2. Describing Data

We consider

1. Graphical methods
2. Numerical methods

General Use of Graphical and Numerical Methods

Graphical methods can be used to visually and qualitatively present data and compare variables and samples.

They **can be used to illustrate conclusions**, but should **not be used to draw conclusions**, since they lack quantitative detail.

Numerical methods can be used to quantitatively present data and compare variables and samples (giving a mental picture).

2.1 Graphical Methods

We consider the following graphical methods.
a) To illustrate the distribution of a variable

1. **For quantitative variables**: Histograms, Box Plots.
2. **For qualitative variables**: Pie Charts and Bar-Charts.

b) To illustrate the relationship between two variables, when we have pairs of observations (e.g. the height and weight of a group of people):**Scatter Plots**.

Box plots will be considered at the end of this chapter. We only consider computer generated boxplots.

Scatter Plots will be considered in the chapters on regression and prediction.

2.1.1 Histograms

Histograms are used to estimate the distribution of a variable based on a sample.

In order to do this data must first be split into intervals (classes).

There is **no one correct way** of drawing a histogram, but the following rules of thumb should be observed, in order to choose **the number of intervals and their length**.

Rules of Thumb for Drawing Histograms

a. If we have n observations, then the number of intervals (classes) used, k, should be approximately $\sqrt{n}$. Hence, if we have 20 observations, it is best to use 4 or 5 classes (i.e $k = 4$ or $k = 5$).

b. The endpoints of the intervals used to define the classes should be nice numbers (i.e. consecutive integers, multiples of 2, 5, 10, 20, 50, 100 etc. depending on the values of the observations).

Rules of Thumb for Drawing Histograms-ctd.

c. Let *a* be the value of the largest "nice" number smaller than any of the observations and *b* the value of the smallest "nice" number greater than any of the observations. Then the width of the intervals w is

$$w = \frac{b - a}{k},$$

where k is the number of classes.

d. The values of a, b, k and w can be fine-tuned.

Rules of Thumb for Drawing Histograms-ctd.

Once the classes are chosen, we count the number of observations in each class.

The relative frequency of a class is the percentage of observations in that class.

Let n_i be the number of observations in class i. The relative frequency of class i is given by f_i , where

$$f_i = \frac{100n_i}{n}.$$

The height of the box representing a class is equal to the class's relative frequency.

Example 2.1

We observe the height of 20 individuals (in cm). The data are given below

172, 165, 188, 162, 178, 183, 171, 158, 174, 184,
167, 175, 192, 170, 179, 187, 163, 156, 178, 182.

Construction of a Histogram

First we choose the number of classes and the corresponding intervals.

There are 20 items of data and $\sqrt{20} \approx 4.5$, thus we should choose 4 or 5 intervals.

All the individuals are between 150cm and 200cm tall (the tallest 192cm and the shortest 156cm).

Hence, if we used five intervals the length of the intervals would be

$$\frac{200 - 150}{5} = 10.$$

Construction of a Histogram-ctd.

Hence, it seems reasonable to use 5 intervals of length 10, starting at 150.

It should be noted that it is necessary to decide whether values at the boundaries of two intervals belong to the lower or upper interval. This choice is not crucial, but should be consistent.

Here, we have intervals $[150, 160], (160, 170], \ldots, (190, 200]$. If we count 160 (the upper end-point of the first interval) as belonging to the interval $[150, 160]$, then 170 (the upper end-point of the second) should be counted as belonging to the interval $(160, 170]$.

Note: Square brackets indicate that the endpoint belongs to the interval, a round bracket indicates that the endpoint does not belong to the interval.

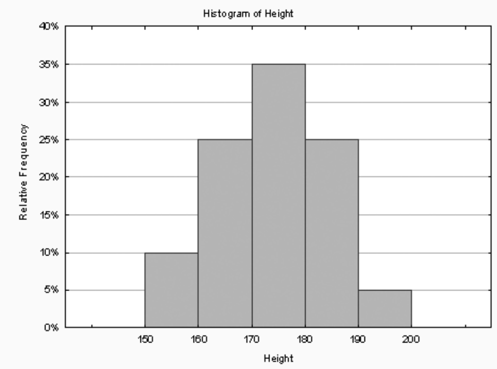
Construction of a Histogram-ctd.

