MONASH University

Information Technology

FIT1006
Business Information Analysis

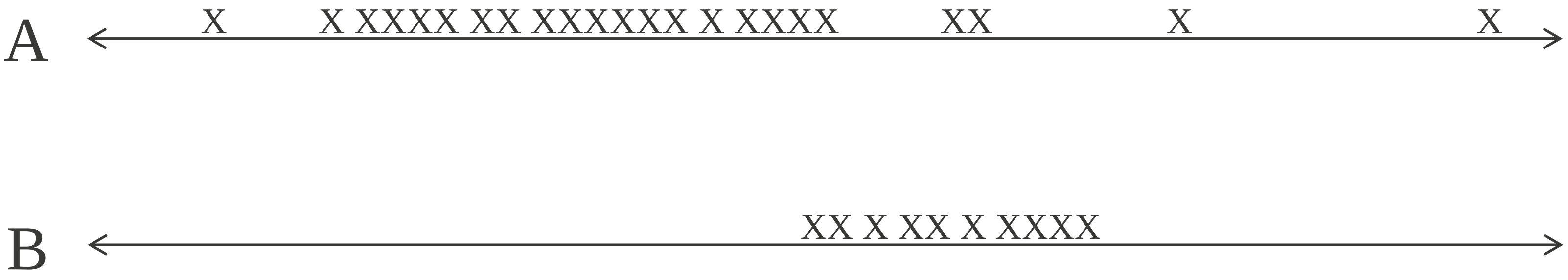
Lecture 4
Descriptive Statistics

Topics covered:

- Measures of Centre
  - Mean, Median, Mode, Trimmed Mean
  - Robust Statistics
- Measures of Spread
  - Variance and Standard Deviation
  - Quartiles and Percentiles, Quick Quartiles
  - Boxplots

Learning Objectives

- This lecture is about how we characterise a data set using some summary statistics.
- A typical problem that could be answered with the techniques covered today is: describe the differences between the two data sets A and B below?



Motivating problem…

- A grocery store wants you to analyse the amount spent by their customers. They also think there might be different types of customers. They have given you the sales history of 10 randomly sampled customers.
- Data is from the Kaggle ‘Dunnhumby’s Shopper Challenge’ which recorded the amount spent and date of the transaction at a supermarket in the US over one year.
  - See: http://www.kaggle.com/c/dunnhumbychallenge
- I have resampled the original data, using approx 20% of the original observations.
- We will use the data for 10 groups of shoppers.

Sample Data – Stem and leaf plot

53, 16, 66, 10, 77, 25,

17, 44, 37, 25, 24, 62,

3, 50, 16, 18, 5, 31, 29.

- Output from SYSTAT on the RHS

- What can you say about that customer?

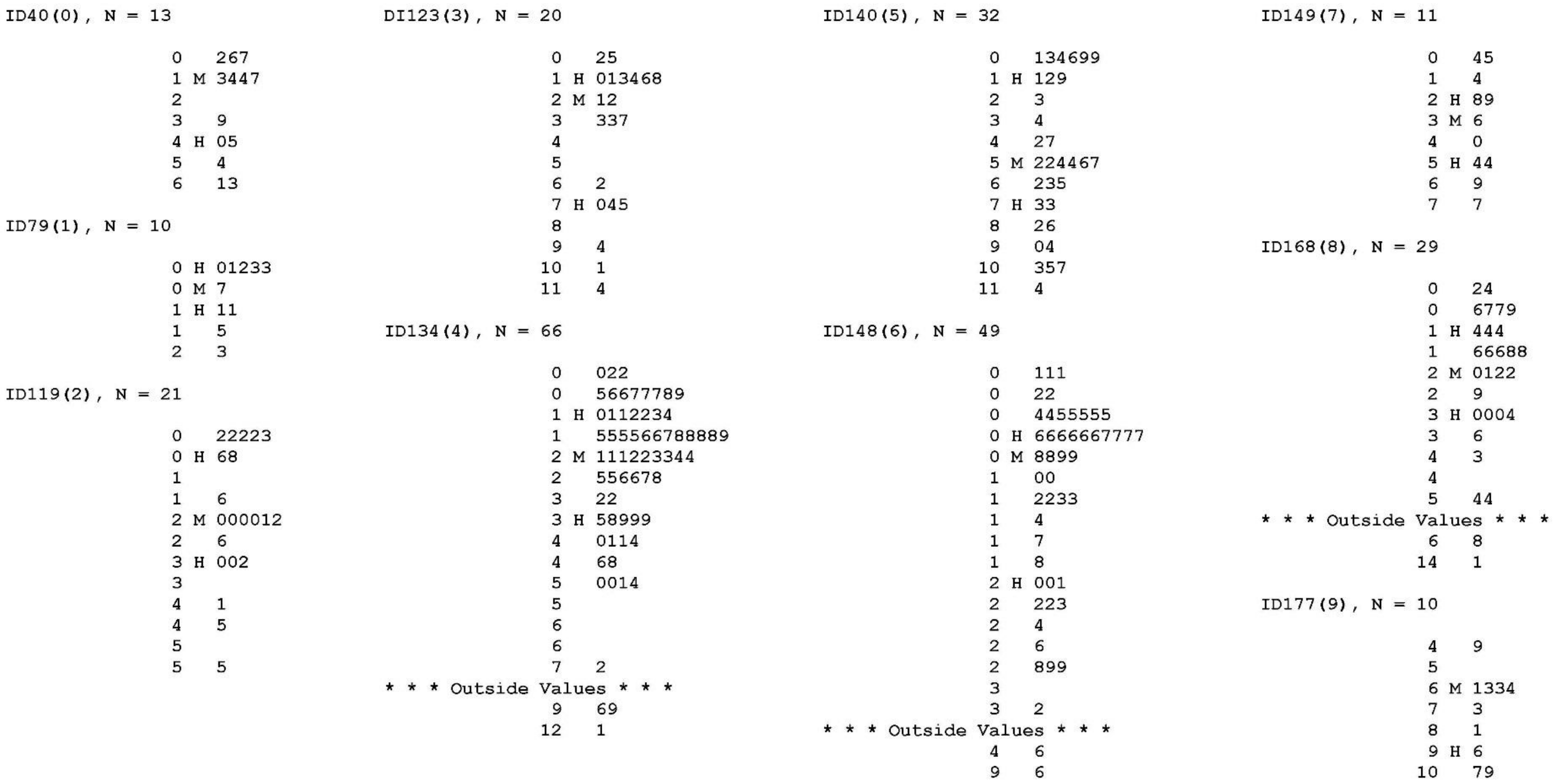
```
Stem and Leaf Plot of Variable:
RSPEND, N = 19


Minimum      :  3.000
Lower Hinge  : 16.500
Median       : 25.000
Upper Hinge  : 47.000
Maximum      : 77.000


               0   35
               1 H 06678
               2 M 4559
               3   17
               4 H 4
               5   03
               6   26
               7   7
```

