**12. Analysing Quantitative Data**

Quantitative data are gathered based on the positivist research philosophy and they are numerical in nature. For those who have adopted the interpretivist philosophy may also use quantitative data in combination with qualitative data. For the data to be useful they need to be analysed and interpreted. It needs quantitative analysis to assist this process. This can range from creating simple tables or diagrams that show the frequency of occurrence through establishing statistical relationships between variable to complex statistical modelling. Computer has been used to analyse the data.

The analysis of quantitative data involves a process consisting of the following stages:

1. preparing your data for analysis;
2. summarizing and presenting your data using tables and graphs;
3. describing your data using suitable statistical methods; and
4. examining relationships and trends between variables.

***1. Preparing data for analysis***

Considerations must be given before obtaining the data for quantitative analysis:

* type of data (level of numerical measurement);
* format in which your data will be input to the analysis software;
* impact of data coding on subsequent analyses (for different data types);
* need to weight cases;
* methods you intend to use to check data for errors.
* Data types - Two distinct groups as shown in the diagram below:

Data

Categorical Quantifiable

Descriptive Ranked Continuous Discrete

Increasing precision

1. Categorical data – refer to data whose values cannot be measured numerically but can be either classified into sets according to characteristics or placed in rank order.

The categorical data are subdivided into (a) descriptive (nominal) data and (b) rank (ordinal) data.

1. *Descriptive (nominal) data* cannot be measured numerically or to rank it. However, it is possible to count them and to establish which category has the most and whether they are spread evenly between categories.
2. *Ranked (or ordinal) data* are more precise as it is possible to know the position of each case within the data set.
3. Quantifiable data – refer to those values that can be actually measured numerically as quantities and they are subdivided into two groups:
4. Continuous data – refer to those data whose values can be theoretically take any value i.e. can be measured accurately e.g. temperature, delivery distance and length of service and these are continuous data.
5. Discrete data – can be measured precisely. Each case takes a finite number of values from a scale that measures changes in discrete units. E.g. the size of a pair of shoes (a non-integer) or a finite number such as the number of mobile phones produced or customers served.

Note: The more precise the level of measurement the greater the range of analytical techniques can be used. E.g. data that have been collected and coded using precise numerical measurements can also be regrouped to a less precise level where they can also be analysed. E.g. a student’s score in a test could be recorded as the actual mark (discrete data) or as the position in their class (ranked data). By contrast, less precise data cannot be made more precise. The implication is if you are not sure of the level of precise you require, then it is better for you to collect data at the highest level possible and to regroup them if necessary.

**Data Layout**

It is possible to use computer-aid programmes to assist in the collection of data such as computer-aided personal interviewing (CAPI), computer-aided telephone interviewing (CATI) and on-line questionnaire can automatically enter and save data to a computer file at the time of collection. Certain survey design and analysis software can help to bring about analysis and integration of the data. It is then possible to specify a data layout compatible with the analysis software. Other data collection methods require the entry of the data and then to proceed to analyse them. In this situation the data layout needs to meet the requirements of the analysis software.

Virtually all analysis software will accept the data if they are entered in table format. This is called a *data matrix*.

A simple data matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | id | age | gender | service | employed |
| Case 1 | 1 | 27 | 1 | 2 | 1 |
| Case 2 | 2 | 19 | 2 | 1 | 2 |
| Case 3 | 3 | 24 | 2 | 3 | 1 |

Within the data matrix, each column usually represents a single *variable* for the data obtained. Each matrix row contains the variables for an individual case. E.g. if the data have been obtained using a survey, each row will contain the data from one survey form.

On the other hand, for longitudinal data such as a company’s share price over time, each row (case) might be a different time period.

In the above table, the first variable (id) is the *survey form identifier*. This means it is possible to link data for each case in the matrix to the survey form when error checking.

The second variable (age) contains quantifiable data, the age of each respondent (case) at the time of the survey.

Subsequent variables contain the remaining data: the third (gender) records this descriptive data using code 1 for male and 2 for female; the fourth (service) records each case’s length of service to the nearest year with their most recent employer.

The final variable (employed) records whether each case is (code 1) or is not (code 2) currently in employment. Codes can have different meaning for different variables.

If you intend to enter data into a spreadsheet, the first variable is in column A, the second in column B and so on. Each cell in the first row (1) should contain a short variable name to enable you to identify each variable. Subsequent rows (2 onwards) will each contain the data for one case. Statistical analysis software follows the same logic, although the variable names are usually displayed ‘above’ the first row, as in the above table.

