2.1 Measures of Location (P.9-11)

2.1.1 Summation Notation

Suppose that we observe n values from an experiment. This collection (or set) of n values is called a *sample*. Let x_1 be the first sample point or observation; x_2 be the second sample point or observation etc and x_n be the n^{th} sample point or observation. The sum of these n values is abbreviated

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

Example: For the values

$$x_1 = 3,\ x_2 = 4,\ x_3 = 5,\ x_4 = 3$$

evaluate the following summation expressions.

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = \underline{\hspace{4cm}}$$

$$\sum_{i=1}^{4} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = \underline{\hspace{5cm}}$$

$$\sum_{i=2}^{3} x_i = x_2 + x_3 = \underline{\hspace{2.5cm}}$$

$$\begin{aligned}\sum_{i=1}^{4} (2x_i + 3) &= (2x_1 + 3) + (2x_2 + 3) + (2x_3 + 3) + (2x_4 + 3) \\ &= \underline{\hspace{8cm}} \\ &= \underline{\hspace{4.5cm}}.\end{aligned}$$

2.1.2 Sample Mean

The sample mean is the simple arithmetic average of the observations. For observations $x_1, x_2, \ldots, x_n$, this is denoted by $\bar{x}$ and is given by

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

Example:

The mean of the sample of 4 values from the previous example is

$$\bar{x} = \frac{\qquad\qquad\qquad}{\qquad\qquad\qquad}.$$

The mean is very sensitive to large or small outliers in the sample. In such cases it is better to use the median as a measure of the "centre" of the data.

2.1.3 The Median

An alternative measure of location is the *median* or, more precisely, the *sample median*. The median, $\tilde{x}$ is a value such that *at least half* the observations are less than or equal to $\tilde{x}$ *and at least half* the observations are greater than or equal to $\tilde{x}$.

To find the median, we arrange the data in ascending order. If the number of data is ODD the median is the middle data point.
Example:

$$3 \quad 5 \quad \underset{\substack{\nearrow \\ \text{The median}}}{7} \quad 7 \quad 8$$

If the number of data is EVEN, then we average the 2 values around the middle:

Example:

$$3 \quad 5 \underset{\substack{\nearrow \\ \text{Middle space}}}{\quad} 7 \quad 7$$

$$\tilde{x} = \frac{5+7}{2} = 6$$

2.1.4 The Mode

The *mode* is the most frequently occurring value among all the observations in a data set (sample).

Example:

A random sample of 12 measurements of interorbital width of domestic pigeons is obtained as follow:

12.2, 12.9, 11.8, 11.9, 10.8, 11.1, 13.3, 11.8, 11.0, 12.2, 10.7, 11.6

Find the mean, median, mode.

Solution: Order the data x_i:

10.7, 10.8, 11.0, 11.1, 11.6, 11.8, ¦ 11.8, 11.9, 12.2, 12.2, 12.9, 13.3

$$\begin{aligned}
\text{mean} &= \frac{\sum\limits_{i=1}^{n} x_i}{n} \\
\text{median} &= \frac{x_6 + x_7}{2} = \underline{\hspace{8cm}} \\
\text{mode} &= \underline{\hspace{3cm}}
\end{aligned}$$

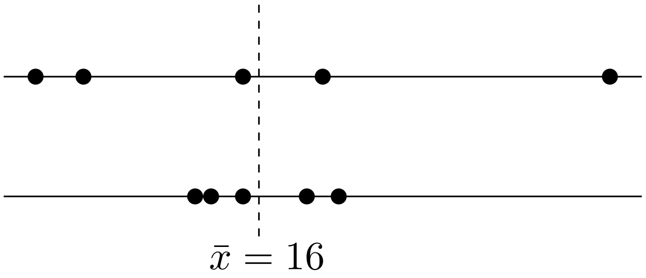
Mode is not unique in this example. Such datasets are also called *bimodal*.

2.2 Measures of Spread (P.11-15)

Consider the following two sets of observations:

2, 5, 15, 20, 38

12, 13, 15, 19, 21



Exercise: Verify that both sets have the same centre: $\bar{x} = 16$.

However, the two samples visually appear radically different. This difference lies in the greater *variability*, or *spread*, or *dispersion* in the first dataset. Therefore, we need a measure of dispersion to find an indication of the amount of variation that a data set exhibits.
We will now describe the most popular measures of spread used in practice.

