STATISTICS AND PROBABILITY

Table des matières

Chapitre 1

Introduction to statistics, charts and graph

1.1 Basic concepts

Statistics can be defined as a science dealing with the collection, analysis, interpretation and presentation of numerical data. It is an important decision-making tool in general.

The key elements : Collecting data, analyzing data, interpreting data and presenting results. There are two main branches : descriptive statistics and inferential statistics.

A population is a collection of persons, objects, or items of interest. For various reasons to be explained later, it could be more convenient to work with a sample of the population instead of the entire population. A sample is a portion of the whole and, if properly taken, is a representative of the whole.

The study of statistic can be subdivided into two main areas. Descriptive and inferential statistics.

Descriptive statistics : here, data gathered on a group are used to describe or reach conclusions about that same group.

Inferential statistics : here, researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken.

A descriptive measure of the population is called a parameter. Example : population mean, population variance, population standard of deviation,...

A descriptive measure of a sample is called a statistic. The basis of inferential statistics is the ability to make decisions about parameters without having to complete a census of the population.

A variable is characteristic of any entity being studied that is capable of taking on different values.

The appropriateness of the data analysis depends on the level of measurement of the data gathered. There are four common level of data measurement :

Nominal level : Numbers representing nominal data can be used only to classify or categorize. Example, employee identification numbers. The numbers are used only to differentiate employees and not to make a value statement about them. The following is another example that would result in nominal data.

a. Educator - b. Construction worker - c. Manufacturing worker - d. Lawyer - e. Doctor - f. Other.

Numbers, if used only differentiate educator with constructor, lawyer and so on...

Other types of variable that often produce nominal levels are : sex, religion, ethnicity, geographic location,...

Ordinal : Ordinal-level data measurement is higher than nominal level. In addition to having the nominal-level capabilities, ordinal level measurement can be used to rank the order objects. For example, using ordinal data a supervisor can evaluate three employees by ranking their productivity with the numbers 1 through 3. The supervisor could not use ordinal data to establish that the intervals between the emplyees ranked 1 and 2 and between the employees ranked 2 and 3 are necessarily the same. Below is an example of questionnaires that are considered to be ordinal in level. This computer tutorial is :

——— (1) ——— (2) ——— (3) —— (4) ———(5)

Not helpful Somewhat helpful Moderately helpful very Extremely helpful

Nominal and ordinal data are sometimes called non metric or qualitative data.

Interval level : Interval-level data is the next to the highest level of data, in which the distances between consecutive numbers have meaning and the data are always numerical. The distances represented by the differences between consecutive numbers are equal, that is, interval data have equal intervals. In addition, with interval-level data, the zero point is a matter of convention or convenience and not a natural or fixed zero point.

Ratio level : Ratio-level data measurement is the highest level of data measurement. Ratio data have the same properties as interval data, but ratio data have an absolute zero and the ratio of two numbers is meaningful. The notion of absolute zero means that zero is fixed, and the zero value in the data represents a fixed point.

Examples of ratio data are height, mass, volume,... With ratio data, a reseracher can state that 180 kg of mass is twice as much as 90 kg. Interval and ratio-level data are sometimes referred to as quantitative data.

1.2 Charts and graph

One particularly useful tool for grouping data is the frequency distribution, which is a summary of data presented in the form of class intervals and frequencies.

Example :

Class midpoint : The midpoint of each class interval is called class midpoint and is sometimes referred to as the class mark. It is calculated as the average of the two class endpoints.

Relative frequency : relative frequency is the proportion of the total frequency that is in any given class interval in a frequency distribution. It is the individual class frequency divided by the total frequency.

Cumulative frequency : It is the running total of frequencies through the classes of a frequency distribution. Precisely, the cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total.

1.2.1 Quantitative data graphs.

- **Histogram** : A histogram is a series of contiguous rectangles that represents the frequency of data in given class intervals.

If the class intervals used along the horizontal axis are equal, then the heights of the rectangles represent the frequency of values in a given class interval.
If the class intervals are unequal, then the areas of rectangles can be used for relative comparisons of class frequencies.

- **Frequency polygons** : Like the histogram, it is a graphical display of class frequencyies. However, instead of using rectangles like a histogram, in a frequency polygon, each class frequency is plotted as a dot at the class midpoint and the dots are connected by series of line segments.
- **Ogives** : An ogive is a cumulative frequency polygon. Construction begins by labelling the x axis with the class endpoints and the y-axis with the frquencies. However, the use of cumulative frequency values requires that the scale along the y axis be great enough to include the frequency total. A dot of zero frequency is plotted at the beginning of the first class and construction proceeds by marking a dot at the end of each class interval for the cumulative value. Connecting the dots then completes the ogive.

