Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part B: One quantitative and one qualitative variable

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For two qualitative variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For two qualitative variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

✓ For one quantitative and one qualitative variable:

- side-by-side box plots
- back-to-back stem & leaf plots

Bivariate Data: Two Variables

Different tables / plots for different data types . . .

For two qualitative variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

For one quantitative and one qualitative variable:

- side-by-side box plots
- back-to-back stem & leaf plots

For two quantitative variable/s:

- scatterplots
- line plots (against time)

Comparing Batches: One quantitative and one qualitative variable

Question: Is there a **difference** between two or more batches of data?

- One quantitative variable
- Two batches (male & female)
- Or many batches (eg brands)

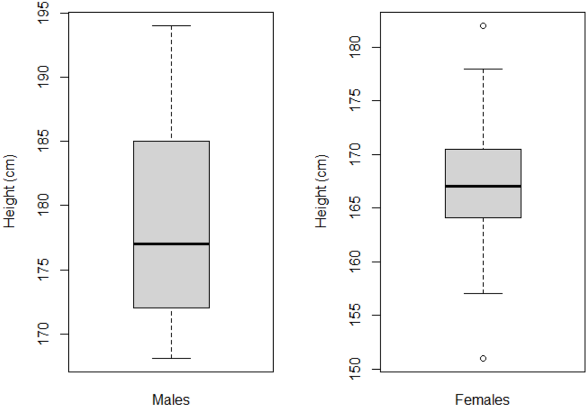
The aim is to turn data into meaningful information AND to **communicate it effectively**

- Plots should be on the same scale
- Do NOT use two separate plots
- Different plots will show different aspects of the data

Later we examine hypothesis tests - eg. are the population means significantly different?

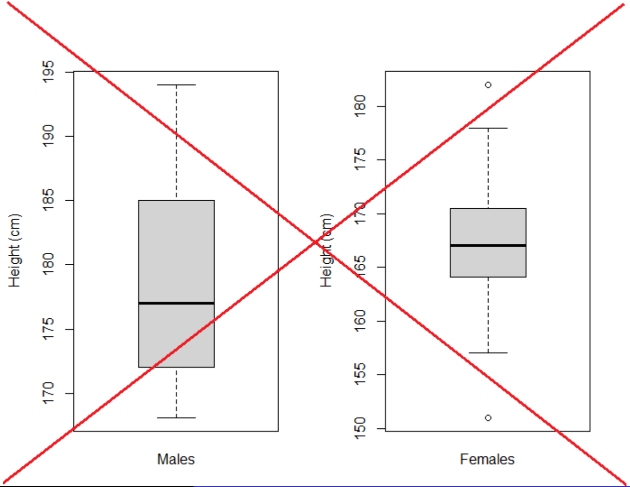
Comparing batches

Example: Measured Heights (cm) for 46 M and 19 F



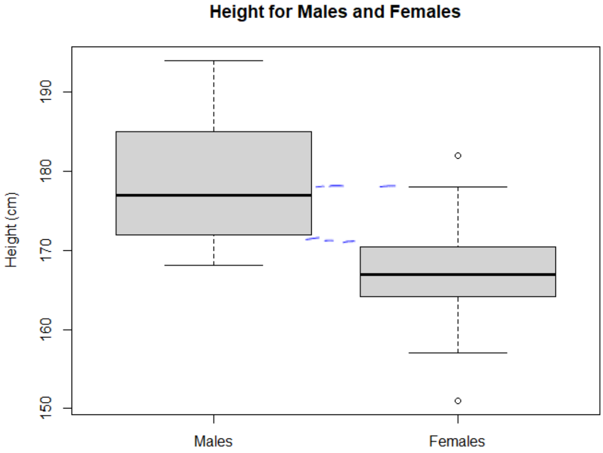
Comparing batches

Example: Measured Heights (cm) for 46 M and 19 F



To compare: Use one plot, one set of axes

Example: Measured Heights (cm) for 46 M and 19 F



Comparing batches: Communication

Key descriptors involve comparison

Based on comparative techniques make **comparative statements**

- Greater than . . .
- Similar to . . .
- Less than . . .

