Data Analysis: Relationships Between Variables Part 1

TU Dublin Tallaght Campus

Department of Computing

Relationships

- The relationships we refer to here are relationships **between two variables**.
- We say that a relationship exists if we can say something about unknown values of one variable from the values of another.
- Relationships can be examined using **visualisation** or expressed through mathematically derived measures called **statistics**.
- The existence of a relationship between two variables **does not imply that there is any causation** involved.

- To measure a relationship, we must have values pertaining to the **same instances** for both variables, for example, instances i_1 to i_n for variables x_1 and x_p in the picture.

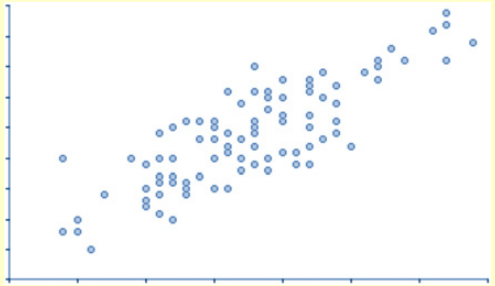
	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\cdots$	$\boldsymbol{x}_{\mathsf{p}}$
i_1	x_{11}	x_{21}	$\cdots$	$x_{\mathsf{p}1}$
i_2	x_{12}	x_{22}	$\cdots$	$x_{\mathsf{p}2}$
.	.	.		.
.	.	.		.
.	.	.		.
i_{n}	$x_{1\mathsf{n}}$	$x_{2\mathsf{n}}$	$\cdots$	$x_{\mathsf{p}\mathsf{n}}$

Relationship visualisation

Relationships can be discovered by examining suitable visual representations of data

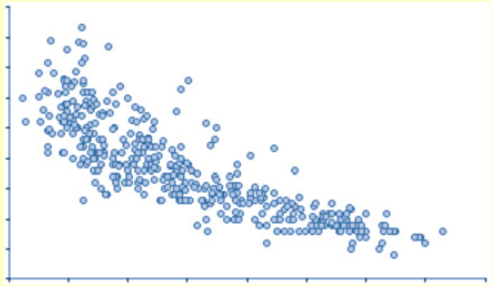
- **Scatterplots (two numeric variables)**

Positive relationship



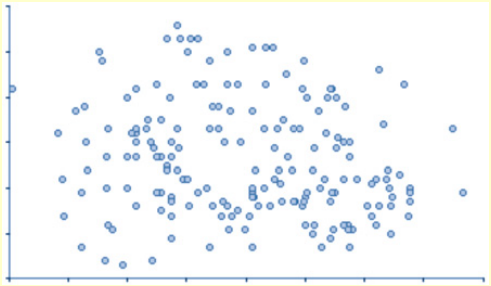
Source: [MSD]

Negative relationship



Source: [MSD]

No relationship



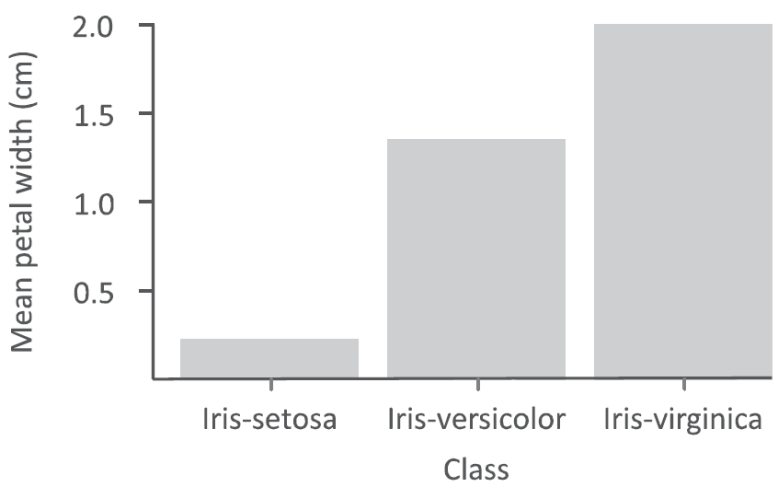
Source: [MSD]

