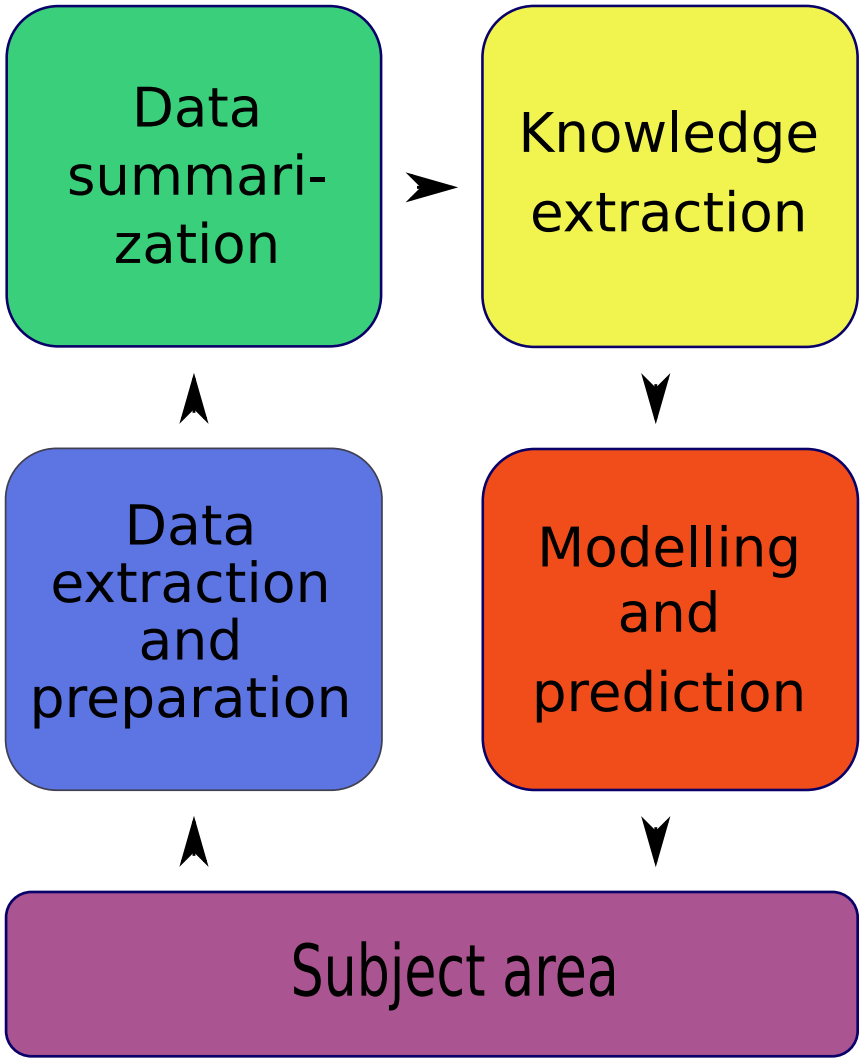
## The data analysis cycle



# Data Analysis: Relationships Between Data

Institute of Technology Tallaght

Department of Computing

# Relationships

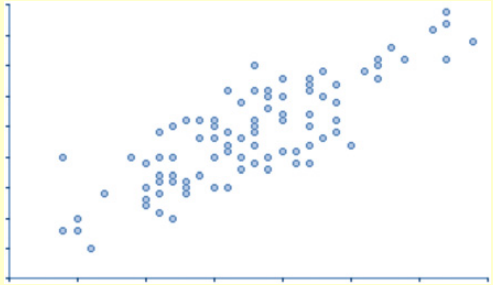
- Different relationships between data are of interest in data analysis:
  - Relationships between **numerical data variables** (on the interval or ratio scale)
  - Relationships between **data variables on the nominal and ordinal scales**
  - Relationships between **data set statistics**, such as the mean or the standard deviation, for different variables
- Relationships can be examined using **visualisation** or expressed through mathematically calculated **measures**
- NOTE: The existence of a relationship between two variables *does not* imply that there is any causation involved.