2.2.1 The Range

The *range* is the difference between the largest and smallest observations in a sample.

Example:

In the set $\{2, 5, 15, 20, 38\}$, the range = 38 - 2 = 36.

In the set $\{12, 13, 15, 19, 21\}$, the range = 21 - 12 = 9.

2.2.2 Quartiles

The range may be inflated by extreme large or small values called *outliers*. Instead one may look at the spread of the middle 50% of observations which is free from the effects of outliers.

The *lower quartile*, Q_1, is a value such that *at least* 25% of the observations are less than or equal to Q_1 *and at least* 75% of the observations are greater than or equal to Q_1.

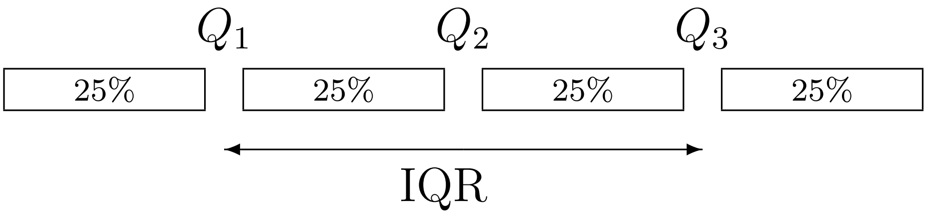
That is, 25% of observations $\leq$ lower quartile, Q_1

Similarly the *upper quartile*, Q_3, separates off the upper 25% of the observations in an ordered array.

Clearly, the median, Q_2, is such that 50% of observations $\leq Q_2$.

2.2.3 Interquartile Range

$IQR = Q_3 - Q_1$. The middle 50% of the observations lie in $[Q_1, Q_3]$.



2.2.4 Five-number Summary

The collection of the following 5 values is called the 5-number summary:

$$(\text{Minimum},\ Q_1,\ \tilde{x} = Q_2,\ Q_3,\ \text{Maximum}).$$

Example: Consider the data set x_i:

$$98, 100, 100, 103, 105, 107, 110, 113, 115.$$

Find the five-number summary. What are the IQR and range of this data set?

Upper half

Solution: $n = 9$. 98, 100, 100, 103, 105, 107, 110, 113, 115.

Lower half

- The middle value of the ordered sample is the median: $Q_2 =$ ______________________________
- Since the sample size is *odd*, we include the median in both the half of the sample. Therefore, the middle value of the lower half is $Q_1 =$ ______________________________
- Similarly, the median is included in the upper half of the sample. Therefore, the middle value of the upper half is $Q_3 =$ ______________________________
- Min= __
- Max= ___
- IQR $= Q_3 - Q_1 =$ __________
- Range = Max - Min = _________

Quantiles or Percentiles

The p^{th} percentile of a data set is the point such that p percent of the sample points are less than or equal to that point.

2.2.5 Boxplots (P.15-16)

Boxplots show the shape of the distribution of data very clearly and are also helpful in identifying any outlying (or extreme) values. In this diagram all outlying observations are identified by using the following threshold values:

Upper threshold value (UT) = upper quartile + 1.5 × IQR

Lower threshold value (LT) = lower quartile - 1.5 × IQR.

Outliers

Observations lying outside the interval (LT,UT) are called *outliers*.

Example

Consider the following data set of 13 observations x_i:

$$4 \quad 6 \quad 6 \quad 7 \quad 7 \quad 9 \quad 10 \quad 11 \quad 13 \quad 15 \quad 22 \quad 24 \quad 30$$

Find the median, lower quartile, upper quartile and IQR. Find LT and UT and identify any outliers.

Solution: $4 \ \ 6 \ \ 6 \ \ 7 \ \ 7 \ \ 9 \ \ 10 \ \ 11 \ \ 13 \ \ 15 \ \ 22 \ \ 24 \ \ 30$

