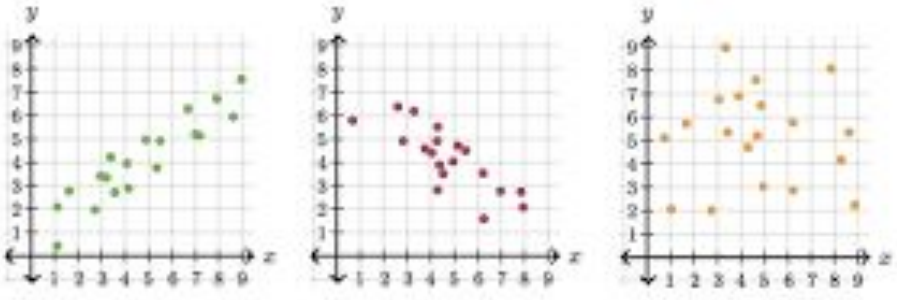
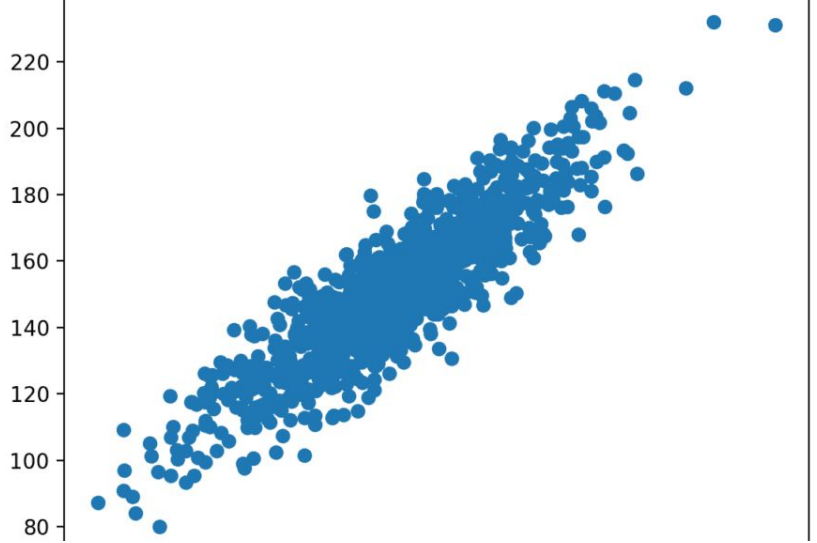


