2. Descriptive Statistics

STAT*2060: Statistics for Business Decisions

Nishan Mudalige

Department of Mathematics and Statistics
University of Guelph

Fall 2014

Table of Contents

1 Numerical Measures

2 Graphical Techniques

3 Homework

Table of Contents

1 Numerical Measures

2 Graphical Techniques

3 Homework

Mean

Suppose we have a sample of n observations:

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_n$$

Definition (Sample Mean)

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

- The symbol $\sum$ is called capital sigma.
- It the symbol for summation (i.e. summing, adding).
- The mean is the typical average that we are all used to.

Median and Mode

Definition (Median)

The middle value of ordered data.

- If n is odd

$$Median = \left(\frac{n+1}{2}\right) \ observation$$

- If n is even

$$Median = average \ of \ \left(\frac{n}{2}\right) \ and \ \left(\frac{n}{2}+1\right) \ observation$$

Definition (Mode)

The most frequent observations relative to the rest of the data.

Variance and Standard Deviation

Definition (Sample Variance)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

Definition (Sample Standard Deviation)

$$s = +\sqrt{s^2}$$

- These are measures of spread relative to the mean.
- Standard deviation is used more often when describing data because with variance the units are squared.
- Both these values are non-negative (i.e. ≥ 0)

Variance and Standard Deviation Ctd...

Recall that

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

An easier way to calculate the variance is:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Example 1

Suppose we draw a sample of the following set of measurements:

$$110, \quad 102, \quad 130, \quad 130, \quad 115$$

mean:

Example 1 Ctd...

110, 102, 130, 130, 115

variance:

Example 1 Ctd...

110, 102, 130, 130, 115

variance:

standard deviation:

Example 1 Ctd...

110, 102, 130, 130, 115

median:

mode:

Example 2

Suppose we draw a sample of the following set of measurements:

$$15,\ \ 10,\ \ 17,\ \ 16,\ \ 10,\ 16$$

mean:

Example 2 Ctd...

15, 10, 17, 16, 10, 16

variance:

Example 2 Ctd...

15, 10, 17, 16, 10, 16

variance:

standard deviation:

Example 2 Ctd...

median:

15, 10, 17, 16, 10, 16

mode:

The Importance of Numerical Measures

- We are interested in these numerical measures because they all give information about our data.
- Each measure on its own gives some information.
- Collectively they describe our data with a fair amount of detail.
- We can get an idea about the distribution of our data using these numerical measures and without using all of the raw data on its own.

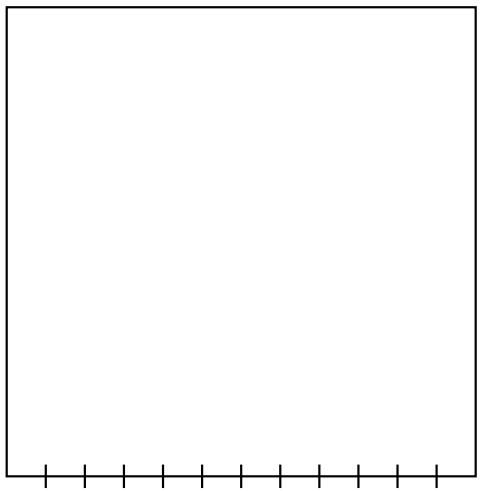
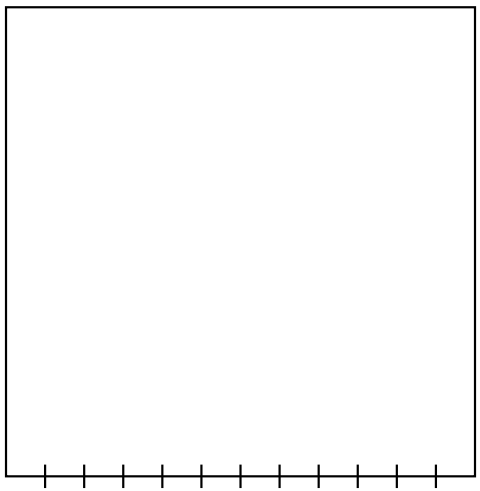
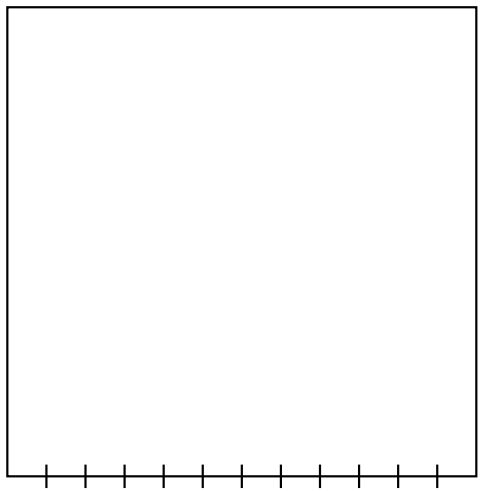
Example 1

Class A

Class B

Example 2

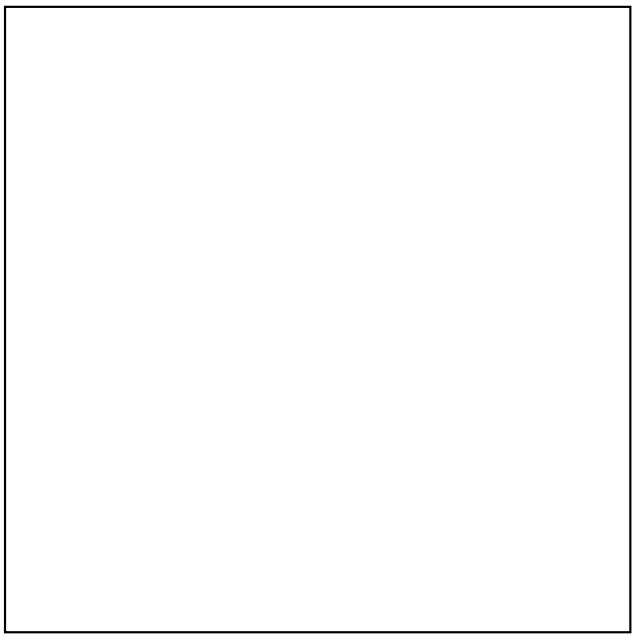
Some background about stock prices

Example 2 Ctd...

