Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part B

Quartiles, IQR & Box Plots

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

How do we Measure Variability?

Variability (spread) can be measured by:

- **Variance σ^2 or s^2**
  - uses all data values but is inflated by outliers
- **Standard deviation σ or s**
  - uses all data values but is inflated by outliers
- **Range** $=$ maximum $-$ minimum = $x_{(n)}$ - $x_{(1)}$
  - unreliable measure, depends on extreme values
- **Interquartile range**: $IQR = Q_3 - Q_1$
  - spans middle 50% of data,
  - unaffected by outliers, ignores variation in tails

Interquartile Range (IQR)

- IQR = Upper quartile (Q_3) - lower quartile (Q_1)
- or IQR = 75^{th} - 25^{th} percentile
- There are different ways of calculating quartiles:
  we will use the repeated median method
  - $Q_1 =$ median of lower half of sorted data.
  - $Q_3 =$ median of upper half of sorted data.
  - For n even, split the data into two halves - find median of lower half to get Q_1; find median of upper half to get Q_3.
  - For n odd, **leave Q_2 in both halves** to find Q_1 and Q_3 .

Five-number summaries

Data for a quantitative variable can be summarised by giving the following five numbers:

1. the minimum value, $x_{(1)}$ (min)
2. the lower quartile, Q_1 (or LQ)
3. the median, Q_2
4. the upper quartile, Q_3 (or UQ)
5. the maximum value. $x_{(n)}$ (max)

The ordered set $(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$ is the **five-number summary** of the data.

They can be used to construct box plots.

Exercise: Handspan Set 1

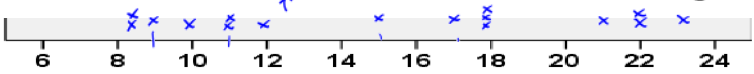
$n = 16$

Draw a dot plot and calculate the 5 number summary for handspan (right) for 16 students in a tutorial class.

8.5, 8.5, 9, 10, 11, 11, 12, 15, 17, 18, 18, 18, 21, 22, 22, 23 (cm)

8th 9th

Dot Plot:



For Q_2

$\frac{n+1}{2} = \frac{17}{2} = 8.5$ th value

$Q_2 = 16$.

For Q_1

$\frac{n+1}{2} = \frac{9}{2} = 4.5$ th value

$Q_1 = 10.5$

Q_3 4.5th from 17.

18 21

$\frac{18+21}{2} = \frac{39}{2} = 19.5$

$Q_3 = 19.5$

5 no. summary is

(8.5, 10.5, 16, 19.5, 23) cm.

Exercise: Handspan Set 2

Draw a dot plot and calculate the 5 number summary for handspan (right) for 17 students in a tutorial class.

$n = 17$.

8.5, 9, 9, 10, 11, 11.5, 12, 15, 17, 18, 19, 20, 21, 22, 22, 22.5, 23 cm.

(1 2 3 4 5)

Dot Plot:

Median Q_2

$\frac{n+1}{2} = \frac{18}{2} = 9$th value

$Q_2 = 17$.

For Q_3 $Q_3 = 21$. (incl 17).

For Q_1

$\frac{n+1}{2} = \frac{9+1}{2} = 5$th value

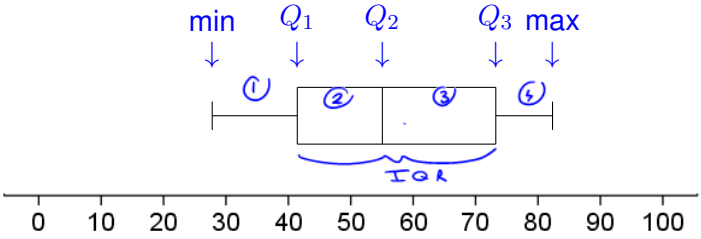
$Q_1 = 11$

5. no. summary is

(8.5, 11, 17, 21, 23) cm.

Basic box plots

The most basic **box plot** is a box-and-whisker diagram drawn alongside a scale to indicate the **five-number summary**.



The interquartile range (=$Q_3 - Q_1$) is the length of the central box.

Question: What proportion of points lie within each section of the box plot? 25%

Box plots

- Centre and spread can be seen at a glance.
- Width of box (if drawn horizontally) or Height of box (if drawn vertically) is IQR.
- Shows whether the distribution is
  - approximately symmetric (equal whiskers, crossbar in the middle of the box) or
  - not symmetric so it is skewed.
- Can plot boxplots side-by-side on same scale when comparing 2 or more groups.
- Outliers can be plotted separately as dots.

Quartiles in R

R code: In R, the repeated median method is implemented in the : fivenum function

- For n even: (n=16)

```
x <- c(8.5, 8.5, 9, 10, 11, 11, 12, 15, 17, 18, 18, 18, 21, 22, 22, 23)
fivenum(x)
[1]  8.5 10.5 16.0 19.5 23.0
```

min Q1 Q2 Q3 max

✓ agree

```
IQRx <- fivenum(x)[4] - fivenum(x)[2]
IQRx
[1] 9
```

↑ 4th element

↑ 2nd element

$IQR = Q_3 - Q_1$

$= 19.5 - 10.5$

$= 9.$ ✓

Quartiles in R cont.

R code: In R, the repeated median method is implemented in the : fivenum function

For n odd: (n=17)

```
y<-c(8.5, 9, 9, 10, 11, 11.5, 12, 15, 17, 18, 19, 20, 21, 22, 22, 22.5, 23)
fivenum(y)
[1]  8.5 11.0 17.0 21.0 23.0
```

min Q_1 Q_2 Q_3 max

$IQR = 21 - 11 = 10$

```
IQRy <- fivenum(y)[4] - fivenum(y)[2]
IQRy
[1] 10
```

Quartiles in R - a different method

For your information, another widely used definition is to use the ranks where:

Q_1 is the $\frac{(n+3)}{4}^{th}$ observation; Q_3 is the $\frac{(3n+1)}{4}^{th}$ observation.

In R, this method is implemented in the quantile function

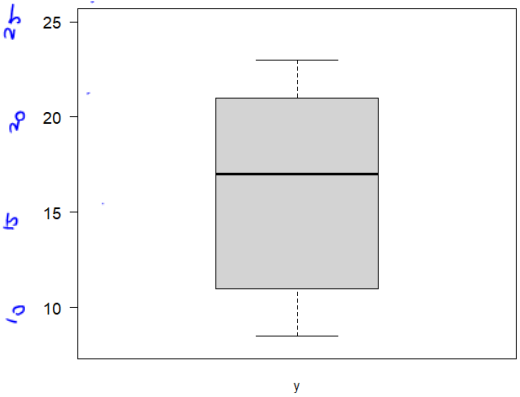
```
quantile(x)
0%   25%   50%   75%  100%
8.50 10.75 16.00 18.75 23.00

quantile(y)
0%  25%  50%  75% 100%
8.5 11.0 17.0 21.0 23.0
```

In R: Box Plots

R code:

boxplot(y) draws a single boxplot of data y.

```
boxplot(y, xlab="y",
        ylim =c(8, 25), las=1) )
```



In R: Box Plots

R code:

Add options for labels:

```
boxplot(Temps_Airport$Temp_Wollo,
ylab="Temperature (degrees Celcius)",
xlab="Wollongong Monthly Average
Temperature")
```