MTH102
Engineering Mathematics II
Academic Year 2017-2018
Semester 2
Chapter 1 Data Analysis

23 February 2018

Giovanni Merola

MTH102: Engineering Mathematics II

- Venue & Time
  - Lecture/1: Science Building SB102 09:00 – 11:00 Mon
  - Lecture/2: Science Building SB102 11:00 – 13:00 Wed

  - Tutorial/1: Science Building SD102 14:00 – 15:00 Fri
  - Tutorial/2: Science Building SD102 15:00 – 16:00 Fri

- Lecturer: Dr Giovanni Merola
  - Email: giovanni.merola@xjtlu.edu.cn
  - Phone Number: 8816 1635
  - Office: Business Building BA 242

MTH102: Engineering Mathematics II

- Recommended Textbook
  - **A first course in probability by S. Ross**
    Can buy (70RMB) on Amazon.cn
  - *Additional Readings*
  - For calculus review

Freely available on OpenStax.com

Calculus Volume 2

Calculus Volume 3

Introductory Statistics

MTH102: Engineering Mathematics II

- Assessment
  - Assignment 1: 5%,
  - Assignment 2: 5%,
  - Midterm Exam: 20%, Duration 1.5 hours (week 7)
  - Final Exam: 70%, Duration 2 hours

- Teaching Materials
  - Most will be uploaded to ICE in due time, so please check the following website regularly:
  - https://ice.xjtlu.edu.cn
  - ***N.B. The slides may be updated;; please check the version number in the file name and date on first page to ensure you get the latest version.***

Attendance

- I will take attendance electronically using ICE
- **You need to have a cell phone or tablet** with which you register your presence.
- You can use the ICE app or a web browser
- Attendance is required by the university and **will not be used for your evaluation**
- However, **attending classes and practicing during the course** will greatly enhance your learning experience because otherwise after a few weeks you will not understand what is being discussed. Needless to say that it will also increase your final result.

Notes

- Printing out the slides, read them before class and bring them to class will help you follow the lesson. Also bring a calculator to class
- Please put your phones to silent during class
- Do ask questions during or after class
- Attend the tutorials: essential for solving exam questions
- Practice during the week, if you don't you will not be able to follow lessons after a few weeks
- Office hours Mon/Wed 14:00-15:00, you can send me quick questions by email or arrange a meeting also by email

Questions?

- Any questions now?

- If you have questions later, please approach me after class or send me mail

Chapter 1.1.1 Data Representation

- 1.1 Sources of data
- 1.2 Four scales of measurement
- 1.3 Data and its Representation
- 1.4 Describing Distributions
  - Central Tendency
  - Spread
  - Shape
- 1.5 Box-and-whisker Plot
- Summary

1.1.1 Sources of data

- Primary:
  - Observational studies (most common)
  - Surveys (record opinions – not directly observable)
  - Experiments (most powerful and expensive)
- Secondary
  - Distributed by organisations (administrative or official data - reliable)
  - Commercial or freely available (books, CDs, Internet - unverified)

Sources of data: Exercise (3 minutes)

Want to find out:

1. customer satisfaction?
2. Is a new drug useful?
3. Population in Italy
4. Import of goods in China from Viet Nam in USD
5. Viewers rating of new “Star Wars” film
6. Percentage of XJTLU students who come with an e-bike

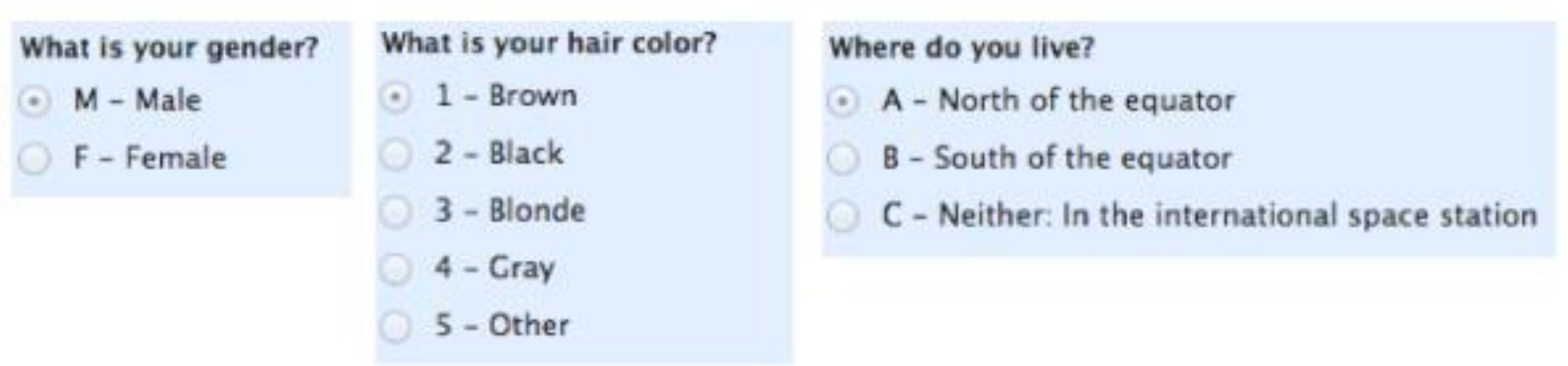
Choices

A. Observational studies
B. Surveys
C. Experiments
D. Distributed by organisation
E. Commercial or free

1.1.2 Four scales of measurement

1) Nominal (Name/Label) (Qualitative = not numerical)

- E.g. Countries in Europe, the number pinned on the T-shirt of sportsperson
- Not a quantitative value