# Relationship visualisation

Relationships can be discovered by examining suitable visual representations of data

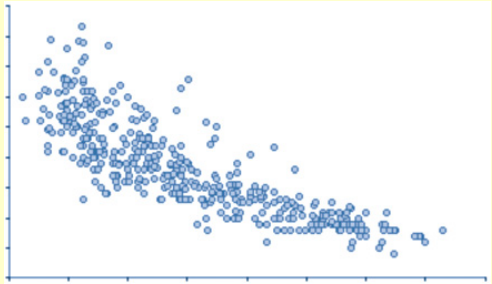
- **Scatterplots**

*Positive relationship*



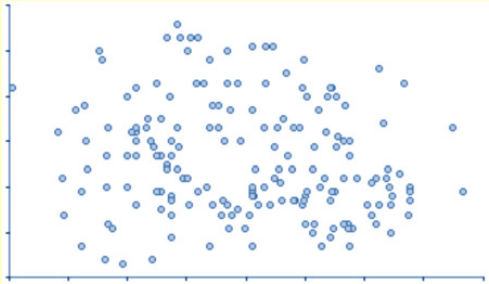
**Source: [MSD]**

*Negative relationship*



**Source: [MSD]**

*No relationship*



**Source: [MSD]**

- **Summary tables and graphs**

  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.

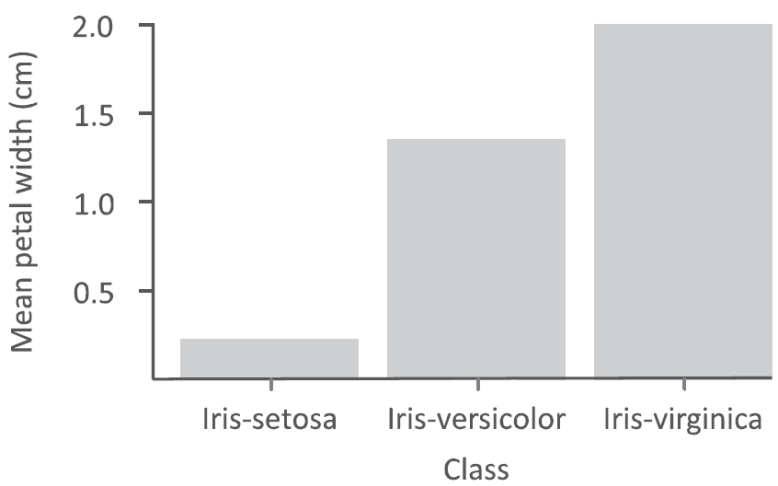
  - Summary table
    * Each data subset is shown in a row of the table
    * One of the columns usually contains observation counts for the subsets
    * The remaining columns represent properties, such as mean or standard deviation, of a numeric variable

| Class | Count | Minimum ( petal width (cm)) | Maximum ( petal width (cm)) | Mean ( petal width (cm)) | Median ( petal width (cm)) | Standard deviation ( petal width (cm)) |
|---|---|---|---|---|---|---|
| Iris-setosa | 50 | 0.1 | 0.6 | 0.244 | 0.2 | 0.107 |
| Iris-versicolor | 50 | 1 | 1.8 | 1.33 | 1.3 | 0.198 |
| Iris-virginica | 50 | 1.4 | 2.5 | 2.03 | 2 | 0.275 |

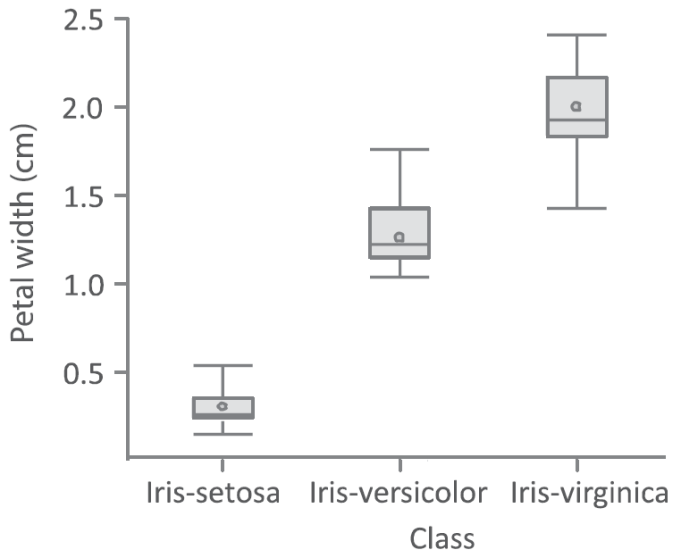
**Source: [MSD]**

*The data in the table above is grouped by type of iris. The columns represent properties of the variable "petal width". 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**
  - 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)

A number of 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 (on interval or ratio scale)
  - 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*.