- **Summary tables and graphs (one numeric and one categorical variable)**
  To show data in a summary table or graph, the data is grouped based on the values of a discrete variable or ranges of a continuous variable. The table or graph summarises properties of a numeric variable for each group.
  - Summary table
    * Each data subset is shown in a row of the table
    * One of the columns usually contains observation counts for the subsets
    * The remaining columns represent properties, such as mean or standard deviation, of a numeric variable

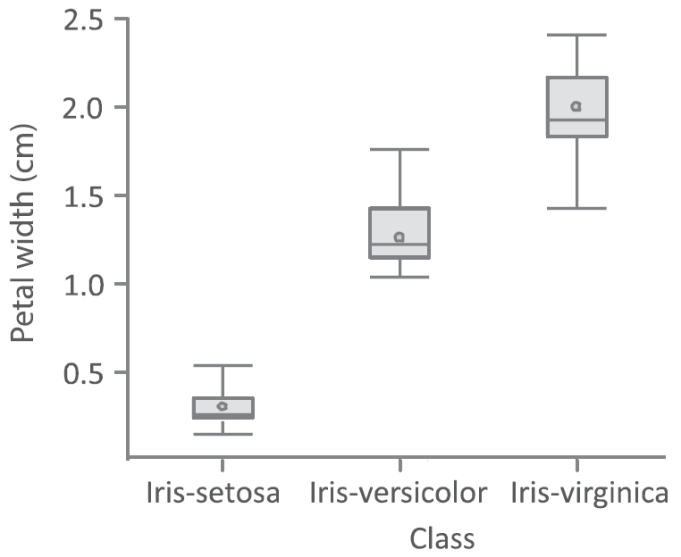
		Petal.Length				
Class	Count	Max (cm)	Min (cm)	Mean (cm)	Median (cm)	SD (cm)
setosa	50	1.9	1.0	1.46	1.5	0.17
versicolor	50	5.1	3.0	4.26	4.35	0.47
virginica	50	6.9	4.5	5.55	5.55	0.55

*The data in the table above is grouped by type of iris. The columns represent properties of the variable **Petal.Length**. The table is immediately informative, showing the marked difference between the petal width of the three classes of iris.*

- Summary graph
  - A graphical representation of the data shown in a summary table
  - The grouping variable from the summary table is on the x-axis
  - The y-axis represents a property of the other variable



Source: [MSD]



Source: [MSD]

These two summary diagrams are of the mean petal width and of petal width box and whisker visualisations for the different iris classes. Comparing the classes is even easier than with the summary table.

- **Contingency table (two categorical variables)**
  - Provides insight into the relationship between two categorical variables (or non-categorical variables converted to categorical).
  - Contains observation counts for each possible pair of values for the two variables
  - Also called *cross-classification* table

		Gender		
		Male	Female	Totals
Age-group	10–19	847	810	1657
	20–29	4878	3176	8054
	30–39	6037	2576	8613
	40–49	5014	2161	7175
	50–59	3191	1227	4418
	60–69	1403	612	2015
	70–79	337	171	508
	80–89	54	24	78
	90–99	29	14	43
	Total	21,790	10,771	32,561

Source: [MSD]

An example of a contingency table, showing counts for pairs of values for variables Gender and Age-group

Measuring relatedness (statistics)

Many different measures are used for quantifying the degree of relatedness between variables.

- **Pearson's product-moment correlation coefficient**
  - Most common correlation coefficient, often referred to simply as 'correlation coefficient'
  - Defined by Karl Pearson (1857-1936), English mathematician and statistician
  - Measures linear correlation between two numeric variables
  - Is always a number between -1.0 and 1.0, inclusive, expressing relatedness on a scale from *perfect negative correlation* (-1.0) to *perfect positive correlation* (1.0). A value of 0 means *no correlation*.

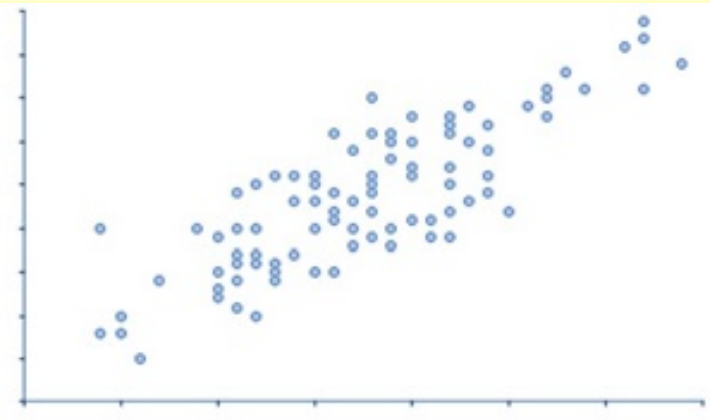
- The formula for calculating the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

