Data Analysis:
Relationships Between Variables

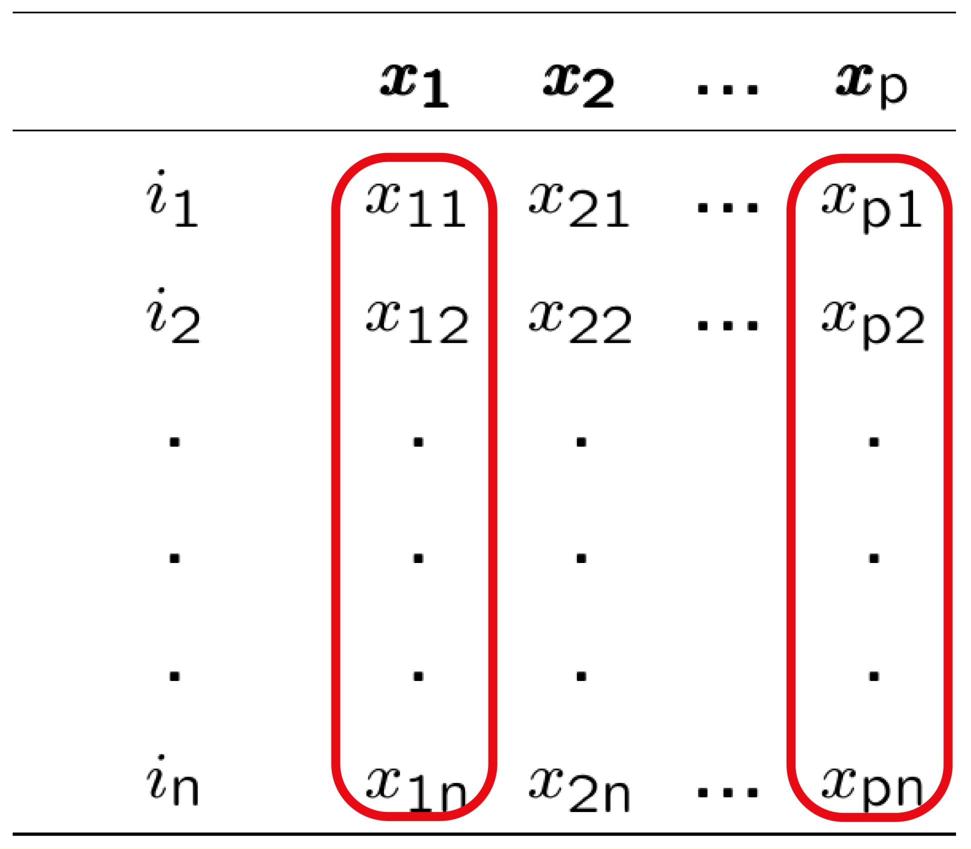
Technological University Dublin Tallaght Campus

School of Enterprise Computing and Digital Transformation

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 statistical measures**.
- The existence of a relationship between two variables **does not imply that there is any causation** involved.
- Studying relationships between variables allows us to better **understand** the world. It also supports the building of **models** for prediction and informed decision making.

- Typically, 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.



	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_p$
i_1	x_{11}	x_{21}	$\cdots$	x_{p1}
i_2	x_{12}	x_{22}	$\cdots$	x_{p2}
.	.	.		.
.	.	.		.
.	.	.		.
i_n	x_{1n}	x_{2n}	$\cdots$	x_{pn}

Intuition

The intuitive understanding of a relationship between two variables depends somewhat on the types of variables involved.

- **Two variables with orderable values:**

  The bigger the value of a variable in an instance, the bigger the value of a related variable; or, conversely, bigger values for one variable may be associated with smaller values for another. Examples: Pearson's P.M.C.C, Spearman's Rho, Kendall's Tau.

- **A numeric and a categorical variable:**

  The values of the numeric variable are substantially different (e.g. bigger) in instances belonging to one category than in instances belonging to another category.

- **Two categorical variables:**

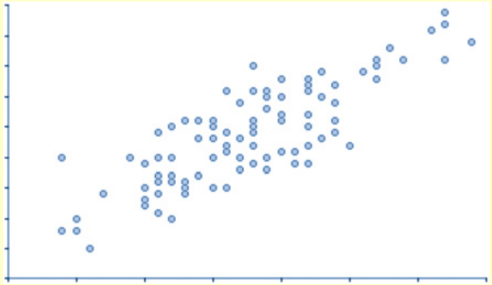
  A particular value of one categorical variable is more likely to co-occur with a particular value of another variable. The probability is higher than if the values were randomly distributed.

Relationship visualisation

Relationships can be discovered by examining suitable visual representations of data

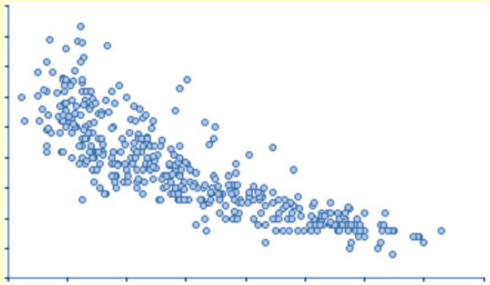
- **Scatterplots (two numeric variables)**

Positive relationship



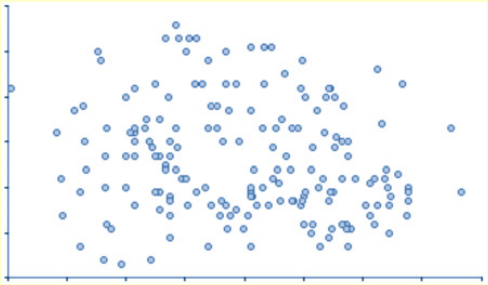
Source: [MSD]

