Aim 1 What is Statistics? What is Data?

Aim 2 Data Types

Aim 3 Graphical summaries

Aim 4 Describing distributions of numerical data

Aim 5 Numerical summaries

Aim 6 Population and Sample

7 exercises as your own homework

Motivation: Big Data

By 2020 – the increasing volume of data:

https://www.newgenapps.com/blog/big-data-statistics-predictions-on-the-future-of-big-data

- the new information generated per second for every human being will approximate amount to 1.7 megabytes.
- the accumulated volume of big data will increase to 44 zettabytes (the impact of IoT (Internet of Things))
- Google Search: 40,000 search queries are performed per second, which makes it 3.46 million searches per day and 1.2 trillion every year.
- Every minute Facebook users send roughly 31.25 million messages and watch 2.77 million videos.
- On YouTube alone, 300 hours of video are uploaded every minute.
- Business transactions via the internet will reach up to 450 billion per day.

1GB=1000MB; 1 TerraB=1000GB; 1PetaB=1000TB; 1ZettaB=1,000,000PB

Aim 1. What is Statistics?

A set of methods for:

- **data collection,**
- **data presentation,**
- **data modelling, • analysis** and
- **decision making** which take proper account of the **variation** and **uncertainty** that occurs in the real world.

What is Data?

- In a study, we collect information—data—from **cases. Cases** can be individuals, companies, animals, plants, or any object of interest.

- A **label** is a special variable used in some data sets to distinguish the different cases.

- A **variable** is any characteristic of an case. A variable **varies** among cases.

- **Examples:** age, height, blood pressure, ethnicity, leaf length, first language

- Different cases can have different **values** of a variable.

- The **distribution of a variable** tells us what values the variable takes and how often it takes these values.

Variables: In-class Exercise 1

- What are the other characteristics, apart from height, that we may wish to record if collecting information about people?

- Write down at least 10 possibilities.

- These characteristics are called ***variables***.

Variability or Uncertainty

- **Variation is everywhere!**

"People are not identical. They have different heights, weights, personalities, hair colours etc."

"What about a single person? Height/weight of a person is not the same over time."

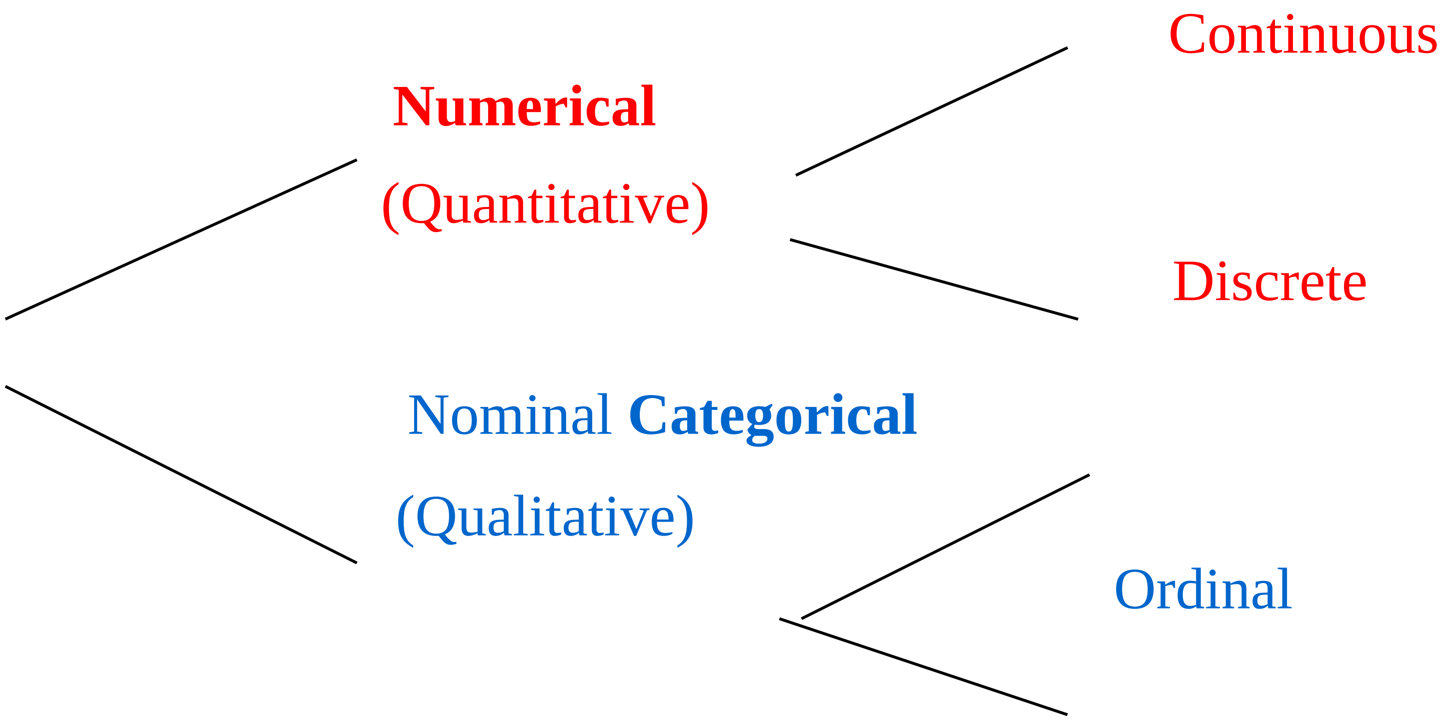
"Let's say at the moment John's height is exactly 180 cm. But we are not sure. Because all measurements have error or uncertainty.

The variation between the numbers might be related to:

- ***actual differences between people***
- ***changes in a person over time or - measurement error.***

Aim 2 Data Types

- Type of data indicates possible tools to use and what analyses are possible



Continuous

Numerical

(Quantitative)

Discrete

Nominal **Categorical**

(Qualitative)

Ordinal

Types of variables; characteristics

Variables can be either

- **numerical / quantitative…**

  Something that takes numerical values for which arithmetic operations, such as adding and averaging, make sense.

  Example 1: How tall you are; your age; your blood cholesterol level; the number of credit cards you own.

- or **categorical…..**

  Something that falls into one of several categories. What can be counted is **the count or proportion of cases** in each category.

  Example 2: Your blood type (A, B, AB, O); your hair color; your ethnicity; whether you paid income tax last tax year or not.

More on Data Types…

- **Data** can be classified as:
- **categorical** (or **qualitative**)
- Nominal (categories with no order)

  Eg: gender - m/f; colour - blue/green/yellow/red; condition - good/bad
- ordinal (categories with order)

  Eg: grades - FF, P, C, D, HD;

  Temperature - Low, Medium, High
- **numerical** (or **quantitative)**
- Continuous: temperature, height, weight, time, speed
- Discrete: number of defects, result of die toss, product count

Numerical - Continuous

- Numerical values that can be measured.
- Observed data take on any value in a given interval.
- The values are ‘measured’.

Example 3:

If a person is assembling a product component, the time it takes to accomplish that task could be any value with a reasonable range such as 3 minutes 36.4218 seconds or 5 minutes 17.5692 seconds.

Once the data is measured and recorded, the data is normally rounded off to a discrete number, however the data is actually continuous.

Numerical - Discrete

- Numerical values that have a finite or a countably finite number.
- The observed data values are **'counted'.**

Example 4:

Sampling 100 voters and determining how many voted for the government in the last election.

Number of Facebook/Twitter/LinkedIn users at Curtin University

In-class Exercise 2

1. Data on number of Facebook users by countryType of data?

Numerical - discrete (counted)

2. Data from student’s eye colour

(use 1=blue, 2=green, 3=brown, 4=hazel, 5=other) Type of data?

Categorical – nominal (no order)

3. Data on time to connect to internet: (use fast (0-3s), medium (3-7s), slow (>7s)) Type of data?

Categorical – ordinal (ordered)

Aim 3

Type of the variable dictates the required type of analysis including graphs

Graphical Summaries

(Moore *et al* Chapter 1.1)

Always, always, always, always graph your data!