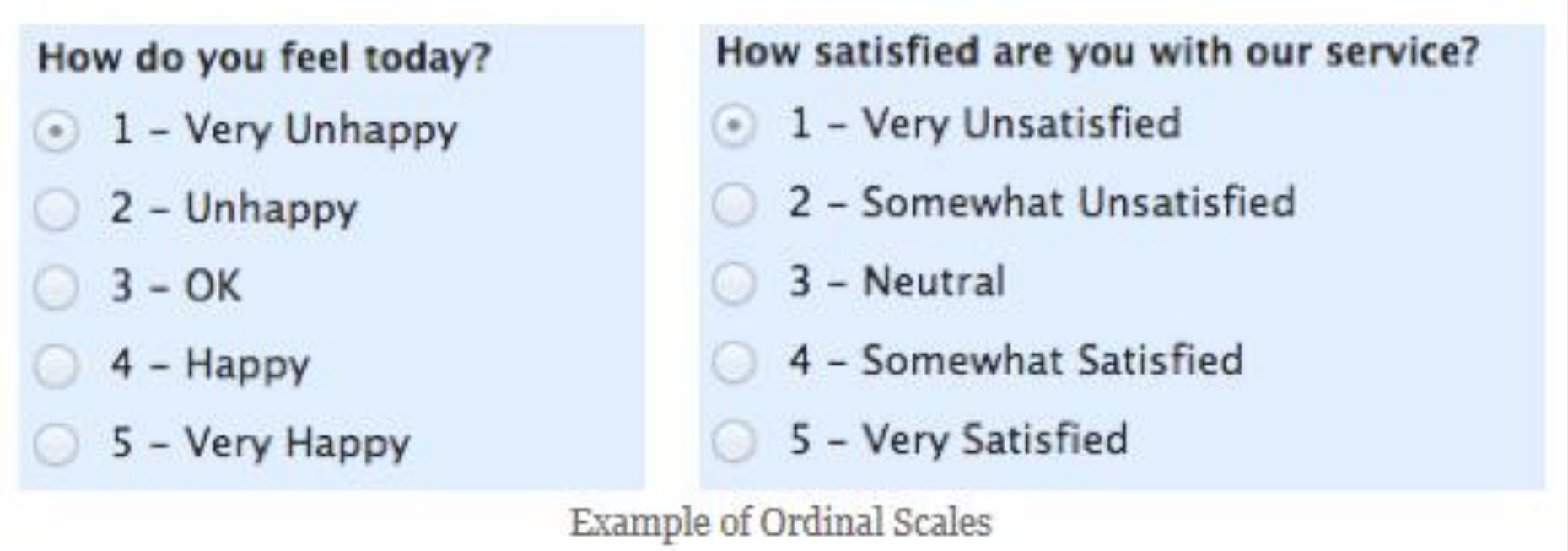


Examples of Nominal Scales

1.1.2 Four scales of measurement

2) Ordinal [Order] (Qualitative)

- E.g. Happiness level on a scale of 1 to 10
- Not a quantitative value



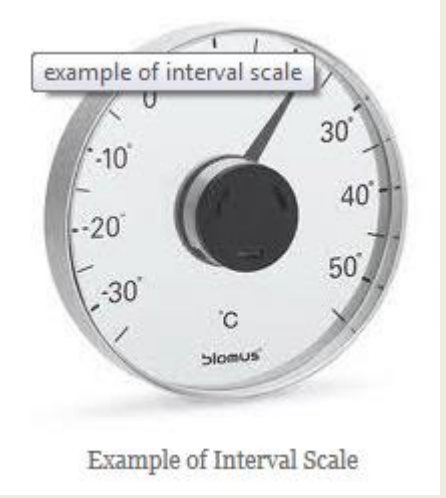
Example of Ordinal Scales

1. 1.2 Four scales of measurement

3) Interval (Quantitative)

- Different “zero”; order known; differences between values known
- E.g. Celsius scale vs Farenheiht; Western or Arabic calendar
- “Difference between 60 and 50 degrees is measurable 10 degrees, as is the difference between 80 and 70 degrees”
- Impossible to obtain ratio. Example

Temperature	Celsius	T_i/T_{i-1}	Farhenheit	T_i/T_{i-1}
T1	10		50	
T2	20	2	68	1.36
T3	30	1.5	86	1.26



Example of Interval Scale

1.1.2 Four scales of measurement

4) Ratio (Quantitative)

- quantitative; order known; differences between values known, “true zero”
- E.g. Weight, height, time passed (age, delay, etc)
- Possible to obtain ratio; “20 kg is twice as heavy as 10 kg” and “20 pounds is twice as heavy as 10 pounds”

This Device Provides Two Examples of Ratio Scales (height and weight)

Types of data: Exercise (3 minutes)

1. Telephone number
2. Intelligence quotient (IQ)
3. Education level (elementary, highschool, ..)
4. First names

Nominal
Ordinal
Interval
Ratio

1.1.2 Four scales of measurement

Summary

In summary, **nominal** variables are used to "*name*," or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median		✔	✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

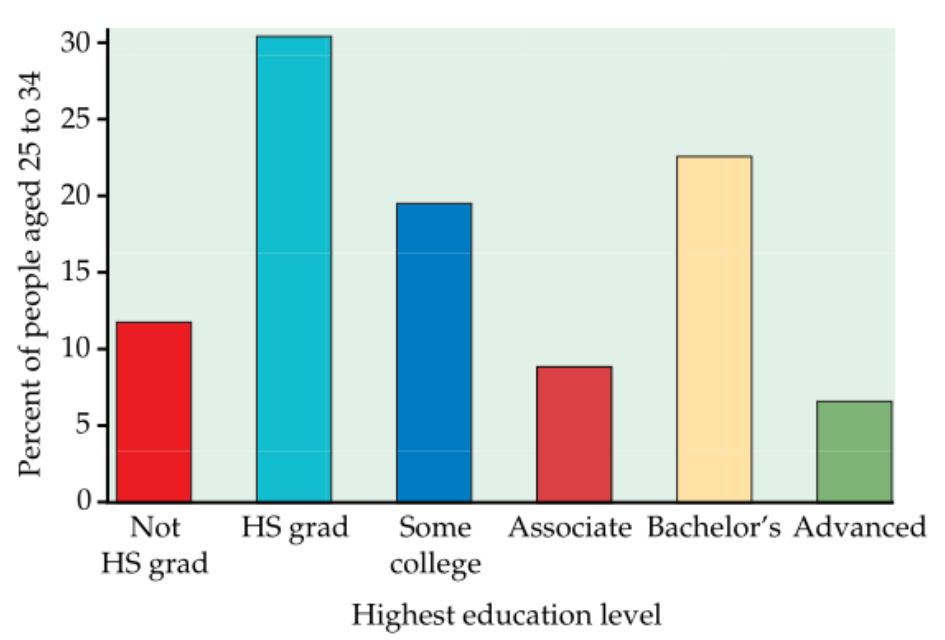
summary of data types and scale measures

