Statistical Computing
Week 3: 18. October 2019
Descriptive Statistics: Measures of location, quantiles, box plots

Lecture 3: Contents

- Mean
- Median
- Mode
- Geometric and harmonic mean
- Quantiles
- Box plot
- Strip chart

Measures of location

If we want to summarise a numeric variable using one number then we describe *where* the “middle” of the data lies. This is a measure of location.

The two most common measures of location are

- the **arithmetic mean** and
- the **median**.

Informally either of these may be referred to as the average.

Arithmetic mean

Assume we have a sample size of n and the n observations for variable X are $x_1, \ldots x_n$.

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} \qquad \text{or equvalently} \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The seven Russian dolls on the next page have the heights
11.3, 3.7, 5.0, 15.5, 6.6, 2.7 und 8.8 cm.
The mean is:

$$\frac{1}{7}(11.3 + 3.7 + 5.0 + 15.5 + 6.6 + 2.7 + 8.8) = \frac{53.6}{7} = 7.657$$

In R:

```
> sum(x)/length(x)
> mean(x)
```

Median

The median splits the data into two halves containing small and large values respectively.

$x_1, \ldots x_n$ are the data values in their original order.

If we sort the data with the smallest value first, we indicate this using round brackets around the index numbers.

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

The median is the middle value which is easy if n is odd, we find the $\frac{n+1}{2}$th orderd value

For odd n $$x_{0.5} = x_{\left(\frac{n+1}{2}\right)}$$

If n is even we need to take the mid point between the two middle values $x_{\left(\frac{n}{2}\right)}$ and $x_{\left(\frac{n}{2}+1\right)}$ so the median is

For even n $$x_{0.5} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

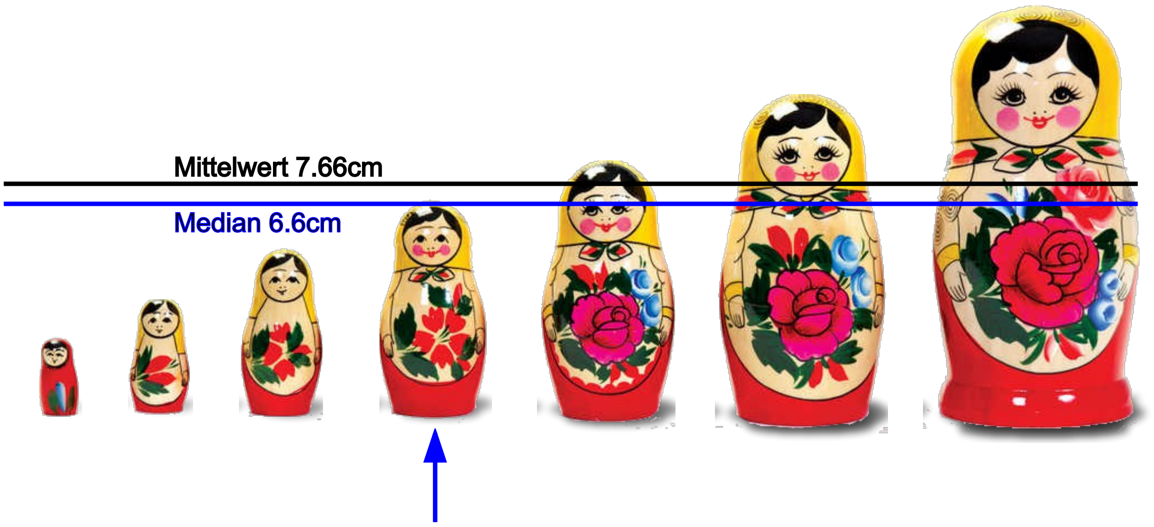
In R:

```
> median(x)
```

To find the median of our Russian doll heights we first order the data: $x_{(1)} = 2.7,\ x_{(2)} = 3.7,\ x_{(3)} = 5.0,\ x_{(4)} = 6.6,\ x_{(5)} = 8.8,\ x_{(6)} = 11.3$ and $x_{(7)} = 15.5$.

n is odd so we find $\frac{n}{2} + 1 = 8$ and so the median is the fourth ordered height $x_{0.5} = x_{(4)} = 6.6$

The height of the 4th smallest doll is the median value



The mode

The mode is the most frequently occurring value. It is usually used for nominal or ordinal variables.

Example from last week: Absolute frequency table for degree subject in the Fake student data set.

Subject	Biotech.	Economics	Elec. Engineering	Maths
Frequency	10	6	5	5

The mode of degree subject is Biotechnology, because it is the most frequently occurring.