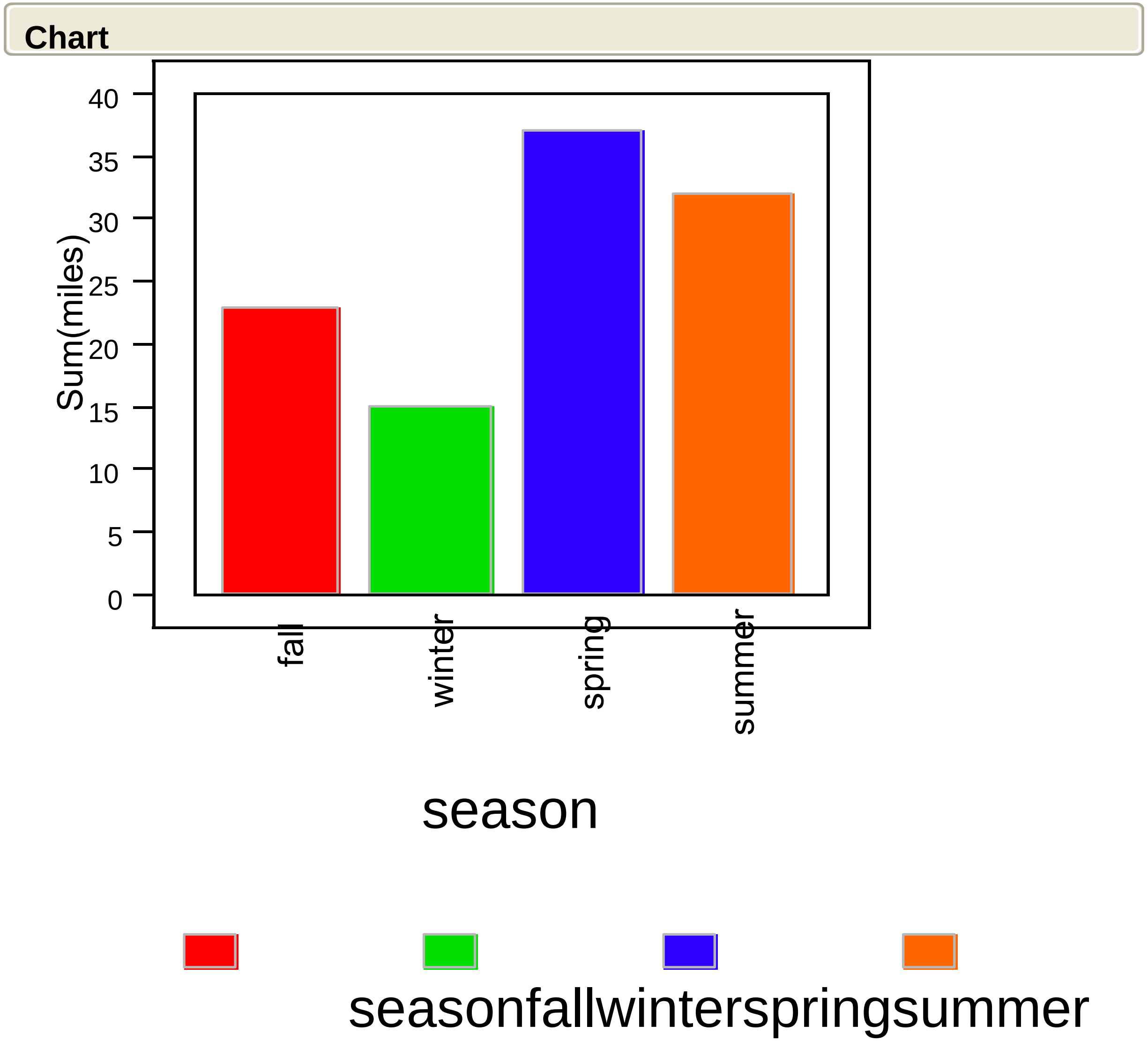
Charts for types of variables

CATEGORICAL

- Ordinal variable Bar chart

- Nominal variable
- Pareto chart
- Pie chart

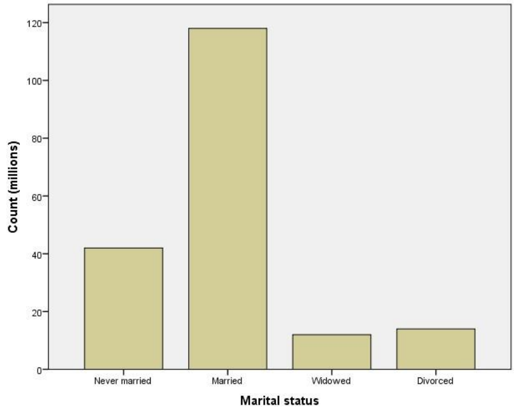
Example 5 Bar chart (categorical – ordinal)



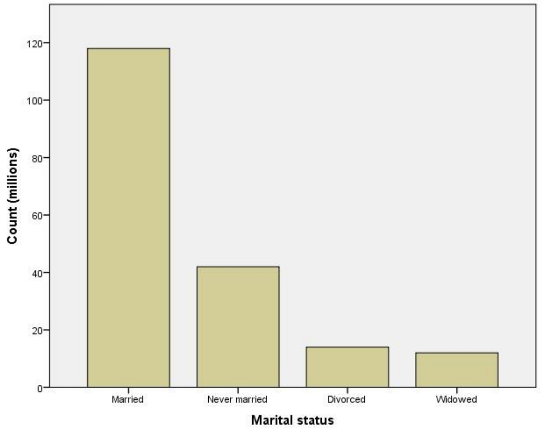
Ways to chart categorical - nominal data

Because the variable is categorical, the data in the graph can be ordered any way we want (alphabetical, by increasing value, by year, by personal preference, etc.)

Pareto chart Simply a bar chart where the bars are based on height.



ordered



Example 6 (Moore et al 2017):

Top 10 causes of death in the United States 2006

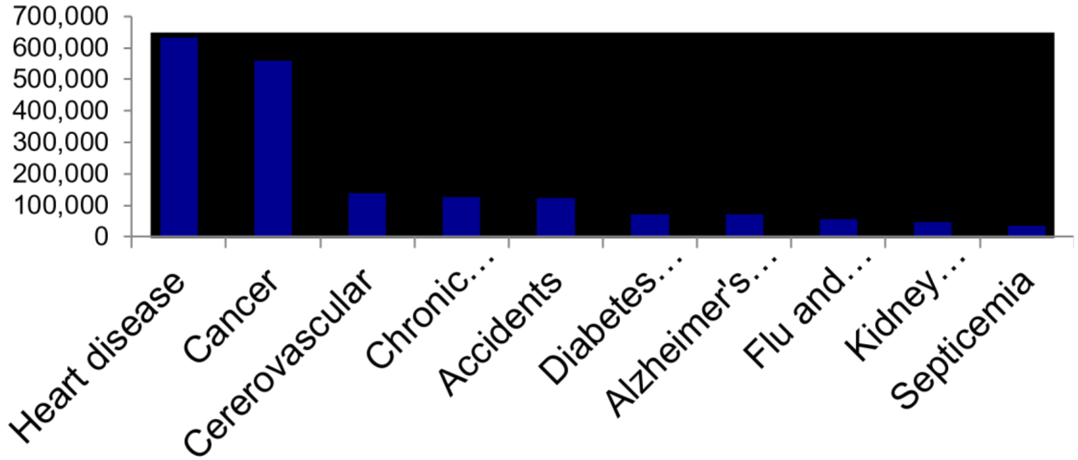
Rank	Causes of death	Counts	% of top 10s	% of total deaths
1	Heart disease	631,636	34%	*26%*
2	Cancer	559,888	30%	*23%*
3	Cerebrovascular	137,119	7%	*6%*
4	Chronic respiratory	124,583	7%	*5%*
5	Accidents	121,599	7%	*5%*
6	Diabetes mellitus	72,449	4%	*3%*
7	Alzheimer's disease	72,432	4%	*3%*
8	Flu and pneumonia	56,326	3%	*2%*

9	Kidney disorders	45,344	2%	*2%*
10	Septicemia	34,234	2%	*1%*
	All other causes	*570,654*		*24%*

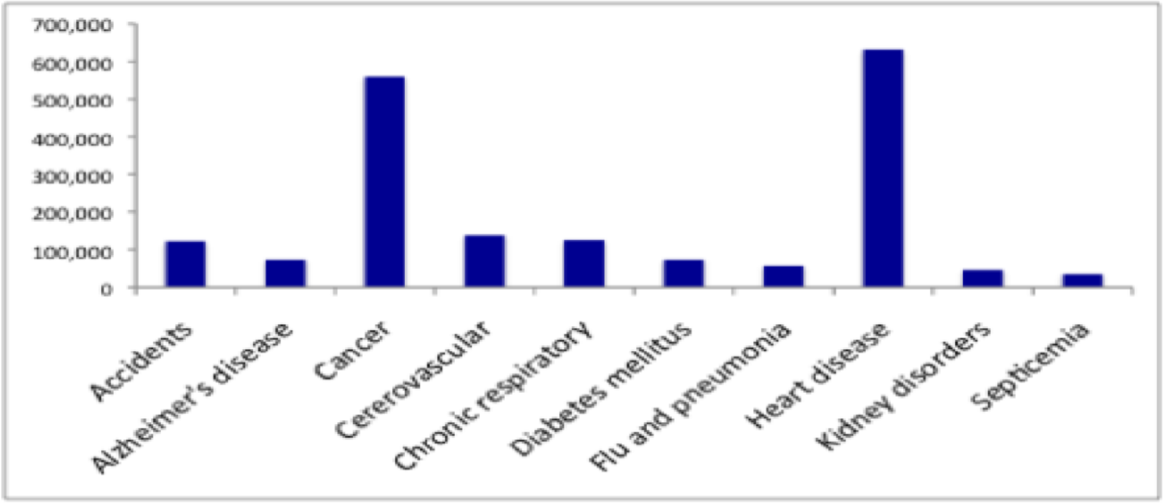
For each individual who died in the United States in 2006, we record what was the cause of death. The table above is a summary of that information.

BPareto charts

Each category is represented by one bar. The bar’s height shows the count (or sometimes the percentage) for that particular category. **Top 10 causes of deaths in the United States 2006**



Sorted by rank
→ Easy to analyze



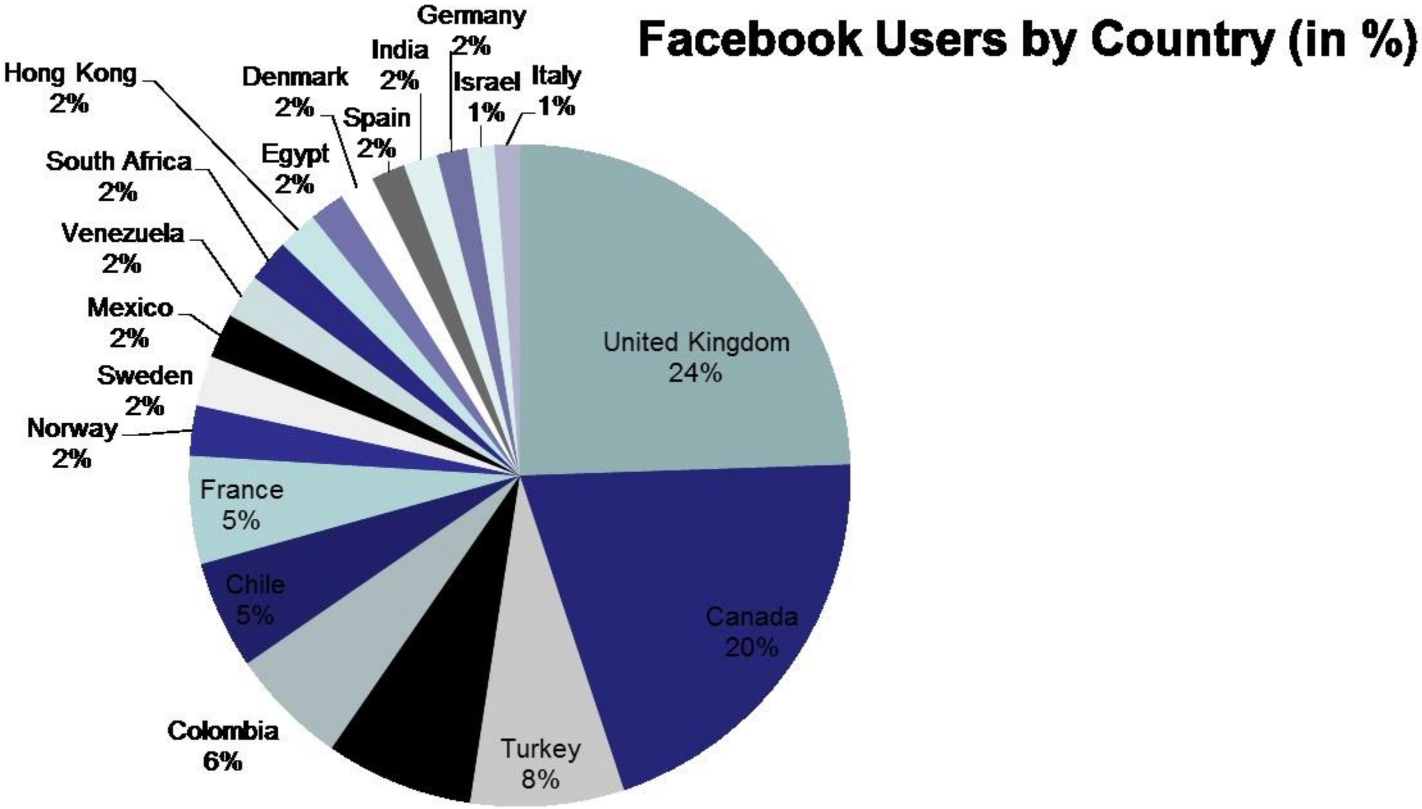
Sorted alphabetically
→ Much less useful

Pie charts

Each slice represents a piece of one whole.