A worked example

Data Input

As part of a market research interview survey you need to discover which of four products (tomato, ketchup, brown sauce, soy sauce, vinegar) have been purchased within the last month by consumers. You therefore need to collect four data items from each respondent:

* Tomato ketchup purchased within the last month? Yes/No
* Brown sauce purchased within the last month? Yes/No
* Soy sauce purchased within last month? Yes/No
* Vinegar purchased within last month? Yes/No

Each of these data items is a separate variable. However, the data are collected using one question:

1. Which of the following items have you purchased within the last month?

Item purchased not purchased not sure

Tomato ketchup 1 2 3

Brown sauce 1 2 3

Soy sauce 1 2 3

Vinegar 1 2 3

The data collected from each respondent will form separate variables in the data matrix using numerical codes (1=purchased, 2= not purchased, 3= not sure). This is known as multiple-dichotomy coding:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | tomato | brown | soy | vinegar |
| respondent | 1 | 1 | 1 | 2 |

Question 2 (below) could theoretically have millions of possible responses for each of the ‘things’. The number that each respondent mentions way also vary,. Our experience suggests that virtually all respondent will select five or less. Space therefore has to be left to code up to five responses after data have been collected.

For office use only

1. List up to five things you like ………………………………

about your current job ……………………………...

……………………………..

……………………………..

……………………………

……………………………….

The multi-dichotomy method uses a separate variable for each different answer. For

Question 2 a separate variable could be used for each ‘thing’’ listed: e.g. salary, location, colleagues, hours, holidays, car and so on. Subsequently it is possible to code each variable as ‘listed’ or ‘not listed’ for each case. The alternative, the multiple-response method, uses the same number of variables as the maximum number of different responses from any one case. For question 2 these might be named ‘liked 1’, ‘like2, ‘liked3’, ‘liked4’, and ‘liked5’. Each of these variables would use the same codes and could include any of the responses as a category. Statistical analysis software often contains special multiple-response procedures to analyse such data.

**Coding**

All data types should, with few exceptions, be recorded using numerical codes. This enables the data to be quickly entered with fewer errors. It also makes subsequent analyses, in particular those that require re-coding of data to create new variables, more straightforward. Unfortunately, analyses of limited meaning are also easier, such as calculating a mean (average) gender from codes 1 and 2! A common exception to using a numerical code for categorical data is where a postcode is used as the code for a geographical reference. If a spreadsheet is used, a list of codes for each variable needs to be kept. Statistical analysis software can store these so that each code is automatically labelled.

*Coding quantifiable data*

Actual numbers are often used as codes for quantifiable data. Once the data have been entered as a matrix it is possible to use analysis software to group or combine data to form additional variables with less detailed categories. The process is referred to as ***re-coding***.

E.g. A person’s salary could be coded to the nearest dollar and entered into the matrix as 23453 (discrete data). Later, re-coding could be used to place it in a group of similar salaries, from $20,000 to $24,999 (categorical data).

*Coding categorical data*

Codes are also applied to categorical data.

Existing coding schemes can be used for many variables. These include industrial classification, occupation, and socioeconomic classification as using them would:

* Save time;
* Are normally well tested;
* Allow comparisons of the results with other (often larger) surveys.

These codes should include on the data collection form as pre-set codes (e.g. questionnaires) unless it is decided to use another code. The important point is that it is possible to compare the data with those already collected e.g. data collected in the earlier questionnaires.

*Coding at data collection* occurs when there is a limited range of well-established categories into which the data can be placed. These are included on the data collection form and the person filling in the form selects the correct category.

*Coding after data collection* is necessary when it is unclear of the likely responses or there are a large number of possible responses, it is better to wait until data from the first 50 to 100 cases are available and then develop the coding scheme. This is called codebook. While the data collection methods are being designed, it is essential to be clear about the intended analyses and in particular ***the level of precision required*** and ***the coding schemes used by surveys with which comparisons are to be made.***

To create the codebook for each variable it is necessary:

1. to examine the data and establish broad groupings;

2. to subdivide the broad groupings into increasingly specific subgroups dependent on the

intended analyses;

3. to allocate codes to all categories at the most precise level of detail required;

4. to note the actual responses that are allocated to each category and produce a codebook;

5. to ensure that those categories that may need to be aggregated together are given adjacent

codes to facilitate re-coding.

