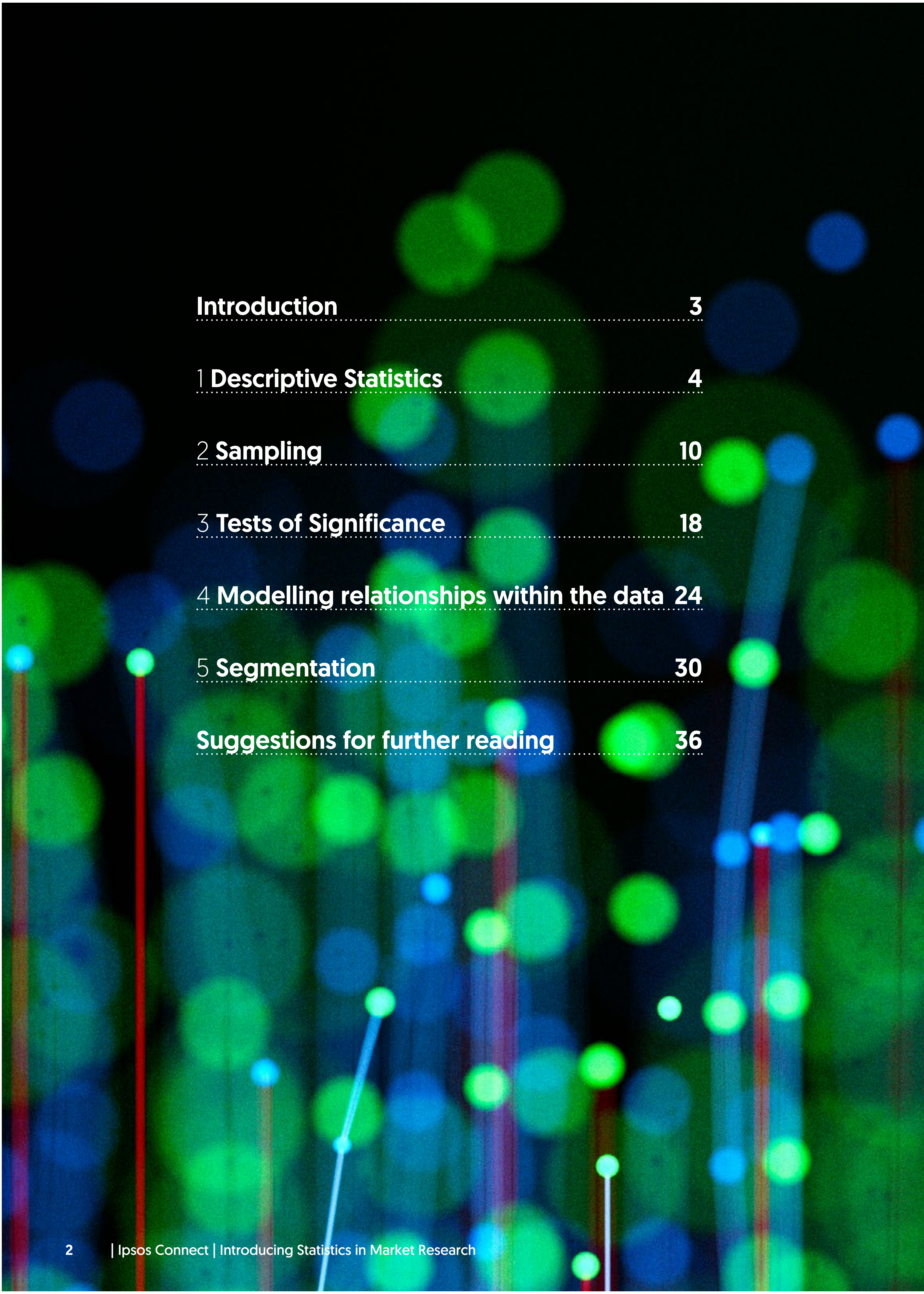
Ipsos Connect

Introducing

Statistics in Market Research

Second Edition

Prepared by
Leo Cremonezi
Statistical Scientist
January 2018



INTRODUCTION

Market research relies heavily on stats techniques in order to bring more insights to the usual deliverables and outputs. Analysing the collected data with basics tools is a fundamental aspect but sometimes a statistical methodology can answer the client's question in a better way.

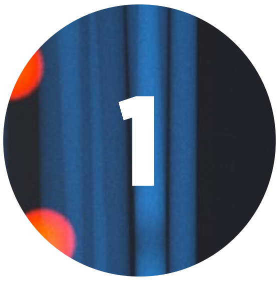
Statistical techniques can be employed in almost all areas of life to draw inference about populations. In the context of market research the researcher samples customers from populations of consumers in order to establish what they think of particular products and services, or to identify purchasing behaviour so as to predict future preferences or buying habits. The information gathered in these surveys can then be used to draw inference about the wider population with a certain level of statistical confidence that the results are accurate.

A necessary prerequisite to conducting a survey, and subsequently to drawing inference about a population, is to decide upon the best method of data collection. Data collection encompasses the fundamental areas of survey design and sampling. These are key elements in the statistical process, a poorly designed survey and an inadequate sample may lead to biased or misleading results which in turn will lead the researcher to draw incorrect inference.

Analysing the collected data is another fundamental aspect and can include any number of statistical techniques. For the newcomer a broad understanding of numerical data and an ability to interpret graphical and numerical descriptive measures is an important starting point for becoming proficient at data collection, analysis and interpretation of results.

The aim of this document is to provide a broad overview of survey design, sampling and statistical techniques commonly used in a market research environment to draw inference from survey data. ●

DESCRIPTIVE STATISTICS



1.1 Data types

Data can be broken down into two broad categories: qualitative and quantitative. Qualitative observations are those which are not characterised by a numerical quantity. Quantitative observations comprise observations which have numeric quantities.

Qualitative data arises when the observations fall into separate distinct categories. The observations are inherently discrete, in that there are a finite number of possible categories into which each observation may fall. Examples of qualitative data include gender (male, female), hair colour, nationality and social class.

Quantitative or numerical data arise when the observations are counts or measurements. The data can be either discrete or continuous. Discrete measures are integers and the possible values which they can take are distinct and separated. They are usually counts, such as the number of people in a household, or some artificial grade, such as the assessment of a number of flavours of ice cream. Continuous measures can take on any value within a given range, for example weight, height, miles per hour.

Within these two broad categories there are three basic data types:

- Nominal (classification); gender
- Ordinal (ranked); social class
- Measured (uninterrupted, interval); weight, height

Qualitative observations can be either nominal or ordinal whilst quantitative data can take any form.

Table 1 lists 20 beers together with a number of variables common to them all. Beer type is qualitative information which does not fall into any of the 3 basic data types. However, the beers could be coded as light and non-light, this would generate a nominal variable in the data set. The remaining variables in the data set are all measures.

1.2 Summary statistics

1.2.1 Measures of location

It is often important to summarise data by means of a single figure which serves as a representative of the whole data set. Numbers exist which convert information about the whole data set into a summary statistic which allows easy comparison of different

TABLE 1

Data set showing calories, sodium and alcohol content for 20 types of beer together with the cost of each beer. (Data source: SPSS Professional Statistics v6.1)

ID	Beer	Calories	Sodium	Alcohol	Cost
1	Budweiser	144	15	4.7	0.43
2	Schlitz	151	19	4.9	0.43
3	Lowenbrau	157	15	4.9	0.48
4	Kronenbourg	170	7	5.2	0.73
5	Heineken	152	11	5.0	0.77
6	Old Milwaukee	145	23	4.6	0.28
7	Augsberger	175	24	5.5	0.40
8	Stroh's-Bohemian Style	149	27	4.7	0.42
9	Miller Lite	99	10	4.3	0.43
10	Budweiser Lite	113	8	3.7	0.44
11	Coors	140	18	4.6	0.44
12	Coors Light	102	15	4.1	0.46
13	Michelob Lite	135	11	4.2	0.50
14	Becks	149	19	4.7	0.76
15	Kirin	149	6	5.0	0.79
16	Pabst Extra Light	68	15	2.3	0.38
17	Hamms	136	19	4.4	0.43
18	Heilemans Old Style	144	24	4.9	0.43
19	Olympia Gold Light	72	6	2.9	0.46
20	Schlitz Light	97	7	4.2	0.47

groups within the data set. Such statistics are referred to as measures of location, measures of central tendency, or an average.

Measures of location include the mean, median, quartiles and the mode. The mean is the most well known and commonly used of these measures. However, the median, quartiles and mode are important statistics and are often neglected in favour of the mean.

The mean is the sum of the observations in a data set divided by the number of observations in the data set. Note that all reference to the mean relate exclusively to the arithmetic mean (there are other means which we are not concerned with here, these include the geometric and harmonic mean). The mean of a set of observations takes the form

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Note that $\bar{x}$ is conventional mathematical notation for representing the mean (spoken as 'x bar'). The x values represent the observations in the data set and n is the total number of observations in the data set. As an example, the mean number of calories for the

20 beers listed in table 1 is

Mean number of calories =

$$\frac{144 + 151 + 157 + 170 + \ldots + 97}{20} = 132.4$$

A characteristic of the mean, which on some occasions is an advantage and on others a disadvantage, is its reliance on every number in the data set. A drawback is that the mean can be greatly affected by one or two atypical or outlying observations.

The median is the middle observation in a data set when the data is arranged in ascending or descending order. If the number of observations in the data set is odd there is a unique median. If the number is even there is strictly speaking no middle observation. The median is conventionally defined as the mean of the two middle observations in these circumstances.

The data below is the calories data from table 1 sorted in ascending order. The central observations (there are 2 as the number of observations is even) in this example are 144 and 144, the mean of these 2 values is obviously 144 and this is also the median.

$$68, 72, 97, 99, 102, 113, 135, 136, 140, \overbrace{144, 144}^{\text{Central Observations}}, 145, 149, 149, 149, 151, 152, 157, 170, 175$$

A disadvantage of the median is that it is based on the central position in the data set rather than its magnitude; therefore it is usually less effective than the mean because it wastes information.

