Topic 7: Correlations, Bivariate and Multivariate Analyses

7.1. Correlation Analysis

Introduction

There are many types of research questions that require us to understand the relationship between two variables. We might want to study how rainfall in a catchment basin is related to regional stream flow. We might want to study the relationship between some measure of soil quality and crop yield. We might want to know how large scale climate processes such as the El Nino-Southern Oscillation (ENSO) affect temperature and precipitation in remote regions of the globe. Or in a paleoceanographic study, we may be interested in one variable, (e.g. phosphate content of an ancient oceans), but may only be able to measure a related or proxy variable, such as the Cd content of shells from fossil benthic foraminifera. C*orrelational methods* can also be applied to observational data to explore the relationships between variables.

Learning outcomes

By completion of this section the learner should be able to:

a. Describe the basic concepts in correlation statistics
b. Identify different types of correlations
c. Apply correlation techniques in statistical analysis

Key Concepts

Description **-** by learning how two variables are related in a quantitative way we learn something about the processes that relate them.

Common variance **-** Variables that are correlated, covary. This explains how to determine how much of the variance in one variable in explained by its correlation to another variable.

Prediction **-** If a correlation is strong enough, we may be able to use it as a predictive statistical tool.

Linear vs. Non-Linear **-** a scatter plot of two linearly correlated variables will follow a straight line with variable degree of noise. In contrast, if two variables follow some arbitrary function they exhibit a non-linear correlation.

Positive vs. negative **-** If high values of one variable occur in conjunction with high values of another variable they are ***positively*** correlated. If high values on one variable occur with low values of another, they are said to be ***negatively*** or ***inversely*** correlated.

Orthogonal **-** two variables that are unrelated or uncorrelated are said to be orthogonal.

Strong vs. weak **-** if much of the variability is explained or shared between two variables, they are said to have a strong correlation. Weak correlations occur between variables that share little common variance.

7.1. 1. The Correlation Coefficient (r)

An important statistic for bivariate numerical data is the correlation coefficient. The sample correlation coefficient of two variables x and y is denoted r and is defined by either of the equivalent formulas. The correlation coefficient r for the data,

x	-2	-1	0	1	2
y	-4	1	0	1	4

is zero, even though x and y satisfy the equation, $y = x^2$.

There are a number of ways that have been devised to quantify the correlation between two variables. We will be primarily concerned with the simple linear correlation coefficient. This statistic is also referred to as the product moment correlation coefficient, or as Pearson's correlation coefficient. For simplicity, we will refer to it as the correlation coefficient. The sample correlation coefficient is denoted with the letter r, the true population correlation is denoted with the greek symbol, rho.

The linear correlation coefficient is defined as:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y} = \frac{\sum xy - n\overline{xy}}{(n-1)s_x s_y} = \frac{\Sigma Z_x Z_y}{(n-1)}$$

The associated degrees of freedom for the linear correlation coefficient is df = n-2.

There are two primary assumptions behind the use of the linear correlation coefficient:

1). The variables must be related in a linear way.

2). The variables must both be normal in distribution.

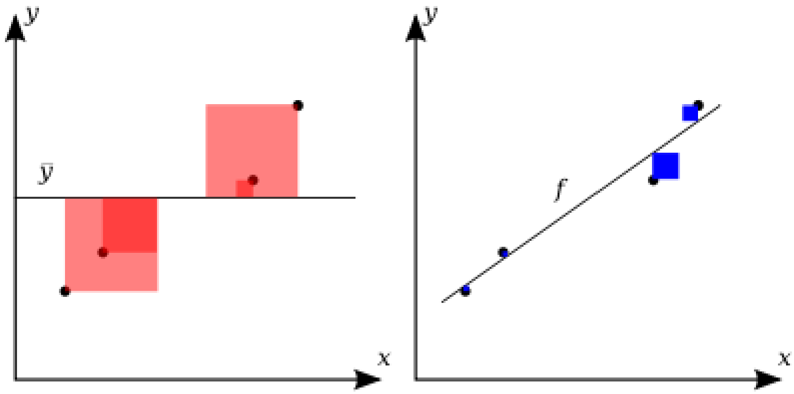
In fact, for the two variables to be correlated, they must exhibit a bivariate normal distribution. This second constraint makes sense when you consider the third way that r is written out above. It states that the r value is equal to the sum of the product of the observed z-scores for the variables x and y, normalized by n-1 observations. (This is also where the term product moment correlation coefficient arises - The mean is referred to as the first product moment, and the z-scores relate each observation to the mean)

