no account of body size: people with bigger bodies have bigger brains irrespective of intellectual ability.

Stephen Jay Gould in 1981 famously reanalyzed Broca’s data and demonstrated that the strong relationship between gender and brain size disappeared when you accounted for body size.

Example of a partial correlation 2

🧠 IQ and Success (Controlling for Rich Parents)

People say smart folks become more successful.

But what if **rich parents** are secretly helping both IQ *and* success?

Partial correlation says:
➤ Being a genius helps, but **dad owning 3 companies helps more.** 💼💸

Example of a partial correlation 3

🏀👟 **Example: Shoe Size & Sports Success (Controlling for Height)**

Observation: Bigger shoe size seems linked to greater success in sports.

But: Taller people tend to have both **bigger feet** *and* do better in sports.

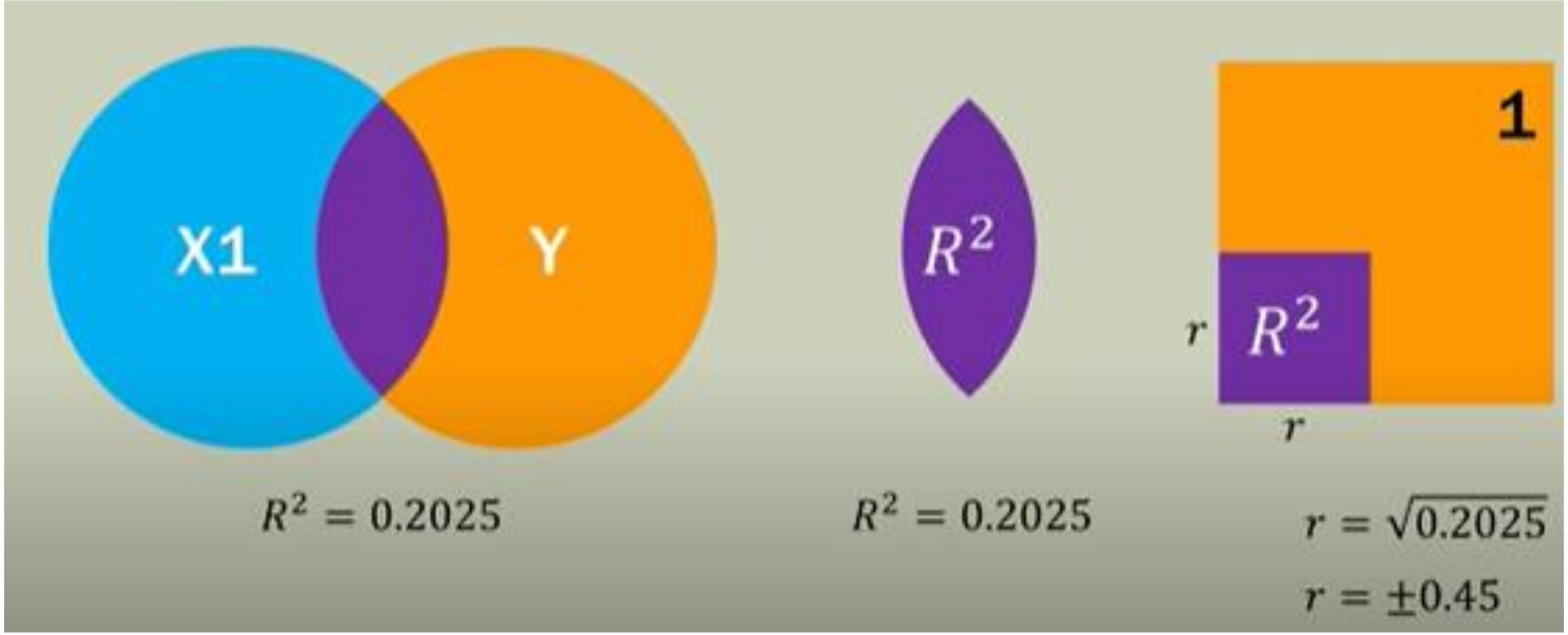
Hidden variable: Height is influencing both shoe size and athletic performance.