Negative relationship



Source: [MSD]

No relationship



Source: [MSD]

- **Summary tables and graphs (one numeric and one categorical variable)**

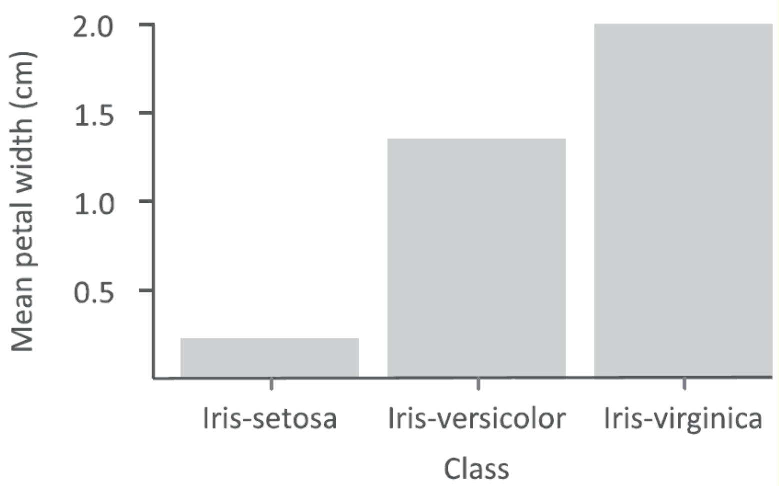
  To show data in a summary table or graph, the data is grouped based on the values of a discrete variable or ranges of a continuous variable. The table or graph summarises properties of a numeric variable for each group.

  - 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

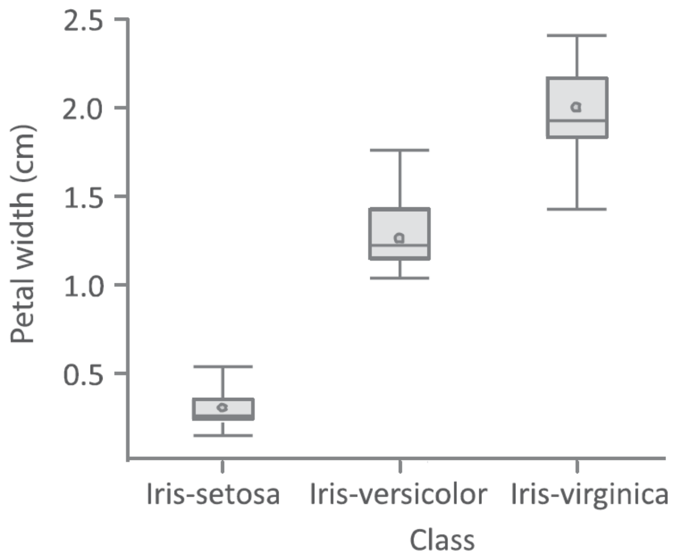
		Petal.Length				
Class	Count	Max (cm)	Min (cm)	Mean (cm)	Median (cm)	SD (cm)
setosa	50	1.9	1.0	1.46	1.5	0.17
versicolor	50	5.1	3.0	4.26	4.35	0.47
virginica	50	6.9	4.5	5.55	5.55	0.55

*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



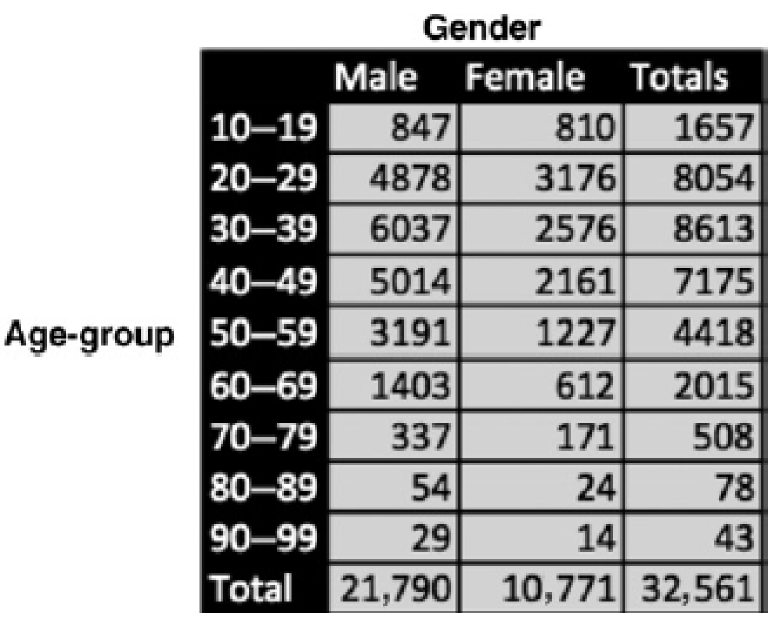
Source: [MSD]



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
	10–19	847	810	1657
	20–29	4878	3176	8054
	30–39	6037	2576	8613
	40–49	5014	2161	7175
Age-group	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

Statistical measures of relatedness

- These provide a measure of relatedness between two variables
- Two important concepts directly related to these measures are *effect size* and *p-value*

Concept: effect size

- The effect size is the strength of the relationship between the variables
- In some cases, the measure statistic is also used to express the effect size
- In other cases, the measure statistic is converted e.g. normalised to a value that is more readily interpretable as an effect size
- More recently, this has become an important aspect of evaluating relatedness between variables in statistics (in the past p-values held centre-stage)

Concept: p-value

- A p-value expresses how statistically significant a relationship between variables is
- The statistic is chosen so that its expected distribution under some 'uninteresting' conditions is known and the probability of the statistic falling into any particular value range under those conditions can be calculated. The p-value corresponding to a value for the statistic is the distribution-derived probability that the statistic should be exactly equal to or more extreme than that particular value for the statistic.
- If the p-value for a statistic value calculated on a sample is very small, it indicates that the calculated value would be very unusual under the 'uninteresting' conditions and is declared to signal that the 'uninteresting' conditions probably do not hold but that something 'interesting' is happening.
- In the context of relationships in data, the 'uninteresting' conditions are those of no relationship, while the statistics corresponding to very small p-values are taken to indicate that a relationship probably exists.
- When performing the relationship 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 considered *significant at the p-value level of significance* (e.g. 'significant at the 0.01 level of significance').
- When performing the relationship tests programmatically, the available 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.