7.1.2. Coefficient of Determination (R^2)

In statistics, the coefficient of determination, R^2, is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

There are several different definitions of R^2 which are only sometimes equivalent. One class of such cases includes that of linear regression. In this case, R^2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values, or in the case of simple linear regression, between the outcome and the values being used for prediction. In such cases, the values vary from 0 to 1. Important cases where the computational definition of R^2 can yield negative values, depending on the definition used, arise where the predictions which are being compared to the corresponding outcome have not derived from a model-fitting procedure using those data.

7.1.3. Definitions



$$R^2 = 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}$$

The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to one. The area of the blue squares represent the squared residuals with respect to the linear regression. The area of the red squares represent the squared residuals with respect to the average value.

A data set has values y_i, each of which has an associated modelled value f_i (also sometimes referred to as $\hat{y}_i$). Here, the values y_i are called the observed values and the modelled values f_i are sometimes called the predicted values.

The "variability" of the data set is measured through different Sums of squares

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$ the total sum of squares (proportional to the sample variance);

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$ the regression sum of squares, also called the explained sums of squares.

, the sum of squares of residuals, $SS_{\text{err}} = \sum_i (y_i - f_i)^2$ also called the residual sums of square

In the above $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n} \sum_i^n y_i$$ where n is the number of observations.

The notations SS_R and SS_E should be avoided, since in some texts their meaning is reversed to **R**esidual sum of squares and **E**xplained sum of squares, respectively.

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\rm err}}{SS_{\rm tot}}.$$

7.1.4. Coefficient of Determination in Relation to Variance

As unexplained variance

In a general form, R^2 can be seen to be related to the unexplained variance, since the second term compares the unexplained variance (variance of the model's errors) with the total variance (of the data).

As explained variance

In some cases the total sum of squares equals the sum of the two other sums of squares defined above,

$$SS_{\rm err} + SS_{\rm reg} = SS_{\rm tot}.$$

When this relation does hold, the above definition of R^2 is equivalent to

$$R^2 = \frac{SS_{\rm reg}}{SS_{\rm tot}}.$$

In this form R^2 is given directly in terms of the explained variace: it compares the explained variance (variance of the model's predictions) with the total variance (of the data).

This partition of the sum of squares holds for instance when the model values f_i have been obtained by linear regression. A milder sufficient conditions reads as follows: The model has the form

$$f_i = \alpha + \beta q_i$$

where the q_i are arbitrary values that may or may not depend on i or on other free parameters (the common choice $q_i = x_i$ is just one special case), and the coefficients α and β are obtained by minimizing the residual sum of squares.

This set of conditions is an important one and it has a number of implications for the properties of the fitted residuals and the modeled values. In particular, under these conditions:

$$\bar{f} = \bar{y}.$$

As squared correlation coefficient

Similarly, after least squares regression with a constant + linear model, R^2 equals the square of the correlation coefficient between the observed and modeled (predicted) data values.

Under general conditions, an R^2 value is sometimes calculated as the square of the correlation coefficient between the original and modeled data values. In this case, the value is not directly a measure of how good the modeled values are, but rather a measure of how good a predictor might be constructed from the modeled values (by creating a revised predictor of the form α + βf_i). According to Everitt (2002, p. 78), this usage is specifically the definition of the term "coefficient of determination": the square of the correlation between two (general) variables.

7.1.5. Interpretation of Coefficient of Determination

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data.

Values of R^2 outside the range 0 to 1 can occur where it is used to measure the agreement between observed and modeled values and where the "modeled" values are not obtained by linear regression and depending on which formulation of R^2 is used. If the first formula above is used, values can never be greater than one. If the second expression is used, there are no constraints on the values obtainable.

In many (but not all) instances where R^2 is used, the predictors are calculated by ordinary least-squares regression: that is, by minimizing SS_{err}. In this case R-squared increases as we increase the number of variables in the model (R^2 will not decrease). This illustrates a drawback to one possible use of R^2, where one might try to include more variables in the model until "there is no more improvement". This leads to the alternative approach of looking at the adjusted R^2. The explanation of this statistic is almost the same as R^2 but it penalizes the statistic as extra variables are included in the model. For cases other than fitting by ordinary least squares, the R^2 statistic can be calculated as above and may still be a useful measure. If fitting is by weighted least squares or generalized least squares, alternative versions of R^2 can be calculated appropriate to those statistical frameworks, while the "raw" R^2 may still be useful if it is more easily interpreted. Values for R^2 can be calculated for any type of predictive model, which need not have a statistical basis.