Now we count how many observations fall in each group.

Height (x)	No. of Obs.	Relative Freq. %
$150 \leq x \leq 160$	2	$100 \times 2/20 = 10$
$160 < x \leq 170$	5	$100 \times 5/20 = 25$
$170 < x \leq 180$	7	$100 \times 7/20 = 35$
$180 < x \leq 190$	5	$100 \times 5/20 = 25$
$190 < x \leq 200$	1	$100 \times 1/20 = 5$

Construction of a Histogram-ctd.

Now we draw the histogram



2.1.2 Bar Charts

These are analogous to histograms, but are used for qualitative data.

Bar charts are simpler than histograms, since the data is already classified.

We only have to calculate the relative frequencies of the classes.

If the scale is nominal, then the bars should be drawn separately.

2.1.3 Pie Charts

Pie charts show the same information as bar charts in a different way.

The section of the pie corresponding to a class occupies the same proportion of the circular pie as its relative frequency.

The angle made by the section corresponding to the ith class is α_i. Since 360^o corresponds to 100% of the data, 3.6^o corresponds to 1% of the data and

$$\alpha_i = 3.6 f_i,$$

where f_i is the relative frequency of class i in percentage terms.

Example 2.2

Three schools made up the old college of Informatics and Electronics: Computer Science, Electronic Engineering and Mathematics.

The number of students in the first year is as follows:

Computer Science 90
Electronic Engineering 150
Mathematics 60

Illustrate this data in the form of i) a bar chart, ii) a pie chart.

Construction of a Bar Chart and a Pie Chart

These two diagrams are based on the relative frequencies of the number of students. The relative frequency of class i is given by

$$f_i = \frac{100n_i}{n},$$

where n is the total number of observations. Here,

$$n = 90 + 150 + 60 = 300.$$

Thus, f_1 , the relative frequency of computer science students, is given by

$$f_1 = \frac{100n_1}{n} = \frac{100 \times 90}{300} = 30$$

Construction of a Bar Chart and a Pie Chart-ctd.

f_2 , the relative frequency of electronic engineering students, is given by

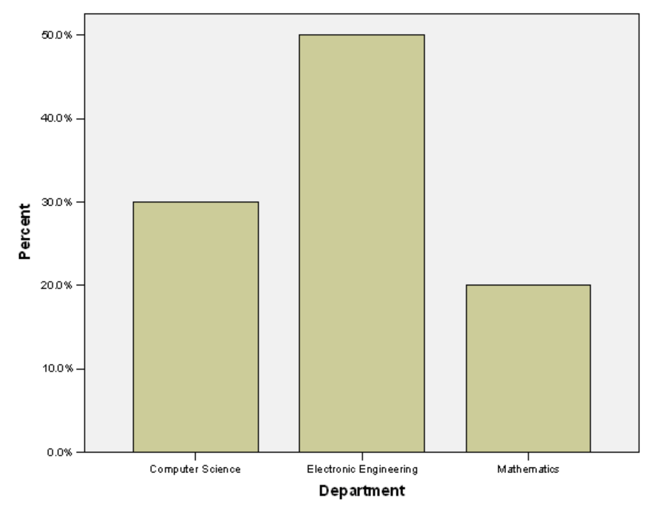
$$f_2 = \frac{100 n_2}{n} = \frac{100 \times 150}{300} = 50.$$

f_3 , the relative frequency of mathematics students, is given by

$$f_3 = \frac{100 n_3}{n} = \frac{100 \times 60}{300} = 20.$$

Construction of a Bar Chart and a Pie Chart-ctd.