1.2.2 Qualitative data graphs.

- **Pie charts** : A pie chart is a circular depiction od data where the area of the whole pie represents 100% of the data and slices represent a percentage breakdown of the sublevels. Pie charts show the relative magnitudes of parts to a whole. Construction of the pie chart is done by determining the proportion of the subunit to the whole. The amount of each category is is represented as a slice of the pie proportionate to the total.
- **Bar chart or bar graph** : A bar chart contains two or more categories along one axis and a series of bars, one for each category, along the other axis. The length of the bar represents the magnitude of the measure for each category.

Chapitre 2

Descriptive statistics

The two types of data are grouped and ungrouped. Grouped data are data organized into a frequency distribution. In general, statistical operations on the two types are computed differently.

2.1 Measures of central tendency : Ungrouped data.

One type of measure that is used to describe a set of data is the measure of central tendency. Measures of central tendency yield information about the centre, or middle part, of a group of members.

- **The mean** : The (arithmetic) mean is the average of a group of numbers and is computed by summing all numbers and dividing by the number of numbers. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the mean.
  The population (size N) mean is represented by μ and the sample (size n)mean is represented by $\overline{x}$.
  $\mu = \frac{1}{N}\sum x$ and $\overline{x} = \frac{1}{n}\sum x = \frac{x_1+x_2+...+x_n}{n}$
- **Median** : The median is the middle value in an ordered array of numbers. For an array with an odd number of terms, the median is the middle number. For any array with an even number of terms, the median is the average of the two middle numbers. For the determination, use the following steps.
  - Arrange the observations in an ordered data array
  - For an odd number of terms, find the middle term of the ordered array. It is the median.
  - For an even number of terms, find the average of the middle two terms. This average is the median.
- **Mode** : It is the most frequently occuring value in a set of data. Data with two modes are said bimodal and data with more than two modes are referred to as multimodal.
- **Percentiles** : Percentiles are measures of central tendency that divide a group of data into 100 parts. There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts.
  The nth percentile is the value such that at least n percent of the data are below that data and at most $(100 - n)$ percent are above that value.
  Steps in determining the Location of a percentile.

- Organize the numbers into an ascending-order array
- Calculate the percentile location (i) by : $i = \frac{p}{100} \times n$ where p is the percentage in interest, i the percentile location and n the number in the data set.
- Determine the location by either (a) or (b).
(a) : If i is a whole number, the pth percentile is the average of the value at the ith location and the value at the $(i+1)$th location.
(b) : If i is not a whole number, then the pth percentile value is located at the whole-number part of $i+1$

– **Quartiles** : Quartiles are measures of central tendency that divide a group of data into four subgroups of parts. The three quartiles denoted as Q_1, Q_2 and Q_3.
$Q_1 = P_{25}$, $Q_2 = P_{50}$ = Median and $Q_3 = P_{75}$

2.2 Measures of variability : Ungrouped data

Measures of central tendency yield information about particular points of a data set. Measures of variability describe the spread or thw dispersion of a set of data. Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

– **Range** : The range is the difference between the largest value of a data set and the smallest value of the set. An advantage of the range is its ease of computation. A disadvantage of the range is that, because it is computed with the value that are on the extremes of the data, it is affected by extreme values, and its application as a measure of variability is limited.

– **Interquartile range** : It is the range of values between the first and third quartiles. Essentially, it is the range of the 50% of the data and is determined by computing the value $Q_3 - Q_1$. It is useful in situations where data users are more interested in values toward the middle and less interested in extremes.

– **Mean absolute deviation, variance and standard deviation** : They are obtained through similar process and are, therefore, presented together.
- The mean absolute deviation (MAD) is the average of the absolute values of the deviations around the mean for a set of numbers. $MAD = \frac{1}{N}\sum |x - \mu|$
- The variance is the average of the squared deviations about the arithmetic mean for a set of numbers. The population variance is denoted by $\sigma^2 := \frac{1}{N}\sum (x-\mu)^2$
- The standard deviation is the square root of the variance. The population standard deviation is denoted by :
$\sigma := \sqrt{\frac{1}{N}\sum (x-\mu)^2}$. Denote in the sequel $SS_x = \sum (x-\mu)^2$
- The sample variance is : $s^2 = \frac{1}{n-1}\sum (x-\overline{x})^2$ and the sample standard of deviation, $s = \sqrt{\frac{1}{n-1}\sum (x-\overline{x})^2}$
Computational formula : $\sigma^2 = \frac{1}{N}\sum x^2 - (\frac{1}{N}\sum x)^2$ and $s^2 = \frac{1}{n-1}\sum x^2 - (\frac{1}{n(n-1)}\sum x)^2$
Example : Problem 3.6 page 72
- **The empirical** rule is an important guideline that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed. The normal distribution will be discussed later (in chapter 6). It is a unimo-

dal, symmetrical distribution that is bell (or mound) shaped. (Representation graphique page 68)

We have the following :

Distance from the mean	$\mu \pm 1\sigma$	$\mu \pm 2\sigma$	$\mu \pm 3\sigma$
Values within Distance	68%	95%	99.7%

If a set of data is normally distributed, or bell-shaped, approximately 68% of the data values are within one standard deviation.