Coding missing data

Each variable for each case in the data set should have a code, even if no data has been collected. A missing data code is used to indicate why data are missing. deVaus (2002) identifies four main reasons for missing data:

* The data were not required form the respondent, perhaps because of a skip generated by a filter question in a survey.
* The respondent refused to answer the question (a non-response).
* The respondent did not know the answer or did not have an opinion. Sometimes this is treated as implying an answer; on other occasions it is treated as missing data.
* The respondent may have missed a question by mistake, or the respondent’s answer may be unclear.
* leaving part of a question in a survey blank implies an answer; in such cases the data are not classified as missing.

Weighing cases

Most data a researcher uses will be a sample.

In stratified random sampling the researcher may have used a different sampling fraction for each stratum (a group of persons performing a certain function such as the management group i.e. the managers, middle managers or technical supervisors). Or the researcher may have obtained a different response rate for each of the strata. To obtain an accurate overall picture it is necessary to take into account these differences in response rates between strata. A common method of achieving this is to use cases from those strata that have lower proportions of responses to represent more than one case in your analysis. Most statistical analysis software allows a person to do this by weighing cases. To weight the cases the person:

1. calculate the percentage of the population responding for each stratum;

2. calculate which stratum had the highest percentage of the population responding;

3. calculate the weight for each stratum using the formula

highest proportion of population responding for any stratum

Weight =

proportion of population responding in stratum for which calculating weight

4. apply the appropriate weight to each case.

An example: Weighting cases

To select your sample for a survey you have used stratified random sampling. The

percentage of each stratum’s population that responded is given below:

Upper stratum: 90%

Lower stratum: 65%

To account for the differences in the response rates between strata you decide to weight

the cases prior to analysis.

The weight for the upper stratum is: 90/90 = 1

This means that each case in the upper stratum will count as 1 case in your analysis.

The weight for the lower stratum is: 90/65 = 1.38

This means that each case in the lower stratum will count for 1.38 cases in your analysis.

You enter these as a separate variable in your data set and use the statistical analysis

software to apply the weights.

Checking for errors

Some errors will occur no matter how carefully you code and subsequently enter data . The main methods to check data for errors are as follows:

* Look for illegitimate codes. In any coding scheme, only certain numbers are allocated. Other numbers are therefore errors. Common errors are the inclusion of betters O and o instead of zero, letters l or 1 instead of 1, and number 7 instead of 1.
* Look for illogical relationships. E.g. if a person is coded to the professional socioeconomic group and their social class is unskilled manual an error has occurred.
* Check that rules in filter questions are followed. Certain responses to filter questions mean that other variables should be coded as missing values. If this has not happened there has been an error.

For each possible error there is a need to discover whether it occurred at coding or data entry and then correct it.

Data checking is very time consuming and so is often not undertaken. For not doing it is a very dangerous thing and can result in incorrect results from which false conclusions are drawn.

***2.* *Exploring and Presenting Data***

Once the data have been entered and checked for errors, the analysis can begin.

Exploratory data analysis approach is useful in these initial stages. It emphasises the use of diagrams to explore and understand the data. Always keep in mind the research questions and objectives when exploring the data. Exploratory data analysis approach looks for relationships in the data. Each diagram or table must be clearly labelled and structured.

A summary checklist of the points to remember when designing a diagram or table:

Checklist for diagrams and tables

For both diagrams and tables:

* Does it have a brief but clear and descriptive title?
* Are the units of measurement used clearly stated?
* Are the sources of data used clearly stated?
* Are there notes to explain abbreviations and unusual terminology?
* Does it state the size of the sample on which the values in the table are based?

For diagrams

* Does it have clear axis labels?
* Are bars and their component in the same logical sequence?
* Is more dense shading used for smaller areas?
* Is a key or legend included (where necessary)?

For tables

* Does it have clear column and row headings?
* Are columns and rows in a logical sequence?

It is best to begin exploratory analysis by looking at individual variables and their components. The key aspects to consider will be guided by the research questions and objectives. and to include:

* specific values;
* highest and lowest values;
* trends over time;
* proportions
* distributions.

After completing the exploratory analysis of the data, it moves to begin comparing and looking for relationships between variables, considering in addition:

* conjunctions (the point where values for two or more variables intersect).
* totals
* Interdependence and relationships.