Suppose we had 2 stocks

Stock A



Stock B

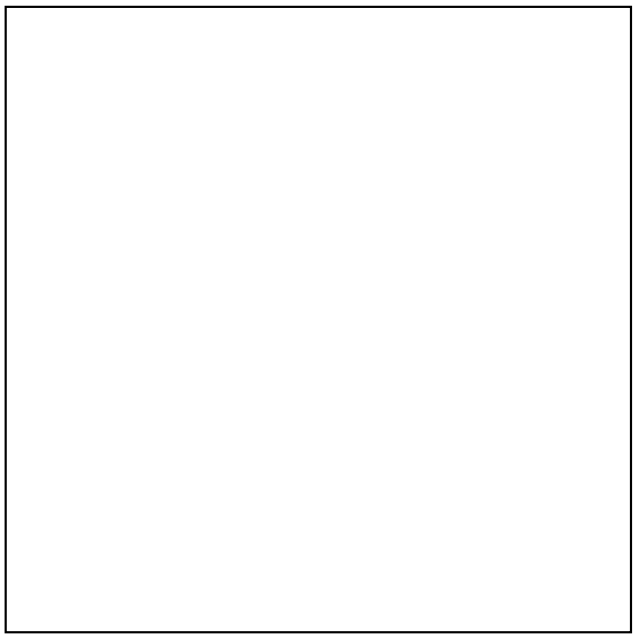


Table of Contents

1 Numerical Measures

2 Graphical Techniques

3 Homework

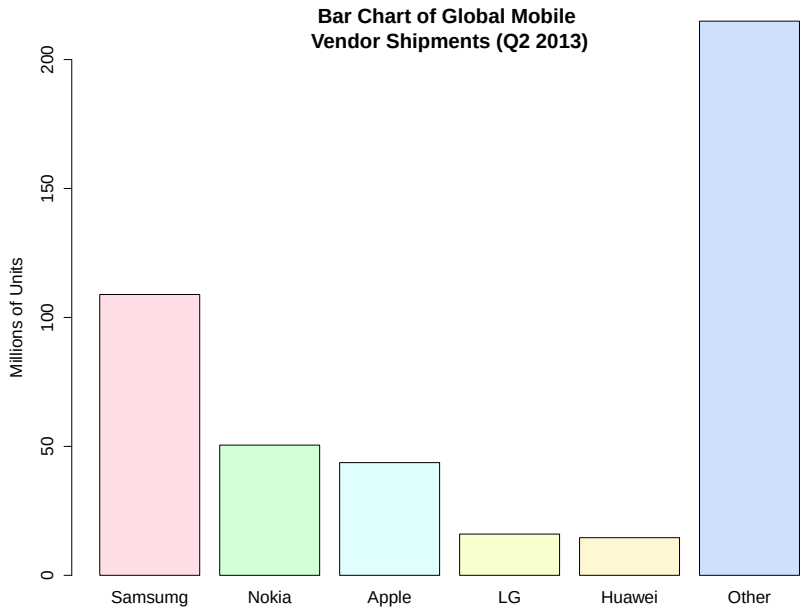
Graphs and Plots

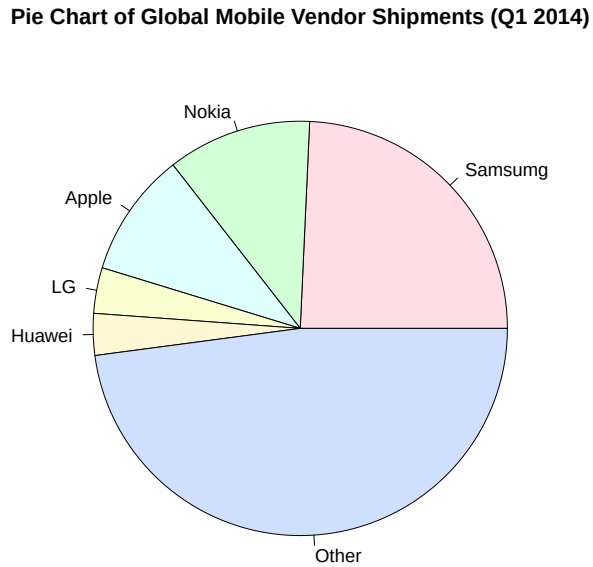
- Raw numbers on their own can be difficult to interpret.
- Pictures and plots can be very useful with representing information.
- Plots used for qualitative data
  - Bar chart
  - Pie chart
- Plots used for quantitative data
  - Histograms
  - Box plots
  - Stem and leaf plots

Bar Charts and Pie Charts

Manufacturer	Q2 2013 Shipments (Millions of units)
Samsung	108.9
Nokia	50.5
Apple	43.7
LG	16.0
Huawei	14.6
Other	214.9

Source: IDC Worldwide Mobile Phone Tracker,2014





Histograms

- A histogram is a graphical way to represent (quantitative) data.
- Similar to a bar chart but only for quantitative data.
- We make class intervals (of equal or varying width) which will contain our data of interest.
- We then count the frequency at which we observe data falling into one of these class intervals and construct a frequency table.

Histograms Ctd...

- A frequency table contains
  - Frequencies which are the counts that fall into an interval.
  - Relative frequencies which are the percentage of counts that fall into an interval.
- In a Frequency Histogram, the class intervals become the width of the bars of the histogram and the frequency will become the heights.
- In a Relative Frequency Histogram, the class intervals become the width of the bars of the histogram and the relative frequencies will become the heights.

Note: When we use the term "histogram" on its own, we usually refer to a frequency histogram.

Histograms Ctd...

General form of a frequency table

Class Interval	Freq.	Relative Freq.	Cumulative Freq.	Cumulative Relative Freq.
$[a_1, b_1)$	f_1	$r_1 = f_1/F$	f_1	r_1
$[a_2, b_2)$	f_2	$r_2 = f_2/F$	$f_1 + f_2$	$r_1 + r_2$
$[a_3, b_3)$	f_3	$r_3 = f_3/F$	$f_1 + f_2 + f_3$	$r_1 + r_2 + r_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_m, b_m]$	f_m	$r_m = f_m/F$	$f_1 + \ldots + f_m = F$	$r_1 + \ldots + r_m = 1$
	$F = \sum_{i=1}^{m} f_i$	1		

It may look overwhelming, but we will do an example.

Example

Suppose we have the following set of 30 observations which represents manufacturing times (in days) for mining equipment:

53	51	92	53	77	78	77	76	53	40
45	60	99	64	44	93	64	45	53	26
58	114	35	64	58	118	74	37	48	39

It is hard to see any obvious patterns with just the raw data alone.

Perhaps a picture will help.

Example Ctd...

First lets sort the data (Optional but very helpful):

26	35	37	39	40	44	45	45	48	51
53	53	53	53	58	58	60	64	64	64
74	76	77	77	78	92	93	99	114	118

After sorting, the following class intervals appear intuitively "nice":

Interval	Mathematical representation
20 to 39	$[20, 40)$
40 to 59	$[40, 60)$
60 to 79	$[60, 80)$
80 to 99	$[80, 100)$
100 to 120	$[100, 120]$

Example Ctd...

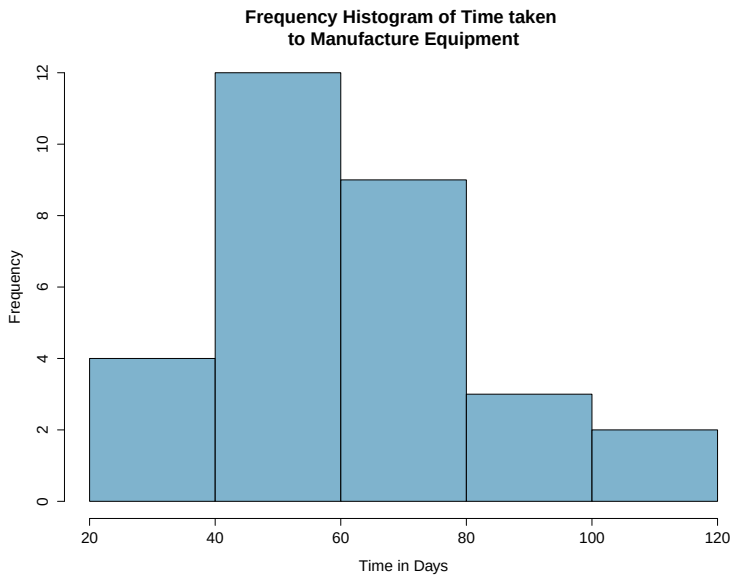
26	35	37	39	40	44	45	45	48	51
53	53	53	53	58	58	60	64	64	64
74	76	77	77	78	92	93	99	114	118