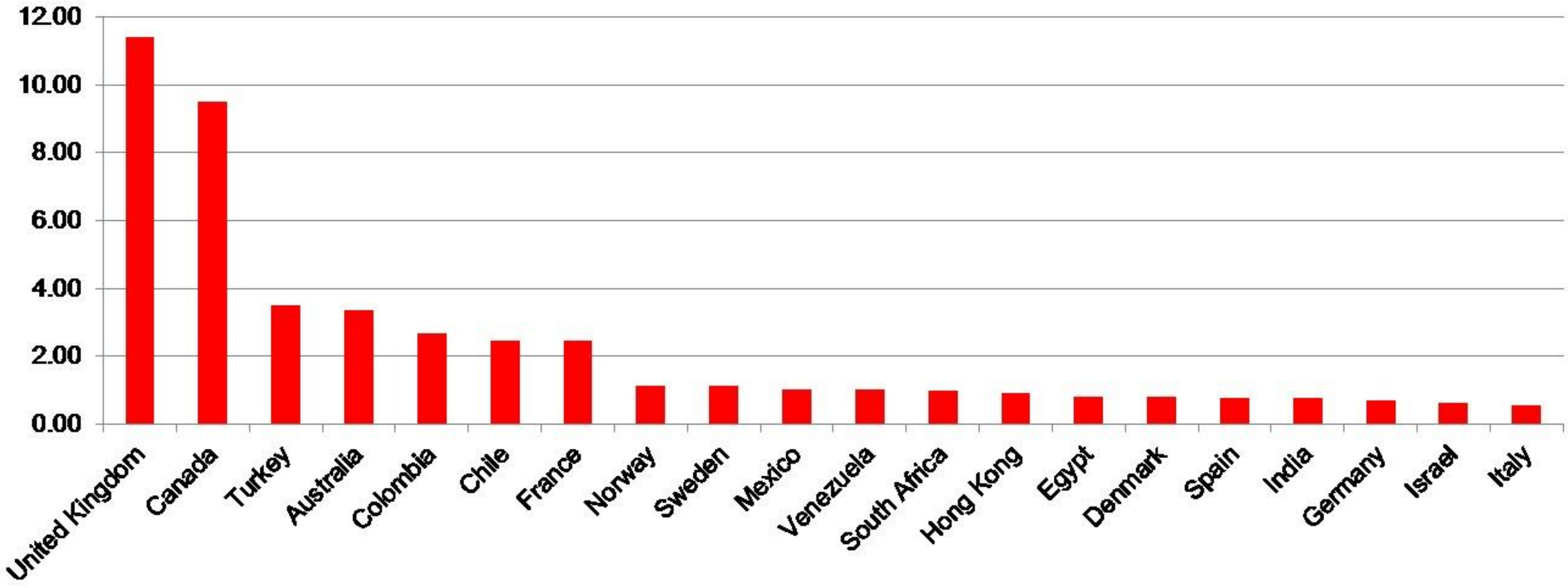
The size of a slice depends on what percent of the whole this category represents.

Percent of Facebook users by country (Moore et al 2017)



Example 7: Facebook users by country (a better graph than a pie chart)

Facebook Users By Country (in %)



Ways to chart quantitative data

- Histograms

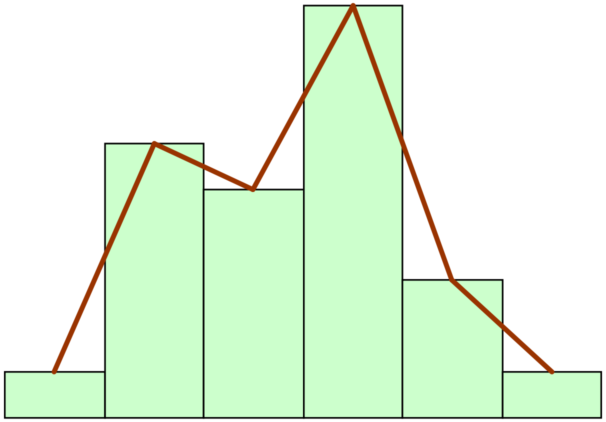
A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class.

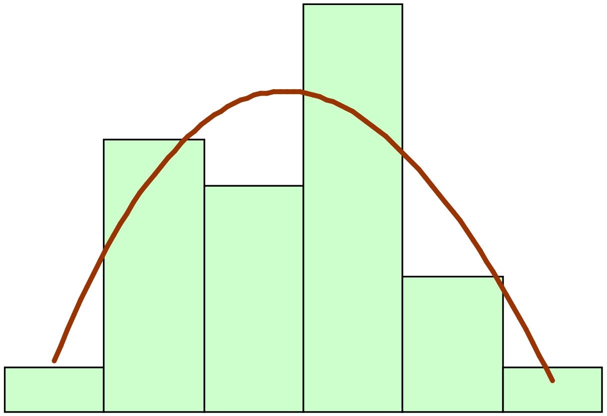
- Boxplot

Provides 5 number summary

Interpreting histograms

- When **describing the distribution** of a quantitative variable, we look for the overall pattern and for striking deviations from that pattern.
- We can **describe** the *overall* pattern of a histogram by its **shape, (s) center,** and **spread (3S).**





Histogram with a line connecting Histogram with a
smoothed curve each column → too detailed
highlighting the overall pattern of the
distribution

How to create a histogram

Divide the possible values into classes or intervals or bins of equal widths.

Count how many observations fall into each interval/bin. Instead of counts, one may also use percents.

Draw a picture representing the distribution—each bar height is equal to the number (or percent) of observations in its interval.

It is an iterative process – try and try again. What bin size should you use?

Not too many bins with either 0 or 1 counts

Not overly summarized that you lose all the information

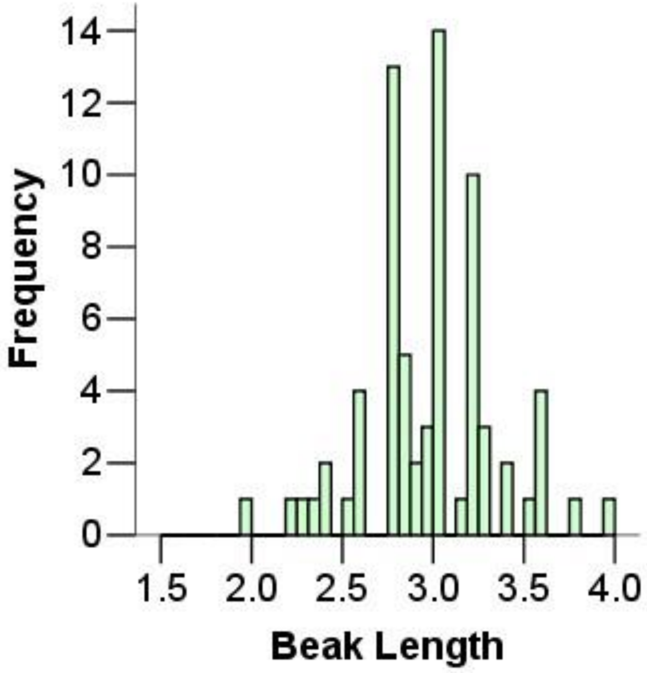
Not so detailed that it is no longer summary

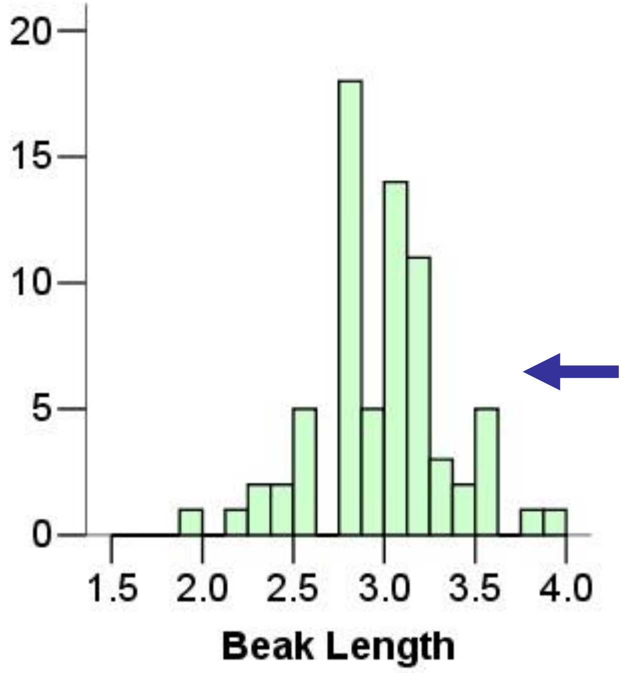
➔ rule of thumb: start with 5 to 10 bins

Look at the distribution and refine your bins

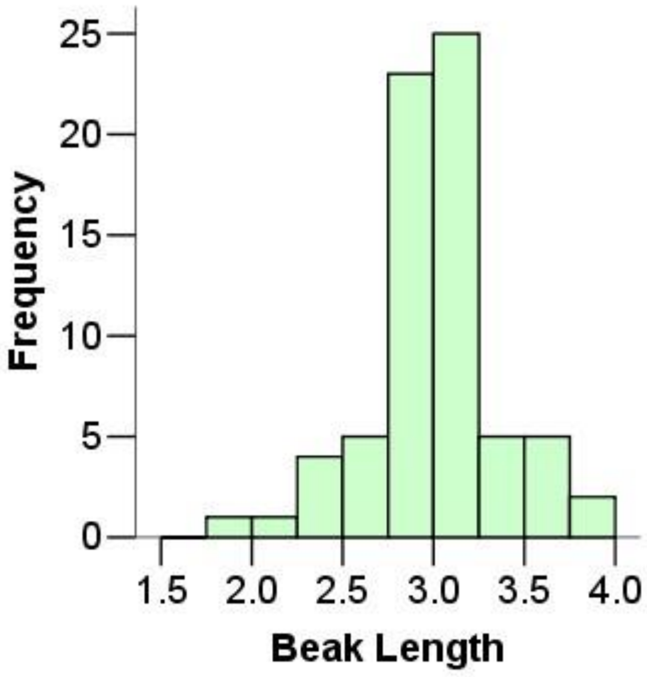
(There isn't a unique or "perfect" solution)

