BIVARIATE DATA

7.1 Scatter plots: A **scatter plot**, **scatterplot**, or **scattergraph** is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data.

The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. This kind of plot is also called a *scatter chart*, *scatter gram*, *scatter diagram*, or *scatter graph*.

Overview[

A scatter plot is used when a variable exists that is below the control of the experimenter. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the *control parameter* or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval. For example, weight and height, weight would be on x axis and height would be on the y axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the pattern of dots slopes from lower left to upper right, it suggests a positive correlation between the variables being studied. If the pattern of dots slopes from upper left to lower right, it suggests a negative correlation.

A line of best fit (alternatively called 'trendline') can be drawn in order to study the correlation between the variables. An equation for the correlation between the variables can be determined by established best-fit procedures. For a linear correlation, the best-fit procedure is known as linear regression and is guaranteed to generate a correct solution in a finite time. No universal best-fit procedure is guaranteed to generate a correct solution for arbitrary relationships. A scatter plot is also very useful when we wish to see how two comparable data sets agree with each other. In this case, an identity line, *i.e.*, a $y=x$ line, or an 1:1 line, is often drawn as a reference. The more the two data sets agree, the more the scatters tend to concentrate in the vicinity of the identity line; if the two data sets are numerically identical, the scatters fall on the identity line exactly.

One of the most powerful aspects of a scatter plot, however, is its ability to show nonlinear relationships between variables. Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

The scatter diagram is one of the seven basic tools of quality control.

Example

For example, to display a link between a person's lung capacity, and how long that person could hold his/her breath, a researcher would choose a group of people to study, then measure each one's lung capacity (first variable) and how long that person could hold his/her breath (second variable). The researcher would then plot the data in a scatter plot, assigning "lung capacity" to the horizontal axis, and "time holding breath" to the vertical axis.

A person with a lung capacity of 400 cl who held his/her breath for 21.7 seconds would be represented by a single dot on the scatter plot at the point (400, 21.7) in the Cartesian coordinates. The scatter plot of all the people in the study would enable the researcher to obtain a visual comparison of the two variables in the data set, and will help to determine what kind of relationship there might be between the two variables

A **plot** is a graphical technique for representing a data set, usually as a graph showing the relationship between two or more variables. The plot can be drawn by hand or by a mechanical or electronic plotter. Graphs are a visual representation of the relationship between variables, very useful for humans who can quickly derive an understanding which would not come from lists of values. Graphs can also be used to read off the value of an unknown variable plotted as a function of a known one. Graphs of functions are used in mathematics, sciences, engineering, technology, finance, and other areas.

Overview

Plots play an important role in statistics and data analysis. The procedures here can broadly be split into two parts: quantitative and graphical. Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis. There are also many statistical tools generally referred to as graphical techniques. These include:

- scatter plots
- histograms
- probability plots
- residual plots
- box plots, and
- block plots

Graphical procedures such as plots are a short path to gaining insight into a data set in terms of testing assumptions, model selection, model validation, estimator selection, relationship identification, factor effect determination, outlier detection. Statistical graphics give insight into aspects of the underlying structure of the data.

Graphs can also be used to solve some mathematical equations, typically by finding where two plots intersect.

Types of plots

- Arrhenius plot : This plot displays the logarithm of a rate ($\ln(k)$, ordinate axis) plotted against inverse temperature ($1/T$, abscissa). Arrhenius plots are often used to analyze the effect of temperature on the rates of chemical reactions.
- Biplot : These are a type of graph used in statistics. A biplot allows information on both samples and variables of a data matrix to be displayed graphically. Samples are displayed as points while variables are displayed either as vectors, linear axes or nonlinear trajectories. In the case of categorical variables, category level points may be used to represent the levels of a categorical variable. A generalised biplot displays information on both continuous and categorical variables.
- Bland-Altman plot : In analytical chemistry and biostatistics this plot is a method of data plotting used in analysing the agreement between two different assays. It is identical to a Tukey mean-difference plot, which is what it is still known as in other fields, but was popularised in medical statistics by Bland and Altman.
- Bode plots are used in control theory.
- Box plot : In descriptive statistics, a box plot, also known as a box-and-whisker diagram or plot, is a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation,

lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation). A box plot may also indicate which observations, if any, might be considered outliers.

- Carpet plot : A two-dimensional plot that illustrates the interaction between two to three independent variables and one to three dependent variables.
- Contour plot : A two-dimensional plot which shows the one-dimensional curves, called contour lines on which the plotted quantity q is a constant. Optionally, the plotted values can be color-coded.
- Dalitz plot : This a scatter plot often used in particle physics to represent the relative frequency of various (kinematically distinct) manners in which the products of certain (otherwise similar) three-body decays may move apart

6.2 Correlation: In statistics, **dependence** is any statistical relationship between two random variables or two sets of data. **Correlation** refers to any of a broad class of statistical relationships involving dependence.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship (i.e., correlation does not imply causation).