The quartiles are the observations in a data set when the data has been split into four equal parts and arranged in ascending or descending order. The lower quartile is the median of the lower half of a data set, the middle quartile is the median and the upper quartile is the median of the upper half of a data set. The inter-quartile range is calculated by subtracting the lower quartile away from the upper quartile.

There is no set method when calculating quartiles. Software packages use different methods when calculating quartiles and will produce different values for each quartile. For example, Excel uses the formula (n-1)/4 for Q_1 (the lower quartile) and 3(n-1)/4 for Q_3 (the upper quartile), whereas SPSS uses the formula (n+1)/4 and 3(n+1)4. There is no ‘correct’ way to calculate quartiles and the differences between the values are often small. Since Excel is primarily a spreadsheet programme and SPSS a statistics programme we advise using the SPSS method. For example, using the data above Q_1 is the 5.25th value in the list and Q_3 is 15.75th value. In order to calculate these figures we need to calculate the difference between the 5th and 6th value and divide by 4, to understand what 0.25th of a value represents for Q_1; and the difference between the 15th and 16th value and divide by 4 then multiply by 3 to understand what 0.75th of a value represents. Q_1 is therefore 104.75 and Q_3 is 150.5

$$68, 72, 97, 99, \overbrace{102, 113}^{Q1}, 135, 136, 140, 144$$

$$144, 145, 149, 149, \overbrace{149, 151}^{Q3}, 152, 157, 170, 175$$

The interquartile range is the difference between these two figures 45.75.

An advantage of the interquartile range is that it is less affected by outliers as it describes the middle values of a data set, therefore it is often more useful than the range (later discussed).

THERE IS NO SET METHOD WHEN CALCULATING QUARTILES. SOFTWARE PACKAGES USE DIFFERENT METHODS WHEN CALCULATING QUARTILES AND WILL PRODUCE DIFFERENT VALUES FOR EACH QUARTILE

The mode is defined as the observation which occurs most frequently in the data set, i.e. the most popular observation. A data set may not have a mode because too many observations have the same value or there are no repeated observations. The mode for the calories data is 149; this is the most frequently occurring observation as it is present in the data set on 3 occasions.

There is no correct measure of location for all data sets as each method has its advantages. A distinct advantage of the mean is that it uses all available data in its calculations and is thus sensitive to small changes in the data. Sensitivity can be a disadvantage where atypical observations are present. In such circumstances the median would provide a better measure of the average. On occasions when the mean and median are unsuitable average measures the mode may be preferred. An instance where the mode is particularly suited is in the manufacture of shoe and clothes sizes.

1.2.2 Measures of dispersion

Once the mean value of a set of observations has been obtained it is usually a matter of considerable interest to establish the degree of dispersion or scatter around the mean. Measures of dispersion include the range, variance and the standard deviation. These measures provide an indication of how much variation there is in the data set.

The range of a group of observations is calculated by simply subtracting the smallest observation in the data set from the largest observation in the data set. The advantage of this measure of variation is that is very easy to calculate. However, it has a major drawback in that it depends on two extreme observations, which may not be representative of the whole data set.

The variance and standard deviation are measures of the average deviation from the mean of all observations in the data set and can be used to assess how dispersed the data is. The variance is calculated by subtracting each observation from the mean, summing the squared values and dividing by the number of observations. This is expressed algebraically as

$$\text{Variance} = \frac{\sum(x-\bar{x})^2}{n}$$

The numerator is often referred to as the sum of squares about the mean. The variance for calories in the 20 beers listed in table 1 is calculated as

$\sum(x-\bar{x})^2 = (144\text{-}132.4)^2 + (151\text{-}132.4)^2 + (157\text{-}132.4)^2 + (170\text{-}132.4)^2 + \dots + (97\text{-}132.4)^2 = 17360$

Variance = 17360/20 = 868

The variance is measured in the square of the unit of measurement. Thus if the unit of measurement is seconds the variance is measured in seconds squared. This may be undesirable and a measure of variance expressed in the original units of measure may be more convenient. A measure of variance in the original units of measure can easily be obtained by taking the square root of the variance. This value is the standard deviation.

$$\text{Standard Deviation} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

The standard deviation for the calories data is

$\text{Standard Deviation} = \sqrt{868} = 29.5$

SAMPLING

2.1 Population and Sample data

The word ‘population’ is usually used to refer to a large collection of data such as the total number of people in a town, city or country. Populations can also refer to inanimate objects such as driving licences, birth certificates or public houses. To study the behaviour of a population it is often necessary to resort to sampling a proportion of the population. The sample is to be regarded as a representative sample of the entire population. Sampling is a necessity as in most instances it is impractical to conduct a census of the population. Unfortunately in most instances the sample will not be fully representative and does not provide the real answer. Inevitably something is lost as a result of the sampling process and any one sample is likely to deviate from a second or third by some degree as a consequence of the observations included in the different samples.

A good sample will employ an appropriate sampling procedure, the correct sampling technique and will have an adequate sampling frame to assist in the collection of accurate and precise measures of the population.

2.2 Sampling techniques

Sampling techniques are broken down into two very broad categories on the basis of how the sample was selected; these are probability and non-probability sampling methods. A probability sample has the characteristic that every member of the population has a known, non-zero, probability of being included in the sample. A non-probability sample, on the other hand, is based on a sampling method which does not incorporate probability into the selection process. Generally, non-probability methods are regarded as less reliable than probability sampling methods. Probability sampling methods have the advantage that selected samples will be representative of the population and allow the use of probability theory to estimate the accuracy of the sample.

2.3 Sampling procedures

2.3.1 Probability sampling

Simple random sampling is the most fundamental probability sampling technique. In a simple random sample every possible sample of a given size from the population has an equal probability of selection.

Systematic sampling involves selecting the members of a population in a systematic fashion from a list which is arranged in some relevant order. For example, in drawing a sample of size 3 from the beer data in table 1 the first beer would be selected by picking a random number between 1 and 20, after which every 3rd beer would be included in the sample. If the initial selection was 5 then Heineken is included in the sample together with Stroh's Bohemian and Coors.

This method is fundamentally equivalent to simple random sampling and in many instances can be more efficient than simple random sampling. There are however situations where systematic sampling is dangerous. If the list of data is arranged in a cyclical fashion a biased estimate of the population measures will be obtained.

Stratified sampling is the process of dividing the population into homogeneous subgroups, or strata, each of which is randomly sampled with a known sample size. The strata are formed on some known characteristic of the population which is known to be related to the variable of interest. This method has the advantage of improving the precision of a sample by reducing the degree of sampling error observed.

The main disadvantage of stratified sampling is that it is no less expensive than simple random sample, indeed it may be more expensive, as detailed information is required about all of the strata before sampling takes place.

Multistage sampling is a more complex sampling technique which clusters enumerate units within the population. The method is most frequently used in circumstances in which a list of all members of the population is not available or does not exist. In this method the population is divided into a number of initial sampling groups. For example, this could comprise all universities in the UK. A random sample is drawn from the initial sampling group. Then, all of the members of the selected cluster are listed. For example, the academic staff working in statistics departments of the randomly selected universities. Finally, the members listed in each of the selected clusters are sub-sampled by either systematic or simple random sampling. This provides the final sample of the population of interest, which in this instance is academic staff working in statistics departments in UK universities.

2.3.2 Non-probability sampling

Purposive sampling is a typical non-probability sampling method which uses existing knowledge of the populations' characteristics to select members for sampling. Quota sampling is an example of this type of sampling. For example, quotas may be set for certain characteristics in a survey (such as age, gender and social class) so that proportions in various subgroups are representative of those in the population.

2.4 Sample frames

In probability sampling the probability of any observation appearing in a sample must be known. For this to be accomplished a list must exist of all members of the population. Such a list is called a sampling frame. Electoral registers, telephone directories and databases are some examples of sampling frames.

The frame must have the property that all members of the population have some known chance of being included in the sample by whatever method is used to list the members of the population. It is the nature of the survey that dictates which sampling should be used. A sampling frame which is adequate for one survey may be entirely inappropriate for another study.

2.5 Sample size

The most frequently asked questions concerning sampling are usually, “What size sample do I need?” or “How many people do I need to ask?”. The answer to these questions is influenced by a number of factors, including the purpose of the study, the size of the population of interest, the desired level of precision, the level of confidence required or acceptable risk and the degree of variability in the collected sample.

The size and characteristics of the population being surveyed are important. Sampling from a heterogeneous (i.e. a population with a large degree of variability between the observations) population is more difficult than one where the characteristics amongst the population are homogeneous (i.e. a population with little variability between the observations in the population). The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The more homogeneous a population, the smaller is the required sample size.

2.5.1 Sampling error

The quality of any given sample will also rely upon the accuracy and precision (sampling error) of the measures derived from the sample. An accurate sample will include observations which are close to the true value of the population. A precise sample, on the other hand, will contain observations which are close to one another but may not be close to the true value (i.e. small standard deviation). A sample that is both accurate and precise is thus desirable.

A common example often used to demonstrate the difference between accuracy and precision is shown in figure 1 below. The example shows a number of targets which can be thought of as rifle shots. In the upper half of the figure the targets show inaccurate data. These observations will provide biased measures of the population values and are undesirable, even if they are precise – as in figure 1a. Accurate data is shown in the lower half of the figure. Figure 1c is both precise and accurate and is the most desirable. Figure 1d is also accurate, even though the scatter between the observations is great. The greater the level of precision in a sample, the smaller is the standard deviation.

If a number of samples are drawn from a population at random on average they will estimate the population average correctly. The accuracy with which these samples reflect the true population average is a function of the size of the samples. The larger the samples, the more accurate the estimates of the population average will be.