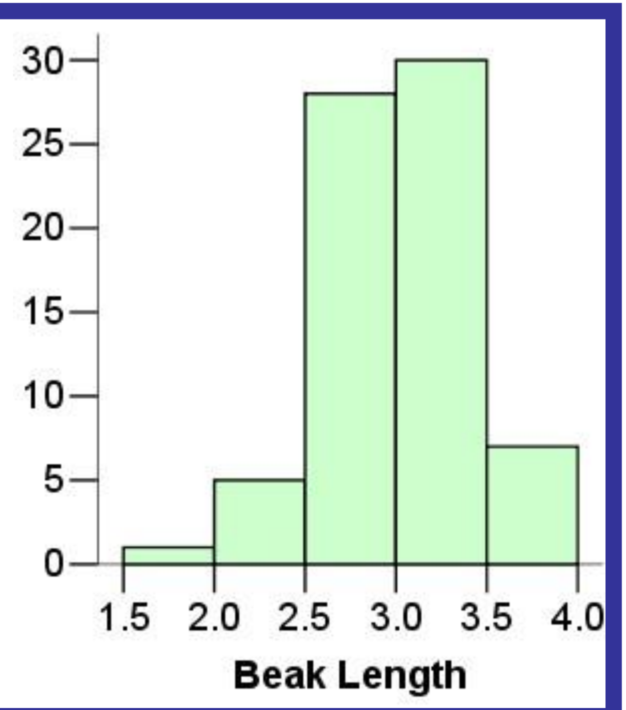
Same data set



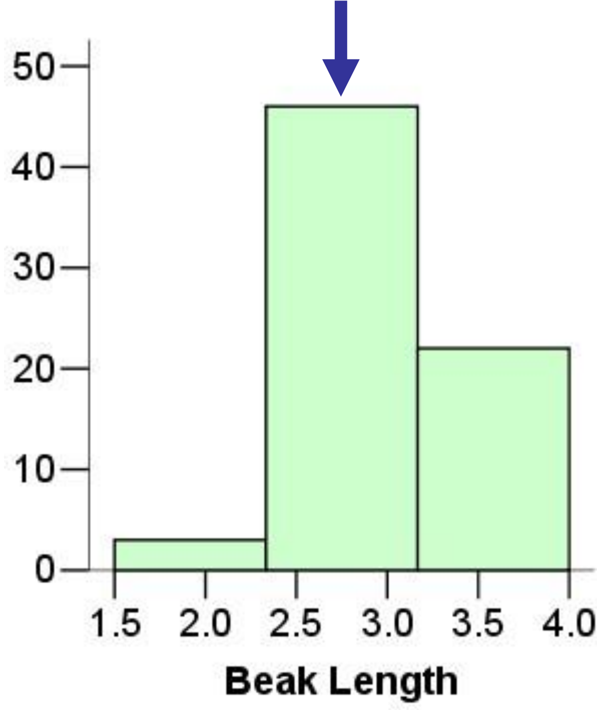


Not summarized enough





Too summarized



Example 8. IQ data
Moore et al 2017 Chapter 1

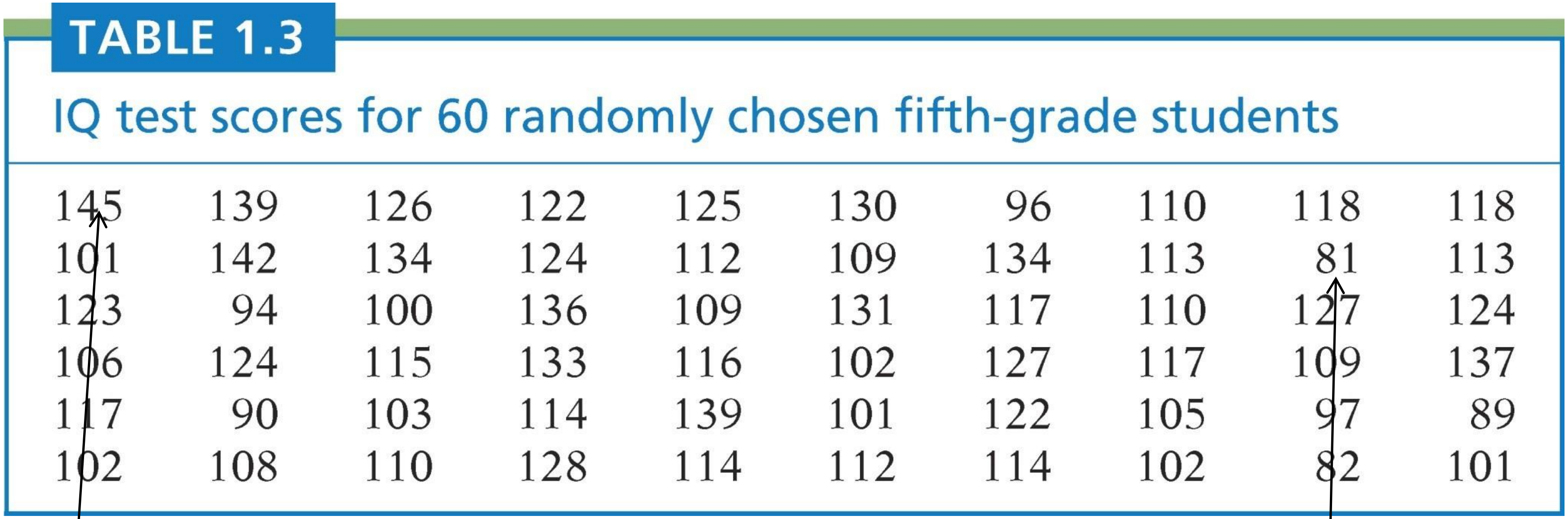


TABLE 1.3

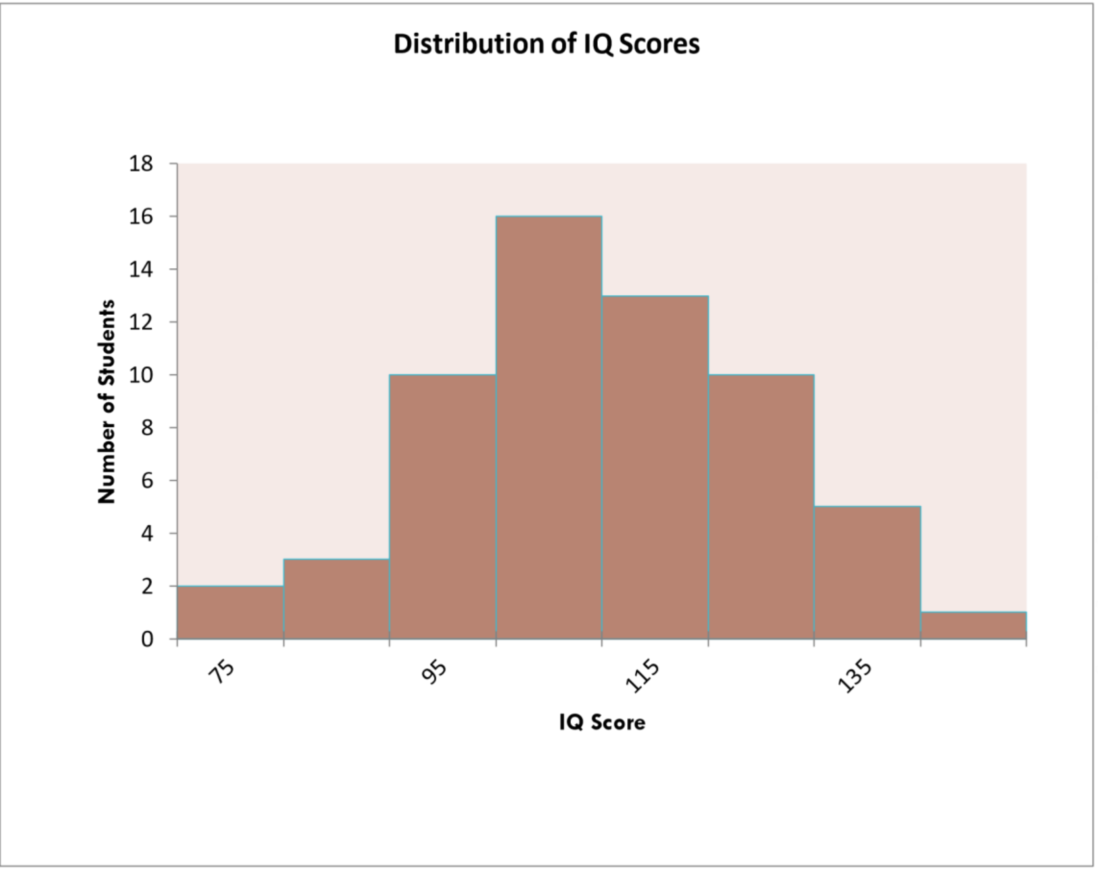
IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Maximum=145

Minimum=81

Histograms: IQ data



Class	Count
75 ≤ IQ Score < 85	2
85 ≤ IQ Score < 95	3
95 ≤ IQ Score < 105	10
105 ≤ IQ Score < 115	16
115 ≤ IQ Score < 125	13
125 ≤ IQ Score < 135	10
135 ≤ IQ Score < 145	5
145 ≤ IQ Score < 155	1

Uses for Graphs

Explore *data* explore distribution of one or more variables explore possible relationships between variables.

Present *data* to *highlight* specific/important information or *answer* a specific question.

Interpreting graphs

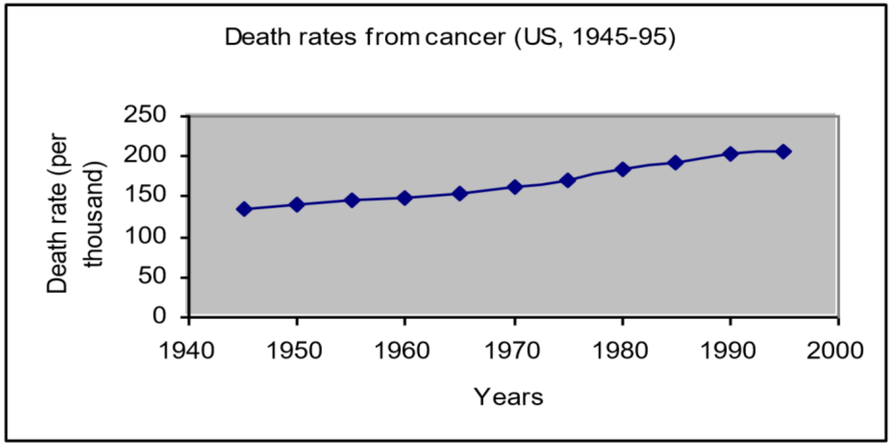
Evaluate critically

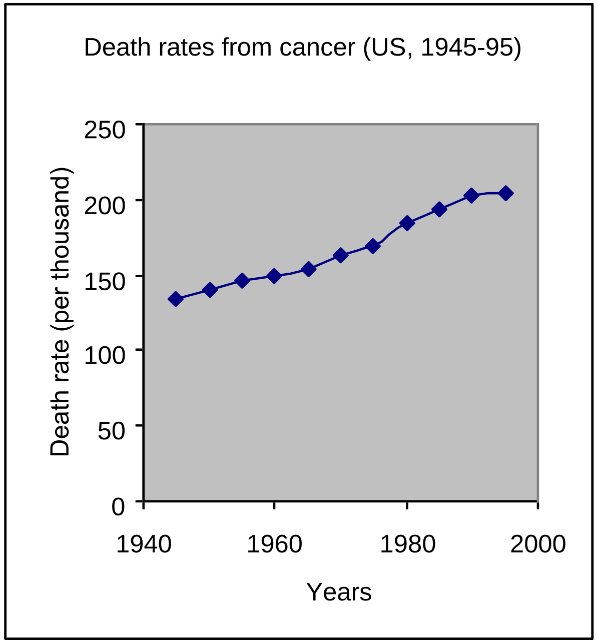
- Is title clear and informative?
- Look at axis labels
- what is being graphed?