Partial correlation result:
➤ After controlling for **height**, the link between **shoe size** and **sports success** mostly disappears.

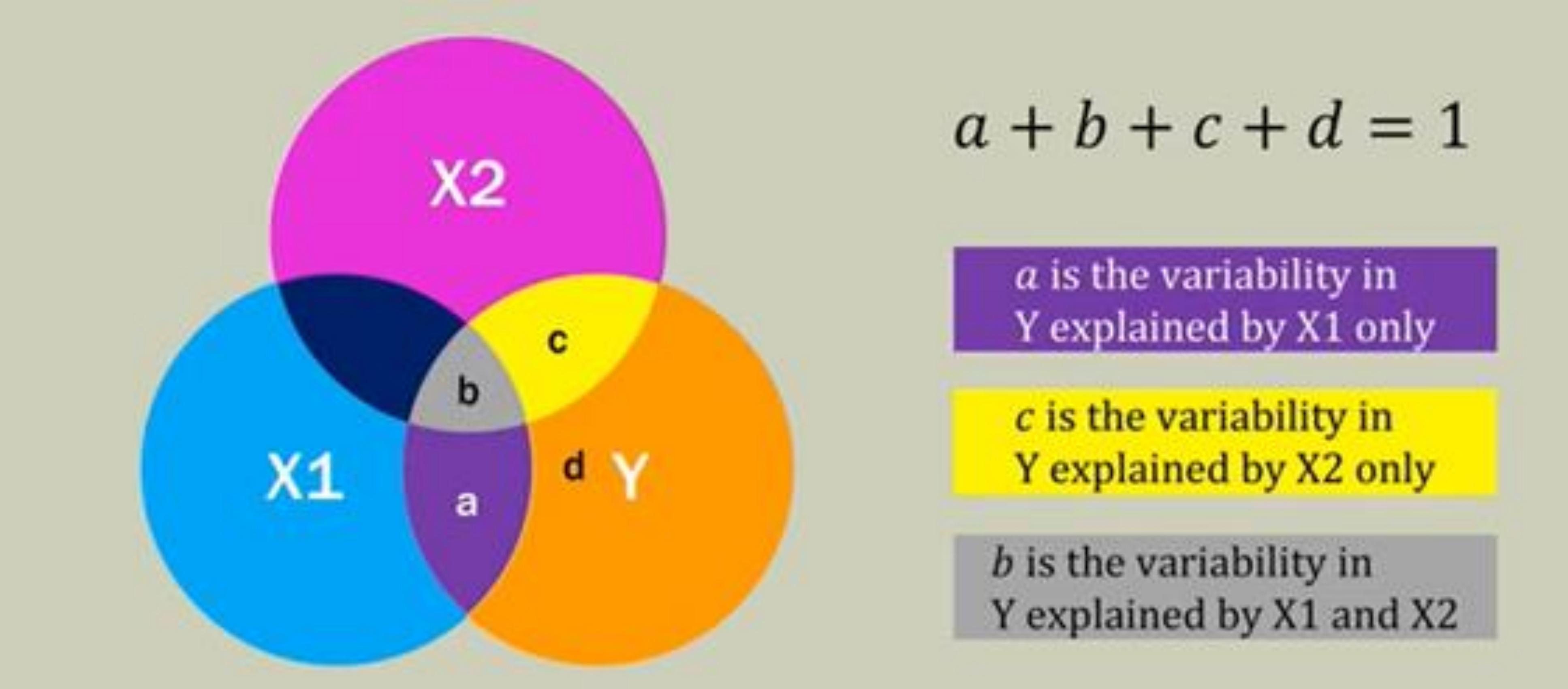
Conclusion:
➤ It's not big feet that make you good at sports — it's the **tall body attached to them**.

Partial correlation

R^2 is an area of overlap. The area of each circle is 1.