R^2 does *not* tell whether:

- the independent variables are a true cause of the changes in the dependent variable;
- Omitted-variable bias exists;
- The correct regression was used;
- The most appropriate set of independent variables has been chosen;
- There is collinearity present in the data on the explanatory variables;
- The model might be improved by using transformed versions of the existing set of independent variables.

In a Linear Model, consider a linear model of the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i,$$

where, for the *i*th case, Y_i is the response variable, $X_{i,1},\ldots,X_{i,p}$are *p* regressors, and $\varepsilon_i$is a mean zero error term. The quantities $\beta_0,\ldots,\beta_p$ are unknown coefficients, whose values are determined by least squares. The coefficient of determination R^2 is a measure of the global fit of the model. Specifically, R^2 is an element of [0, 1] and represents the proportion of variability in Y_i that may be attributed to some linear combination of the regressors in *X*.

R^2 is often interpreted as the proportion of response variation "explained" by the regressors in the model. Thus, $R^2 = 1$ indicates that the fitted model explains all variability in *y*, while $R^2 = 0$ indicates no 'linear' relationship (for straight line regression, this means that the straight line model is a constant line (slope=0, intercept=$\bar{y}$) between the response variable and regressors. An interior value such as $R^2 = 0.7$ may be interpreted as follows: "Approximately seventy percent of the variation in the response variable can be explained by the explanatory variable. The remaining thirty percent can be explained by unknown, inherent variability."

A caution that applies to R^2, as to other statistical descriptions of correlation and association is that "correlation does not imply causation." In other words, while correlations may provide valuable clues regarding causal relationships among variables, a high correlation between two variables does not represent adequate evidence that changing one variable has resulted, or may result, from changes of other variables.

In case of a single regressor, fitted by least squares, R^2 is the square of the Pearson product-moment correlation coefficient relating the regressor and the response variable. More generally, R^2 is the square of the correlation between the constructed predictor and the response variable.

Inflation of coefficient of determination

In least squares regression, R^2 is weakly increasing in the number of regressors in the model. As such, R^2 alone cannot be used as a meaningful comparison of models with different numbers of independent variables. For a meaningful comparison between two models, an F-test can be performed on the residual of squares

To demonstrate this property, first recall that the objective of least squares regression is:

$$\min_b SS_{\text{err}}(b) \Rightarrow \min_b \sum_i (y_i - X_i b)^2$$

The optimal value of the objective is weakly smaller as additional columns of *X* are added, by the fact that relatively unconstrained minimization leads to a solution which is weakly smaller than relatively constrained minimization. Given the previous conclusion and noting that SS_{tot} depends only on *y*, the non-decreasing property of R^2 follows directly from the definition above.

The intuitive reason that using an additional explanatory variable cannot lower the R^2 is this: Minimizing SS_{err} is equivalent to maximizing R^2. When the extra variable is included, the data always have the option of giving it an estimated coefficient of zero, leaving the predicted values and the R^2 unchanged. The only way that the optimization problem will give a non-zero coefficient is if doing so improves the R^2.

7.1.6. Adjusted Coefficient of Determination

Adjusted R^2 (sometimes written as $\bar{R}^2$) is a modification of R^2 that adjusts for the number of explanatory terms in a model. Unlike R^2, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative, and will always be less than or equal to R^2. The adjusted R^2 is defined as

$$1-(1-R^2)\frac{n-1}{n-p-1}=1-\frac{SS_{\text{err}}}{SS_{\text{tot}}}\frac{df_t}{df_e}$$

where p is the total number of regressors in the linear model (but not counting the constant term), and n is sample size.

The principle behind the Adjusted R^2 statistic can be seen by rewriting the ordinary R^2 as

$$R^2=1-\frac{VAR_{\text{err}}}{VAR_{\text{tot}}}$$

where $VAR_{\text{err}} = SS_{\text{err}} / n$ and $VAR_{\text{tot}} = SS_{\text{tot}} / n$ are estimates of the variances of the errors and of the observations, respectively. These estimates are replaced by notionally "unbiased" versions: $VAR_{\text{err}} = SS_{\text{err}} / (n - p - 1)$ and $VAR_{\text{tot}} = SS_{\text{tot}} / (n - 1)$.