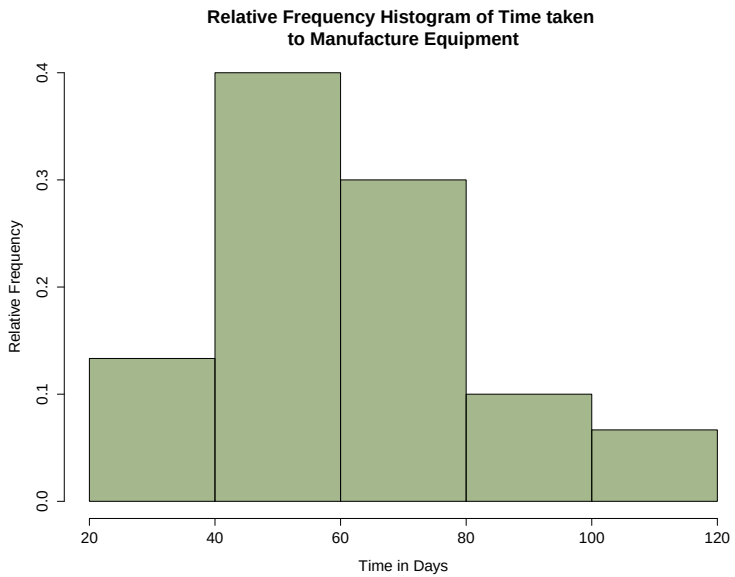
Complete the frequency table:

Class Interval	Freq.	Relative Freq.	Cumulative Freq.	Cumulative Relative Freq.
$[20, 40)$				
$[40, 60)$				
$[60, 80)$				
$[80, 100)$				
$[100, 120]$				

Example Ctd...

Class Interval	Freq.	Relative Freq.	Cumulative Freq.	Cumulative Relative Freq.
$[20, 40)$	4	0.13333333	4	0.13333333
$[40, 60)$	12	0.40	16	0.53333333
$[60, 80)$	9	0.30	25	0.83333333
$[80, 100)$	3	0.10	28	0.93333333
$[100, 120]$	2	0.06666667	30	1
	30	1		





Estimating the Mean using a Histogram

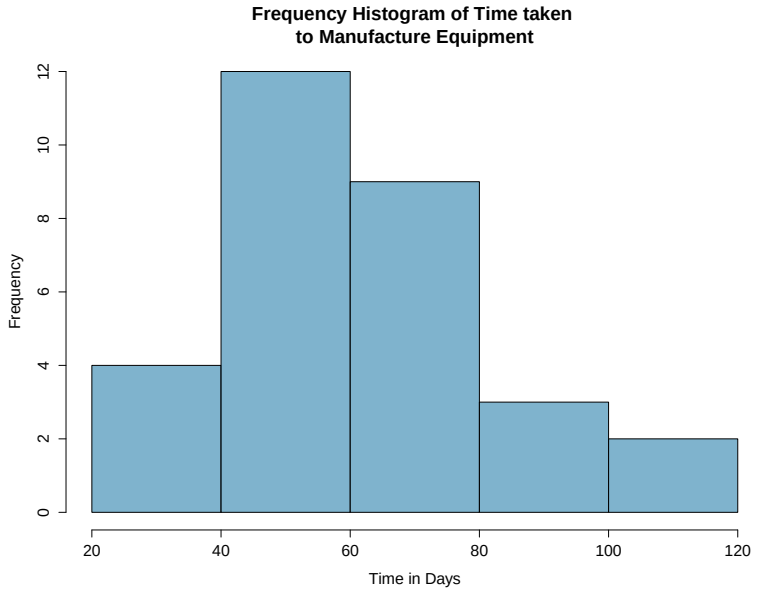
- We estimate the mean of a histogram using

$$\textit{Estimated mean} = \frac{\sum(\textit{midpoint of class interval}) \cdot (\textit{frequency})}{\textit{total}}$$

- Note that this estimation assumes a uniform spread of the date in each class interval.

Example

Estimate the mean of the following histogram



Skewness

- Skewness is a measure of how much a distribution leans towards a particular side or whether it is symmetric.

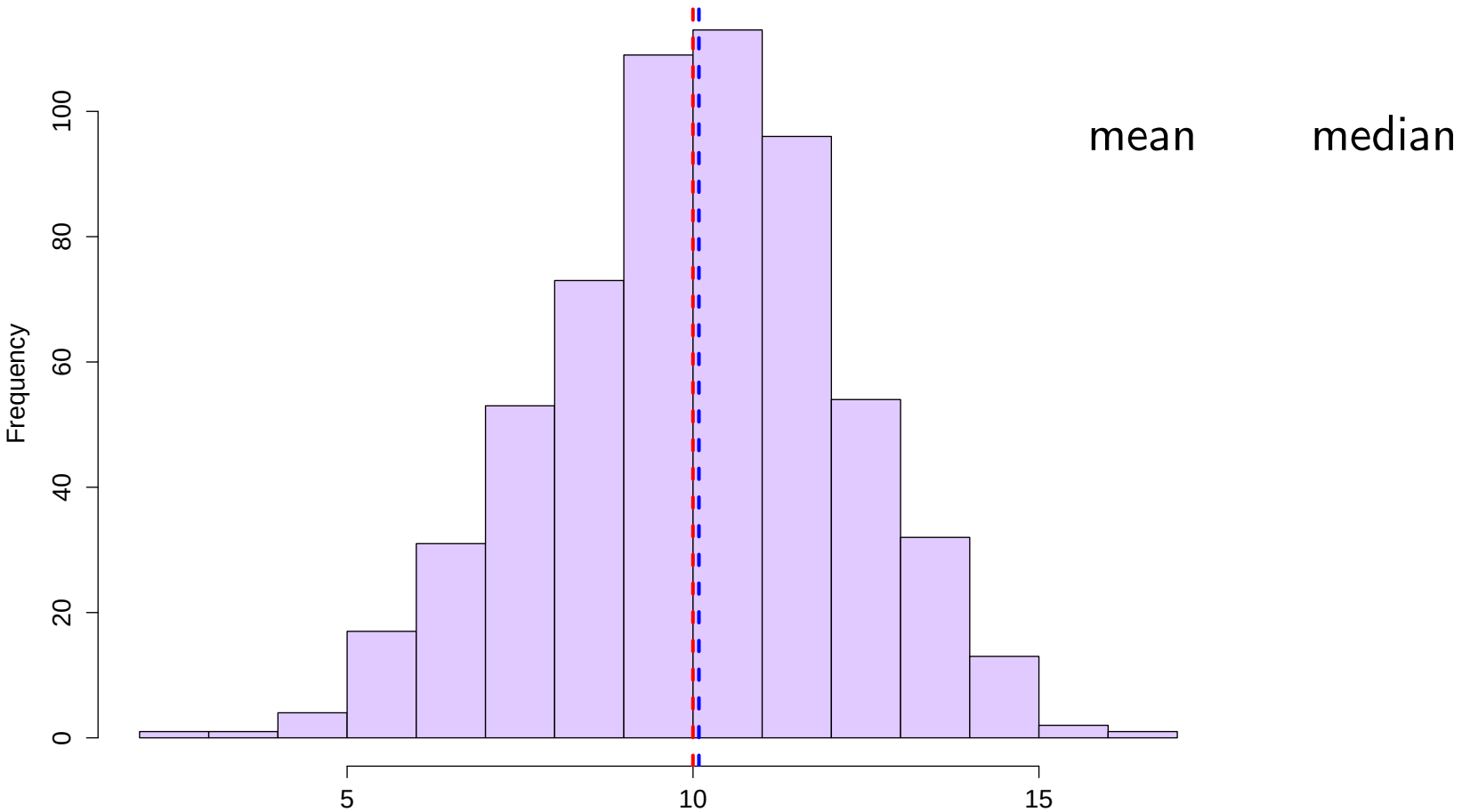
Symmetric	:	Most observations are concentrated around the mean and tail off fairly evenly on both sides of the mean.
Right skewed (Positively skewed)	:	We observe a tail to the right side of the mean. More observations are concentrated on smaller values.
Left skewed (Negatively skewed)	:	We observe a tail to the left side of the mean. More observations are concentrated on larger values.