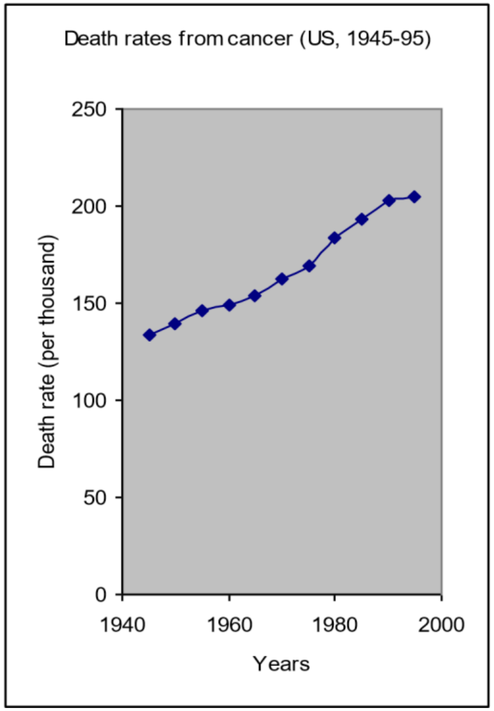
- Are axes clearly labeled?
- Look carefully at scales.
- Do they start at zero?
- are they linear?
- Is there misleading chart junk, effects or perspective?
- Is the graphical message relevant?

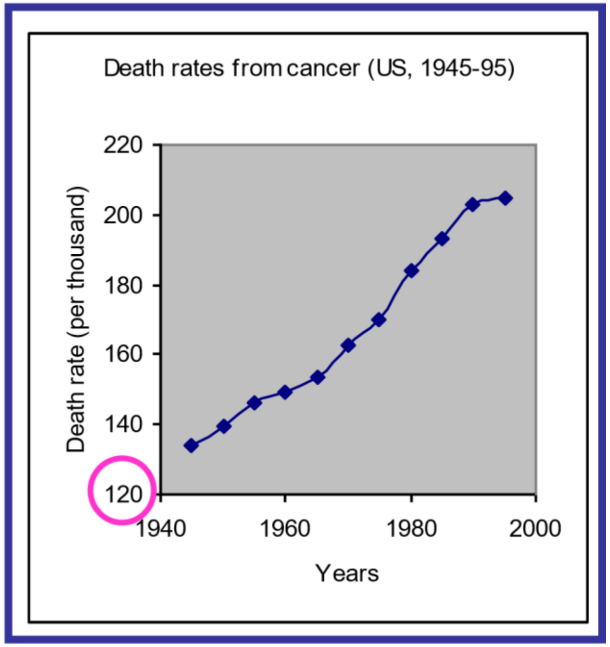
Scales matter

How you stretch the axes and choose your scales can give a different impression.









A picture is worth a thousand words,

BUT

There is nothing like hard numbers.

→ **Look at the scales.**

In-class Exercise 3

Q3. Variables measured in a study considering potential childhood experiences that affect an adult's eyesight:

GLASSES : Whether or not person currently wears glasses (1='Yes', 2='No')

TV_HOURS : Measuring the number of hours of TV viewed per week as a child

NIGHTLIGHT : Whether person slept with a nightlight as a child (1='Yes', 2='No')

EDUCATION : A person's greatest educational level

Responses: School Cert, HSC, TAFE, Uni Degree, Hons, PhD.

Which one of the following sets of statements about the data types of the above variables is most correct?

(a) GLASSES is continuous; TV_HOURS is nominal; NIGHTLIGHT is ordinal

(b)NIGHTLIGHT is quantitative; TV_HOURS is quantitative; EDUCATION is categorical

(c)TV_HOURS is continuous; EDUCATION is discrete; NIGHTLIGHT is ordinal
(d)TV_HOURS is quantitative; EDUCATION is ordinal; NIGHTLIGHT is nominal

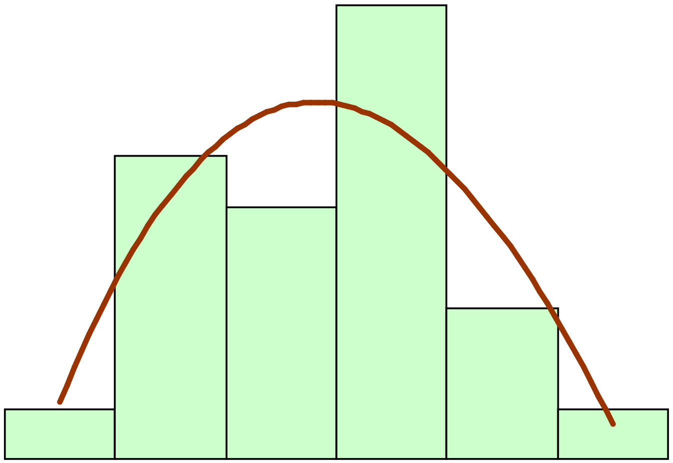
ANSWER: Glasses (Categorical-nominal); TV_Hours (Numerical-Continuous);

Nightlight (Categorical – Nominal); Education (Categorical –Ordinal) ------- (d)

Aim 4 Describing Distributions – 3S Numerical Data

When **describing the distribution** of a **quantitative** variable, we look for the **overall pattern** and for **striking deviations** from that pattern.

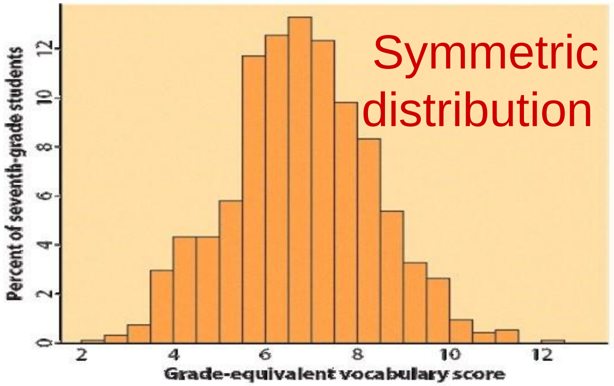
We can describe the **overall pattern** of a histogram by its **Shape, (S)center,** and **Spread (3S).**

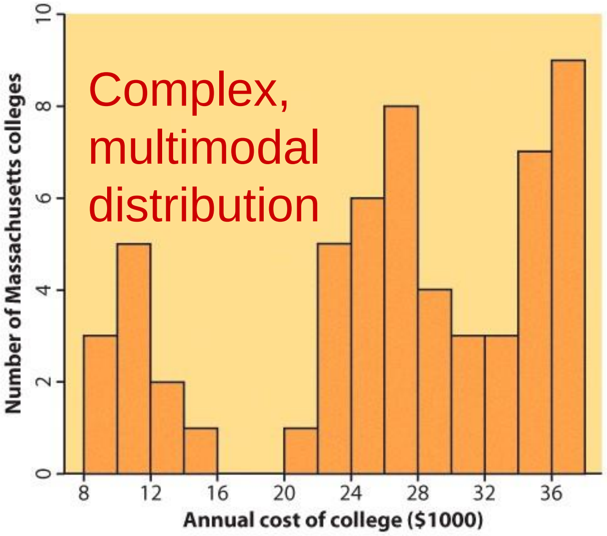


Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

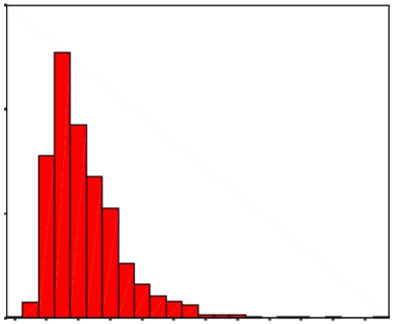
A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.



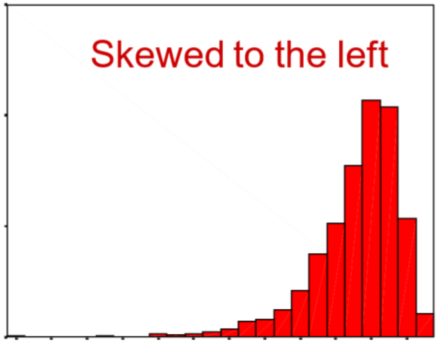


- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side.

Skewed to the right



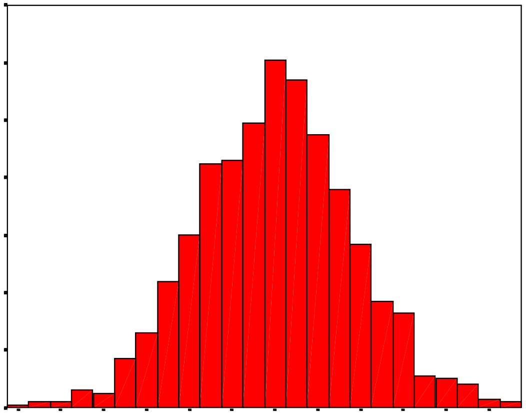
- It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.



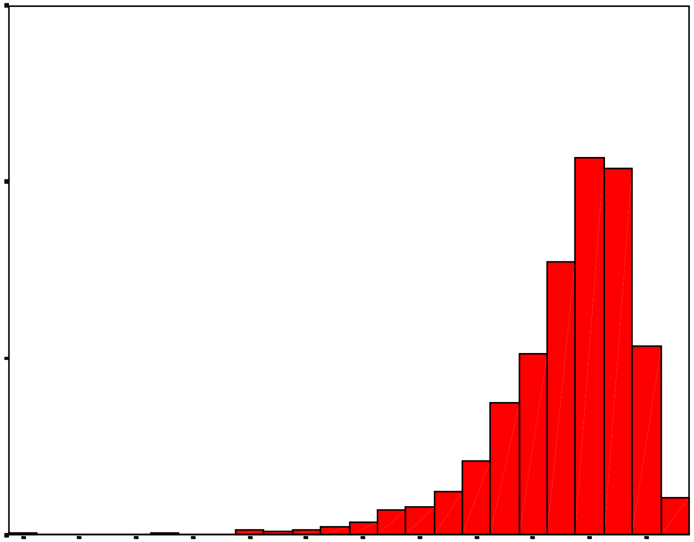
□ Not all distributions have a simple overall shape, especially when there are few observations.

Distributional Shape

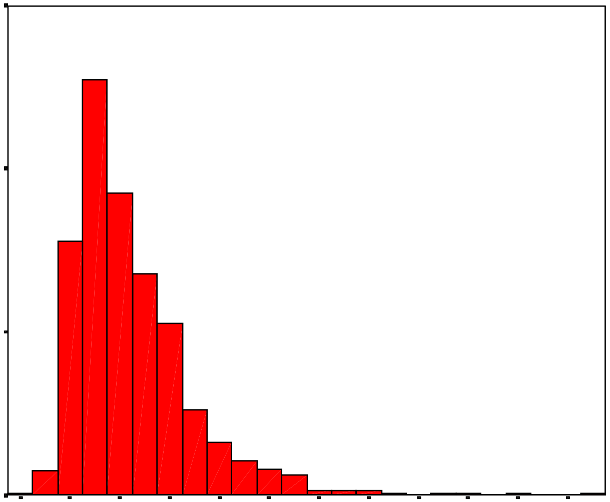
Symmetric



Skewed to the left



Skewed to the right



Aim 5 Numerical summaries for *numerical* data

- In Statistics we often use summary statistics and graphs to represent samples of data

- This allows us to efficiently present information and provides a basis for comparison and tentative conclusions **Numerical summaries for *numerical* data**

Numerical summaries (statistics) for **‘center’** or **location**

1. mode
2. median
3. mean

Numerical summaries (statistics) for **spread**

1. **range**

2. inter-quartile range (IQR)
3. standard deviation

Measure of center / location 1: Mode

The value of the variable that occurs most frequently.

