Data and Statistics

Animal Health Risk Analysis
Canadian Food Inspection Agency
3851 Fallowfield Road
Ottawa, Ontario, Canada K2H 8P9

Analyse des risques de la santé des animaux
Agence canadienne d'inspection des aliments
3851, chemin Fallowfield
Ottawa (Ontario) Canada K2H 8P9

Canadian Food Inspection Agency (CFIA)

Agence canadienne d'inspection des aliments (ACIA)

STATISTICS

is never having to say

you're certain.

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

H.G. Wells

"There are three kinds of lies; lies, damned lies, and statistics." (Benjamin Disraeli, 1804-1881)

"He uses statistics as a drunken man uses lamp posts - for support rather than for illumination." (Andrew Lang, 1844-1912)

"Statistics are mendacious truths." (Lionel Strachey 1864-1927)

"Get your facts first, and then you can distort them as much as you please." (Mark Twain, 1835-1910)

ORIGINS OF STATISTICS

POLITICAL ARITHMETIC

1.1 ***And the Lord spoke unto Moses in the wilderness ofSinai, in the tabernacle of the congregation, on the first day of the second month, in the second year after they were come out of the land of Egypt, saying, Take ye the sum of all the congregation to the children of Israel, after their families, by the house of their fathers, with the number of their names, every male by their polls; From twenty years old and upward, all that are able to go forth to war in Israel; thou and Aaron shall number them by their armies.***

The fourth book of the Old Testament begins with this instruction to Moses to conduct a census of the fighting men of Israel. In the passages that follow we are given the results of that early census, conducted about 1500 B.C. Actually, we know that censuses were carried out in much earlier times for purposes of taxation. Censuses in ancient Babylonia, China, and Egypt apparently were taken as early as 3000 B.C.

Origins of Statistics

1.2 One of the most interesting accounts of an early census is given in the twenty-fourth chapter of the Book of Second Samuel. "And again the anger of the Lord was kindled against Israel and He moved David against them to say, Go, number Israel and Judah." David instructed a reluctant Joab to make a census of the people to determine the number of fighting men. It is recorded that because David did this, divine wrath was visited on Israel and that 70,000 men died of a pestilence.

The census of King David (c. 1500 B.C.) and the resulting punishment seem to have provided a basis for future public resistance of censuses. Governor Hunter of New York reported in 1712 [2], "I have issued out orders to the several counties and cities for an account, of the numbers of their inhabitants and slaves but have never been able to obtain it compleat, the people being deterr'd by a simple superstition and observation, that the sickness follow'd upon the last numbering of the people."[1] When we consider that the early censuses were precursors of military drafts and tax collectors, it is not surprising - that they were resisted by the populace.

[1] From E.B. O'Callaghan, ed. Documents Relative to the Colonial History of the State of New York Vol. V. (Albany, N.Y.: Weed. Parsons & Co., 1855, p. 339

Origins of Statistics

1.3 The word census itself is derived from the Latin word censere, which means to tax. The Roman census was established by the sixth king of Rome, Servius Tullius (534-378 B.C.). Under this system Roman officials called censors made a register at 5-year intervals of the people and their property for taxation purposes and for determining the number of able-bodied fighting men [4]. In 5 B.C. Caesar Augustus extended the census to include the entire Roman Empire. Thus it is that we have the beginning verse in the beautiful and traditional Christmas story: "And it came to pass in those days, that there went out a decree from Caesar Augustus, that all the world should be taxed." To register for such a taxation, Joseph and Mary journeyed to Bethlehem, where the infant Jesus was born. The last regular Roman census was conducted in 74 A.D. With the collapse of the Roman Empire, regular periodic censuses were not conducted in the Western world until the seventeenth century.

Origins of Statistics

1.4 The name statistics can be traced to the Latin words status, meaning state, and statista, meaning statesman. Aristotle (384-322 B.C.) [5] was born in Macedonia, studied under Plato in Athens, and was a tutor to Alexander at the request of King Phillip. He established his own school in Athens when Alexander inherited the throne. The Politeiai of Aristotle contained a description of 158 states. This initial attempt at the comparative description of states was subsequently developed by Italian and German authors into a subject called statistics (Staatenkunde in German). Westergaard [6] traces the development of the description of states.