Median, $Q_2 = \tilde{x} = x_7 =$ ________________________________

Lower quartile, $Q_1 = x_4 =$ ________________________________

Upper quartile, $Q_3 = x_{10} =$ ______________________________

IQR = $Q_3 - Q_1$ = ________

LT = Q_1 - 1.5 × IQR = _____________

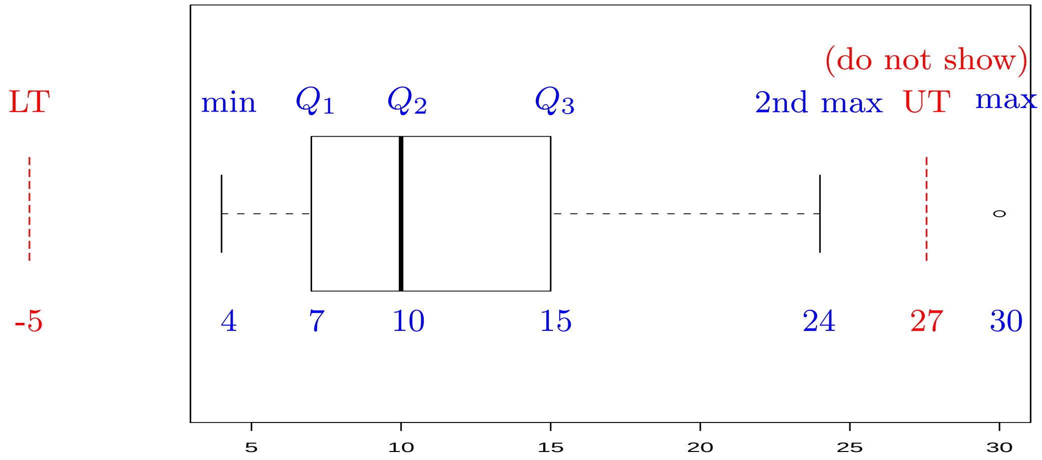
UT = Q_3 + 1.5 × IQR = _______________

All observations in the interval (-5,27) are considered "legitimate". Clearly, there is only one data point outside this interval. Therefore, the last observation 30 is considered as abnormally high. This is an outlier.

A *boxplot* summarises the above information as a graph.

Steps to draw a boxplot:

1. Draw a rectangle (horizontal or vertical) of arbitrary width from Q_1 to Q_3.
2. Draw a line across the rectangle at Q_2.
3. Draw two lines (called, Whiskers) to and from the observations in (LT,UT) from the above rectangle.
4. Mark any identified outliers by ∘



2.2.6 Sample Variance and Standard Deviation

Another measure of spread is the *sample variance*, denoted s^2. This is defined as follows:

For data $x_1, x_2, \ldots, x_n$, let

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

In a sense, the above s^2 can be considered as the average squared deviation of all observations from their mean.

The *sample standard deviation* is then defined as

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

Note: It is easier to use the following calculation formula in practice (It can be shown after expanding the square bracket that it is equivalent to the above definition of s^2):

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] \overset{\text{or}}{=} \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right].$$

This, and many other important formulas, is provided on a formula sheet available from the course web site.

Example: Find the sample standard deviation (sd) of

$$55, 48, 59, 64, 65, 57, 58, 41, 57, 59, 64, 62$$

Solution: $n =$ __. First calculate

$$\sum_{i=1}^{12} x_i = \rule{6cm}{0.4pt}$$

$$\sum_{i=1}^{12} x_i^2 = \rule{7cm}{0.4pt}$$

Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{12} x_i = \rule{3cm}{0.4pt}$

Variance:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \rule{6cm}{0.4pt}$$

Standard Deviation:

$$s = \rule{4cm}{0.4pt}$$

Note: Many scientific calculators and computer packages (including `R`) can be used to find the standard deviation of a given dataset.

Note: It can be proved that after a change in origin of a data set, the variance and standard deviation remain the same. If the sample points change in scale by a factor c, then the variance changes by a factor of c^2 and the sd changes by a factor of c.

2.3 The Coefficient of Variation

The *coefficient of variation*, denoted CV, is a *normalized* version of the standard deviation, expressed as a *proportion* of the mean.

For a dataset with $\bar{x} \neq 0$, we define

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

Example: The CV for the previous dataset is

$$\mathrm{CV} = \frac{s}{\bar{x}} = \underline{\hspace{5cm}}$$

or the s.d. accounts for 12% of the mean.

Note: It is clear that the CV is *dimensionless* as it is a proportion. For example, it is not affected by multiplicative changes of scale. Therefore, the CV is a useful measure for comparing the dispersions of two or more variables that are measured on different scales.

Exercise:

Do problems 1 and 3 from Chapter 1 on p. 26, 27 of the textbook.