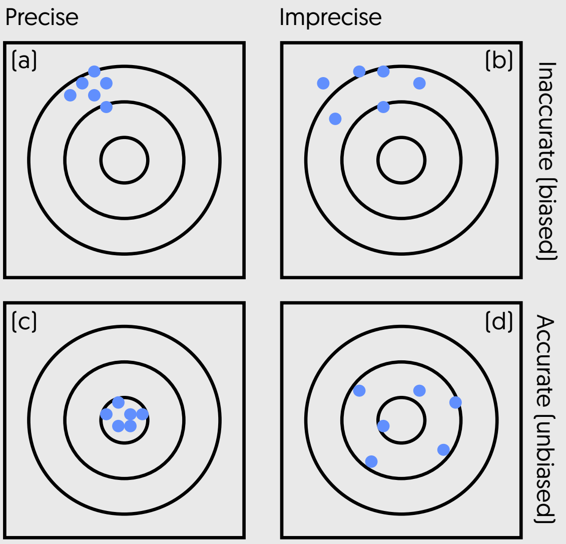
If a sample is large (n>100) the sampling distribution of the observations will be ‘bell shaped’. Many, but not all, bell shaped sampling distributions are examples of the most important distribution in statistics, known as the Normal or Gaussian distribution.

A normal distribution is fully described with just two parameters: its mean and standard deviation. The distribution is continuous and symmetric about its mean and can take on values from negative infinity to positive infinity. Figure 2 shows an example of a normally distributed sample of 200 (the histogram) with the theoretical normal curve overlaid. The figure shows the actual amount of beer consumed by 200 men in a week in the UK. The mean beer consumption for this set of data is 5.67 and has a standard deviation of 2.09.

From this survey of 200 men we can tentatively assume that the mean consumption of beer for

FIGURE 1

Biased and unbiased estimates of a population measure from a sample.



the whole population of men in any given week is approximately 5.67 pints. How accurate is this statement? This sample of beer drinkers may not be representative of the population, if this is the case the assumption that the mean beer consumption is 5.67 pints will be incorrect. It should be noted that it is not possible to obtain the true mean value in this case as the population of beer drinkers is very large. To obtain the true population mean all beer drinkers would have to be surveyed. This is clearly impractical. See Figure 2.

2.5.2 Confidence intervals

From the data that has been collected one will want to infer that the population mean is identical to or similar to the sample mean. In other words, it is hoped that the sample data reflects the population data.

The precision of a sample mean can be measured by the spread (standard deviation) of the normal distribution curve. An estimate of the precision of the standard deviation of the sample mean, which describes the spread of all possible sample means, is called the standard error. The standard error of the mean of a sample is given by

$$\text{Standard Error} = \frac{SD}{\sqrt{n}}$$

where SD is the standard deviation of the sample and n in the number of observations in the sample. The beer sample data has a standard error of

$$\text{Standard Error} = \frac{2.09}{\sqrt{200}} = 0.15$$

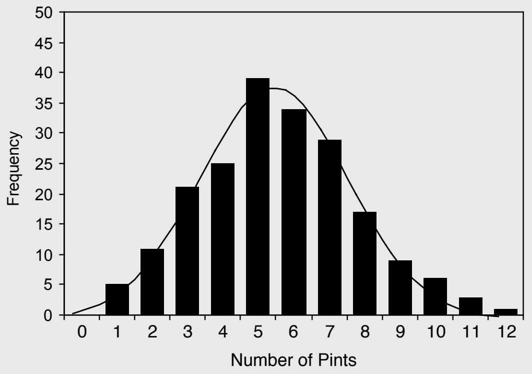
The standard error can be used to estimate confidence levels or intervals for the sample mean which will give an indication of how close our estimate is to the true value of the population. Generally, a confidence level is set at either 90%, 95% or 99%. At a 95% confidence level, 95 times out of 100 the true values will fall within the confidence interval. The 95% confidence interval for a sample mean is calculated by

Confidence Interval = Mean $\pm$ 1.96 x Standard Error

The level of risk is reduced by setting wider confidence intervals. The 99% confidence interval will only have a 1% chance of being incorrect. Confidence intervals are affected by sample size. A large sample size can improve the precision of the confidence intervals.

FIGURE 2

The distribution of average beer consumption for 200 adult men in a period of 7 days. The histogram shows the actual amount of beer consumed by the men. The superimposed bell-shaped curve shows the theoretical distribution.



A confidence interval for the mean number of pints consumed in the UK on any given week is: 95% Confidence Interval = 5.67 ± 1.96 x 0.15 = 5.67 ± 0.29

This translates to a confidence interval of 5.38 pints to 5.96 pints. Thus we can conclude that, although the true population mean for beer consumption is unknown for men in the UK, we are 95% confident that the true value lies somewhere between 5.38 and 5.96 pints*. There is a 1 in 20 chance that this statement is incorrect.

If a greater level of precision were required the risk of making an incorrect statement would be reduced by setting wider confidence intervals. For example, the 99% confidence interval for the mean number of pints consumed is: 99% Confidence Interval = 5.67 ± 2.58 x 0.15 = 5.67 ± 0.39

Using these figures we are 99% confident that the mean number of pints drunk in a given week lies between 5.28 and 6.06 pints of beer*. There is now only a 1 in 100 chance that this statement is incorrect.

[* While it is common to say one can be 95% (or 99%) confident that the confidence interval contains the true value it is not technically correct. Rather, it is correct to say: If we took an infinite number of samples of the same size, on average 95% of them would produce confidence intervals containing the true population value.]

In a market research environment sample survey data is often analysed using proportions or percentage scores. For example, the proportion of customers who recommend a product or service to a friend, or the proportion of customers who are satisfied with the product or service purchased. The standard error and subsequent confidence intervals can be derived for percentage scores. However, the distribution of such data is not normally distributed and hence cannot be calculated using the formulae described above.

The distribution of a proportion follows the binomial distribution (discussion of this distribution not included there). If there is a trial in which there are two possible outcomes occurring with probability p for the first outcome and 1-p for the second the standard error of these proportions is calculated as

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}}$$

For example, if the percentage of customers, from a sample of 200, who state that they would re-purchase a product is 72%, the standard error of this value is

$$\text{Standard Error} = \sqrt{\frac{0.72(1-0.72)}{200}} = 0.032$$

which gives a standard error of 3.2%.

The confidence intervals are calculated using the formula described above. Thus the 95% confidence intervals for this sample of 200 customers are

95% Confidence Interval = 0.72 ± 1.96 x 0.032 = 0.72 ± 0.063

The researcher can conclude from this analysis that between 65.7% and 78.3% of the population of customers will re-purchase on the evidence in the survey.

2.5.3 Determining sample size

The formula for deriving the standard error presented above can be rearranged to determine sample size for a set of categorical variables. The formula for estimating sample size takes the form

$n = p (p – 1)/ (\text{Standard Error})^2$

where the standard error is defined as

Standard Error = d / (z coefficient)

The value d is defined as the acceptable level of error present in the sample. For example, ±5% may be an acceptable level of deviation from the true population parameter which is being estimated (the proportion is entered into the formula, thus for 5% acceptable error 0.05 is entered). The z coefficient is a statistical value that represents the level of chosen confidence, which is 1.96 for 95% confidence, 2.58 for 99% confidence.

It is clear from the section above that the standard error is a function of n, the sample size of a survey. In turn, the width of the confidence intervals depends upon the magnitude of the standard error. If the standard error is large the confidence intervals around the estimated mean will also be large.

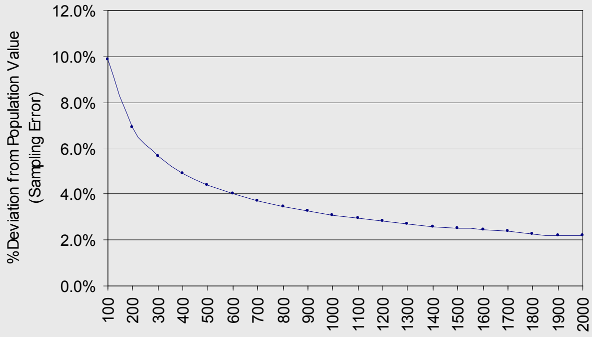
Thus the greater the sample size, the smaller the confidence intervals will be.

As an example, consider the customer data above. In this data set 200 customers were surveyed on their likelihood to re-purchase. Let us assume the sample size in this survey is increased to 800 and that the outcome remains unchanged – 72% will re-purchase. The confidence intervals for this sample are ±3.1%, which would lead the researcher to conclude that the true proportion of customers who will re-purchase lies between 68.9% and 75.1%.

An important point to observe in this example is that the confidence intervals have been reduced by half – discounting rounding – from ±6.3% to ±3.1%. This 50% reduction in error has only been achieved by increasing the sample size fourfold. This is because it

FIGURE 3

Sample size for an estimate drawn from a simple random sample. Sampling error is a function of sample size (n=infinity; p=0.5), as a sample size increases the level of measurement error decreases.



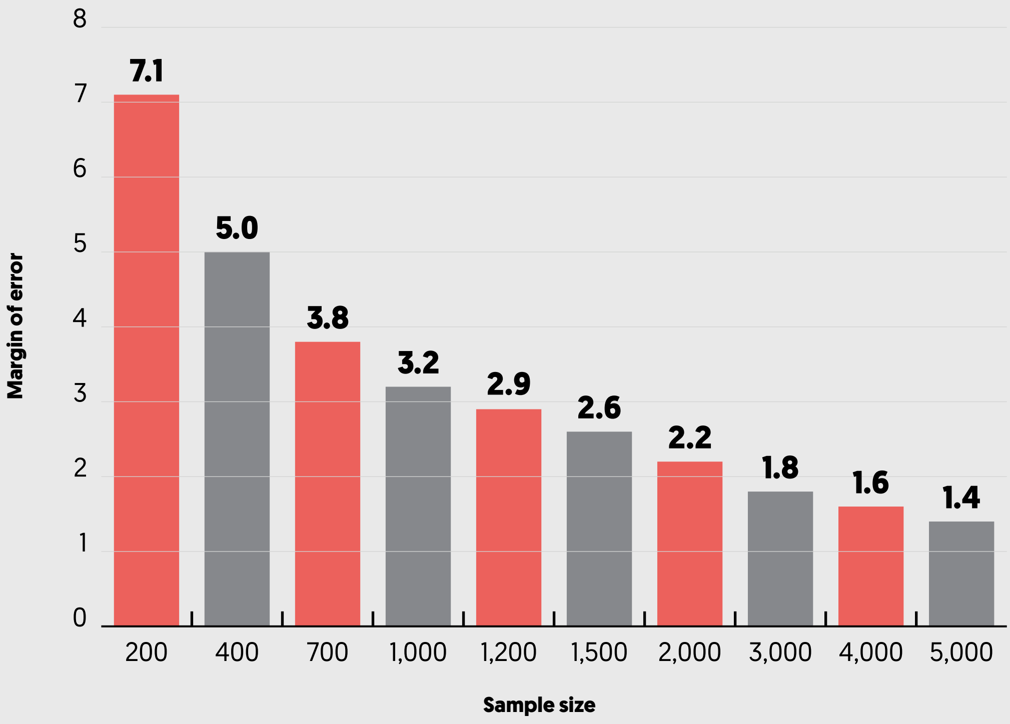
is not the sample size itself which is important but the square root of the sample size. Thus if there is a desire to double the precision of the survey from say +5% to +2.5% the sample size will have to be increased fourfold to achieve this desired level.