1.5 During the Middle Ages the system of feudalism more or less rendered national censuses impossible, although there were attempts to revive them. One notable example is the breviary of Charlemagne in 808 A.D. At Christmas in 1085, William the Conqueror ordered a statistical survey of England. The record of this survey is contained in the Domesday Book [4]. The survey collected information of land, landowners, land use, tenants and servants, and livestock and served as the basis for taxes until 1522 when a new Domesday Book was completed.

1.6 Early in the sixteenth century Bills of Mortality (published summaries) began to appear in London. David [3] traces the beginning of these bills to an order from Thomas Cromwell, acting on behalf of Henry VIII. It has been speculated that the king desired these summaries because of his great fear of the plague. However, Cassedy [2] indicates that it is very difficult to determine the beginning of the bills and that the precise date is unknown. In the beginning the Bills of Mortality recorded only deaths from the plague. Over the years they were expanded to include christenings and, around the end of the sixteenth century, data on deaths from other diseases.

Origins of Statistics

1.7 The Spaniards conducted very early censuses in the Americas. A census of Peru in 1548 was carried out by the Spanish viceroy, Don Pedro de la Fasca. This census is described by Carlos A. Uriarte in the March 1949 issue of Estadistica. Before the Spaniards came, the Incas had their own system of recording statistics. This system used intertwined colored strings and knots known as quipus. Cassedy [2, p. 3] quotes historian William H. Prescott in his description of the system of quipus as a method "which has scarcely a counterpart in the annals of a sernicivilized people. A register was kept of all the births and deaths throughout the country and exact returns of the actual population were made to government every year by means of the quipus."[2]

[2]Quoted in Handbook of Vital Statistics (N.Y. Statistical Officer of the United Nations, 1955, p. 4).

Origins of Statistics

1.8 In seventeenth-century England there was great interest in the so-called political arithmetic, which consisted largely of analyses of recorded births and deaths. In 1662 John Graunt published his first and only book, a remarkable manuscript entitled Natural and Political Observations upon the Bills of Mortality. Despite the unreliable nature of the data contained in the Bills of Mortality, Graunt made an exhaustive study of the information contained therein and noted many regularities and irregularities. For example, he noted that the fraction of male births is almost exactly that of female births, the fraction of male births being slightly greater. This observation, well known in our day, seems to have been new and surprising in 1662. Over a long series of years, he counted the christenings of 139,782 boys compared with 130,866 girls. Graunt made a determined if awkward attempt to develop a mortality table of the type used by insurance companies today.

1.9 The word statistics was coined by the German scholar Gottfried Achenwall about the middle of the eighteenth century. Of course, it was derived from the word status and the German counterpart of political arithmetic. The word was apparently used for the first time in Great Britain by Sir John Sinclair who, in a series of volumes published between 1791 and 1799, gave a statistical account of Scotland drawn up from the communications of the ministers of the different parishes. Yule [7] quotes Sinclair as saying, "Many people were at first surprised at my using the new words Statistics and Statistical, as it was supposed that some term in our own language might have expressed the same meaning." [3] It is hard to believe today that such a short time ago the word statistics was considered new.

[3] Reproduced by permission of the publishers. Charles Griffin & Company, Ltd., of London and High Wycombe, from Yule, An Introduction to the Theory of Statistics, 5th Ed., 1919 (15th Ed., Yule & Kendall, 1950)

Origins of Statistics

Boorstin [1] points out that statistics appeared in the Encyclopaedia Britannica in 1797. He also mentions that for a time the word publicistics competed in literary use. It is interesting to speculate about the course of events if publicistics had won. Would there have been an American Publicistical Association or a Royal Publicistical Society?

1.10 When the Constitution of the United States was written, the census became a regular and vital part of the government. The census was provided for in Article 1, Section 2, which states, "The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of ten years, in such manner as they shall by law direct." The first decennial census in the United States was taken in 1790, and other censuses have followed every 10 years since. In addition, censuses in certain fields have been taken at more frequent intervals.

1.11 In this lecture we have sketched a few of the origins of census data and the analysis of such data. The subject of statistics, which owes its name to the description of states, has expanded far beyond these original boundaries, but the modern versions of political arithmetic and description of states constitute an important part of statistics.

Origins of Statistics

SUMMARY

The root word of statistics is state, and the description of states constitutes one of the important roots of modern statistics. Partial descriptions of nations and states were provided by the early censuses, which were conducted in order to raise armies and establish tax rolls. Many examples of censuses can be found in the Old Testament, and the New Testament begins with the account of a census of the Roman Empire conducted by Caesar Augustus.