2.4 Grouped Data (P.16-17)

Recall from last week that large datasets are sometimes summarised with a frequency distribution table, like this:

Group/Class interval	Class center	Frequency	Relative frequency
$y_1 < x \leq y_2$	$u_1 = (y_1 + y_2)/2$	f_1	f_1/n
$y_2 < x \leq y_3$	$u_2 = (y_2 + y_3)/2$	f_2	f_2/n
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_k < x \leq y_{k+1}$	$u_k = (y_k + y_{k+1})/2$	f_k	f_k/n
TOTAL		n	1.000

2.4.1 The mean of Grouped Data

If we only have the information provided by a grouped frequency table, for example, we only have access to the published report and not the original data set, then we can approximate the sample mean by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i u_i,$$

where k are is the number of bins, the *interval centres* are $u_1, u_2, \ldots, u_k$ with corresponding *frequencies* $f_1, f_2, \ldots, f_k$.

Example: Consider the data on weight in pounds (recorded to the nearest pound) of 92 students from week 1.

This is the frequency distribution that we obtained last week:

CLASS INTERVAL	CLASS CENTER	FREQUENCY
94-104	99	2
104-114	109	5
114-124	119	11
124-134	129	10
134-144	139	3
144-154	149	4
TOTAL		35

Find the grouped mean.

Solution:

$$\sum_{i=1}^{6} f_i\, u_i \quad = \quad \underline{\hspace{10cm}}$$

$$\Rightarrow \bar{x} \quad = \quad \frac{1}{n} \sum_{i=1}^{6} f_i\, u_i = \underline{\hspace{4cm}}$$

Exercise: Find the exact mean of the data and compare it to the above approximation.

Answer: Using the complete data, the exact $\bar{x} =$ _____.

2.4.2 Sample Variance of Grouped Data

For data from a frequency table, the grouped sample variance is:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j(u_j - \bar{x})^2$$

or equivalently

$$s^2 = \frac{1}{n-1}\left[\sum_{j=1}^{k} f_j u_j^2 - \frac{1}{n}(\sum_{j=1}^{k} f_j u_j)^2\right] \overset{\text{or}}{=} \frac{1}{n-1}\left[\sum_{j=1}^{k} f_j u_j^2 - n(\bar{x}^2)\right].$$

Example:

Find the sample variance from the previous example frequency distribution.

Solution:

$$\begin{aligned} \sum_{i=1}^{6} f_i\, u_i^2 &= \underline{\hspace{10cm}} \\ \Rightarrow s^2 &= \underline{\hspace{7.5cm}} \\ \Rightarrow s &= \underline{\hspace{4.5cm}} \end{aligned}$$

compared to 13.372 using the complete data set.

Exercise:

Do problem 2 from Chapter 1 on p. 26 of the textbook.

Additional example:

Consider the two samples:

Sample 1: 1.76, 1.45, 1.03, 1.53, 2.34, 1.96, 1.79, 1.21
Sample 2: 0.49, 0.85, 1.00, 1.54, 1.01, 0.75, 2.11, 0.92

For each of the two samples, calculate the mean and the standard deviation and draw a boxplot.

Solution: In ascending order:

Sample 1 x_i: 1.03, 1.21, Q_1 1.45, 1.53, Q_2 1.76, 1.79, Q_3 1.96, 2.34
Sample 2 y_i: 0.49, 0.75, Q_1 0.85, 0.92, Q_2 1.00, 1.01, Q_3 1.54, 2.11

We have $n = 8$ is even and
$$\sum_{i=1}^{8} x_i = 13.07,\ \sum_{i=1}^{8} x_i^2 = 22.5873,\ \sum_{i=1}^{8} y_i = 8.67,\ \sum_{i=1}^{8} y_i^2 = 11.2153$$

Sample 1:

$$\text{The mean } \bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{13.07}{8} = 1.63$$

$$\text{The sd } s_x = \sqrt{\frac{1}{8-1}\left[\sum_{i=1}^{8} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{8} x_i\right)^2\right]}$$

$$= \sqrt{\frac{1}{7}\left[22.5873 - \frac{13.07^2}{8}\right]} = 0.42$$

$$Q_1 = \frac{x_2 + x_3}{2} = \frac{1.21 + 1.45}{2} = 1.330 \text{ (pos.} = \frac{n_m + 1}{2} = \frac{4+1}{2} = 2.5);$$

