EMS702P Statistical Thinking and Applied Machine Learning

Week 2.1 – Descriptive Statistics & Exploring Data

Jun Chen

Introduction to Statistics and Machine Learning

Edition: v1.1

Table of Contents

1 Descriptive Statistics **1**
1.1 Frequency Tables 1
1.2 Measures of Location 4
1.3 Measures of Spread 7
2 Exploratory Data Analysis & Visualisation **10**
2.1 Five-number Summary Statistics 10
2.2 Bar Charts 11
2.3 Histograms 12
2.4 Pie Charts 13
2.5 The Box-and-Whisker Diagram 13
3 Irregularities **14**
3.1 Outliers 14
3.2 Skewness 17
3.3 Multimodal Distribution 18
4 Further Reading **18**

1 Descriptive Statistics

Descriptive statistics are brief informational coefficients that summarise a given data set, which is *often* a **sample** of a population. If a **representation** of the entire population is available (not very often), techniques similar to descriptive statistics can also be used to summarise the quantities of the population.

Descriptive Statistics concentrate on the basic tabular and diagrammatic techniques for displaying data and the calculation of elementary statistics. *A statistic is a numerical summary of a sample of data.* Descriptive statistics consist of three basic categories of measures: measures of **frequency distribution**, measures of **central tendency** (or location), and measures of **variability** (or spread).

- Measures of frequency distribution describe the occurrence of data within the data set (count).
- Measures of central tendency describe the centre of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).

In Week 3, we will consider formal methods for making *inferences* about a population based on observation of a sample. Before doing this we need to consider simple methods for describing and summarising the observations in a sample, both *numerically* and *graphically*. These descriptive methods are used

(1) as an important first step before we go on to apply more formal analyses,
(2) to report results of learning algorithms, both numerically and graphically.

1.1 Frequency Tables

Frequency: The number of times of a certain *value*, in the case of categorical or discrete variables, or *class* defined over a *class interval*, in the case of continuous variable, is observed in the sample.

Total Frequency: The total frequency is the total number of observations.

Relative Frequency: The relative frequency is the proportion of the observations which fall in a particular category, value, or class. It is the frequency divided by the total frequency.

Example 1

Suppose that we observe vehicles at a crossroads for thirty minutes. Consider just the vehicles arriving at the junction from the South. Such vehicles can carry straight on to go North, turn left to go West or turn right to go East. Suppose that we see 147 of these vehicles go North, 85 go West and 43 go East.

Solution

Example 2

The numbers of vehicles passing a point on a road in each of 100 intervals of length one minute are recorded in the frequency table below. Calculate the cumulative frequency, relative frequency and cumulative relative frequency and fill them out in the table below.

Solution

Number of Vehicles x	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
0	0			
1	0			
2	0			
3	4			
4	3			
5	5			
6	8			
7	10			
8	19			
9	12			
10	13			
11	7			
12	4			
13	11			
14	0			
15	1			
16	2			
17	0			
18	1			
$x > 18$	0			
Total	100			

Example 3

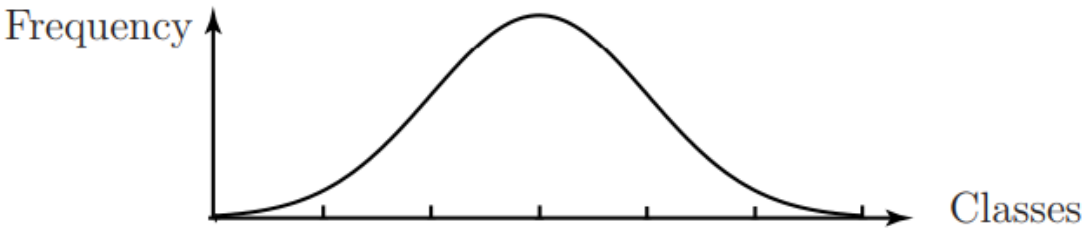
The following data are the heights of a second sample of 30 students studying engineering statistics. Organise the data into classes using class intervals $[145, 150)$, $[150, 155)$, . . . , $[185, 190)$ and construct a frequency table of the data.

155.3	177.3	146.2	163.1	161.8	146.3	167.9	165.4	172.3	188.2
178.8	151.1	189.4	164.9	174.8	160.2	187.1	163.2	147.1	182.2
178.2	172.8	164.4	177.8	154.6	154.9	176.3	148.5	161.8	178.4

Solution

1.2 Measures of Location

The statistical properties of the observations in a sample are often described as a frequency distribution, referring to the fact that some values, or some ranges of values, may have greater frequencies than others, as shown below.



There are three widely used measures of location.

Mean: the *arithmetic* average of the data.