Formally, *dependence* refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, *correlation* can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several **correlation coefficients**, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation – that is, more sensitive to nonlinear relationships. Mutual information can also be applied to measure dependence between two variables.

Rank correlation coefficients

Rank correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient (τ) measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. If, as the one variable increases, the other *decreases*, the rank correlation coefficients will be negative. It is common to regard these rank correlation coefficients as alternatives to Pearson's coefficient, used either to reduce the amount of calculation or to make the coefficient less sensitive to non-normality in distributions. However, this view has little mathematical basis, as rank correlation coefficients measure a different type of relationship than the Pearson product-moment correlation coefficient, and are best seen as measures of a different type of association, rather than as alternative measure of the population correlation coefficient.

To illustrate the nature of rank correlation, and its difference from linear correlation, consider the following four pairs of numbers (x, y):

(0, 1), (10, 100), (101, 500), (102, 2000).

As we go from each pair to the next pair x increases, and so does y. This relationship is perfect, in the sense that an increase in x is *always* accompanied by an increase in y. This means that we have a perfect rank correlation, and both Spearman's and Kendall's correlation coefficients are 1, whereas in this example Pearson product-moment correlation coefficient is 0.7544, indicating that the points are far from lying on a straight line. In the same way if y always *decreases* when x *increases*, the rank correlation coefficients will be −1, while the Pearson product-moment correlation coefficient may or may not be close to −1, depending on how close the points are to a straight line. Although in the extreme cases of perfect rank correlation the two coefficients are both equal (being both +1 or both −1) this is not in general so, and values of the two coefficients cannot meaningfully be compared. For example, for the three pairs (1, 1) (2, 3) (3, 2) Spearman's coefficient is 1/2, while Kendall's coefficient is 1/3.

Other measures of dependence among random variables

The information given by a correlation coefficient is not enough to define the dependence structure between random variables. The correlation coefficient completely defines the dependence structure only in very particular cases, for example when the distribution is a multivariate normal distribution. (See diagram above.) In the case of elliptical distributions it characterizes the (hyper-)ellipses of equal density, however, it does not completely characterize the

dependence structure (for example, a multivariate t-distribution's degrees of freedom determine the level of tail dependence).

Distance correlation and Brownian covariance / Brownian correlation[10][11] were introduced to address the deficiency of Pearson's correlation that it can be zero for dependent random variables; zero distance correlation and zero Brownian correlation imply independence.

The Randomized Dependence Coefficient is a computationally efficient, copula-based measure of dependence between multivariate random variables. RDC is invariant with respect to non-linear scalings of random variables, is capable of discovering a wide range of functional association patterns and takes value zero at independence.

The correlation ratio is able to detect almost any functional dependency and the entropy-based mutual information, total correlation and dual total correlation are capable of detecting even more general dependencies. These are sometimes referred to as multi-moment correlation measures, in comparison to those that consider only second moment (pairwise or quadratic) dependence.

The polychoric correlation is another correlation applied to ordinal data that aims to estimate the correlation between theorized latent variables.

One way to capture a more complete view of dependence structure is to consider a copula between them.

The coefficient of determination generalizes the correlation coefficient for relationships beyond simple linear regression.

Sensitivity to the data distribution

The degree of dependence between variables X and Y does not depend on the scale on which the variables are expressed. That is, if we are analyzing the relationship between X and Y, most correlation measures are unaffected by transforming X to $a + bX$ and Y to $c + dY$, where a, b, c, and d are constants. This is true of some correlation statistics as well as their population analogues. Some correlation statistics, such as the rank correlation coefficient, are also invariant to monotone transformations of the marginal distributions of X and/or Y.

Most correlation measures are sensitive to the manner in which X and Y are sampled. Dependencies tend to be stronger if viewed over a wider range of values. Thus, if we consider the correlation coefficient between the heights of fathers and their sons over all adult males, and compare it to the same correlation coefficient calculated when the fathers are selected to be between 165 cm and 170 cm in height, the correlation will be weaker in the latter case. Several techniques have

been developed that attempt to correct for range restriction in one or both variables, and are commonly used in meta-analysis; the most common are Thorndike's case II and case III equations.[13]

Various correlation measures in use may be undefined for certain joint distributions of *X* and *Y*. For example, the Pearson correlation coefficient is defined in terms of moments, and hence will be undefined if the moments are undefined. Measures of dependence based on quantiles are always defined. Sample-based statistics intended to estimate population measures of dependence may or may not have desirable statistical properties such as being unbiased, or asymptotically consistent, based on the spatial structure of the population from which the data were sampled.

Sensitivity to the data distribution can be used to an advantage. For example, scaled correlation is designed to use the sensitivity to the range in order to pick out correlations between fast components of time series.[14] By reducing the range of values in a controlled manner, the correlations on long time scale are filtered out and only the correlations on short time scales are revealed.