Curtin College

DIPLOMA OF ENGINEERING

LINEAR ALGEBRA & STATISTICS EMTH1019
WEEK 1 – STATISTICS DATA HANDLING

Curtin College

TEACHING STAFF

Unit Coordinator

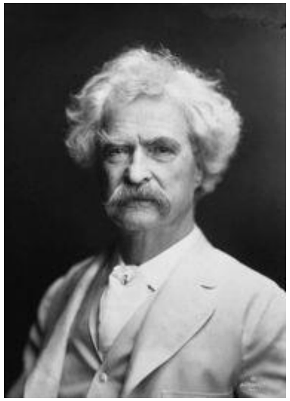
- Mary Jane (MJ) O’Callaghan
    - moca@study.curtincollege.edu.au
    - Appointments by arrangement

Curtin College

Statistics Week 1

STATISTICS

“There are three types of lies -- lies, damn lies, and statistics.”
— Benjamin Disraeli



“Facts are stubborn things, but statistics are pliable.”
— Mark Twain

“A single death is a tragedy; a million deaths is a statistic.”
— Joseph Stalin

Source https://www.goodreads.com/quotes/tag/statistics

STATISTICS- WEEK 1

- Overview
- Infographics
- Histograms
- Measures of central tendency
- Measures of dispersion
- Five number summary
- Box plot and outliers

WHAT IS STATISTICS?

The science of collecting, describing and interpreting data

Descriptive statistics

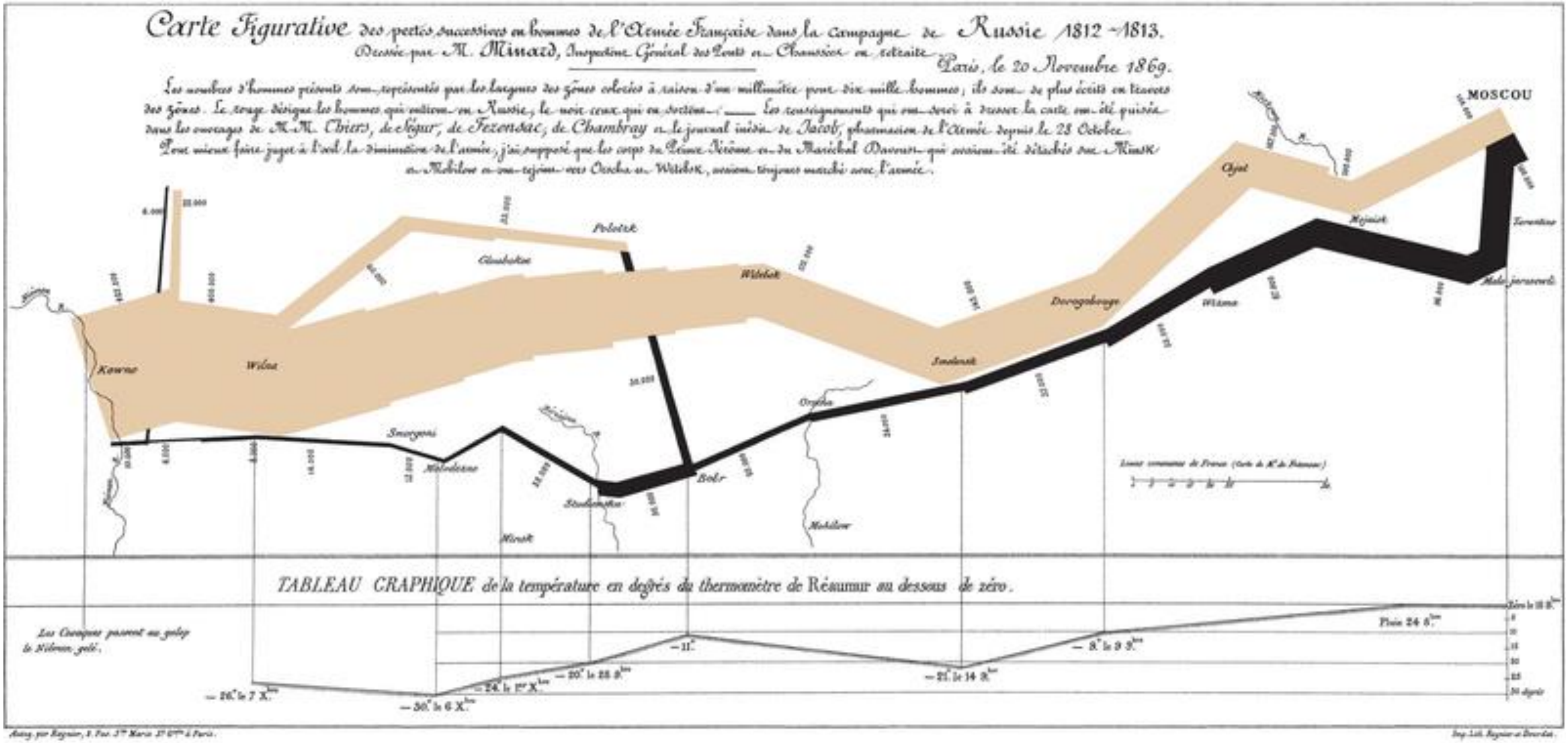
- Collection, presentation and description of data

Inferential Statistics

- **Interpreting** the values resulting from **descriptive** techniques and **drawing conclusions** about a population of data

CHARLES JOSEPH MINARD (1781- 1870)

This infographic of Napoleon’s failed 1812 invasion of Russia combines: troop numbers, distance, temperature, latitude & longitude, direction of travel and location relative to specific dates.



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

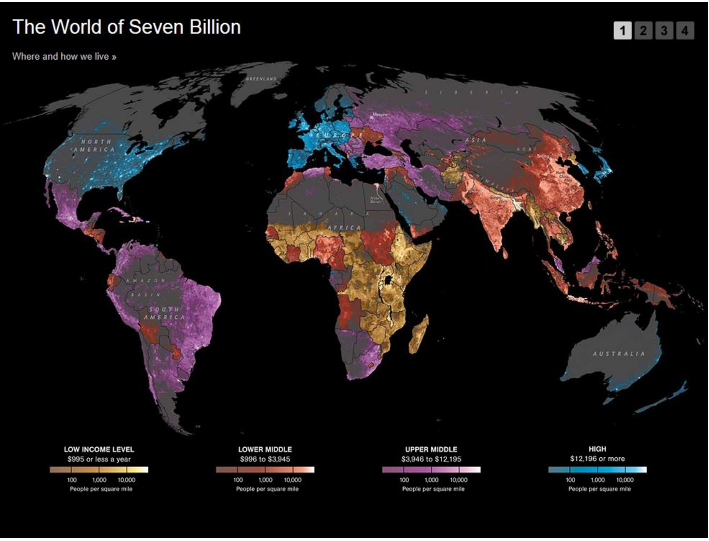
Les Cosaques passent au galop le Niémen gelé.

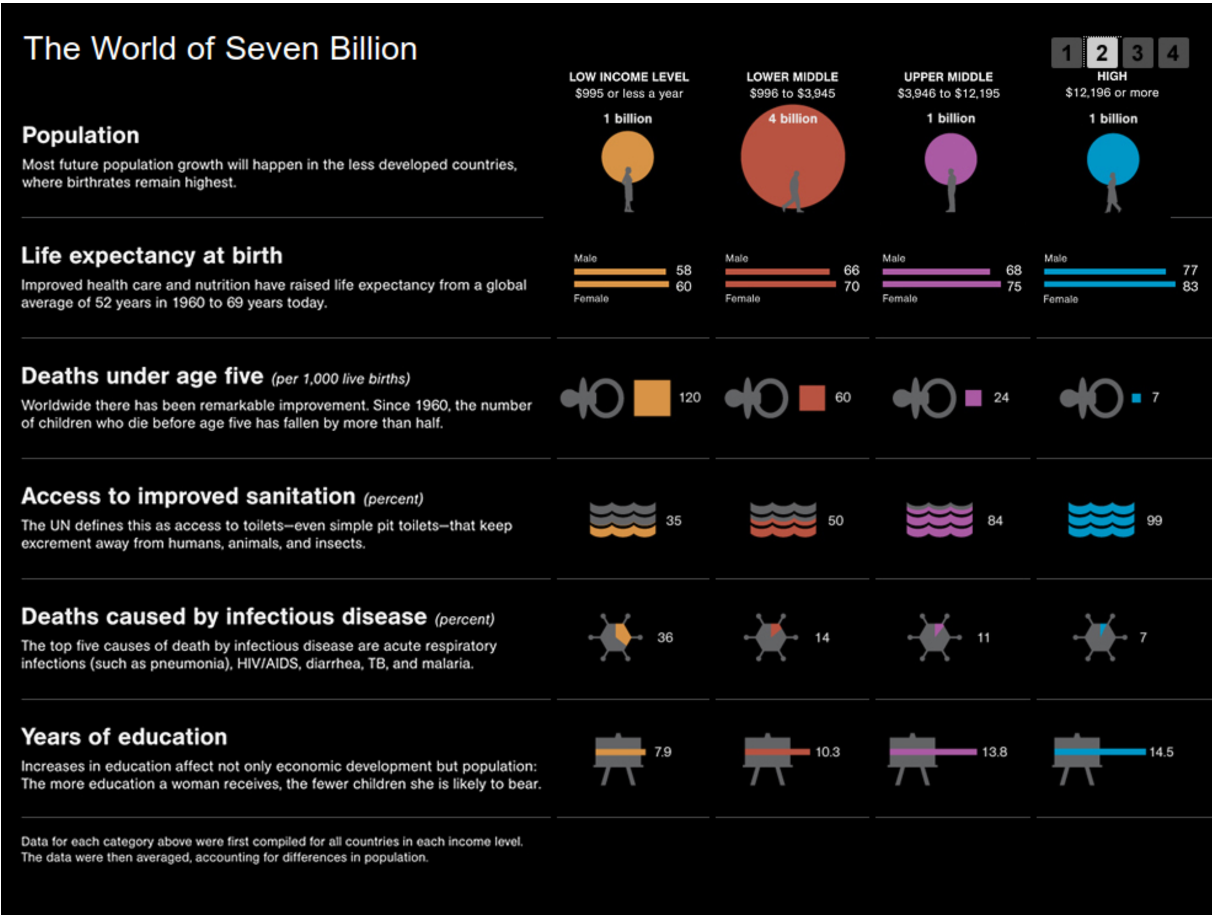
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

Imp. Lith. Regnier et Dourdet.