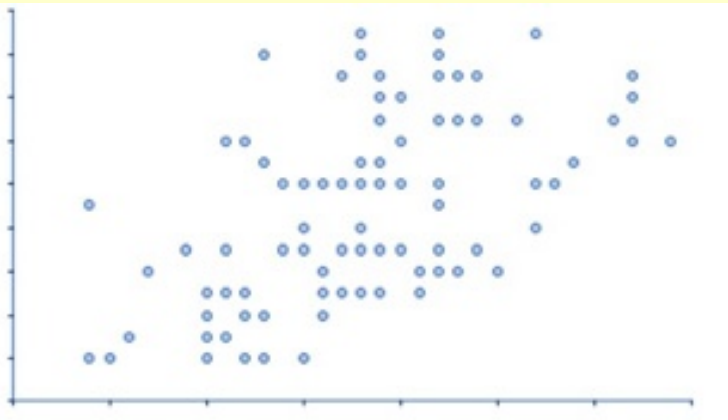
where x and y are variables, x_i are the individual values of x, y_i are the individual values of y, $\bar{x}$ is the mean of the x variable, $\bar{y}$ is the mean of the y variable, s_x and s_y are the standard deviations of variables x and y, respectively, and n is the number of observations.

- Examples of two scatterplots with different values of correlation coefficient:

$r = 0.83$



$r = 0.59$



Source: [MSD]

- Testing the significance of Pearson's product-moment correlation coefficient
  * For a pair of perfectly independent variables, the coefficient that is calculated on a pair of samples has a distribution with mean equal to 0 (see here).
  * The smaller the sample size, the greater the correlation value needs to be in order to be considered significant (Wikipedia picture illustrating this).
  * The critical values are always given in a table:

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
4	.950	.990	7	.754	.874	10	.632	.765	13	.553	.684
5	.878	.959	8	.707	.834	11	.602	.735	14	.532	.661
6	.811	.917	9	.666	.798	12	.576	.708	15	.514	.641

Source: [US]

A larger table can be found here. The degrees of freedom value is calculated as $df = N - 2$.

- **Kendall Tau**
  - Measures association betwen a pair of numeric variables for which values can be ordered i.e. ranked (numeric or ordinal)
  - Defined by English statistician Maurice Kendall (1907-1983)
  - Also called *Kendall rank correlation coefficient*
  - This is a non-parametric test of correlation, meaning that it does not use distribution parameters
  - Calculated using value *ranks* rather than values, as follows:
    1. Let's say that the variables are X and Y, with values $x_1, x_2, ...x_n$ and $y_1, y_2, ...y_n$, respectively, and pairs $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ corresponding to observations.
    2. Each value is given a *rank* within the context of its variable, based on numeric value (the highest value receives rank 1 etc.). This means that the values for X will have unique rank values between 1 and n, as will the values for Y.
    3. The pairs are tested with other pairs for *concordance* vs. *discordance*. The following table shows tested conditions and corresponding rank relationship designations.

Fulfilled condition	**Concordance value for (x_i, y_i) and (x_j, y_j)**
$r(x_i) - r(x_j) > 0$ AND $r(y_i) - r(y_j) > 0$	concordant
$r(x_i) - r(x_j) < 0$ AND $r(y_i) - r(y_j) < 0$	concordant
$r(x_i) - r(x_j) > 0$ AND $r(y_i) - r(y_j) < 0$	discordant
$r(x_i) - r(x_j) < 0$ AND $r(y_i) - r(y_j) > 0$	discordant
$r(x_i) = r(x_j)$	x-tied
$r(y_i) = r(y_j)$	y-tied

NOTE: $r(x_i)$ denotes the rank of value x_i in the context of variable X etc.