Achieving an accurate and reliable sample size is always a compromise between the ideal size and the size which can be realistically attained. The size of most samples is usually dictated by the survey period, the cost of the survey, the homogeneity of data and number of subgroups of interest within the survey data.

In figure 3 below the plotted line shows the expected level of sampling error which would be expected for sample sizes ranging from 100 to 2000. As the sample size is increased the accuracy of the estimate will increase and the degree of measurement or sampling error will decrease. For example, if a sample of 100 is drawn from an infinitely large population the sampling error of the measured variable or attribute will be approximately +10%. If the sample was increased to 2000 the measurement error would be just over +2%.

The larger the sample size the more accurate results will reflect true measures. If there are a number of small subgroups to be analysed, a large sample which adequately reflects these groups will provide more accurate results for all subgroups of interest.

CALCULATED MARGINS OF ERROR FOR SELECTED SAMPLE SIZES



TESTS OF SIGNIFICANCE

When a sample of measurements is taken from a population, the researcher carrying out the survey may wish to establish whether the results of the survey are consistent with the population values. Using statistical inference the researcher wishes to establish whether the sampled measurements are consistent with those of the entire population, i.e. does the sample mean differ significantly from the population mean? One of the most important techniques of statistical inference is the significance test. Tests for statistical significance tell us what the probability is that the relationship we think we have found is due only to random chance. They tell us what the probability is that we would be making an error if we assume that we have found that a relationship exists.

In using statistics to answer questions about differences between sample measures, the convention is to initially assume that the survey result being tested does not differ significantly from the population measure. This assumption is called the null hypothesis and is usually stated in words before any calculations are done. For example, if a survey is conducted the researcher may state that the data collected and measured does not differ significantly from the population measures for the same set of parameters, this is the null hypothesis. If this statement is rejected, the researcher would adopt an alternative hypothesis, which would state that a real difference does in fact exist between your calculated measures of the sample and the population.

There are many statistical methods for testing the hypotheses that significant differences exist between various measured comparisons. These tests are broadly broken down into parametric and non-parametric methods. There is much disagreement amongst statisticians as to what exactly constitutes a non-parametric test but in broad terms a statistical test can be defined as non-parametric if the method is used with nominal, ordinal, interval or ratio data.

For nominal and ordinal data, chi-square is a family of distributions commonly used for significance testing, of which Pearson's chi-square is the most commonly used test. If simply chi-square is mentioned, it is probably Pearson's chi-square. This statistic is used to test the hypothesis of no association of columns and rows in tabular data. A chi-square probability of .05 or less is commonly interpreted as justification for rejecting the null hypothesis that the row variable is unrelated to the column variable.

TABLE 2

The results of improved employee efficiency for two training plans implemented by a company.

Improved Efficiency	Training plan A	Training plan B	Total
Yes	175	100	275
No	25	100	125
Total	200	200	400

TABLE 3

The expected frequency scores for improved employee efficiency for two training plans implemented by a company.

Improved Efficiency	Training plan A	Training plan B	Total
Yes	137.5	137.5	275
No	62.5	62.5	125
Total	200	200	400

As an example, consider the data presented in Table 2 above. The data shows the results of two training plans provided to a sample of 400 employees for a company. The company wishes to establish which of the plans is most efficient at improving employee efficiency as it wishes to provide the rest of its employees with training.

A simple chi-square test can be used to assess whether the differences between the two training plans are due to chance alone or whether there is a difference between the two methods of training which can improve efficiency.

Chi-square is computed by looking at the different parts of the table. The cells contain the frequencies that occur in the joint distribution of the two variables. The frequencies that we actually find in the data are called the observed frequencies.

The total columns and rows of the table show the marginal frequencies. The marginal frequencies are the frequencies that we would find if we looked at each variable separately by itself. For example, we can see in the total column that there were 275 people who have experienced an improvement in efficiency and 125 people who have not. Finally, there is the total number of observations in the whole table, called N. In this table, N=400.

The first thing that chi-square does is to calculate the expected frequencies for each cell. The expected frequency is the frequency that we would have expected to appear in each cell if there was no relationship between type of training plan and improved performance.

The way to calculate the expected cell frequency is to multiply the column total for that cell, by the row total for that cell, and divide by the total number of observations for the whole table. Table 3 below shows the expected frequencies for the observed counts shown in Table 2.

Table 3 shows the distribution of expected frequencies, that is, the cell frequencies we would expect to find if there was no relationship between type of training and improved performance. Note that chi square is not reliable if any cell in the contingency table has an expected frequency of less than 5.

To calculate Chi-square, we need to compare the original, observed frequencies with the new, expected frequencies, the formula for this calculation is:

$$x^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

the summation being over the 4 inner cells of the tables. The contribution of the 4 cells to the chi square is thus calculated as

$$x^2 = \frac{(175 - 137.5)^2}{137.5} + \frac{(100 - 137.5)^2}{137.5} + \frac{(25 - 62.5)^2}{62.5} + \frac{(100 - 62.5)^2}{62.5}$$

$$= 65.45$$

The value calculated can be read off from the chi-square distribution tables to establish the significance of this result. This chi square value has a p value (i.e. probability) less than 0.05 which suggests that the null hypothesis that the two training methods are the same should be rejected. Note that the calculations shown above do not usually need to be manually calculated by the researcher as most statistical applications (SPSS, SAS) and spreadsheets (Excel) have routines which calculate chi square and the relevant p values.

The chi-square test is a classic example of a significance test for use with nominal or ordinal data. There are many other tests which can also be used to test for significant variability of this type of data, Kolmogorov-Smirnov, Mann-Whiney, Wilcoxon and Kriskal-Wallis are but a few others which are described as non-parametric and can be used for nominal, ordinal and interval data types.

When data is continuous the tests of significance are usually described as parametric. As with the non-parametric variety there are many tests which can be used to compare data. Two of the most common tests are the t-test and analysis of variance (ANOVA). The t-test is used for testing for differences between mean scores for two groups of data and is one of the most commonly used tests applied to normally distributed data. The ANOVA test is a multivariate test used to test for differences between 2 or more groups of data.

There a number of t-tests which can be applied to parametric data, in the example below the paired t-test is used to test for differences in the flavour of two confectionery products. Ten consumers were asked to taste the products and rate the overall flavour on a scale of 1 to 20 (note: in theory this scale may not necessarily be regarded as continuous thus rendering the use of the t-test as inappropriate, however for this example the scale will be assumed

TABLE 4

Overall flavour scores of 2 confectionery products given by 10 consumers.

Consumer	Sweet A	Sweet B	Difference (A-B)
1	16	20	-4
2	11	18	-7
3	14	17	-3
4	15	19	-4
5	14	20	-6
6	11	12	-1
7	15	14	1
8	19	11	8
9	11	19	-8
10	8	7	1
Means	13.4	15.7	-2.3

to be approximately continuous). The score given to each sweet by the 10 consumers is shown in table 4.

Using the standard error and the mean difference (-2.3) calculated for the paired observations the t-test statistic can be derived. The formula for the paired t-test is

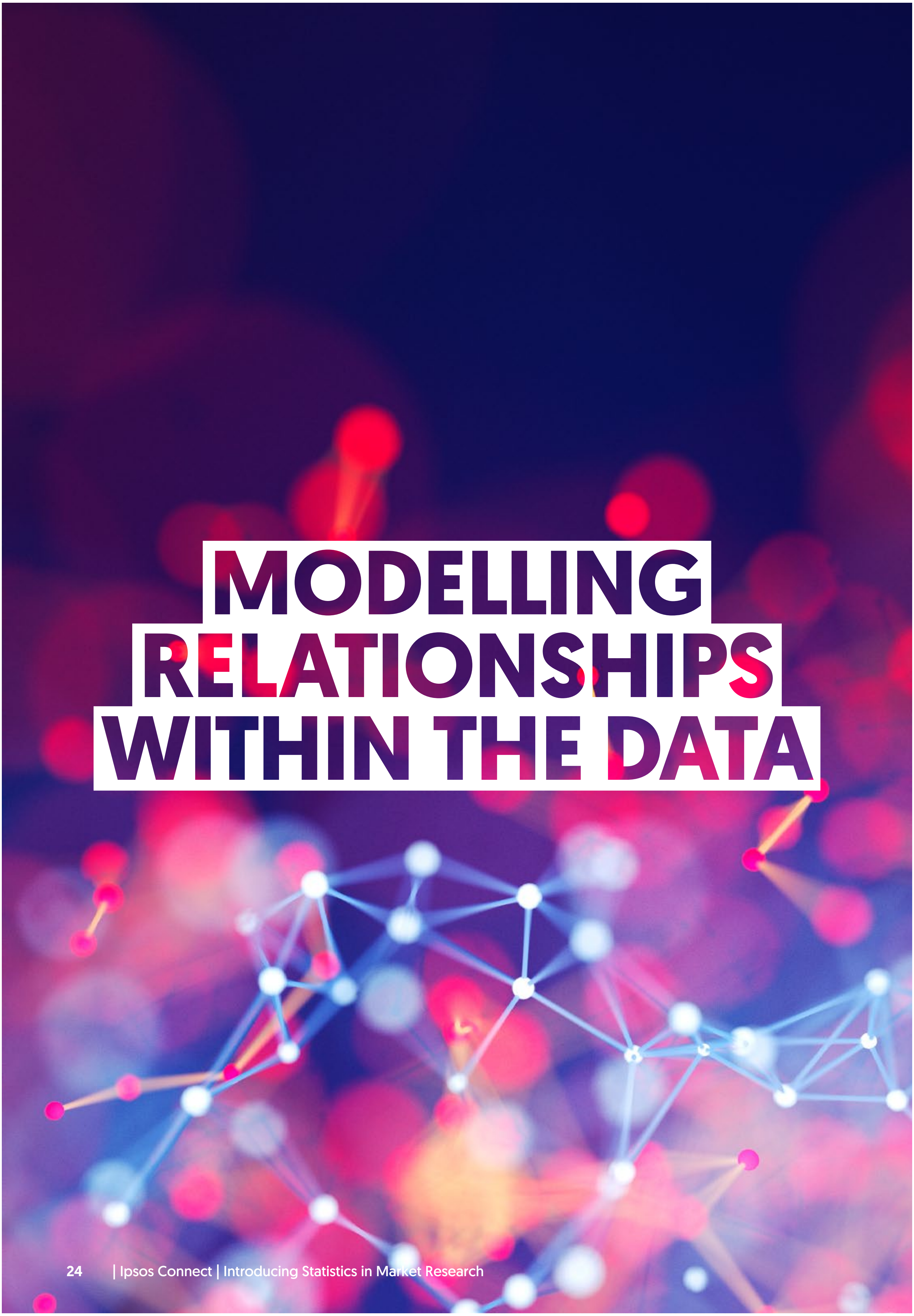
$$t = \frac{\text{Difference}}{\text{Standard Error}}$$