Pearson’s product-moment correlation coefficient

[two numeric variables]

- 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*.
- A parametric measure of correlation.

- The formula for calculating the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} = \frac{S_{xy}^2}{S_x S_y}$$

  where x, y, x_i, y_i, $\bar{x}$, $\bar{y}$ are the two variables, their ith values and their means, S_x and S_y are the sample standard deviations and S_{xy}^2 the sample covariance of the two variables.

  – For the measure to be meaningful:

    ∗ x and y must be either

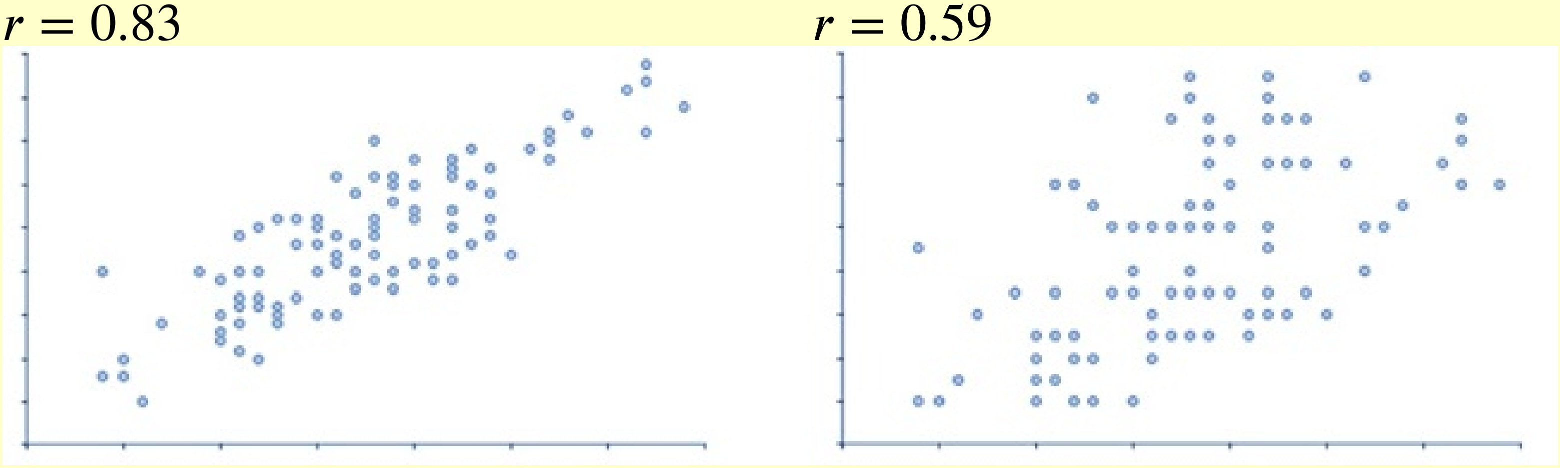
      · both variables on the interval or ratio scale

      · one a variable on the interval or ratio scale and the other a dichotomous variable (in which case a *point-biserial* or *biserial* correlation coefficient is calculated)

    ∗ each pair of values x_i and y_i must pertain to the same instance in the analysis domain

    ∗ there are no outliers in the data

- It only measures the extent of linear relatedness. If the relationship has a non-linear component, it will not be 'picked up' by the coefficient. This linear component may not be what we expect (click here for Anscombe's quartet on Wikipedia).
- The square of the correlation coefficient, r^2, is called the coefficient of determination and it measures the amount of variance that the two variables have in common.
- Examples of two scatterplots with different values of correlation coefficient:



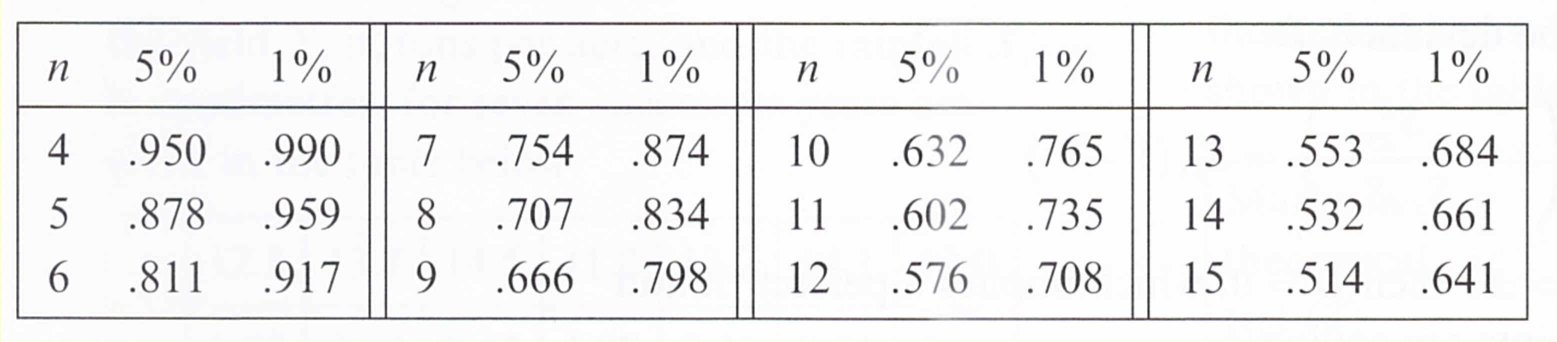
$r = 0.83$ $r = 0.59$

Source: [MSD]