4. The counts of concordant, discordant, x-tied and y-tied pairs (n_c, n_d, t_x and t_y, respectively) are determined (note that a pair can be both x-tied and y-tied at the same time i.e. counted both in t_x and in t_y). The counts are used in the calculation of the Tau value.
5. There are two Tau calculations, τ_A and τ_B, each resulting in values between -1.0 and 1.0. The former is simpler and is used when there are no ties. The Tau forumalae are:

$$\tau_A = \frac{n_c - n_d}{n(n-1)/2}$$

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_x)(n_c + n_d + t_y)}}$$

– The significance of Kendall's Tau also depends on the sample size. We test it by looking it up in the table of Kendall Tau critical values:

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
4	1.000	*	7	.619	.810	10	.467	.600	13	.359	.513
5	.800	1.000	8	.571	.714	11	.418	.564	14	.363	.473
6	.733	.867	9	.500	.667	12	.394	.545	15	.333	.467

Source: [US]

HOWTO: Calculating Kendall's Tau

Assumptions: The two variables we want to calculate Kendall's Tau for have orderable values (numeric or ordinal).

Example scenario: We want to calculate Kendall's Tau for variables **x** and **y**, with values shown on the right.

x	y
80	70
65	99
77	80
80	65
81	77
88	70

1) Assign ranks to the values

index	r(x)	x	y	r(y)
1	**3**	80	70	**4**
2	**6**	65	99	**1**
3	**5**	77	80	**2**
4	**3**	80	65	**6**
5	**2**	81	77	**3**
6	**1**	88	70	**4**

2) Determine concordance for all instance pairs, as done for pairs 1, 2 and 4, 6 in the picture

index	r(x)	x	y	r(y)
1	**3**	80	70	**4**
2	**6**	65	99	**1**
3	**5**	77	80	**2**
4	**3**	80	65	**6**
5	**2**	81	77	**3**
6	**1**	88	70	**4**

instance pair 1, 2: **discordant** (3 < 6 but 4 > 1)

instance pair 4, 6: **concordant** (3 > 1 and 6 > 4)

3) Count the number of concordant (n_c) and discordant (n_d) pairs

(1, 2): discordant
(1, 3): discordant
(1, 4): x-tied
(1, 5): concordant
(1, 6): y-tied
(2, 3): discordant
(2, 4): discordant
(2, 5): discordant
(2, 6): discordant
(3, 4): discordant
(3, 5): discordant
(3, 6): discordant
(4, 5): concordant
(4, 6): concordant
(5, 6): discordant

$\mathbf{n_c = 3}$, $\mathbf{n_d = 10}$

4) Calculate the number of pairs, using the formula

$$n_p = \frac{n(n-1)}{2}$$

where n is the number of instances and n_p is the number of pairs

$$n_p = \frac{6 \times 5}{2} = 15$$

5) Calculate Kendall's Tau using the formula

$$\tau_K = \frac{n_c - n_d}{n_p}$$

$$\tau_K = \frac{3 - 10}{15} = -0.47$$

HOWTO: Calculating Kendall's Tau - alternative for case with no ties

Assumptions: The two variables we want to calculate Kendall's Tau for have strictly orderable values (numeric or ordinal) i.e. neither variable has any duplicate values.

Example scenario: We want to calculate Kendall's Tau for variables **x** and **y**, with values shown on the right.

x	y
80	69
65	99
77	80
79	65
81	77
88	70

1) Assign ranks to the values

index	r(x)	x	y	r(y)
1	**3**	80	69	**5**
2	**6**	65	99	**1**
3	**5**	77	80	**2**
4	**4**	79	65	**6**
5	**2**	81	77	**3**
6	**1**	88	70	**4**

2) Order the values by rank, taking note of the original index

index of x	r(x)	x	y	r(y)	index of y
6	**1**	88	99	**1**	2
5	**2**	81	80	**2**	3
1	**3**	80	77	**3**	5
4	**4**	79	70	**4**	6
3	**5**	77	69	**5**	1
2	**6**	65	65	**6**	4

3) Draw lines between same-instance values of x and y, making sure that no pair of lines crosses more than once. Count the number of intersections. This is the number of discordant pairs.

index of x	x	y	index of y
6	88	99	2
5	81	80	3
1	80	77	5
4	79	70	6
3	77	69	1
2	65	65	4

$\mathbf{n_d = 10}$

4) Calculate the number of pairs, using the formula

$$n_p = \frac{n(n-1)}{2}$$

where n is the number of instances and n_p is the number of pairs

$$n_p = \frac{6 \times 5}{2} = 15$$

5) Calculate Kendall's Tau using the formula

$$\tau_K = \frac{n_p - 2n_d}{n_p}$$

Note that this is equivalent to the formula used in the previously presented method as the number of concordant pairs is $n_c = n_p - n_d$ and subsequently $n_c - n_d = n_p - 2n_d$.