Skewness Ctd...

- We can usually determine the skewness of a distribution by visually observing a histogram.
- By noting skewness, we get even more information about our data.
- Note however that we may not always be able to tell the skewness by visually observing a histogram.

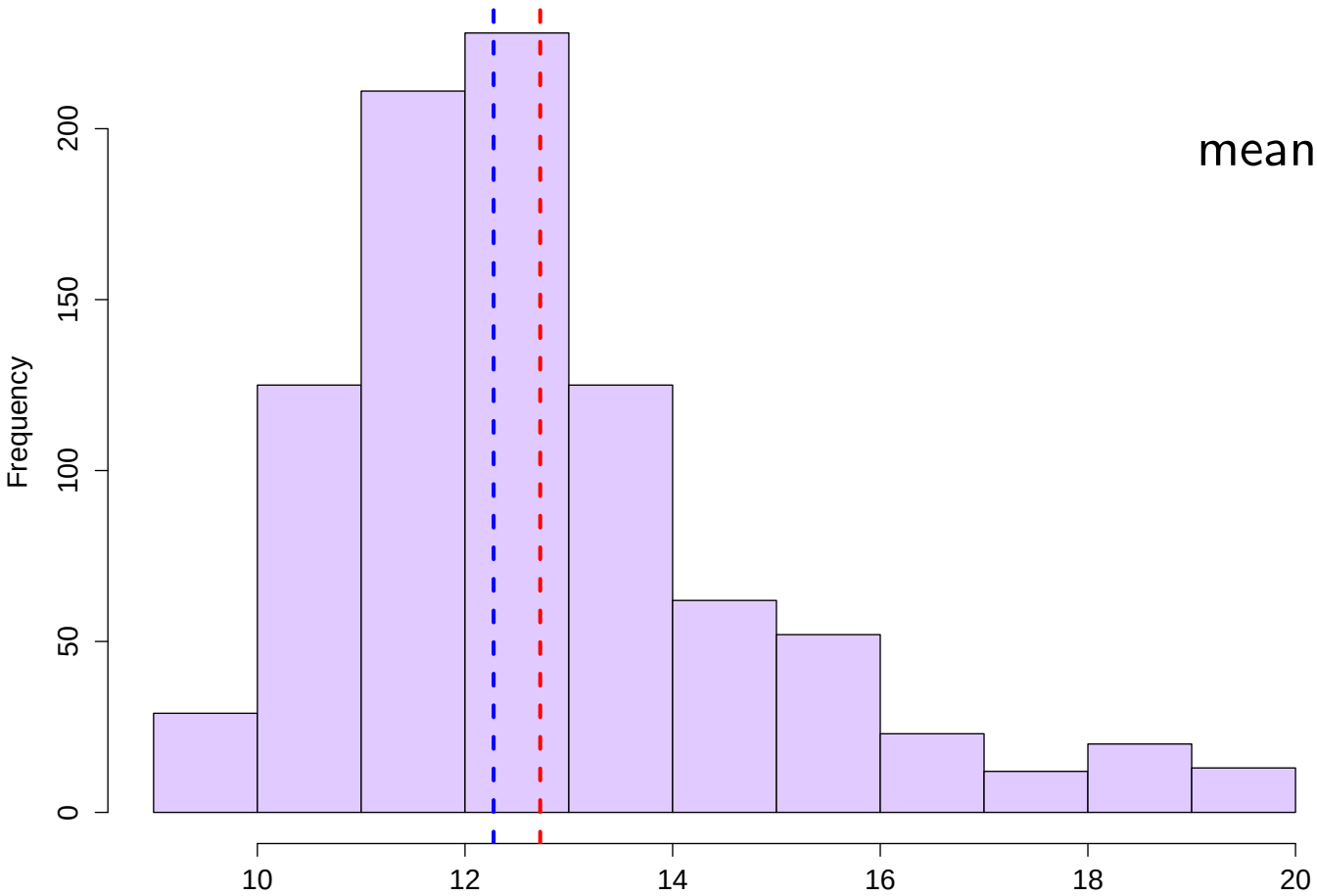
Skewness Ctd...

Symmetric



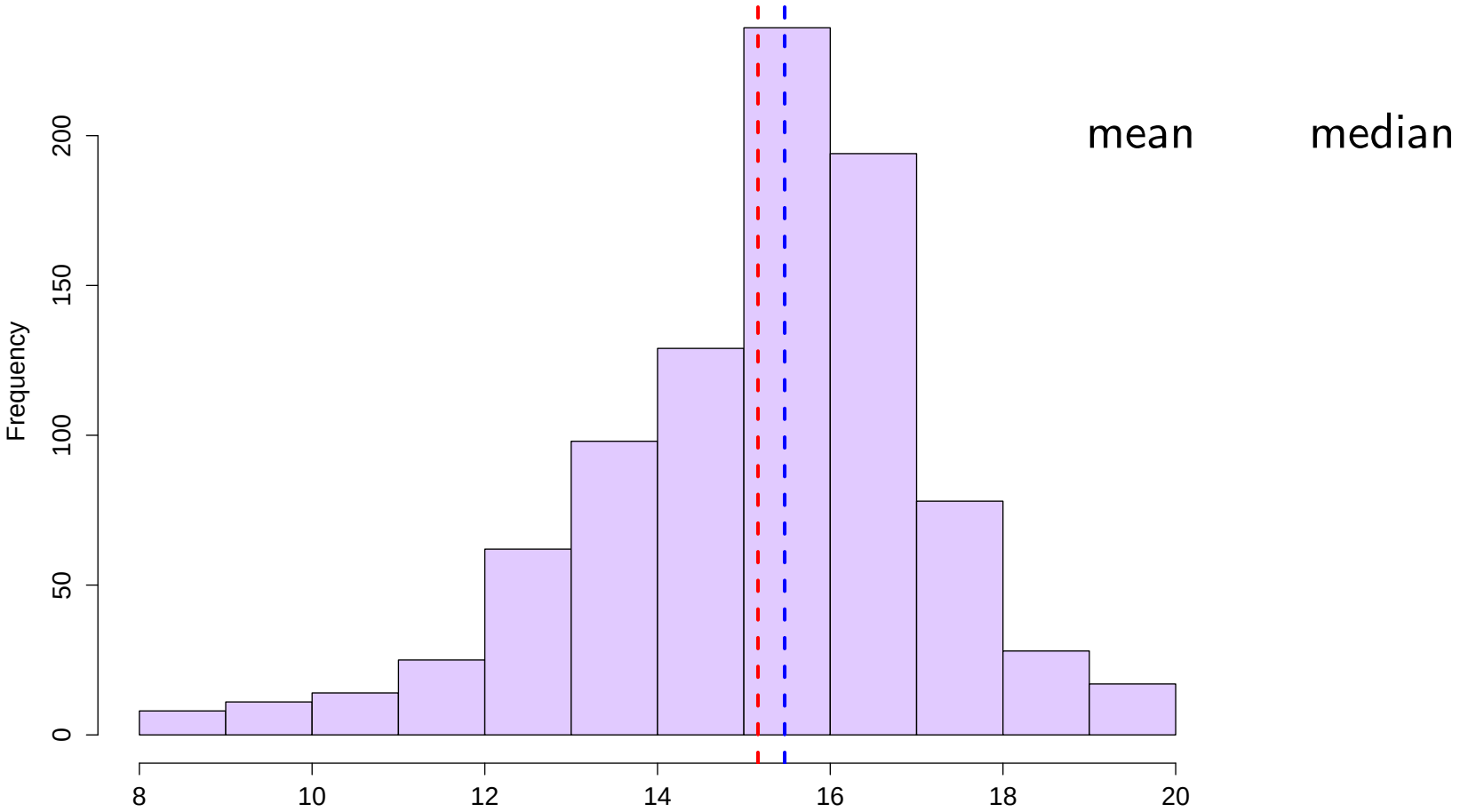
Skewness Ctd...