- Effect size: this is equal to r (see row 1 in table here for magnitutes)

- Testing the significance of Pearson's product-moment correlation coefficient
  - Additional assumptions (see those listed for meaningfulness of Pearson's correlation coefficient, on the previous page) need to be met for this test to be valid (these will be dealt with separately later in the lecture):
    - the data must follow a bivariate normal distribution (but this is hard to test hence each variable is tested for normality separately and this is sufficient in most cases)
    - the variables must be homoskedastic
  - Under the above conditions (click for a very good article on Pearson's correlation coefficient assumptions), a pair of perfectly independent variables (the 'uninteresting case'), the coefficient calculated on a pair of samples has a distribution in the range -1 to 1, with mean equal to 0.
  - 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 given in this 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. Alternatively, the following t-distributed statistic can be used: $t_r = \dfrac{r\sqrt{N-2}}{\sqrt{1-r^2}}$. In both cases degrees of freedom are: $df = N - 2$.

- In the case that the assumptions for validity of the significance test are not met, the coefficient value can be tested using an empirical distribution with the bootstrap method. Alternatively, one of the non-parametric tests (Spearman's Rank Coefficient or Kendall's Tau) can be used.

Spearman's Rho (Rank Correlation Coefficient)

[two variables with orderable values]

- Measures relatedness betwen a pair of variables the values of which can be ordered i.e. ranked (numeric or ordinal)
- Defined by English psychologist and statistician Charles Edward Spearman (1863 - 1945)
- A non-parametric test of correlation - no assumptions are made about the data except that the values must be orderable (rankable)
- The value of ρ is computed in two steps:
  1. all values are ranked and the ranks form two new variables (the rank variables)
  2. Pearson's correlation coefficient is calculated for the rank variables
- Measures how close to a monotonic function the relationship between two variables gets (as opposed to Pearson's coefficient, which tries to fit the data to a straight line)
- Significance is read from a table, for which values originate from a discrete distribution (owing to the discreteness of the ranks)
- Effect size is assessed in the same way as for Pearson's PMCC

Kendall's Tau

[two variables with orderable values]

- Like Spearman's Rho, this measure is based on ranks rather than data values
- Defined by English statistician Maurice Kendall (1907-1983)
- Also called *Kendall rank correlation coefficient*
- This is a non-parametric test of correlation and does not impose strict conditions on data, except that they must be orderable
- 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.

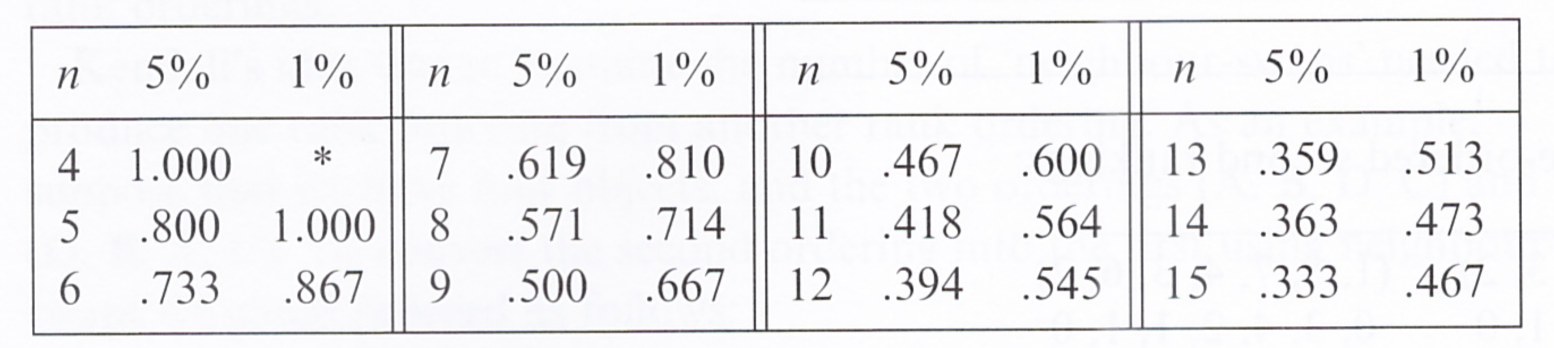
4. The counts of concordant, discordant, x-tied and y-tied pairs (n_c, n_d, t_x and t_y, respectively) are determined (note that a pair can be both x-tied and y-tied at the same time i.e. counted both in t_x and in t_y). The counts are used in the calculation of the Tau value.

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 can 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]

  The critical values are calculated from a discrete distribution for the no-relationship case (the 'uninteresting' scenario).

- The effect size of Kendall's Tau is assessed as the other correlation measures (see here).

HOWTO: Calculating Kendall's Tau

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

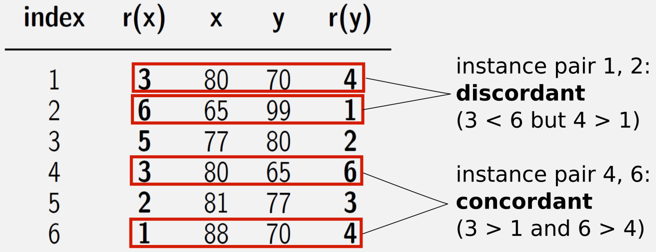
Example scenario: We want to calculate Kendall's Tau for variables **x** and **y**, with values shown on the right.