In this example the p value derived from the t-test statistic is 0.16. As this value is somewhat greater than 0.05 we must concluded that the flavour in the two types of confectionery is much the same and the differences between the two products mean scores is not statistically significant.

An example of the uses of the ANOVA test is not provided here but a useful application of the test would be in a situation where the number of sweets being tested exceeded 2 and we wished to establish if there are differences between all the products on offer.

Please contact the statisticians that you're working with to find out about the excel tools available to calculate tests of significance.

NOTES



MODELLING RELATIONSHIPS WITHIN THE DATA

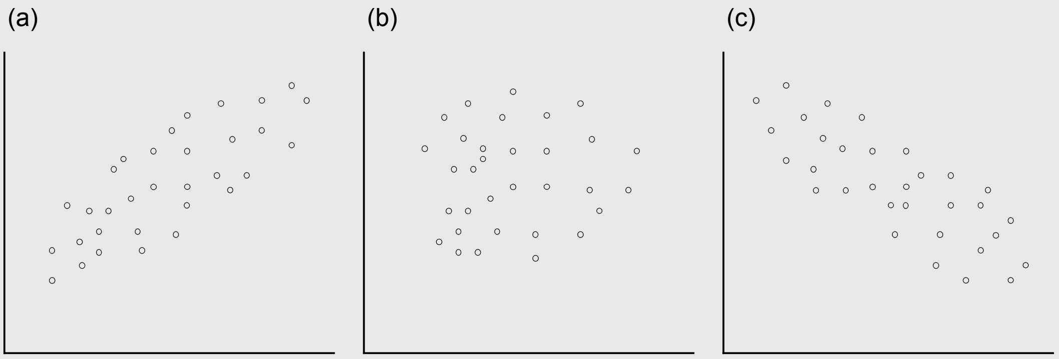
4.1 Correlation

The correlation coefficient (r) for a paired set of observations is a measure of the degree of linear relationship between two variables being measured. The correlation coefficient measures the reliability of the relationship between the two variables or attributes under investigation. The correlation coefficient may take any value between plus and minus one. The sign of the correlation coefficient (+/-) defines the direction of the relationship, either positive or negative. A positive correlation coefficient means that as the value of one variable increases, the value of the other variable increases. If the two attributes being measured are perfectly correlated the coefficient takes the value +1. A negative correlation coefficient indicates that as one variable increases, the other decreases, and vice-versa. A perfect relationship in this case will provide a coefficient of –1.

The correlation coefficient may be interpreted by

FIGURE 4

Three scatter plots showing varying degrees of correlation between pairs of attributes. In (a) there is a positive correlation between the attributes (r>0); (b) there is no correlation between the attributes (r=0); (c) there is a negative correlation between the attributes (r<0).



EXAMPLE OF CORRELATION

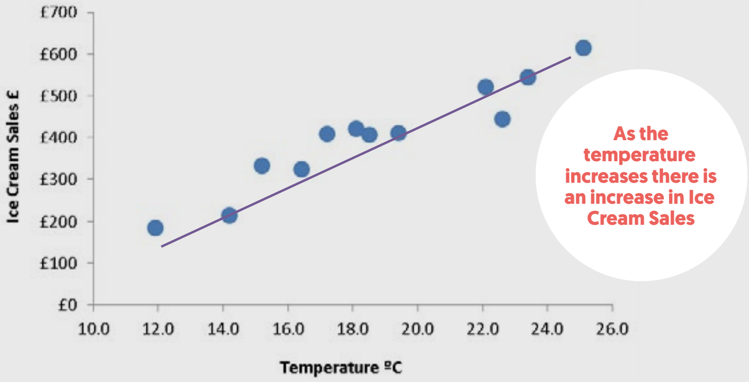
The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. Here are their figures for the last 12 days (Table (a))

Temperature °C (x)											
14.2°	16.4°	11.9°	15.2°	18.5°	22.1°	19.4°	25.1°	23.4°	18.1°	22.6°	17.2°
Ice Cream Sales (y)											
£215	£325	£185	£332	£406	£522	£412	£614	£544	£421	£445	£408

First rank the x data from 1 to n. (Here, n = 12). The results are in Table (b)

Temperature °C (x)											
11.9°	14.2°	15.2°	16.4°	17.2°	18.1°	18.5°	19.4°	22.1°	22.6°	23.4°	25.1°
Ice Cream Sales (y)											
£185	£215	£332	£325	£408	£421	£406	£412	£522	£445	£544	£614

The Correlation Coefficient is +0.96. So, there is a strong positive relationship between the temperature and Ice Cream Sales for this local shop



As the temperature increases there is an increase in Ice Cream Sales

various means. The scatter plot best illustrates how the correlation coefficient changes as the linear relationship between the two attributes is altered. When there is no correlation between the paired observations the points scatter widely about the plot and the correlation coefficient is approximately 0 (see figure 4b below). As the linear relationship increases points fall along an approximate straight line (see figures 4a and 4c). If the observations fell along a perfect straight line the correlation coefficient would be either +1 or –1.

Perfect positive or negative correlations are very unusual and rarely occur in the real world because there are usually many other factors influencing the attributes which have been measured. For example, a supermarket chain would hope to see a positive correlation between the sales of a product and the amount spent on promoting the product. However, they would not expect the correlation between sales and advertising to be perfect as there are undoubtedly many other factors influencing the individuals who purchase the product.

No discussion of correlation would be complete without a brief reference to causation. It is not unusual for two variables to be related but that one is not the cause of the other. For example, there is likely to be a high correlation between the number of fire-fighters and fire vehicle attending a fire and the size of the fire. However, fire-fighters do not start fires and they have no influence over the initial size of a fire!

4.2 Regression

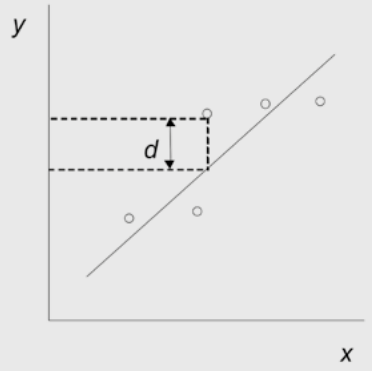
The discussion in the previous section centred on establishing the degree to which two variables may be correlated. The correlation coefficient provides a measure of the relationship between measures, however it does not provide a measure of the impact one variable has on another. Measures of impact can be derived from the slope of a line which has been fit to a group of paired observations. When two measures are plotted against each other a line can be fitted to the observations. This straight line is referred to as the line of 'best fit' and it expresses the relationship between the two measures as a mathematical formula which is the basis for regression analysis.

Regression analysis attempts to discover the nature of the association between the variables, and does this in the form of an equation. We can use this equation to predict one variable given that we have sufficient information about the other variable. The variable we are trying to predict is called the dependent variable, and in a scatter plot it is conventional to plot this on the y (vertical) axis. The variable(s) we use as the basis for prediction are called the independent variable(s), and, when there is only one independent variable, it is customary to plot this on the x (horizontal) axis.

As already stated, correlation attempts to express the degree of the association between two variables. When measuring correlation, it does not matter which variable is dependent and which is independent. Regression analysis implies a 'cause and effect' relationship, i.e. smoking causes lung cancer. But variables can be correlated even though they in no way affect each other – because they are both influenced by a third variable (causation). For example, it has been shown that there is a high degree of correlation between infant mortality and overcrowding – but both variables are closely affected by the level of income.

FIGURE 5

An illustration of a regression line takes the classical form. In this model the method of least squares is used to minimise the error (d)

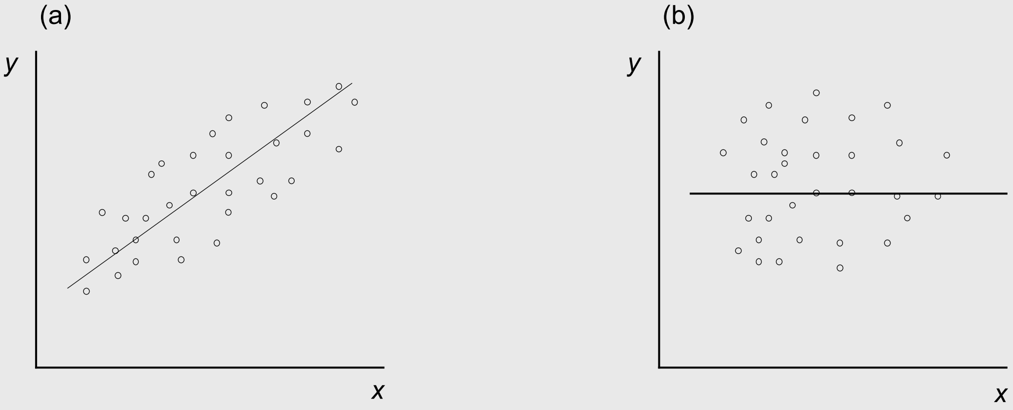


The regression line expresses the nature of the relationship between variables in the form of an equation by drawing the line of best fit. The problem is how to decide which straight line has the best fit? The method used is called the least squares line. This is done by drawing the line which minimises the sum of the vertical distances squared, between the line and each of the points, as seen in figure 5 above.

