Topic: Exploratory Data Analysis (EDA)

Shape & Reporting of Univariate Data

School of Mathematics and Applied Statistics

UNIVERSITY OF WOLLONGONG AUSTRALIA

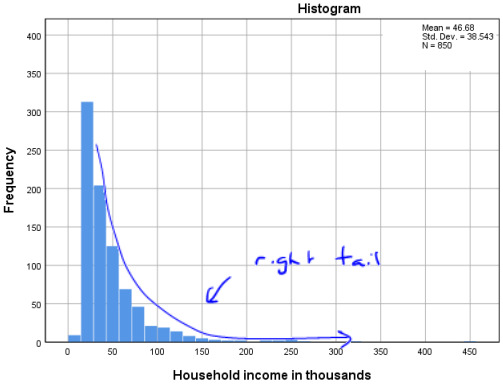
Meaningful Reporting - Univariate Data

1. **Context**
   - Units, sampling, design of collection, research findings, daily recommended exposure
2. **Shape**
   - Bell, normal, uniform, symmetric, unimodal, skewed, bimodal
3. **Outliers/Extremes** - different software packages may define differently
4. **Centre** - Mean, median, mode, trimmed mean
5. **Spread** - Range, IQR, Variance, Standard Deviation
6. **Patterns** - are there any?

Tails of Distribution

- The **left-hand tail** is the region of **lowest** data values.
- The **right-hand tail** is the region of **highest** data values (don't confuse with highest frequency).

E.g. income distributions typically have a **long** right-hand tail; a minority have much higher incomes than the majority.



Skewness

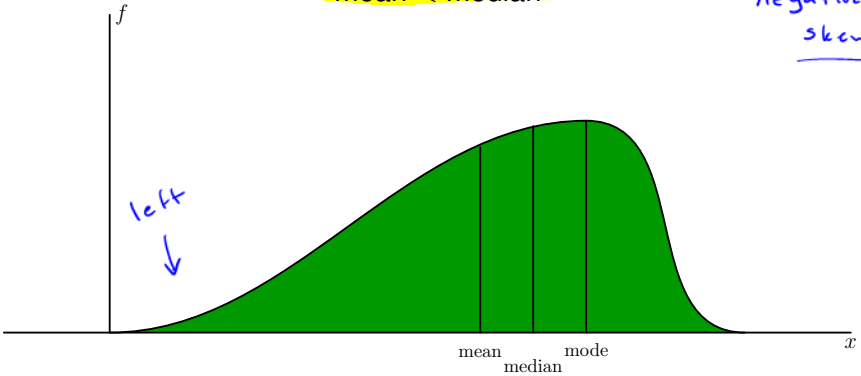
The direction of the **skew** is determined by the **location of the tail**

If the tail is on left then the distribution is *skewed to the left*

The mean is dragged down by unusually small values in the left tail,

mean $<$ median

negatively skewed



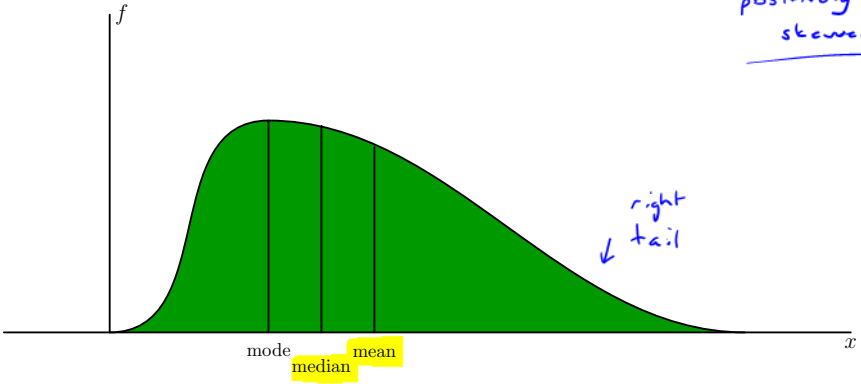
Skewed Distribution

If the tail is on right then the distribution is *skewed to the right*

The mean is inflated due to unusually large data values in the right tail

mean $>$ median

'positively skewed'