German and Italian scholars developed the description of states into a subject that more nearly resembles modern statistics. Under William the Conqueror, a detailed statistical survey was made of England and recorded in the Domesday Book. In seventeenth-century England political arithmetic flourished. Consisting largely of analyses of recorded births and deaths, this provided some of the first mortality tables.

The census was required by the Constitution of the United States, although the word statistics itself was not firmly established until 1797, when it first appeared in the Encyclopaedia Britannica.

Data and Statistics *CFIA/ACIA*

Origins of Statistics

REFERENCES

1. Boorstin, Daniel. (1973). The Americans: The Democratic Experience, New York: Random House.

2. Cassedy, James H. (1969). Demography in Early America, Cambridge, Mass.: Harvard University Press.

3. David, F. N. (1962). Gaines, Gods and Gambling, New York: Hafner.

4. Dudley, Lavinia P., Exec. Ed. (1957). Encyclopedia Americana, New York: Americana Corporation.

5. Farrington, Benjamin. (1965). Aristotle, Founder of Scientific Philosophy, London: Weidenfeld & Nicolson.

6. Westergaard, Harald. (1968). Contributions to the History of'Statistics, New York: Agathon Press.

7. Yule, G. Udny. (1919). An Introduction to the Theory of Statistics, London: Charles Griffin.

Statistics

Introduction

It is no more possible for a business executive or an economist to function without a knowledge of statistics than for a physicist to function without a knowledge of mathematics. In order to answer a question or problem arising in a business situation, one has to gather information or data. This data must be collected properly, analyzed efficiently, presented in an orderly and coherent fashion and interpreted correctly. All of these functions fall into the realm of statistics, whether it relates to stock market activity, weather forecasts, unemployment rates, opinion poll results, etc.

A company planning to market a new brand of soap needs to know, among other things, the market potential, the expected sales, and what its competitors are doing. Much of this information is quantitative in nature and must, therefore, be analyzed with those tools and techniques specifically developed and designed to handle quantitative information. Furthermore, the information is often incomplete, in that the full market potential is rarely completely known, the expected sales are calculated based on assumptions only, and one cannot always know what one's competitors are doing. In brief, the businessman must make decisions in the face of uncertainty.

Definition: Statistics is a discipline developed to extract the relevant facts from a large body of information and help people make decisions when uncertainty exists concerning the information.

Statistics

Statistics is a tool to assist you in your area of study and interest. It deals primarily with occurrences or events which cannot be predicted with certainty! The origins of statistics stem as far back as during the Greek and Roman empires when information was collected on its citizenry, often for tax purposes, which today our government refers to as a census, while the rudiments of probability theory stem from the 1600's when mathematicians of renown such as Pascal, Descartes, Bernoulli, De Moivre and Gauss forged the concept of likelihood in games of chance. In our own century, the development of statistics, has mushroomed and become one of the fastest growing areas of research at the hand of numerous scientists such as Fisher, Pearson, Gosset, Tukey, Neyman, Wald, Box, etc.

The field of statistics comprises basically two areas: descriptive statistics and inferential or analytical statistics. Descriptive statistics is concerned with summarizing and describing a set of data in a concise, clear, useful and informative manner. This can be achieved through such techniques as graphing, tabular presentation and calculation of averages and standard deviations. For instance, census or survey data can be so immense that it becomes necessary to condense and interpret this information to make it meaningful and useful. Since this process may distort the results, an important task of descriptive statistics is to carefully limit such distortions. Inferential statistics is the process of drawing conclusions from data under uncertainty and attempting to generalize these conclusions. Although we cannot make correct statements with certainty, we can make statements which have a high probability of being correct.

Basic Components of Statistics

Collection of Data

Question?
Hypothesis?

Definition and Design

Experiment or Survey

Organization

Results

graphs
tables
charts
summary calculations
- **median**
- **mean**
- **mode**

variance
range
coefficent of variation

Analysis

Testing and inference
- **z test**
- **t test**
- **F test**
- **χ^2**
- **ANOVA**
- **regression**
- **correlation**

Interpretation

Prediction and estimation to population

Statistics

Example: The sales of a particular toothpaste are affected by its flavour. The manufacturer wishes to know if sales can be increased by doubling the amount of flavouring that goes into each tube; the company believes that increased production costs will be more than offset by increased sales. To test this idea, a small amount of the more highly flavoured toothpaste is produced and given to a sample of 10,000 families for their reaction. Based on the number of favourable reactions to the new product, the manufacturer will be helped in making his decision to market the new brand of toothpaste.