Skewed to the Right



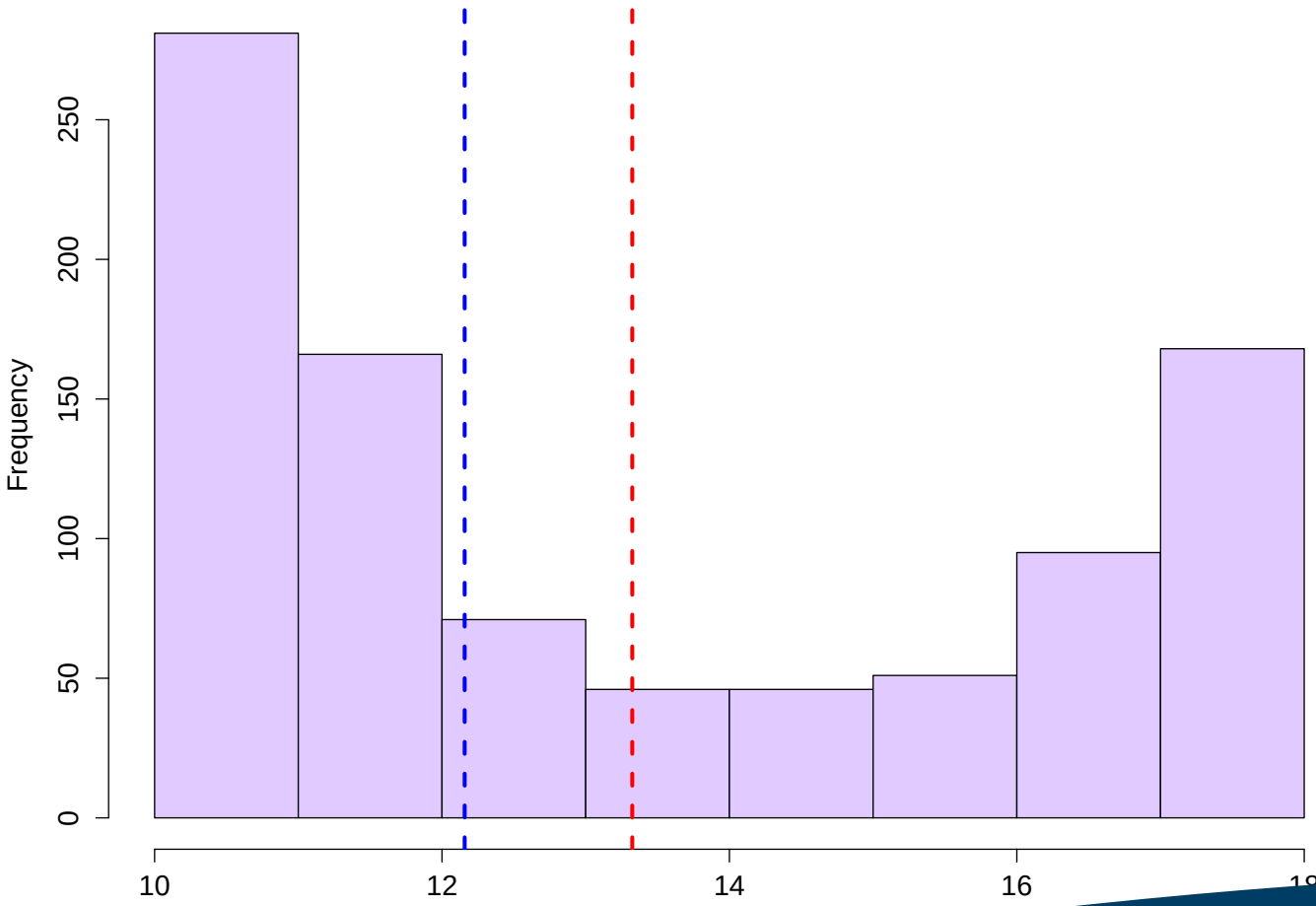
Skewness Ctd...

Skewed to the Left



Skewness Ctd...

None of the above (i.e. can not determine skewness visually)



Percentiles

Definition (Percentile)

For ordered data the p^{th} percentile is a value such that p% of observer data fall below it.

- We are particularly interested in:
  - The 25^{th} percentile
  - The 50^{th} percentile (median)
  - The 75^{th} percentile
- We call these values quartiles.

Quartiles

First Quartile (Q_1) : A value such that 25% of observations lie below it. (a quarter) (25^{th} percentile)

Second Quartile (Q_2) (median) : A value such that 50% of observations lie below it. (two quarters) (50^{th} percentile)

Third Quartile (Q_3) : A value such that 75% of observations lie below it. (three quarters) (75^{th} percentile)

The inter-quartile range (IQR) is:

$$IQR = Q_3 - Q_1$$

Quartiles Ctd...

Note:

- The "interquartile range" should not be confused with the "range".
- The range is defined as

$$\text{Range} = \max - \min$$

Example

Find the quartiles and interquartile range for the set of data below:

109, 112, 114, 120, 126, 132, 141, 142, 147, 150, 152

Boxplots

- Boxplots are another visual aid to present data.
- We use the quartiles as well as:

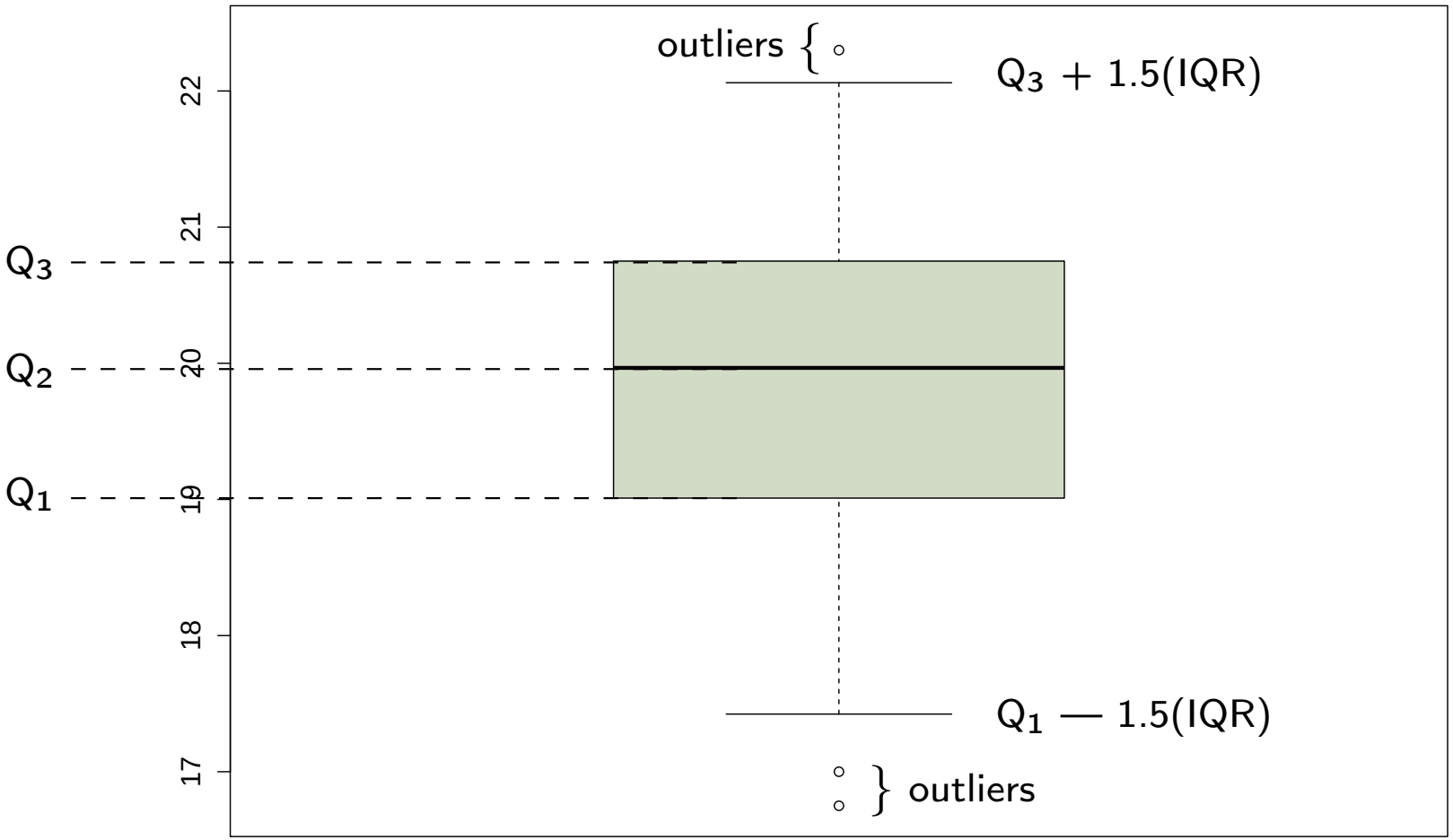
  Lower Whisker : $Q_1 - 1.5(IQR)$

  Upper Whisker : $Q_3 + 1.5(IQR)$

- Boxplots are useful in helping us identify outliers.
- An outlier is an unusual data point that appears to be far away from the rest of the data. (i.e. it appears outside the range that we would expect to see "typical" values of the data we are studying.

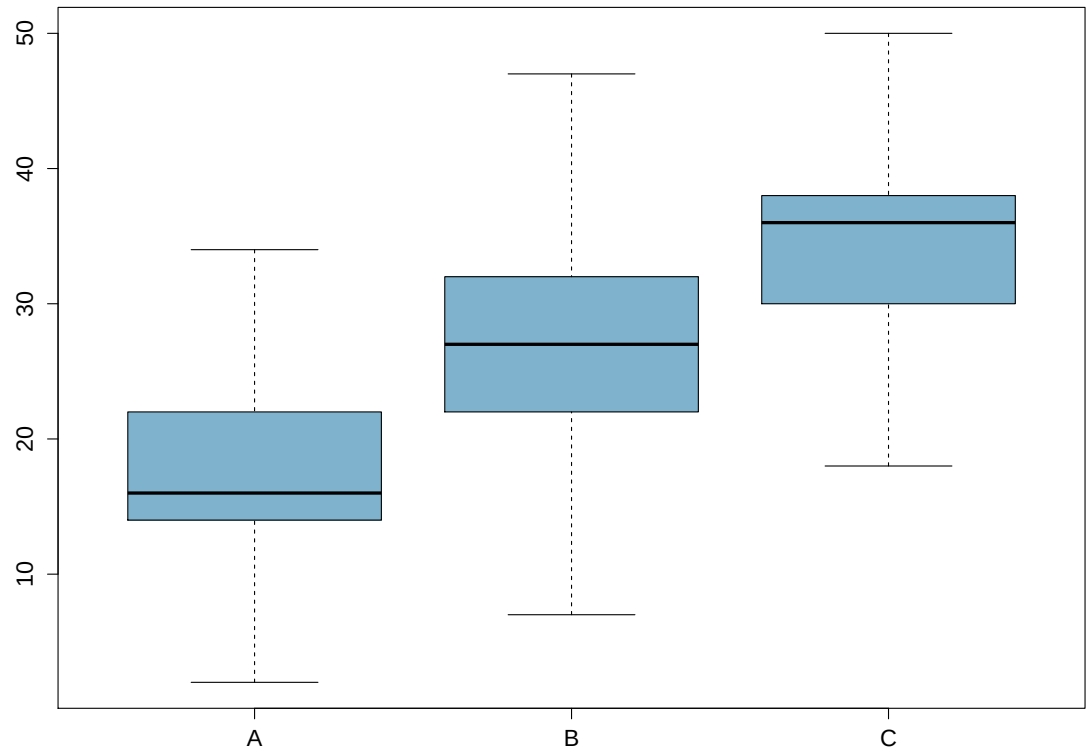
Boxplots Ctd...

How to interpret a boxplot



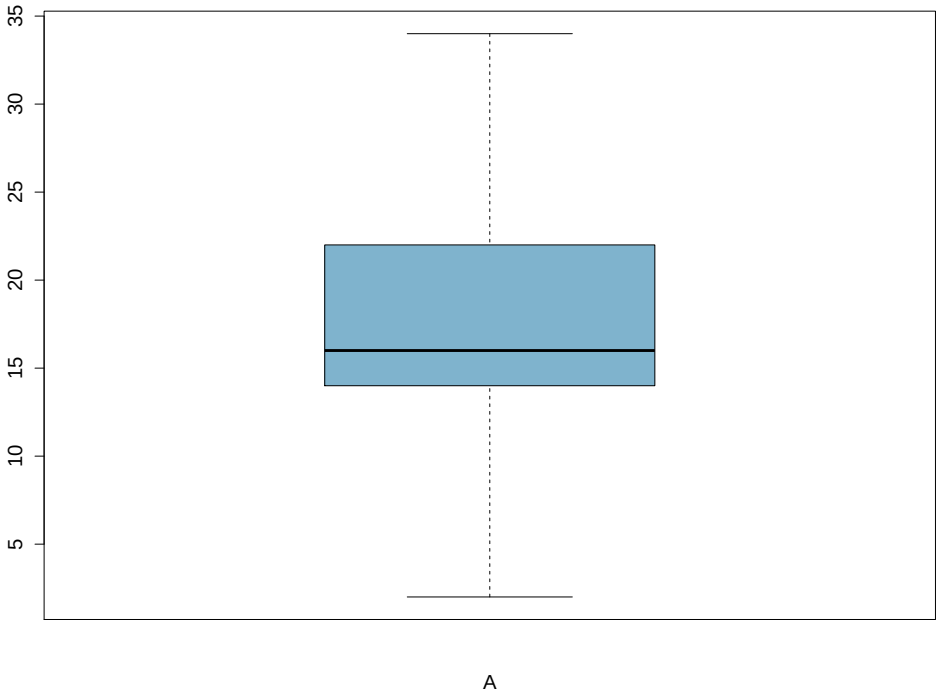
Boxplots Ctd...