Curtin College

THE WORLD OF SEVEN BILLION





The World of Seven Billion

1 2 3 4

	LOW INCOME LEVEL $995 or less a year	LOWER MIDDLE $996 to $3,945	UPPER MIDDLE $3,946 to $12,195	HIGH $12,196 or more
Population Most future population growth will happen in the less developed countries, where birthrates remain highest.	1 billion	4 billion	1 billion	1 billion
Life expectancy at birth Improved health care and nutrition have raised life expectancy from a global average of 52 years in 1960 to 69 years today.	Male 58 Female 60	Male 66 Female 70	Male 68 Female 75	Male 77 Female 83
Deaths under age five *(per 1,000 live births)* Worldwide there has been remarkable improvement. Since 1960, the number of children who die before age five has fallen by more than half.	120	60	24	7
Access to improved sanitation *(percent)* The UN defines this as access to toilets—even simple pit toilets—that keep excrement away from humans, animals, and insects.	35	50	84	99
Deaths caused by infectious disease *(percent)* The top five causes of death by infectious disease are acute respiratory infections (such as pneumonia), HIV/AIDS, diarrhea, TB, and malaria.	36	14	11	7
Years of education Increases in education affect not only economic development but population: The more education a woman receives, the fewer children she is likely to bear.	7.9	10.3	13.8	14.5

Data for each category above were first compiled for all countries in each income level. The data were then averaged, accounting for differences in population.

https://www.google.com.au/search?q=the+world+of+seven+billion&tbm=isch&source=iu&ictx=1&fir=3bDMEc2gT9mS2M%253A%252COH2G3-p9I2V2uM%252C_&usg=__8RjNNWjH1LjfKdcADhXT_tMvhjc%3D&sa=X&ved=0ahUKEwiNo_vVr6XcAhXFGJQKHX-YCqQQ9QEIPjAH#imgrc=3bDMEc2gT9mS2M:

STATISTICAL REASONING

Concrete

- Comparing the effect on concrete strength of various combinations of drying times and amount of aggregate
- Need to have a plan and experimental method
- Use **inferential** statistics to analyse which combination of drying time and amount of aggregate is optimal.
- **How many samples of each mix will you need?**

Quality control

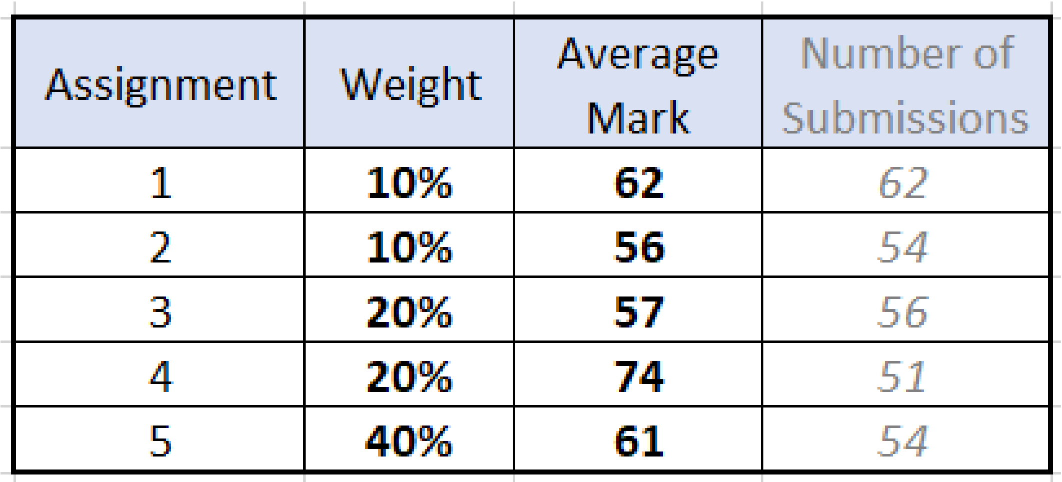
- Periodically sample a few chips from a production and count the number of defectives
- Use statistics to decide if failure rate is within acceptable standards
- **What is an acceptable failure rate? Who decides this?**

SIMPSON’S PARADOX

Numbers and “logical” calculations do not necessarily give the correct answer.

Below are the average marks for each assignment in a course, with weightings and the number of submissions.

Determine the average final mark?



Assignment	Weight	Average Mark	Number of Submissions
1	**10%**	**62**	*62*
2	**10%**	**56**	*54*
3	**20%**	**57**	*56*
4	**20%**	**74**	*51*
5	**40%**	**61**	*54*

WAS THIS YOUR CALCULATION?

Was this your final mark calculation?

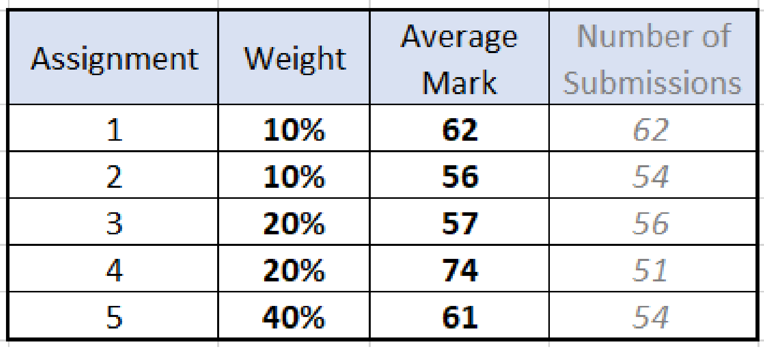
Final Mark Calculation = 0.1*62+ 0.1*56 + 0.2*57 +0.2*74 + 61*0.4

= 6.2 + 5.6 +11.4 + 14.8 + 24.4 = **62.4**

Estimated average final mark = 62

The ACTUAL average final mark was… 55

What is wrong with the calculation?



Assignment	Weight	Average Mark	Number of Submissions
1	**10%**	**62**	*62*
2	**10%**	**56**	*54*
3	**20%**	**57**	*56*
4	**20%**	**74**	*51*
5	**40%**	**61**	*54*

AVERAGES HIDE VARIATION

In this calculation you probably assumed that the population for each assessment was the same, **but whilst similar the populations were not identical.**

- In each assessment the number of submissions varied and who submitted varied.
- 16 people missed submissions but only 6 people missed 1 submission
- The average mark per assessment does not include non-submissions.
  - Should I have given all the non - submissions zero and then calculated the average mark for each assessment?
- The final mark for each student does include non-submissions.

What assumptions did you make about the data when you did your calculation?

Did you have enough information?

KEY STATISTICAL ISSUES AND QUESTIONS

What do you want to know?

What question(s) are you asking?

How was the data collected?

How do we an analyse it?

What test do we use?

What information does it give us?

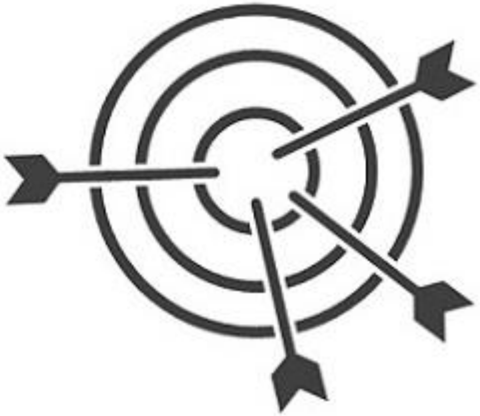
What conclusions can we draw?

Curtin College

JARGON - PRECISION AND ACCURACY

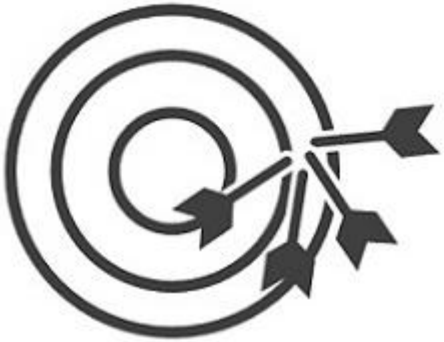
Accuracy

- The degree of closeness of an observation to its true value
- Correctness



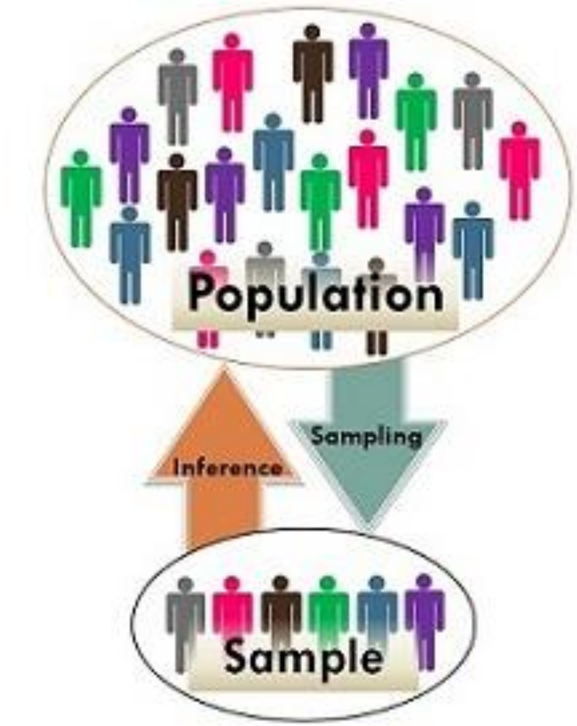
Precision

- The degree of closeness between multiple measures
- Exactness



TERMINOLGY (MORE JARGON)

- **Population** – the **entire** set of observations that can be made.
- **Sample** – a **subset** of a population
- **Statistic** – a fact or piece of data obtained from a study of a large quantity of numerical data. e.g. average income, hottest day
- **Variable** – a characteristic, number quantity that can be measured or counted e.g. income, rainfall, age…



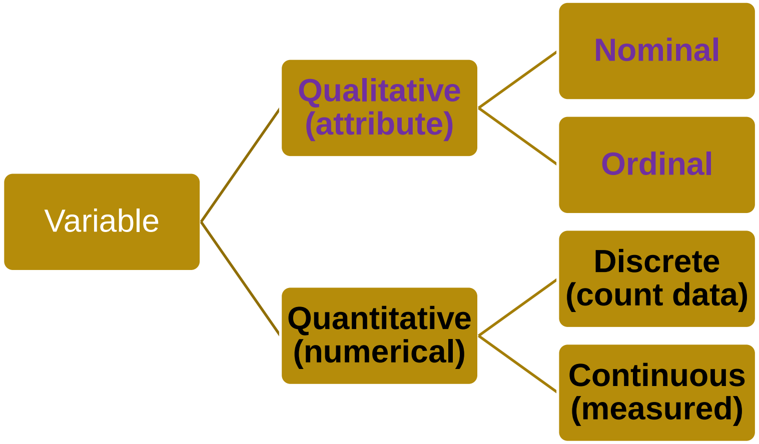
https://www.google.com.au/search?biw=1366&bih=635&tbm=isch&sa=1&ei=WY9NW_W3HqymmAXfk5CQAQ&q=sample+versus+population&oq=sample+versus+population&gs_l=img.3..0i30k1j0i8i30k1j0i24k1j0i10i24k1j0i24k1l3.19384.26359.0.26793.24.23.0.0.0.0.304.1805.8j0j3j2.13.0....0...1c.1.64.img..11.13.1798...0j0i67k1.0.J7OF_Ziu6e0#imgrc=Er8URn5drFfV8M:

PROCESSING DATA

The best data is so obvious you don't need to analyse it.

- Unfortunately raw data is not always informative.
  - The raw data still needs to be available,
- Statistics **aims** to get meaningful information from data,
  - But having a statistic does not make it meaningful!
- You can present data numerically and graphically.
  - You should do both.
- The conclusion has to be consistent with the objectives as well as the limitations of how you collected the data.

DATA TYPES


Variable

Qualitative (attribute)

Nominal

Ordinal

Quantitative (numerical)

Discrete (count data)

Continuous (measured)

Quantitative

- **Continuous** – weight, length 1.1235813213455 kg
- **Discrete** – number of students (no fractional students!)