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

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

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

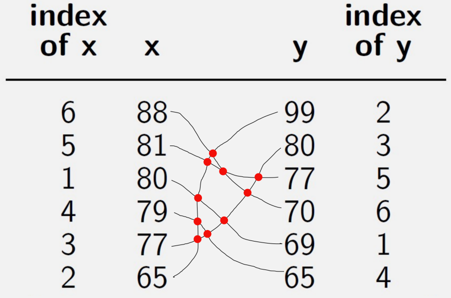
1) Assign ranks to the values

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

[a numeric variable and a dichotomous categorical variable]

- 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 (one-sample t-test). Now we look at how the 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). This is called a two-sample t-test.
- Defined by William Sealy Gosset (1876-1937), who had to publish the t-test under the pseudonym Student, because of rules imposed by his employer, 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)
  - **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)

Unpaired two-sample t-test

- Assumptions
  - One variable is numeric (at least interval scale), the other binary.
  - Either the samples are large (> 30) or the data are normally distributed.
  - The number of values 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.

- As with many other hypothesis tests, the t-test uses a statistic that is in fact a standardised residual* quantity:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_E} = \frac{\bar{x}_1 - \bar{x}_2}{S_E}$$

where x is the numeric variable, $\bar{x}_1$ and $\bar{x}_2$ are the sample means for the two groups defined by instance membership of the two categories, $\mu_1 - \mu_2$ is the hypothesised difference between the means (equal to 0) and $S_{E(\text{difference of means})}$ is the standard error of the difference of means.

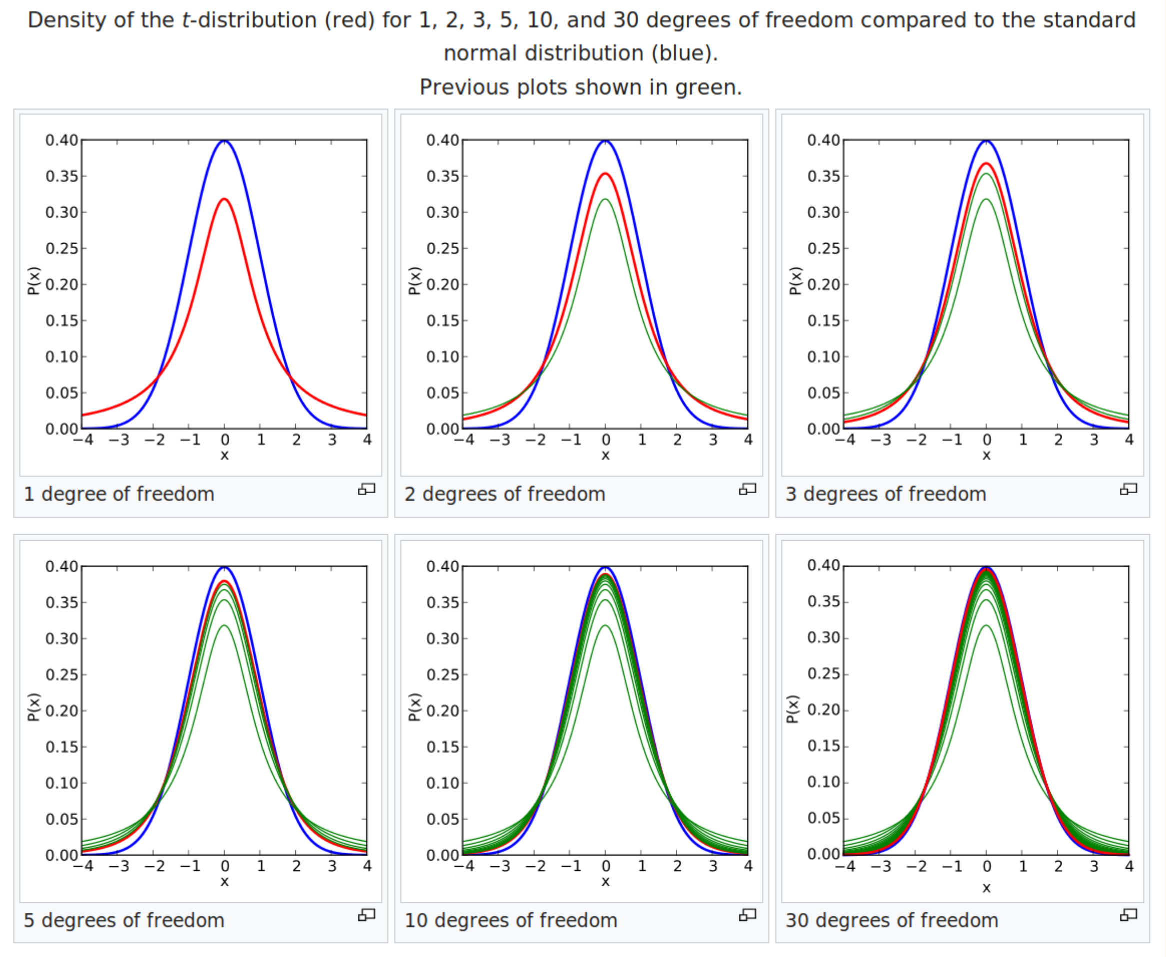
* Residuals represent the difference between experimentally obtained values (sample statistics) and hypothesised values (i.e. model values).

- 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.

- 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.

1 degree of freedom

2 degrees of freedom

3 degrees of freedom

5 degrees of freedom

10 degrees of freedom

30 degrees of freedom

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%).