Mode ✓
Median ✗

1.1.3 Data and its Representation

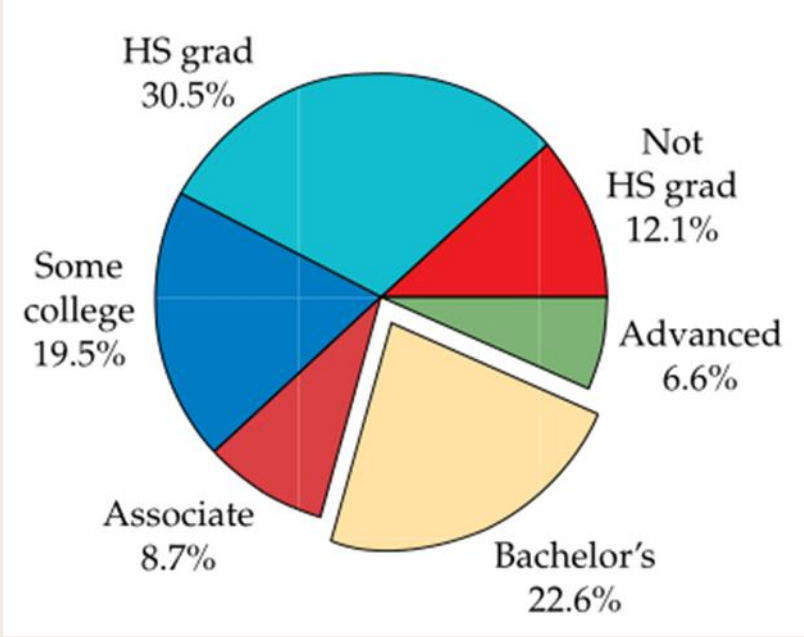
For Qualitative data

Bar graph



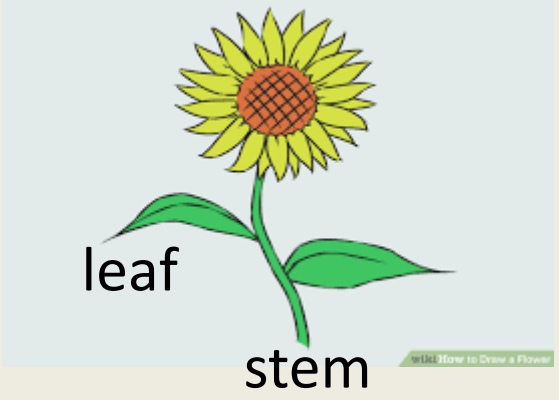
Bars have gaps. Always add axis labels

Pie Charts



Pretty but difficult to read. Always add labels with % values

1.1.3 Data and its Representation



Quantitative data Stem and Leaf Plot

TABLE 1.2

Literacy rates (percent) in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

Order the data.

Tens are stems

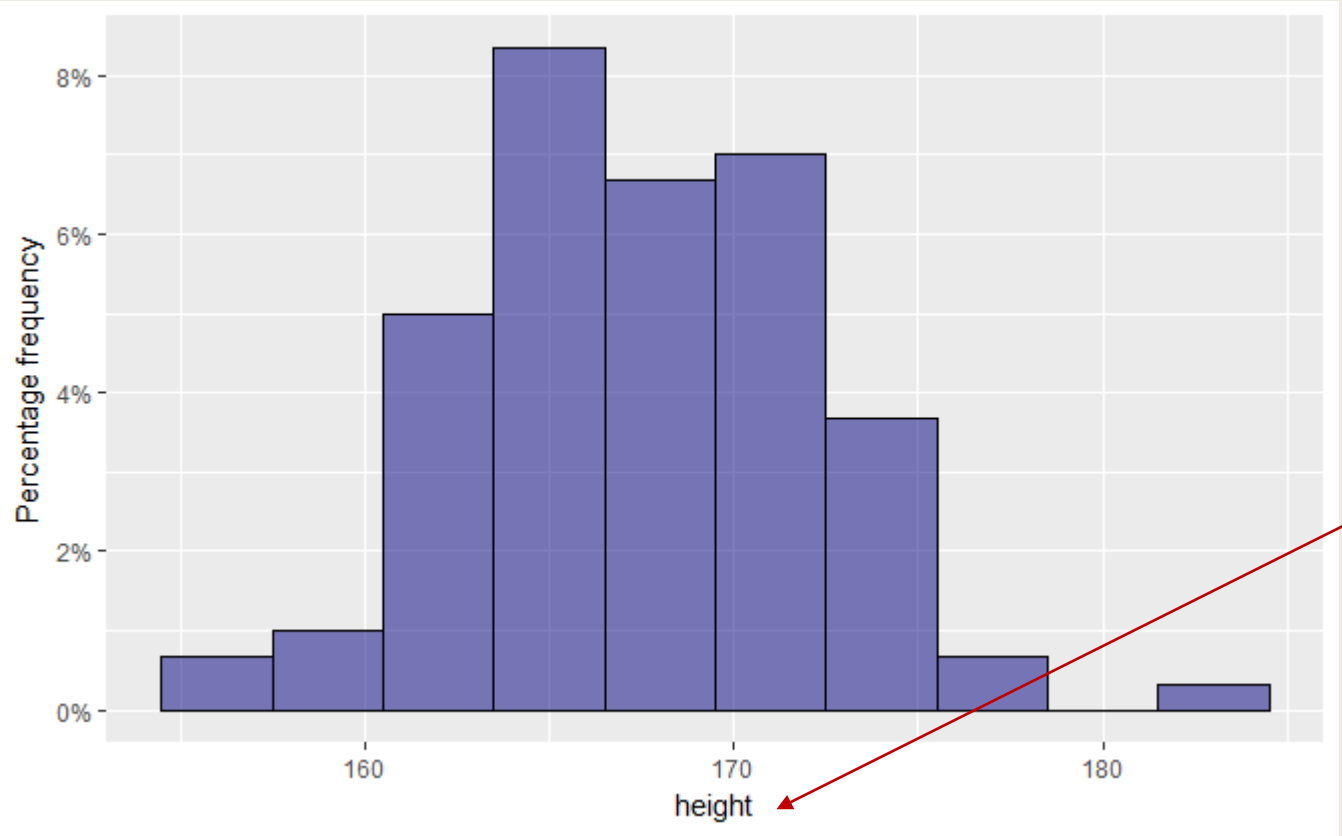
Units are leaves

Female		Male
9	2	
81	3	
6	4	
	5	0
330	6	88
8110	7	08
652	8	3459
999	9	22456
	10	000

- Most frequent % ten?
- More in the 60s or 90s?
- How do you describe the data?