In-class Exercise 4

Data: 7, 2, 5, 1, 5, 5, 3, 2, 12

Mode = ?

Measure of center 2: the median

The **median** is the midpoint of a distribution—the number such that half of the observations are smaller and half are larger.

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6 7	2.1
7	8	2.3
8	9	2.3
9	10	2.5
10	11	2.8
11	12	2.9
12		3.3
13	3.4	

12		3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6

Example 4: Years until death for a certain disease

14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6
25	**12**	**6.1**

1. Sort observations by size.
n = number of observations

2.a. If *n* is **odd,** the median is observation (*n*+1)/2 down the list

← *n* = 25
(*n*+1)/2 = 26/2 = 13
Median = 3.4

n = 24
→ *n*/2 = 12
Median = (3.3+3.4) /2 = 3.35

2.b. If *n* is **even,** the median is the mean of the two middle observations.

Measure of center 3: the mean

Example 5 Women’s height

The mean or arithmetic average

To calculate the *average,* or **mean,** add all values, then divide by the number of cases. It is the “center of mass.”

Sum of heights is 1598.3 divided by 25 women = 63.9 inches

In-Class Exercise 5. What is the median?

58.2	64.0
59.5	64.5
60.7	64.1
60.9	64.8
61.9	65.2
61.9	65.7
62.2	66.2
62.2	66.7
62.4	67.1
62.9	67.8
63.9	68.9
63.1	69.6
63.9	

woman (i)	height (x)		woman (i)	height (x)
i = 1	x_1= 58.2		i = 14	x_{14}= 64.0
i = 2	x_2= 59.5		i = 15	x_{15}= 64.5
i = 3	x_3= 60.7		i = 16	x_{16}= 64.1
i = 4	x_4= 60.9		i = 17	x_{17}= 64.8
i = 5	x_5= 61.9		i = 18	x_{18}= 65.2
i = 6	x_6= 61.9		i = 19	x_{19}= 65.7
i = 7	x_7= 62.2		i = 20	x_{20}= 66.2
i = 8	x_8= 62.2		i = 21	x_{21}= 66.7
i = 9	x_9= 62.4		i = 22	x_{22}= 67.1
i = 10	x_{10}= 62.9		i = 23	x_{23}= 67.8
i = 11	x_{11}= 63.9		i = 24	x_{24}= 68.9
i = 12	x_{12}= 63.1		i = 25	x_{25}= 69.6
i = 13	x_{13}= 63.9		n=25	Σ=1598.3

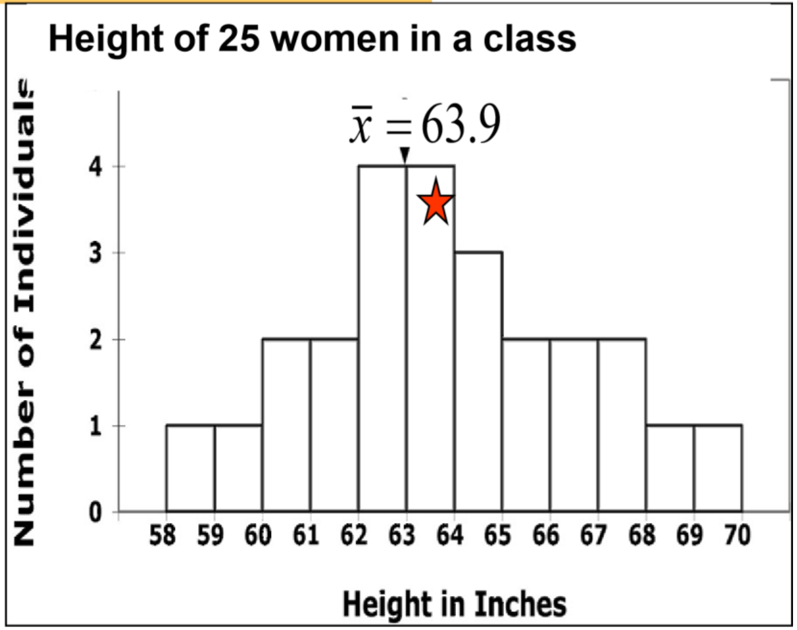
Mathematical notation:

Data x_i , i=1,2, …, n

Sample mean $\bar{x}$

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

$$\bar{x} = \frac{1598.3}{25} = 63.9$$

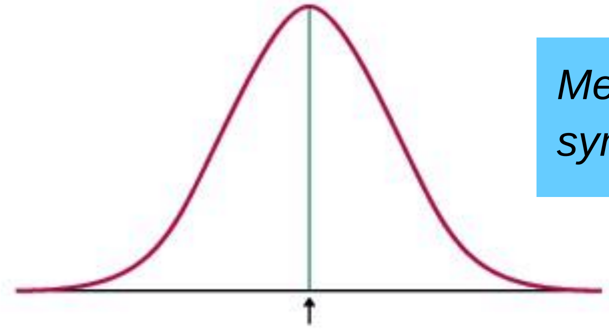


Height of 25 women in a class

Learn right away how to get the mean using calculator.

Comparing the mean and the median

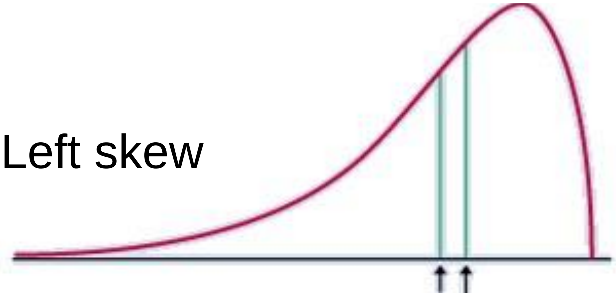
- The mean and the median are the same only if the distribution is symmetrical.
- The median is a measure of center that is resistant or robust to skew and outliers. The mean is not.



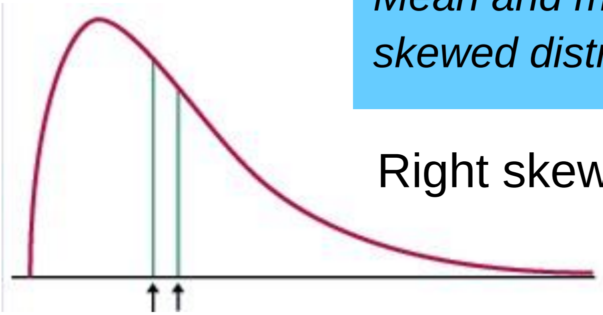
Mean and median for a symmetric distribution

Mean and median for skewed distributions

Left skew



Mean < Median



Right skew

Median < Mean

Comparison between mean and median: Which one is better?

- Both are useful for indicating the center of a data set.
- Mean is more commonly used but is affected by extreme values (outliers) and skewness
- Median may be a better representation of the ‘typical’ value for skewed data OR data with extreme values because the sample is split in half.

Measure of Spread 1: Range

Range is the difference between largest (maximum) and smallest (minimum) values in the data set.

Sensitive to unusually extreme values (i.e., values at the ends of distribution)

In-class Exercise 6 Data values:

21, 25, 23, 28, 16, 19, 17, 21, 15, 22

maximum = ? minimum = ? range = ?

Quartiles

- First 25% of data are less than first quartile Q1 (and 75% of data are greater than Q1)

- Second quartile Q2 is the median, with 50% of data on either side

- First 75% of data are below the third quartile Q3 (and 25% of data are greater than Q3)

The quartiles

Example 7 Years until death for a certain disease

The **first quartile, *Q*₁,** is the value in the sample that has 25% of the data at or below it (it is the median of the lower half of the sorted data, excluding *M*).

The **third quartile, *Q*₃,** is the value in the sample that has 75% of the data at or below it (it is the median of the

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	1	2.3
9	2	2.5
10	3	2.8
11	4	2.9
12	5	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1

upper half of the sorted data,

excluding *M*).

M = median = *3.4*

***Q_1*= first quartile**

=(2.1+2.3)/2== 2.2

***Q_3*= third quartile**

=(4.2+4.5)/2= = 4.35

Measure of spread 2:

Inter-Quartile Range (IQR)

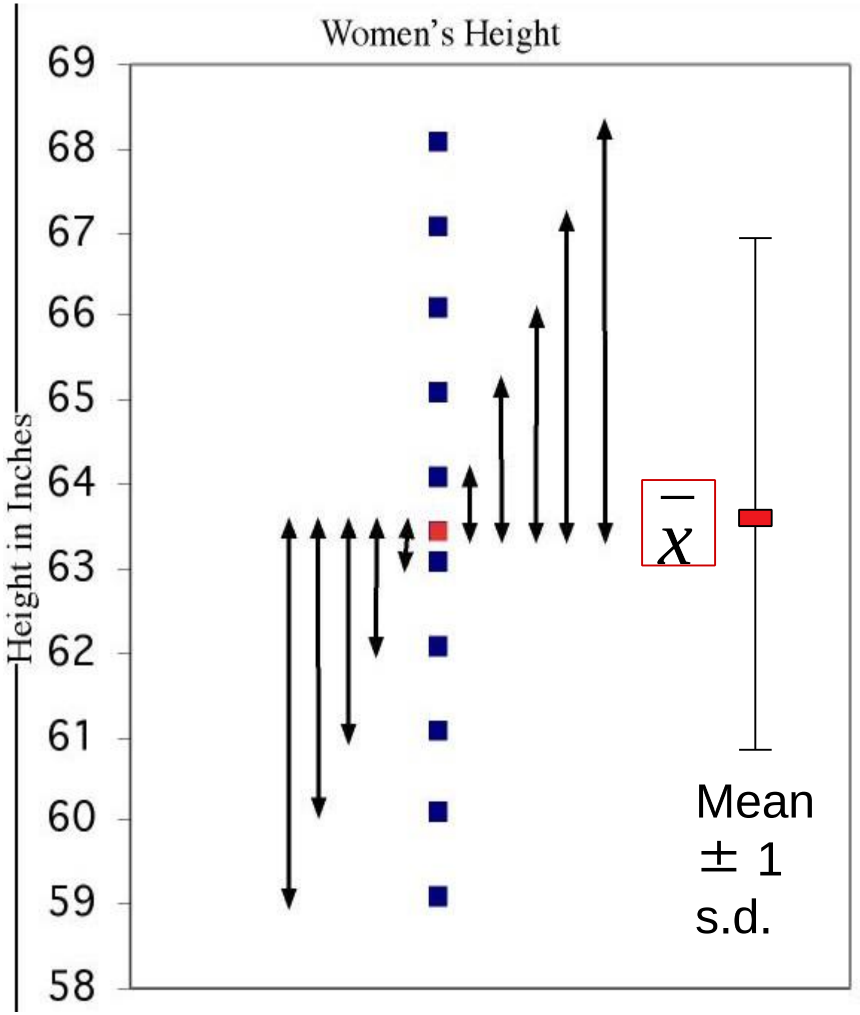
- The IQR is the difference between Q1 and Q3: **IQR=Q3-Q1**

- For the previous example • Q1 = ? and Q3 = ?
- IQR = ?
- IQR measures the spread of the middle 50% of the data.
- It is not sensitive to extreme values. Why?