- to summarize the collected data and evaluate the percentage of sampled families favouring the new product descriptive statistics

- can the results be probabilistically generalized to all the families in the nation? inferential statistics

- how to select the 10,000 families who are to receive the experimental brand sampling

Statistics

Basic Components of Statistics

Collection of data:

The quality of the output of a statistical analysis depends as much on the quality of the input data as in the computational, analytical and interpretive processes. For this reason, careful consideration must be given to experimental or survey design, sampling and clearly defined objectives.

Organization of data:

Collected data must be presented in a suitable and effective manner, conducive for deriving logical considerations. The tabular method.of presentation facilitates the reference, comparison and interpretation of data while the graphic presentation points to trends and presents data quickly and easily comprehended.

Analysis of data:

The data is suitably submitted to various mathematical processes to yield averages, proportions, deviations, etc. for comparison and interpretation.

Statistics

Interpretation of data:

Conclusions are drawn from the analyses performed to extrapolate meaning present in the collected data.

All areas of study possess their own special vocabulary and statistics is no exception.

Definition: A population refers to the universe or totality of items or units under consideration to which the results will be generalized. Populations can be finite or infinite.

Definition: A frame is a listing of all the elements or units in a population.

Definition: A sample consists of a group of units or items chosen from the population because measurements on the entire population cannot or will not be made.

Statistics

Note: Sampling is of utmost importance as it saves time, money and energy. For example, firms continually have to test whether the materials they receive from suppliers conform to specifications of quality and performance. Since it is frequently too expensive to test all incoming materials, firms must base their decisions on testing only a sample of them. It is often even impossible to access the entire population such as when testing requires destruction of the materials or a political poll is taken. A survey refers to a sample while a census refers to a population as it attempts to include all the units or elements in the frame. A sample should be representative of the population from which it is derived!

Definition: A characteristic is a variable which can be measured for each unit or member of a population.

For the above example, the population constitutes all toothpaste users in the country while the sample is the group of 10,000 families on which the data is to be obtained. The characteristic measured is their opinion of the new toothpaste.

Definition: Data comprise outcomes or responses observed from the sample units.

Definition: A parameter is a constant expressed as a function of the population values.

Definition: A statistic is a number obtained as a function of the sample data.

In the above example, the data are the number of people in favour and the number not in favour of the new brand of toothpaste in the sampled 10,000 families. The statistic of interest could be the proportion of people in the sample who would prefer the new toothpaste while the corresponding parameter would be the proportion of people in the population, or country, who would choose the new brand of toothpaste.

Note: Descriptive statistics attempt to summarize and describe the sample data information while inferential statistics use sample statistics to draw conclusions about the true population parameters with some probability, which is the likelihood that sample results reflect those of the population.

Definition: Experimental or sampling errors occur because of a large number of uncontrolled factors, such as instrument measurement limitations, human error or inconsistency, lack of homogeneity among the units, etc., which we can subsume under the term chance or random error. These types of errors can reasonably be expected to cancel each other out over a period of time or over a large number of experiments or samples. In contrast, bias consists of a systematic and persistent type of error which will not tend to cancel out over time or as the sample or the experiment is repeated a large number of times, such as human bias, equipment maladjustment, etc.

Note: With the advent of the computer, available statistical methods and procedures have skyrocketted. Now, a multitude of statistical packages are available on mainframe and micro-computers sixch as SPSS, SAS, Minitab, IPS, etc. one must, however, properly understand the concepts, techniques, and their underlying assumptions before effectively using sophisticated calculator or computer programs to perform statistical analyses.

Counting Techniques

Multiplication Principle:

If there are k operations, and if the first can performed in n_1 ways, and If no matter how the first operation is performed a second operation can be performed in n_2 ways, and if no matter how the first and second operations are performed a third operation can be performed in n_3 ways, and so on for k operations, then the k operations can be performed in $n_1 \cdot n_2 \cdot n_3 \cdot ... \cdot n_k$ ways.

Example - A license plate is to consist of three digits followed by two letters other than 0 and U. How many possible distinct license plates can be produced in this fashion? 10x10x(10+26)x24x24

Example - A restaurant menu offers four appetizers, ten entrees, three beverages and six desserts. How many possible dinners can this restaurant serve if an item from each category must be selected? 4x10x3x6

Definition: Given the positive integer n, the product of all positive integers less than or equal to n is called n factorial and denoted by n ! so that n ! = n(n-1)(n-2)(n-3) ... 1.