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

- Effect size (see magnitude ranges):

Cohen's d

$$d = \frac{2t}{\sqrt{df}}$$

$\boldsymbol{\eta^2}$ (eta squared)

$$\eta^2 = \frac{t^2}{t^2 + df}$$

- T-TEST AS A TEST OF RELATEDNESS

  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.

Paired two-sample t-test

- Assumptions:

  These are the same as for the unpaired test, with the additional assumption that each value in the first group has a *pair* value in the second group, in that they are both associated with the same subject or instance, but based on measurements taken at different times, typically after a *treatment*. This implies that the two groups are of the same size.

- Test statistic:

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}}$$

  where $\bar{D}$ is the mean of the sample value pair differences, S_D is the sample value pair difference standard deviation and n is the number of value pairs in the sample

  Being calculated on the standard error of the mean of the value pair difference as variable (rather than the standard error of the difference of means as is the case with the unpaired test), this test has better *power* than the unpaired test.

- significance is derived as for the upaired test
- effect size is Cohen's d: $d = \frac{t}{\sqrt{df}}$

Non-parametric alternatives to the t-test

[a numeric variable and a dichotomous categorical variable]

Man-Whitney (or Wilcoxon's rank sum) test

- These are used as a non-parametric alternative to the unpaired two sample t-test when the assumption of normality is not met and the sample size is small.
- As we have seen, heteroskedacity can be handled by Welch's variant of the t-test
- The two named tests differ in details but their methods are equivalent.
- Idea:
  - data belonging to both groups are ranked in a single sequence
  - the ranks for each group are summed up
  - if the data are centred similarly, the sums would be similar (provided an adjustment is made for group size)
  - if there is a difference between how the data are centred, the sums are different

  - distributions for these sum statistics in the case of similarity ('uninteresting' case) are available (having been compiled using Monte Carlo methods i.e. with the use of random numbers for values)

- significance is derived from these approximately normal distributions
- effect size can be expressed using Rosenthal's r:

$$r = \frac{z}{\sqrt{N}}$$

For magnitude interpretation see here.

Wilcoxon's signed rank test

- This test is a non-parametric alternative to the paired two sample t-test.
- Similarly to the Man-Whitney test, it has an approximately normal distribution and its effect size can be expressed using Rosenthal's r

ANOVA

[a numeric variable and a categorical variable with $>$ 2 values]

- 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*
- 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
- Assumptions:
  - data in the groups are independent: there should be no correlation between the data in the groups and this should be ensured at experiment design time
  - if the group sizes are different then also
    - $*$ the group populations must have similar variances (homoskedasticity)
    - $*$ the distribution of each group must be normal (without significant skewness or kurtosis) **OR** the distribution of the residuals (all in one big group) should be normal

- Calculation of the F-statistic:

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

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

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

  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; SSB, SSW, MSB, MSW are the sums of squares and mean squares between and within groups.

- The F-distribution depends on two **degrees of freedom** values, are used in looking up the F-table:

$$df_B = k - 1, \quad df_W = N - k$$

  Can you tell where these numbers come from? We have had similar examples earlier in the module.

- The F-statistic differs from the t and z statistics in that it is a ratio of variances rather than a ratio in the variable dimension (i.e. residual/standard error). For the case of two group comparison, $F = t^2$, and is in essence the same test, albeit using a different statistic.

- Significance: F-statistic sampling distribution for the correct degrees of freedom (w.r.t. sample size and number of groups) should be used. A file containing an F-table can be found here.

- Effect size (for more information see here):

r^2

$$r^2 = \eta^2 = \frac{SSB}{SST} = \frac{SSB}{SSB + SSW}$$

Somewhat biased.

$\boldsymbol{\omega^2}$

$$\omega^2 = \frac{SSB - df_W MSW}{SSB + SSW + MSW}$$

Less biased.

- **Comparison of groups individually**
  - ANOVA is an **omnibus test**, which means that it tests for the existence of an effect within several entities (groups), without giving detailed information as to the existence of the effect (difference) between pairs of entities.
  - Performing a series of t-tests would have been time consuming in the past (when these calculations were done by hand), but pairwise testing has another problem: the inflation of the real type I error probability.

    $$p_{TypeIErr}(m, \alpha) = 1 - (1 - \alpha)^m$$

    In the case of 5 tests $(m = 5)$ and 0.05 significance $(\alpha = 0.05)$, $p_{TypeIErr}(5, 0.05) = 0.226$.
  - The way this problem is generally dealt with is to use a corrected critical value, which in the simplest case is $\alpha_B = \frac{\alpha}{m}$. This is the Bonferroni correction - an easy to understand approach.

- The Bonferroni correction is problematic because it controls the type I error probability but at the same time reduces the power for each test. Many more elaborate correction schemes have been designed. Some examples are in the table.