Adjusted R^2 *does not have the same interpretation as* R^2. As such, care must be taken in interpreting and reporting this statistic. Adjusted R^2 is particularly useful in the Feature selection stage of model building. Adjusted R^2 is not always *better* than R^2: adjusted R^2 will be more useful only if the R^2 is calculated based on a sample, not the entire population. For example, if our unit analysis is a state, and we have data for all counties, then adjusted R^2 will not yield any more useful information than R^2. The use of an adjusted R^2 is an attempt to take account of the phenomenon of statistical shrinkage.

7.1.7. Generalized Coefficient of Determination

Nagelkerke (1991) generalizes the definition of the coefficient of determination:

1. A generalized coefficient of determination should be consistent with the classical coefficient of determination when both can be computed;
2. Its value should also be maximised by the maximum likelihood estimation of a model;
3. It should be, at least asymptotically, independent of the sample size;
4. Its interpretation should be the proportion of the variation explained by the model;
5. It should be between 0 and 1, with 0 denoting that model does not explain any variation and 1 denoting that it perfectly explains the observed variation;
6. It should not have any unit.

7.2. Partial Correlation

In probability theory and statistics, **partial correlation** measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

7.2.1. Formal definition

Formally, the partial correlation between X and Y given a set of n controlling variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$, written $\rho_{XY\cdot\mathbf{Z}}$, is the correlation between the residual R_X and R_Y resulting from the linear regression of X with $\mathbf{Z}$ and of Y with $\mathbf{Z}$, respectively. In fact, the first-order partial correlation is nothing else than a difference between a correlation and the product of the removable correlations divided by the product of the coefficients of alienation of the removable correlations.

7.2.2. Computation

Using linear regression

A simple way to compute the partial correlation for some data is to solve the two associated linear regression problems, get the residuals, and calculate the correlation between the residuals. If we write x_i, y_i and $\mathbf{z}_i$ to denote i.i.d samples of some joint probability distribution over X, Y and $\mathbf{Z}$, solving the linear regression problem amounts to finding

$$\mathbf{w}_X^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$
$$\mathbf{w}_Y^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

with N being the number of samples and $\langle \mathbf{v}, \mathbf{w} \rangle$ the scalar product between the vectors **v** and **w**. The residuals are then

$$r_{X,i} = x_i - \langle \mathbf{w}_X^*, \mathbf{z}_i \rangle$$
$$r_{Y,i} = y_i - \langle \mathbf{w}_Y^*, \mathbf{z}_i \rangle$$

and the sample partial correlation is

$$\hat{\rho}_{XY\cdot\mathbf{Z}} = \frac{N\sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum r_{Y,i}}{\sqrt{N\sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i}\right)^2}\ \sqrt{N\sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i}\right)^2}}.$$

Using recursive formula

It can be computationally expensive to solve the linear regression problems. Actually, the *n*th-order partial correlation (i.e., with |**Z**| = *n*) can be easily computed from three (*n* - 1)th-order partial correlations. The zeroth-order partial correlation $\rho_{XY\cdot\text{Ø}}$ is defined to be the regular correlation coefficient ρ_{XY}.

It holds, for any $Z_0 \in \mathbf{Z}$:

$$\rho_{XY\cdot\mathbf{Z}} = \frac{\rho_{XY\cdot\mathbf{Z}\setminus\{Z_0\}} - \rho_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}\rho_{YZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}\sqrt{1-\rho^2_{YZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}}.$$

This computation as a recursive algorithm yields an exponential time complexity. However, this computation has the overlapping subproblem property, such that using dynamic programming or simply caching the results of the recursive calls yields a complexity of $\mathcal{O}(n^3)$.