1.1.3 Data and its Representation

- Numerical data: Histograms

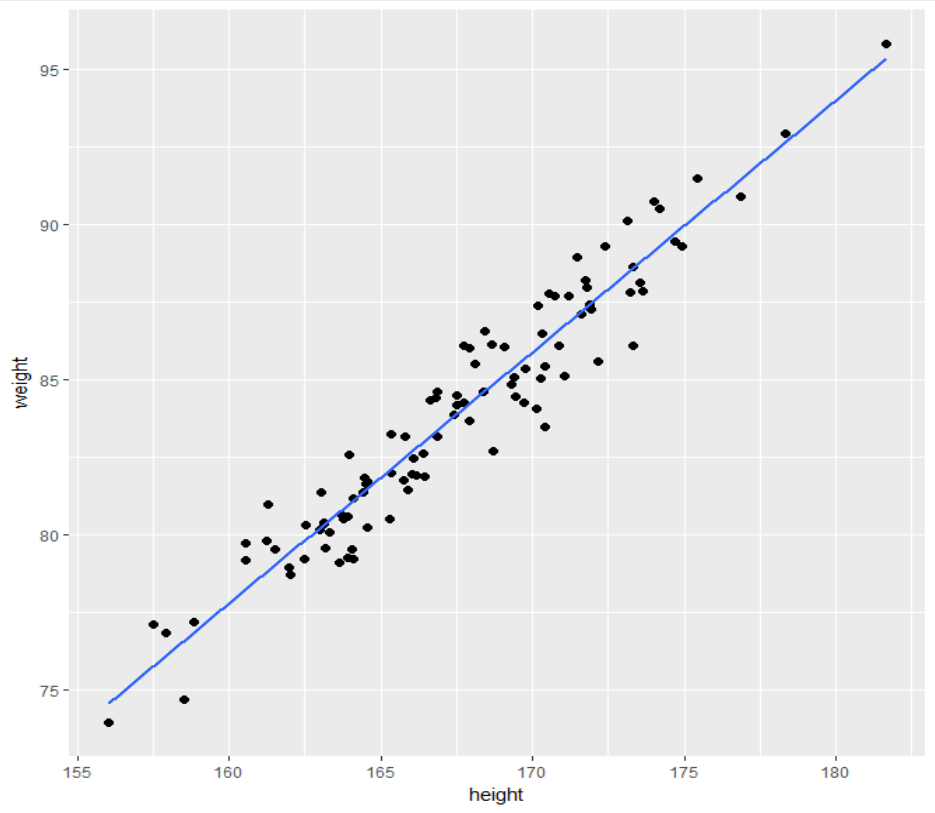


Bars have no gaps
(numerical data)
Always add axis labels

- How do you describe the data?

1.1.3 Data and its Representation

- Numerical data: Scatter plot 2 variables: x and y



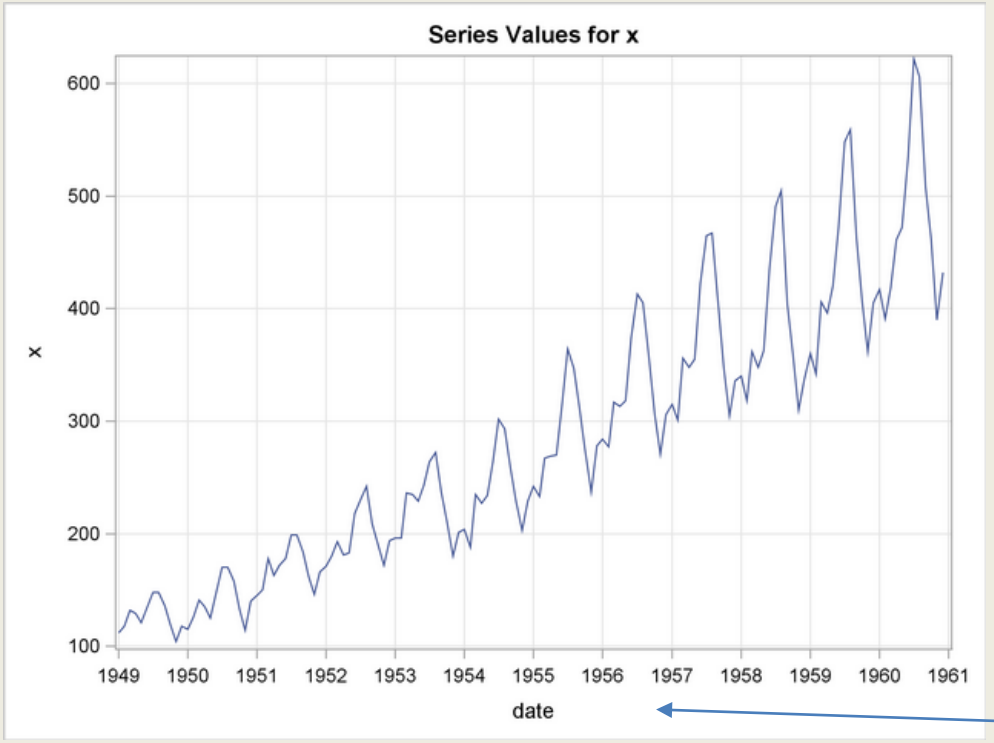
Data are points.
The fitted line shows linear relationship – very strong here-

Try to make the plotting area almost square (aspect ratio 1:1)
Always add axis labels

- How do you describe the data?