In the case of discrete data, the mean is calculated as follows

$$\bar{x} = \frac{\sum_{j=1}^{J} f_j x_{(j)}}{\sum_{j=1}^{J} f_j} = \frac{\sum_{j=1}^{J} f_j x_{(j)}}{n}$$

Where f_j is the frequency of the possible value $x_{(j)}$ of the variable and $J \to \infty$.

In the case of continuous data, the mean is calculated as follows

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In both cases, n is the size of the sample.

Example 4

Calculate the sample means using the samples from Examples 2 and 3.

Solution

Order Statistics: some statistics are derived from quantitative data which are placed in *rank order* (i.e. from the smallest to the largest). Median, Lower Quartile and Upper Quartile are order statistics which are used as measures of location.

- **The Median** divides a sample of data into two equally sized groups, one containing the smaller observations and the other containing the larger observations.
- **The Lower Quartile** divides the observations into a lower 25% and an upper 75%.
- **The Upper Quartile** divides the observations into a lower 75% and an upper 25%.

The three *r*-values are ranks that can be used to find the corresponding median, lower quartile and upper quartile.

$$r_1 = \frac{n+2}{4}, r_2 = \frac{2n+2}{4}, r_3 = \frac{3n+2}{4}$$

Example 5

Calculate the median, lower quartile and upper quartile of the sample from Example 3.

Solution

In the case of larger samples, the quantities can be approximated by using a *cumulative frequency curve* or *ogive*.

Mode: the most frequently occurring *value*, in the case of discrete data, or *class*, in the case of continuous data, in the data set.

Example 6

Calculate the mode of the sample from Example 3.

Solution

If a frequency distribution is *symmetric* and *unimodal*, the mean, median and mode will be very similar. If a frequency distribution is not symmetric (i.e. the distribution is *skewed* as will be discussed in **Section** 3.3), then these three values can be quite different. In the latter case, the mean is not a **resistant statistic** comparing to the median. A resistant statistic is relatively unaffected by unusual observations (i.e. extreme values).

1.3 Measures of Spread

There are three widely used ways to measure the spread of a distribution about a central value (e.g. mean or median).

Range: the difference between the greatest and least values.

Inter-quartile range (IQR): the difference between the upper and lower quartiles.

Sample Variance and Sample standard deviation: variance s^2 is the average squared deviation of an observation from the sample mean. The standard deviation s is the square root of the variance. They can be calculated as below

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- We divide by $n-1$ rather than n as there are only $n-1$ **degrees of freedom**. The degrees of freedom are the number of independent pieces of information that are used to estimate a parameter or calculate a statistic from the data sample.
- Alternatively, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ can be calculated using $\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$ or $\sum_{i=1}^{n} {x_i}^2 - \frac{1}{n}[\sum_{i=1}^{n} x_i]^2$.

Example 7

Prove the above two statements.

Solution

The degrees of freedom will always be equal to or less than the size of the sample, often denoted as n. If the statistic being calculated makes use of another statistic in an intermediate step, then the number of degrees of freedom must be corrected via making a subtraction.

Example 8

Calculate the sample variance and sample standard deviation of the sample from Example 2.

Solution

If a frequency distribution is symmetric and unimodal, the sample variance and sample standard deviation are useful to provide a concise summary of the spread of the observed frequency distribution. If a frequency distribution is not symmetric (i.e. the distribution is skewed as will be discussed in **Section** 3.3), then a more resistant statistic to describe the spread is IQR.

2 Exploratory Data Analysis & Visualisation

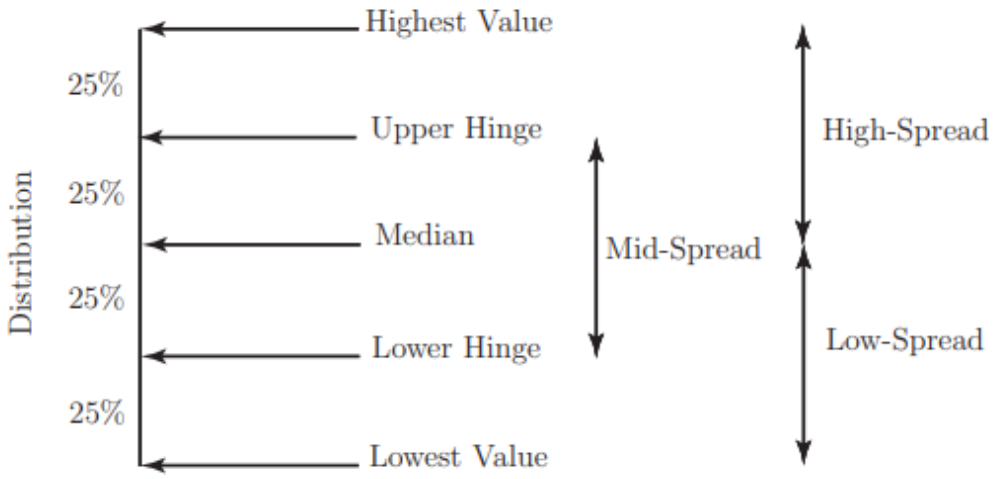
Exploratory Data Analysis (EDA) is the activity by which a data set is explored and organised *in order* that information it contains is made clear.