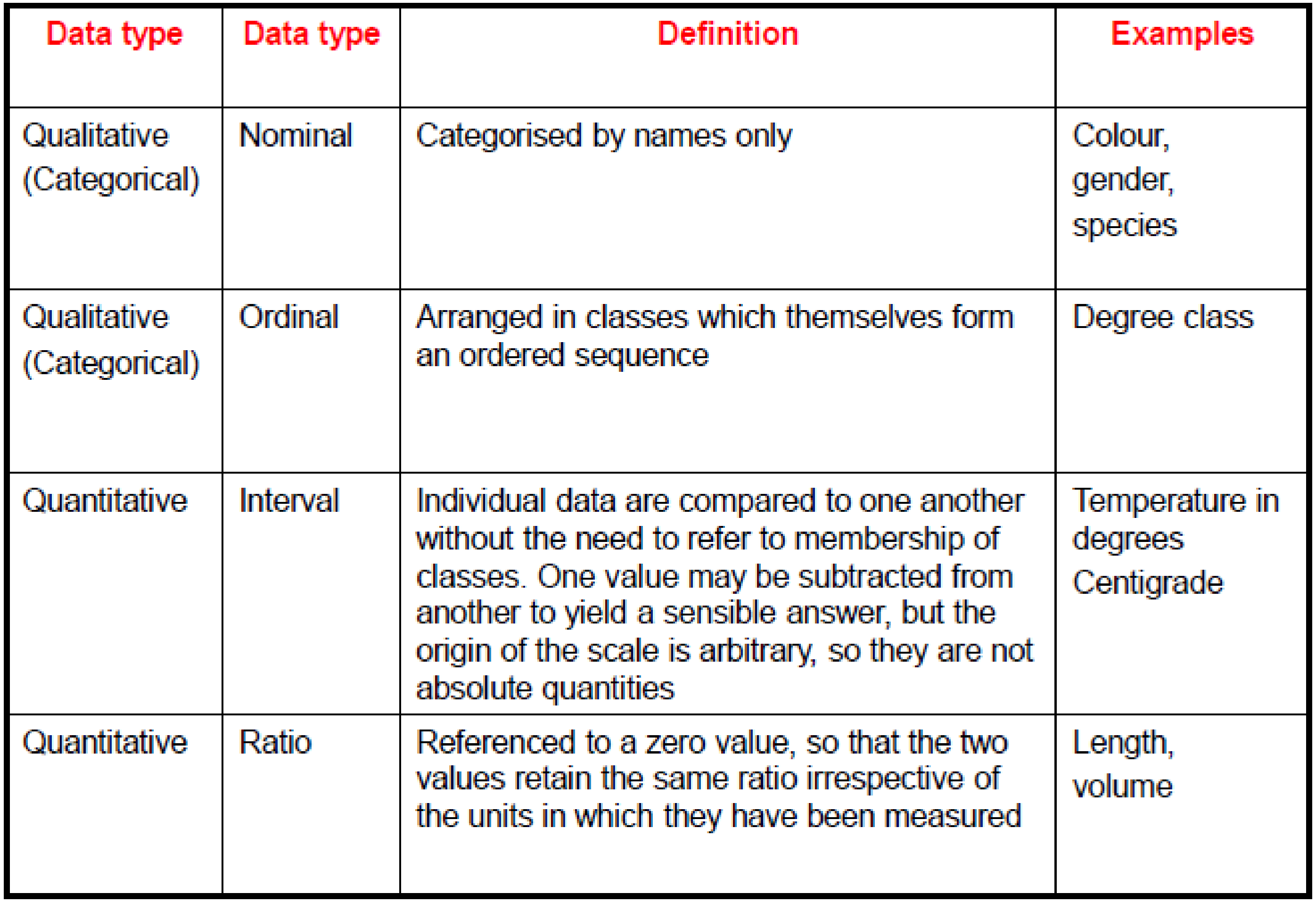
https://www.theguardian.com/politics/2013/nov/25/mapping-redheads-which-country-has-the-most

Qualitative

- **Nominal** - categorise by name e.g. Hair colour, gender…Mapping redheads: which country has the most?

- **Ordinal** – arranged in natural ordered categories but the distances between categories is not known e.g. job satisfaction,

DATA TYPES



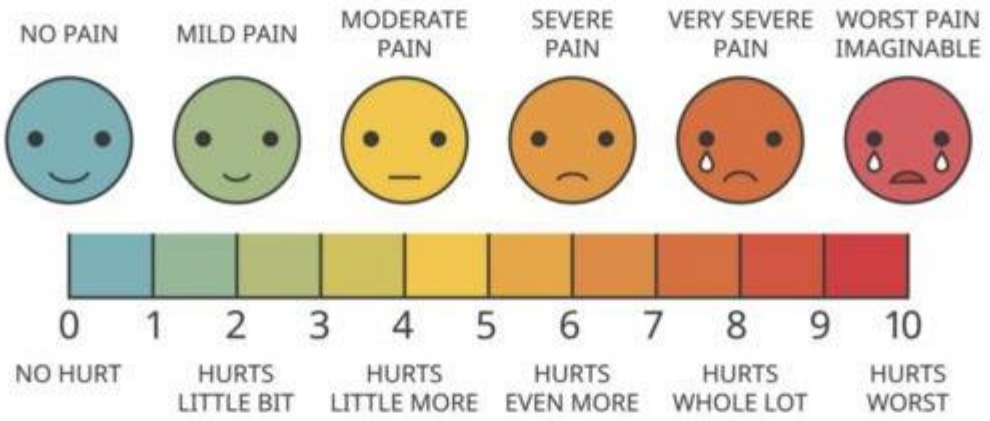
Data type	Data type	Definition	Examples
Qualitative (Categorical)	Nominal	Categorised by names only	Colour, gender, species
Qualitative (Categorical)	Ordinal	Arranged in classes which themselves form an ordered sequence	Degree class
Quantitative	Interval	Individual data are compared to one another without the need to refer to membership of classes. One value may be subtracted from another to yield a sensible answer, but the origin of the scale is arbitrary, so they are not absolute quantities	Temperature in degrees Centigrade
Quantitative	Ratio	Referenced to a zero value, so that the two values retain the same ratio irrespective of the units in which they have been measured	Length, volume

Curtin College

LIKERT SCALE

This scale below is used by doctors to measure the amount of pain a patient is feeling.

Is this data quantitative or qualitative?



CAN YOU USE STATISTICS TO ANALYZE ORDINAL DATA THAT HAS NUMBERS?

Sometimes but be very, very, very cautious.

We will also be referring to the documents below in later weeks and they are available under the week 1 lecture.

- Berg Balance Scale with instructions Berg_balance_scale_with_instructions.pdf (physio-pedia.com)
- Hackney, M. E., & Earhart, G. M. (2010). Effects of dance on gait and balance in Parkinson's disease: a comparison of partnered and nonpartnered dance movement. *Neurorehabilitation and neural repair*, *24*(4), 384–392. https://doi.org/10.1177/1545968309353329

QUESTIONS FOR YOU TO ANSWER

What is Parkinson Disease?

What is the Berg Balance Scale?

- Who uses this data and how?
- Is it measured?
- Is it qualitative, quantitative or something else?

In the paper by *Hackney and Earhart* they refer to:

- Power calculations
- Effect size
- *“Data was analyzed using Sigmastat software (Systat, Richmond, VA)*”

What do these things mean?

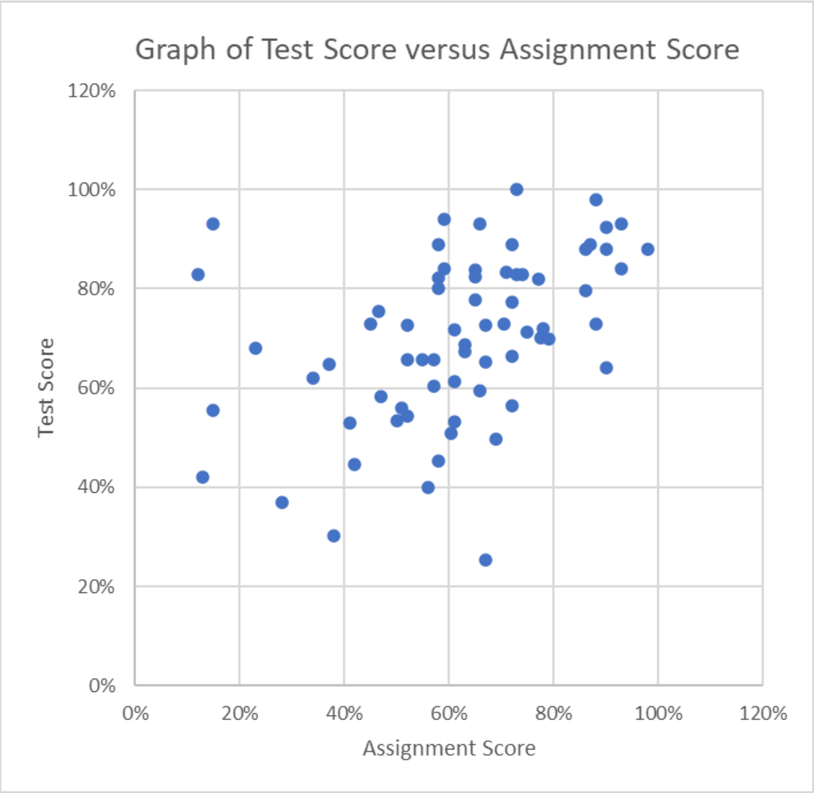
- Where is the raw data?

DON'T IGNORE QUALITATIVE

The important things in life cannot always be measured or counted.

SIMPLE GRAPH: WHAT SHOULD I DO NEXT?

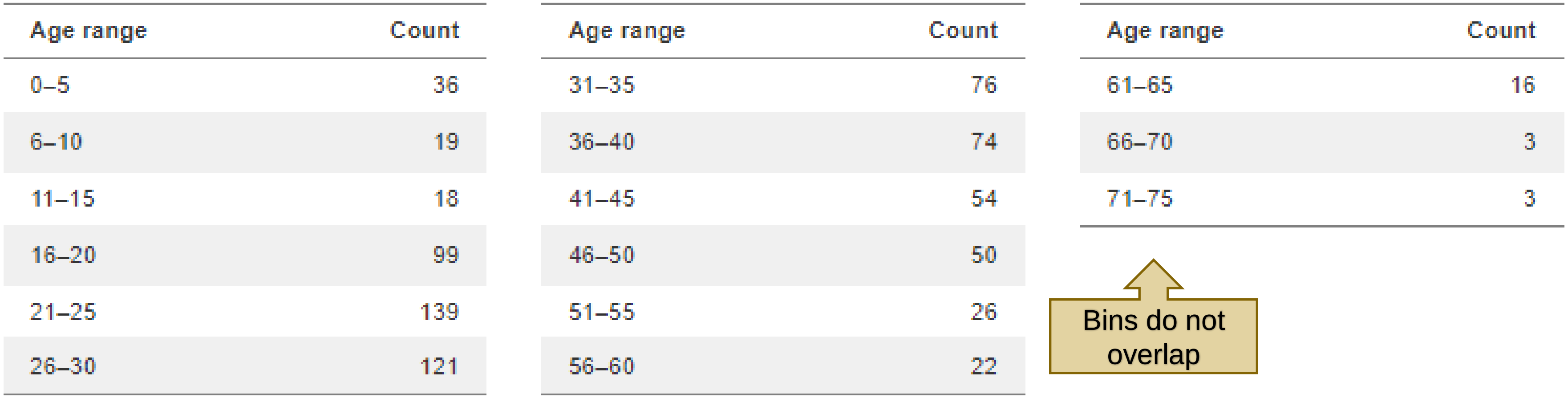
I have just finished marking an assessment that had 2 parts: a test and an assignment on the same topic. Each dot represents a student's results.