The mode can also be used for discrete values when there are not to many unique values in that variable. E.g. for the variable *number of siblings* in the Fake student data, the mode is 1 sibling, which occurs 11 times.

Other measures of location

Just for completeness:

Geometric mean:

$$\overline{x}_G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

Usefull when the values span over many orders of magnitude.

Harmonic mean

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}$$

Used in certain applications such as average speed calulations.

Linear Transformation of a mean and median

Suppose the variable X is transformed into a new variable Y using the formula $Y = aX + b$, where a and b are known constants.

$$\begin{array}{lcl} \text{If} & y_i = ax_i + b & \text{for } i = 1, \ldots, n, \\ \text{then} & \overline{y} = a\overline{x} + b & \\ \text{and} & y_{0.5} = ax_{0.5} + b. & \end{array}$$

This means that if the arithmetic mean or median value of the variable is known, then it is not necessary to transform all the n data values, but just the mean or median.

For an example see the exercise in the work sheet.

The sum of two variables

When a new variable Z is computed as the sum of two existing variables X and Y then

$$\begin{array}{lll} \text{If} & z_i = x_i + y_i & \text{for} \quad i = i, \ldots, n \\ & \overline{z} = \overline{x} + \overline{y} & \\ \text{and} & z_{0.5} = x_{0.5} + y_{0.5} & \end{array}$$

An example is income for married partners: x_i is the husband's income, y_i is the wife's income, and z_i is the joint income for each couple.

You only need to know the two individual means $\overline{x}$ and $\overline{y}$ to calculate the mean joint income $\overline{z}$

Quantiles

The median X is the value of X which splits the data in two groups with half the data in one group and half the data in the other group, and has the notation $x_{0.5}$

The point which splits the data into 10% and 90% is called the 0.1-quantile.

The point which splits the data into a proportion p and $(1 - p)$ is called the pth-quantile.

The formula to find a the pth-quantile by hand is depends on whether pn is a whole number.

$$\begin{array}{ll} \text{If } pn \text{ is an integer} & x_p = \frac{1}{2}\left(x_{(pn)} + x_{(pn+1)}\right) \\ \text{If } pn \text{ is } \textbf{not} \text{ an integer} & x_p = x_{(\lceil pn \rceil)} \end{array}$$

$\lceil pn \rceil$ means round pn **up** to the next integer. See exercise sheet.

Note that there are many very similar ways of calculating quantiles. The above method is chosen to be easy to calculate by hand. R and other software use slightly different methods of calculating quantiles.

In R:

```
> x<-5:15
> quantile(x,0.5)
50%
 10
> quantile(x,0.1)
10%
  6
> quantile(x,c(0.1,0.11))
10% 11%
6.0 6.1
```

Quartiles

The median is a particular name for the 0.5-quantile.

There are two further quantiles with particular names

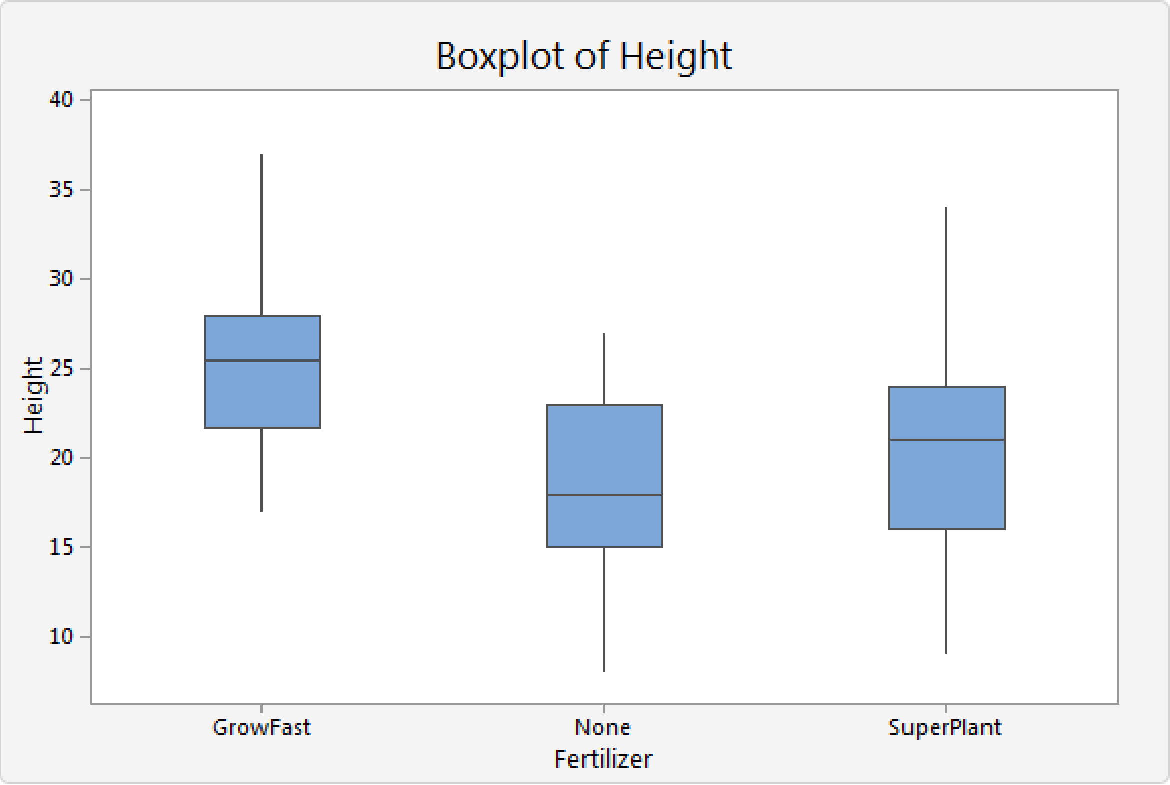
$Q_1 = x_{0.25}$ the 1st or lower quartile is the 0.25-quantile.