Note: 0 ! = 1.

Counting Techniques

Factorial Principle: The number of ways n objects can be arranged in order is n ! (factorial notation).

Example - In how many ways can a supermarket manager display seven brands of cereal on a shelf? 7x6x5x4x3x2x1

Example - How many displays can this supermarket manager create if he has only three spaces on his shelf now for his seven brands of cereal? 7x6x5

Note that

$$7 \times 6 \times 5 = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} = \frac{7!}{4!}$$

At such times when we have more objects than we have positions to fill, we seek the number of possible ordered arrangements of these n objects in r positions, where r < n, which we call permutations.

Counting techniques

Permutation Rule: If we have n distinct objects from which an ordered arrangement of r objects is to be derived, the number of such permutations is given by

$$_nP_r = \frac{n!}{(n-r)!}.$$

The total number of arrangements of n objects (n !) only has the factors (n) (n-1) (n-2) ... (n-r+1) since only r positions are to be filled, the first having n possible slots, the second n-1 slots, the third n-2 slots, ..., the rth. n-r+1 slots available. Hence, is the number of permutations of n things taken r at a time.

Note: The number of permutations of n objects taken all at a time then reduces to the factorial principle since

$$_nP_n = \frac{n!}{0!} = n!.$$

Counting Techniques

Example - In how many ways can the first, second, third and fourth prize be awarded to ten entries in a high tech exhibition?

$$_{10}P_4 = \frac{10!}{6!}$$

A department store manager is interested in arranging a window display with two different colours by choosing from three different available colours: green, yellow and red. The number of permutations of three different colours taken two at a time is

$$_3P_2 = \frac{3!}{1!} = 3 \times 2 = 6,$$

namely, GY GR TR
YG RG RY

However, the three columns of permutations each represent the same set of colours; these are called combinations, arrangements of objects without regard to order. Thus, the number of combinations of three different colours taken two at a time (3) is the number of permutations of three different colours taken two at a time (6) divided by the number of permutations of colours in the selection (2).

Counting Techniques

Combination Rule: The number of combinations of n objects taken r at a time in given by

$$_nC_r = \frac{n!}{(n-r)!r!}.$$

Since, for r objects taken all together at a time, we obtain one combination and r ! permutations, it is obvious then that

$$_nC_r = \frac{_nP_r}{r!} = \frac{n!}{(n-r)!r!}.$$

Example - In how many ways can Revenue Canada choose three of 36 tax returns for a special audit?

$$_{36}C_3 = \frac{36!}{33!3!}.$$

"A mean by any other name is still a mean."

Weighted Mean

For instance, the Consumer Price Index, a prominent measure of the rate of inflation, is a weighted mean of the relative changes in the prices of various goods and services.

Example: A sample of five firms exhibit profit rates of 10, 12, 15, 16 and 18%. The firms with 15, 16 and 18% profit rates have assets of $1.25 billion, $1 billion and $1.5 billion respectively while the other two firms each have assets of $2 billion. Find the mean of the firms' profit rates.

Geometric Mean

Example: Consider the following population figures for a small community:

		Percentage Rate of Change	Population x	
Year	Population	(of previous period)	122.7%	122.56%
1940	5000			
1950	6000	120	6135	6128
1960	7800	130	7528	7510
1970	9204	118	9237	9204

Calculate the suitable average rate of change in population for these three times periods.

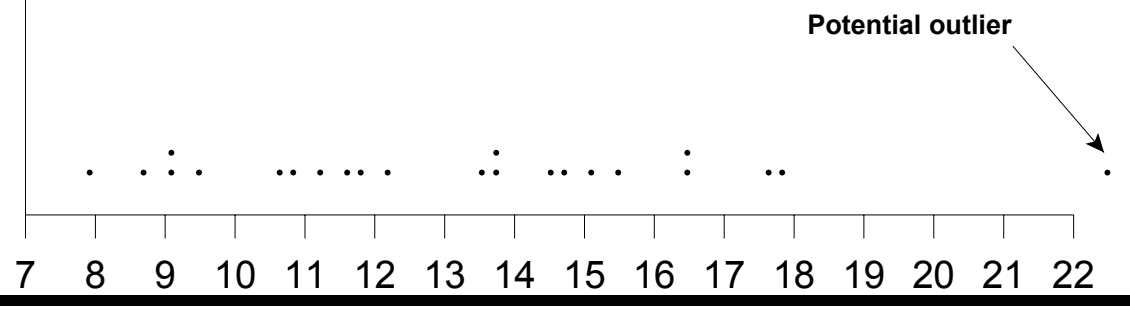
$$\overline{X} = \frac{1}{3}\sum_{i-1}^{3} X_i = \frac{120 + 130 + 118}{3} \approx 122.7\%$$