TITANIC

On 14 April 1912 the Titanic sank with ~ 1300 passengers (not including the crew ~908) on the Titanic with reported ages for 756 of them.

Table: Number of passengers with known age on the Titanic.



Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3

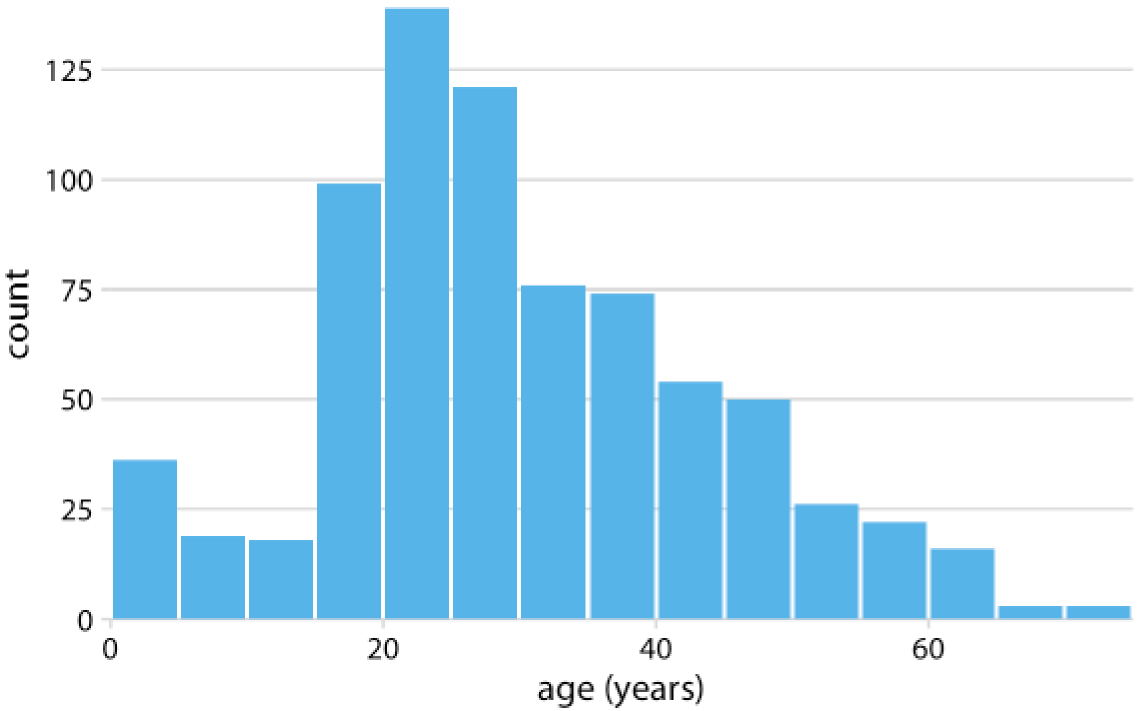
Bins do not overlap

Bin width = 5

Wilke C.O., Fundamentals of Data Visualization, 7.1 Visualizing a single distribution. https://clauswilke.com/dataviz/

TITANIC PASSENGER HISTOGRAM

Figure: Histogram of the ages of Titanic passengers



Wilke C.O., Fundamentals of Data Visualization, 7.1 Visualizing a single distribution. https://clauswilke.com/dataviz/

BIN WIDTH EFFECTS APPEARANCE OF HISTOGRAM

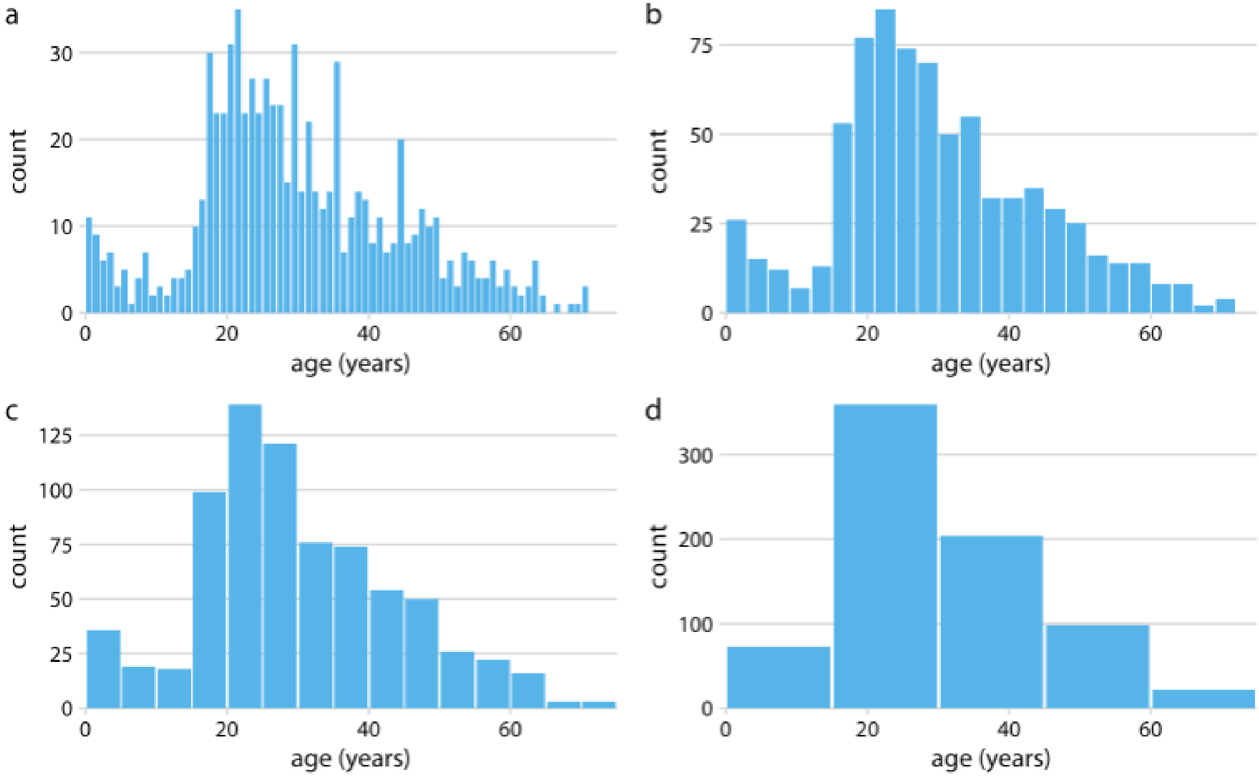
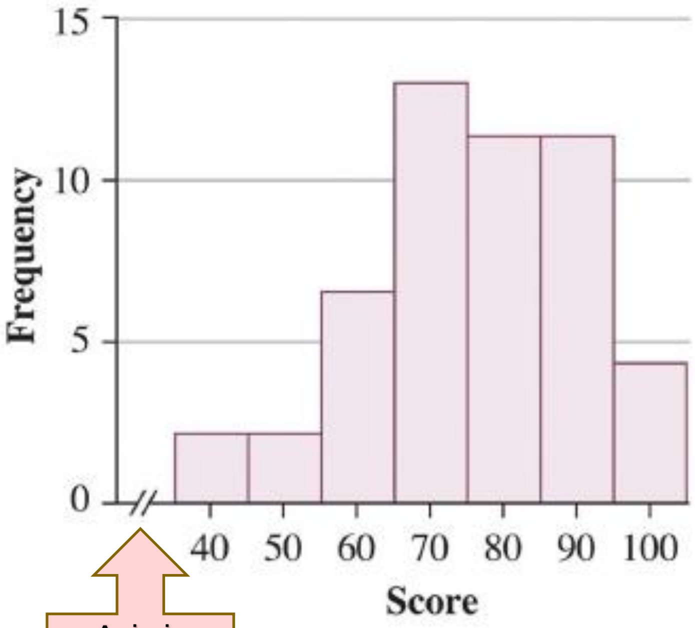


Figure 7.2: Histograms depend on the chosen bin width. Here, the same age distribution of Titanic passengers is shown with four different bin widths: (a) one year; (b) three years; (c) five years; (d) fifteen years.

Wilke C.O., Fundamentals of Data Visualization, 7.1 Visualizing a single distribution. https://clauswilke.com/dataviz/

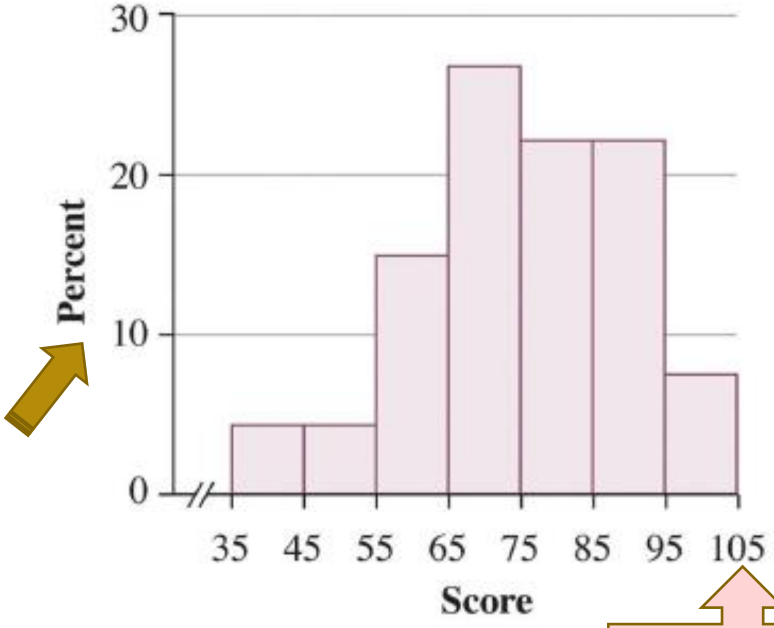
HISTOGRAMS TYPES

Frequency Histogram
50 final exam scores in Elementary Statistics



Axis is cut. Is this good or bad?

Relative Frequency Histogram
50 final exam scores in Elementary Statistics
Relative frequency = (class frequency) / (total of all frequencies)



Do the numbers make sense?

THINGS TO CONSIDER WHEN CREATING A HISTOGRAM

Context of the data

Sample size

Bin widths

- Explore different widths.

