0.1 Boxplots

$\hat{3}$. Percentiles

A percentile is defined as a point below which a certain per cent of the observations lie e.g. the 50th percentile is the point below which half the observations lie. The percentiles that divide the data into four quarters are called:

Q1 - 25th percentile or lower quartile

Q2 - 50th percentile or median

Q3 - 75th percentile or upper quartile

0.2 Interquartile Range (IQR)

The Interquartile Range (Q3 Q1) is a measure of variability commonly used for skewed data. The IQR the difference between the point below which 25To find the value of the quartiles, think of Q1 as the middle of the data less than or equal to the median, and of Q3 as the middle of the data greater than or equal to the median. Use the same technique for calculating the mean to find these values.

Example

- Find the three quartiles and the IQR of the following data

$$153471218914256281324 35$$

- First sort the data set into ascending order

$$17912131518242834354256$$

$$17912131518242834354256$$

- Count how many items are in the data set (answer 13 items)
- Which value is the second quartile, which is the median (answer: the 7th item, which is 18)
- Q1 median of data less than or equal to median (7 items)

Boxplots

$$\{1, 7, 9, 12, 13, 15, 18\}$$

Answer: the 4th item, which is 12 Q3 median of data greater than or equal to median (also 7 items)

$$\{18, 24, 28, 34, 35, 42, 56\}$$

Answer: the 4th item of this 7, which is 34

The three quartiles are therefore Q1 = 12, Q2 = 18, Q3 = 34 The interquartile range is therefore Q3 Q1 = 22

0.3 Boxplots

A graphical representation of the quartiles is called a Box plot (Figure 1.3). It displays

(a) lower quartile

(b) median

(c) upper quartile

(d) interquartile range (IQR)

(e) whiskers of length = 1.5 IQR

(f) outlying observations

The Coefficient of Variation
The Coefficient of Variation [page 26]
What happens if you have two sets of data with two different means and two different standard deviations? How do you decide which set is more spread out? Remember the size of the standard deviation is relative to the mean it is associated with.

The coefficient of variation (cv) is often used to compare the relative dispersion between two or more sets of data. It is formed by dividing the standard deviation by the mean and is usually expressed as a percentage i.e. (multiplied by 100). Again we distinguish between the population and sample coefficient of variation.
Population C.V.
Population standard deviation Population mean
Sample C.V.
(i.e. Sample standard deviation divided by Sample mean)
The coefficient of variation for different distribution are compared and the distribution with the largest CV value has the greatest spread.

0.4 Dispersion

The centre of the distribution is only part of the story. We also need to know how spread out - *dispersed* - the data values are.

Consider the following:

A software engineer is offered a job with an annual salary of €45,000. The employer says that this is a very attractive salary as it is above the average for this type of job (€40,000).

Is this a good offer?... We don't know. Not without knowing how much the data *varies* about the central value.

0.4.1 Dispersion

If the distribution of salaries is highly variable, there are many posts available with a better salary. On the other hand, if variability is very low, we have been offered one of the highest salaries in the field.

0.4.2 The Range

0.4.3 The Range

The most basic measure of dispersion is the range of the data.

$$\boxed{\text{range} = \max(x) - \min(x)},$$

i.e., the largest value minus the smallest value in the set of data.

Disadvantage: It only tells us the *overall spread* of the data. But what we really want to know is how the data varies about its centre.

We mainly focus on other techniques (standard deviation and inter-quartile range).

0.4.4 The Variance

The variance is the *average squared distance from the mean* and is given by:

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \\ &= \frac{\sum (x_i - \bar{x})^2}{n-1} \end{aligned}$$

In words, subtract the mean from each value, square the results and then add them all together. Finally, divide by $n-1$.
(For technical reasons, in the case of variance, we divide by $n-1$ rather than n)

The units of variance are *squared-units*, for example, if we were looking at income (in euro) then the variance would be in euros-squared.

0.4.5 The Variance

It turns out that the previous formula can be rewritten as:
(if you're good with sums, i.e., Σ, then you can show this)

$$\boxed{s^2 = \frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}$$

This version of the formula involves less computation so we will use it.

0.4.6 The Standard Deviation

0.4.7 The Standard Deviation

The standard deviation is a *very* important quantity in statistics (as we will see later in this course).

The standard deviation is the *square root* of the variance:

$$\boxed{s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}}.$$

Since the variance is in squared-units, the standard deviation has *the same units as the data* (as a result of taking the square root).

0.4.8 The Standard Deviation: Example

0.4.9 The Standard Deviation: Example

We return to our earlier example - the incomes of 5 individuals.

						$\sum$
x_i	25	29	33	35	40	162
x_i^2	625	841	1089	1225	1600	5380

Using the above and the mean value, $\bar{x} = \frac{162}{5} = 32.4$, we then calculate the variance:

$$\begin{aligned} s^2 = \frac{\sum x_i^2 - n\,\bar{x}^2}{n-1} = \frac{5380 - 5\,(32.4^2)}{5-1} &= \frac{5380 - 5\,(1049.76)}{4} \\ &= \frac{5380 - 524.8}{4} \\ &= \frac{131.2}{4} = 32.8 = 32,800\,€^2. \end{aligned}$$

0.4.10 The Standard Deviation: Example

0.4.11 The Standard Deviation: Example

The standard deviation is then

$$s = \sqrt{s^2} = \sqrt{32.8} = 5.727 = 5,727\,€.$$

Note the units are euros.

0.4.12 Symbols

It is worth preparing ourselves for things to come:

- For the *sample standard deviation* we use the symbol s as shown.
- For the true *population standard deviation* we use σ (the Greek letter "sigma").

Naturally we have s^2 and σ^2 for the sample variance and population variance.

Don't forget, sample *statistics* estimate the true population *parameters*.

0.4.13 Important Note on Dispersion Measures

Variance and standard deviation are *always positive numbers*.

In fact *all* measures of dispersion are positive numbers.

Consider the following four numbers: $\boxed{-10,\ -9,\ -5,\ -4}$.

Clearly the centre is negative; however, standard deviation will not be. Show that $\bar{x} = -7$ and $s = 2.94$.