$Q_2 = \dfrac{x_4 + x_5}{2} = \dfrac{1.53 + 1.76}{2} = 1.645$ (pos. $= \dfrac{n+1}{2} = \dfrac{8+1}{2} = 4.5$);
$Q_3 = \dfrac{x_6 + x_7}{2} = \dfrac{1.01 + 1.54}{2} = 1.875$ (pos. $= n_m + n_{Q1} = 4 + 2.5 = 6.5$);
IQR $= Q_3 - Q_1 = 1.875 - 1.330 = 0.545$;
LT $= Q_1 - 1.5 \times IQR = 1.330 - 1.5(0.545) = 0.5125$ (min=1.03);
UT $= Q_3 + 1.5 \times IQR = 1.875 + 1.5(0.545) = 2.6925$ (max=2.34)
There is no outlier.

Sample 2 :

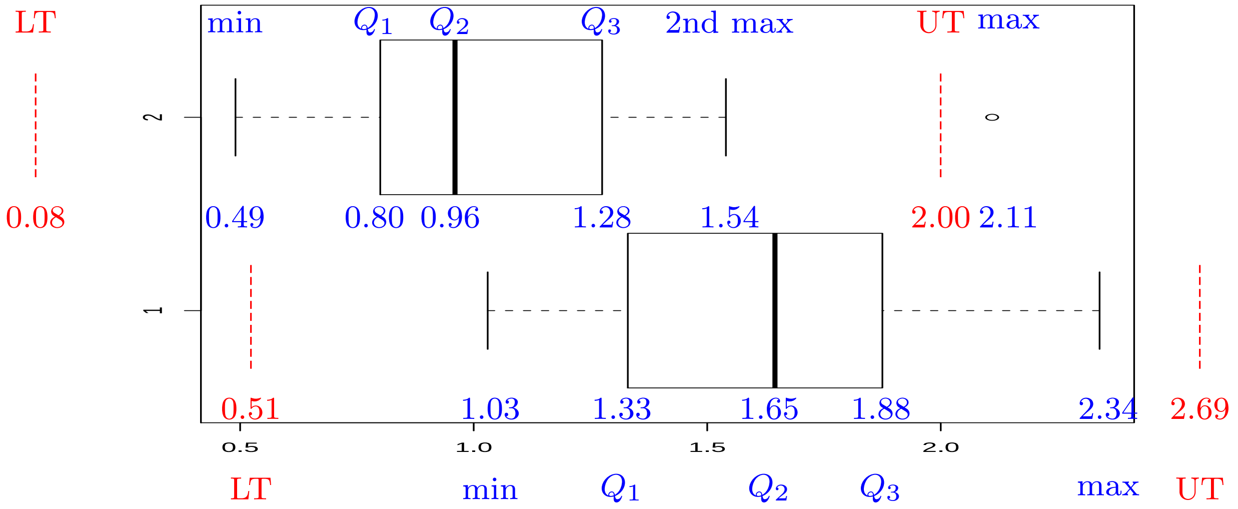
$$\text{The mean } \bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{8.67}{8} = 1.08$$

$$\text{The sd } s_y = \sqrt{\frac{1}{8-1}\left[\sum_{i=1}^{8} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{8} y_i\right)^2\right]}$$

$$= \sqrt{\frac{1}{7}\left[11.2153 - \frac{8.67^2}{8}\right]} = 0.51$$

$Q_1 = \dfrac{y_2 + y_3}{2} = \dfrac{0.75 + 0.85}{2} = 0.80$;
$Q_2 = \dfrac{y_4 + y_5}{2} = \dfrac{0.92 + 1.00}{2} = 0.96$;
$Q_3 = \dfrac{y_6 + y_7}{2} = \dfrac{1.01 + 1.54}{2} = 1.28$;
IQR $= Q_3 - Q_1 = 1.28 - 0.80 = 0.48$;
LT $= Q_1 - 1.5 \times IQR = 0.80 - 1.5(0.48) = 0.08$ (min=0.49);
UT $= Q_3 + 1.5 \times IQR = 1.28 + 1.5(0.48) = 2.00$ (max=2.11)

Since the max $= 2.11$ lies outside (LT,UT) $=$ (0.08,2.00).



Useful R commands:

```
mean(x)
sd(x)
sort(x)
median(x)
sd(x)/mean(x) #cv
fivenum(x)
boxplot(x,y) #2 boxplots side by side
```

where `x` and `y` are vectors of measurements.