$$\tau_K = \frac{15 - 2 \times 10}{15} = -0.33$$

- **t-Test for comparing two groups**
  - We have already encountered the t-test, in the section on hypothesis tests, where it was used to decide if a population had a hypothesised parameter value. Here we look at how the same t-distribution is used to decide if the difference of means for two groups of observations is statistically significant (i.e. not likely to be due to statistical variation).
  - Defined by William Sealy Gosset (1876-1937), who had to publish the t-test under the pseudonim Student, because of regulations at his place of employment, the Guinness Brewery
  - There are two types of t-test:
    * **with paired samples**: applicable when the values in the two groups of observations are closely connected, instance by instance (e.g. one group of observations are blood pressure measurements for a group of patients before treatment and the other group of observations are blood pressure measurements for *the same group of patients* after treatment) - ***we will not be studying this test type in further detail***
    * **with unpaired samples**: applicable when the grouped values come from different instances, grouped by some attribute (e.g. one group of observations are blood pressure measurements for the patients in hospital A and the other group are blood pressure measurements for patients in hospital B)

– Different formulae are used for the T-statistic, depending on whether the standard deviations differ a lot.

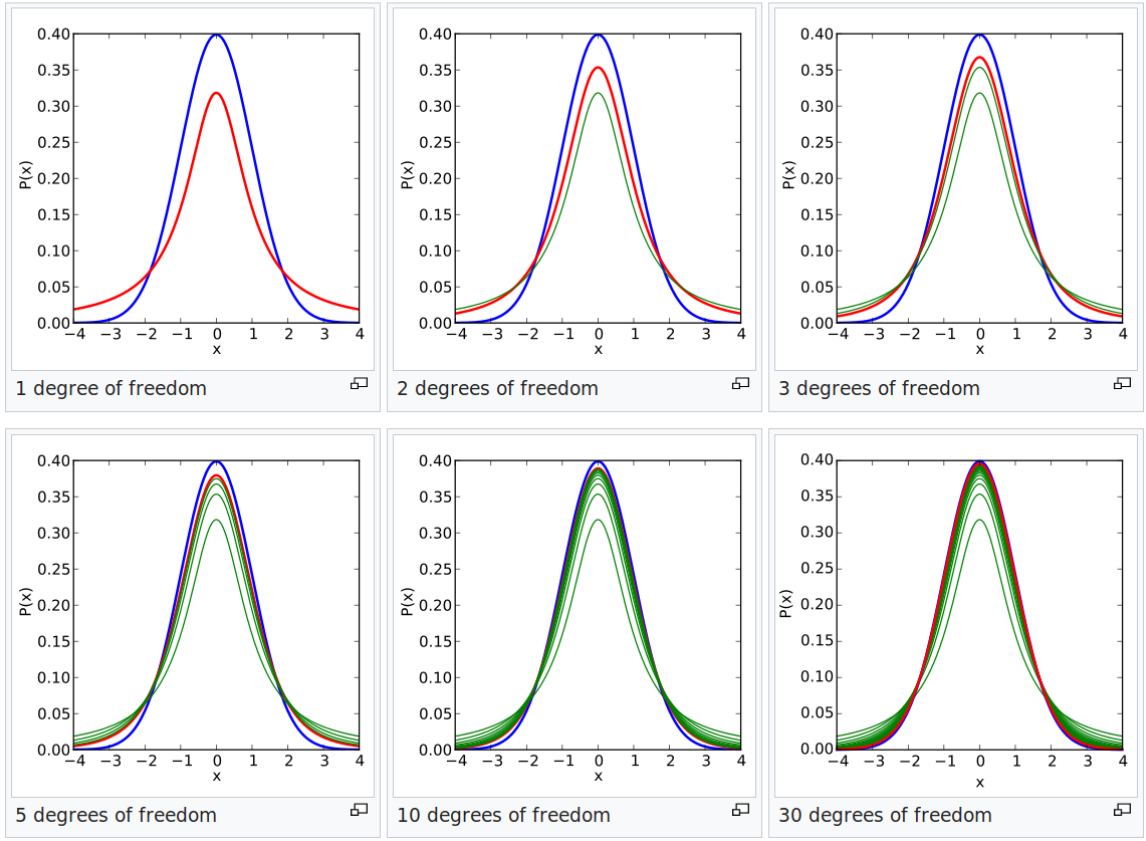
SD ratio range	**Student's t-test:** Standard deviations differ by a factor of 2 or less $\left(\frac{1}{2} \leq \frac{S_1}{S_2} \leq 2\right)$	**Welch's test:** Standard deviations differ by more than a factor of 2 $\left(\frac{S_1}{S_2} < \frac{1}{2} \quad \text{OR} \quad \frac{S_1}{S_2} > 2\right)$
T-statistic	$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$	$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Degrees of freedom (df)	$df = n_1 + n_2 - 2$	$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$