Partial correlation



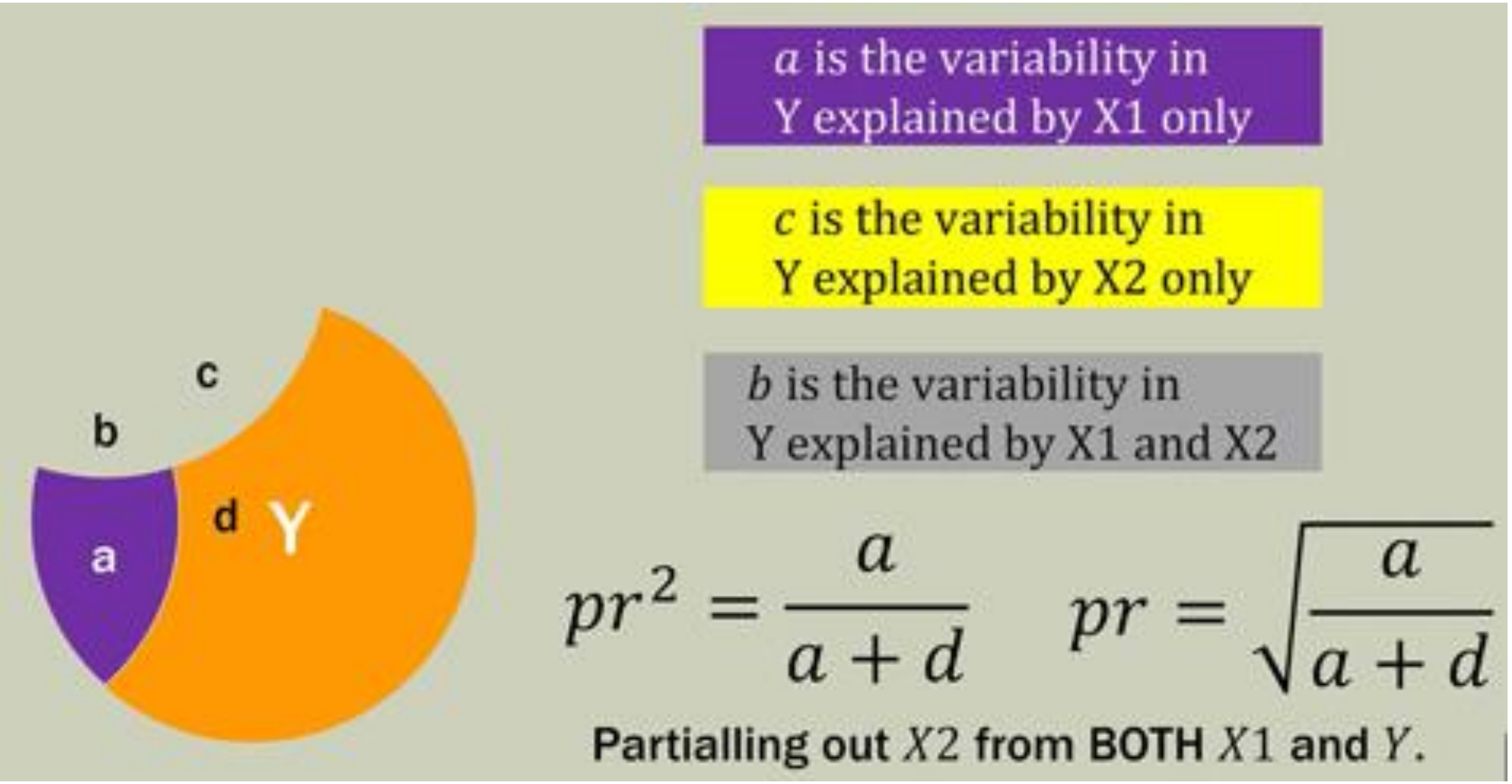
$$a + b + c + d = 1$$

a is the variability in Y explained by X1 only

c is the variability in Y explained by X2 only

b is the variability in Y explained by X1 and X2

Partial correlation



a is the variability in Y explained by X1 only

c is the variability in Y explained by X2 only

b is the variability in Y explained by X1 and X2

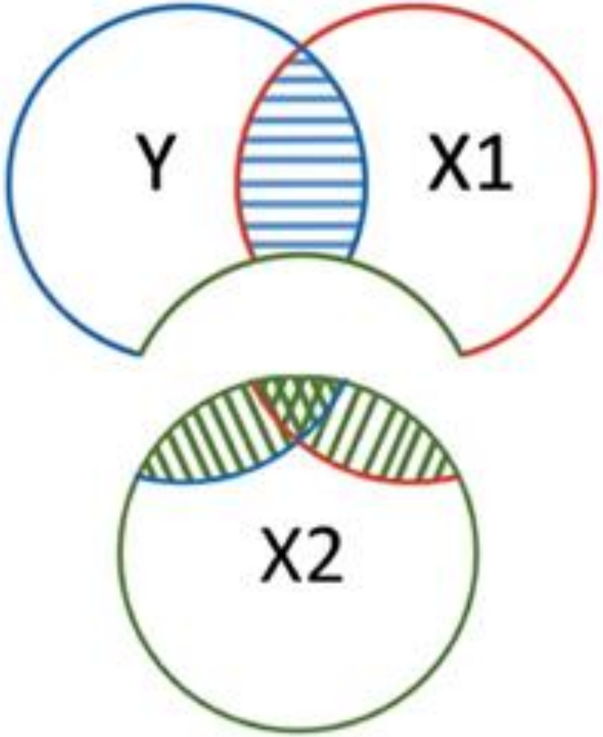
$$pr^2 = \frac{a}{a+d} \qquad pr = \sqrt{\frac{a}{a+d}}$$

Partialling out $X2$ from BOTH $X1$ and Y.

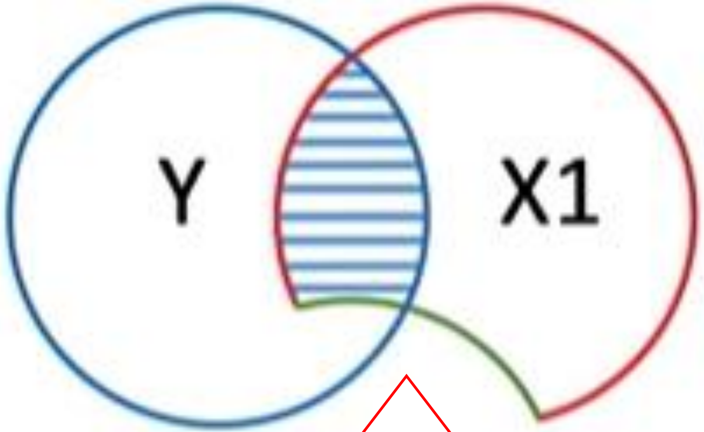
Source: https://www.youtube.com/watch?v=UyyWsctkXaw

Part (semi-partial) and partial correlation

Partial Correlation between Y and X1 controlling for X2



Part Correlation between Y and X1 controlling for X2



The effect of X2 on Y has not been removed

Semi-partial (or part) correlation

When we do a *partial* correlation, we control for the effect that the third variable has on *both* variables in the correlation. In a *semi-partial* correlation, we control for the effect that the third variable has on only one of the variables in the correlation.

Semi-partial (or part) correlation

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^{2}}} \; and \; r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^{2}}}$$

Correlation between 1 and 2 controlling the effect of 3 on 2.

Correlation between 1 and 2 controlling the effect of 3 on 1.

Other (less common) correlation coefficients

Correlation coefficient	Variable 1	Variable 2
Point-biserial Correlation	Binary	Metric
Phi (φ) Correlation	Binary	Binary