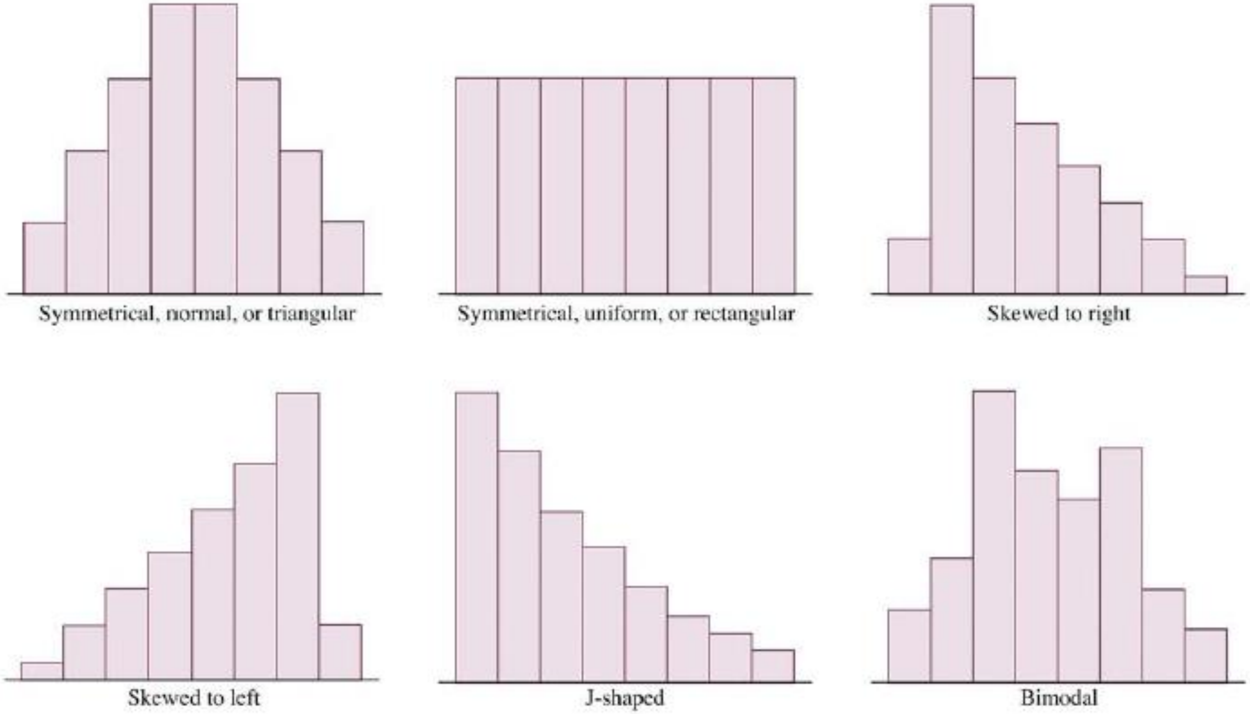
Frequency table

- Bin groupings do not overlap.
- Number of data points in each bin.

Graph

- Frequency (number of) or Relative Frequency (%) or both?
- Horizontal scale.
- Titles and legends.

SOME SHAPES OF HISTOGRAMS



Symmetrical, normal, or triangular

Symmetrical, uniform, or rectangular

Skewed to right

Skewed to left

J-shaped

Bimodal

MEASURES OF CENTRAL TENDENCY

Values that locate, in some way, the centre of a set of data

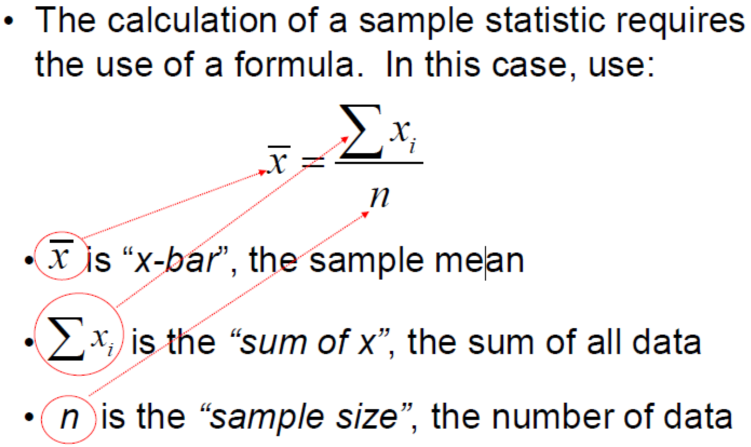
Given a set of data consisting of five values: 6, 3, 8, 6, 4

Determine the:

- Sample mean $\bar{x}$ (average value)
- Mode (most common value)
- Median (middle value)

SAMPLE MEAN FORMULA

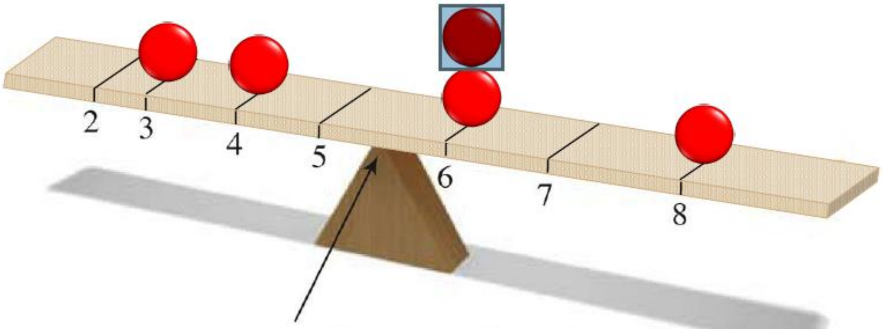
Even though this is a simple formula you need to get used to the notation.



- The calculation of a sample statistic requires the use of a formula. In this case, use:

$$\bar{x} = \frac{\sum x_i}{n}$$

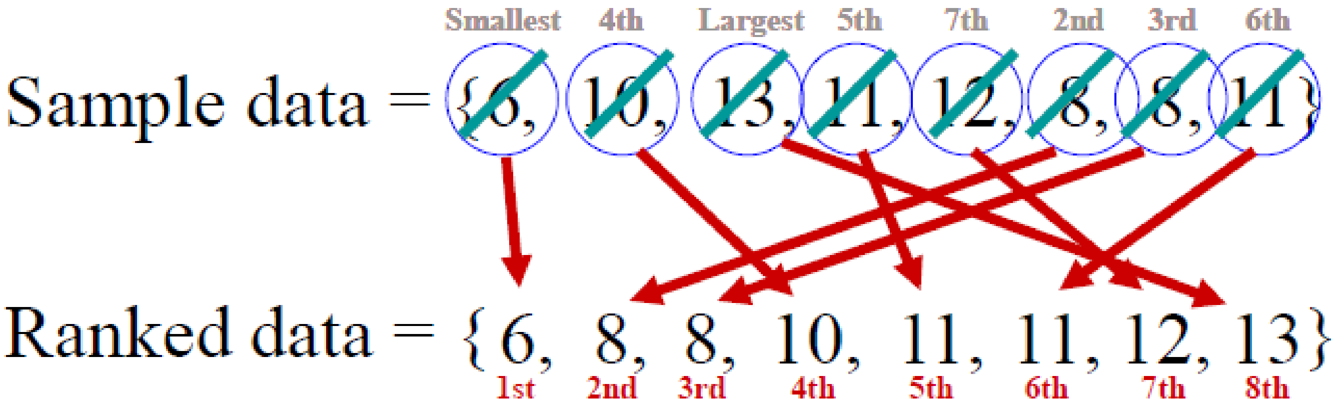
- $\bar{x}$ is "*x-bar*", the sample mean
- $\sum x_i$ is the *"sum of x"*, the sum of all data
- n is the *"sample size"*, the number of data



$\bar{x} = 5.4$ (the center of gravity, or balance point)

RANKED OR ORDERED DATA

- Since the median is the “middle value”, the data must first be ranked in order of value
- Typically, ranking is smallest value first and largest value last:



Sample data = {6, 10, 13, 11, 12, 8, 8, 11}

Ranked data = {6, 8, 8, 10, 11, 11, 12, 13}

SAMPLE MEDIAN FORMULA

1. Order the data from smallest to largest
2. For an odd number of data values in the distribution,

   Median = Middle data value

3. For an even number of data values in the distribution,

$$\text{Median} = \frac{\text{Sum of middle two values}}{2}$$

- The position of the middle value (or depth) of the median is determined using the formula

$$\text{Position of middle value} = \frac{n+1}{2}$$

SOLUTION TO PROBLEM

Data $\{6, 3, 8, 6, 4\}$

Number of data points $n = 5$

Sample mean $\bar{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5.4$

Median

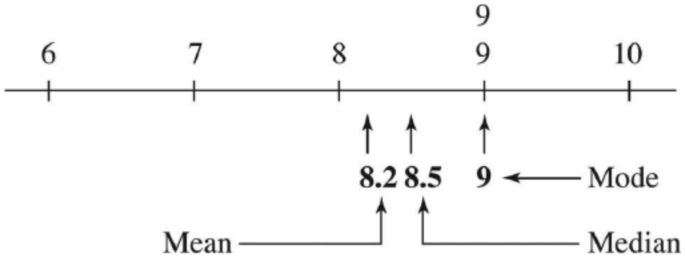
- Ranked data {3, 4, 6, 6, 8}
- n=5
- Median = 6

Mode = 6 (it occurs twice the other values only occur once)

QUESTIONS

- Calculate the measures of central tendency (mean, mode and median) for {6, 7, 8, 9, 9, 10}
- Plot and label these statistics on a number line.
  - Even though mean, mode and median are all measures of centre they do not always have the same value.

Measures of Central Tendency for {6,7,8,9,9,10)



WORLD CENTRAL TENDENCIES DATA

Global Wealth Report 2018

	Mean Wealth US$	Median Wealth US$
Australia	411,060	191,453
China	47,810	16,333
United Kingdom	279,048	97,169
India	214,893	35,169
USA	403,974	61,667

https://www.credit-suisse.com/corporate/en/research/research-institute/global-wealth-report.html

What does this tell us?

The average wealth for the USA is large but the median is 15% of the average.

If USA had a population of 100 people with the same median:

- 50 people have wealth < $61,667
- 50 people have wealth > $61,667
- The extreme wealth of a few has an enormous impact on the mean wealth.

The median wealth in Australia is 46% of the average.

ELON MUSK WALKS IN TO A CAFÉ

Elon Musk has wealth of ~ US$224 billion (US$224,000,000,000).

Top 10 Richest People in the World (June 2022) (investopedia.com)

In 2020 the:

- USA Mean wealth US$505,420 (Number 2 in the world)
- USA Median wealth US$79,274 (Number 23 in the world)

Example

There are 5 people in a café in they all have wealth of $505,420.

- The mean and the median are the same.

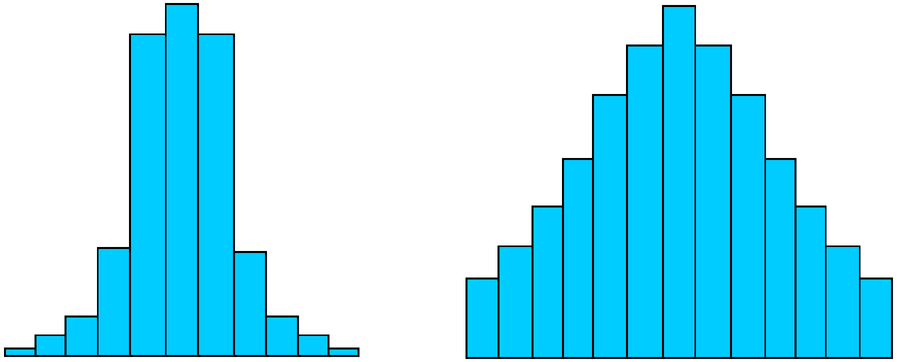
One person leaves but Elon Musk walks into to get a skinny soy latte.

- The median wealth is still $79,274.
- The mean wealth is now $44,800,404,336 (or $$44.8 * 10^9$)

MEASURES OF DISPERSION

Descriptive statistics that:

- Measure the spread of the data
- Can give us a feel for how much variation there is in the data
- Lets us know if values are clustered close to the mean or median, or spread out.