1.1.3 Data and its Representation

- Time Series Plot



Join observations,
time is ordered!

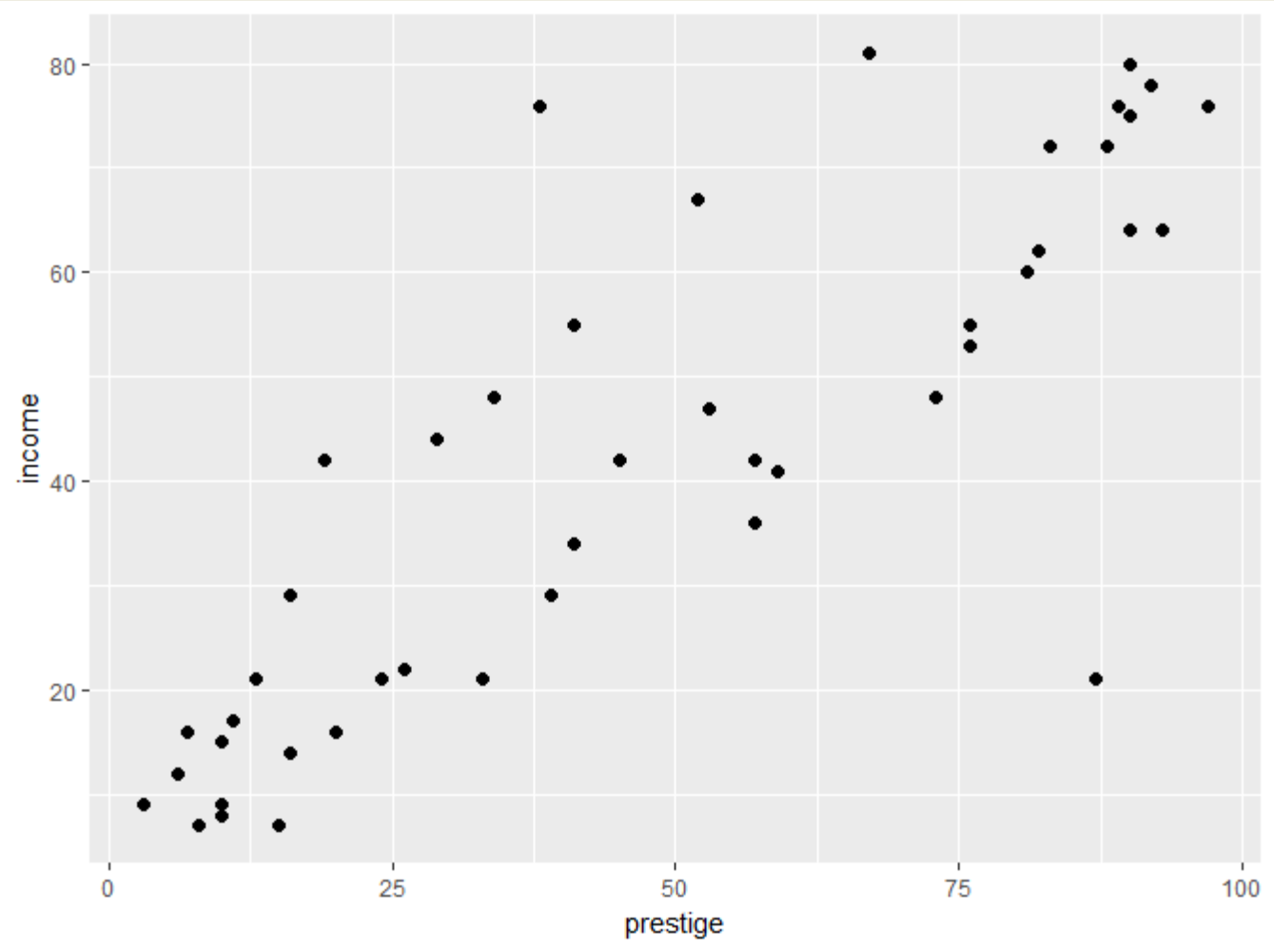
Time always on horizontal axis

- How do you describe the data?

1.1.3 Data and its Representation: problem (3 minutes)

Comment this scatter plot of income versus job prestige

job name	income	prestige
accountant	62	82
pilot	72	83
architect	75	90
author	55	76
.	.	.
.	.	.
.	.	.



1.1.4 Describing Distributions

We want to describe asset of numbers with few “statistics”, that is summaries of the data. What can be describe?

1) Central Tendency: the single numerical value considered “the most *typical* of data”. A central value.

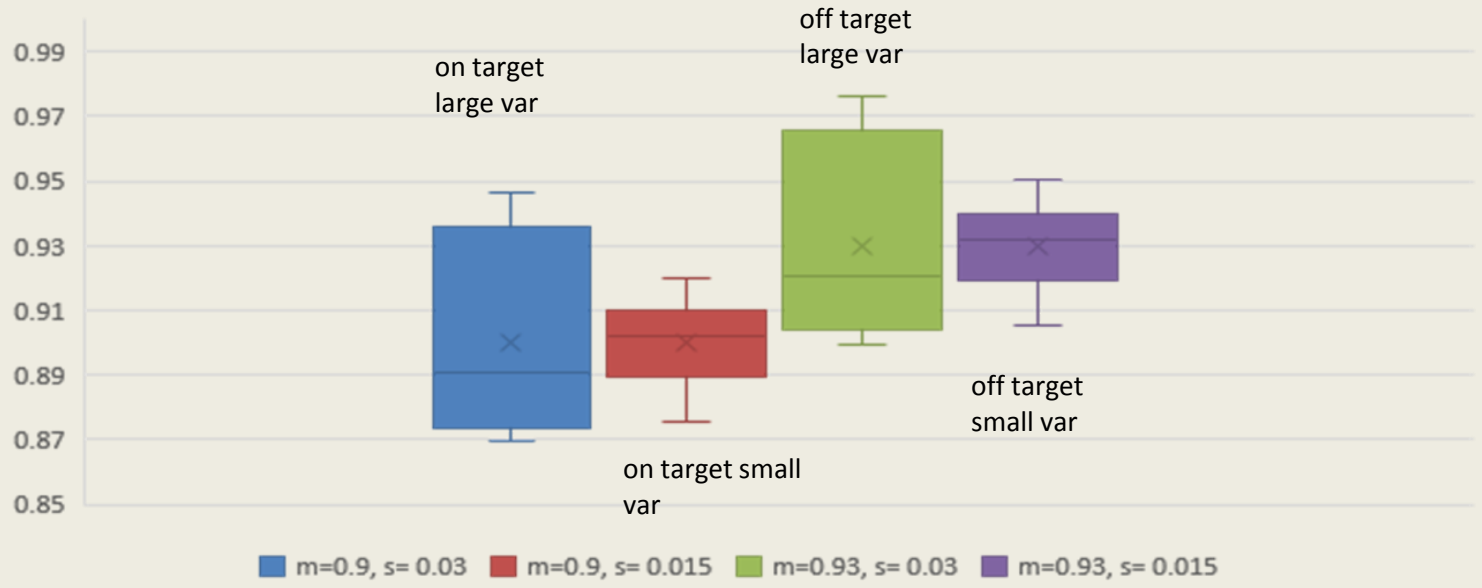
2) Spread: how much data are different from a central value.