Note in the case where Z is a single variable, this reduces to:

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho^2_{XZ}}\sqrt{1-\rho^2_{YZ}}}.$$

Using matrix inversion

In $\mathcal{O}(n^3)$time, another approach allows *all* partial correlations to be computed between any two variables X_i and X_j of a set **V** of cardinality *n*, given all others, i.e., $\mathbf{V} \setminus \{X_i, X_j\}$, if the correlation matrix (or alternatively covariance matrix) **Ω** = (ω_{ij}), where $\omega_{ij} = \rho_{XiXj}$, is invertible[] . If we define **P** = $\mathbf{\Omega}^{-1}$, we have:

$$\rho_{X_iX_j\cdot\mathbf{V}\setminus\{X_i,X_j\}} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}.$$

As conditional independence test

With the assumption that all involved variables are multivariate Gaussian, the partial correlation $\rho_{XY\cdot\mathbf{Z}}$ is zero if and only if *X* is conditions independent from *Y* given **Z**. This property does not hold in the general case.

To test if a sample partial correlation $\hat{\rho}_{XY\cdot\mathbf{Z}}$vanishes, Fisher's *z-transform of the partial correlation* can be used:

$$z(\hat{\rho}_{XY\cdot\mathbf{Z}}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_{XY\cdot\mathbf{Z}}}{1-\hat{\rho}_{XY\cdot\mathbf{Z}}}\right).$$

The null hypothesis is $H_0 : \hat{\rho}_{XY\cdot\mathbf{Z}} = 0$, to be tested against the two-tail alternative $H_A : \hat{\rho}_{XY\cdot\mathbf{Z}} \neq 0$. We reject H_0 with significant level α if:

$$\sqrt{N - |\mathbf{Z}| - 3} \cdot |z(\hat{\rho}_{XY\cdot\mathbf{Z}})| > \Phi^{-1}(1-\alpha/2),$$

where Φ(·) is the cumulative distribution function of a Gaussian distribution with zero mean and unit standard deviation, and *N* is the sample SIZE. Note that this *z*-transform is approximate and that the actual distribution of the sample (partial) correlation coefficient is not straightforward. However, an exact t-test based on a combination of the partial regression coefficient, the partial correlation coefficient and the partial variances is available.

Summary

Correlation is a way to measure how associated or related two variables are. The researcher looks at things that already exist and determines if and in what way those things are related to each other. The purpose of doing correlations is to allow us to make a prediction about one variable based on what we know about another variable. For example, there is a correlation between income and education. We find that people with higher income have more years of education. Partial correlation is a procedure that allows us to measure the region of three-way overlap precisely, and then to remove it from the picture in order to determine what the correlation between any two of the variables would be (hypothetically) if they were not each correlated with the third variable. Alternatively, you can say that partial correlation allows us to determine what the correlation between any two of the variables would be (hypothetically) if the third variable were held constant.

Learning Activities

One

a. Prepare the dataset provided for correlation analyses
b. Use the dataset to carry out bivariate and partial correlation analyses in SPSS, SAS or Genstat
c. Compare the outputs from the different software
d. Interpret the outputs

Two

a. Sample research reports where correlation and partial correlation analyses have been used, then do the following:
b. Identify the hypotheses being tested
c. Identify the conclusions made
d. Identify the type of types of data subjected to correlation and partial correlation analyses

7.3. Bivariate analysis

Introduction

This method is most useful when two different variables work together to affect the acceptability of a process or part thereof. Correlation, a measure of association often ranges between –1 and 1 is commonly used in bivariate analysis. Where the sign of the integer represents the "direction" of correlation (negative or positive relationships) and the distance away from 0 represents the degree or extent of correlation – the farther the number away from 0, the higher or "more perfect" the relationship is between the independent and dependent variations.

Statistical significance relates to the generalizability of the relationship and, more importantly, the likelihood the observed relationship occurred by chance. Often significance levels, when n (total number of cases in a sample) is large, can approach .001 (only 1/1000 times will the observed association occur). Measures of association and statistical significance that are used vary by the level of measurement of the variables analyzed.

Learning Objectives

By completion of this topic the learner should be able to:

- Describe basic concepts of bivariate analysis
- Explain the steps of bivariate analysis
- Carry out analysis of biviate data