Meaningful Reporting - Univariate Data

1. **Context**
   - Units, sampling, design of collection, research findings, daily recommended exposure
2. **Shape**
   - Bell, normal, uniform, symmetric, unimodal, skewed, bimodal
3. **Outliers/Extremes** -defined differently with different plots/data
4. **Centre** - Mean, median, mode, trimmed mean
5. **Spread** - Range, IQR, Variance, Standard Deviation
6. **Patterns**

Meaningful Paragraph: Describing Data

Univariate analysis - one variable at a time.

The aim is to turn data into meaningful information covering all major aspects of the data, with precision AND to communicate that information (paragraphs).

- Example:
  - Each row represents one item of food ⇒ can of soup
  - Two columns of data ⇒ 2 variables: fat and sodium

    - both of type quantitive and ratio.

Fat	Sodium
0.5	120
9.0	20
3.0	140
1.0	65
0.5	110
2.0	300
2.0	160
0.0	150
6.0	240
3.0	320
0.5	210
0.5	220
1.5	200
2.0	280
3.5	210
1.0	190
1.0	270
0.5	230
0.0	300
0.0	300
0.0	120
6.0	170
0.0	170
0.0	210
1.0	140
1.0	210
1.0	170
1.0	150
1.5	210

Meaningful Paragraph: Context

Here is nutrition information (fat) taken from cans of soup.
To make sense of this data we need to know:

- How is fat measured? What units are used? (g)
- How were they sampled? eg. Brands? Flavours? Shelves in supermarket/s?
- Sample information? Sample size? $n = 76$
- Previous study findings on similar products:- general context
  - what is the maximum or minimum recommended daily intake for adults? for children?
  - mean?
  - or spread?
  - or...?
  - groupings?

Fat
0.5
9.0
3.0
1.0
0.5
2.0
2.0
0.0
6.0

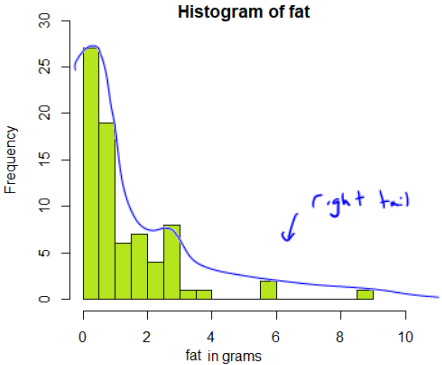
Meaningful Paragraph: Shape

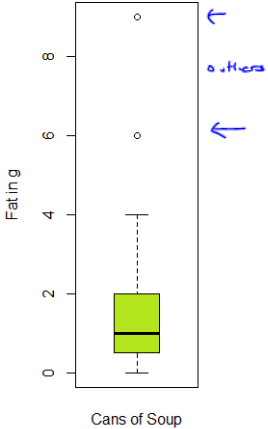
- Quantitative – look at different plots
  - stem & leaf
  - histogram ⇒ grouped discrete or continuous
  - bar chart ⇒ a small no. of discrete values
  - boxplot ⇒ 5-number summary
  - dot plot
- Different versions of the same plot
  - Drawing by hand with different stems
  - Different scales or bin widths
- Why?
  - Helps determine analysis – which
    - measure of centre
    - and spread to use
  - Helps to find unusual values

Meaningful Paragraph: Shape

Quantitative – look at 3 different plots of same data

```
> stem(fat)

  The decimal point is at the |

  0 | 000000000000000000555555555555
  1 | 000000000000000000000555555
  2 | 00000005555
  3 | 000000005
  4 | 0
  5 |
  6 | 00
  7 |
  8 |
  9 | 0
```