- We can sometimes the get information about skewness of data from a boxplot.
- Consider the following 3 boxplots

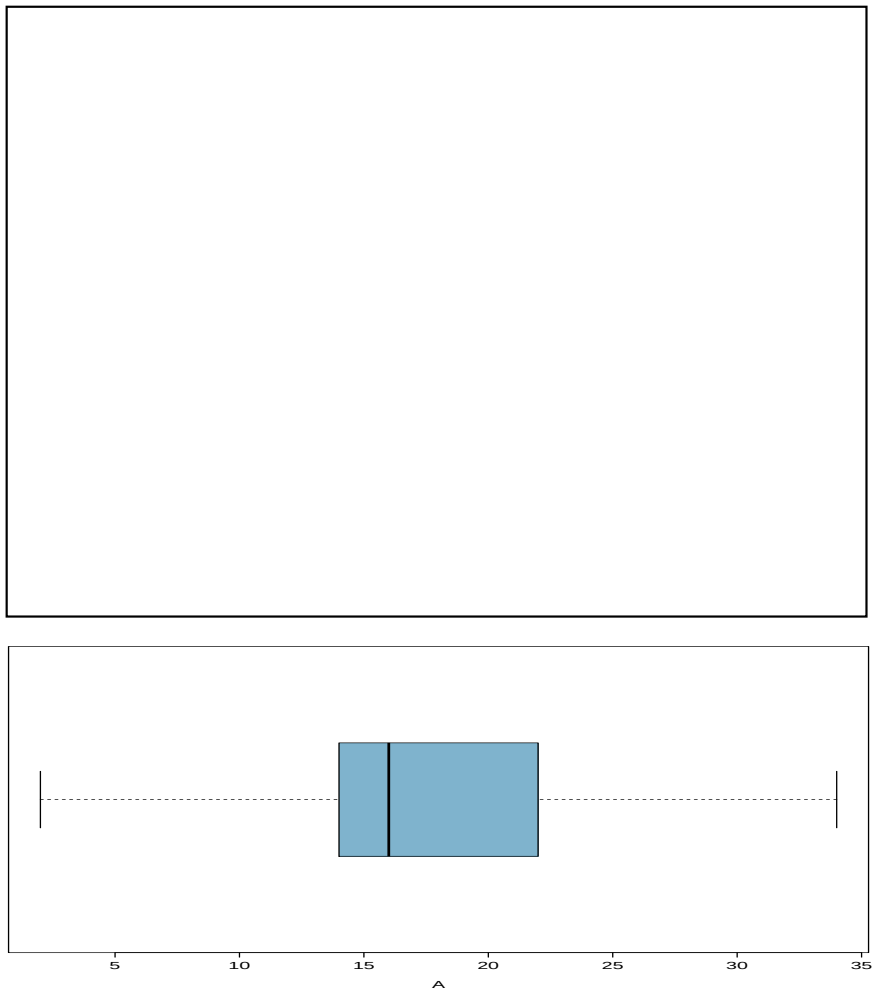


Boxplots Ctd...

Boxplot

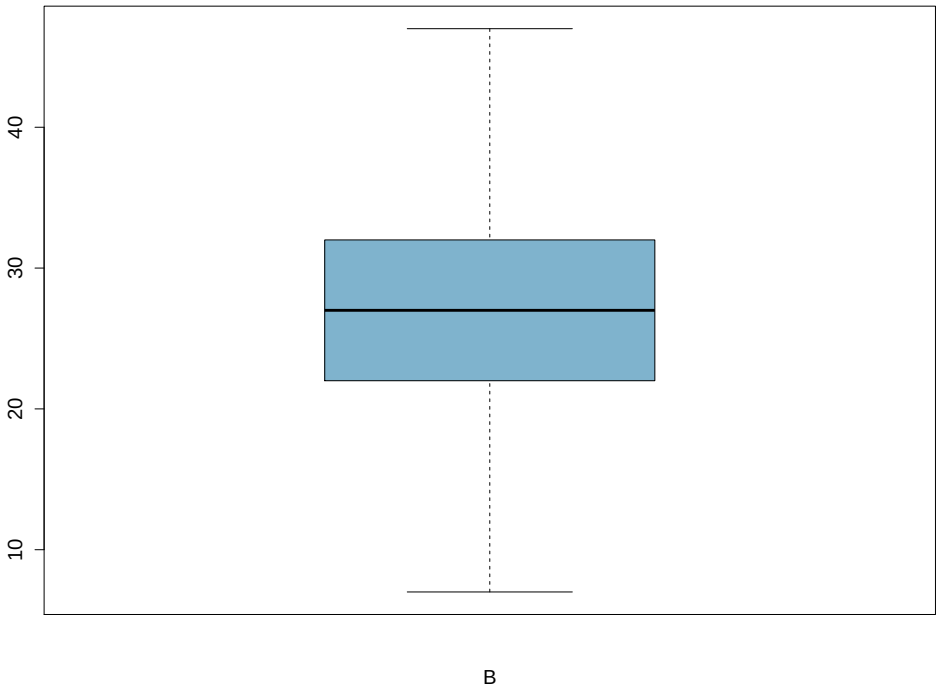


Distribution

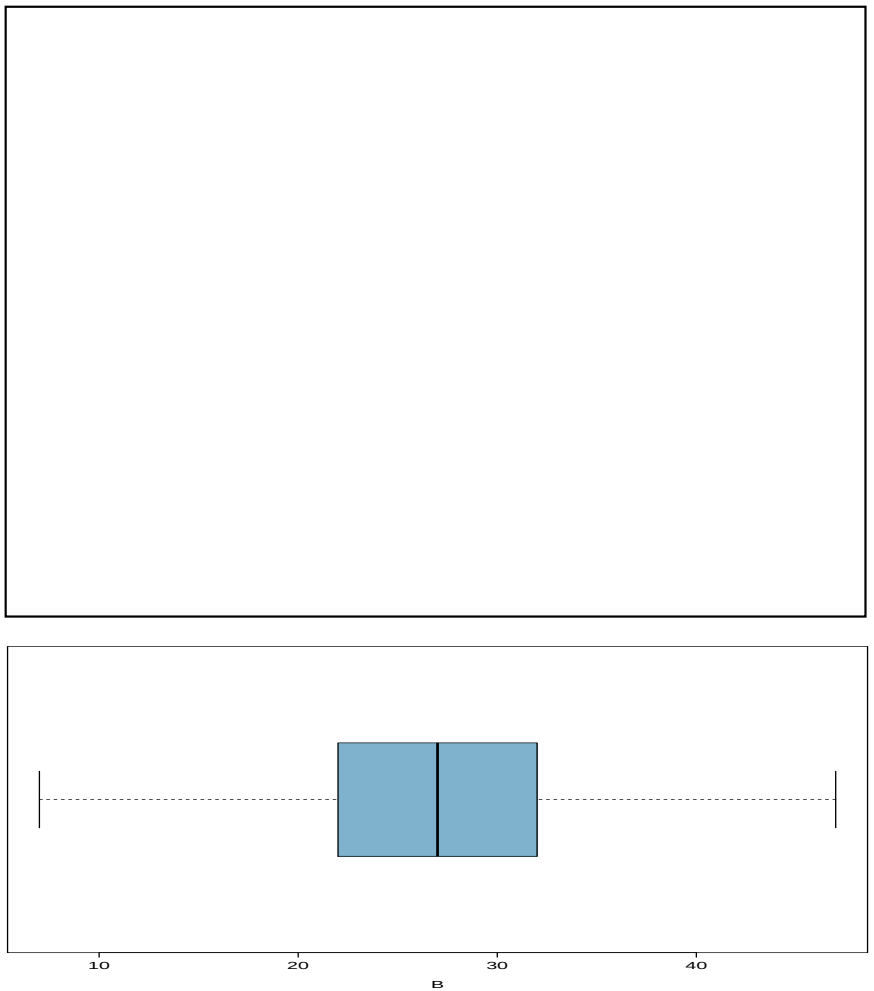


Boxplots Ctd...

