Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part C

Identifying Outliers & Errors

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Outliers on box plots

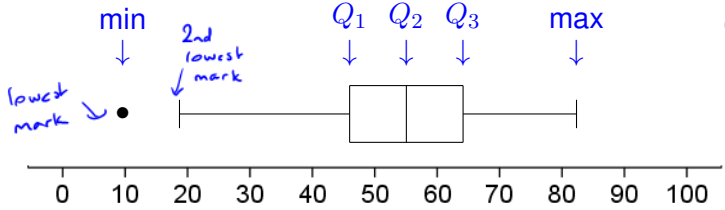
Outliers

A data point is identified as an outlier

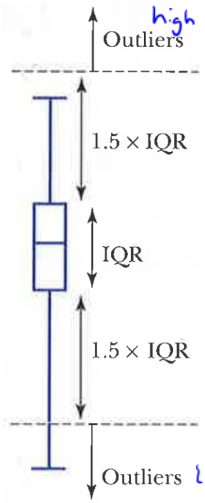
if it is more than $1.5 \times$ IQR beyond the upper or lower quartiles

It is marked separately with a dot (usually) or a cross.

The whiskers are then drawn only as far as the most extreme points which are not outliers.



Identifying Outliers

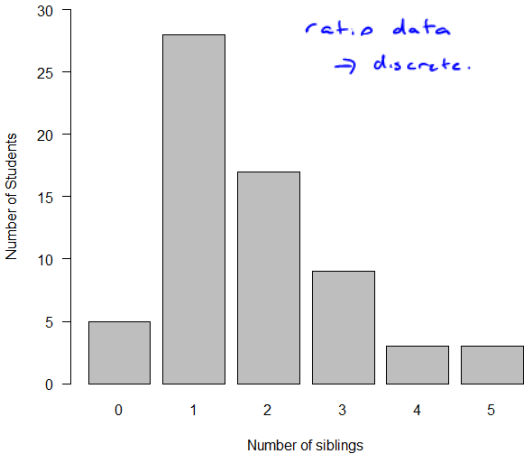


Steps to identifying outliers:

1. **Sort** the n data values in ascending order or create a stem-and-leaf plot
2. Determine the five number summary
3. Calculate $IQR = Q_3 - Q_1$
4. Calculate bounds for low/high outliers
   - Low: bound is $Q_1 - 1.5 \times IQR$
   - High: bound is $Q_3 + 1.5 \times IQR$
5. Check for data values outside these bounds:
   - Are there any $x_{(i)} < Q_1 - 1.5 \times IQR$ $\Rightarrow$ low outliers
   - Are there any $x_{(i)} > Q_3 + 1.5 \times IQR$ $\Rightarrow$ high outliers

Reference: p57 Griffiths Stirling & Weldon (1998) Understanding Data

Example 1: Number of Siblings



No. of Siblings	0	1	2	3	4	5
Frequency	5	28	17	9	3	3
Cum. Freq.	5	33	50	59	62	65

- Range $= x_{(n)} - x_{(1)} =$ 5−0 = 5
- Median
  $Q_2 =$ 1

  $\frac{n+1}{2} = \frac{66}{2} = 33$rd value.
- Quartiles:
  $Q_1 =$ 1

  $\frac{n+1}{2} = \frac{34}{2} = 17$th value

  $Q_3 =$ 2

Example 1: Number of Siblings cont.

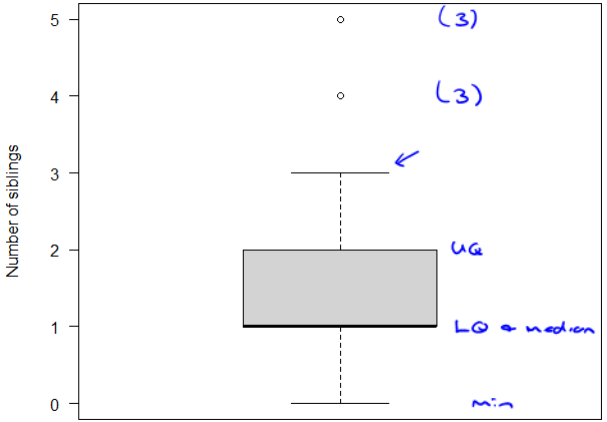
- 5-number summary: $(0, 1, 1, 2, 5)$ (Q_1, Q_3)
- Interquartile Range:
  - $IQR = Q_3 - Q_1 = 2 - 1 = 1$
- Calculate bounds:
  - Low bound is: $Q_1 - 1.5 \times IQR = 1 - 1.5 \times 1 = -0.5$ N/A.
  - High bound is: $Q_3 + 1.5 \times IQR = 2 + 1.5 \times 1 = 3.5$
- Identify outliers:

No low outliers

High outliers : 4, 4, 4, 5, 5, 5.

Example 1: Number of Siblings cont.

