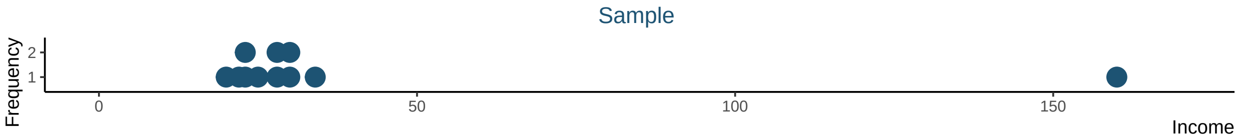
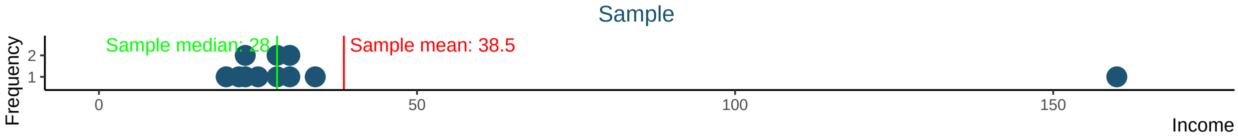
Tutorial 1

Relevant material: Unit 1.

1. (a) We start by plotting the data.



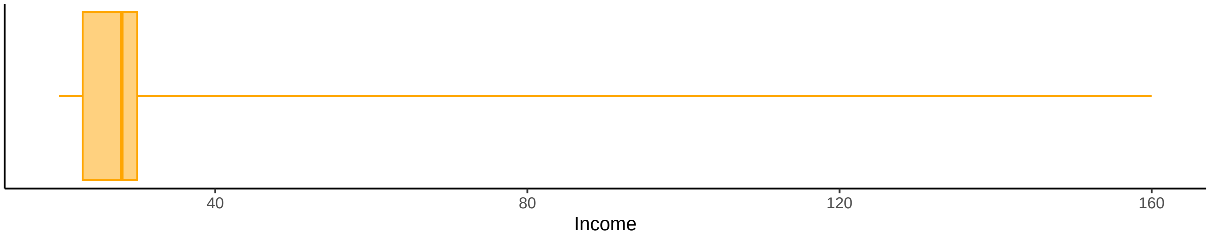
Let's plot the mean and the median with the data.



The distribution is characterised by a cluster of observations around 30 and a single large outlier. Because of this outlier, sample mean is significantly larger than the values in the main cluster of observations. Sample median gives a more accurate description of the main cluster of observations.

(b) The sample standard deviation depends on the sample mean. Since the sample mean is affected by the single outlier, so is the sample standard deviation. As a consequence, the sample standard deviation is high, which indicates a high degree of variability in the data. However, in the main cluster of observations, values are relatively close to each other. This is more informatively summarized by the interquartile range (7), which indicates a relatively low degree of variability in the data.

(c) The box-and-whisker plot usually displays 5 values: The minimum, the 25th percentile, the median, the 75th percentile, and the maximum. (Sometimes a box plot is drawn excluding outliers - here, we include the outliers too.)



(d) If the largest value in the sample is 360 instead of 160, our summary statistics change as follows:

- Sample average ≈ 56.6
- Sample median 28
- Sample standard deviation ≈ 100.7
- Interquartile range 7 (from 23 to 30)

Sample median and interquartile range are not affected by the value of the outlier.

2. (a) *Solution*

Let $h(c) = \sum_{i=1}^{n} (X_i - c)^2 = \sum_{i=1}^{n} \big(X_i^2 - 2X_i c + c^2\big)$

To find the minimum, we take the first-order condition w.r.t. c:

$$h'(c) = \sum_{i=1}^{n} \big(-2X_i + 2c\big) = -2\sum_{i=1}^{n} \big(X_i - c\big)$$

We wish to solve for c such that $h'(c) = 0$: $h'(c) = 0 \iff \sum_{i=1}^{n} \big(X_i - c\big) = 0$

$$\begin{aligned}
&\sum_{i=1}^{n} \big(X_i - c\big) = 0 \\
&(X_1 - c) + (X_2 - c) + ... + (X_n - c) = 0 \\
&-nc + \sum_{i=1}^{n} X_i = 0 \\
&c = \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

Which is what we wanted to show.

(b) The term $X_i - c$ is the *difference* between each observation and some constant c. These differences are sometimes positive and sometimes negative - but if we square them, they are always positive. The summation $\sum_{i=1}^{n} (X_i - c)^2$ gives us *the sum of squared differences*. When we solve the minimization, we are looking for the value of c such that the sum of squared differences between each observation and c is the smallest possible.

In other words, we are looking for c such that c is closest possible to different values of X_i, when "closest possible" is measured by the squared differences.

Thus, the sample mean is the quantity which minimizes the sum of squared differences in the data.

Remark.

Notice that the sample median is the solution of the problem

$$\min_c \sum_{i=1}^{n} \big|X_i - c\big| \tag{1}$$

Thus, while sample mean minimizes sum of squared differences, sample median minimizes the sum of absolute differences.

The "best" measure of central tendency depends on the measure of distance that we want to minimize.

3. (a) Show that $\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2$.

Solution

We are showing a property related to *sample variance*.

$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2) \\
&= \sum_{i=1}^n X_i^2 - 2\bar{X}_n \sum_{i=1}^n X_i + n\bar{X}_n^2 \\
&= \sum_{i=1}^n X_i^2 - 2\bar{X}_n \cdot n\bar{X}_n + n\bar{X}_n^2 \\
&= \sum_{i=1}^n X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2 \\
&= \sum_{i=1}^n X_i^2 - n\bar{X}_n^2
\end{aligned}$$

proving the result.

(b) Show that $\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \sum_{i=1}^n (X_i - \bar{X}_n)Y_i = \sum_{i=1}^n X_iY_i - n\bar{X}_n\bar{Y}_n$.

Solution

We are showing a property related to *sample covariance*.

$$\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \sum_{i=1}^n (X_i - \bar{X}_n)Y_i - \sum_{i=1}^n (X_i - \bar{X}_n)\bar{Y}_n = \sum_{i=1}^n (X_i - \bar{X}_n)Y_i$$

because

$$\sum_{i=1}^n (X_i - \bar{X}_n) = \sum_{i=1}^n X_i - n\bar{X}_n = 0$$

It follows that

$$\sum_{i=1}^n (X_i - \bar{X}_n)Y_i = \sum_{i=1}^n X_iY_i - \bar{X}_n \sum_{i=1}^n Y_i = \sum_{i=1}^n X_iY_i - n\bar{X}_n\bar{Y}_n$$