Motivating Problem

- Working in groups of 3, each group will draw a stem and leaf plot for one of the 10 customers. Your customer is based on the first letter of your last name indicated in the worksheet.

- Describe the shape of the distribution of data.

Motivating Problem – SYSTAT



```
ID40(0), N = 13

              0   267
              1 M 3447
              2
              3   9
              4 H 05
              5   4
              6   13

ID79(1), N = 10

              0 H 01233
              0 M 7
              1 H 11
              1   5
              2   3

ID119(2), N = 21

              0   22223
              0 H 68
              1
              1   6
              2 M 000012
              2   6
              3 H 002
              3
              4   1
              4   5
              5
              5   5

DI123(3), N = 20

              0   25
              1 H 013468
              2 M 12
              3   337
              4
              5
              6   2
              7 H 045
              8
              9   4
             10   1
             11   4

ID134(4), N = 66

              0   022
              0   56677789
              1 H 0112234
              1   555566788889
              2 M 111223344
              2   556678
              3   22
              3 H 58999
              4   0114
              4   68
              5   0014
              5
              6
              6
              7   2
 * * * Outside Values * * *
              9   69
             12   1

ID140(5), N = 32

              0   134699
              1 H 129
              2   3
              3   4
              4   27
              5 M 224467
              6   235
              7 H 33
              8   26
              9   04
             10   357
             11   4

ID148(6), N = 49

              0   111
              0   22
              0   4455555
              0 H 6666667777
              0 M 8899
              1   00
              1   2233
              1   4
              1   7
              1   8
              2 H 001
              2   223
              2   4
              2   6
              2   899
              3
              3   2
 * * * Outside Values * * *
              4   6
              9   6

ID149(7), N = 11

              0   45
              1   4
              2 H 89
              3 M 6
              4   0
              5 H 44
              6   9
              7   7

ID168(8), N = 29

              0   24
              0   6779
              1 H 444
              1   66688
              2 M 0122
              2   9
              3 H 0004
              3   6
              4   3
              4
              5   44
 * * * Outside Values * * *
              6   8
             14   1

ID177(9), N = 10

              4   9
              5
              6 M 1334
              7   3
              8   1
              9 H 6
             10   79
```

Describe the different types of customers…

Visualising Data

- Why do we want to ‘see’ our data?
  - Visual inspection is the fastest way to get an overview of data.
  - Visual inspection enables a description of the distribution of the data to be made.
  - The distribution of data determines which statistics are appropriate.
  - To make comparisons between data from different groups.

Sample Data

- The data below is for Customer #208

53, 16, 66, 10, 77, 25, 17, 44, 37, 25, 24, 62, 3, 50, 16, 18, 5, 31, 29

- What can you say about that customer?

- Using the Stem and leaf plot…

Question 1

For the sample data

Which is most likely true?