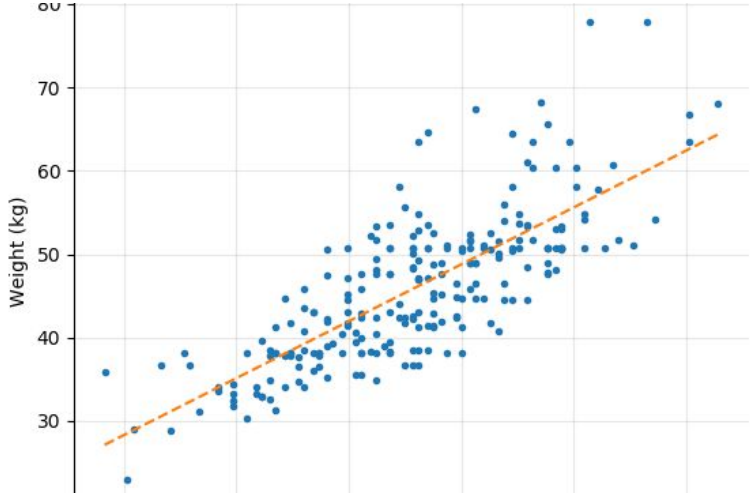


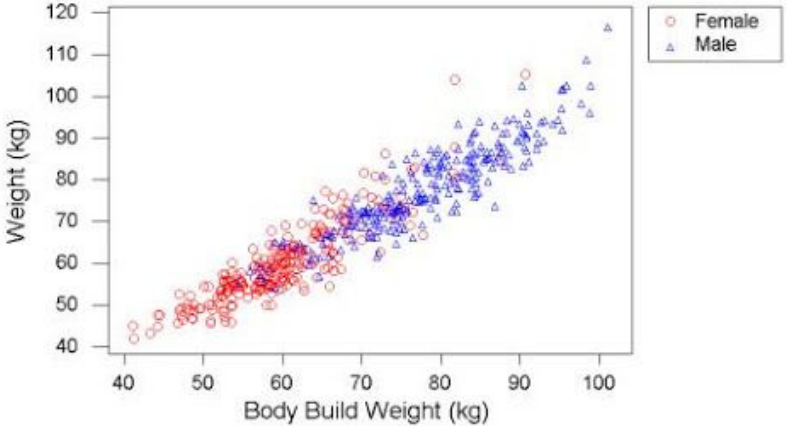
Correlation

Dr.Suresh Kumar Murugesan PhD

About me

- Dr.Suresh Kumar Murugesan is a passionate Professor, researcher and Mental Health Practitioner from Madurai, Tamil Nadu, India
- At present he is heading the department of Psychology, The American College, Madurai
- He is very keen in learning new research studies in behavioural Sciences and open to learn.
- His ultimate aim is to make impression in the field of Knowledge
- His area of specializations are Psychomentry, Psychotherapy, Positive wellbeing, Education Psychology, Cognitive Psychology
- WhatsApp +91 975040 6463 email - sureshkumar800@yahoo.com





Correlation

Correlation is a statistic that measures the degree to which two variables move in relation to each other.

Correlation

When two sets of data are strongly linked together we say they have a **High Correlation**.

The word Correlation is made of **Co-** (meaning "together"), and **Relation**

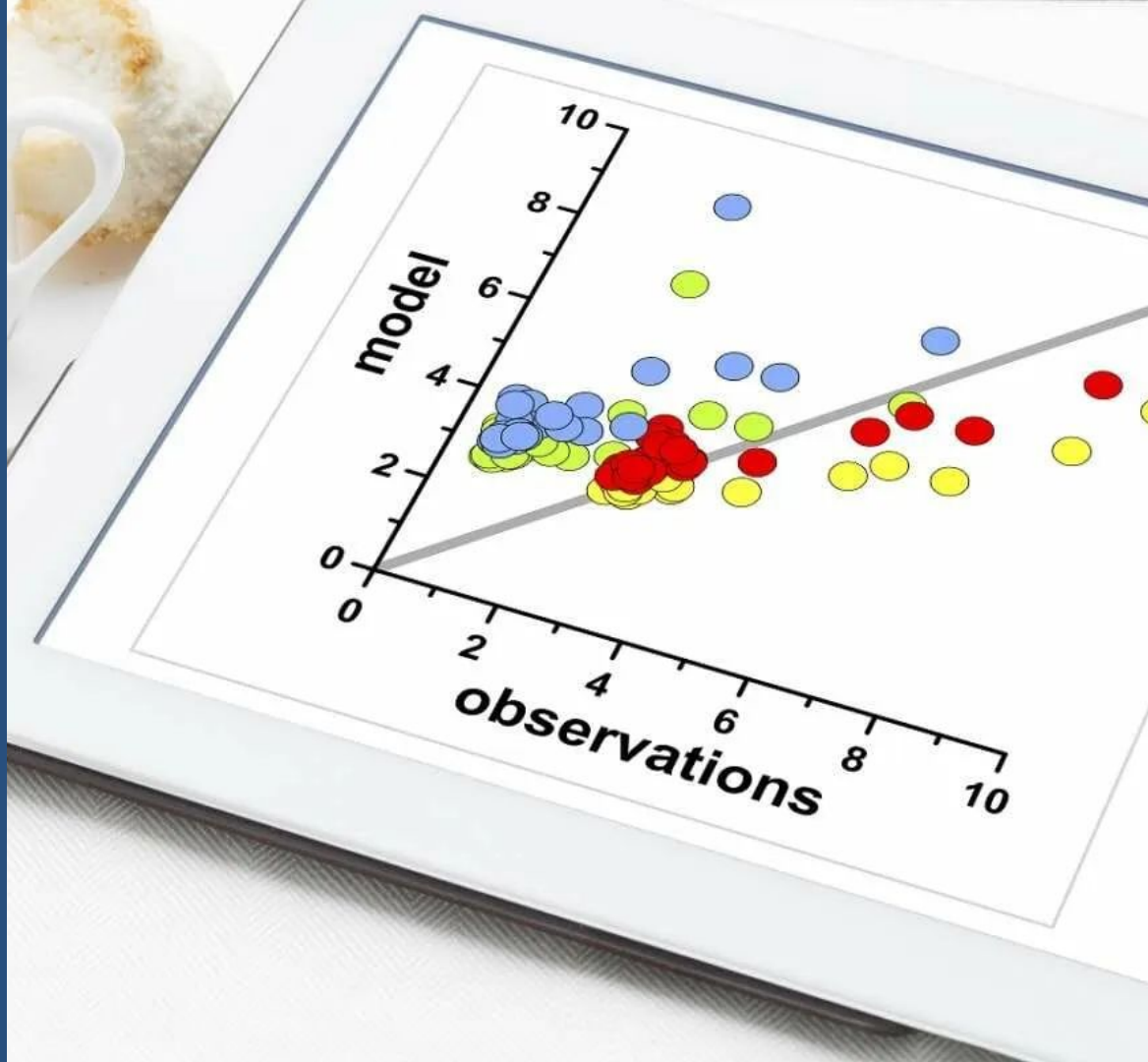
- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