$$\log \overline{X}_G = \frac{1}{3}\sum_{i-1}^{3} \log X_i = \frac{1}{3}[2.079 + 2.114 + 2.072] \approx 2.0883 \qquad \Rightarrow \overline{X}_G = 10^{2.0833} \approx 122.56\%$$

If each year's population figure is multiplied by 122.7%, the 1970 figure of 9237 disagrees with the actual value of 9204, indicating that the arithmetic mean rate of change does not accurately reflect the growth situation whereas the geometric mean does as the 1970 population figure agrees with the actual one when each previous year's population is multiplied by 122.56%. Hence, the suitable average rate of change is best expressed in terms of the geometric mean. $\overline{X}_G = 122.56\%$

Dot Diagram

Range is a single value, the positive difference between the smallest and largest valure. The range is not an interval, it is one value. The minimum value is 7.9, the maximum value is 22.3. The range is (22.3-7.9) = 14.4.



Quantiles or Percentiles or Fractiles

Quantiles are measures of noncentral location often utilized in summarizing or describing properties of large samples of quantitative data and can assist in determining its empirical distribution. They are quantities which divide the ordered data into equal portions. Percentiles divide it into hundredths, deciles into tenths and quartiles into quarters.

We first arrange our observations in ascending order of magnitude. If p is then a fraction where O<p<1, the 100pth percentile is that piece of data such that at least 100p% of the observations are at or below this value and at least 100(1-p)% are at or above this value. As with the median, if two observations fit this definition, their mean is used as the percentile value.

Of these percentiles, the 25th, 50th and 75th are most routinely used in applications and are by definition the first, second and third quartiles, denoted as Q1, Q2 and Q3, respectively. Consequently, Q1 is the value for which 25% of the observations are smaller or equal to it and 75% of the observations are larger or equal to it while Q2 is the median.

Oil Firm Profit Rate Data (%)

Descriptive Statistics

Ordered Data
7.9, 8.7, 9.0, 9.0, 9.6, 10.5, 10.7, 11.2, 11.7, 11.9, 12.4, 13.3, 13.4, 13.4, 14.4, 14.5, 14.9, 15.5, 16.2, 16.2, 17.7, 17.8, 22.3

Mean = μ = 13.14%

Median = Q2 = 13.3%

Mode(s) = 9%, 13.4%, 16.2%

Midrange = 15.1%

1st Quartile = Q1 = 10.5%

3rd Quartile = Q3 = 15.5%

6th Decile = 60th Percentile = 13.4%

Trimmed Mean = 12.91%

Winsorized Mean = 12.95%

Range = R = 14.4%

Interquartile Range = I.R. = 5%

Variance = s^2 = 12.3743 (% 2)

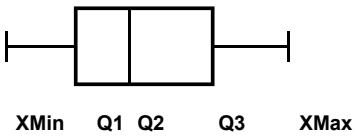
Standard Deviation = s = 3.52%

Coefficent of Variation = C.V. = 26.8%

Coefficent of Skewness = SK = -0.136

Box-and-Whisker Plot

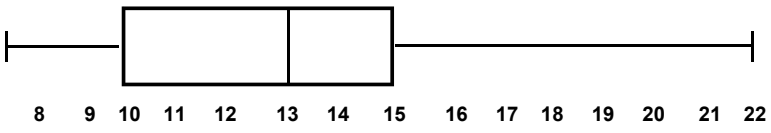
A single plot, known as the Box-and-Whisker plot, allows a graphical perspective of the type and shape of the underlying distribution by visually summarizing the range and quartiles of the sampled data. The box represents the middle 50% of the observations in the dataset while the dashed lines, or whiskers, represent the lower and upper 25 % of the data as follows:



The above indicates that the data follows a positively skewed distribution.

Example: Construct a box-and-whisker plot for the oil firm profit rate data.
Here, X Min = 7.9, X Max = 22.3, Q2 = median = 13.3, Q,1 = 10.5, and Q3 = 15.5.