These are summarised in a table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Categorical | | Quantifiable | |
|  | Descriptive | Ranked | Continuous | Discrete |
| To show one variable so that any specific value can be read easily | Table/frequency distribution (data often grouped) | | | |
| To show the frequency of occurrence of categories or values for one variable so that highest and lowest are clear | Bar chart (data may need grouping) | | Histogram or frequency polygon (data must be grouped) | Bar chart of pictogram (data may need grouping |
| To show the trend for a variable |  | Line graph or bar chart | Line graph or histogram | Line graph or bar chart |
| To show the proportion of occurrences of categories or values for one variable | Pie chart or bar chart (data may need grouping) | | Histogram or pie chart (data must be grouped | Pie or bar chart (data may need grouping |
| To show the distribution of values for one variable |  | | Frequency polygon, histogram (data must be grouped) or box plot | Frequency polygon, bar chart (data may need grouping) or box plot |
| To show the interdependence between two or more variables so that any specific value can be read easily | Contingency table/cross-tabulation (data often grouped) | | | |
| To compare the frequency of occurrence of categories or values for two or more variables so that highest and lowest are clear | Multiple bar chart (continuous data must be grouped, other data may need grouping) | | | |
| To compare the trends for two or more variables so that conjunctions are clear |  | Multiple line graph or multiple bar chart | | |
| To compare the proportions of occurrences of categories or values for two or more variables. | Comparative pie charts or percentage component bar chart (continuous data must be grouped, other fata may need grouping) | | | |
| To compare the distribution of values for two or more variables |  | | Multiple box plot | |
| To compare the frequency of occurrence of categories or values for two or more variables so that totals are clear. | Stacked bar chart (continuous data must be grouped, other data may need grouping) | | | |
| To compare the proportions and totals of occurrences of categories or values for two or more variables | Comparative proportional pie charts (continuous data must be grouped, other data may need grouping) | | | |
| To show the relationship between cases for two variables |  | Scatter graph/scatter plot | | |

**Exploring and presenting individual variables**

*To show specific values*

Using a table (frequency distribution) is the simplest way of summarising data for individual variables so that specific values can be read.

For descriptive data, the table summarises the number of cases (frequency) in each category.

For variables where a large number of categories (or values for quantifiable data), there is a need to group the data into categories that reflect the research question(s) and objectives.

*To show highest and lowest values*

As a table, there is no visual significance to highest or lowest values unless emphasized by some other forms. Diagrams are used to provide visual clues. For categorical and discrete data, bar charts and pictograms are used. However, bar charts provide a more accurate representation. Bar charts are sometimes known as column charts

For continuous data, a histogram is used. In a histogram the area of each bar represents the frequency of occurrence and the continuous nature of the data is emphasised by the absence of gaps between the bars.

Figure 1 An example of a stacked bar/column chart

Frequency

Amounts spent per visit in $

Figure 2: An example of a histogram

Figure 3: An example of frequency polygons/line graph

*To show the trend*

Trends can only be presented for variables containing quantifiable (and occasionally ranked) longitudinal data. The most suitable diagram for exploring the trend is a line graph. In which the data values for each time period are joined with a lone to represent the trend.(See above figure the line graph. Bar charts can also be used to show trends between discrete time periods and histograms for continuous time periods.

To show proportions

Research has shown that the most frequently used diagram to emphasise the proportion or share occurrences is the pie chart. A pie chart is divided into proportional segments according to the share each has of the total value.

Figure 4: An example of a Pie Chart

*To show the distribution of value*

This can be done by plotting either a frequency polygon or a histogram for continuous data or a frequency polygon or bar chart for discrete data. If the diagram shows a bunching to the left and a long tail to the right the data is positively skewed. If the converse is true, the data are negatively skewed.

If the data are equally distributed either side of the highest frequency then they are symmetrically distributed. A special form of the symmetric distribution, in which the data can be plotted as a bell-shaped curve, is known as the normal distribution.

Figure 5: Example of a bell-shaped curve

The box plot is a more advanced statistical analysis. This diagram provides a pictorial representation of the distribution of the data for a variable. The plot shows where the middle value or median is, how this relates to the middle 50% of the data or inter-quartile range, and highest and lowest values or extremes.

This represents the middle value or median (c. 16600)

This represents the This represents the

This represents the lower value of the upper value of the This represents the

lowest value or inter-quartile range inter-quartile range highest value or

extreme (c.11200) (c. 13600) (c. 22200) extreme (c. 25600)

10 15 20 25

Sales in $’’000

This represents the middle 50% or

Inter-quartile range of the data (c. 8000)

This represents the full range of the data (c. 14000)

Annotated sketch of box plot

Figure 6: An example of a box ploy

**Comparing variables**

*To show specific values and interdependence*

The best method is a table known as the contingency table or cross-tabulation as shown in the example below. Such a table enables the examination of interdependence between the variables. For variables where there are likely to be a large number of categories (or values for quantifiable data), there is a need to group the data to prevent the table from becoming too large.