We then draw the bar chart



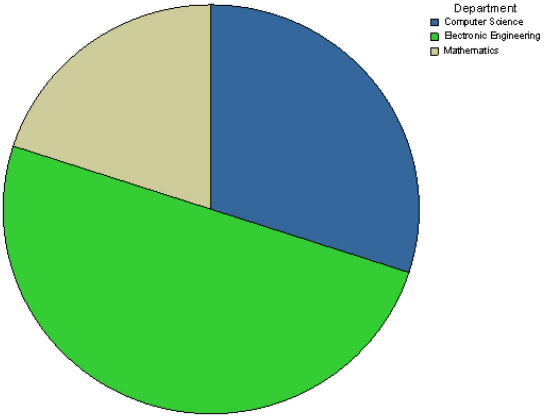
Construction of a Bar Chart and a Pie Chart-ctd.

The angles made by the pieces of pie representing these classes in the pie chart are

$$\begin{aligned}\alpha_1 &= 3.6 \times f_1 = 108\\ \alpha_2 &= 3.6 \times f_2 = 180\\ \alpha_3 &= 3.6 \times f_3 = 72\end{aligned}$$

Construction of a Bar Chart and a Pie Chart-ctd.

We now draw the pie chart



2.2 Numerical Methods

The most common measures (statistics) used to describe numerical variables can be classified into two groups.

1. Measures of **centrality** i.e. measure of the centre of the distribution.
2. Measures of **variability** i.e. the spread, dispersion of the data.

2.2.1 Notation and Measures of Centrality

Let x denote a particular numerical variable e.g. height.

The n observations of this variable (in the same sequence as they are observed) are denoted

$$x_1, x_2, \ldots, x_n$$

The n ordered observations are denoted

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. This is simply a list of the observations from the smallest to the largest.

If a value appears k times in the sample, it must appear k times in this ordered list. In this case the index is the rank of the observation.

The Sample and Population Means

The sum of the data is denoted by

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$$

The **sample mean** (which we observe) is given by $\overline{x}$, where

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

The **population mean** (which we do not observe) is given by μ, where

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

where N is the population size.

The Sample Median

1. When n is **odd**, the sample median, denoted *Med*, is the observation that appears in the middle of the list of ordered observations.
2. When n is **even**, the sample median is the average of the two observations in the middle of this list.

That is to say

1. For n odd, the median has rank $\frac{(n+1)}{2}$.
2. For n even, the median is the average of the two observations with ranks $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Note: At least half the data are less than or equal to the median and at least half the data are greater or equal to the median.

The Sample Mode

The mode is useful when we are dealing with discrete or categorised data (**not continuous** data).

The mode is the observation (or category, as appropriate) that occurs most frequently in a sample.

Example 2.3

The number of children in 14 families is given below:

$$4,\ 2,\ 0,\ 0,\ 4,\ 3,\ 0,\ 1,\ 3,\ 2,\ 1,\ 2,\ 0,\ 6$$

Calculate

a. the mean

b. the median

c. the mode

Calculation of the Mean, Median and Mode

$$a)\overline{x} = \frac{4+2+0+0+4+3+0+1+3+2+1+2+0+6}{14} = 2$$

b) Since $n = 14$, the median will be the average of the observations with rank 7 and 8. Ordering these observations, we obtain

$$0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 4,\ 6.$$

The 7th and 8th observations in this list are both 2. Thus, the Median is 2.

0 appears most frequently (4 times). Thus, the mode is 0.