3) Skew: are there more values on the left or on the right of the central value? Or are they more or less the same?

Could use more but we need to be concise

1.1.4 Describing distributions

- Consider a chemical reactor producing a polymer. The target density is 9.0g/cc
- Every day the density of the product is slightly different, it is a random variable depending on many factors. During the month the density is distributed according to some distribution.



1.1.4 Describing Distributions: Central tendency

1) Measures of Central Tendency: the single numerical value that is considered to be the most *typical* of data.

The equality value: if all observations are equal they are equal to the mean

i. **Mean $\bar{x}$**: arithmetic average $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

E.g. the mean of the observations 2, 3, 3, and 4, is equal to 12/4 = 3.

ii. **Median** m: the centre point in a set of **ordered numbers**; it is also the fiftieth percentile.

$$m := \begin{cases} x_{\frac{n+1}{2}}, \text{when } n \text{ is odd.} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \text{when } n \text{ is even.} \end{cases}$$

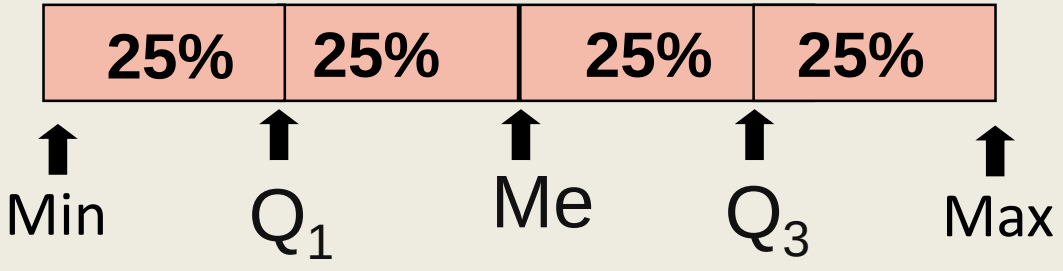
E.g. the median of 1, 2, 3, 4, 5 is equal to 3;
the median of 1, 2, 3, 4, 5, 6 is equal to (3 + 4)/2 = 3.5.

Splits the observations into lower and upper half.

iii. **Mode**: the most frequently occurring number

1.1.4 Describing Distributions: quartiles

- **Quartiles** split the data into 4 quarters. The middle one is the median: Q_2 = Median
- Given a set of data $\{x_1, x_2, \cdots, x_n\}$ in ascending order,
  - If n is odd: Q_1 is the median of $\{x_1, x_2, \cdots, x_{\frac{n+1}{2}-1}\}$ and Q_3 is the median of $\{x_{\frac{n+1}{2}+1}, \cdots, x_n\}$.
  - If n is even: Q_1 is the median of $\{x_1, x_2, \cdots, x_{\frac{n}{2}}\}$ and Q_3 is the median of $\{x_{\frac{n}{2}+1}, \cdots, x_n\}$.

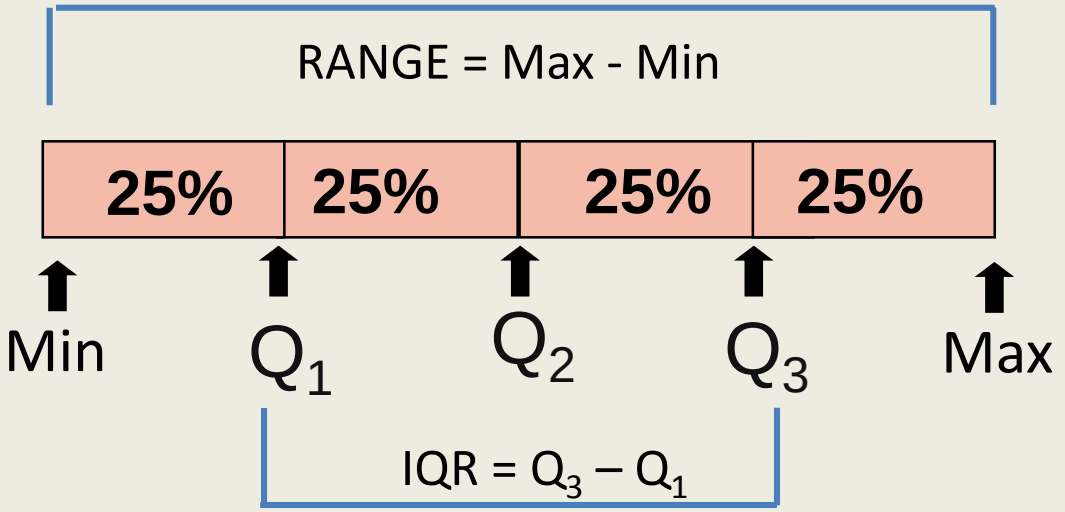


1.1.4 Describing Distributions: spread

2) Measures of Spread: information regarding the variability of the data.

i. **Range**: difference between the biggest and smallest numbers

ii. **Interquartile range IQR**: difference between upper quartile Q_3 and lower quartile Q_1



1.1.4 Describing Distributions: spread 2

iii. **Variance s^2 and Standard deviation s**: The standard deviation tells you, on average, how far the numbers vary from the mean.