Correlation

- The word correlation is used in everyday life to denote some form of association.
- We might say that we have noticed a correlation between hot day and anger.
- However, in statistical terms we use correlation to denote association between two quantitative variables.
- We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.
- The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

Correlation coefficient

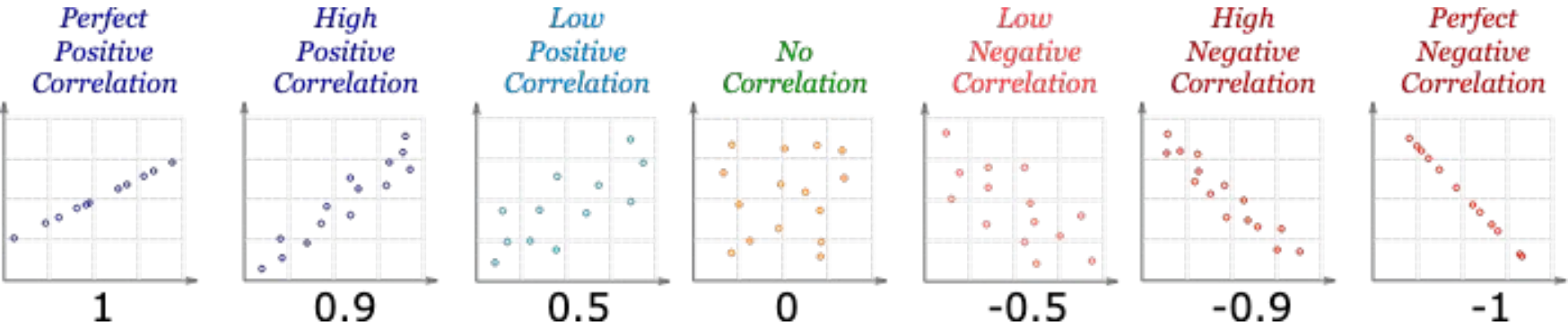
- The degree of association is measured by a correlation coefficient, denoted by r.
- It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association.
- If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.