This figure shows that the lower values of the dependent variable (y) are associated with the lower values of the independent variable (x). This means that the data are positively correlated and the highest

FIGURE 6

An illustration of two lines fitted to different data sets. In the first example (a) the line of best fit is significantly different from 0; x is thus a reasonable predictor of y. In the second example (b) the line is not significantly different from 0; any value of x will provide the same predictor for y.



values of the dependent variable are associated with the highest values of the independent variable.

Having obtained a regression line ‘y on x’ (as it is sometimes called) it shows that the values of y depend on the values of x. Given values of x we can now estimate or predict the values of y – the dependent variable. The next step is to establish how well the estimated regression line fits the data, i.e. how well does the independent variable predict the dependent variable. Once a linear regression line has been fit to the observations the next step to test that the slope is significantly different from 0. In figure 6 below least squares regression lines have been fitted to two sets of data. Figure 6a shows a line which is significantly greater than 0, thus values of x (independent variable) provide reasonable estimates of y (dependent variable). Figure 6b show a line which is not significantly different from zero. In this example the fitted line does not provide a good model for estimating the dependent variable, thus x in this example is a poor predictor of y.

The closer the value of the slope is to 1 the better the fit of the line and the more significant the independent variable will be. Having determined the significance of the slope there are a number of other statistics of importance which assist in determining how well the line fits the data. The squared value of the correlation coefficient, r^2 (r–squared), is a useful statistic as it explains the proportion of the variability in the dependent variable which is accounted for by changes in the independent variable(s). Another useful statistic is the ‘adjusted sum of squares of fit’ calculation: this is the predictive capability of the model for the whole population not just for the sample of values on which the model has been derived. This adjusted measure of fit is usually, as would be expected, not as close to one as the standard measure.

The illustrations so far have used regression with only one independent variable – however, in many instances several independent variables may be available for use in constructing a model to assess the behaviour of a dependent variable. This is called multiple linear regression. For example IQ might be the dependent variable and information on mathematical ability, spatial awareness and language comprehension might be used as independent variables to predict it. Software packages test if each coefficient for the variables in a multiple linear regression model are significantly different from zero. If any are not they can be omitted from the model as a predictor of the dependent variable.

Regression is a very powerful modelling tool and can provide very useful insight into a predictive (dependent) variable and insight into which of the independent measures have the greatest impact upon the dependent variable, i.e. which variables are the most significant. However, as with many statistical methods the technique is often misused and the methods are subject to error. A linear model may be an inadequate description of the relationship between the dependent and independent variable(s) and should not be fitted to a data set which is known not to be linear over the range of the independent variable(s). The danger of using the wrong model is particularly severe if an attempt is made to extrapolate beyond the range of the independent variables. For example, in figure 6a above it would be dangerous to use the regression of y on x to predict beyond the point where the fitted line stops, as there are no observations beyond this point and we cannot be sure that the relationship would continue to be linear.

EXAMPLE OF REGRESSION

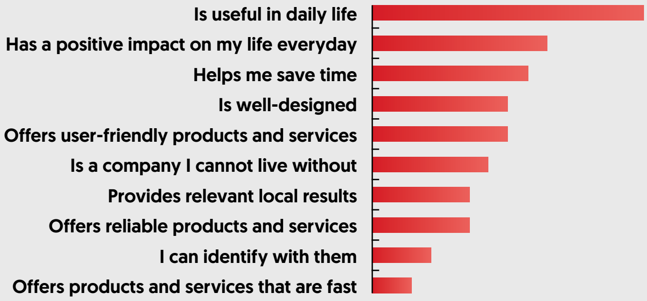
A client commissioned a survey to assess what's driving Favourability of a brand.

There are a number of statements related to Favourability in the survey and with Regression Analysis we can check the Top10 attributes (in order of importance) driving Favourability.

The attributes encompass areas such as Functional Aspects, Brand Personality and Privacy.

After running the Regression Analysis, the attributes received a weight which measures the importance of each of them in impacting on Favourability

The Top 10 attributes in order of importance are:



The most important attribute in driving Favourability is "Is useful in daily life". Because there is a causation involved in this technique, any actions or changes in this aspect will impact on Favourability as well as the importance of the other attributes.

SEGMENTATION

5

Segmentation techniques are used for assessing to what extent and in what way individuals (or organisations, products, services) are grouped together according to their scores on one or more variables, attributes or components. Segmentation is primarily concerned with identifying differences between such groups of individuals and capitalising on differences by identifying to which groups the individuals belong. Through the use of segmentation analysis it is possible to identify distinct segments with similar needs, values, perceptions, purchase patterns, loyalty and usage profiles – to name a few.

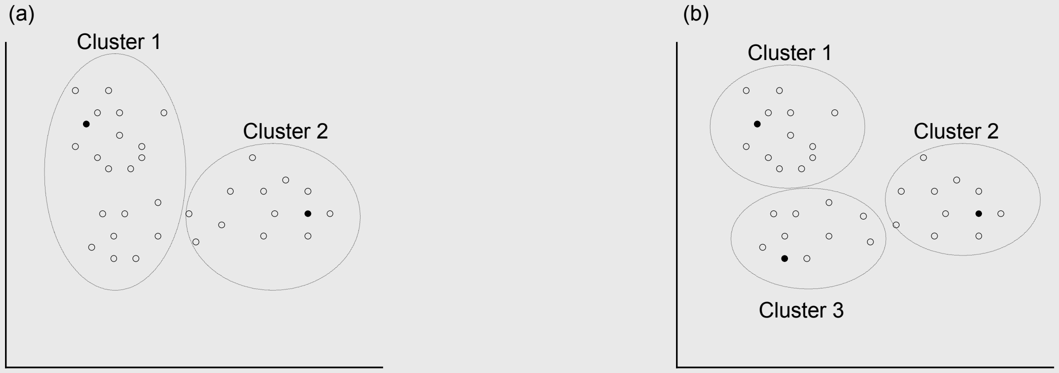
In this section the discussion will centre on two statistical processes which fall into a branch of statistics known as multivariate analysis. These methods are cluster analysis and factor analysis. The two methods are similar in so far as they are grouping techniques, generally cluster analysis groups individuals – but can also be used to classify attributes – while factor analysis groups characteristics or attributes.

5.1 Factor analysis

Factor analysis is a technique used to find a small number

FIGURE 7

Two cluster solutions for a hypothetical data set. The solid black circles represent the initial cluster centres (centroids) for the two solutions.



of underlying dimensions from amongst a larger number of attributes or variables. The method reduces the total number of attributes in a survey from a larger number of correlated attributes to only a few fundamental factors or components which are uncorrelated. The derived factors represent key dimensions in the data which can represent underlying perceptions of a product or service. Thus individuals responses to a battery of attributes can be reduced to a few underlying factors. These factors may be concepts which are difficult to isolate within a single attribute or statement but make sense once a group of attributes are grouped to form a factor.

Factor analysis works by examining the paired correlations between all the attributes measured in a survey. Attributes which are strongly correlated will be grouped together to form new factors. Defining the number of factors to be used of any given battery of questions can be somewhat subjective. A balance must be struck between the number of factors and the variance explained by the factors. The variance explained is a measure of the detail which is retained by the factors used. The greater the number of factors, the greater is the proportion of the variance explained and more information about the data is retained.

Once an appropriate factor solution has been obtained each respondent can be assigned a score for each of the derived factors. Principal components analysis (PCA) is a common technique used to assign scores to the defined factors. These scores or weights estimate the relative importance of each component to the respondent and can be used as new variables for additional analysis. A respondent who scores a high score on the attributes which contributes most to a factor will also have a high score for the factor. Thus high scores for a factor will generally imply agreement with the factor and low scores will generally imply disagreement.

EXAMPLE OF FACTOR ANALYSIS

TV is likely to hold a deeper emotional attachment with people compared to other media. With the objective to better understand this emotional attachment with TV, 15 attitudinal questions from a TV Survey were used to find out an index to represent the concept “TV - Essential Part of Life”.

Question
To what extent, if at all, do you agree with each of the following statements about watching TV? (5 pt. scale)

Running the Factor Analysis, the attributes were grouped into three themes:

TV – Content Advocates
- I often recommend my favourite TV programmes to other people
- I often learn new things whilst watching TV
- I'm always on the look out for new things to watch on TV

TV – Inspiration
- I aspire to be more like some of the people I admire on TV
- TV often inspires me to find out more about products/brands
- TV has inspired me to try something new or get involved in a new activity
- TV helps bring my household/ family/friends together
- TV programmes are regularly the hot topic of conversation
- Adverts on TV are becoming increasingly better quality
- TV gives me something in common with other people

TV – Essential Part of Life
- TV is one of my favourite forms of entertainment
- Watching TV is one of my favourite ways of relaxing
- TV provides me with company when I am alone
- TV can move me emotionally in a way that other media can't
- TV can grip my attention in a way that other media can't

The questions from this factor were combined to create the index “TV – Essential part of Life”.

Using other stats techniques we can assess the impact of the other attributes on this emotional attachment with TV.

5.2 Cluster analysis

Cluster analysis is a classification method which uses a number of mathematical techniques to arrange sets of individuals into clusters. The aim is to establish a set of clusters such that individuals within a given cluster are more similar to each other than they are to individuals in other clusters. Cluster analysis sorts the individuals into groups, or clusters, so that the degree of association is strong between members of the same group and weak between members of different groups. Each such cluster thus describes, in terms of the data collected, the class into which the individuals fall and reveals associations and structures in the data which were previously undefined.