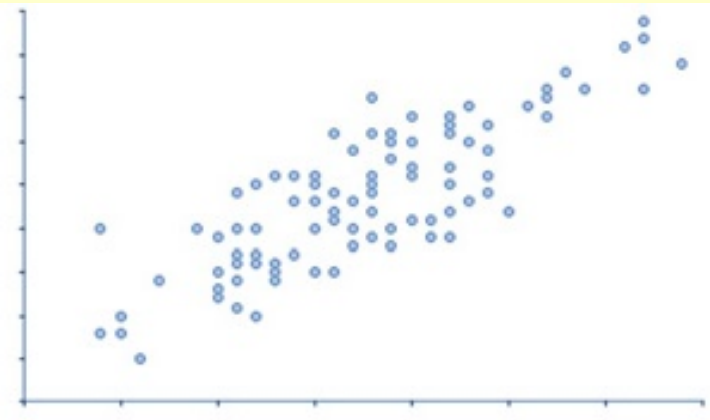
- The formula for calculating the correlation coefficient is:

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

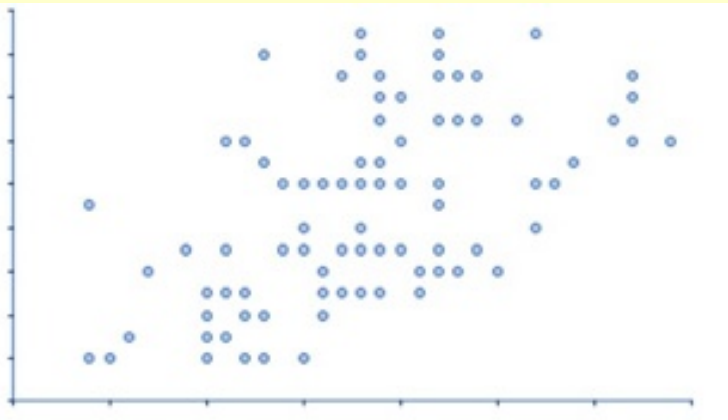
where $x$ and $y$ are variables, $x_i$ are the individual values of $x$, $y_i$ are the individual values of $y$, $\bar{x}$ is the mean of the $x$ variable, $\bar{y}$ is the mean of the $y$ variable, $s_x$ and $s_y$ are the standard deviations of variables $x$ and $y$, respectively, and $n$ is the number 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 (on the interval or ratio scale)
  - Defined by English statistician Maurice Kendall (1907-1983)
  - Also called *Kendall rank correlation coefficient*
  - 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 |
| $r(x_i) - r(x_j) = 0$ | x-tied |
| $r(y_i) - r(y_j) = 0$ | 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, $\tau_A$ and $\tau_B$, 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]**

- **t-Tests for comparing two groups**
  - Test the means of two groups of observations for whether the difference between their means is statistically significant (not likely to be due to expected variation in values)
  - Take the form of hypothesis tests
  - Is typically applied in cases where the number of instances is small (less than 30)
  - The statistic used in a t-test is called a T value and is the difference between the two means, normalised to the standard error of the difference distribution
  - The T value follows a t-distribution with a number of degrees of freedom that depends on the properties of the two variables
  - The formula for this value, in the case that the two groups are independently normally distributed and have similar variances, is:

$$T = \frac{\bar{x_1} - \bar{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