Comparing batches: Communication

Key descriptors involve comparison

Based on comparative techniques make **comparative statements**

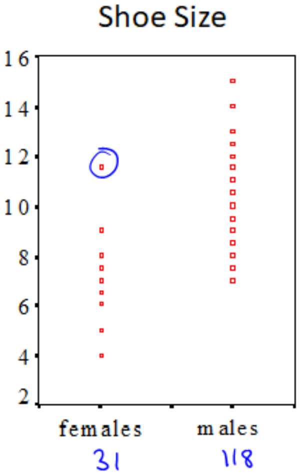
- Greater than . . .
- Similar to . . .
- Less than . . .

For all key features

- Contexts
- Shape of distribution
- Outliers/Extremes
- Centre
- Spread
- Patterns

Comparison - Dot plots

Comparison of male and female shoe size



We can easily see:

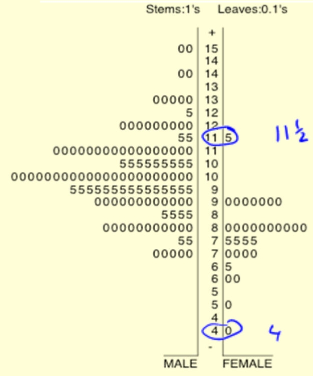
- spread
- possible outliers

What can't we see?

- shape of the distribution
- centre of data
- density of dots (ie how many people with one shoe size) as dots overlayed

Comparison: Back-to-back stem-and-leaf plots

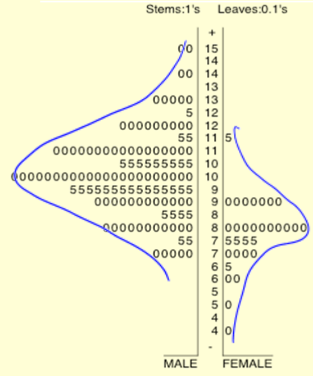
What does the data reveal?



```
                    Stems:1's    Leaves:0.1's
                       +
                    00|15|
                      |14|
                    00|14|
                      |13|
                 00000|13|
                     5|12|
             000000000|12|
                    55|11|5          11½
     00000000000000000|11|
             555555555|10|
0000000000000000000000|10|
       555555555555555| 9|
          000000000000| 9|0000000
                  5555| 8|
           00000000000| 8|0000000000
                    55| 7|5555
                 00000| 7|0000
                      | 6|5
                      | 6|00
                      | 5|
                      | 5|0
                      | 4|
                      | 4|0            4
                       -
                  MALE    FEMALE
```

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?

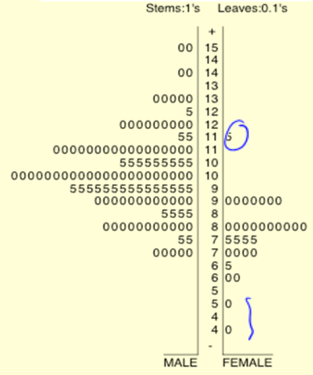


```
              Stems:1's   Leaves:0.1's
                        |  +  |
                      00| 15  |
                        | 14  |
                      00| 14  |
                        | 13  |
                   00000| 13  |
                       5| 12  |
               000000000| 12  |
                      55| 11  |5
       00000000000000000| 11  |
               555555555| 10  |
  0000000000000000000000| 10  |
         555555555555555|  9  |
            000000000000|  9  |0000000
                    5555|  8  |
             00000000000|  8  |0000000000
                      55|  7  |5555
                   00000|  7  |0000
                        |  6  |5
                        |  6  |00
                        |  5  |
                        |  5  |0
                        |  4  |
                        |  4  |0
                        |  -  |
                    MALE       FEMALE
```