For n observations $x_1, x_2, \cdots, x_n$ obtained

- Variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ (n-1 gives better estimation properties)
- Standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

s^2 and s are large if the observations are widely spread about $\bar{x}$, and small if the observations are all close to $\bar{x}$.

1.1.4 Describing Distributions: Putting all together

n = 16, values = 1 3 4 5 5 6 8 10 15 15 16 18 18 20 20 26

Mean = (1 + 3 + …. + 26)/16 = 11.875

Median: n is even, so n/2 = 8, Me = $\frac{x_8+x_9}{2} = \frac{25}{2} = 12.5$

1st quartile, Q_1 is median of $(x_1, .., x_8) = \frac{5+5}{2} = 5$

3rd quartile, Q_3 is median of $(x_9, .., x_{16}) = \frac{18+18}{2} = 18$

Range = Max – Min = 26 – 1 = 25

IQR = $Q_3 - Q_1 = 18 - 5 = 13$

$S^2 = 56.65, \qquad S = \sqrt{56.65} = 7.53$

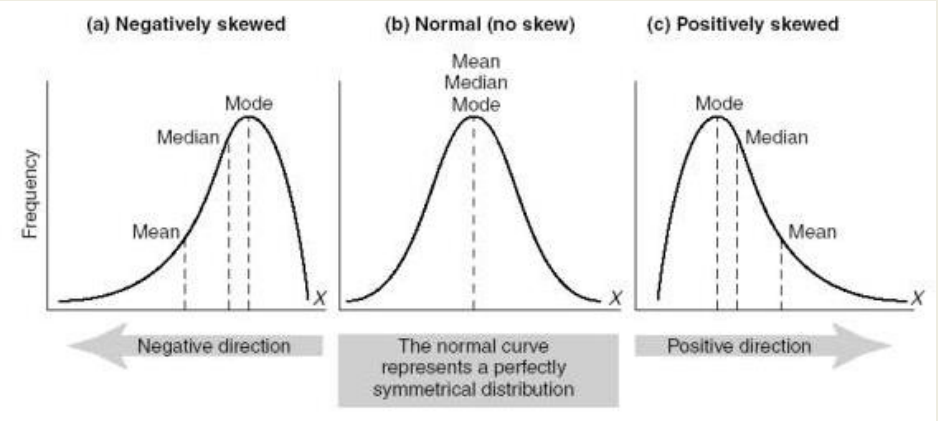
1.1.4 Describing Distributions

3) Shape:

No skew: symmetric
Mean ≈ median

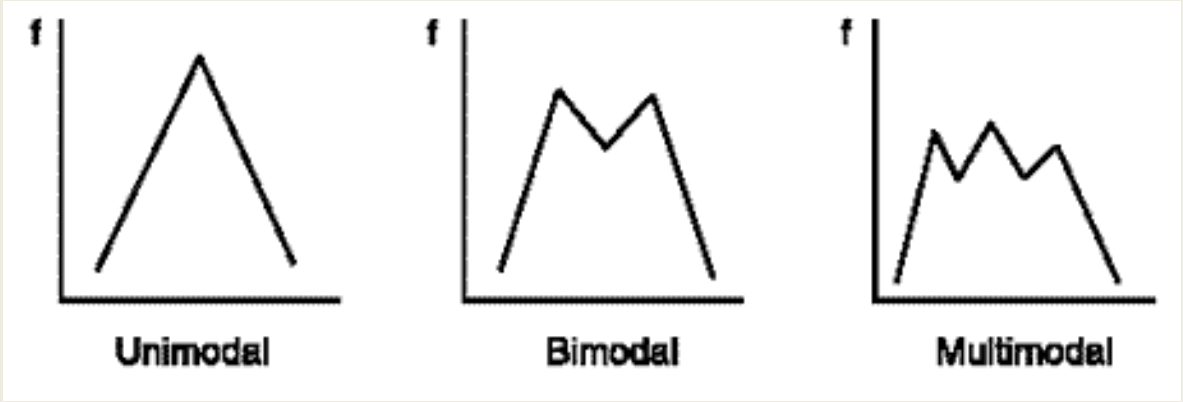
i. **Skewness**: Lack of symmetry

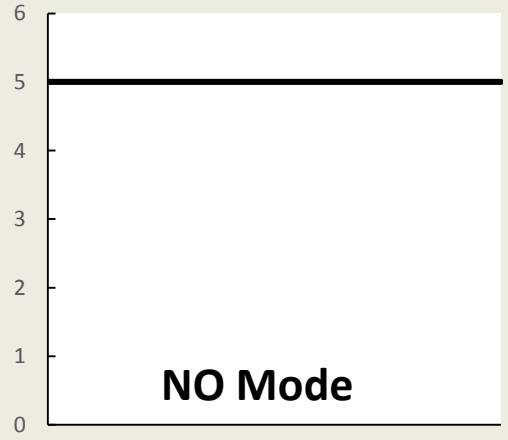
Left skew: long tail on the left
mean < median