2.2.2 Advantages and Disadvantages of the Mean and the Median

1. The mean is much more sensitive to extreme values (or errors in data) than the median.

e.g. **Data Set 1**: 3.7, 4.6, 3.9, 3.7. 4.1

median: 3.9 mean: 4

Data Set 2: 37, 4.6, 3.9, 3.7, 4.1

median 4.1 mean 10.66

Use of Median for Asymmetric distributions

If the distribution (histogram) is **symmetric** or close to being symmetric, then

$$\overline{x} \approx Med.$$

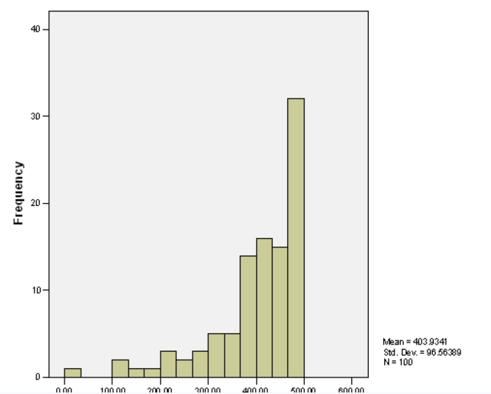
In this case, there will tend to be few outliers. Either measure may be used (the mean is normally used as it has nicer statistical properties).

If the distribution is **clearly asymmetric** these measures will be significantly different.

In this case the median should be used as a measure of centrality, since there will be outliers which have a large influence on the mean.

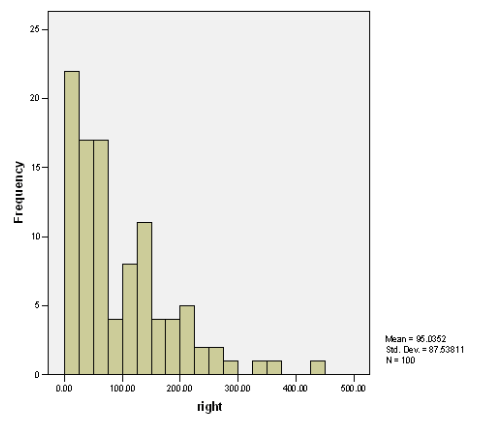
Conclusion: The median is always a good measure. The mean is a good measure if the distribution is symmetric

A left skewed distribution



There is a "tail" to the left of the "centre" of the distribution.

A right skewed distribution



There is a "tail" to the right of the "centre" of the distribution.

Skewness and the Relation between Mean and Median

If the distribution is **left skewed** (asymmetric), then $\overline{x} < Med$

If the distribution is **right skewed** (asymmetric), then $\overline{x} > Med$.

Right skewed distributions are met quite commonly e.g. income, weight (somewhat less skewed).

When reading media articles, we should be aware of the lack of precision in the terminology used when quoting statistics.

For example, the "average man" (intuitively understood as the median earner) earns less than the mean salary. Due the skewness of the distribution of salaries, the median wage will be lower than the "average" (i.e. mean) wage.

Around 25% of individuals will earn above the mean wage.

2.2.2 Measures of Variability (Dispersion)

These measures are used to measure how scattered the data are.

1. **Range.**

$$Range = x_{(n)} - x_{(1)},$$

i.e. The range is the largest observation minus the smallest observation.

This is simple to calculate, but is very sensitive to extreme values (mistakes in the data).

Since it is based only on two items of data, it may give very little information about the data as a whole.

The Sample Variance

2. **Sample Variance**: The sample variance measures how closely the data are placed around the mean.

If all the data are equal, then the variance will be zero. The more dispersed around the mean the data are, the greater the variance is. The sample variance is given by s^2, where

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

It is a measure of the average square distance of an observation from the mean.

Standard Deviation

3. **Standard deviation.** The standard deviation s is the square root of the variance.

It is preferred to the variance as a desciptive measure, since it is a measure of the average distance of an observation from the mean.

It is thus measured in the same units as the observations themselves.

Note: On scientific calculators the standard deviation is denoted as s_{n-1} or σ_{n-1}.

Coefficient of Variation

4. **Coefficient of variation,** *cv*: The coefficient of variation is often used to compare the relative dispersion of two or more **positive** variables, e.g. height and weight, height of males and height of females, especially if the sample means are significantly different.

It is calculated by dividing the standard deviation by the sample mean. The sample with the largest cv has the largest relative spread.