1. Distribution shape:
   - Male: bell-shaped
   - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?

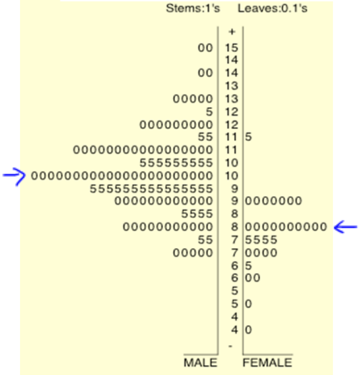


```
                        Stems:1's    Leaves:0.1's
                                  +
                            00 | 15 |
                               | 14 |
                            00 | 14 |
                               | 13 |
                         00000 | 13 |
                             5 | 12 |
                     000000000 | 12 |
                            55 | 11 |5
             00000000000000000 | 11 |
                     555555555 | 10 |
        0000000000000000000000 | 10 |
               555555555555555 |  9 |
                  000000000000 |  9 |0000000
                          5555 |  8 |
                   00000000000 |  8 |0000000000
                            55 |  7 |5555
                         00000 |  7 |0000
                               |  6 |5
                               |  6 |00
                               |  5 |
                               |  5 |0
                               |  4 |
                               |  4 |0
                                  -
                          MALE        FEMALE
```

1. Distribution shape:
   - Male: bell-shaped
   - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
2. Outliers - Possible - but not shown here

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?



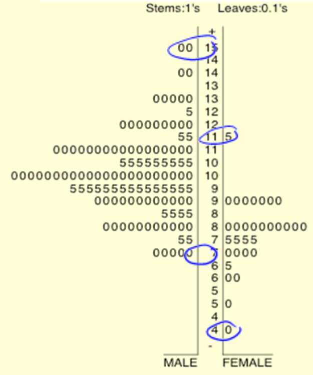
```
                    Stems:1's    Leaves:0.1's

                        | + |
                      00|15 |
                        |14 |
                      00|14 |
                        |13 |
                   00000|13 |
                       5|12 |
               000000000|12 |
                      55|11 |5
       00000000000000000|11 |
               555555555|10 |
  0000000000000000000000|10 |
         555555555555555| 9 |
            000000000000| 9 |0000000
                    5555| 8 |
             00000000000| 8 |0000000000
                      55| 7 |5555
                   00000| 7 |0000
                        | 6 |5
                        | 6 |00
                        | 5 |
                        | 5 |0
                        | 4 |
                        | 4 |0
                        | - |
                    MALE     FEMALE
```

1. Distribution shape:
   - Male: bell-shaped
   - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
2. Outliers - Possible - but not shown here
3. Centre - can only see mode
   - mode for males is 10 &
   - is higher than mode for females (8)

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?

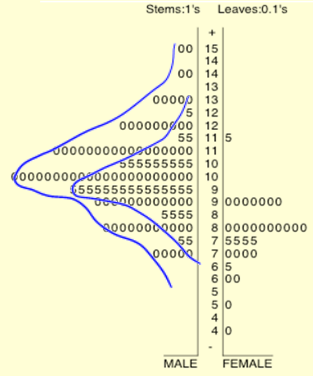


```
                     Stems:1's   Leaves:0.1's
                            +
                        00 |15|
                           |14|
                        00 |14|
                           |13|
                     00000 |13|
                         5 |12|
                 000000000 |12|
                        55 |11|5
        000000000000000000 |11|
                 555555555 |10|
   00000000000000000000000 |10|
           555555555555555 | 9|
              000000000000 | 9|0000000
                      5555 | 8|
               00000000000 | 8|0000000000
                        55 | 7|5555
                     00000 | 7|0000
                           | 6|5
                           | 6|00
                           | 5|
                           | 5|0
                           | 4|
                           | 4|0
                            -
                      MALE    FEMALE
```

1. Distribution shape:
   - Male: bell-shaped
   - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
2. Outliers - Possible - but not shown here
3. Centre - can only see mode
   - mode for males is 10 &
   - is higher than mode for females (8)
4. Spread can determine range
   - for males is 7-15 and females 4-11.5
   - so range is a little wider for males 8 than females 7.5

Comparison: Back-to-back stem-and-leaf plots

What does the data reveal?