The basic principles followed in EDA are:

- To measure the location and spread of a distribution we use statistics which are resistant to departures from normality;
- To summarise shape location and spread we use several statistics rather than just two;
- Visual displays as well as numerical displays are used to summarise information obtained about shape, location and spread.

Data visualisation is an important skill in applied statistics and machine learning. Data visualisation can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualisations can be used to express and demonstrate key relationships in plots and charts that are more visceral to yourself and stakeholders than measures of association or significance.

2.1 Five-number Summary Statistics



The lower and upper hinges are the lower and upper sample quartiles. Thus the mid-spread is the inter-quartile range (IQR). The five-number summary, especially when used in conjunction with the three spreads shown in above gives an adequate representation of a non-symmetrical distribution. The median and the hinges are unaffected by changes in extreme values.

Example 9

Find the five number summary and the mid-spread, high-spread and low-spread for the sample from Example 3.

Solution

2.2 Bar Charts

Bar charts can be used to represent the frequencies for categorical or discrete data. In a bar chart, we draw a bar for each value of the variable. The length of the bar is proportional to the frequency for that value. Bars can be drawn vertically or horizontally.

Example 10

Construct a bar chart for the sample from Example 2.

Solution

2.3 Histograms

Histogram is used to represent the frequency distribution of a sample of continuous data. In a histogram, the base of each block or column is the class interval on the x-axis. If the class intervals are of equal width, the height of the columns are proportional to the frequencies.

Example 11

Construct a histogram for the sample from Example 3.

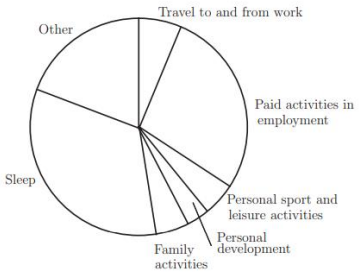
Solution

The approximate shape of the distribution of data is indicated by a **frequency polygon** which is formed by joining the mid-points of the tops of the blocks forming the histogram with straight lines.

2.4 Pei Charts

A pie chart is simply a circular diagram where the circle is divided into sectors and the angles of the sectors are proportional to the quantity represented. Since the total area of the circle is fixed, pie charts are considered to be useful for representing proportions of a total.

Hours spent on:	Males	Females
Travel to and from work	10.5	8.4
Paid activities in employment	47.0	37.0
Personal sport and leisure activities	8.2	3.6
Personal development	5.6	6.4
Family activities	8.4	18.2
Sleep	56.0	56.0
Other	32.3	38.4



Hours spent on:	Males	Proportion of Time	Sector Angle
Travel to and from work	10.5	$\frac{10.5}{168}$	$\frac{10.5}{168} \times 360 = 22.5$

2.5 The Box-and-Whisker Diagram

Boxplots are useful to summarise the distribution of a data sample as an alternative to the histogram. The diagram is constructed as follows.

The Box

a) The left-hand vertical is placed at the lower hinge;

b) The right-hand vertical is placed at the upper hinge;

c) The vertical in the box is placed at the median.

The Whiskers

a) Find the greatest value which is within 1.5 mid-spread of the upper hinge;

b) Find the least value which is within 1.5 mid-spread of the lower hinge;

c) Connect the greatest and least values to the box by means of dashed lines.

The Outlying Values

Mark as large dots any values which are more than 1.5 mid-spreads from the hinges.

Example 12

Construct a box-and-whisker diagram representing the sample from Example 3. Does the box-and-whisker diagram tell you that the data set that you are working with is symmetrical?

Solution

3 Irregularities

Irregularities refers to something in a sample distribution that departures from normality, e.g. extreme values lie well outside the range of most of a sample, asymmetric distribution, and multiple peaks.

3.1 Outliers

There is no standard precise definition but some simple criteria do exist which may be used to detect outliers and accept or reject outliers. Outliers can be extremely important for the following reasons:

- They can have misleading effect on statistics such as the mean and standard deviation;
- Their occurrence may be due to incorrect observation, measurement or recording. In this case it is often possible to correct the data;
- Their presence can induce a false skewness (see **Section** 3.2) in a data set;
- They may actually be members of a population not under consideration. For example, data on road traffic speeds collected at a point in a highway may be intended to provide information on the speeds of motor vehicles but the data are likely to include some observations for bicycles and other slower-moving types of traffic.