One major advantage of the coefficient of variation is that is insensitive to the units used and so may be used to compare the spread of e.g. height and weight.

$$cv = \frac{s}{\overline{x}}$$

Example 2.4

Calculate the sample variance, standard deviation, range and coefficient of variance of the data given in Example 2.3.

Calculation of Sample Variance

We calculated $\overline{x} = 2$. Hence, the sample variance is given by

$$\begin{aligned} s^2 =& \frac{\sum_{i=1}^{n}(x - \overline{x})^2}{n - 1} \\ =& \frac{(4-2)^2 + (2-2)^2 + (0-2)^2 + \ldots + (6-2)^2}{13} \\ =& \frac{44}{13} \approx 3.3846 \end{aligned}$$

Calculation of the Standard Deviation

The standard deviation is given by

$$s = \sqrt{s^2} = \sqrt{3.3846} \approx 1.8397.$$

Calculation of the Coefficient of Variation and the Range

The coefficient of variance is given by

$$cv = \frac{s}{\overline{x}} = \frac{1.8397}{2} \approx 0.9199.$$

The range is given by $Range = 6 - 0 = 6$ (the largest observation is 6 the smallest is 0).

2.2.3 Grouped Data

Data from a discrete distribution may well be displayed in tabular form, e.g. the data from Example 2.3 may be presented as:

Number of children	0	1	2	3	4	5	6
Number of families	4	2	3	2	2	0	1

Grouped Data

The first row gives the value of the observation x (here no. of children).

The second row gives the frequency of each observation n_x (the number of observations of x children). The sum of these numbers is the total number of observations.

Since 3 families have 2 children, these families account for $3 \times 2 = 6$ children (in mathematical notation $n_2 \times 2$, where n_2 is the number of times 2 children were observed).

The Mean of Grouped Data

It can be seen that summing xn_x over all the possible observations of the variable (number of children), we obtain the sum of the observations.

It follows that the sample mean is given by

$$\overline{x} = \frac{\sum[xn_x]}{n},$$

where n is the total number of observations, i.e. $n = \sum n_x$.

The Sample Variance of Grouped Data

The sample variance is given by

$$s^2 = \frac{\sum [n_x (x - \overline{x})^2]}{n - 1}.$$

Calculation of the Mean for Grouped Data

These formulae can be used to calculate the mean and the sample variance in the following way.

If we are not given the total number of observations, n, we first calculate it

$$n = \sum n_x = 4 + 2 + 3 + 2 + 2 + 1 = 14.$$

The sample mean is given by

$$\overline{x} = \frac{\sum [x n_x]}{n} = \frac{0 \times 4 + 1 \times 2 + 2 \times 3 + 3 \times 2 + 4 \times 2 + 6 \times 1}{14} = 2$$

Calculation of the Sample Variance for Grouped Data

The sample variance is given by

$$
\begin{aligned}
s^2 &= \frac{\sum[n_x(x-\overline{x})^2]}{n-1} \\
&= \frac{4 \times (0-2)^2 + 2 \times (1-2)^2 + \ldots + 1 \times (6-2)^2}{13} \\
&= \frac{44}{13}
\end{aligned}
$$

Calculation of the Median for Grouped Data

The median is calculated in the same way as for data given individually, by ranking (ordering) the data.

Since there are 14 observations, the median is the average of the 7th and 8th ranked observations.

Starting from the smallest observations.

1. There are 4 observations of 0 (ranks 1 to 4).
2. There are 2 observations of 1 (ranks 5 and 6).
3. There are 3 observations of 2 (ranks 7, 8, 9).

It follows that the median is 2.

2.3 Percentiles and Box Plots

The r percentile (or quartile) P_r is defined such that $r\%$ of the data are less than P_r.

The percentiles which split the ordered data into 4 subsamples of equal size are called quartiles.

Q_1 is the 25% percentile (lower quartile) - 25% of the data are smaller than Q_1.

Q_2 or Med is the 50% percentile (median) - 50% of the data are smaller than Q_2.

