Topic: Exploratory Data Analysis (EDA)

Measures of Centre

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Statistical Process

- Ethics
- Nature of the question to be answered
- Context/Expertise
- Design:
  - Experiment vs. observational study
  - Sampling
  - Measurement
- **Description and analysis**
- Conclusions and decision making

Measures of Centre/Location Statistics

- **Population Mean:**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + \ldots + x_N)$$

- **Sample Mean:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

- **Trimmed Mean:** Average after eliminating a percentage of the highest and lowest. Eg. some packages use 5%
- **Median:** Middle score when data values arranged in order from smallest to largest
- **Mode:** Most frequent score

Quantitative Data: Sample Mean

- Consider n quantitative data values $x_1, x_2 \ldots, x_n$
- The **mean** is the *average* value:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

capital 'sigma'

- **Notation:** x_i denotes the ith value in the list (with no order), $i = 1, \ldots, n$
- In R:

```
x <- c(3,1,4,5,9)
mean(x)
```

$n = 5$

$\bar{x} = \frac{1}{5} \times 22 = \frac{22}{5} = 4.4$

4.4

Sample Median

To find the median, Q_2, the *middle* score

1. **Sort** the n data values in ascending order: $x_{(1)}, x_{(2)} \ldots, x_{(n)}$
   Note $x_{(i)}$ denotes the ith value in the list (with ascending order) $i = 1, \ldots, n$.
2. Calculate $\frac{n+1}{2} \Rightarrow$ position / rank (of middle position)
3. Determine the observation which is in the $\left(\frac{n+1}{2}\right)$th position
4. Unlike the mean which can be pulled up or down by unusual data values (outliers), the median is only affected slightly by outliers: It is more **robust.**
5. In R:

```
x <- c(3,1,4,5,9)

median(x)
```

$x_{(1)} = \min$

$x_{(n)} = \max.$

① ② ③

1, 3, 4, 5, 9

$Q_2 = 4.$

$n = 5$

$\frac{n+1}{2} = \frac{6}{2} = 3$rd value.

Sample Median - Exercise

Determine median using the calculated rank:

- for **odd** n, the median is the *middle* sorted data value.
  Ex: determine median of $\{6, 5, 8, 2, 9\}$ $\Rightarrow$ $2, 5, \boxed{6}, 8, 9$ (3rd value.)

  $n = 5$ $\quad \frac{n+1}{2} = 3$rd value

  $Q_2 = 6.$

- for **even** n, the median is the average of the middle 2 data values.
  Ex: determine median of $\{2, 5, \boxed{6, 8}, 9, 11\}$

  $n = 6.$

  $\frac{n+1}{2} = \frac{7}{2} = 3.5$th value

  3rd $\downarrow$ 3.5th value $\downarrow$ 4th

  6 $\quad$ 7 $\quad$ 8

  $\frac{6+8}{2} = 7$ $\quad Q_2 = 7.$

Exercises

Determine the mean, median and mode of the following:

1. **Set 1:** 1, 8, 4, 2, 7, 8

in order $\Rightarrow$ 1, 2, 4, 7, 8, 8 $\quad n = 6$.

$\bar{x}_1 = \frac{30}{6} = 5$ $\quad$ mode = 8 $\quad$ $\frac{6+1}{2} = 3.5$th value

3rd 4th
4 7

$Q_2 = \frac{4+7}{2} = 5.5$

2. **Set 2:** 1, 8, 4, 2, 7, 8, 40 $\quad n = 7$

1 2 4 7 8 8 40

$\bar{x}_2 = \frac{70}{7} = 10$. $\quad$ median

$\frac{n+1}{2} = \frac{8}{2} = 4$th value.

mode = 8

$Q_2 = 7$. $\quad Q_2 < \bar{x}_2$

Exercise: Wollongong Monthly Average Temp: Jan 2009 - Jun 2018

Exercise: Find the median temperature from the stem-and-leaf plot $n = 114$:

Wollongong Temperature (Monthly Average)
Stem-and-Leaf Plot

CF	Frequency	Stem &	Leaf
1	1.00	16 .	6
12	11.00	17 .	11333357789
24	12.00	18 .	122333556789
28	4.00	19 .	0348
39	11.00	20 .	00244445789
45	6.00	21 .	012499
55	10.00	22 .	0033444689
68	13.00	23 .	1224567777889
	4.00	24 .	1679
	22.00	25 .	0123334455666677889999
	9.00	26 .	001256789
	6.00	27 .	014468
	4.00	28 .	0113
	1.00	29 .	8

Stem width: 1.0
Each leaf: 1 case(s)

16.6°C

22.9°C
55th value

29.8°C

$\frac{n+1}{2} = \frac{115}{2} =$ 57.5 th value

57th: 23.2°C
58th: 23.2°C

$Q_2 = 23.2°C$

Leaf unit $= \frac{1}{10}$ th Stem unit
$= 0.1$

In R: Wollongong Monthly Average Temp: Jan 2009 - Jun 2018

↓ column name.

```
> median(Temps_Airport$Temp_Wollo)
[1] 23.2
```

✓

```
> stem(Temps_Airport$Temp_Wollo)
```

The decimal point is at the |

16 | 6 16.6°C
17 | 11333357789
18 | 122333556789
19 | 0348
20 | 00244445789
21 | 012499
22 | 0033444689
23 | 1224567777889
24 | 1679
25 | 0123334455666677889999
26 | 001256789
27 | 014468
28 | 0113
29 | 8