____

A. Mode ≈ 16

B. Mode ≈ 25

[✓ C. Median ≈ 25]

D. Mean ≈ 25

E. None of the above.

```
Stem and Leaf Plot of
Variable:
RSPEND, N = 19

          0   35
          1 H 06678
          2 M 4559
          3   17
          4 H 4
          5   03
          6   26
          7   7
```

Question 2

For customer #148

Which is most likely true?

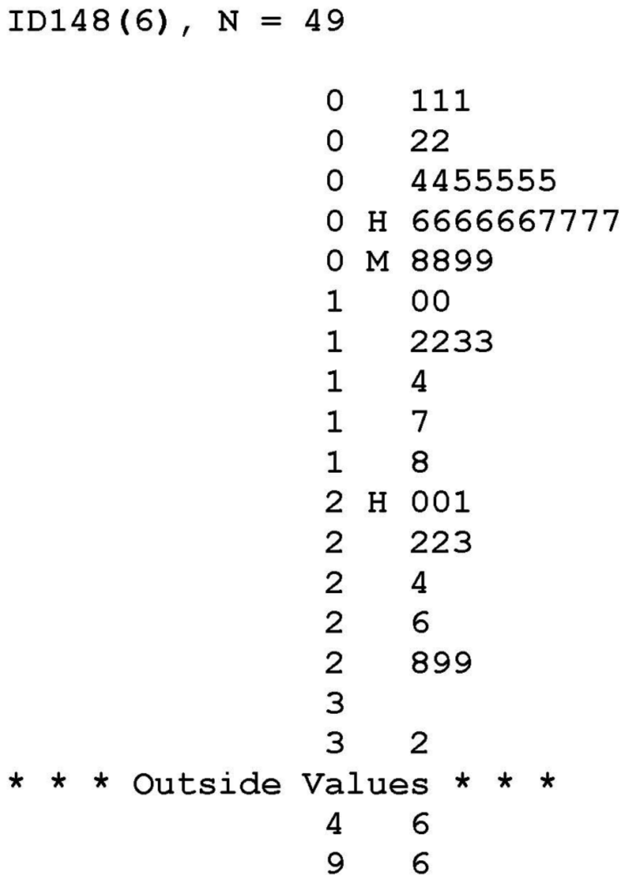
A. Mode < Mean < Median

B. Median < Mode < Mean

[✓ C. Mode < Median < Mean]

D. Mean < Mode < Median

E. Something else!



```
ID148(6), N = 49

               0   111
               0   22
               0   4455555
               0 H 6666667777
               0 M 8899
               1   00
               1   2233
               1   4
               1   7
               1   8
               2 H 001
               2   223
               2   4
               2   6
               2   899
               3
               3   2
* * * Outside Values * * *
               4   6
               9   6
```

Measures of centre

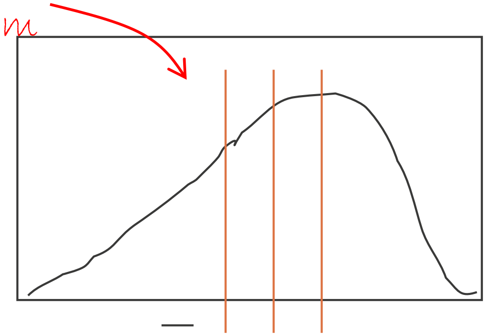
- The mean or average is the most well known. It is the sum of data divided by the number of observations.
- The median is the central observation in an ordered data set.
- The mode is the most frequently occurring observation.
- The *a%* trimmed mean provides some compromise between the mean and the median. The highest and lowest *a%* of values are trimmed from the data. The mean of the remainder is then calculated.

Mean *v* Median *v* Mode

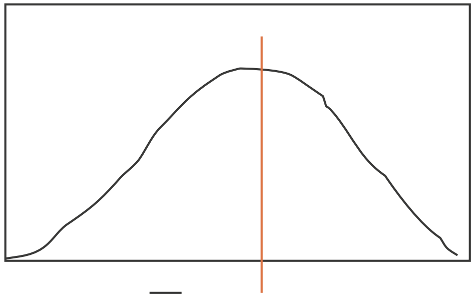
- The mean and median provide the most usual measures of centre for quantitative data.
- If the data is symmetrically distributed then either the mean or median are acceptable and the mean is usually preferred.
- If the data is skewed or contains exceptionally high or low values then the median is usually preferred.