$Q_3 = x_{0.75}$ the 3rd or lower quartile is the 0.75-quantile.

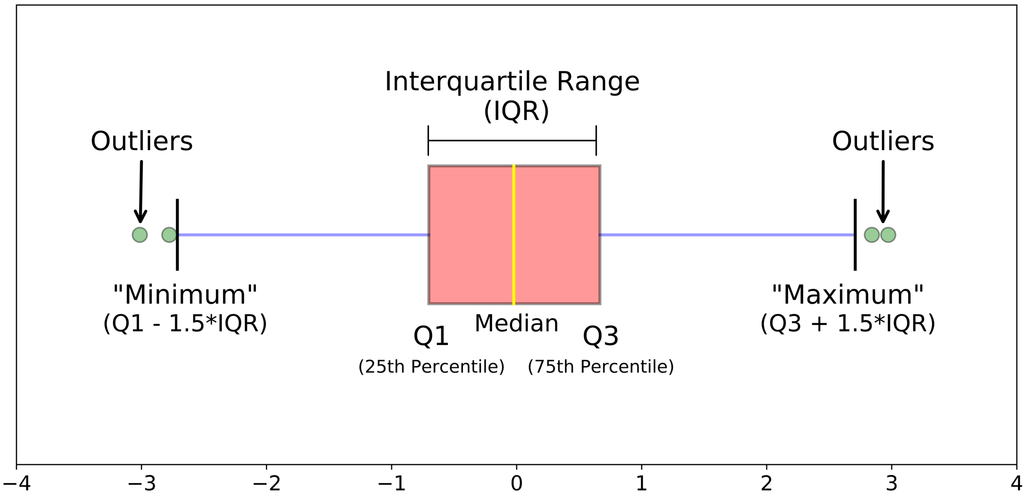
You occasionally see the word “decile” which means x_p where p is one of $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$

Box plots

Example



- A box plot gives a quick visualisation of the distribution of the values in a continuous variable. It can also be used for discrete variable when it contains many different values.
- As in the previous slide, you can easily compare the distribution of a continuous variable split into two or more groups.
- The official name is *Box and Whisker Plot*, due to the lines to either side of the box looking like cat's whiskers.



- The box contains the middle 50% of the data values.
- **The whisker ends always point to a value in the data.** The exact position of the whisker is
  - Calculate Q1-1.5*IQR, then find the next data value *in the direction of the box*.
  - Calculate Q3+1.5*IQR, then find the next data value *in the direction of the box*.
- Values which are outside of the whiskers are plotted as points and are called *box plot outliers*.

Box plot: worked example

The Body-Mass Index BMI was measured for 25 adults in the UK. The units are kg/m^2.