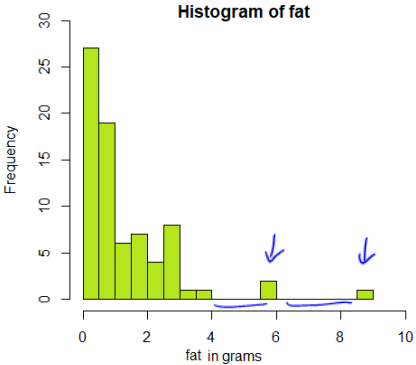


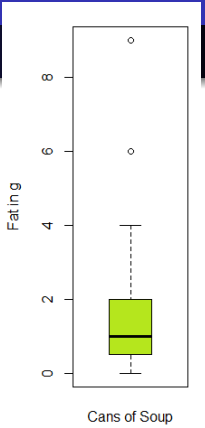
Description

- longer tail of high values $\Longrightarrow$ high fat content for a small no. cans.
- skewed distribution
  - Positively skewed or skewed to the right

Meaningful Paragraph: Shape

```
> stem(fat)

  The decimal point is at the |

  0 | 000000000000000000555555555555
  1 | 000000000000000000000555555555
  2 | 00000005555
  3 | 000000005
  4 | 0
  5 |
  6 | 00
  7 |
  8 |
  9 | 0
```





Stem & leaf

See Detail:

- Patterns
- Centre
- Spread
- Extremes not so easily

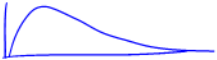
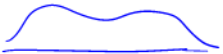
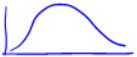
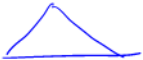
Histogram

- Loss of detail
- See outliers
- See gaps

Boxplot: summary plot

- See outliers – not all packages do so
- Symmetry, skewness
- Poorer shape and detail

Describing Shape

- Bell-shaped 
- Normal 
- Skewed to the right (or positively skewed) $\Rightarrow$ Longer tail of high values 
- Skewed to the left (negative) $\Rightarrow$ Longer tail of low values 
- Uniform 
- Unimodal versus Bimodal  
- Exponential 
- Symmetric - two halves same about centre   

Activity: Sketch an example for each

Different versions of one type of plot

Take care when describing shape

- Different no. of stems in a stem & leaf plot
- or different no. of bins in histogram
- Whether or not the plot shows outliers

Spot the differences:

fat Stem-and-Leaf Plot SPSS

```
 Frequency    Stem &  Leaf

    16.00        0 .  0000000000000000
    11.00        0 .  55555555555
    19.00        1 .  0000000000000000000
     6.00        1 .  555555
     7.00        2 .  0000000
     4.00        2 .  5555
     8.00        3 .  00000000
     1.00        3 .  5
     1.00        4 .  0
     3.00 Extremes    (>=6.0)

 Stem width:      1.00
 Each leaf:       1 case(s)
```

0 – 4
5 – 9

```
> stem(fat)

  The decimal point is at the |

  0 | 00000000000000000555555555555
  1 | 0000000000000000000555555
  2 | 00000005555
  3 | 000000005
  4 | 0
  5 |
  6 | 00
  7 |
  8 |
  9 | 0
```

R.

Meaningful Paragraph: Centre

- The mean uses all information in the sample because each value is added to the sum
- Mean is subject to error if spurious values are entered
- Median is less affected by "wild" values i.e. it is robust.

If the median is similar to the mean:

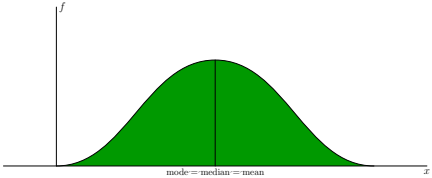
- Use the mean as it uses all data
- It is easier to work with means

If they are different because of non-symmetric distributions

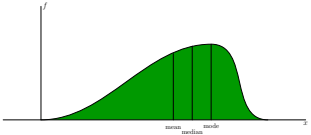
- Can be useful to report both
- The context of what the data are used for may also determine which is the appropriate statistic

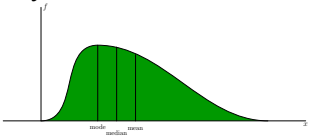
Centre: which measure to use?