Negative (left skewed) histogram

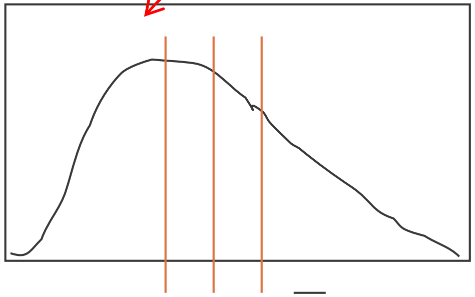
Positive (right skewed) histogram



$\overline{X}$<Me<Mo



$\overline{X}$ = Me = Mo



Mo<Me<$\overline{X}$

Question 3

For the BUS1234 data:

Approx 95% of obs are between:

Remember that 95% of data lies within ± 2 standard deviations of the mean

A. 17 and 91

B. 26 and 76

C. 48 and 63

[✓ D. 24 and 85]

E. None of the above.

```
N of Cases          ¦       50
Minimum             ¦   17.000
Maximum             ¦   91.000
Arithmetic Mean     ¦   54.640
Standard Deviation  ¦   15.147

                  1    7
                  2    4
* * * Outside Values * * *
                  2    67
                  3
                  3    77788
                  4    22
                  4  H 7889
                  5  M 1122233334
                  5    55778889
                  6  H 01233
                  6    5789
                  7    133
                  7    566
                  8    1
* * * Outside Values * * *
                  9    1
```

Variance & Standard Deviation - *s*

- The standard deviation is perhaps the most commonly used measure of the spread of data.
- The variance of a sample is calculated as the average of the squared deviations from the mean, adjusted for the fact that we are considering a sample.
- The standard deviation is the square root of the variance.
- Two standard formulas are used in practice:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n - 1}}$$

Question 4

For the BUS1234 data:

Approx 50% of obs are between:

____

A. 17 and 91

B. 26 and 76

C. 48 and 63

D. 24 and 85

E. None of the above.

```
N of Cases         ¦       50
Minimum            ¦   17.000
Maximum            ¦   91.000
Arithmetic Mean    ¦   54.640
Standard Deviation ¦   15.147

               1   7
               2   4
* * * Outside Values * * *
               2   67
               3
               3   77788
               4   22
               4 H 7889
               5 M 1122233334
               5   55778889
               6 H 01233
               6   5789
               7   133
               7   566
               8   1
* * * Outside Values * * *
               9   1
```

Question 5

For the BUS1234 data:

Approx 50% of obs are between:

____

A. 17 and 91

B. 26 and 76

[✓ C. 48 and 63]

D. 24 and 85

E. None of the above.

```
Minimum             ¦ 17.000
Lower Hinge         ¦ 48.000
Median              ¦ 54.500
Upper Hinge         ¦ 63.000
Maximum             ¦ 91.000

               1    7
               2    4
* * * Outside Values * * *
               2    67
               3
               3    77788
               4    22
               4 H  7889
               5 M  1122233334
               5    55778889
               6 H  01233
               6    5789
               7    133
               7    566
               8    1
* * * Outside Values * * *
               9    1
```

Quartiles

- Ranked data can be divided into four quartiles.
- 25% of the data is less than the first - or lower quartile,
- 50% lower than the second quartile - or median,
- 75% of data is less than the third - or upper quartile.
- In SYSTAT, the upper and lower quartiles are referred to as 'Hinges'.

For a data set $\mathrm{x}_1, x_2 \cdots x_n$ arranged in ascending order

We wish to find the Qth quartile, Q=1, 2 or 3

$$q = (n+1)\frac{Q}{4} \text{ and } \mathrm{Q} = \mathrm{x}_\mathrm{q} \quad \text{(the required value of x)}$$

When q is non-integer we we calculate

$$Q = x_q + r(x_{q+1} - x_q) \text{ where r is the fractional part of q}$$

Sample Data – Stem and leaf plot

53, 16, 66, 10, 77, 25,

17, 44, 37, 25, 24, 62,

3, 50, 16, 18, 5, 31, 29.

```
Stem and Leaf Plot of Variable:
RSPEND, N = 19


Minimum       :   3.000
Lower Hinge   :  16.500
Median        :  25.000
Upper Hinge   :  47.000
Maximum       :  77.000


     0      35
     1  H   06678
     2  M   4559
     3      17
     4  H   4
     5      03
     6      26
     7      7
```

- Output from SYSTAT on the RHS
- What can you say about that customer?

Median : $\frac{19+1}{2}$ = 10th value

~~LH : $\frac{10+1}{2}$ = 5.5 value~~

$\frac{16+17}{2}$ = 16.5

LQ = $\frac{19+1}{4}$ = 5th value → 16

~~UH = $\frac{44+50}{2}$ = 47~~

UQ = $\frac{3(19+1)}{4}$ = 15th value → 50

Percentiles

- In the same way that quartiles divide the data into four, we can use percentiles to divide our data into one hundredths.
- We can calculate the *Cth* percentile and say that *C*% of data lies below this value. The median is the 50th percentile.