- **Chebyshev's theorem :** The empirical rule applies only when data are known to be approximately normally distributed. Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is nonnormal.

The theorem : Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least $1 - \frac{1}{k^2}$ proportion of the value, with the assumption $k > 1$.

Chebyshev's theorem states that at least $1 - 1/k^2$ values will fall within $\pm k$ standard deviations of the mean regardless of the shape of the distribution.

Example : problem 3.5 page 70.

Example : Problem 3.4 page 69.

– **z scores** : A z-score represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed.

For population, we have : $z = \frac{x-\mu}{\sigma}$ and for sample, $z = \frac{x-\overline{x}}{s}$.

If a z-score is positive then the raw value x is above the mean and if it is negative then the raw value (x) is below the mean.

– **Coefficient of variation** : It is a statistic that is the ratio of the standard of deviation to the mean expressed in percentage and is denoted CV. It is essentially a relative comparison of standard deviation to its mean. $CV = \frac{\mu}{\sigma}(100)$

Example : Suppose five weeks of average prices for stock A are 57, 68, 64, 71 and 62. To compute a coefficient of variation for these prices, firts determine the mean and standard deviation :

$\mu_A = 64,4$, $\sigma_A = 4,84$ and $CV_A = \frac{64,4}{4,84} = 0.075 = 7,5\%$. The standard deviation is $7,5\%$ of the mean.

Exercise : Problem 3.11 page 75.

2.3 Measures of central tendency and variability : grouped data

Grouped data do not provide information about individual values. Hence, measures of central tendency and variability must be computed differently.

– **Mean** : The mean is obtained by summing the data values divided by the number of values. The midpoint of each class interval is used to represent all the value in a class interval. Each midpoint is weighted by the frequency of values in that class interval.

$\mu = \frac{\sum fM}{N} = \frac{f_1M_1+f_2M_2+...+f_iM_i}{N}$ where : i is the number of classes, f class frequency, M class midpoint and N the total number of data values ie total frequency.

Example : Table 3.7 page 80

– **Median** : The calculation of the median for grouped data is done by using the following formula.
$Me = l + \left(\frac{\frac{N}{2}-F}{f}\right) w$ where : l is the lower endpoint of the class containing the median, w is the width of the class containing the median, f the frequency of the class containing the median, F is the cumulative frequency of classes preceding the class containing the median and N the total frequencies (total number of data values).
The first step is then to determine the value $N/2$, which is the location of the median term. This can be done by determining the cumulative frequencies. The class interval containing the median value is referred to as the median class interval.

Example : Table 3.6, fin de page 79 et debut page 80.

– **Mode** : The mode for grouped data is the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.

Example : juste au dessus de table 3.6 page 80

– **Measures of variabilities** : We present here two measures of variabilities : the variance and the standard deviation.
- For Population variance and standard deviation :
$\sigma^2 = \frac{\sum f(M-\mu)^2}{N}$, computational formula : $\sigma^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{N}}{N}$ and $\sigma = \sqrt{\sigma^2}$where :
f is the frequency, M the class midpoint, $N = \sum f$ the total frequencies of the population and μ the grouped mean for the population.
- For sample variance and standard deviation.
$s^2 = \frac{\sum f(M-\overline{x})^2}{n-1}$, computational formula : $s^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{n}}{n-1}$ and $s = \sqrt{s^2}$ where :
f is the frequency, M the class midpoint, $n = \sum f$ the total frequencies of the sample and $\overline{x}$ the grouped mean for the sample.

Example : Table 3.8 et 3.9 page 81 and 82.

2.4 Measure of shapes

– Measures of shapes are tools that can be used to describe the shape of a distribution of data. In this section we examine two measures of shape, skewness and kurtosis.

A distribution of data in which the right half is a mirror image of the left half is said to be symmetrical. One example of a symmetrical distribution is the normal distribution, or bell curve, to be presented later.

Skewness is when a distribution is asymmetrical or lacks symmetry.

Donner ici les trois types : page 86.

Symmetrical distribution, distribution skewed left or negatively skewed, distribution skewed right or positively skewed.

– Skewness and relationship with the mean, median and mode.
Page 87, les trois figures

– **Coefficient of skewness**. The following coefficient defined by Karl Pearson can be used to determine the degree of skewness in a distribution. This coefficent compares the mean and median in light of the magnitude of the standard deviation.

$S_k = \frac{3(\mu - M_d)}{\sigma}$ where S_k is the coefficient of skewness, M_d is the median and μ the mean.

Remark : If the distribution is symmetrical, then the mean and median are the same value and hence the coefficient of skewness is equal to zero.

A positive S_k means that the distribution is positively skewed and a negative S_k means that the distribution is negatively skewed. The greater the magnitude of S_k, the more skewed is the distribution.