Correlation coefficient

- The correlation coefficient is measured on a scale that varies from + 1 through 0 to – 1.
- Complete correlation between two variables is expressed by either + 1 or -1.
- When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative.
- Complete absence of correlation is represented by 0.

A correlation is assumed to be **linear** (following a line).



Correlation

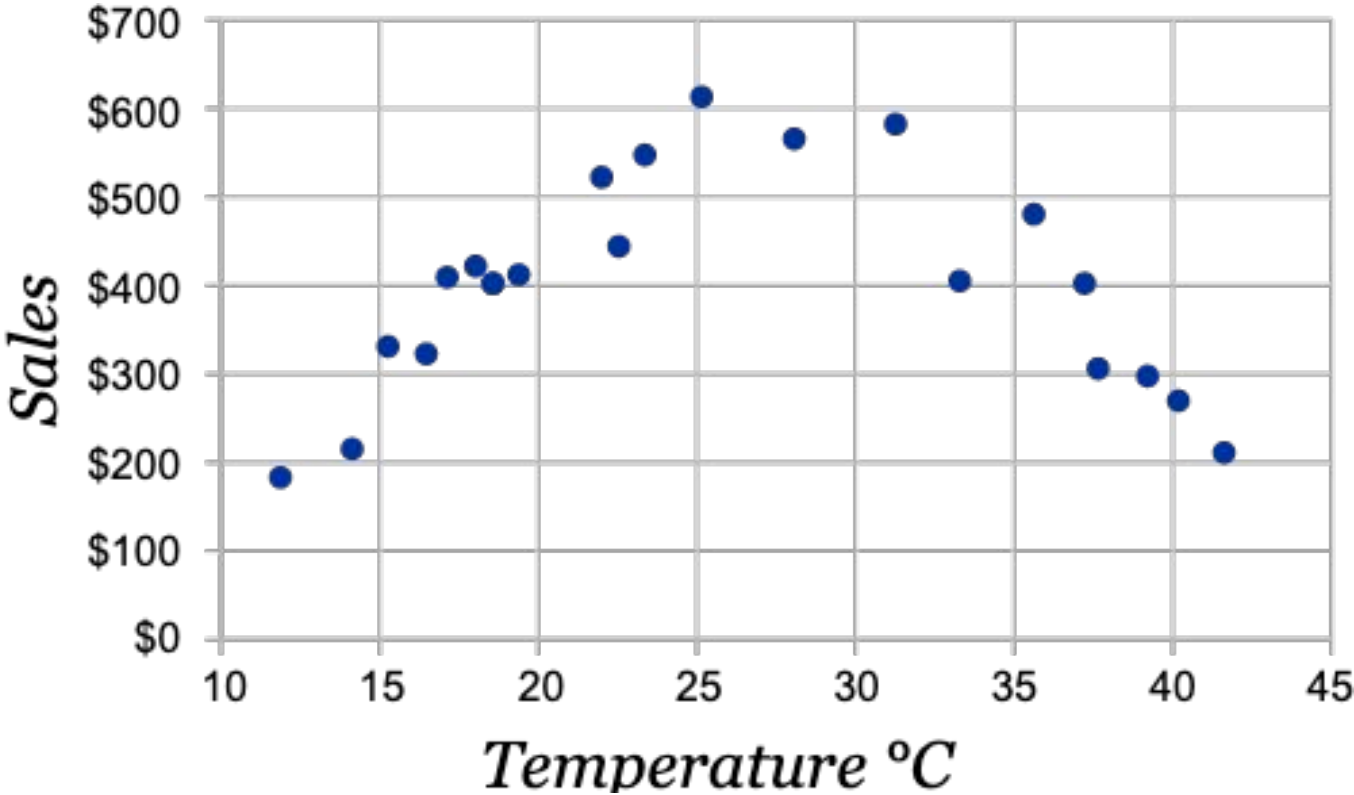
Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

Correlation Is Not Good at Curves

The correlation calculation only works properly for straight line relationships.