For a data set $\mathrm{x}_1, x_2 \cdots x_n$ arranged in ascending order

We wish to find the Cth percentile, C = 0, 1, 2 ... 100

$$p = (n+1)\frac{C}{100} \text{ and } \mathrm{P_C} = \mathrm{x_p} \quad \text{(the required value of x)}$$

When p is non-integer we we calculate

$$P_C = x_p + r(x_{p+1} - x_p) \text{ where r is the fractional part of p}$$

- (Note: see textbook P. 159)

MONASH University

Measures of spread

- The variance is the average of the squared deviations adjusted for estimation of the mean.
- The standard deviation is the most well known. It is the square root of the variance.
- The range is largest – smallest observation.
- The interquartile range is Q3 – Q1 it contains the middle 50% of observations.

Sample Data

- The data below is for Customer #208

53, 16, 66, 10, 77, 25, 17, 44, 37, 25, 24, 62, 3,
50, 16, 18, 5, 31, 29

- Q1 = ? Lower quartile

$$q=(n+1)\frac{Q}{4}$$

$q = \frac{19+1}{4}$
= 5th value
Q_1 → 16

$$Q = x_q + r(x_{q+1} - x_q)$$

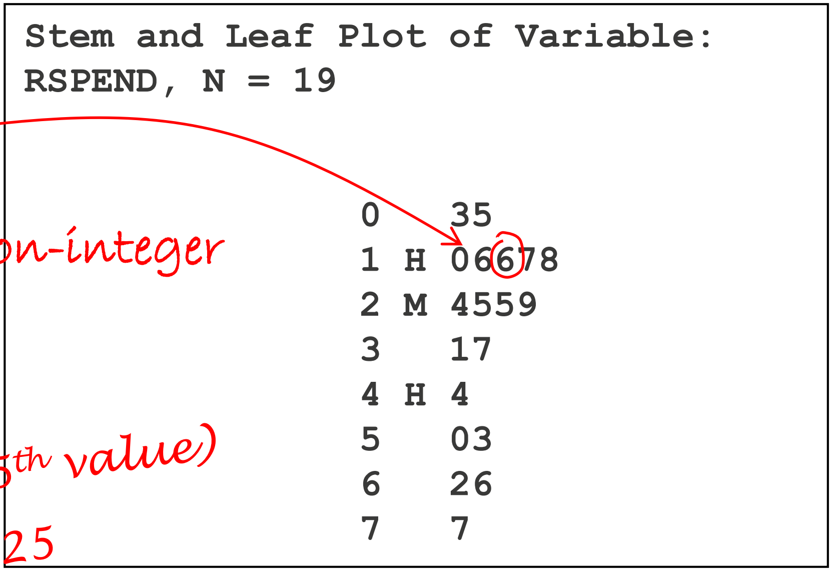
When q is non-integer

If N = 20:

$q = \frac{20+1}{4}$ = 5.25th value

Q_1 → 5th value + 0.25 (6th value - 5th value)

Q_1 → 16+ 0.25 (17 - 16) = 16.25



```
Stem and Leaf Plot of Variable:
RSPEND, N = 19


                  0     35
                  1  H  06678
                  2  M  4559
                  3     17
                  4  H  4
                  5     03
                  6     26
                  7     7
```

Sample Data

```
Stem and Leaf Plot of Variable:
RSPEND, N = 19


Minimum      :   3.000
Lower Hinge  :  16.500
Median       :  25.000
Upper Hinge  :  47.000
Maximum      :  77.000


                 0    35
                 1 H  06678
                 2 M  4559
                 3    17
                 4 H  4
                 5    03
                 6    26
                 7    7
```