Measure of spread 3: the standard deviation

Example 8 Women’s height

The standard deviation “s” is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.



1. First calculate the **variance** ***s*².**

1

2

n

2

–

s □ ______□(*x*ᵢ □ *x*)

n □1 1

Data xi , i=1,2, ..., n • *n*: sample

size Sample mean *x*пђ • ∑ : sum

of

2. Then take the square root to get the **standard deviation s.**

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

Calculations …

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

Mean = x = 63.4 n=14

Sum of squared deviations from mean = 85.2

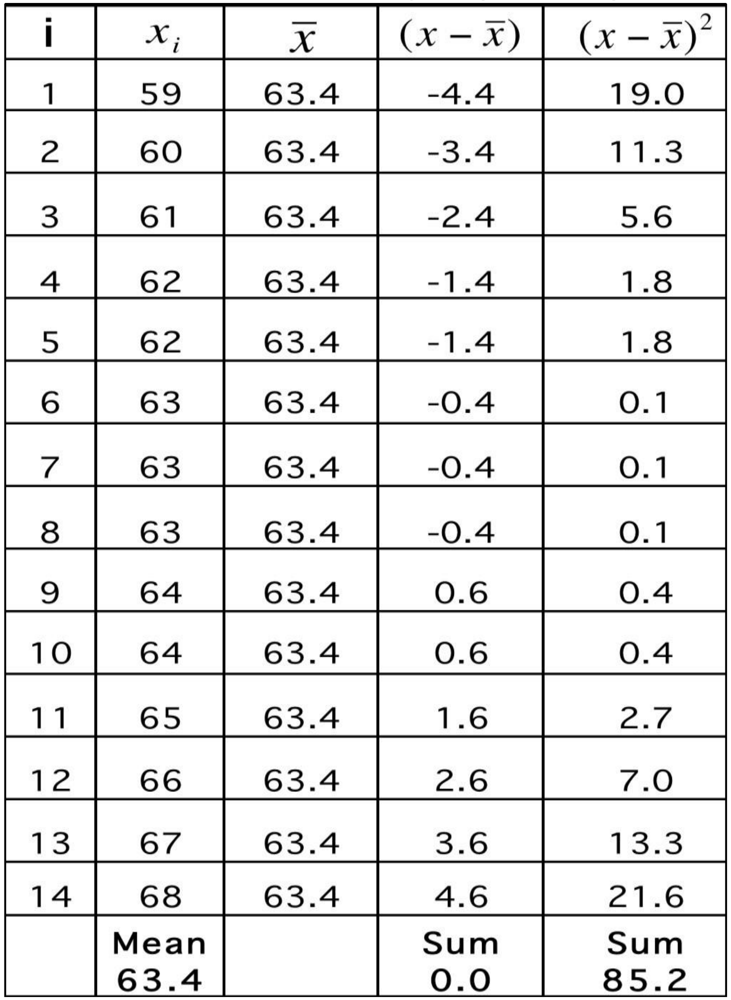
Degrees freedom (df) = (n − 1) = 14-1=13

s^2 = variance = 85.2/13 = 6.55 inches squared

s = standard deviation = √6.55 = 2.56 inches

We'll rarely calculate these by hand, so make sure to know how to get the standard deviation using your calculator or Excel.

Women's height (inches)



i	x_i	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
1	59	63.4	-4.4	19.0
2	60	63.4	-3.4	11.3
3	61	63.4	-2.4	5.6
4	62	63.4	-1.4	1.8
5	62	63.4	-1.4	1.8
6	63	63.4	-0.4	0.1
7	63	63.4	-0.4	0.1
8	63	63.4	-0.4	0.1
9	64	63.4	0.6	0.4
10	64	63.4	0.6	0.4
11	65	63.4	1.6	2.7
12	66	63.4	2.6	7.0
13	67	63.4	3.6	13.3
14	68	63.4	4.6	21.6
	Mean 63.4		Sum 0.0	Sum 85.2

Properties of Standard Deviation

- *s* measures spread about the mean and should be used only when the mean is the measure of center.

- *s* = 0 only when all observations have the same value and there is no spread. Otherwise, *s* > 0.

- *s* is not resistant to outliers.

- *s* has the same units of measurement as the original observations.

Interpreting measure of spread

- Small standard deviation implies the data is concentrated around the mean.
- Large standard deviation implies the data is widely spread around the mean.
- Can examine spread of data using histograms or box plots.

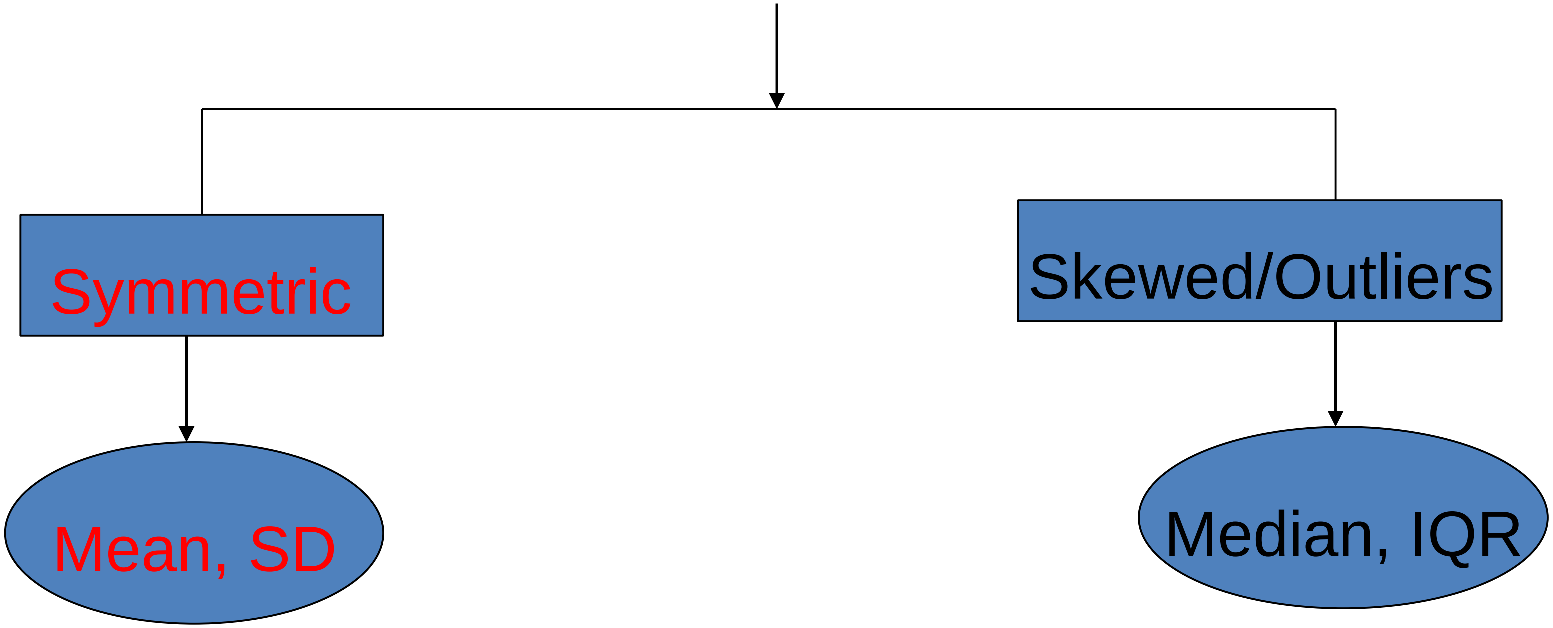
Comparison between IQR and SD

- Both are useful for indicating the spread of a data set.

- SD is more commonly used but is affected by outliers (ie. SD is sensitive to outliers)
- IQR is the best measure of spread for skewed data or data with extreme values because outliers have little effect on the IQR (ie IQR is insensitive to outliers)