In the table s_1^2 and s_2^2 are group variances, $\bar{x}_1$ and $\bar{x}_2$ are the group means, n_1 and n_2 are group observation counts.

- **T-test assumptions:**
  * Either the sample is large (> 30) or the data for both groups is normally distributed.
  * The number of samples in the two groups does not have to be the same, although equal group sizes allow the use of Student's t-test (rather than Welch's test) regardless of the difference between standard deviations.

– The t-distribution is not a single probability density function (PDF) but a series of distributions corresponding to different *degrees of freedom*.

Density of the *t*-distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).
Previous plots shown in green.



Source: Wikipedia

- Once the correct t-distribution is identified based on the degrees of freedom, the T statistic value is tested in the same way as any other statistic, with respect to a certain level of significance (e.g. 5% or 1%).
- t-statistic critical value table for lookup can be found here.

– <u>T-TEST AS A TEST OF CORRELATION</u>

The t-test can be viewed as a **test for whether there is a relationship between a numeric variable (that for which the values are grouped) and a categorical, grouping variable**.

For example, if we are testing the difference of means between blood pressure measured for patients of hospital A on the one hand and the blood pressure measured for patients of hospital B, the hospital (A or B) is the categorical variable that groups the numeric blood pressure values. If it is found that the difference between the means is statistically different, then which hospital the measurement is taken in can tell us something about the value of blood pressure that we can expect for new patients and vice versa - this equates to a correlation.

- **ANOVA**
  - Tests the variance of *three or more groups of observations* for whether there is a significant difference between their means (i.e. probably not due to normal variation in the samples) - it is like the t-test but is performed on 3 or more groups of numeric values
  - The name stands for *completely randomized one-way analysis of variance*
  - Can be applied to cases where the groups are independent and random, the distributions are normal and the populations have similar variances
  - A hypothesis test, where the null hypothesis is that the means of the groups are equal
  - Central to ANOVA is a number called the F statistic, which is essentially the ratio between inter-group variation and intra-group variation
  - The F-statistic is compared with the critical value from a table called the F-table (based on the F-statistic distribution) as a means of deciding whether the null hypothesis should be accepted. The critical value depends on the required confidence level, the number of groups and the number of observations.

- Calculation of the F-statistic:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}, \quad MSW = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N-k}, \quad F = \frac{MSB}{MSW}$$

  where k is the number of groups, N is the overall number of observations, n_i is the number of observations in group i and s_i is the standard deviation within group i, $\bar{x}_i$ is the mean within group i and $\bar{\bar{x}}$ is the overall mean.

- The F-distribution depends on two **degrees of freedom** values, which also need to be used when looking up the F-table:

$$df_{between} = k - 1, \quad df_{within} = N - k$$

- A file containing an F-table can be found here.

– <u>ANOVA AS A TEST OF CORRELATION</u>

As with the t-test, ANOVA can be viewed as **a test for correlation between the numeric variable that contributes the values and a categorical variable that groups the values**.

For example, the test could be performed on the heights of trees sampled from three different forests. In this case the forest (with values e.g. *North Forest*, *Big Forest*, *Old Forest*) is the categorical variable and the tree height is the numeric variable.