Remember from slide 21

$LH = \frac{10+1}{2}$

$=5.5$ value

$\frac{16+17}{2}$

$=16.5$

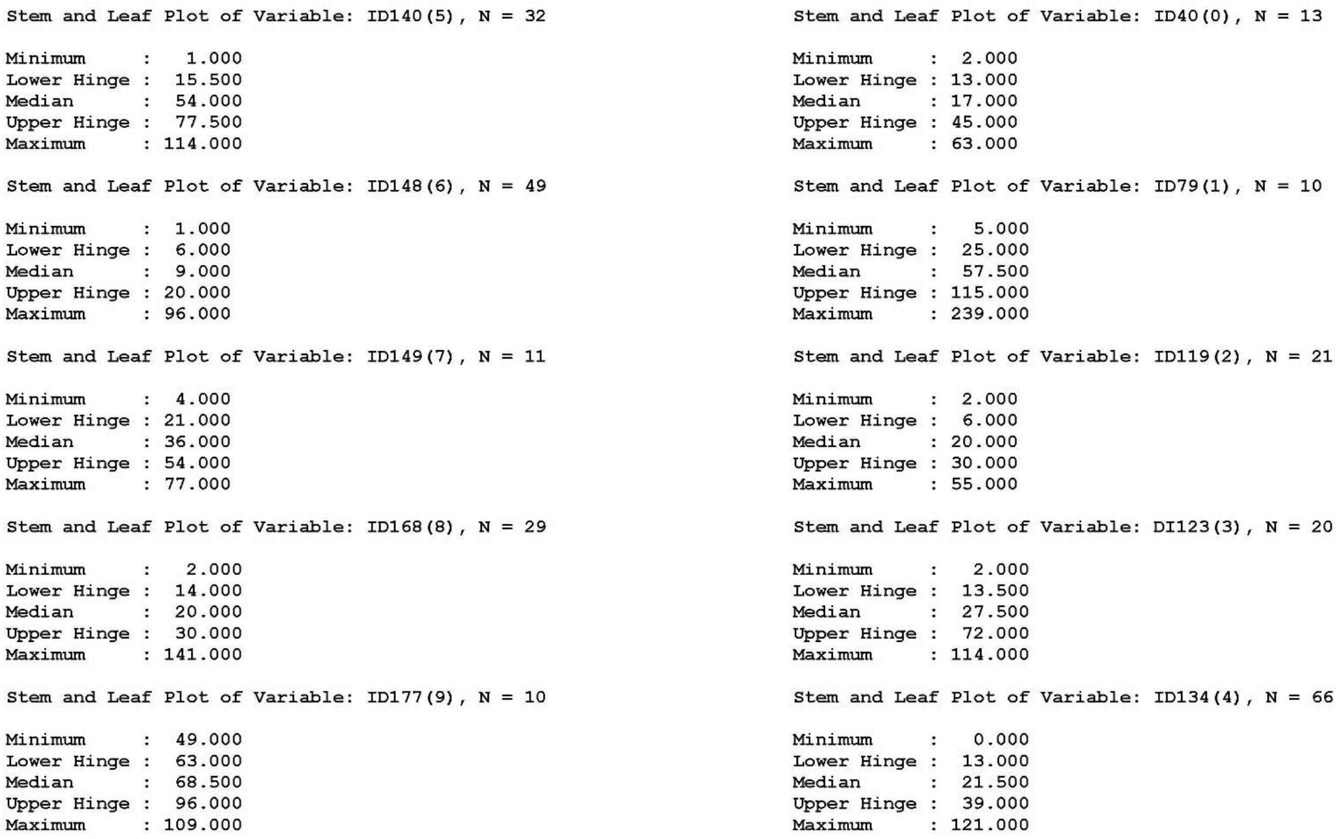
Note that SYSTAT results for Q1 and Q3 are slightly different to hand calculations. SYSTAT interpolates values from smoothed distribution.

Motivating Problem

Homework for you!

- ~~Working in groups of 3, each group will draw a stem and leaf plot for one of the 10 customers. Your customer is based on your last name as indicated on your worksheet.~~
- Using today's data sheet… calculate the quartiles.
- (Try the 5th and 95th percentiles if you're keen.)

Motivating Problem - SYSTAT

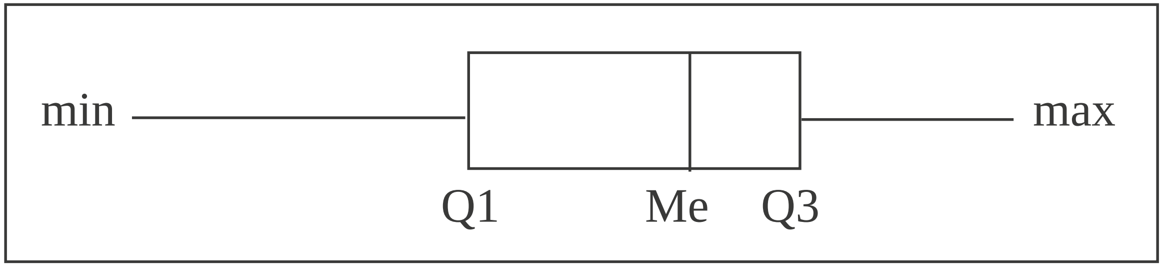


```
Stem and Leaf Plot of Variable: ID140(5), N = 32

Minimum     :   1.000
Lower Hinge :  15.500
Median      :  54.000
Upper Hinge :  77.500
Maximum     : 114.000

Stem and Leaf Plot of Variable: ID148(6), N = 49

Minimum     :  1.000
Lower Hinge :  6.000
Median      :  9.000
Upper Hinge : 20.000
Maximum     : 96.000

Stem and Leaf Plot of Variable: ID149(7), N = 11

Minimum     :  4.000
Lower Hinge : 21.000
Median      : 36.000
Upper Hinge : 54.000
Maximum     : 77.000

Stem and Leaf Plot of Variable: ID168(8), N = 29

Minimum     :   2.000
Lower Hinge :  14.000
Median      :  20.000
Upper Hinge :  30.000
Maximum     : 141.000

Stem and Leaf Plot of Variable: ID177(9), N = 10

Minimum     :  49.000
Lower Hinge :  63.000
Median      :  68.500
Upper Hinge :  96.000
Maximum     : 109.000
```

```
Stem and Leaf Plot of Variable: ID40(0), N = 13

Minimum     :  2.000
Lower Hinge : 13.000
Median      : 17.000
Upper Hinge : 45.000
Maximum     : 63.000

Stem and Leaf Plot of Variable: ID79(1), N = 10

Minimum     :   5.000
Lower Hinge :  25.000
Median      :  57.500
Upper Hinge : 115.000
Maximum     : 239.000

Stem and Leaf Plot of Variable: ID119(2), N = 21

Minimum     :  2.000
Lower Hinge :  6.000
Median      : 20.000
Upper Hinge : 30.000
Maximum     : 55.000

Stem and Leaf Plot of Variable: DI123(3), N = 20

Minimum     :   2.000
Lower Hinge :  13.500
Median      :  27.500
Upper Hinge :  72.000
Maximum     : 114.000

Stem and Leaf Plot of Variable: ID134(4), N = 66

Minimum     :   0.000
Lower Hinge :  13.000
Median      :  21.500
Upper Hinge :  39.000
Maximum     : 121.000
```