Correlation Is Not Causation

A common saying is "Correlation Is Not Causation".

What it **really** means is that a correlation does **not prove** one thing causes the other:

- One thing **might** cause the other
- The other **might** cause the first to happen
- They may be linked by a different thing
- Or it could be random chance!

There can be many reasons the data has a good correlation.

Types of Correlation

Type of Correlation

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

1. A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
2. A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
3. A **zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

Positive Correlation

- *Positive correlation* is a relationship between two variables in which both variables move in the same direction.
- This is when one variable increases while the other increases and visa versa.
- For example, *positive* correlation may be that the more you exercise, the more calories you will burn.

Pearson’s Product-Moment Correlation

The most common measure of correlation is *Pearson’s product-moment correlation*, which is commonly referred to simply as the *correlation*, the *correlation coefficient,* or just the letter *r* (always written in italics). The *correlation* coefficient r measures the strength and direction of a linear relationship, for instance:

- 1 indicates a perfect positive correlation.
- -1 indicates a perfect negative correlation.
- 0 indicates that there is no relationship between the different variables.

Values between -1 and 1 denote the strength of the correlation

How to Find the Correlation?

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
- x_i – the values of the x-variable in a sample
- $\bar{x}$ – the mean of the values of the x-variable
- y_i – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the y-variable

Steps in Correlation

In order to calculate the correlation coefficient using the formula above, you must undertake the following steps:

1. Obtain a data sample with the values of x-variable and y-variable.
2. Calculate the means (averages) $\overline{x}$ for the x-variable and $\bar{y}$ for the y-variable.
3. For the x-variable, subtract the mean from each value of the x-variable (let's call this new variable "a"). Do the same for the y-variable (let's call this variable "b").
4. Multiply each a-value by the corresponding b-value and find the sum of these multiplications (the final value is the numerator in the formula).
5. Square each a-value and calculate the sum of the result
6. Find the square root of the value obtained in the previous step (this is the denominator in the formula).
7. Divide the value obtained in ***step 4*** by the value obtained in ***step 7***.

Misinterpreting correlations

- Just about all the common problems that can render statistical analysis meaningless can occur with correlations.
- One example of a common problem is that with small samples, correlations can be unreliable.
- The smaller the sample size, the more likely we are to observe a correlation that is further from 0, even if the true correlation (obtained if we had data for the entire population) was 0.
- The standard way of quantifying this is to use *p-values.* In academic research, a common rule of thumb is that when *p* is greater than 0.05, the correlation should not be trusted.

Uses of Correlation

Prediction

- If there is a relationship between two variables, we can make predictions about one from another.

Validity

- Concurrent validity (correlation between a new measure and an established measure).

Reliability

- Test-retest reliability (are measures consistent).
- Inter-rater reliability (are observers consistent).

Theory verification

- Predictive validity.

Strengths of Correlations

1. Correlation allows the researcher to investigate naturally occurring variables that maybe unethical or impractical to test experimentally. For example, it would be unethical to conduct an experiment on whether smoking causes lung cancer.

2. Correlation allows the researcher to clearly and easily see if there is a relationship between variables. This can then be displayed in a graphical form.

Limitations of Correlations

1. Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.
For example suppose we found a positive correlation between watching violence on T.V. and violent behavior in adolescence. It could be that the cause of both these is a third (extraneous) variable - say for example, growing up in a violent home - and that both the watching of T.V. and the violent behavior are the outcome of this.

2. Correlation does not allow us to go beyond the data that is given. For example suppose it was found that there was an association between time spent on homework (1/2 hour to 3 hours) and number of G.C.S.E. passes (1 to 6). It would not be legitimate to infer from this that spending 6 hours on homework would be likely to generate 12 G.C.S.E. passes.