Table 1: Contingency table: number of insurance claims by gender, 2002

|  |  |  |  |
| --- | --- | --- | --- |
| Number of claims | Male | Female | Total |
| 0  1  2  3 | 10032  2156  120  13 | 13478  1430  25  4 | 23500  3586  145  17 |
| Total | 12321 | 14937 | 27258 |

*To compare highest and lowest values*

The best approach is to use the multiple bar chart (also known as the compound bar chart).

See example given below.

Figure 7: An example of a multiple bar chart

*To compare proportions*

Comparison of proportions between variables uses either a percentage compound bar chart or two or more pie charts.

Figure 8: An example of a percentage compound bar chart

*To compare trends and conjunctions*

The most suitable diagram to compare trends for two or more quantifiable (or occasionally ranked) variable is a multiple line graph where one line represents each variable or use multiple bar charts in which bars for the same time period are placed adjacent.

Looking for conjunctions in the trends - that is, where values for two or more variables intersect - this is where the lines on a multiple line graph cross.

E.g. Breakeven analysis chart.

*To compare totals*

Comparison of totals between variables uses a variation of the bar chart. A stacked bar chart can be used for all data types provide that continuous data and data where there are more than six possible values or categories are grouped.

Figure 9: An example of a stacked bar chart

*To compare proportions and totals*

To compare both proportions of each category or value and the totals for two or more variables it is best to use *comparative proportional pie charts* for all types of data.

Figure 10: an example of a comparative proportional pie chart

*To compare the distribution of values*

Often it is used to compare the distribution of values for two or more variables. Plotting multiple frequency polygons or bar charts will enable the comparison of distribution for up to three or four variables (see figures 3 and 7). An alternative is to use a diagram of multiple box plots (see figure 6).

*To show the relationship between cases for variables*

To explore possible relationships between ranked and quantifiable data variables, it is done by plotting one variable against another. This is called scatter graph or scatter plot.

Convention dictates that the dependent variable i.e. the variable that changes in response to changes in the other (independent) variable - on the vertical axis.

Figure 11: An example of scatter plots.

The strength of the relationship is indicated by the closeness of the points to an imaginary line. If, as the values for one variable increase, so do those for the other, then a positive relationship occurs. If, as the values for one variable decreased, those for the other variable increased, then a negative relationship is indicated.

Furthermore the strength of this relationship can be assessed statistically using techniques such as correlation or regression.

***3. Describing Data using Statistics***

In the exploratory data analysis approach the emphasis is on the use of diagrams in order to understand the data collected.

On the other hand, descriptive statistics enable enables a person to describe and compare variables numerically. The research question(s) and objectives should guide the choice of statistics. Statistics focus on two aspects to describe a variable:

1) the central tendency

2) the dispersion.

These are summarised in table 2.

Table 2: Descriptive statistics by data type: a summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| To calculate a measure of: | | Categorical | | Quantifiable | |
| Descriptive | Ranked | Continuous | Discrete |
| Central tendency that …… | … represents the value that occurs most frequently | Mode | | | |
| … represents the middle value |  | | Median | |
| … includes all data values (average) |  | | Mean | |
| Dispersion that …… | … states the difference between the highest and lowest values |  | | Range (data need not be normally distributed but must be placed in rank order) | |
| … states the difference within the middle 50% of values |  | | Inter-quartile range (data need not be normally distributed but must be placed in rank order) | |
| … states the difference within another fraction of the values |  | | Deciles or percentiles (data need not be normally distributed but must be placed in rank order) | |
| … describes the extent to which data values differ from the mean |  | | Variance, or more usually, the standard deviation (data should be normally distributed) | |
| … compares the extent to which data values differ from the mean between variables |  | | Coefficient of variation (data should be normally distributed) | |

*Describing the central tendency*

There are three ways of measuring the central tendency:

* Value that occurs most frequently (mode)
* Middle value or mid-point after the data have been ranked (median)
* Value, often known as the average, that includes all data values in its calculation (mean)

The mode is the value that occurs most frequently. For descriptive data, the mode is the only measure of central tendency that can be sensibly interpreted. E.g. the most common colour of motor cars sold last year was red or that the most common makes were Ford and Vauxhall.

Median value is obtained by ranking all the values in ascending order and finding the mid-point (or 50th percentile) in the distribution. For variables that have an even number of data values the median will occur halfway between the two middle data values. The median has the advantage that it is not affected by extreme values in the distribution.

Mean or average, is the arithmetic mean. It is defined for a set of n observations as the sum of all the values divided by n. The mean \ is indispensable in statistics and easy to work with mathematically, but it is not ideal for all situations. If there are values a long way apart from the others, then the median may be preferred.