- **Chi-squared**
  - Tests the independence of categorical variables (on the nominal or ordinal scale)
  - Chi-squared can also be used to test goodnes of fit for a distribution (but we are not looking at that here)
  - A hypothesis test where the hypothesis is that there is no relationship between the variables
  - The statistic on which the hypothesis test is based is:

$$\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where k is the number of cells (categories) in the contingency table for the two variables, O_i is the observed frequency in cell i and E_i is the expected frequency for cell i. The expected cell frequency is calculated as:

$$E_{ij} = \frac{n_{Ri} \times n_{Cj}}{n}$$

where E_{ij} is the expected frequency for cell in row i and column j, n_{Ri} is the sum of frequencies in row i, n_{Cj} is the sum of frequencies in column j and n is the sum of frequencies across the entire table.

- The calculated χ^2 value is compared with the critical value for the required confidence level and number of degrees of freedom $(df = (r-1) \times (c-1)$, where r and c are the number of rows and columns, respectively, in the contingency table) from the standard chi-squared table. If the calculated χ^2 is greater than the critical value, the null hypothesis is rejected and it is taken that *there is a relationship* between the categorical variables.
- A Chi-squared table of critical values is shown on the next page.

Table of Chi-squared critical values

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	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
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

HOWTO: Calculating Chi-squared

Assumptions: We have the contingency table for two categorical variables.

Example scenario: We want to calculate Chi-squared for variables `car colour` and `accident in 10 years`, with contingency table shown on the right.

car colour → accident ↓	silver/white	other
YES	22	27
NO	222	243

1) Calculate the row, column and overall counts

car colour→ accident ↓	silver or white	other	ROW TOTALS
YES	22	27	$n_{r1} = 49$
NO	222	243	$n_{r2} = 465$
COLUMN TOTALS	$n_{c1} = 244$	$n_{c2} = 270$	$n = 514$

2) Using the formula for expected frequency ($E_{ij} = \frac{n_{Ri} \times n_{Cj}}{n}$), populate an expected frequencies table

For example: $E_{12} = \frac{49 \times 270}{514} = 25.74$

car colour → accident ↓	silver or white	other
YES	23.26	25.74
NO	220.74	244.26

3) Calculate the term of the Chi-squared sum for each cell and place in a table; the formula is $T_{ij} = \frac{(E_{ij} - O_{ij})^2}{O_{ij}}$, where O_{ij} is the observed value for cell in row i and column j

For example: $T_{12} = \frac{(25.74 - 27)^2}{27} = 0.059$

car colour → accident ↓	silver or white	other
YES	0.072	0.059
NO	0.007	0.007

4) Calculate Chi-squared by adding all the terms:

$$\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1,j=1}^{k_R,k_C} T_{ij}$$

$$\chi^2 = 0.072 + 0.059 + 0.07 + 0.07 = 0.145$$

This value can now be looked up in the Chi-squared table of critical values.

Summary of covered measures of relatedness

Table: Measures of relatedness by variable types

variable type/variable type	**categorical**	**ordinal/ interval-scale numeric**	**numeric**
categorical	Chi-square [Information Gain] [Lift]		t-test ANOVA
ordinal/ interval-scale numeric		Kendall's Tau	Kendall's Tau
numeric			Pearson's correlation coefficient

The p-value

- A **p-value** expresses the level of significance of a calculated statistic.
- **Any value for a statistic** (such as the measures of relatedness covered here) is associated with a p-value. For example, Pearson's correlation coefficient of 0.27, calculated on a sample size of 100, has the p-value 0.007.
- The p-value is derived from the distribution for the statistic and represents the **probability that the statistic will be equal to or greater than the calculated value**. Following the previous example, the probability of obtaining 0.27 or greater for Pearson's correlation coefficient, r, with a sample of 100 on *unrelated variables* is about 0.07 (the distribution used to derive this is the distribution of r when calculated on samples of 100 for two unrelated variables).
- A smaller p-value corresponds to **more extreme and less common** values of the calculated statistic.
- When performing the tests by hand, we do not calculate the p-value. Rather, we look up the statistic values corresponding to some standard p-values (e.g. 0.05 or 0.01) to see how our calculated statistic compares. If our calculated statistic value is more extreme (usually greater) than the statistic value for a particular standard p-value (e.g. 0.01), then our statistic value is *significant at the p-value level of significance* (e.g. 'significant at the 0.01 level of significance').
- When performing the tests programmatically, the available functions (e.g. in Python with scipy.stats functions) return the p-value together with the statistic and we can test for significance directly by comparing the p-value with the required level of significance. Even if no level of significance is stated, the p-value stands on its own as an indication of how significant our results is.

References Some pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.