Variability and its measures

Variability

Variability (or heterogeneity) is the tendency of natural and social phenomena to manifest themselves in different ways.

Main issues

- Measuring variability
- Control or explain variability in statistical analysis

Measures of variability

Type of measures for numerical variables

1. Measures expressing the variability in the data as difference between specific observations.
2. Measures expressing the variability in the data as the average “distance” of observations from their mean.
3. Measures expressing the variability in the data as the average of all the pairwise absolute differences between observations (less usual not considered here).

Note that all measures of variability must be positive to make interpretation and comparisons possible among groups of statistical units.

Measure of variability cont'd

Let X be a numerical variable and let $x_1, \ldots, x_n$ be the values it takes in the data.

Relevant measures of variability are

- The range (type 1.)
- The interquartile range (type 1.)

- The mean absolute deviance (type 2)
- The variance and the standard deviation (type 2)

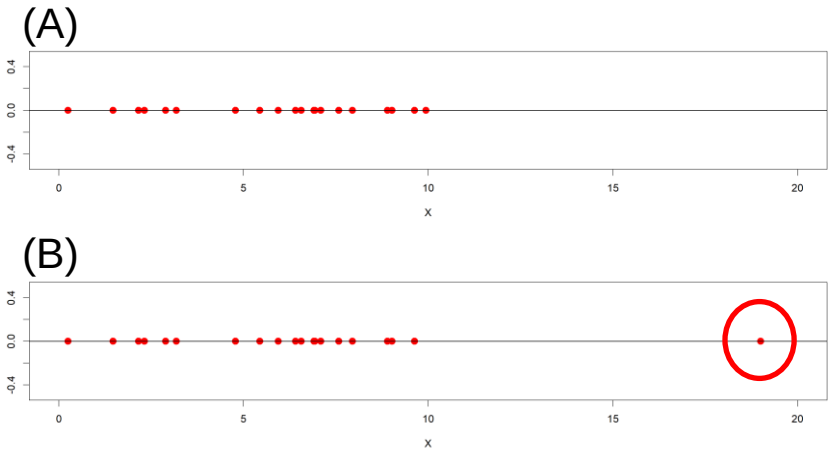
Range and interquartile range

Let $x_{(1)}, \ldots, x_{(n)}$ be the values arranged in ascending order

Range $R = x_{(n)} - x_{(1)} = Max - min$

Interquartile range $IQR = Q_3 - Q_1$

IQR is less sensitive to anomalous observations (outliers)



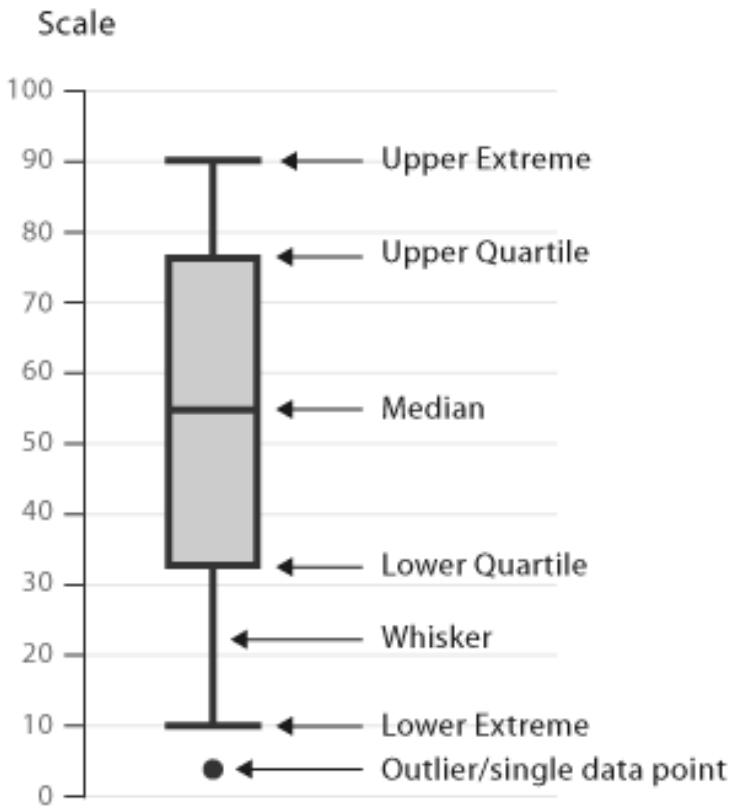
Data points differ only for one value.

IQR =4.65 is unchanged.

The range in case (B) is doubled.

The huge number in case (B) can be due to error (19 was typed instead of 9).