*Describing the dispersion*

It is important to describe how the data values are dispersed around the central tendency. This is possible only for quantitative data. Two of the most frequently used ways of describing the dispersion are the:

1. difference within the middle 50% of values (inter-quartile range);
2. extent to which values differ from the mean (standard deviation).

Although these measures of dispersion are suitable only for quantitative data most statistical analysis software will also calculate them for categorical data if numerical codes are used.

*To state the difference between values*

It is possible to calculate the difference between the lowest and the highest values i.e. the range. However in statistics this is rarely used. A most frequently used statistics, is the inter-quartile range. The range is divided into 4 equal sections called quartiles. The lower quartile is the value below which is a quarter of the data values will fall, the upper quartile is the value above which a quarter of the data values will fall. The remaining half of the data values will fall between the lower and upper quartiles. The difference between the upper and lower quartiles is the inter-quartile range. The concern is the middle 50% of data values and ignores extreme values.

*To describe and compare the extent by which values differ from the mean*

It is important to look at the extreme to which the data values for a variable are spread around their means, as this is what is needed to know in order to assess its usefulness as a typical value for the distribution. If the data values are very close to the mean, then the mean is more typical than if they vary widely. To describe the extent of spread of quantifiable data it is to use the ***standard of deviation***. (If your data are a sample, this is calculated using a slightly different formula than if your data are a population, although if your sample is larger than about ***30 cases*** there is little difference in the two situations (Morris, 1999).

If there is a need to compare the relative spread of data between distributions of different magnitudes (e.g. one may be measured in hundreds of tonnes, the other in billions of tonnes). To make a meaningful comparison then a common way of doing this is:

1. to divide the standard deviation by the mean;

2. then to multiply the answer by 100.

This result in statistic called the ***coefficient of variation***. The values of this statistics can then be compared. The distribution with the largest coefficient of variation has the largest relative spread of data.

A worked example

A bank collects data on the total value of transactions at each of its main and sub-branches. The mean value of total transactions at the main branches is 5 times as high as that for the sub-branches. This makes it difficult to compare the relative spread in total value of transactions between the two types of branches. Calculating the coefficients of variation reveals that there is relatively more variation in the total value of transactions at the main branches than the sub-branches:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Branch type Mean total Standard deviation Coefficient of variation

transaction value\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Main $6 000 000 $1 417 000 23.62

Sub $1 200 000 $217 000 18.08\_\_\_\_\_\_\_\_\_\_

This is because the coefficient of variation for the main braches is larger (23.62) than the coefficient for the sub-branches (18.08).

***4. Examining relationships, differences and trends using statistics***

Question: How does a variable relate to another variable? In statistic the test is known as ***statistical significance***. This is testing the likelihood of the relationship occurring by change alone, if there was no difference in the population from which the sample was drawn. The way in which this significance is tested can be thought of as answering one from a series of questions, dependent on the data type:

* Is the association statistically significant?
* Are the differences statistically significant?
* What is the strength of the relationship, and is it statistically significant?
* Are the predicted values statistically significance?

These are summarised in table 3 below along with statistics used to help examine trends.

Table 3: Statistics to examine relationships, differences and trends by data type: a summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Categorical | | Quantifiable | |
|  | Descriptive | Ranked | Continuous | Discrete |
| To test whether two variables are associated | Chi square (data may need grouping) | | Chi square if variables grouped into discrete classes | |
| To test whether two groups (categories) are different | Kolmogorov-Smirnov (data may need grouping) | | Independent t-test or paired t test (often used to test for changes over time) | |
| To test whether three or more groups (categories) are different |  | | Analysis of variance (ANOVA) | |
| To assess the strength of relationship between two variables | Spearman’s rank  correlation  coefficient | | Pearson’s product moment correlation coefficient (PMCC) | |
| To assess the strength of a relationship between one dependent and one or more independent variables |  | | Regression coefficient | |
| To predict the value of a dependent variable from one or more independent variables |  | | Regression coefficient | |
| To compare relative changes over time |  | | Index number | |
| To determine the trend over time of a series of data |  | | Times series: moving averages Regression equation | |

*Testing for significant relationships and differences*

Testing the probability of a relationship between variables occurring by chance alone if there really was no difference in the population from which that sample was drawn is known as ***significant testing***.

If the probability of the test statistic or one more extreme having occurred by chance alone is very low (usually p = 0.05 or lower), then it is a statistically significant relationship. If the probability of obtaining the test statistic or one more extreme by chance alone is higher than 0.05, then it can conclude that the relationship is not statistically significant. There may still be a relationship between the variables under such circumstances, but you cannot make the conclusion with any certainty.