Rule of thumb for choosing between Moment-based and Quantile-based measures

Shape



Symmetric

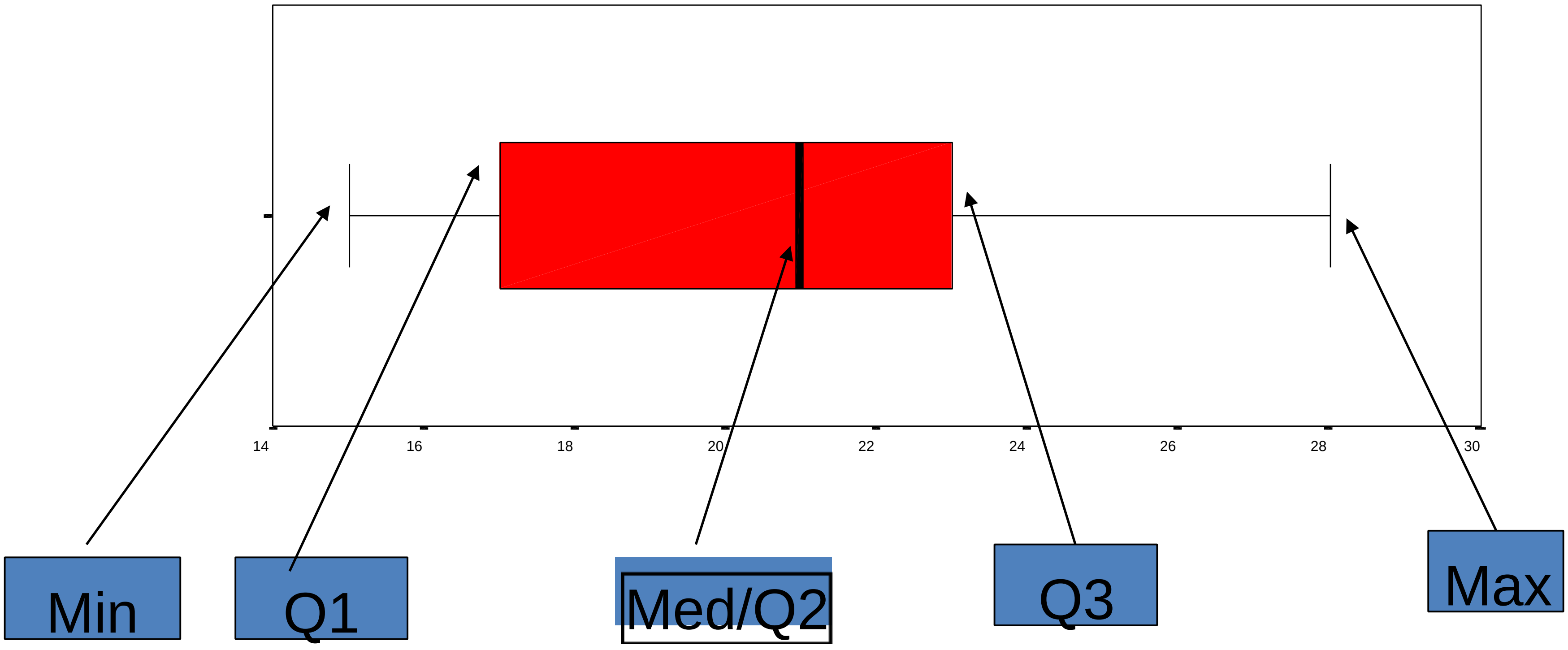
Skewed/Outliers

Mean, SD

Median, IQR

Box plot

Provides **5-number summary**


14
16
18
20
22
24
26
28
30
Min
Q1
Med/Q2
Q3
Max

Five-number summary and boxplot

25	65	6.1
24	43	5.6
23	2	5.3
22		4.9
21		4.7
20	1	4.5
19	6	4.2

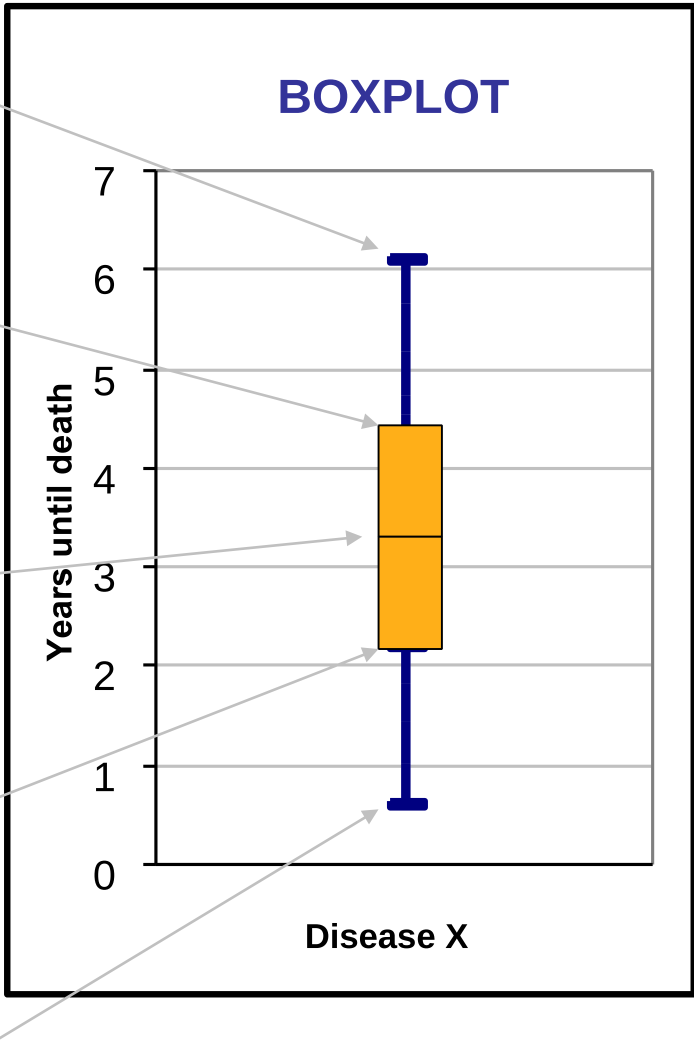
Largest = max = 6.1

Q_3= third quartile
= 4.35

M = median = 3.4

Q_1= first quartile
= 2.2

Smallest = min = 0.6



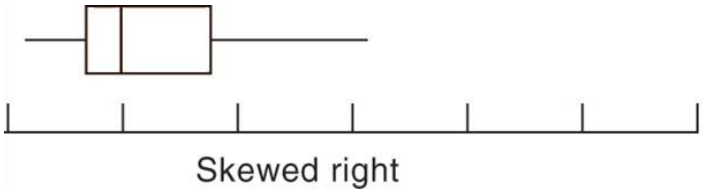
Five-number summary:
min Q_1 M Q_3 max

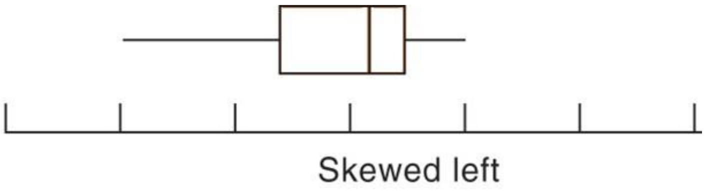
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9

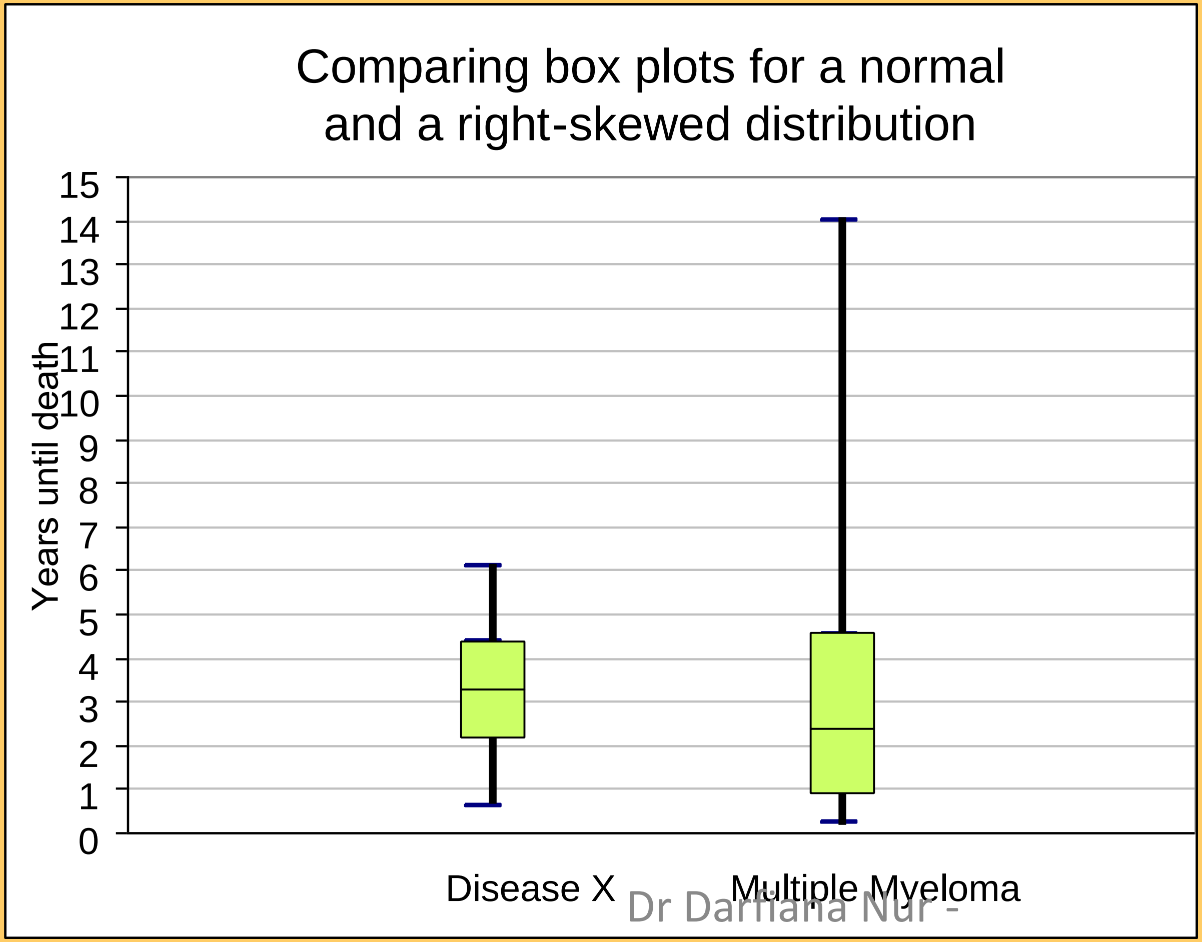
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2

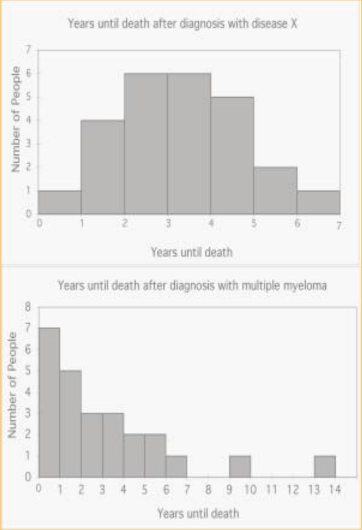
1	1	0.6

Boxplots for skewed data: Example 9









Boxplots remain true to the data and depict clearly symmetry or skew.

In Class Exercise 7.

If a distribution is skewed to the right, data taken from the distribution will tend to have a larger mean than median.

a) TRUE

b) FALSE

ANSWER

Skewed to the right

Some large values

Large values don’t affect for calculation of median

Large values will be used for mean (hence larger)

=>larger mean. TRUE

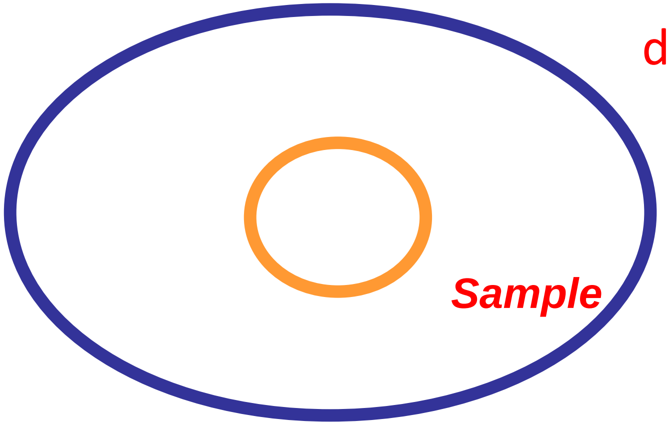
Aim 6 Population versus sample

- **Population:** The entire group of individuals in which we are interested but which we do can’t usually assess directly.

Sample: The part of the population we actually examine and for which we have data.

- Example: All humans, all working-age people in SA, all tertiary students in South Australia

How well the sample represents the population depends on the sample design.



- A **parameter** is a number describing a characteristic of the population.

A **statistic** is a number describing a characteristic of a sample.

Sampling

- The idea of *sampling* is to study a part (the sample) in order to gain information about the whole (the *population*).

- A *census* is where we study the whole *population*.

- *Sample* is a collection of individual observations selected from the *population*. Ideally our *sample* will be representative of the entire *population*.

Example 10