The boxplot

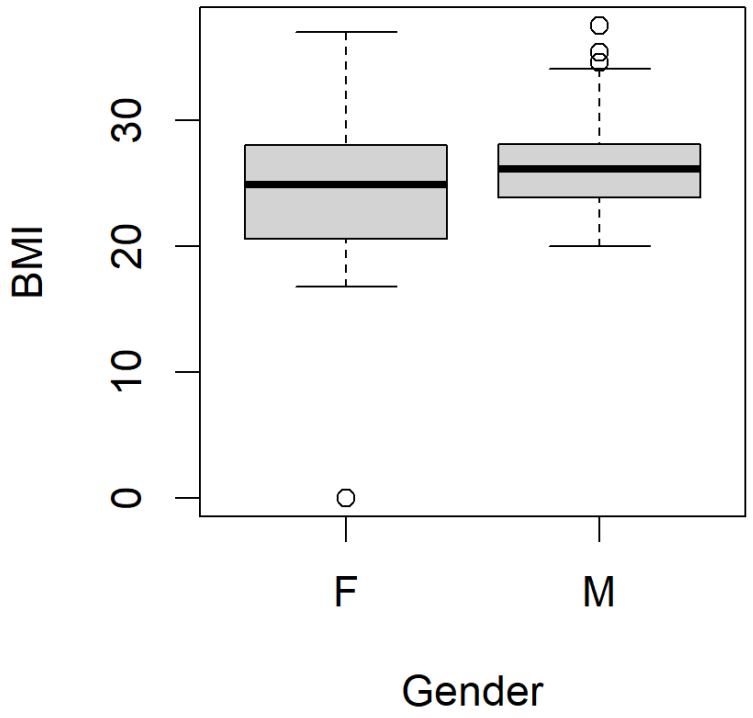


The box is drawn from Q_1 to Q_3 with a horizontal line drawn inside it to denote the median.

The whiskers are as long as 1.5 times the IQR value but are cut at the nearest sample point.

Because the whiskers must end at an observed data point, the whisker lengths can appear unequal

The box plot cont’d



The box plot is also useful for comparing distributions across different groups representing the location and the variability of the data as well as anomalous values by groups.

Mean Absolute Deviation

The Mean Absolute Deviation from the mean is given by

$$MAD = \frac{1}{n}(|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Note the use of the deviation modulus. This arise from

- $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ consequently signed deviances are not informative
- The focus is on the deviation from the centre whether positive or negative, therefore deviances with different sign should not compensate each other

Variance and standard deviation

The *variance* of X is given by

$$\tilde{S}^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The quantity $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is called *deviance*

The *standard deviation* of X is given by

$$\tilde{S} = \sqrt{\frac{1}{n}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note that the standard deviation has the same unit of measurement as X

Coefficient of variation

The coefficient of variation of X is given by

$$CV = \tilde{S}/\bar{X}$$

By calculating a relative measure of variability we eliminate, in some sense, the influence that the variable's average level has on the measure of variability considered.

CV is a pure number and does not depend on the unit of measurement of the variable under consideration.

Therefore, CV allows comparison of the variability of two or more sets of data expressed in different units of measurement.

Measure of heterogeneity for nominal variables

Let X be a nominal variable and let $f_1, \ldots, f_k$ be the relative frequency distribution of X where $f_j = \frac{n_j}{n}$ and $\sum_{j=1}^{k} f_j = 1$.

The frequency distribution is *perfectly heterogeneous* when all categories have the same frequency.

X	f
$\tilde{x}_1$	$1/k$
$\vdots$	$\vdots$
$\tilde{x}_j$	$1/k$
$\vdots$	$\vdots$
$\tilde{x}_k$	$1/k$

The frequency distribution is *perfectly homogeneous* when all observations fall into a single category.

X	f
$\tilde{x}_1$	0
$\vdots$	$\vdots$
$\tilde{x}_j$	1
$\vdots$	$\vdots$
$\tilde{x}_k$	0

Measure of heterogeneity for nominal variables cont'd

Let X be a nominal variable and let $f_1, \ldots, f_k$ be the relative frequency distribution of X where $f_j = \frac{n_j}{n}$ and $\sum_{j=1}^{k} f_j = 1$.

$$e = 1 - \sum_{j=1}^{k} f_j^2 \quad \textit{Gini's Index}$$

$$H = -\sum_{j=1}^{k} f_j \log f_j \quad \textit{Shannon's Index (hentropy)}$$

Both indexes take a value of 0 when heterogeneity is at its minimum, $f_j = 1$ for some j.

Both are limited upwards when heterogeneity is at its maximum, $f_j = 1/k$ for all j.