Boxplot

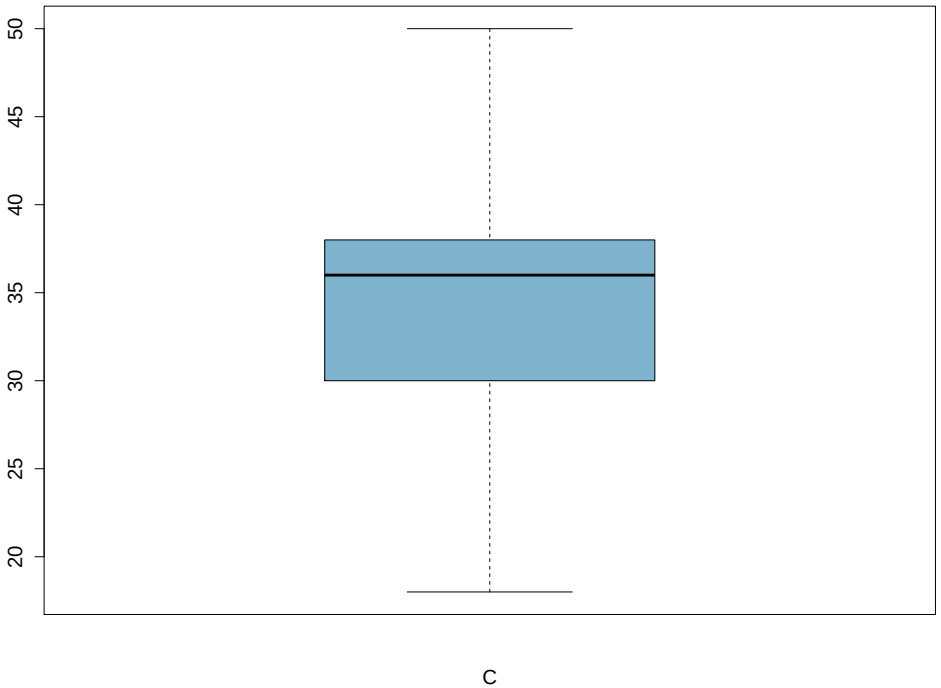


Distribution

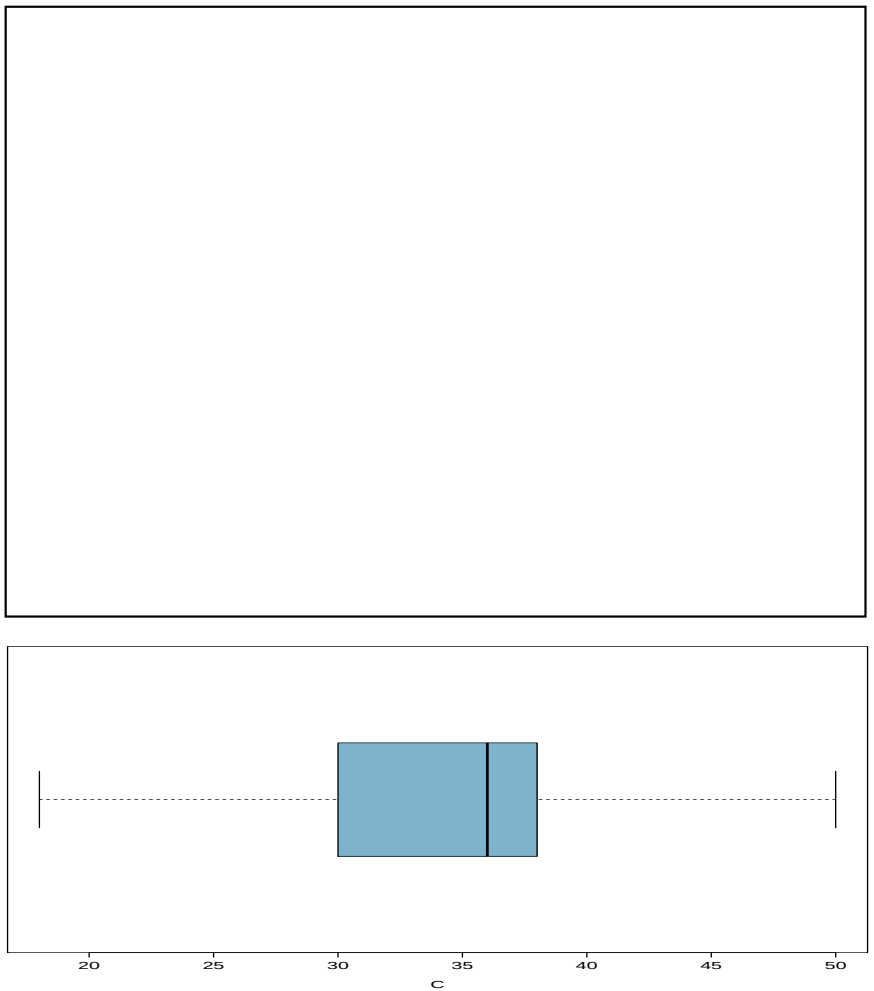


Boxplots Ctd...

Boxplot



Distribution



Stem and Leaf Plots

- A Stem and leaf plot (also known as a stem plot), is a tabular method to display data.
- Each observation is split into a stem (the "larger" part of an observation) and a "leaf" (the smaller part of an observation).
- One common partition is to use the stem as a whole number and the leaves as the decimal part of the number. Another common partition is to consider the stem to be the integer part of some power of 10 (ex. 10, 100, 1000 etc.).
- Stem and leaf plots are easier to understand with the aid of an example.

Stem and Leaf Plots ctd.

- For example consider the following set of (sorted) data points:

10.1	10.1	10.2	10.3	11.1	11.2	11.2	11.2	11.3	12.0
12.0	12.3	12.4	12.5	12.5	12.6	14.2	14.2	14.3	14.3
14.4	15.1	15.2	15.3	15.3	16.0	16.1	16.2	18.0	18.1

- Let's group terms and rewrite our data as follows:

10.1	10.1	10.2	10.3			
11.1	11.2	11.2	11.2	11.3		
12.0	12.0	12.3	12.4	12.5	12.5	12.6
14.2	14.2	14.3	14.3	14.4		
15.1	15.2	15.3	15.3			
16.0	16.1	16.2				
18.0	18.1					

- We can now construct out stem and leaf plot.

Stem and Leaf Plots ctd.

10.1	10.1	10.2	10.3			
11.1	11.2	11.2	11.2	11.3		
12.0	12.0	12.3	12.4	12.5	12.5	12.6
14.2	14.2	14.3	14.3	14.4		
15.1	15.2	15.3	15.3			
16.0	16.1	16.2				
18.0	18.1					

- Let's choose our partition (|) to be at the decimal point:

  Stem and Leaf Plot:

10	|	1123
11	|	12223
12	|	0034556
13	|	
14	|	22334
15	|	1233
16	|	012
17	|	
18	|	01

Table of Contents

1 Numerical Measures

2 Graphical Techniques

3 Homework

Homework

Readings

- Introduction to Probability and Statistics:
  Read 2.1.1 — 2.2.2 (Pages 23 — 34)

Exercises

- Custom Edition of Statistics for Business and Economics (McClave, Benson, Sincich):
  Exercises 2.35 — 2.41

- 11^{th} Edition of Statistics for Business and Economics (McClave, Benson, Sincich):
  Exercises 2.33 — 2.36