There are a number of algorithms and procedures used for performing cluster analysis. Typically, a cluster analysis begins by selecting a number of cluster centres. The number of centres at the beginning is dictated by the number of clusters required. Usually the optimum number of clusters is not known before the analysis begins, thus it may be necessary to run a number of cluster groupings in order to establish the most appropriate solution. For example, consider the solutions defined in figure 7. In this example a two cluster solution was used in the first instance to group each observation into one of two clusters (figure 7a). The solid black dots represent the initial cluster centroids for the two clusters. The remaining observations are then assigned to one of the cluster centres. A three cluster solution was then generated (figure 7b). Clearly the groupings in the three cluster solution appear more distinct, however this does not always indicate that the 'best' cluster solution has been achieved.

Once an appropriate cluster solution has been defined and all observations have been allocated

EXAMPLE OF CLUSTER ANALYSIS

INTERNET USAGE

▼

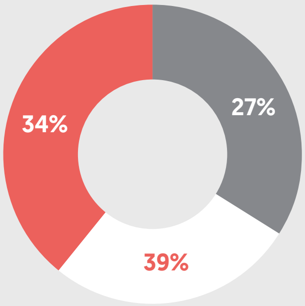
Segment customers based on how they use the internet

▼

7 statements about internet familiarity were asked in 5 pt. scale of agreement to 1,890 customers

1 Strongly disagree
▶ 5 Strongly agree

		Heavy Users	En-gaged	Disen-gaged
1	I feel that I am restricted in what I can and can't do on the internet due to my lack of skill	1	2	5
2	I feel that I am being left behind with my lack of knowledge and use of the internet	1	2	4
3	I need to rely on help from others, be it friends, family, courses or books, in order to be able to learn new skills on the internet	1	1	5
4	I want to learn more about the internet but I don't know where to start	1	1	5
5	I don't like to try new things on internet because I am worried I will do something wrong	1	1	5
6	I can't live without the internet	5	4	1
7	I feel I spend too much time on the internet	5	3	1



We can use other stats techniques to better understand and profile the segments found.

to one of the segments the next step is to establish whether the results are intuitively sound and actionable. This can be achieved by using the clusters as cross breaks on attributes, demographics and behavioural tabulations. Demographic data such as age, gender, social class and marital status can be used to establish the average age of a cluster group, the proportion of the cluster who are male, the predominant social group and so on. For example, if the average age of cluster 1 in figure 7b is 20 years, 87% of the group are single, and 71% of the cluster are male, this information can be used not only to name the cluster (i.e. young single men) but also to profile purchasing behaviour, identify key drivers and to target specific products and services at the group. ●

FOR MORE INFORMATION PLEASE CONTACT:

Leo Cremonezi
Senior Statistician
E Leo.Cremonezi@ipsos.com
T +44 (0)20 8080 6112

Leo is a senior statistician within Ipsos Connect and a chartered member of the Royal Statistical Society. He has been responsible for developing statistical insights for different types of Adhoc and Tracker studies. He is also responsible for teaching & training with the objective to make Stats accessible to all.

SUGGESTIONS FOR FURTHER READING

INTRODUCTORY TEXTS

Devore, J and Roxy P (1994). Introductory Statistics (2nd ed). West Publishing Company

Hannagan, T (1997). Mastering Statistics (3rd ed). MACMILLAN

ADVANCED TEXTS

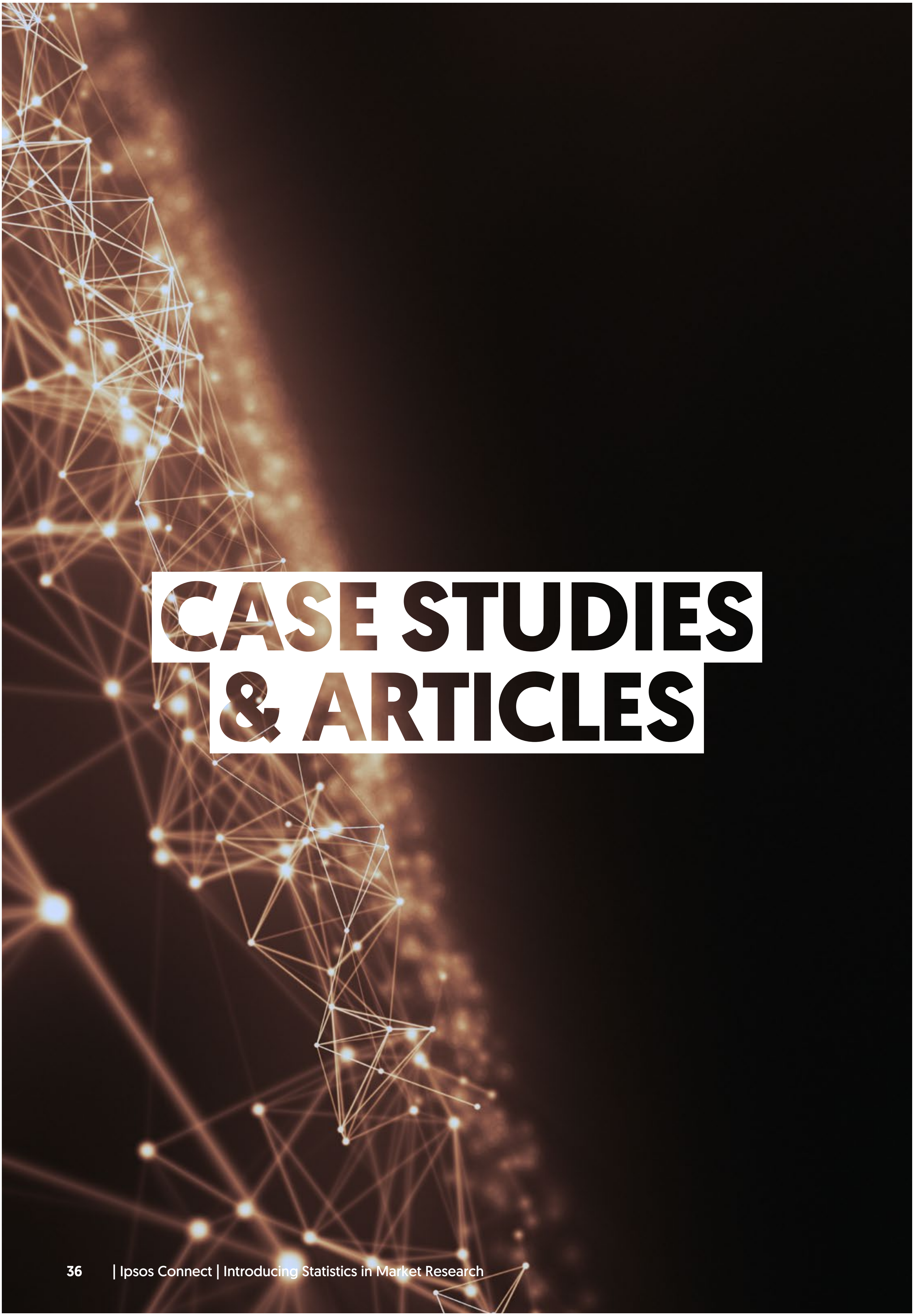
Hair, J F et al (2006). Multivariate Data Analysis (6th ed). Pearson Prentice Hall

GENERAL TEXTS

Vogt, W P and Johnson, R B (2001). Dictionary of Statistics & Methodology: a nontechnical guide for the social sciences (4td ed. SAGE

STATISTICAL TABLES

Murdoch, J and Barnes J a (1986). Statistical Tables for Science, Engineering, Management and Business Studies (3rd ed). MACMILLAN



CASE STUDIES & ARTICLES

LINKAGE ANALYSIS: DATA'S HIDDEN STORIES

In the age of big data there never seems to be a shortage of stats and figures. Linkage analysis combines survey data with a client's in-house statistics to create richer insights.

Linkage analysis combines data sets to extract more tangible and actionable insights. Using the example of fast food restaurants this thought piece investigates how the process works and what the key benefits of the technique are. By translating data into stories, linkage modelling can help interpret complex data sets and allow clients to make the most of their data.

VIEW ARTICLE

HIGH DEFINITION CUSTOMERS – A POWERFUL SEGMENTATION: UNLOCKING VALUE WITH DATA SCIENCE

Just like the best films, data can tell a story too – you just need to know where to look.

This thought piece explores how three types of advanced statistical analysis – Factor, Cluster and CHAID analysis – can help us unlock additional value from market segmentation.

By better understanding the variables in the survey and defining the segments they help us see our customers in high definition.

VIEW ARTICLE

Is Data really a currency?

When people discuss currencies they tend to think of paper notes—American dollars, Japanese yen, euros or pounds. Printed money, however, is only one kind of currency. Currencies have evolved over time from stones and seashells to the sophisticated forms of legal tender that enable today's global financial transactions. The evolution of currency continues today as new, alternative currencies grow in popularity.

In this new digital world expectations are increasing. Today's customers demand products and services that are personalised, reliable and durable, at the times and in places they want them. Thanks to the large amounts of data being made available by the billions of connected devices out there, it's easier than ever for businesses to meet these expectations. It's because of these developments that data is believed to be the new currency.

Data has become more valuable than ever, with the increased use of social media and the internet for everything from: research, connecting with others, and shopping to running and promoting a business, and data mining. Companies are utilizing data from several sources to target people with advertising that appears coincidental, yet is based upon hours of data. In this internet savvy world data is the newest currency on the market and it is creating value from what some may perceive as junk.

To understand how data fits into this evolution we must rethink our conception of currencies. Currency is how we create and exchange economic value across geography and through time. It is anything that can serve as a medium of exchange, something that can be "cashed out" for goods and services, or used to pay a debt or to store value for future use. Data has each of these essential characteristics. Because many business transactions involve buying and selling data, it can serve as a medium of exchange – as a young musician said, instead of sending her royalties, streaming music services should provide her with data about her listeners.

Web browsers, such as Google, are analysing the habits of its users constantly. The search engine then stores this data, relating to people's browsing habits, and sells it to their advertisers. The advertisers are then offered the opportunity to have their ads appear on web pages that support advertisers. For instance, if a user is shopping for a child's bedroom set and then decides to head over to Facebook to ask a question to the mum group she belongs to, an advertisement for a children's furniture store could appear on her Facebook account. The user will think it is just another random ad, however, it is a strategically placed, targeted ad, based on her online shopping habits that day.