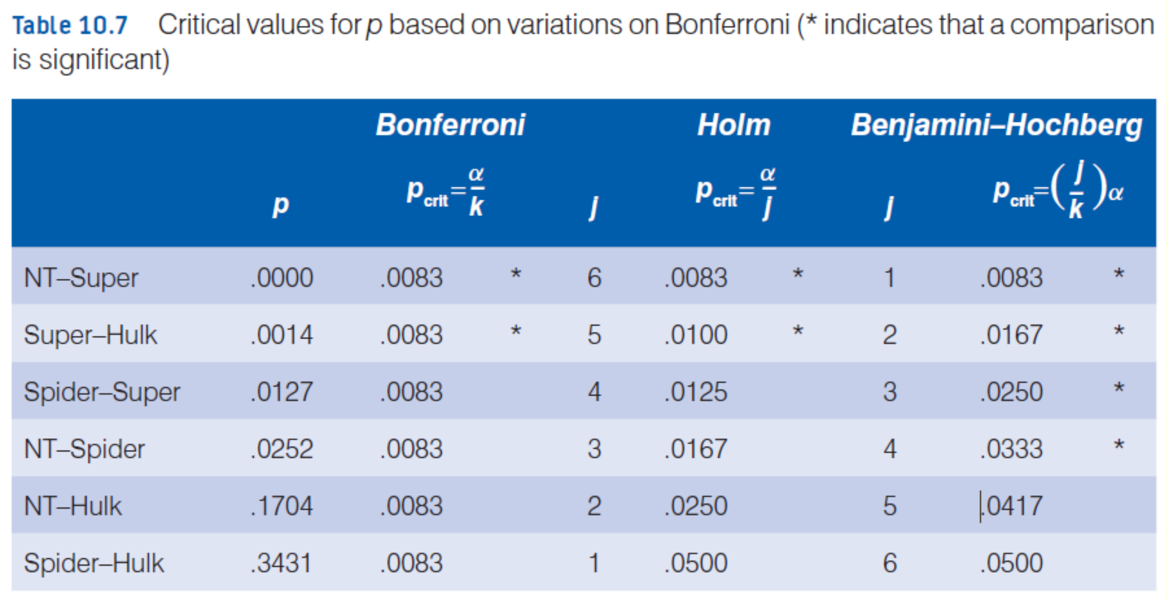


Table 10.7 Critical values for *p* based on variations on Bonferroni (* indicates that a comparison is significant)

		Bonferroni			Holm			Benjamini–Hochberg	
	p	$p_{crit}=\frac{\alpha}{k}$		j	$p_{crit}=\frac{\alpha}{j}$		j	$p_{crit}=\left(\frac{j}{k}\right)\alpha$	
NT–Super	.0000	.0083	*	6	.0083	*	1	.0083	*
Super–Hulk	.0014	.0083	*	5	.0100	*	2	.0167	*
Spider–Super	.0127	.0083		4	.0125		3	.0250	*
NT–Spider	.0252	.0083		3	.0167		4	.0333	*
NT–Hulk	.1704	.0083		2	.0250		5	.0417	
Spider–Hulk	.3431	.0083		1	.0500		6	.0500	

Source: DSR

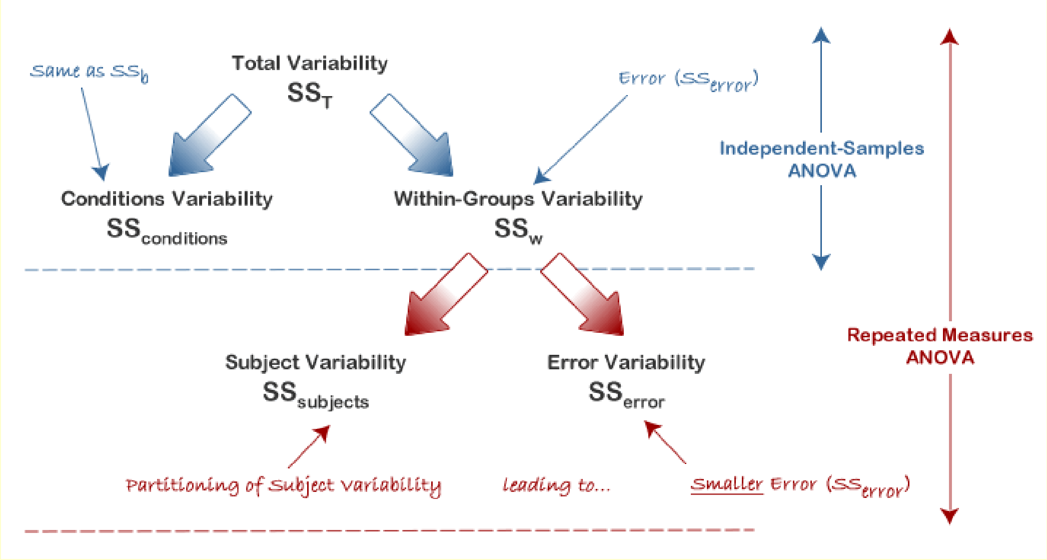
- **Post-hoc testing** involves performing all possible t-tests between pairs of groups (numbering $k(k-1)/2$ for k groups), finding the p-values and then deciding significance based on one of several patterns of correction.

- ANOVA AS A TEST OF CORRELATION

  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.

ANOVA with repeated measures

- Similarly to paired t-tests, ANOVA with repeated measures operates on groups of data with strictly related values by subject, having one value per subject in each group
- In general, this test has greater power than plain one-way ANOVA, resulting from the fact that the effect we are trying to model (the difference between groups) can be compared against a more focused variability that excludes differences between subjects (this could not be isolated in the general case ANOVA, where we did not know of any connections between the subjects from group to group)
- 

Source: laerd.com

- Assumptions
  - Normality of residuals - the test is robust so only approximate normality is ok
  - Sphericity - this is a requirement for all the variances and cross-variances to be approximately the same
    - sphericity can be tested using Mauchly's test
    - if Mauchly's test yields a significant result i.e. the data is not spherical, then corrections of the degrees of freedom (either Greenhouse–Geisser or Huynh–Feldt) allow a re-calculation of a more accurate p-value for the ANOVA test

Non-parametric alternatives to ANOVA

[a numeric variable and a categorical variable with $>$ 2 values]

- These alternatives are to be used when the assumptions of normality and homoskedasticity are not fulfilled.
- For unpaired groups of numeric data, if a non-parametric approach is required then the **Kuskal-Wallis** test, in principle this test is similar to the Man-Whitney/Wilcoxon rank sum test, can be used
- For paired groups, **Friedman's** test can be used as the non-parametric alternative

Chi-squared

[two categorical variables]

- 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 $\big(df = (r-1) \times (c-1)$, where r and c are the number of rows and columns, respectively, in the contingency table$\big)$ 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.
- The significance of the Chi-squared statistic can be deduced from the table of critical values shown on the next page.