Right skew: long tail on the right
mean > median

ii. **Number of modes**





1.1.4 Describing Distributions

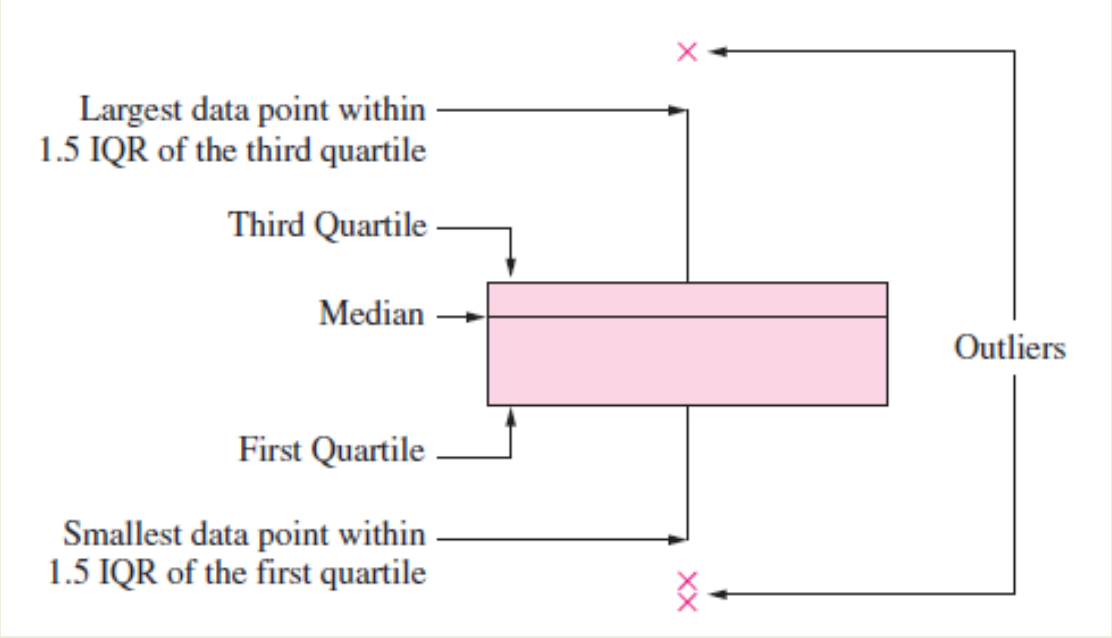
iii. **Outlier**: An observation that is considered to be unusually far from the bulk of the data. It may be due to experimental error; the point is sometimes excluded from the data set

- an observation maybe an outlier if it falls more than $1.5IQR$ above Q_3 or below Q_1.

1.1.4 Box-and-whisker Plot

A Box-and-whisker plot is a graphic that presents the median, the Q_1 and Q_3, and any (potential) outliers that are present in a sample.

The visual information in the box-and-whisker plot or box plot is **not intended** to be a formal test for outliers. Rather, it is viewed as a diagnostic tool.

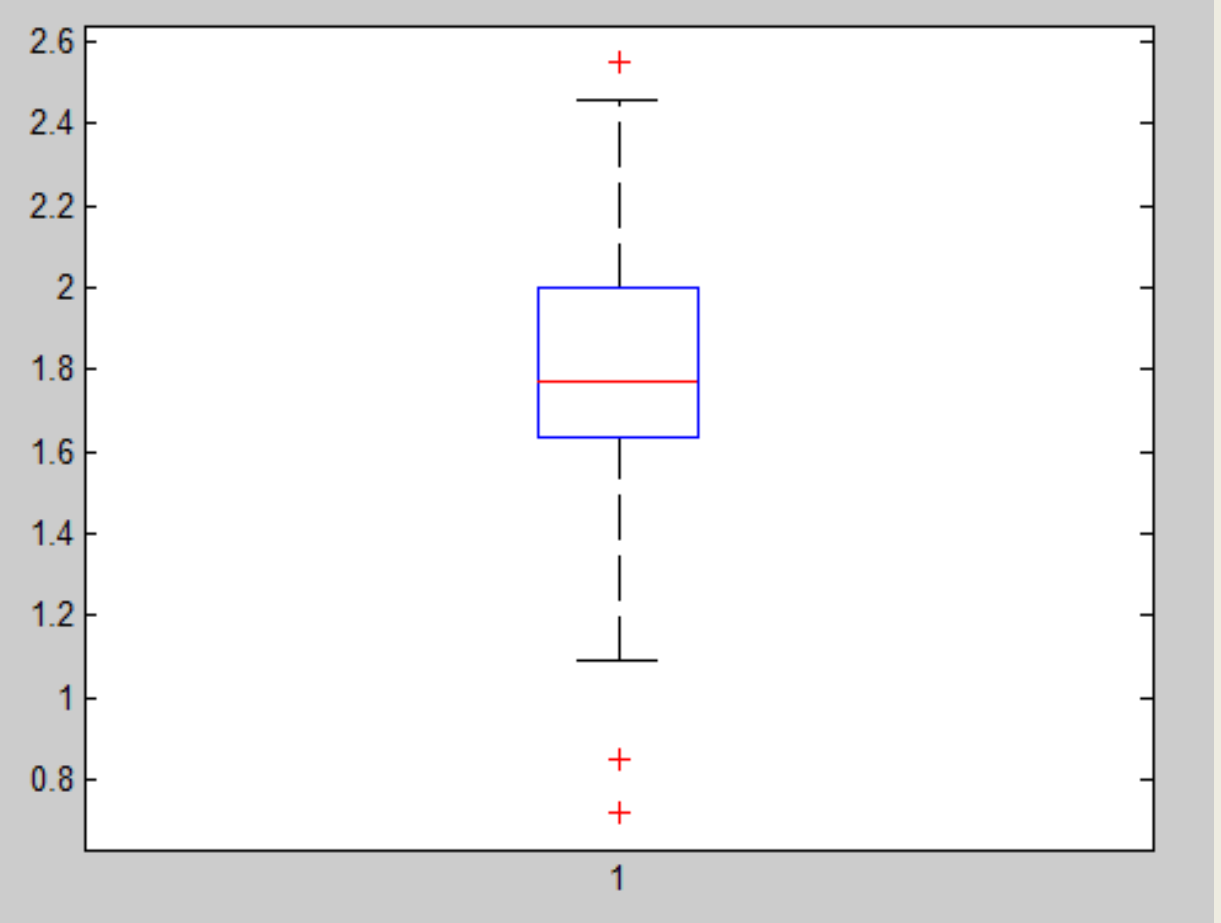


1.1.4 Box-and-whisker Implementation

Can be plotted with Matlab, R, Excel (new version) and more

Is this distribution skewed?

Do you see outliers?



Chapter 1.1 Summary

- Four scales of measurements
  - nominal, ordinal, interval and ratio
- Measures of central tendency and spread
  - Mean, median and mode + quartiles
  - Variance, range, IQR
- Graphs:
  - Qualitative: Barplots and piecharts (and more)
  - Quantitative: stem-and-leaf, histograms, scatter, timeseries (and more)
- How to identify observations that are possible outliers
  - boxplots