The value of data can also be measured easily, as many of today's most successful companies have demonstrated, as data appreciates in value when translated into meaningful information. For instance, it was estimated that retailers would be paying major US banks $1.7 billion a year in 2015 to send targeted discount offers to customers based on information on shopping habits gleaned from credit card records.

Aggregated data from internet browsing and social media profiles has become a line of revenue for many companies. The advertising arms of search engines and social media sites can sell consumer data to advertisers as this ensures increased traffic to the advertiser's website. Advertisers in turn use this data to raise their revenue because they can ensure more qualified leads. All the while the user is being guided, based on their usage and habits, to websites and retailers that appear to be at the top of the buyer's needs and wants list.

Beyond reviewing your customer base's search history and buying patterns, social media companies are aggregating personal data from their users to help direct your marketing efforts. Users of Facebook, Twitter, Instagram, and other social media sites provide data including: age, location, family size, employment, education, interests and likes to the companies to be more transparent. However, this data becomes a tool with which the advertiser can target their audience specifically, saving time and money. This data has become a large stream of revenue for social media, allowing marketers to target the specific audience required for each campaign.

The simplest mobile phone call is a miniature data gold mine. Two people talking on their mobiles only think about their conversation, but as they chat data is being collected about their location, the time and length of their call; possibly even the way that this data correlates to their other digital activities, online or social Everywhere we turn we leave behind "digital breadcrumbs," and advertisers and service providers are "churning through thousands of bits of data to figure out what offers to give each of us next, and at some point, your content and experiences begin to be managed."

We should face the fact that if humans are in the process chain, it is not going to be possible to keep financial track of all the services created on the fly by systems in real time. One of the main growth factors for the internet of things is that data will come mostly from devices and not humans, although humans set the policies for gathering the data in the first place. Similarly, we can predict that computers that receive data will be enabled to make decisions based on that information, albeit under guidance and constraints. In most cases those decisions will be executed by computers who send requests for additional actions out to other devices, again mediated by human-generated policies. Driving the implementation of a billing plan, rates, discounts and commitments could well be executed by policy-driven computers too.

A universal truth in business is that all roads lead to data. In an increasingly complex and connected world the ability of an organisation to collect, manage and analyse data effectively separates the winners from the losers. This new wave of data will become the future of online advertising. However, while it is necessary to direct your marketing efforts, this increased availability of data can come at a hefty price. ●

Defining a Data Scientist

It's been said that Data Scientist is the "sexiest job title of the 21st century." But, why is it such a demanded position? The short answer is that over the last decade there's been a massive explosion in both the data generated and retained by companies, and individuals. Sometimes we call this "big data," and like a pile of lumber we'd like to build something with it. Data scientists are the people who make sense out of all this data and figure out just what can be done with it.

Data science is a multidisciplinary field that combines the latest innovations in advanced analytics, including machine learning and artificial intelligence, with high-performance computing and visualisations. The tools of data science originated in the scientific community. Over the past decade computing costs have shrunk and software has become more sophisticated, causing data science to gradually enter business, government, and other sectors.

Any company, in any industry that crunches large volumes of numbers possesses lots of operational and customer data. They may also have access to that data generated by social media streams, credit data, consumer research, or third-party data sets. Companies with data on this scale can benefit from having a data scientist, or a data science team.

Most data scientists have advanced degrees and training in maths, statistics, and/or computer science. Most likely they have experience in data mining, data visualisation, and/or information management. Previous work with cloud computing, infrastructure design, and data warehousing is also common. On a personal level, they are highly curious and passionate about problem solving and accuracy.

Put simply, data scientists apply powerful tools and advanced statistical modelling techniques to provide solutions and insights about business problems, processes, and platforms. But, let's be clear: big data is not a science project. Rather, it must be operationalised in specific ways through more personalised offers to customers and prospects, better insight into pricing trends, and closer tracking of customer behaviours across channels. However, to do this effectively and efficiently at a larger scale requires that someone continuously seek the highest performance and rethink the possibilities afforded by the data.

Therefore, data scientists are the ones experimenting with intelligence-gathering technologies, developing sophisticated models and algorithms, and combining disparate data sets. They will ask the biggest most improbable-seeming questions. They will lead the deepest data mining expeditions and boldest explorations into the largest and most diverse data sets. Or maybe just help you identify the whiskies you might like best.

They also enrich data's value, going beyond what the data says to what it means for your organisation. In other words, it turns raw data into actionable insights that empower everyone in your organization to discover new innovations, increase sales, and become more cost-effective. Data science is not just about the algorithm, but about deriving value.

But, what do the capabilities of data science mean for businesses? Businesses are continually seeking competitive advantage, where there are multiple ways of using data to underpin strategic, operational, and execution practices. Business personnel today, especially with millennials (comfortable with the open-ended capacities of Siri, Google Assistant, etc.) entering the workforce, expect an intelligent and personalised experience that can help them create value for their organisation. In short, data science drives innovation by arming everyone in an organisation, from frontline employees to the board, with actionable insights that connect the dots in data. This brings the power of new analytics to existing business applications and enables new intelligent applications.

Data scientists bring a critical set of problem-solving skills companies need to win with data, but they are just one piece of the puzzle that must be complemented by executive sponsors, marketing data experts, and business analysts, each of which have similarly important roles to play. ●

Data Massaging: the benefits of a good massage

So, we have a database and we need to come up with a data visualisation of what it contains. Sound familiar? This may be a straightforward task, but what if the database is not formatted in the way you expect? Or the data is completely unstructured? Sounds like you may need to massage the data.

The term data massaging, also referred to as “data cleansing” or “data scrubbing”, may sound a bit naughty. But, it’s commonly used to describe the process of extracting data to remove -unnecessary information, cleaning up a dataset to make it useable. Databases come in different shapes and sizes and each must be treated as unique. A few data massaging techniques are required to adapt the data to the algorithms we are working with. Common tasks include stripping unwanted characters and whitespace, converting number and date values into desired formats, and organising data into a meaningful structure. Simply put, massaging the data is usually the “transform” step.

Big organisations can automate this process and use some data massaging tools to systematically examine data for flaws by using rules, algorithms, and simple tables. Typically, a database massaging tool includes programs that can correct several specific mistakes, such as adding missing codes, or finding duplicate records. Using these tools can save a data scientist a significant amount of time and can be less costly than manually fixing errors.

The transformational steps

Things we do to massage the data include:

- Change formats from the standard source system emissions to the target system requirements, e.g. change date format from m/d/y to d/m/y.

- Replace missing values with defaults, e.g. “0” when a quantity is not given.

- Filter out records that are not needed in the target system.

- Check validity of records and ignore or report on rows that would cause an error.

- Normalise data to remove variations that should be the same, e.g. replace upper case with lower case, replace “01” with “1”.

Beyond the initial hypothesis

Basic Exploratory Analysis and Data Crunching are also included in Data Massaging; these techniques can help us explore the stories behind our data.

Using exploratory analysis, we can summarise the data's main characteristics often with visual methods. Some statistical modelling can be used, but exploratory techniques are primarily for seeing what the data can tell us beyond the initial hypothesis; it's the first "taste" of the data.

Data crunching follows the initial exploratory analysis if the data we need to analyse is completely unknown (usually a client's raw data). So, basically all the techniques are allowed here and the main objective is to find a story in the data that could raise more hypotheses and, consequently, more interesting statistical analysis.

Crunching data also involves creation of a system, which will be implemented to carry out specific analysis. The result is data that is processed, structured, and sorted to have algorithms and program sequences run on it. So, crunched data means data that has already been imported and processed in a system. "Data wrangling" and "munging" are two other words used to describe the same process. The latter two terms are used to refer to the initial semi-manual and manual processing of data.

But, what does this mean in practice? Let's say analytics for a company's website shows that it has 10,000 visitors a day. That data, while impressive, is meaningless because it doesn't tell the company what it's doing right to achieve those numbers and what it can do to increase them?

Now let's say the company can separate the number of visitors into those looking for product information, those looking for business partnerships, those looking for jobs, and casual visitors. This begins to make more sense to the company. If it sells directly to consumers, it knows it needs to strengthen its product information and ordering pages. If it depends on resellers, it knows it needs to strengthen its partnership offerings and create a channel for potential business partners to approach it. If it finds that the cost of recruitment through campus interviews, etc. is high, it can tweak the job offers pages so that it gets the right candidates to apply. Doing all this is data crunching in a nutshell.

A bad massage

Unfortunately, there is such a thing as a bad massage. The term "data massaging" is also associated with the practice of "cherry-picking", selectively excluding or altering data based on what researchers want (or don't want) it to reflect. This is the worst and most harmful example of being dishonest with data. Cherry-picking changes the message that the final visualisation communicates to the audience. Though it's illegal there are companies still doing it, mainly with the objective of pleasing clients or to impact publications. For data scientists, facts are sacred and we must always respect what the data is telling us, even though it hurts.

Finally, there is also the less savoury practice of massaging the data by throwing out data (or adjusting the numbers) where they impact on data quality, these are usually outliers. This is absolutely acceptable, but a topic for another article.

As you can see, a lot can be gained from a good massage. Ipsos Connect Data Science is always keen to help and will be running several external data science training sessions during 2018. If you want to know more about Data Science get in touch and we can run sessions specific to your company's needs. ●

Ipsos Connect

ABOUT IPSOS CONNECT

Ipsos Connect are experts in brand, media, content and communications research. We help brands and media owners to reach and engage audiences in today's hyper-competitive media environment.

Our services include:

- **Brand & Campaign Performance:** Evaluation and optimisation of in-market activities to drive communications effectiveness and brand growth.
- **Content and Communications Development:** Communications, content and creative development from early stage idea development through to quantitative pre-testing alongside media & touchpoint planning.
- **Media Measurement:** Audience measurement and understanding.

Ipsos Connect are specialists in people-based insight, employing qualitative and quantitative techniques including surveys, neuro, observation, social media and other data sources. Our philosophy and framework centre on building successful businesses through understanding brands, media, content and communications at the point of impact with people.

FSC
Mixed Sources
Product group from well-managed forests, and other controlled sources
www.fsc.org Cert no. TT-COC-002639
© 1996 Forest Stewardship Council