Key Concepts

Scatter plots
Contingency table

7.3.1. Steps in Bivariate Analysis

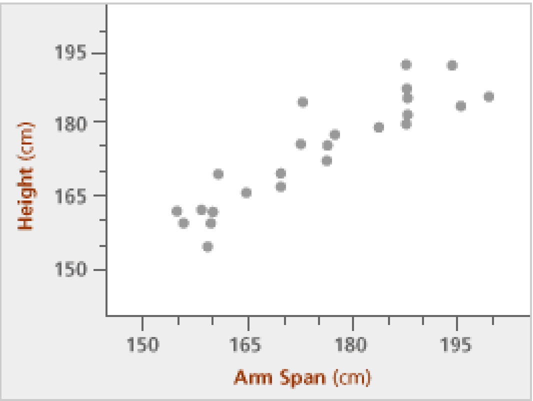
- Discern how a variable is distributed among the cases in one group, and then, discerning how that variable is distributed among the cases in another group,
- Deciding *whether* the two distributions *differ* from each other, in *which ways* they differ (if any),

- Deciding what it is about the two groups that could "account for" the difference in the variable's distribution – that is, *theorizing*.

When a data file contains many variables, there are often several pairs of variables to which bivariate methods may productively be applied. The most common methods include contingency tables, scatterplots, least squares lines, and correlation coefficients.

Example a bivariate data that can be used for scatter diagram

Person #	Arm Span	Height
1	156	162
2	157	160
3	159	162
4	160	155
5	161	160
6	161	162
7	162	170
8	165	166
9	170	170
10	170	167
11	173	185
12	173	176
13	177	173
14	177	176
15	178	178
16	184	180
17	188	188
18	188	187
19	188	182
20	188	181
21	188	192
22	194	193
23	196	184
24	200	186



7.3.2. Bivariate Descriptive

Bivariate data is data that occupies two columns of a data file and comes from two variables. Bivariate descriptive are tables, visual displays, or statistics that reveal or measure some aspect of the relationship between two variables. When a data file contains many variables, there are often several pairs of variables to which bivariate methods may productively be applied. In this section we will consider contingency tables, scatterplots, least squares lines, and correlation coefficients.

Case Study

Two-Way Frequency Tables and Scatterplots

We begin with an example of a data file with several variables. Consider the following scenario. The visitors to a certain Nairobi Animal Orphane zoo, purchase tickets upon arrival at the zoo on Sunday . As they leave the zoo later in the day, their tickets are collected and the variables, age, gender, arrival time, and departure time are recorded. Age is recorded in years and arrival and departure times are recorded to the nearest minute. Zoo visitors usually come in groups: couples, families, one adult supervising several children, etc. Solitary visitors can be considered to be groups of size 1. In the present study, this group phenomenon is recorded by means of a grouping variable called, appropriately, group. The values of the grouping variable are positive integers giving the order in which the groups arrived. Presumably the order of arrival is not as important to us as recording who was grouped with whom, and that is the main information the grouping variable allows us to preserve. Since each person in a group is given the same group number, that number shows who was in which group.

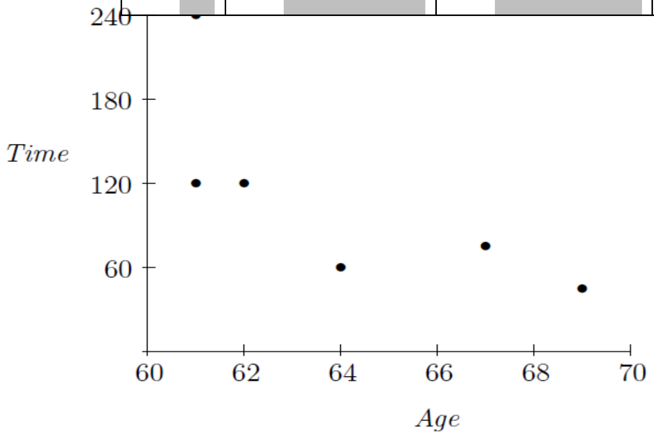
The first few lines of the data file might look something like the following:

Name	Group	Age	Gender	arrival	depart
Ken	001	52	M	1:05	2:15
Terry	001	75	F	1:05	2:15
William	002	45	M	1:15	2:35
Emily	002	36	F	1:15	2:35
Steve	003	43	M	1:15	2:45
Lorna	004	36	F	1:12	3:42
Susan	004	8	F	1:12	3:42
Mary	004	8	F	1:12	3:42
Julie	004	7	F	1:11	3:43
Ann	004	8	F	1:12	3:43
Shame	005	19	M	1:24	3:39
Chris	005	22	M	1:24	3.39

One way to explore the relationship between these two variables is to examine a bivariate frequency table. If we group age values into several intervals, and consider the entire data file, not just the few lines reproduced above, we will obtain the following kind of table:

Age

Gender		Under 18	[18,40]	[41,60]	Over 60
	M	241	109	97	6
	F	237	385	103	11



A bivariate frequency table like this one is also called a contingency table. The present table would be called a 2 by 4 contingency table to express the fact that the body of the table consists of two rows and four columns. This gives 8 cells in the body of the table. An m by n contingency table will have mn cells. The numbers in the individual cells are frequencies. Thus we see that there were 385 females in the age range 18 to 40. A contingency table can be very powerful at revealing relationships between variables. In the present example we see at once that in the age range [18,40], females outnumber males more than three to one, and that this does not come close to happening for the other age ranges.