Different Methods of Correlation

Different Methods for Correlations

Correlations tests are arguably one of the most commonly used statistical procedures, and are used as a basis in many applications such as exploratory data analysis, structural modelling, data engineering etc. Many methods of of correlations are

- **Pearson’s correlation**
- **Spearman’s rank correlation**
- **Kendall’s rank correlation**
- **Biweight midcorrelation**
- **Distance correlation**
- **Percentage bend correlation**
- **Shepherd’s Pi correlation**
- **Blomqvist’s coefficient**
- **Hoeffding’s D**
- **Gamma correlation**
- **Gaussian rank correlation**
- **Point-Biserial and biserial correlation**
- **Winsorized correlation**
- **Polychoric correlation**
- **Tetrachoric correlation**
- **Partial correlation**
- **Multilevel correlation**

Pearson’s correlation

Pearson’s correlation: This is the most common correlation method. It corresponds to the covariance of the two variables normalized (i.e., divided) by the product of their standard deviations.

Spearman’s rank correlation:

A non-parametric measure of correlation, the Spearman correlation between two variables is equal to the Pearson correlation between the rank scores of those two variables; while Pearson’s correlation assesses linear relationships, Spearman’s correlation assesses monotonic relationships (whether linear or not). Confidence Intervals (CI) for Spearman’s correlations are computed using the Fieller, Hartley, and Pearson (1957) correction (see Bishara and Hittner 2017).

Kendall’s rank correlation:

In the normal case, the Kendall correlation is preferred to the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient. However, the interpretation of Kendall’s tau is less direct compared to that of the Spearman’s rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events. Confidence Intervals (CI) for Kendall’s correlations are computed using the Fieller, Hartley, and Pearson (1957) correction (see Bishara and Hittner 2017). For each pair of observations (i ,j) of two variables (x, y)

Biweight midcorrelation:

A measure of similarity that is median-based, instead of the traditional mean-based, thus being less sensitive to outliers. It can be used as a robust alternative to other similarity metrics, such as Pearson correlation (Langfelder and Horvath 2012).

Distance correlation:

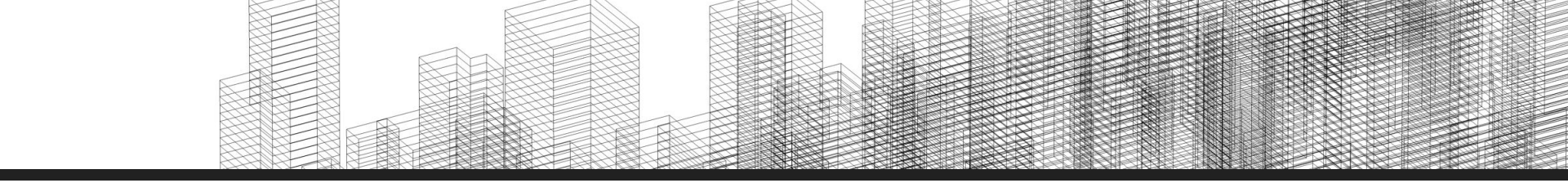
Distance correlation measures both linear and non-linear association between two random variables or random vectors. This is in contrast to Pearson’s correlation, which can only detect linear association between two random variables.

Percentage bend correlation:

Introduced by Wilcox (1994), it is based on a down-weight of a specified percentage of marginal observations deviating from the median (by default, 20 percent).

Shepherd’s Pi correlation

Equivalent to a Spearman’s rank correlation after outliers removal (by means of bootstrapped Mahalanobis distance).



Blomqvist’s coefficient:

The Blomqvist’s coefficient (also referred to as Blomqvist’s Beta or medial correlation; Blomqvist, 1950) is a median-based non-parametric correlation that has some advantages over measures such as Spearman’s or Kendall’s estimates (see Shmid and Schimdt, 2006).

Hoeffding’s D:

The Hoeffding’s D statistic is a non-parametric rank based measure of association that detects more general departures from independence (Hoeffding 1948), including non-linear associations. Hoeffding’s D varies between -0.5 and 1 (if there are no tied ranks, otherwise it can have lower values), with larger values indicating a stronger relationship between the variables.