- Symmetrical mean & median similar



- Skewed - use mean & median as they will differ





- When outliers present - use a robust measure, so outliers do not influence eg. median
- Use mode for nominal data

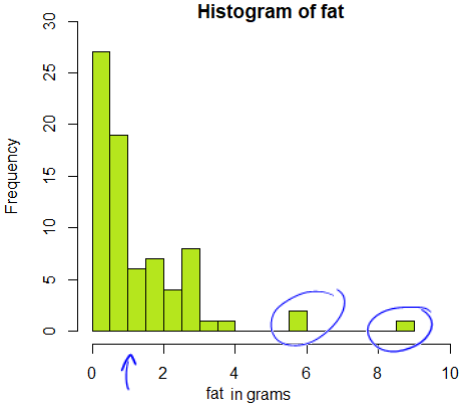
Meaningful Paragraph: Spread or Variability

Think: Which measure of spread is more appropriate for the variable Fat content?

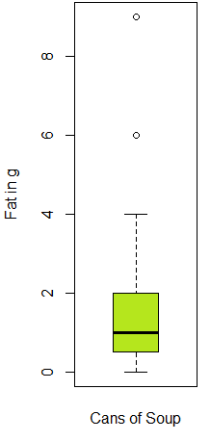
```
> stem(fat)

  The decimal point is at the |

  0 | 000000000000000055555555555
  1 | 000000000000000000555555
  2 | 00000005555
  3 | 000000005
  4 | 0
  5 |
  6 | 00
  7 |
  8 |
  9 | 0
```

mean = 1.447 g

median = 1.0 g



IQR ✓



Example: which measure of variability to use?

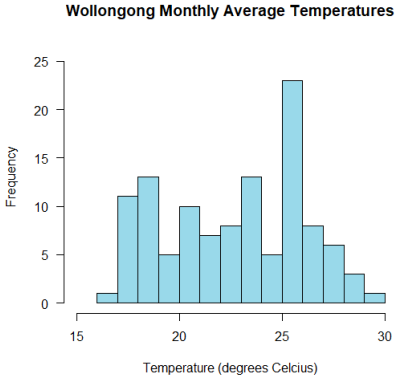
Think: which measure of variability might be appropriate for the variable Monthly Average Temperature?

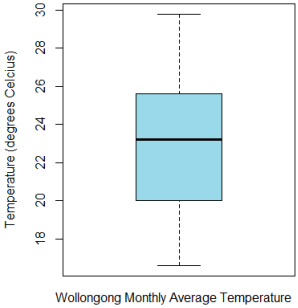
n=114.

```
The decimal point is at the |

16 | 6
17 | 11333357789
18 | 122333556789
19 | 0348
20 | 00244445789
21 | 012499
22 | 0033444689
23 | 1224567777889
24 | 1679
25 | 01233344556666778899 99
26 | 001256789
27 | 014468
28 | 0113
29 | 8
```

mean = 22.77° C

median = 23.2° C





Meaningful Paragraph: Patterns

Why Look?

- There may be something unusual detected about measurement
- Eg 1. Blood pressure taken in different countries may have used different instruments
- Eg 2. those measuring may be more or less prone to rounding numbers, for example at border between grades in exams
- There may be some importance attached to the pattern
  eg. ECG heart attack

Meaningful Paragraph: Patterns

Patterns are not often seen

Fat variable:

- Within this data set the measurement is rounded to the whole or half gram

```
> stem(fat, scale = 2)

  The decimal point is at the |

  0 | 0000000000000000
  0 | 55555555555
  1 | 00000000000000000000
  1 | 555555
  2 | 0000000
  2 | 5555
  3 | 00000000
  3 | 5
  4 | 0
  4 |
  5 |
  5 |
  6 | 00
  6 |
  7 |
  7 |
  8 |
  8 |
  9 | 0
```

Summary

When reporting, need to consider

- type of variable
- different types of plots - reveal different characteristics of data
- shape of distribution: centre & spread
- whether there are outliers: centre & spread
- patterns