Summary

Bivariate descriptive statistics involves simultaneously analyzing (comparing) two variables to determine if there is a relationship between the variables. Generally by convention, the independent variable is represented by the columns and the dependent variable is represented by the rows. As in the case of a Univaraite Distribution, we need to construct the frequency distribution for bivariate data. Such a distribution takes into account the classification in respect of both variables simultaneously. Scatter plot is used to look at pattern in bivariate data. These patterns are described in provisions of linearity, slope, and strength. Linearity defined as whether a records pattern is linear (instantly) or nonlinear (rounded). Slope refers to the way of alteration in variable Y when variable X gets large

Learning Activities

One

a. Prepare the dataset provided for bivariate analyses

b. Use the dataset to carry out bivariate analyses in SPSS, SAS or Genstat

c. Compare the outputs from the different software

d. Interpret the outputs

Two

a. Sample research reports where bivariate analyses have been used, then do the following:

b. Identify the hypotheses being tested

c. Identify the conclusions made

d. Identify the type of types of data subjected to bivariate analyses

REFERENCES

1. Kim H. Esbensen. Multivariate Data Analysis in Practice: An Introduction to Multivariate Data Analysis and Experimental Design (5th Edition).
2. Mardia, KV., JT Kent, and JM. Bibby. (1979). Multivariate Analysis. Academic Press.
3. Gerry Quinn and Michael Keough. (2002). Experimental Design and Data Analysis for Biologists. Cambridge University Press.

Useful Links:

www.tutorvista.com/topic/**bivariate**. Accessed on 24th June 2011

References

1. Baba, Kunihiro; Ritei Shibata & Masaaki Sibuya (2004). "Partial correlation and conditional correlation as measures of conditional independence". *Australian and New Zealand Journal of Statistics* 46 (4): 657–664.
2. Fisher, R.A. (1924). "The distribution of the partial correlation coefficient". *Metron* **3** (3–4): 329–332. *Regression and Analysis of Variance*. McGraw-Hill.
3. Guilford J. P., Fruchter B. (1973). *Fundamental statistics in psychology and education*. Tokyo: MacGraw-Hill Kogakusha, LTD.
4. Kendall MG, Stuart A. (1973) *The Advanced Theory of Statistics*, Volume 2 (3rd Edition), ISBN 0-85264-215-6, Section 27.22
5. Nagelkerke, "A Note on a General Definition of the Coefficient of Determination," Biometrika, vol. 78, no. 3, pp. 691–692, 1991

6. Rummel, R. J. (1976). Understanding Correlation.
7. Steel, R. G. D. and Torrie, J. H., *Principles and Procedures of Statistics,* New York: McGraw-Hill, 1960, pp. 187, 287.

External Links

http://digital.library.adelaide.edu.au/dspace/handle/2440/15182.

http://www.hawaii.edu/powerkills/UC.HTM.

7.4. Multivariate Analysis (MANOVA)

Introduction

Multivariate analysis is a form of quantitative analysis which examines three or more variables at the same time, in order to understand the relationships among them. The simplest form of multivariate analysis is one in which the researcher, interested in the relationship between an independent variable and a dependent variable (eg: gender and political attitudes), introduces an extraneous variable (eg: age) to ensure that a correlation between the two main variables is not spurious.

Learning objectives

a. By completion of this topic the learner should be able to:
b. Describe basic concepts of multivariate analysis
c. Apply multivariate analysis to a dataset
d. Explain the principles behind discriminant function analysis

Key Concepts

- MANOVA,
- Discriminant analysis,
- Spartial data analysis

7.4.1. General Principles of Multivariate Analysis

Multivariate analysis of variance (MANOVA) is simply an extension of the univariate Analysis of variance. In analysis of variance, we examine the one metric dependent variable with the grouping independent variable. Analysis of variance fails to compare the group when the dependent variables become more than one dependent metric variable. To account for multiple dependent variables, MANOVA bundles them together into a weighted linear combination or composite variable. These linear combinations are also called canonical variates, roots, Eigenvalues, vectors or discriminant functions.