Box-and-Whisker Plot of Profit Rates (%) for 23 Major Oil Firms



Trimmed and Winsorized Means

i) Trimmed Mean:

Remove all the observations below the first quartile and those above the third quartile and calculate the arithmetic mean of the remaining data.

ii) Winsorized Mean:

Replace each observation below the first quartile by the first quartile value and each observation above the third quartile with the third quartile value keeping all other observations unchanged and then calculate the mean of all these observations so modified.

For the oil firm profit rate data, compute the trimmed mean and the Winsorized mean.

Trimmed Mean

$$= \frac{1}{13}\{10.5 + 10.7 + 11.2 + 11.7 + 11.9 + 12.4 + 13.3 + 13.4 + 13.4 + 14.4 + 14.5 + 14.9 + 15.5\}$$
$$= \frac{167.8}{13}$$
$$= 12.91\%$$

Winsorized Mean $= \frac{1}{23}\{5x10.5 + 5x15.5 + 167.8\} = \frac{297.8}{23} = 12.95\%$

Coefficient of Variation

When comparing the dispersion of two data sets, the observations being expressed in different units of measurement renders direct comparison impossible. For instance, do young female executives for a firm vary more in weight than in height? One variable is measured in centimetres while the other is in kilograms! Hence, we require a measure of relative dispersion, a single quantity measuring dispersion without being affected by the unit of measurement used for the actual data.

The coefficient of variation answers this requirement and gives a comparison of the average variability to the mean in a data set while being unit free. The standard deviation is expressed as a percentage of the mean and a small percentage indicates a rather homogeneous group of observations. The coefficient of variation is thus expressed as: $C.V. = \frac{s}{\bar{x}} \times 100\%$

Example - For the profit rate data found in the example above, evaluate its coefficient of variation.

$$C.V. = \frac{s}{\bar{x}} \times 100\% = \frac{3.52}{13.14} x100 = 26.8\%$$

Therefore, the variability in the profit rates is slightly over one quarter of the mean which is relatively homogeneous. The coefficient of variation can also assist in comparing the relative variability of distributions expressed in the same units of measurement such as say the variability of morning and evening sales at several large department stores.

Coefficient of Variation

Example - The mean and standard deviation of the average morning and evening sales at twelve branches of a large department store are as follows:

Morning sales: $\bar{x} = \$43.22 \quad s = \$3.260$
Evening sales: $\bar{x} = \$68.46 \quad s = \$2.254$

Which time of day are.the sales less variable?

For morning sales, $C.V. = \frac{3.26}{43.22} \times 100 = 7.54\%$ **while**

for evening sales, $C.V. = \frac{2.254}{68.46} \times 100 = 3.29\%$.

Therefore, the evening sales are more homogeneous or less variable.

There also exist many other measures such as skewness and kurtosis which deal primarily in describing the shape and form of the data distribution since absolute measures for all distributions are rare indeed.

Coefficient of Skewness

We know that the mean and median values coincide for a perfectly symmetrical distribution while they move further apart as extreme, values or outliers skew the distribution more and more. Consequently, we can use this relationship between the man and the median to define a relatively simple measure of the extent to which a distribution is skewed, known as the Pearsonian coefficient of skewness and given as:

$$SK = \frac{3(mean - median)}{s\tan dard\ deviation}$$

For a perfectly symmetrical distribution, the value of SK is 0 while generally its values must fall between -3 and 3.

Example - What is the value and interpretation of the coefficient of skewness for our profit rate data in example 17?

This value is sufficiently small to conclude that the distribution of our profit rates for the 23 major oil firms is nearly symmetrical.

$$SK = \frac{3(\bar{x} - Q_2)}{s} = \frac{3(13.14 - 13.3)}{3.52} = \frac{-.48}{3.52} = -0.136$$

Profit Rates of 23 Major Petroleum Companies, 1973

Firm	Profit rate [a] (percent)
Exxon	17.8
Texaco	16.2
Mobil	14.9
Gulf	14.4
Standard Oil of California	14.5
Standard Oil of Indiana	12.4
Shell	10.7
Continental	13.4
Atlantic Richfield	8.7
Occidental	9.0
Phillips	11.7
Union	10.5
Sun	11.9
Ashland	15.5
Cities Service	9.6
Getty	9.0
Marathon	16.2
Standard Oil of Ohio	7.9
Pennzoil	13.3
Kerr McGee	11.2
Murphy	22.3
Commonwealth	17.7
American Petrofina	13.4

SOURCE: Fortune, 1974. [a] Profit rate is defined here as net income as a percent of stockholders' equity.