Gamma correlation:

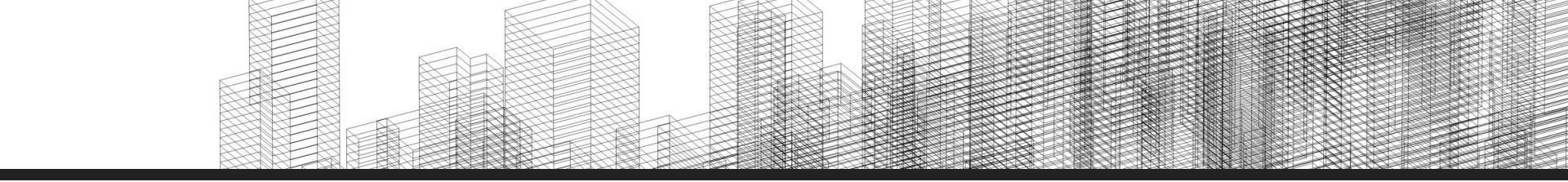
The Goodman-Kruskal gamma statistic is similar to Kendall’s Tau coefficient. It is relatively robust to outliers and deals well with data that have many ties.

Gaussian rank correlation:

The Gaussian rank correlation estimator is a simple and well-performing alternative for robust rank correlations (Boudt et al., 2012). It is based on the Gaussian quantiles of the ranks.

Point-Biserial and biserial correlation:

Correlation coefficient used when one variable is continuous and the other is dichotomous (binary). Point-Biserial is equivalent to a Pearson’s correlation, while Biserial should be used when the binary variable is assumed to have an underlying continuity. For example, anxiety level can be measured on a continuous scale, but can be classified dichotomously as high/low.



Winsorized correlation:

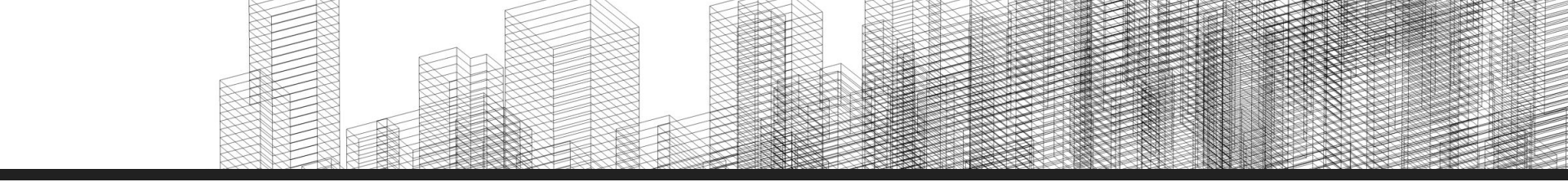
Correlation of variables that have been formerly Winsorized, i.e., transformed by limiting extreme values to reduce the effect of possibly spurious outliers.

Polychoric correlation:

Correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables.

Partial correlation:

Correlation between two variables after adjusting for the (linear) the effect of one or more variables. The correlation test is here run after having partialized the dataset, independently from it. In other words, it considers partialization as an independent step generating a different dataset, rather than belonging to the same model. This is why some discrepancies are to be expected for the *t*- and the *p*-values (but not the correlation coefficient) compared to other implementations



Multilevel correlation:

Multilevel correlations are a special case of partial correlations where the variable to be adjusted for is a factor and is included as a random effect in a mixed model.

References

1. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression
2. https://www.mathsisfun.com/data/correlation.html
3. https://www.displayr.com/what-is-correlation/
4. https://corporatefinanceinstitute.com/resources/knowledge/finance/correlation/
5. https://www.simplypsychology.org/correlation.html
6. https://byjus.com/maths/correlation/
7. https://easystats.github.io/correlation/articles/types.html