Once the dependent variable combines into a canonical variate, MANOVA can be performed, such as univariate ANOVA. Now, MANOVA will compare whether or not the independent variable group differs from the newly created group. In this way, MANOVA essentially tests whether or not the independent grouping variable explains a significant amount of variance in the canonical variate.

8.2. Assumptions in MANOVA

1). Independent Random Sampling: MANOVA normally assumes that the observations are independent of one another. There is not any pattern in MANOVA for the selection of the sample. The sample is completely random.

2). Level and Measurement of the Variables: MANOVA assumes that the independent variables are categorical in nature and the dependent variables are continuous variables. MANOVA also assumes that homogeneity is present between the variables that are taken for covariates.

3). Linearity of dependent variable: In MANOVA, the dependent variables can be correlated to each other, or may be independent of each other. Study shows that in MANOVA, a moderately correlated dependent variable is preferred. In MANOVA, if the dependent variables are independent of each other, then we have to sacrifice the degrees of freedom and it will decrease the power of the analysis.

4). Multivariate Normality: MANOVA is very sensitive with outliers and missing value. Thus, it is assumed that multivariate normality is present in the data.

5). Multivariate Homogeneity of Variance: Like test analysis of variance, MANOVA also assumes that the variance between groups is equal.

Key concepts

Power: Power shows the probability of correctly accepting the null hypothesis.

Post hoc test: In MANOVA, when there is a significant difference between groups, then the post hoc test is performed to know the exact group means, which significantly differ from each other.

Significance: Like ANOVA, probability value is used to make statistical decisions as to whether or not the group means are equal, or if they differ from each other.

Multivariate F-statistics: F- statistics is simply derived by dividing the means sum of the square for the source variable by the source variable mean error.

Comparison between ANOVA and MANOVA:

Computation of MANOVA is more complex compared to the ANOVA. In ANOVA, we compute univariate F statistic but in MANOVA, we compute multivariate F statistics. In ANOVA, we compare grouping independent variables with one dependent variable, but in MANOVA, we compare many dependent variables with the grouping variable.

Summary

The term “multivariate statistics” is appropriately used to include all statistics where there are more than two variables simultaneously analyzed. Because MANOVA is used when there are two or more dependent variables. It can be used to test whether changes in the independent variable(s) have significant effects on the dependent variables. It also helps identify any interactions among the dependent and independent variables

Learning Activities

One

a. Prepare the dataset provided for multivariate analyses
b. Use the dataset to carry out multivariate analyses in SPSS, SAS or Genstat
c. Compare the outputs from the different software

d. Interpret the outputs

Two

a. Sample research reports where multivariate analyses have been used, then do the following:

b. Identify the hypotheses being tested

Identify the conclusions made

c. Identify the type of types of data subjected to multivariate analyses

REFERENCES

1. Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance*. Newbury Park, CA: Sage Publications.
2. de Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika, 53*(4), 437-454.
3. Gill, J. (2001). *Generalized Linear Models: A Unified Approach*. Thousand Oaks, CA: Sage Publications.
4. Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures*. London: Chapman and Hall.
5. Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses *Psychological Bulletin, 105*(2), 302-308.
6. Huynh, H., & Mandeville, G. K. (1979). Validity conditions in a repeated measures design. *Psychological Bulletin, 86*(5), 964-973.
7. Olson, C. L. (1976). On choosing a test statistic in multivariate analyses of variance. *Psychological Bulletin, 83*(4), 579-586.
8. Powell, R. S., & Lane, D. M. (1979). CANCOR: A general least-squares program for univariate and multivariate analysis of variance and covariance. *Behavior Research Methods & Instrumentation, 11*(1), 87-89.
9. Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333-343.
10. Smith, H. F. (1958). A multivariate analysis of covariance. *Biometrics, 14,* 107-127.

External Links

http://www.statsoft.com. Accessed 25th June, 2011

http://www2.chass.ncsu.edu. Accessed 25th June, 2011

http://www.visualstatistics.net. Accessed 25th June, 2011