Kinds of Data

In studying a process, we necessarily acquire information about it by making observations on the set of items of interest. we thus accumulate a set of data. Data and variables may be qualitative or quantitative in nature.

Definition: A qualitative variable yields categorical responses such as yes, no or satisfactory, mediocre, unacceptable while a quantitative variable provides numerical responses such as $50,000, 300 grams, 1000F, 25% or 50 questions per survey. Characteristics leading to qualitative data are called attributes while those providing quantitative results are referred to as variables.

Scales of measurement

Definition: The nominal or labelling scale assigns numbers to items which are really only labels and their magnitude has no real significance.

Example - Males and females are assigned respectively the numbers 1 and 2 as responses to a question on a questionnaire.

Definition: The ordinal scale allows the rank ordering of a set of conditions or characteristics where the assigned numbers imply relative magnitude.

Example - The dryness of champagnes can be rank-ordered from sweet to dry as doux, demi-sec, dry, extra dry and brut or as number 1, 2, 3, 4, 5 or as number 8, 22, 23, 38, 45. Note: There is no implication that a demi-sec (#2) is twice as dry as a doux (#1) !

Kinds of Data

Definition: The interval scale has all the properties of an ordinal scale but with the exact distance between any two numbers known. Equal differences in the magnitude of events are associated with equal intervals between the assigned numbers but the scale has no real zero point.

Example - The temperatures 30 ° F - 40 ° F and 80 ° F - 90 ° F differ by the same magnitude but there is no true zero for the Fahrenheit scale.

Definition: The ratio scale is the strongest scale of measurement and has all the properties of the interval scale with a meaningful zero point in addition.

Example - Weight is measured on a ratio scale where 0 g is the zero point and 40 g is four times as much as 10 g.

Kinds of Data

Type of Characteristics:	**Attributes**	**Variables**

Type of Data:	**Qualitative (Categorical responses)**		**Quantitative (Numerical responses)**	
Scale of measurement:	**Nominal** -no ordering implied among categories Example: Liberals, N.D.P., Conservatives, others	**Ordinal** - ordering implied but no clear existent measure of differences between categories Highly Satisfactory, Satisfactory, Mediocre, Unsatisfactory grades of meat, eggs, etc.	**Interval** - has no true zero point, temperature in °C, calendar time	**Ratio** - has a true zero point, bank interest, weight, length, profit rate | \ | \ | \ | \ discrete continuous (counting) (measured)

(under interval and ratio measurements the order is obvious and the differences in magnitude of values are easily evaluated)

Variables

Definition: A variable is called discrete if the number of its possible values is finite or countably infinite, that is, as many as there are positive integers. Discrete data is generated from counting.

Example - Consider the experiment of tossing a coin twice. Then the frame is {HH, HT, TH, TT}. If X = the No. of heads obtained, then X is a discrete variable with a finite no. of possible values 0, 1 and 2.

Example - Consider the experiment of tossing a coin until a head appears. Then the frame is expressed as {H, TH, TTH, TTTH, ...}. If X = the no. of tosses required to get a first head, then X is a discrete variable with a countably infinite no. of possible values 1, 2, 3, ...

Definition: A continuous variable has its possible values form an interval or a collection of intervals. Continuous data is measured as any value within certain limits.

Example - The projected return on an investment is a continuous variable since it can take on any value within a particular interval.

THE GREEK ALPHABET

SYMBOL		NAME
SMALL	CAPITAL	
α	Α	alpha
β	Β	beta
γ	Γ	gamma
δ	Δ	delta
ε	Ε	epsilon
ζ	Ζ	zeta
η	Η	eta
θ	Θ	theta
ι	Ι	iota
κ	Κ	kappa
λ	Λ	lambda
μ	Μ	mu
ν	Ν	nu
ξ	Ξ	xi
ο	Ο	omicron
π	Π	pi
ρ	Ρ	rho
σ	Σ	sigma
τ	Τ	tau
υ	Υ	upsilon
φ	Φ	phi
χ	Χ	chi
ψ	Ψ	psi
ω	Ω	omega

Acknowledgements

All of the material in this program was taken from the course notes of Roger Trudel, Senior Statistician with the Canadian Food Inspection Agency, 59 Camelot Drive, Ottawa, Ontario, K1A 0Y9