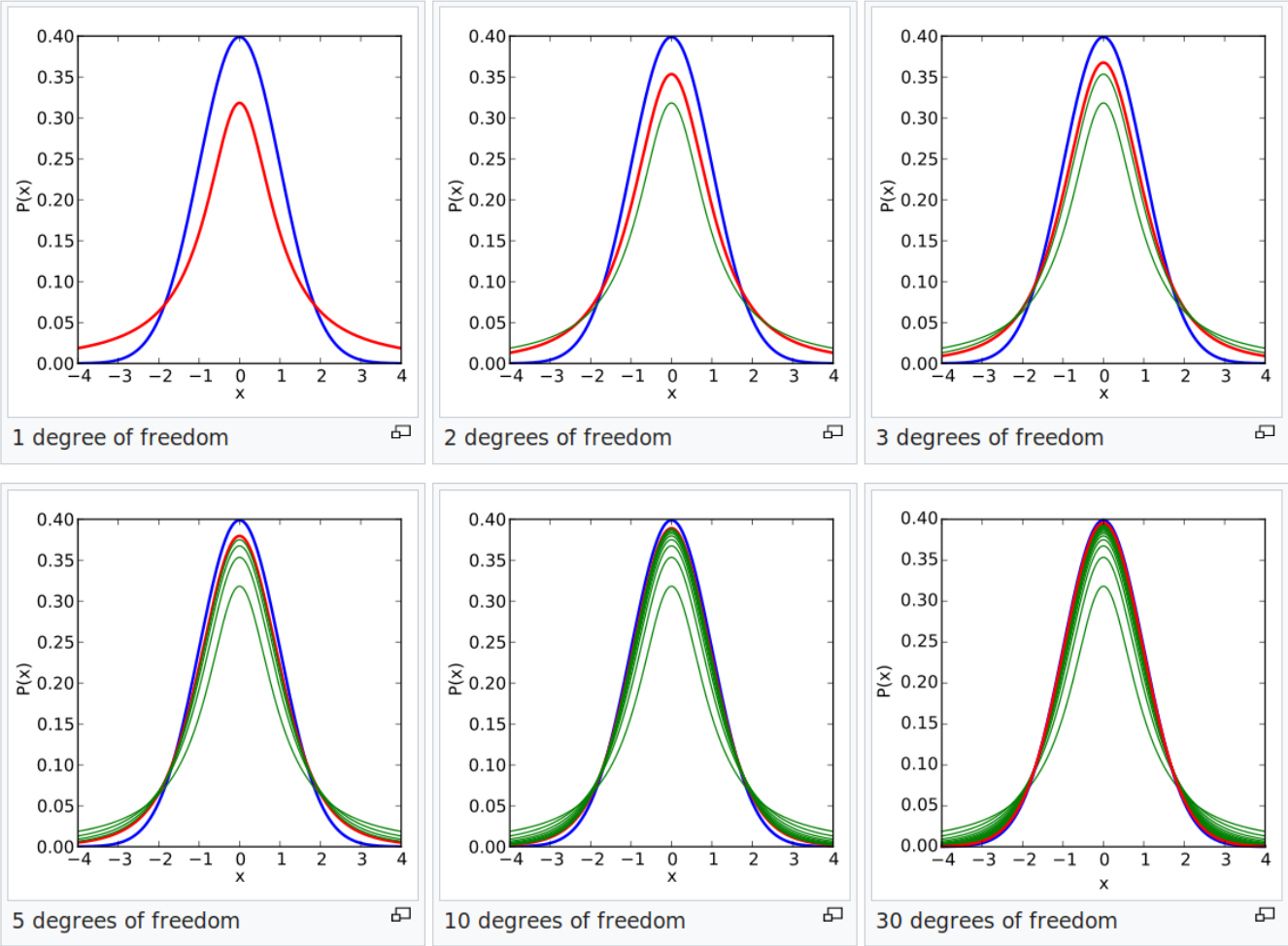
where $\bar{x_1}$ and $\bar{x_2}$ are the group means, $n_1$ and $n_2$ are group observation count and $s_p$ is an estimate of the standard deviation, calculated as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where $s_1^2$ and $s_2^2$ are group variances.
In this case the number of degrees of freedom is $df = n_1 + n_2 - 2$.

- In the case that the group variances are not similar, a different, more complicated formula is used.
- Testing the significance of a t statistic
  * If the calculated value of T falls within a certain range in the middle of the relevant t-distribution (usually a range that includes 95% of the distribution's values), it is taken that the difference between the two group means is not statistically significant i.e. that the means of the populations they represent are highly likely to be the same.
  * t-statistic distribution graphs:

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**

∗ t-statistic critical value table for lookup can be found here.

- **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)
  - 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:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$MSW = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{N-k}$$

$$F = \frac{MSB}{MSW}$$

  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:

$$df_{between} = k - 1$$

$$df_{within} = N - k$$

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

- **Chi-square**
  - Tests the independence of categorical variables (on the nominal or ordinal scale)
  - Chi-square 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 value on which the hypothesis test is based is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

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:

$$E_i = \frac{n_{Ri} \times n_{Ci}}{n}$$

where $n_{Ri}$ is the sum of frequencies in the entire row to which cell $i$ belongs, $n_{Ci}$ is the sum of frequencies in the column to which cell $i$ belongs and $n$ is the sum of frequencies across the entire table.

- The calculated $\chi^2$ 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-square table. If the calculated $\chi^2$ 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-square table can be found here.

**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.