- https://www.universalclass.com/articles/math/statistics/understanding-measures-of-dispersion-in-statistics.htm

SAMPLE VARIANCE & SAMPLE STANDARD DEVIATION

***S*²** is called the **Sample** Variance

S is called the **Sample** Standard Deviation

Notice that the denominator is *n-1*

- When we calculate the variance of the entire population we divide by n
- When we calculate the variance of a sample (very important distinction) we divide by n-1 as we do not want to under estimate the variance

Why is there a √ ?

- In the ***s*²** calculation the differences between the sample mean and the data point are squared. By taking the square root means that ***s*** has the same units as the data.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

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

EXAMPLE

- **Given**: The times, in seconds, required for a sample of students to perform a required task were:

$$6, \quad 10, \quad 13, \quad 11, \quad 12, \quad 8$$

- **Find**: **a)** The sample variance, s^2

  b) The sample standard deviation, s

- Sample size n = 6
- The mean $\bar{x} = 10$

(a) Sample Variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(6-10)^2 + (10-10)^2 + (13-10)^2 + (11-10)^2 + (12-10)^2 + (8-10)^2}{6-1}$$

$$= \frac{34}{5} = 6.8$$

(Variance has no unit of measure, it's a number only)

(b) Standard deviation is

$$s = \sqrt{s^2} = \sqrt{6.8} = 2.60768 = 2.6 \text{ seconds}$$

ADVICE

The standard deviation calculation is fiddly to do by hand.

Learn how to use the statistical functions on your calculator.

- Then when you are doing these sorts of calculations in an assignment I expect you to:
  - Write the correct formula,
  - Use correct notation,
  - State key values: $n\ and, \bar{x}$
  - Show sufficient logical working.
  - Summarise your answers. The marker shouldn't have to search for the answer.

The correct number is meaningless if you do not have sufficient logical working that supports the number.

DO STATISTICS TELL US WHAT THE RAW DATA LOOKS LIKE?

Sometimes yes, sometimes no.

If I am comparing 13 sets of data of the form (x, y) and they ALL have:

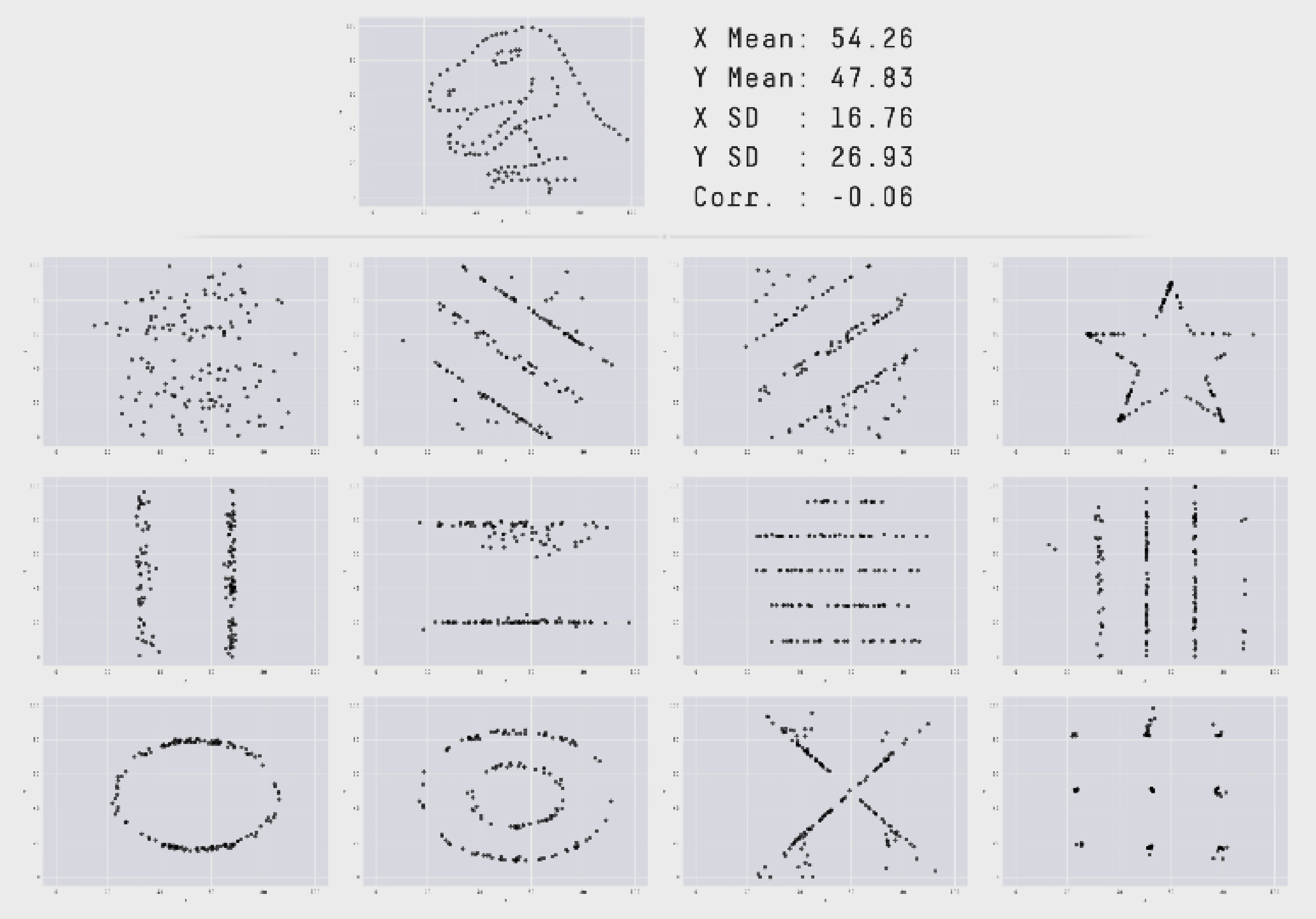
$$\bar{x} = 54.3, \bar{y} = 47.8, s_x = 16.8, s_y = 26.9$$
$$Pearson\text{'}s\ Correlation\ =\ -0.06$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Will the graphs of $x\ versus\ y$ look the same for each of the 13 data sets?