The statistical significance of the relationship indicated by the test statistic is determined in part by the sample size. One consequence of this is that it is very difficult to obtain a significant test statistic with a small sample. Conversely, because the impact of sample size declines rapidly for samples over about 30 cases, if an extremely large sample is chosen it is relatively easy to obtain a significant test statistic for a relationship that is in reality not that obvious.

*Type I and Type II errors*

Errors can occur when making inferences from samples. Statisticians refer to these as Type I and Type II errors. An error made by wrongly coming to a decision that something is true when in reality it is not, is known as a ***Type I error***. This could be coming to a conclusion that the two variables are related when in actual fact they are not.

A ***Type II error*** involves the opposite occurring i.e. a conclusion that something is not true, when in reality it is. For example, you conclude that the two variables are not related when they are.

Researchers consider Type I error to be more serious than Type II error. It is more important to minimise Type I error than Type II error.

*To test whether two variables are associated*

Often descriptive or quantifiable data will be summarised as a two-way contingence (see Table 1). The ***chi square test*** enables the determination of how likely that the two variables are associated. It is based on a comparison of the observed values in the table with what might be expected if the two distributions were entirely independent. It is in fact, assessing the likelihood of the data in the table occurring by chance alone by comparing it with what would be expected if the two variables were independent of each other.

The test relies on:

* + the categories used in the contingency table being mutually exclusive, so that each observation falls into only one category or class interval;
  + no more than 20% of the cells in the table having expected values of less than 5. For contingency tables of two rows and two columns no expected values of less than 10 are preferable.

If the latter assumption is not met, the accepted solution is to combine rows and columns.

The chi square test calculates the probability that the data in your table could occur by change alone.

*To test whether two groups are different*

Categorical data

Sometimes it is necessary to see whether the distribution of an observed set of values for each category of a variable differs from a specified distribution, e.g. whether your sample differs from the population from which it was selected. The Kolmogrov-Smirnov test enables you to establish this. It is based on a comparison of the cumulative proportions in the same categories for the specified population. Therefore you are testing the likelihood of the distribution of your observed data differing from that of the specified population by chance alone.

Quantifiable data

If a quantifiable variable can be divided into two distinct groups using a descriptive variable you can assess the likelihood of these groups being different using an ***independent groups t-test.*** This compares the difference in the means of the two groups using a measure of the spread of the scores. If the likelihood of any difference between these two groups occurring by chance alone is low this will be represented by a large *t* statistic with a probability less than 0.05. This is termed statistically significant.

*To test whether three or more groups are different*

If a quantifiable variable is divided into three or more distinct groups using a descriptive variable, you can assess the likelihood of these groups being different occurring by chance alone by using one-way analysis or one-way ANOVA.

ANOVA analyses the variation within and between groups of data by comparing means. The *F*-ratio or *F* statistic represents these differences. If the likelihood of any difference between groups occurring by chance alone is low, this will be represented by a large *F* ratio with a probability of less than 0.05. This is termed statistically significant.

Hays (1994) lists the following assumptions that need to be met before using one way ANOVA:

* Each data value is independent and does not relate to any of the other data values. This means that you should not use one-way ANOVA where data values are related in some way, such as the same person being tested repeatedly.
* The data for each group are normally distributed. This assumption is not particularly important provided that the number of cases in each group is large.
* The data for each group have the same variance (standard deviation squared). However, provided that the number of cases in the largest group is not more than 1.5 times that of the smallest group, this appears to have very little effect on the test results.

*Assessing the strength of relationship*

As part of your exploratory data analysis you will already have plotted the relationship between cases for two ranked or quantifiable variables using a scatter graph. Such relationships might include those between weekly sales of a new product and those of a similar established product, or age of employees and their length of service with the company. These examples emphasise the fact that your data can contain two sorts of relationship:

* those where a change in one variable is accompanied by a change in another variable but it is not clear which variable caused the other to change, a correlation;
* those where a change in one or more (independent) variables causes a change in another (dependent) variable, a *cause-and-effect relationship*.

*To assess the strength of relationship between pairs of variables*

A correlation coefficient enables you to quantify the strength of the relationship between two ranked or quantifiable variables. This coefficient can take on any value between -1 to +1. A value of +1 represents a perfect positive correlation. This means the two variables are precisely related and that, as values of one variable increase, values of the other variable will increase. By contrast a value of -1 represents a perfect negative correlation. This means that the two variables are precisely related, however, as the values of one variable increase those of the other decrease. Correlation coefficients between +1 and -1 represent weaker positive and negative correlations, a value of 0 meaning the variables are perfectly independent. Within business research it is extremely unusual to obtain perfect correlation.