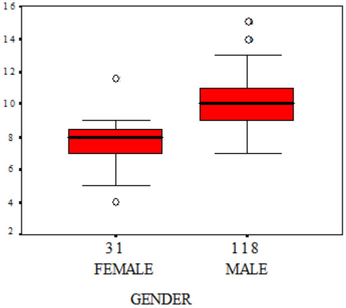


```
                     Stems:1's    Leaves:0.1's
                                + 
                             00|15|
                               |14|
                             00|14|
                               |13|
                          00000|13|
                              5|12|
                     0000000000|12|
                             55|11|5
            0000000000000000000|11|
                     5555555555|10|
      0000000000000000000000000|10|
              55555555555555555| 9|
                  0000000000000| 9|0000000
                           5555| 8|
                   000000000000| 8|0000000000
                             55| 7|5555
                          00000| 7|0000
                               | 6|5
                               | 6|00
                               | 5|
                               | 5|0
                               | 4|
                               | 4|0
                                - 
                           MALE    FEMALE
```

1. Distribution shape:
   - Male: bell-shaped
   - Female: skewed with a longer tail of small shoe sizes (negative skew, skewed to left)
2. Outliers - Possible - but not shown here
3. Centre - can only see mode
   - mode for males is 10 &
   - is higher than mode for females (8)
4. Spread can determine range
   - for males is 7-15 and females 4-11.5
   - so range is a little wider for males 8 than females 7.5
5. Pattern
   - M: bell within a bell; F: not so clear
   - M & F: fewer half sizes

Comparison: Box Plots

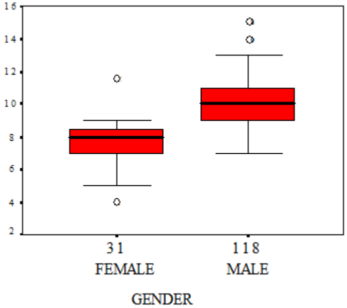
What does the data reveal?



1. Context: shoe size
    - 118 Males & 31 Females

Comparison: Box Plots

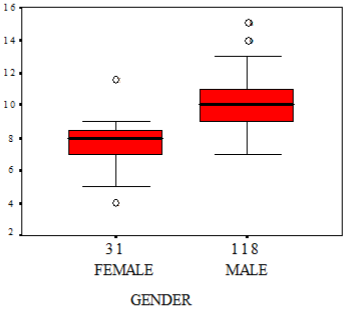
What does the data reveal?



1. Context: shoe size
   - 118 Males & 31 Females

Comparison: Box Plots

What does the data reveal?

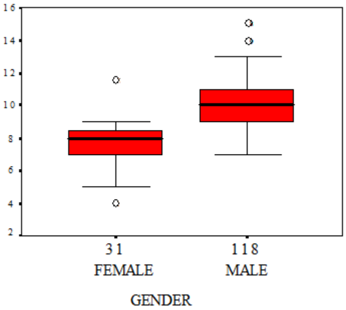


1. Context: shoe size
   - 118 Males & 31 Females

2. Distribution shape:
   - F is more asymmetric than M with relatively shorter tail of upper values

Comparison: Box Plots

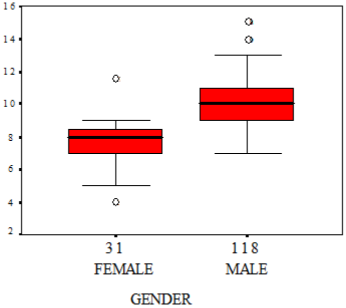
What does the data reveal?



1. Context: shoe size
   - 118 Males & 31 Females
2. Distribution shape:
   - F is more asymmetric than M with relatively shorter tail of upper values
3. Outliers
   - M: two high (sizes 14 & 15)
   - F: a low (size 4) & a high (size 11.5)

Comparison: Box Plots

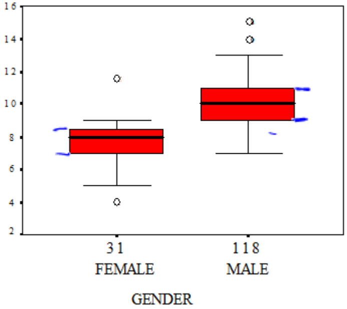
What does the data reveal?