Two criteria for the detection of outliers are given below.

Criterion 1

For variables where the distribution has a "**normal**" shape, we can expect only about 1 in 1000 observations to lie more than 3.3 standard deviations away from the mean. So we could treat any value further than 3.3 standard deviations from the mean as an outlier.

An observation x would be regarded as an outlier if

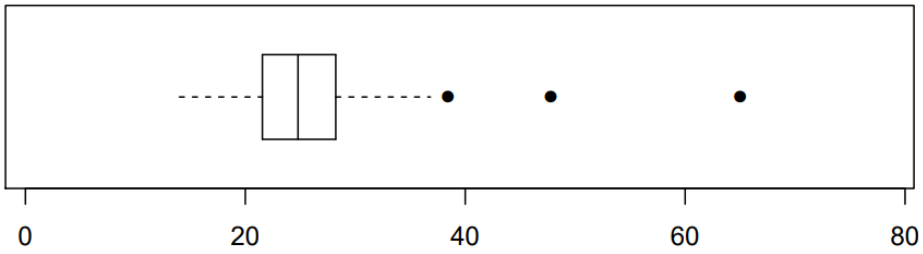
$$\left|\frac{x-\bar{x}}{s}\right| > 3.3$$

Criterion 1 essentially implies that a value has less than 1 in a 1000 of chance of occurring naturally outside the range defined by 3.3 standard deviations from the mean. Note that the property that 0.1% of observations are more than 3.3

standard deviations from the mean really refers to the population mean and standard deviation. Here, we have used $\bar{x}$ and s the sample mean and standard deviation. However, this will be a reasonable approximation in reasonably large samples, say $n > 30$ (see **W2.1**).

Criterion 2

Outliers observations can be defined to be more than 1.5 mid-spreads (or IQRs) from the hinges (or quartiles). Extreme outliers can be defined to be more than 3 mid-spreads from the hinges.



While all values classified as outliers should be investigated, this is particularly true of those classified as extreme outliers.

Example 13

Manufacturing processes generally result in a certain amount of wasted material. For reasons of cost, companies need to keep such wastage to a minimum. The following data were gathered over a five-week period by a manufacturing company whose production lines run seven days per week. The numbers given represent the percentage wastage of the amount of material used in the manufacturing process.

Daily Losses (%)

17	6	8	17	23	18	10	15	17	4
17	18	15	19	11	15	22	12	15	16
11	18	17	17	13	15	9	21	17	16
14	13	15	11	12					

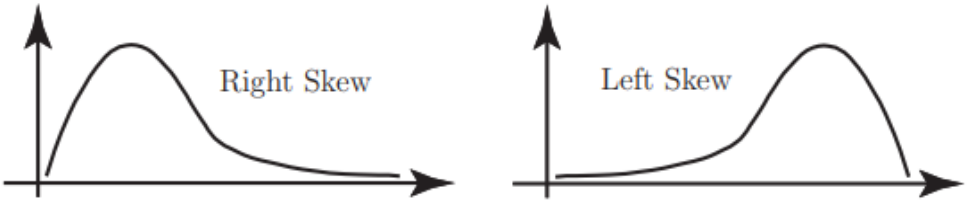
1. Find the mean and standard deviation of the percentage losses of material over the two -week period.
2. Assuming that the losses are roughly normally distributed, apply an appropriate criterion to decide whether any of the losses are smaller or larger than might be expected by chance.

Solution

3.2 Skewness

The regions on either side of the distribution where the frequencies die out are called the *left* and *right* **tails** of the distribution. In a symmetric (or Normal) distribution, the left and right tails are like mirror images of each other. Symmetric unimodal distributions are common but there are exceptions. The distributions of some variables tend to be asymmetric (i.e. skewed), with one tail longer than the other.

If the longer tail is on the right/left, this is called *right* (*positive)/left (negative) skew*.



A skewed distribution cannot be represented purely by two numbers, say the mean and standard deviation. Five-number summary statistics can be used in this case to describe such a distribution.

3.3 Multimodal Distribution

A distribution with more than one peak is called a multimodal distribution. A distribution with exactly two peaks is called a bimodal distribution. Such distributions can be very difficult to summarise. In this case, the stem-and-leaf plot (not covered in this module and please refer to [1] in Further Readings) is more informative than the box-and-whisker plot.

4 Further Readings

[1] HELM Workbook 36 Descriptive Statistics

https://nucinkis-lab.cc.ic.ac.uk/HELM/HELM_Workbooks_36-40/WB36-all.pdf