THE DATASAURUS DOZEN

HTTPS://WWW.AUTODESK.COM/RESEARCH/PUBLICATIONS/SAME-STATS-DIFFERENT-GRAPHS



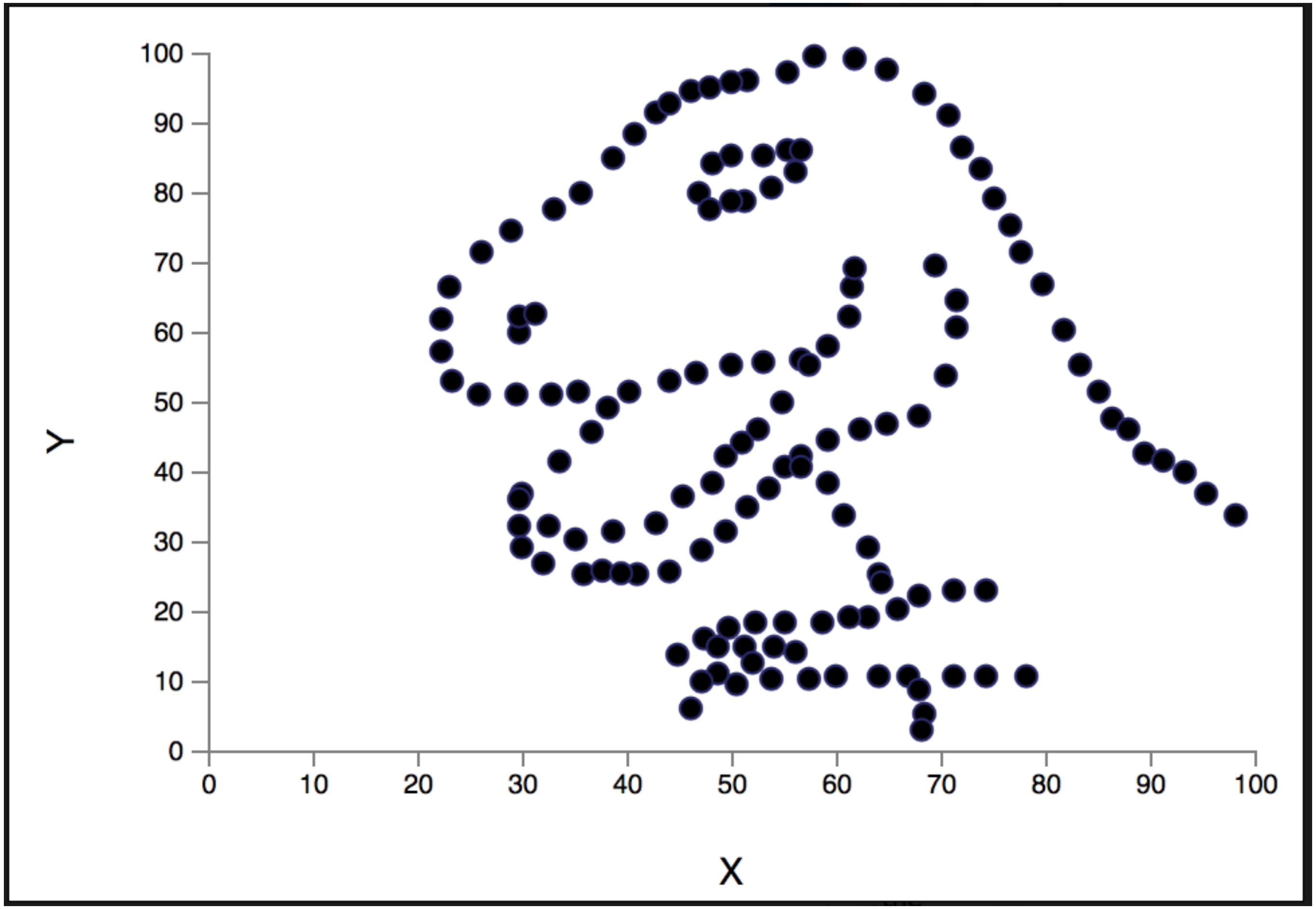
```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

“NEVER TRUST SUMMARY STATISTICS ALONE; ALWAYS VISUALIZE YOUR DATA” *ALBERTO CAIRO THE CREATOR OF DATASAURUS*

- http://www.robertgrantstats.co.uk/drawmydata.html Draw my data app.
- https://www.autodesk.com/research/publications/same-stats-different-graphs Source of the Datasaurus dozen.
- http://www.thefunctionalart.com/ Alberto Cairo’s website
- https://clauswilke.com/dataviz/ Fundamentals of Data Visualization. This is an amazing online text.

Curtin College

REMEMBER TO VISUALIZE YOUR DATA



RESISTANT & NOT RESISTANT MEASURES.

A Resistant Measure is one not affected by outliers or highly skewed data

When providing a numerical summary of data we should provided at least 2 values: a measure of centre and a measure of spread (or dispersion)

Resistant measures – use if data is skewed

- Use the Median and IQR (inter quartile range).

Not resistant measures – data is approximately symmetrical

- Use the Sample Mean and Sample Standard Deviation

PERCENTILES

Ordered data set {3,4,4,5,7,7,8,8,10,13,15,17}, n=12

You may have heard someone state the value of the 90th percentile, but do you know how to calculate it?

- Ordered data ✓
- **Location** of 90th percentile
  - $= \frac{90}{100} * (n + 1) = 0.9 * (12 + 1) = 11.7$
- **Value** of 90th percentile = **11th** value + 0.7*(difference 11th & 12th values)
  - $= 15 + 0.7 * (17 - 15) = 15 + 1.4 = 16.4$

Calculate the 70th percentile.

QUARTILES - ARE ALSO PERCENTILES

Given a sample of n observations

- Sort them in order

If $0 < p < 1$, then the $(100p)$ sample percentile has ~ np observations less than it, and $n(1 - p)$ sample observations greater than it.

- Q1 – the lower quartile or the 25th percentile.
  - One quarter of observations lie **below** Q1
- Q2 – the median or the 50th percentile
- Q3 – the upper quartile or 75th percentile
  - One quarter of observations lie **above** Q3
- IQR – Inter Quartile Range = Q3 – Q1
  - Half of observations lie within IQR

WHY IS MY CALCULATOR GIVING ME A DIFFERENT ANSWER FOR Q1 AND Q3?

The 25th percentile or P_{25} has the same meaning as $Q_{1,}$ 1st quarter.

- Quartiles are a subset of Percentiles
- There has been a trend to say Q_1 and Q_3 are either data points or half way between 2 data points; like the median.

Example

- If n = 20 the location of Q_1 is ¼*(20+1)=5.25. Some calculators round this up to 5.5 and then work out the halfway value.

Does it matter?

- If n=20 and I need to determine the 26th Percentile then the location is determined by 26/100*(20+1)= 5.46.
  - You now have a situation where the value of P_{26} at location 5.46 < Q_1 using the rounding method
  - **This is not mathematically consistent.**

But who cares about this little difference?

- I do.
- Court cases for billions are fought over what a number means.

FINDING Q1 AND Q3 FORMULAE

Ordered data set {3,4,5,7,7,8,8,10,13,15}, n=10

Position of $Q1$ $= \frac{n+1}{4} = \frac{10+1}{4} = 2.75$

Value of Q1

- The 2.75 means that the value of Q1 is 0.75 of the way between the 2nd and 3rd numbers in the ordered data set.
- $= 4 + 0.75(5-4) = 4 + 0.75 = \mathbf{4.75 = Q1}$

Position of $Q3$ $= \frac{3}{4}(n+1) = \frac{3}{4}(10+1) = \frac{33}{4} = 8.25$

Value of Q3

- The 8.25 means that the value of Q3 lies 0.25 of the way between the 8th and 9th numbers of the ordered data set.
- $= 10 + 0.25(13-10) = 10 + 0.75 = \mathbf{10.75 = Q3}$

START WITH THE LOCATION OF Q2

Ordered data set {3,4,4,5,7,7,8,8,10,13,15,17}, n=12

Position of median or $Q2$

- is halfway between 7 and 8 so this is **position 6.5**
- $= 7 + 0.5(8 - 7) = 7 + 0.5$

Q1

- If the position of Q2 is 6.5 then the position of Q1 $= \frac{6.5}{2} = 3.25$ = **position of Q1**

Value of Q1

- The 3.25 means that the value of Q1 is 0.25 of the way between the 3rd and 4th numbers in the ordered data set.
- $= 4 + 0.25(5 - 4) = 4 + 0.25 = \mathbf{4.25} = \boldsymbol{Q1}$ **value**

Q3

- Position of Q3 = Position of Q1 + Position of Q2 = 6.5 + 3.25 = 9.75 **Position of Q3**

Value of Q3

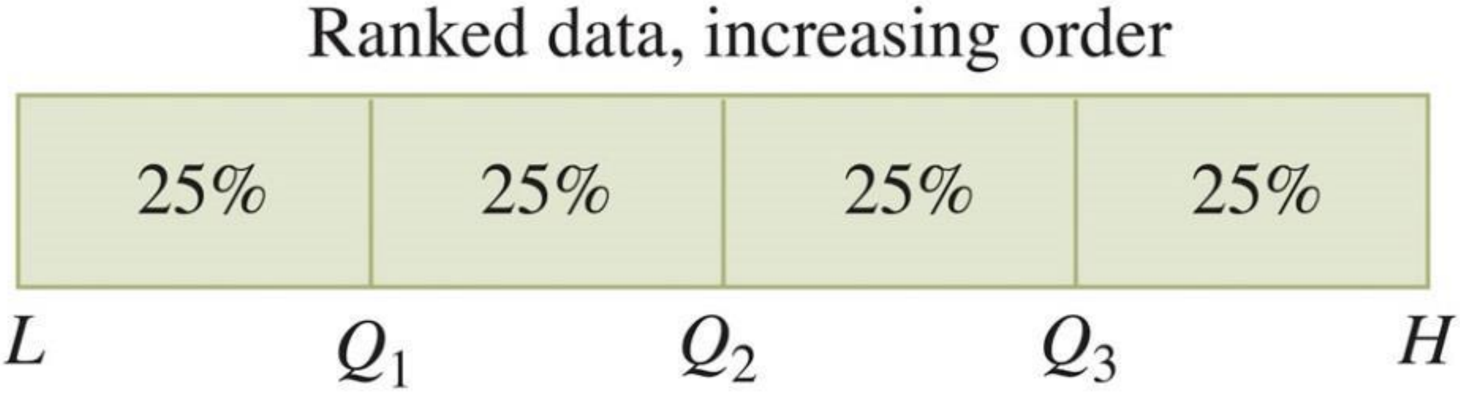
- The 9.75 means that the value of Q3 lies 0.75 of the way between the 9th and 10th numbers of the ordered data set.
- $= 10 + 0.75(13 - 10) = 10 + 2.25 = \mathbf{12.25 = Q3\ value}$

RANGE & INTERQUARTILE RANGE

Range = Maximum value – Minimum Value

IQR = Q3 –Q1