Box plot for Number of Siblings for 65 Students



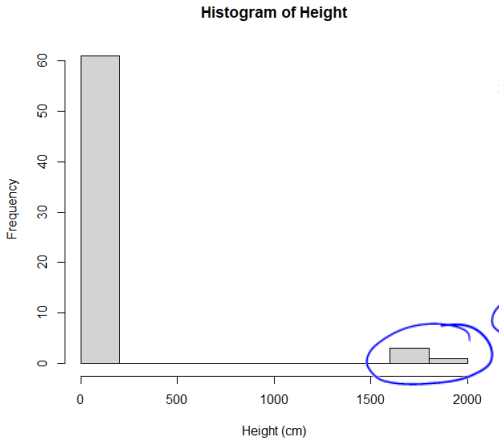
Identifying Errors

Checks

- Outliers are unusual values in the data set
- Check if they are
  - errors- correct if possible - go back to source
  - valid observations - do not usually discard - can check the affect on analysis

Example 2: Height (cm)

The heights in cm were measured for 65 students in a Maths subject.
Check the data by creating a stem & plot and/or a histogram.




```
> stem(Height)

  The decimal point is 2 digit(s) to the right of the |

   0 | 566667777777777777777777777777888888888888888888899999999999
   2 |
   4 |
   6 |
   8 |
  10 |
  12 |
  14 |
  16 | 048
  18 | 7
```

Example 2: Height (cm) cont.

Check data: Identify any errors:

```
Heightsort<-sort(Height)
print(Heightsort[55:65])
[1]  187.0  188.0  188.5  190.0  190.0  194.0  194.0 1700.0
[9] 1740.0 1780.0 1866.0
```

mm

What should we do?

correct the obs from mm into cm.

Example 2: Height (cm) cont.

Correct those observations that were recorded in mm to be in cm:

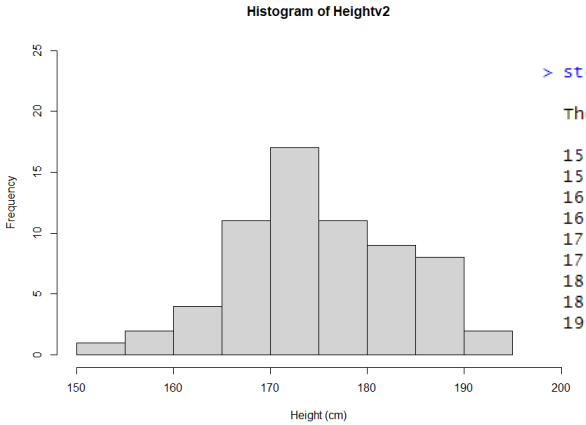
```
Heightv2 <- c(Heightsort[1:61], Heightsort[62:65]/10)

print(Heightv2)
[1] 151.0 157.0 157.0 161.4 163.2 165.0 165.0 166.0 166.0 167.0 168.0
[12] 168.0 168.1 169.0 170.0 170.0 170.0 170.5 171.0 171.0 171.0 171.0
[23] 171.0 171.0 171.0 172.0 172.0 172.0 172.5 173.0 174.0 175.0 175.0
[34] 176.0 176.8 177.0 177.0 177.0 177.0 178.0 178.0 179.0 180.0 180.5
[45] 182.0 182.0 182.0 182.0 183.0 184.2 185.0 185.0 186.0 186.2 187.0
[56] 188.0 188.5 190.0 190.0 194.0 194.0 170.0 174.0 178.0 186.6
```

cm

Now corrected.

Example 2: Height (cm)- corrected

Redo the plots:




```
> stem(Heightv2)

  The decimal point is 1 digit(s) to the right of the |

  15 | 1
  15 | 77
  16 | 13
  16 | 556678889
  17 | 00001111111112223344
  17 | 556777778889
  18 | 01222234
  18 | 55667789
  19 | 0044
```

151.

190 190 194 194 cm

Example 2: Height (cm)- corrected

```
fivenum(Heightv2)
[1] 151 170 174 182 194
```

min Q_1 Q_2 Q_3 max.

```
IQRht <- fivenum(Heightv2)[4] - fivenum(Heightv2)[2]
IQRht
[1] 12
```

```
rangeHt <- fivenum(Heightv2)[5] - fivenum(Heightv2)[1]
rangeHt
[1] 43
```

Example 2: Height - corrected

- Range $= x_{(n)} - x_{(1)} =$ 194 - 151 = 43 cm
- Median
  $Q_2 =$ 174 cm
- Quartiles:
  $Q_1 =$ 170 cm

  $Q_3 =$ 182 cm

confirm the R values using hand calcs.

Example 2: Height - corrected

- 5-number summary: (151, 170, 174, 182, 194) cm. (min)
- Interquartile Range:
  - $IQR = Q_3 - Q_1$ = 182 − 170 = 12 cm ✓
- Calculate bounds:
  - Low bound is: 170 − 1.5 × 12 = 152 cm.
  - High bound is: 182 + 18 = 200 cm ⇒ ×
- Identify outliers:

  1 low outlier = 151 cm

  No high outliers.

Example 2: Height - corrected cont.

Box plot for Height for 65 Students