Other correlation coefficients

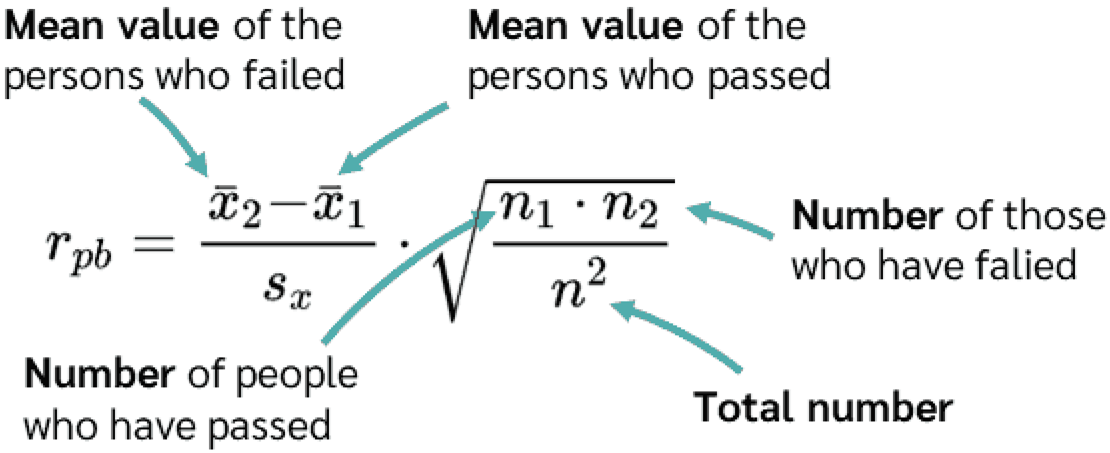
1. Point-biserial Correlation – is a special case of Pearson correlation coefficient that allows to estimate the relationship between a dichotomous and a metric variable

$$r = \frac{Mean_{group\ 1} - Mean_{group\ 2}}{SD_{sample}} \sqrt{\frac{n_{group\ 1}\, n_{group\ 2}}{n^2}}$$

Mean – mean of the metric variable in a group
SD – Standard deviation
n – number of observations

Other correlation coefficients

⌛	🎭	
2	failed	0
3	passed	1
16	failed	0
17	passed	1
5	passed	1
6	passed	1
14	failed	0
7	passed	1



Mean value of the persons who failed

Mean value of the persons who passed

$$r_{pb} = \frac{\bar{x}_2 - \bar{x}_1}{s_x} \cdot \sqrt{\frac{n_1 \cdot n_2}{n^2}}$$

Number of those who have falied

Number of people who have passed

Total number

$$r = \frac{10.6 - 7.6}{5.6} \sqrt{\frac{5 * 3}{8^2}} = 0.25$$

https://datatab.net/tutorial/point-biserial-correlation

Other correlation coefficients

2. Phi (φ) correlation – is a special case of Pearson correlation coefficient that allows to estimate the relationship between two binary variables

	Y = 0	Y = 1	Total
X = 0	n_{00}	n_{01}	n_{0*}
X = 1	n_{01}	n_{11}	n_{1*}
Total	n_{*0}	n_{*1}	n

$$\varphi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1*}n_{*1}n_{0*}n_{*0}}}$$

Summary

Test Name	Purpose	Data Type	Groups Compared	Key Assumptions	When to Use
Chi-Square Test	Test for association between categorical variables	Categorical (nominal/ordinal)	2 categorical variables	Expected frequency ≥ 5 in each cell	Comparing observed vs expected frequencies (e.g., gender vs voting preference)
Pearson Correlation	Measure linear relationship between variables	Continuous	2 variables	Normality, linearity	Examining correlation between height and weight
Spearman Correlation	Measure monotonic relationship	Ordinal / Continuous	2 variables	Many tied ranks might affect the accuracy of the correlation coefficient.	Non-parametric alternative to Pearson correlation
Kendall Correlation	Measure ordinal association	Ordinal	2 variables	Presence of many ties can reduce the power and affect the exact calculation. Kendall's tau-b is specifically designed to handle ties better than tau-a.	Small datasets or when many tied ranks exist
Point-Biserial Correlation	Measure relationship between continuous and binary variable	Continuous + Binary	1 continuous, 1 binary	Binary variable must be truly dichotomous	Correlation between test scores and pass/fail outcome
Phi Correlation Coefficient	Measure strength of association between two binary variables	Binary (dichotomous)	2 binary variables	Variables must be truly binary	Correlation between two yes/no variables.

Thank you for your attention!