1. Context: shoe size
   - 118 Males & 31 Females
2. Distribution shape:
   - F is more asymmetric than M with relatively shorter tail of upper values
3. Outliers
   - M: two high (sizes 14 & 15)
   - F: a low (size 4) & a high (size 11.5)
4. Centre - can only see median
   - median for M is size 10 &
   - is higher than for F (size 8)

Comparison: Box Plots

What does the data reveal?

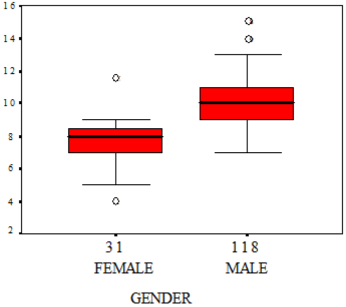


1. Context: shoe size
   - 118 Males & 31 Females

2. Distribution shape:
   - F is more asymmetric than M with relatively shorter tail of upper values
3. Outliers
   - M: two high (sizes 14 & 15)
   - F: a low (size 4) & a high (size 11.5)
4. Centre - can only see median
   - median for M is size 10 &
   - is higher than for F (size 8)
5. Spread: can determine IQR and range
   - IQR for M is 11-9=2 and F 8.5-7=1.5
   - IQR is slightly greater for M than F
   - range is a little wider for M 8 than F 7.5

Comparison: Box Plots

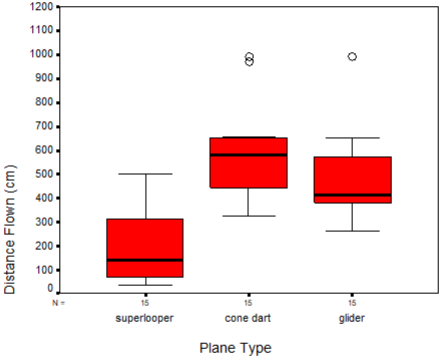
What does the data reveal?



1. Context: shoe size
   - 118 Males & 31 Females
2. Distribution shape:
   - F is more asymmetric than M with relatively shorter tail of upper values
3. Outliers
   - M: two high (sizes 14 & 15)
   - F: a low (size 4) & a high (size 11.5)
4. Centre - can only see median
   - median for M is size 10 &
   - is higher than for F (size 8)
5. Spread: can determine IQR and range
   - IQR for M is 11-9=2 and F 8.5-7=1.5
   - IQR is slightly greater for M than F
   - range is a little wider for M 8 than F 7.5
6. Pattern - cannot be seen in this plot

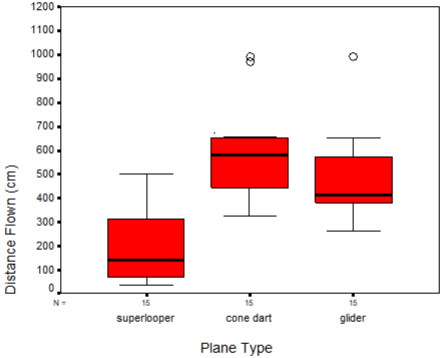
Utility: Boxplots versus Stem-and-leaf Plots

- **Boxplots**
  - are especially useful for comparing ≥ 2 samples or batches.
  - show the 5-number summary and outliers
  - but not the individual values.



Utility: Boxplots versus Stem-and-leaf Plots

- **Boxplots**
  - are especially useful for comparing ≥ 2 ~~samples or~~ batches.
  - show the 5-number summary and outliers
  - but not the individual values.



- **Stem-and-leaf plots**
  - show individual values, and
  - give a better picture of the shape of the spread,
  - but their detail makes them unsuitable for comparing more than two groups (back-to-back)
  - not suitable when a large no. of observations