- IQR gives us a feel for how the middle 50% of data is spread out



5 NUMBER SUMMARY

Minimum, Q1, Median, Q3, Maximum

For the ordered data set {15,16,16,17,17,17,17,18,18,18,19,20,20,21}

n=14

Minimum = 15

Q1 position = (n+1)/4=15/4=3.75

Q1 value = 16+0.75(17-16)=16.75

Position of Median (Q2) = (n+1)*2/4= (n+1)/2=15/2=7.5

Value of Median (Q2)= 17+0.5(18-17)=17.5

Q3 position = ¾ *(n+1)= ¾ *15=11.25

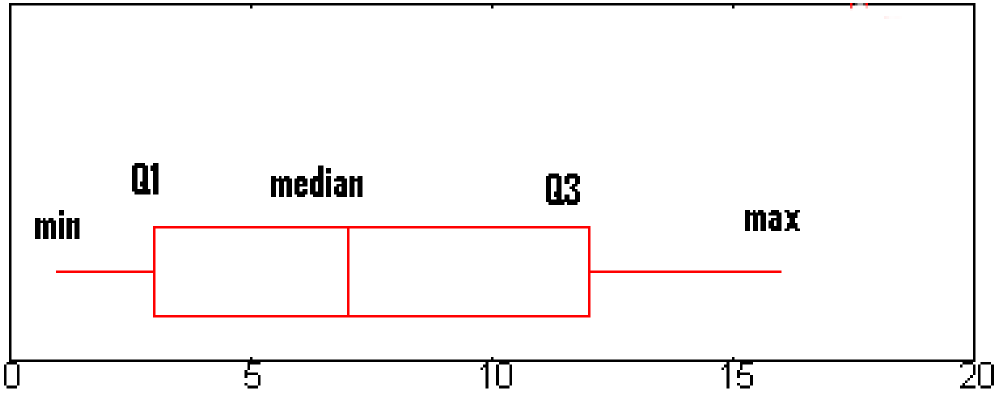
Q3 value = 19+0.25(20-19)=19.25

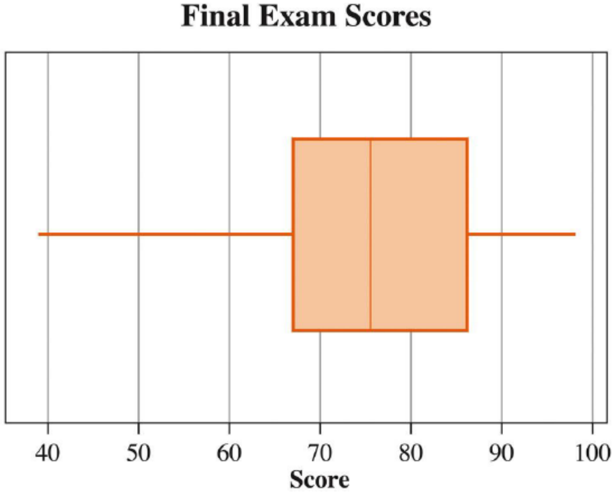
Maximum = 21

BOXPLOT

In a boxplot:

- Q1 and Q3 are the ends of the box
- The vertical line in the middle is the median (Q2)
- The "whiskers" on the box extend to the minimum and maximum values provided there are no *outliers*.





OUTLIERS

- Outliers are always at the extreme ends of any variable data set.
- Do not automatically remove outliers from your data set.
  - Keep or remove outliers? They could be valid observations or the result of an error in data collection or entry.
  - If you choose to remove outliers you must document why you removed each outlier.

In a Box Plot outliers are any observation with:

- Values $< Q1 - 1.5 * IQR$ or Values $> Q3 + 1.5 * IQR$.
  - Remember $IQR = Q3 - Q1$

In a boxplot

- Every outlier is marked with a star *
- The whiskers extend to the largest /smallest value in the data set that lies within $1.5 * IQR$ from the ends of the box.

CALCULATIONS IF N IS ODD

Ordered data = {199,201,236,269,271,278,283,291,301,373,403}

N=11

Min = 199

Max = 403

Q1 position = (11+1)/4 =3,
Q1 value = 236

Median position = (11+1)/2=6
Median = 278

Q3 position = (11+1)*3/4=9
Q3 value = 301

IQR= Q3-Q1= 301-236=65

1.5*IQR = 1.5*65 = 97.5

Q1-1.5* IQR = 236-97.5= 138.5

Q3+1.5*IQR= 301+97.5=398.5

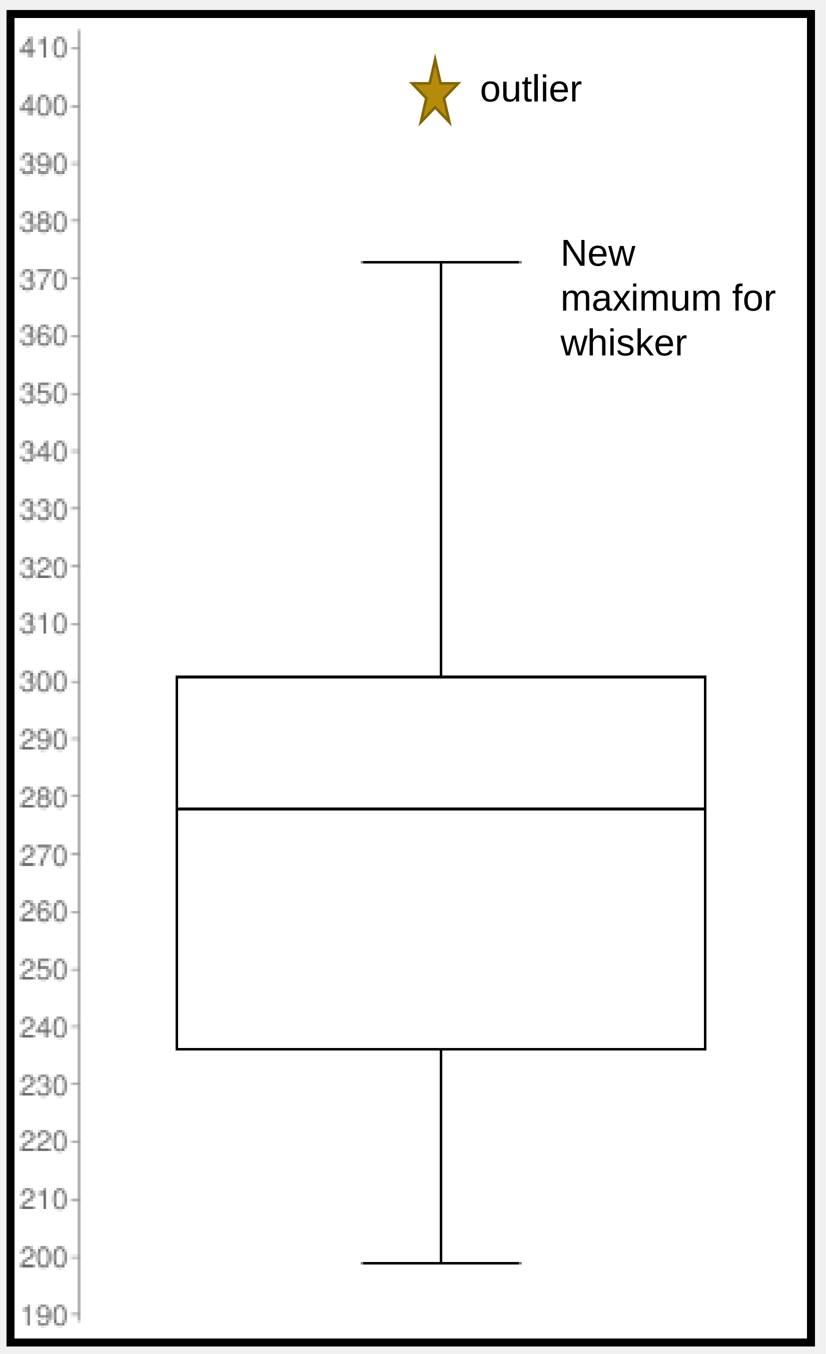
Check for outliers

No lower outliers

- No values < 138.5

1 Upper outlier 403 as it is > 398.5

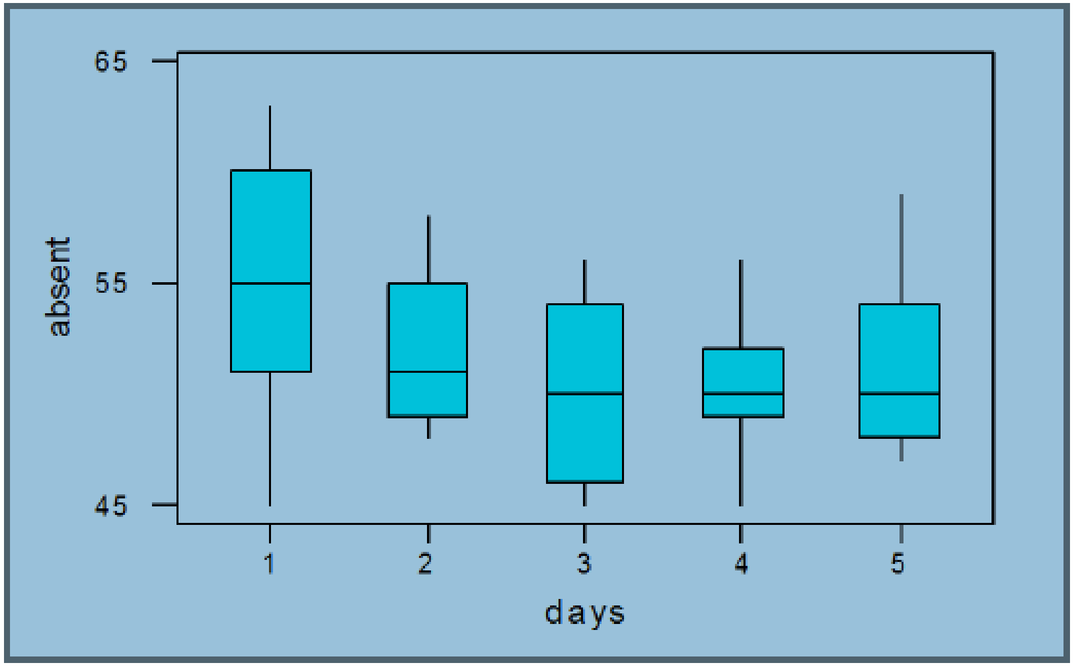
The whisker maximum is now 373, the next highest value.


410
400
390
380
370
360
350
340
330
320
310
300
290
280
270
260
250
240
230
220
210
200
190
outlier
New maximum for whisker

BOXPLOTS FOR COMPARISON

Boxplots are a basic way of summarising data.

Side-by-side boxplots of absenteeism by day of week



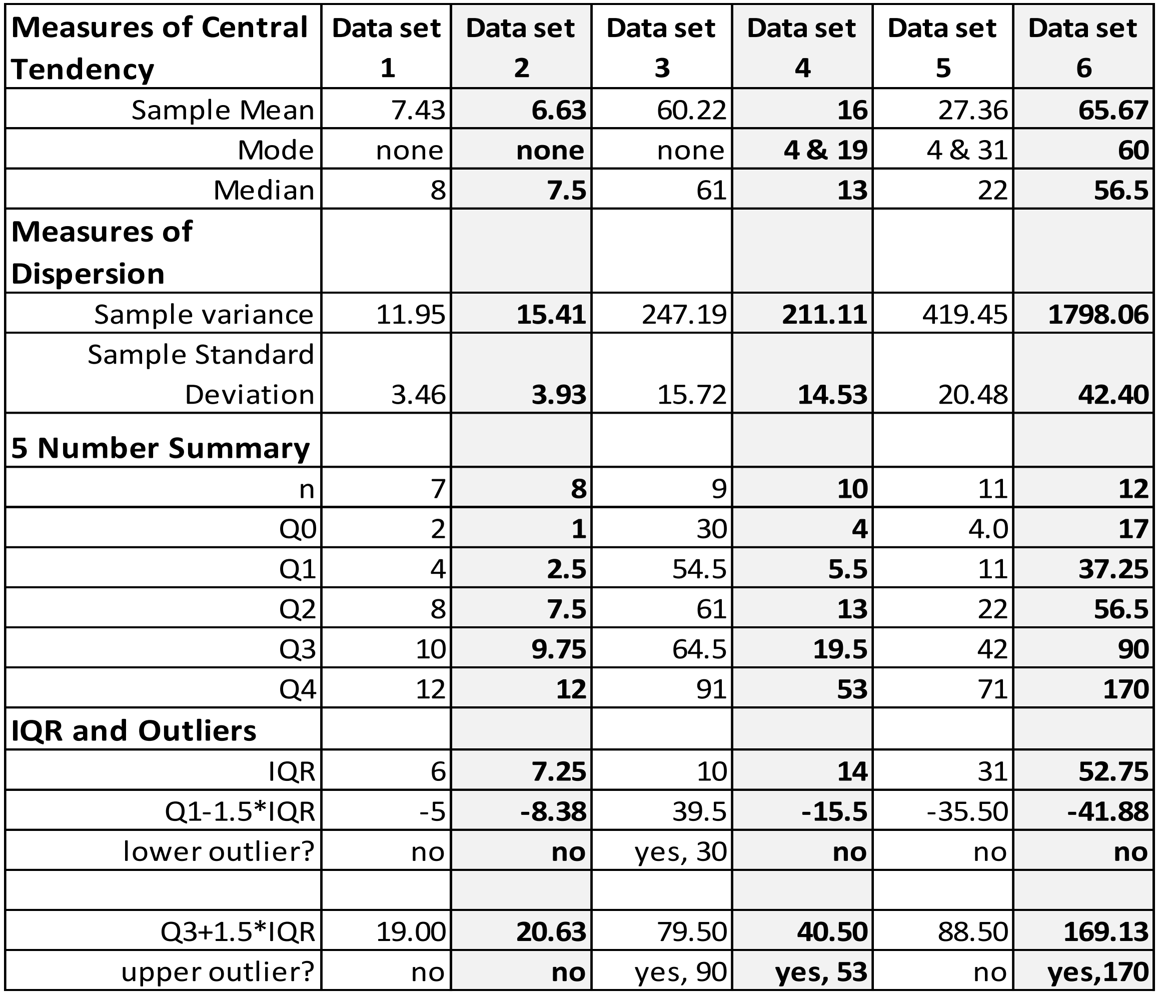
HOMEWORK - QUESTIONS

For each of the sample data sets determine:

1. Measures of Central Tendency - Sample Mean, Mode and Median
2. Measures of Dispersion - Sample Standard Deviation and Variance
3. 5 Number Summary
4. Inter Quartile Range (IQR) and Outliers (if any)

Data set 1	2	4	7	8	9	10	12					
Data set 2	1	2	4	7	8	9	10	12				
Data set 3	30	53	56	59	61	63	64	65	91			
Data set 4	4	4	6	8	10	16	19	19	21	53		
Data set 5	4	4	11	15	21	22	31	31	42	49	71	
Data set 6	17	26	35	44	53	55	58	60	60	100	110	170

HOMEWORK - ANSWERS



Measures of Central Tendency	Data set 1	Data set 2	Data set 3	Data set 4	Data set 5	Data set 6
Sample Mean	7.43	**6.63**	60.22	**16**	27.36	**65.67**
Mode	none	**none**	none	**4 & 19**	4 & 31	**60**
Median	8	**7.5**	61	**13**	22	**56.5**
Measures of Dispersion						
Sample variance	11.95	**15.41**	247.19	**211.11**	419.45	**1798.06**
Sample Standard Deviation	3.46	**3.93**	15.72	**14.53**	20.48	**42.40**
5 Number Summary						
n	7	**8**	9	**10**	11	**12**
Q0	2	**1**	30	**4**	4.0	**17**
Q1	4	**2.5**	54.5	**5.5**	11	**37.25**
Q2	8	**7.5**	61	**13**	22	**56.5**
Q3	10	**9.75**	64.5	**19.5**	42	**90**
Q4	12	**12**	91	**53**	71	**170**
IQR and Outliers						
IQR	6	**7.25**	10	**14**	31	**52.75**
Q1-1.5*IQR	-5	**-8.38**	39.5	**-15.5**	-35.50	**-41.88**
lower outlier?	no	**no**	yes, 30	**no**	no	**no**
Q3+1.5*IQR	19.00	**20.63**	79.50	**40.50**	88.50	**169.13**
upper outlier?	no	**no**	yes, 90	**yes, 53**	no	**yes,170**

OLD ASSESSMENT QUESTION

Boris Johnson has a food van at Curtin University. He collected data for the first 2 weeks of sales in October in his new business. (AU$)

50	50	100	200	350	370	400	400	420	480	550	570	1700	2000

1. Is this population or sample data?
2. Determine the mean, median and standard deviation of the data.
3. Determine the 5-number summary and identify any outliers.
4. Draw a boxplot clearly identifying all key information.
5. Boris wants to get a loan from the bank and has to decide how he will present his sales to the bank manager. He can choose: *Option 1: Sample mean and standard deviation* or *Option 2: Boxplot.*
   Which option should Boris choose? Justify your recommendation with reference to the data set and the previous analysis.