18.13	18.53	20.75	21.86	22.65
22.93	22.95	23.75	23.82	24.01
24.68	24.89	25.25	25.75	25.85
25.90	26.11	26.73	27.20	27.67
27.94	28.19	29.29	31.22	32.37

Median $Q_2 = 25.25$

Lower quartile $Q_1 = x_{(7)} = 22.95$ Upper quartile $Q_3 = x_{(19)} = 27.20$

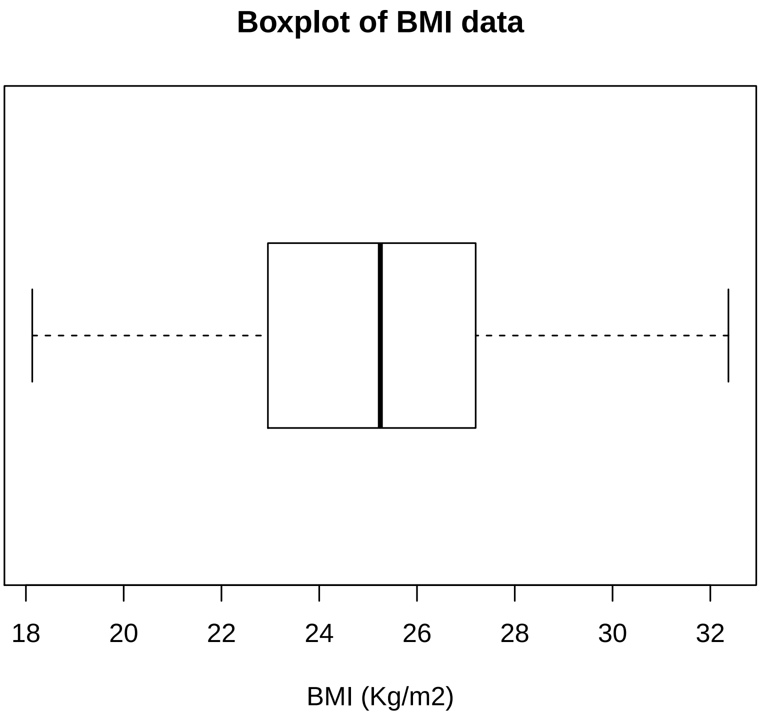
worked example (ctd.)

Interquartile range $IQR = 27.20 - 22.95 = 4.25$

Lower whisker value first step: $Q_1 - 1.5 \cdot IQR = 16.575$
Second step: the lower whisker must point to a data value.
Start from 16.575 and move in the direction of the box.
In this case it is the minimum value 18.13.

Upper whisker value first step: $Q_3 - 1.5 \cdot IQR = 33.575$
Secondstep: the lower whisker must point to a data value.
Start from 33.575 and move in the direction of the box.
In this case it is the maximum value 32.37.

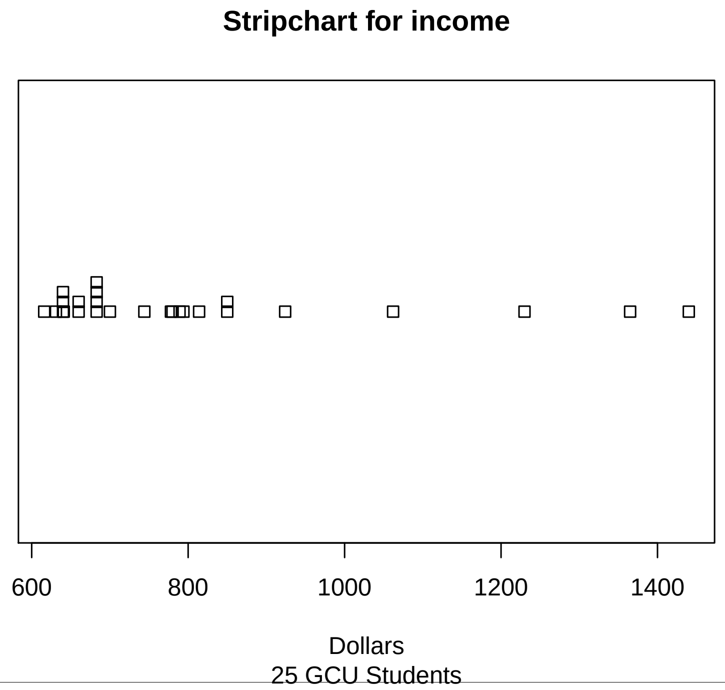
For these BMI data no point lies outside the whisker values, so there are no box plot outliers.



The box plot shows that the BMI data are approximately symmetrical with no extreme values.

Strip chart

If the sample size is small then a *strip chart* is better. Also known as dotplot or 1-dimensional scatterplot. Each data value is plotted with *ties* placed on top of one another.


Stripchart for income
Dollars
25 GCU Students