- The effect is calculated as one of the following (see here for ranges):

Phi

$$\varphi = \frac{\chi^2}{n}$$

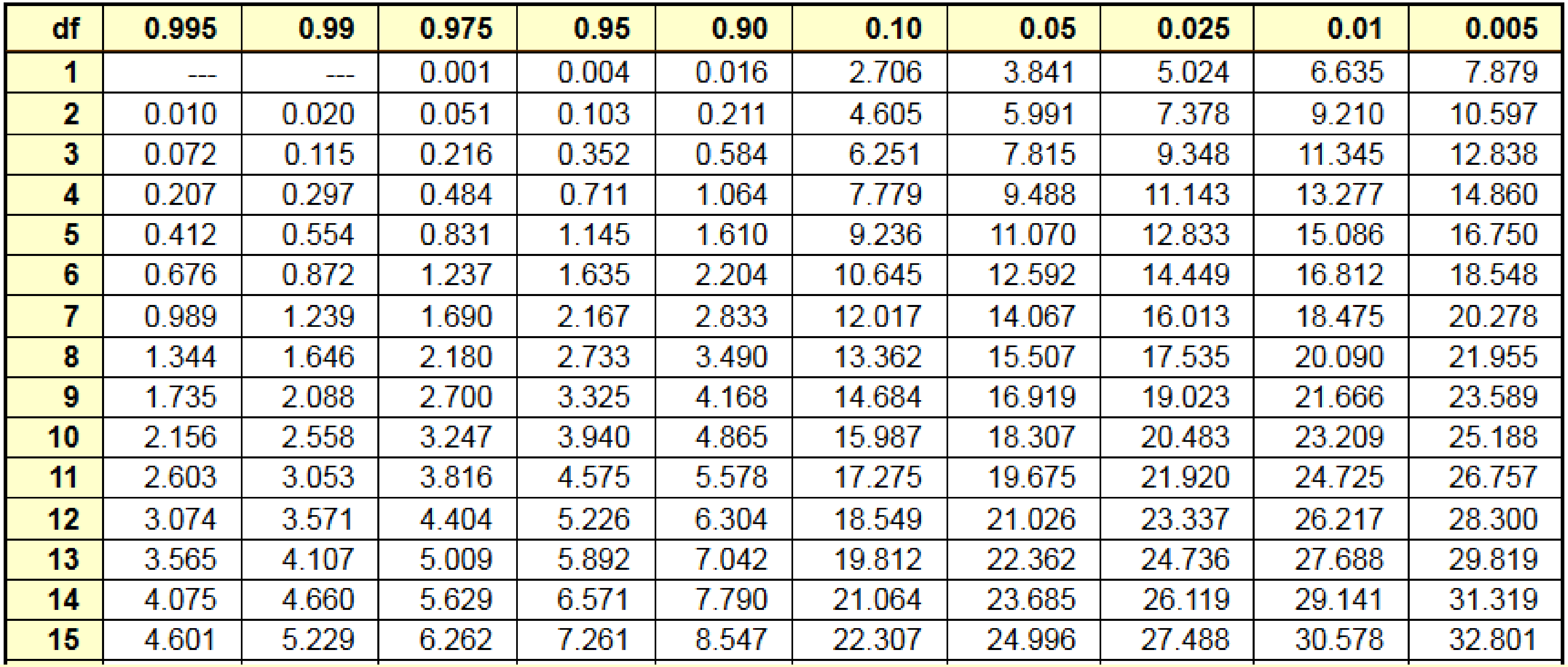
where χ^2 is calculated without Yates's correction.

Cramer's V

$$V = \frac{\chi^2}{n \times min(r-1, c-1)}$$

where r is the number of rows and c the number of columns in the table.

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 ($E_{ij} = \dfrac{n_{Ri} \times n_{Cj}}{n}$), populate an expected frequencies table

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

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 $T_{ij} = \dfrac{(E_{ij} - O_{ij})^2}{E_{ij}}$, where O_{ij} is the observed value for cell in row i and column j

For example: $T_{12} = \dfrac{(25.74 - 27)^2}{25.74} = 0.062$

car colour → accident ↓	silver or white	other
YES	0.068	0.062
NO	0.007	0.006

4) Calculate Chi-squared by adding all the terms: $\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1,j=1}^{k_R,k_C} T_{ij}$

$\chi^2 = 0.068 + 0.062 + 0.07 + 0.06 = 0.143$

This value can now be looked up in the Chi-squared table of critical values for df $= (r-1)(c-1) = (2-1)(2-1) = 1$.

Yates's correction

The Chi-squared distribution is only an approximation of the discrete frequency distribution arising from contingency tables. The difference is particularly pronounced for small tables (2x2), causing higher Type I error rates. This can be improved with the application of a correction suggested by English statistician Frank Yates, which reduces each expected-observed difference by 0.5:

$$\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

Chi-squared alternatives

- **McNemar's test for paired data**

  This is performed in the specific case of two paired measurements of a dichotomous variable.

- **Fisher exact test**

  This is a test with an exact distribution that can be used as an alternative to Chi-squared when the table is small. For larger tables, the calculations take too long and Chi-squared is more feasible. A rule of thumb is to use this test when any of the expected frequencies are 5 or lower.

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.