Boxplots 1

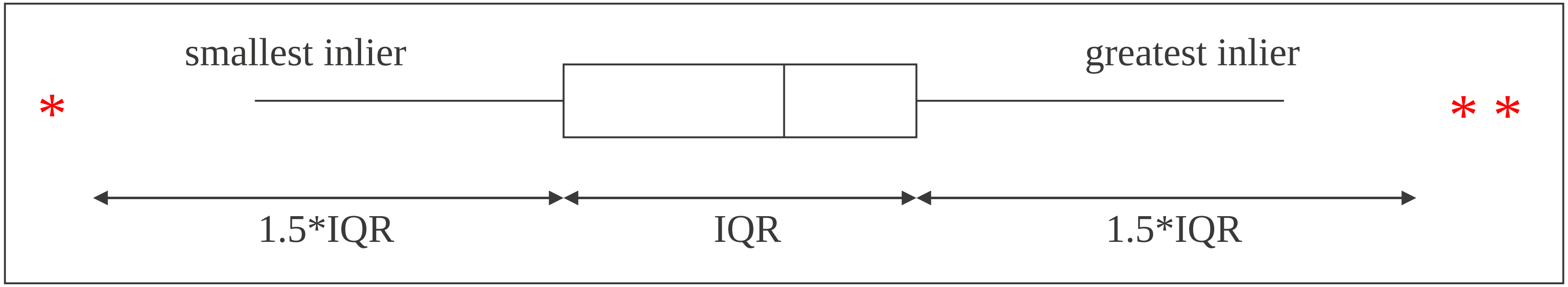
- A boxplot, otherwise known as a box and whisker diagram is, in its simplest form, a five point data summary showing the minimum, maximum, lower quartile, upper quartile and median.



- Boxplots are the most useful tools for comparing data sets, and can be drawn horizontally or vertically.

Boxplots 2

- An alternative form of the boxplot is to extend the whiskers of the boxplot to include only the inlying values. These can be thought of as data that falls within the main central cluster (roughly speaking +/- 2 standard deviations).
- We know that IQR = Q3 - Q1.
- Inliers are all the values greater than Q1 - 1.5*IQR and less than Q3 + 1.5*IQR
- Outliers are values outside this range denoted by ‘*’



Drawing Boxplot

- Using data for Customer #208:

  53, 16, 66, 10, 77, 25, 17, 44, 37, 25, 24, 62, 3, 50, 16, 18, 5, 31, 29

Rearranging the data:

3, 5, 10, 16, 16, 17, 18, 24, 25, 25, 29, 31, 37, 44, 50, 53, 62, 66, 77

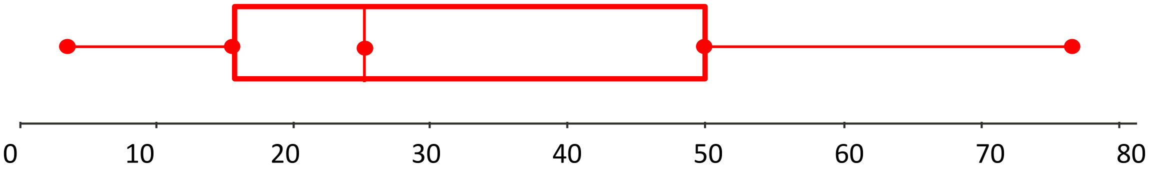
Minimum = 3

Q1 = 16

Median = 25

Q3 = 50

Maximum = 77



IQR = Q3 – Q1 = 50 – 16 = 34

Smallest inlier: Q1 – 1.5 IQR = 16 – 1.5 x 34 = -35

Largest inlier: Q3 + 1.5 IQR = 50 + 1.5 x 34 = 101

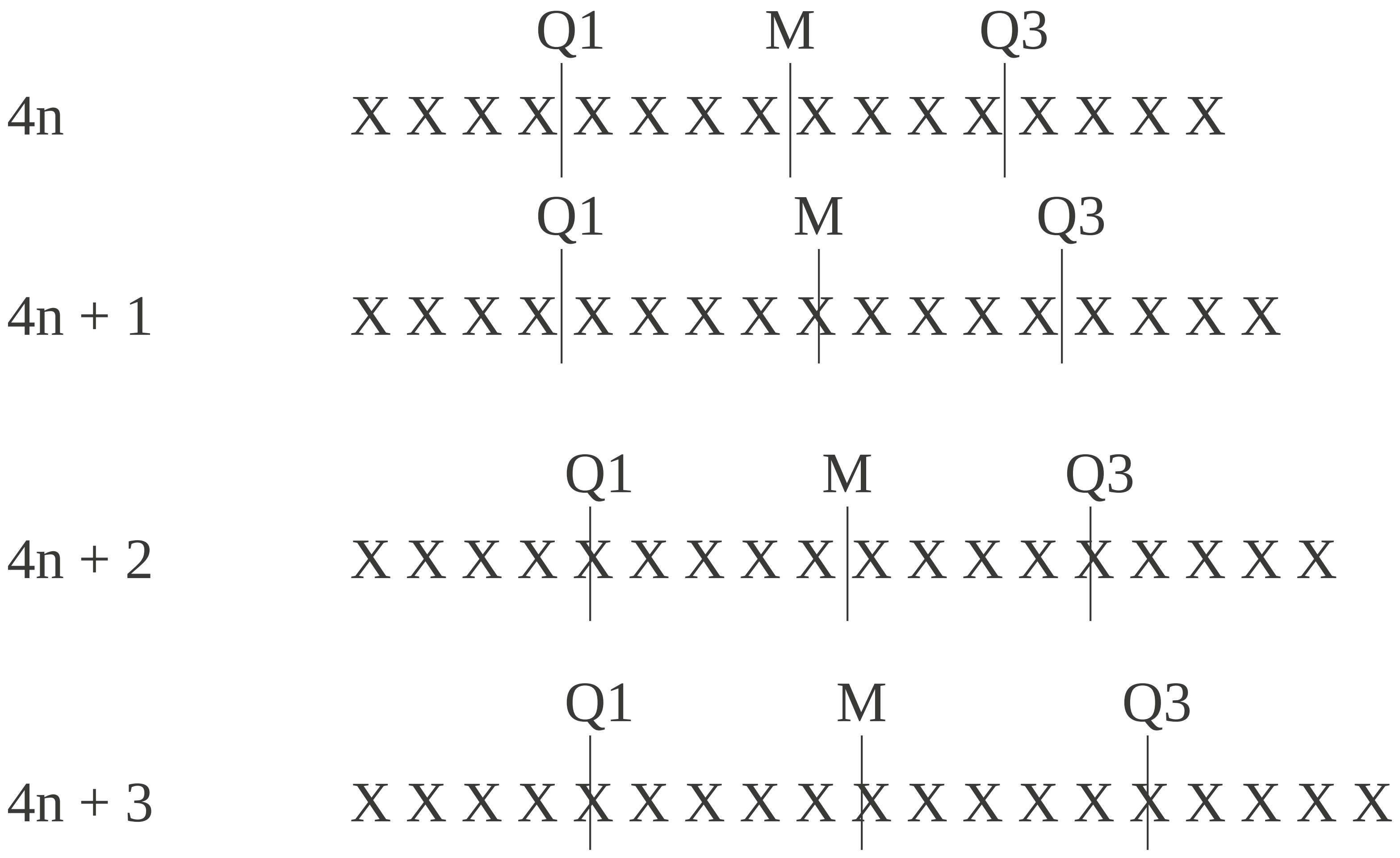
Data fall within this range so no outliers

Key Ideas

- Hand calculations for small data sets.
- Measures of Centre: Mean, Median, Mode, Trimmed Mean. Median *vs* Mean.
- Measures of Spread: Variance and Standard Deviation.
- Quartiles and Percentiles, Quick Quartiles, Boxplots.
- Next week: larger data sets using Excel and SYSTAT.

By the way, this slide is useful to help you work out the quartiles quickly - only for small datasets tho’’

Quick Quartiles



Q1 M Q3

4n X X X X X X X X X X X X X X X X

Q1 M Q3

4n + 1 X X X X X X X X X X X X X X X X X

Q1 M Q3

4n + 2 X X X X X X X X X X X X X X X X X X

Q1 M Q3

4n + 3 X X X X X X X X X X X X X X X X X X X

Reading/Questions (Selvanathan)

- Numerical Descriptive Methods
    - 7th Ed. Sections 5.1 - 5.3.
- Numerical Descriptive Methods
    - 7th Ed. Questions 5.3, 5.4, 5.6, 5.9, 5.12, 5.24, 5.25, 5.26, 5.40, 5.42, 5.62, 5.63.
- Tutorial 2 Questions.