If your variables contain quantifiable data you should use **Pearson’s product moment correlation coefficient (PMCC)** to assess the strength of relationship. However, if one of your variables contains rank data you will need to rank the other variable and use **Spearman’s rank correlation coefficient**. Although this uses a different formula to calculate the correlation coefficient it is interpreted in the same way.

*To assess the strength of a cause-and-effect relationship between variables*

In contrast to the correlation coefficient the ***regression coefficient (also known as coefficient of determination)*** enables you to assess the strength of relationship between a quantifiable dependent variable and one or more quantifiable independent variables. For a dependent variable and one (or perhaps two) independent variables you will have probably already plotted this relationship on a scatter graph. If you have more than two independent variables this is unlikely as it is very difficult to represent four or more scatter graph axes visually.

The regression coefficient (represented by r2) can take on any value between 0 and +1. It measures the proportion of the variation in a dependent variable (amount of sales) that can be explained statistically by the independent variable (marketing expenditure) or variables (marketing expenditure, number of sales, staff etc). This means that if all the variation in amount of sales can be explained by the marketing expenditure and the number of sales staff the regression coefficient will be 1. If 50% of the variation can be explained the regression coefficient be will be 0.5, and if none of the variation can be explained the coefficient will be 0. Within our research we have rarely obtained a regression coefficient above 0.8.

The process of calculating a regression coefficient and regression equation using one independent variable is normally termed ***regression analysis***. Calculating the regression coefficient and regression equation using two or more independent variables is termed ***multiple regression analysis***.

*To predict the value of a variable from one or more other variables*

Regression analysis can also be used to predict the values of a dependent variable given the values of one or more independent variables by calculating a regression equation. You may wish to predict the amount of sales for a specified marketing expenditure and number of sales staff. This can be presented in a regression equation:

Amount of sales = a + (b1 x marketing expenditure) + (b2 x number of sales staff

Using regression analysis you would calculate the values of a, b1 and b2 from data you had already collected on amount of sales, marketing expenditure and number of sales staff. A specified marketing expenditure and number of sales staff could then be substituted into the regression equation to predict the amount of sales that would be generated.

If your equation is a perfect predictor then the regression coefficient will be 1. If the equation can predict only 50% of the variation then the regression coefficient will be 0.5, and if the equation predicts none of the variation the coefficient will be 0.

*Examining trends*

When examining longitudinal data the first thing is to draw a line graph to obtain a visual represent of the trend. Subsequent to this, statistical analysis can be undertaken. Two of the more common uses of such analyses are:

* to compare trends for variables measured in different units or of different magnitudes;
* to determine the long-term trend and forecast future values for a variable.

*To compare trends*

To answer some research question (s) and to meet some objectives you may need to compare trends between two or more variables measured in different unit or at different magnitudes. To compare changes in prices of fuel oil and coal over time is difficult as the prices are recorded for different units (litres and tonnes). One way of overcoming this is to use index numbers and compare the relative changes in prices rather than actual figures. Index numbers are also widely used in business publication and by organisations.

Although each indices can involve quite complex calculations they all compare change over time against a base period. The base period is normally given the value of 100 (or 1000 in the case of many share indices), and change is calculated relative to this. Thus a value greater than 100 would represent an increase relative to the base period, and a value less than 100 a decrease.

To calculate simple index numbers for each case of a longitudinal variable you use the following formulae:

data value for case

Index number for case = ---------------------------------- x 100

data value for base period

Thus if a company’s sales were 125 000 units in 1999 (base period) and 150 000 units in 2000 the index number for 1999 would be 100 and for 2000 it would be 120.

To determine the trend and forecasting

The trend can be estimated by drawing a freehand line through the data on a line graph. However, these data are often subject to variations such as seasonal variations and so this method is not very accurate. A straight forward way of overcoming this is to calculate a moving average for the time series of data values. Calculating a ***moving average*** involves replacing each value in the time series with the mean of that value and those values directly preceding and following it. This smooths out the variation in the data so that you can see the trend more clearly. The calculation of a moving average is relatively straightforward using either a spreadsheet or statistical analysis software.

Once the trend has been established it is possible to forecast future values by continuing the trend forward for time periods for which data have not been collected. This involves calculating the long-term i.e. the amount by which values are changing each time period after variations have been smoothed out. Once again this is relatively straightforward to calculate using analysis software. Forecasting can also be undertaken using other statistical methods including regression.