Q_3 is the 75% percentile (upper quartile) - 75% of the data are smaller than Q_3.

The Interquartile Range

The interquartile range IR is a measure of the spread of the data.

It is the difference between the upper and lower quartile.

$$IR = Q_3 - Q_1$$

The Normal Distribution

The normal distribution is a bell shaped distribution which is often met in nature.

If a numerical trait results from a "sum" of a large number of random factors, none of which is dominating, then it will have close to a normal distribution.

For example, height depends on many random factors (genetic, dietary and environmental) and has a bell shaped distribution.

Box Plot

A box plot is a visual representation of the distribution of a numerical variable (as is a histogram).

It is based on the quartiles of a distribution.

It is normally used to compare 2 or more distributions.

It can also be used to see whether a distribution is symmetric or skewed (although a histogram may well show this more clearly).

The box plot of a symmetric distribution will be symmetric around the median.

A long upper whisker indicates a right skewed distribution. A long lower whisker indicates a left skewed distribution.

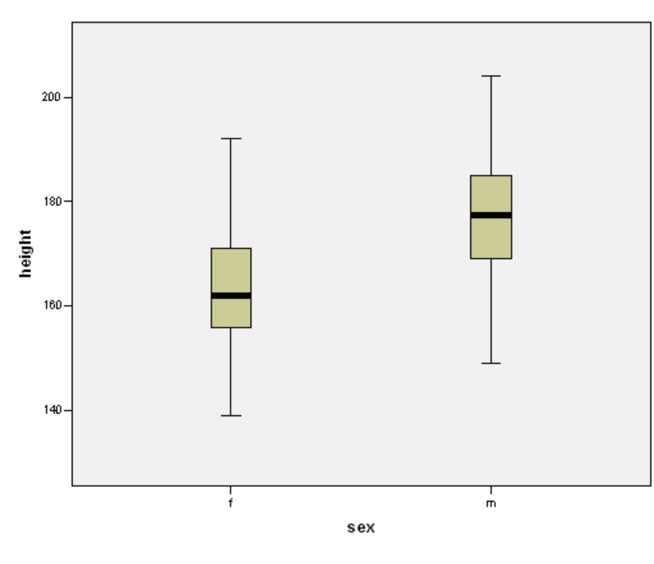
Box Plot

The top and bottom of the box represent the upper and lower quartiles respectively. The central line represents the median.

In SPSS the whiskers represent the interval in which we would expect to see 95% of the data if they came from a normal distribution.

Outliers are represented by points outside the whiskers.

Box Plot



Description and Comparison of Box Plots

The distribution of the height of males is reasonably symmetric (the median lies at the centre of the box and the upper and lower whiskers are of similar lengths).

The distribution of the height of females is slightly right skewed (the upper "half" of the box and the upper whisker are longer than the lower "half" and whisker, respectively).

Males are on average taller than females (the right hand side boxplot is shifted upwards in relation to the left hand boxplot).

The dispersions of male and female height are comparable (the boxes and whiskers are of similar lengths).

2.4 Describing relative frequencies for qualitative data

One should give the total number of observations, together with the number of missing observations.

The number of observations in each category should be given, together with the relative frequency as a percentage of the valid observations.

We comparing groups (samples), the relative frequencies should be compared. However, such comparisons are valid only if there are a reasonable number of observations in both groups (samples).

Describing Relative Frequencies

For example, suppose 200 people were asked if they preferred Sting or Radiohead. There were 200 people in the sample, of which 18 (9%) did not reveal a preference.

Of the 182 who revealed a preference 100 preferred Radiohead (54.95%) and 82 Sting (45.05%).

Of these 182, 95 were male (52.20%) and 87 female (47.8%).

Comparing the prefences of the sexes, 65 males preferred Radiohead (68.42% of males) compared with 35 females (40.23% of females). Hence, comparing the percentages, it seems that males are more likely to prefer Radiohead.