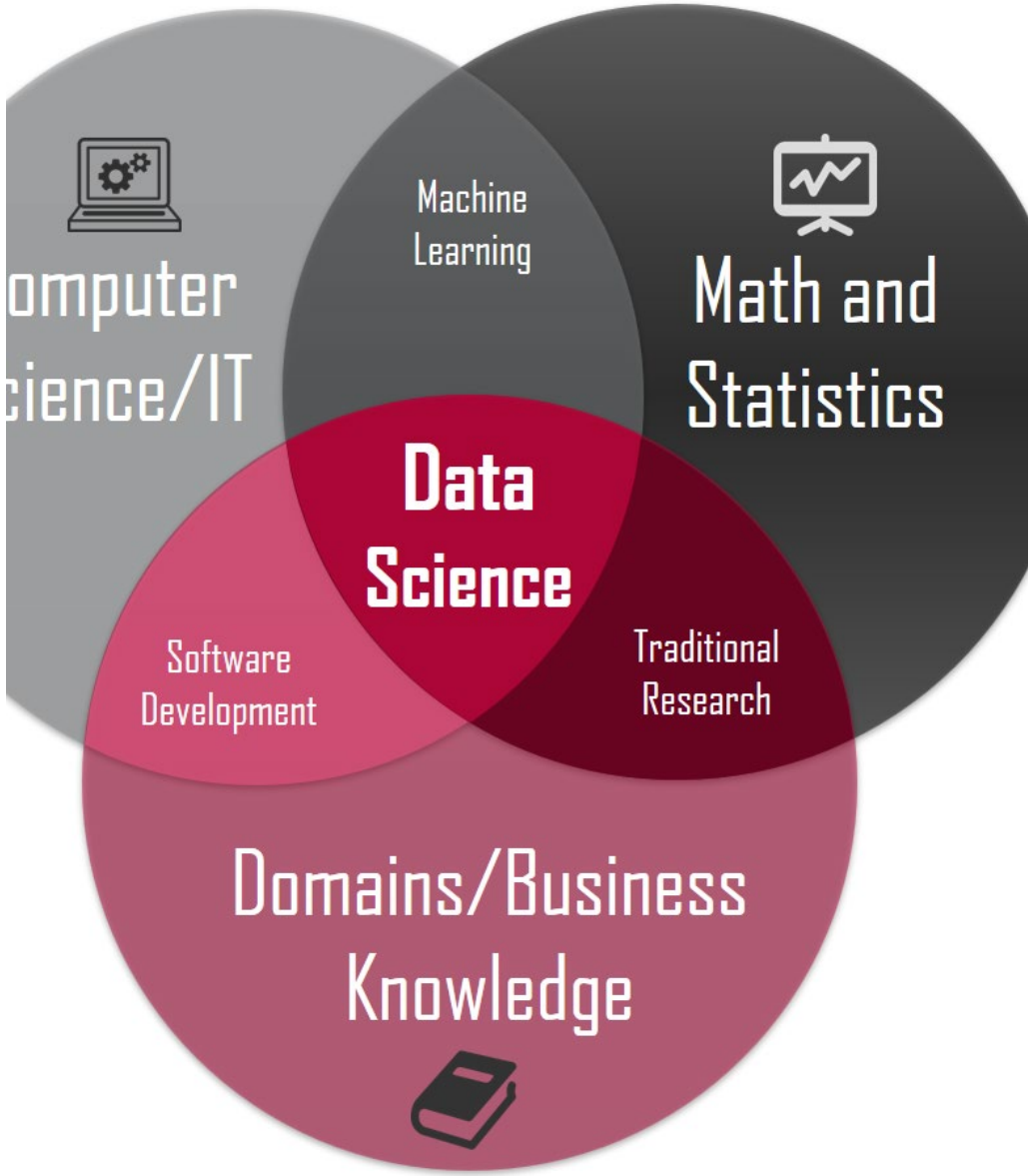


https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

THE UNIVERSITY OF WESTERN AUSTRALIA

STAT 1400

Statistics for Science

Lecture Week 2

Dr Darfiana Nur

Department Mathematics and Statistics

Aims of this week

- Aim 1 Data or Information
  - Variation
  - Types
  - Sampling
- Aim 2 Graphical Summaries
- Aim 3 Describing Distributions – 3S (Histogram)
- Aim 4 Introduction to R (continued) – tidyverse; ggplot

References: Moore et al (2021) Chapter 1

Aim 1
Data or Information

- 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***.

Your answer

1

2

3

4

.

.

10

Variability or Uncertainty

THE UNIVERSITY OF WESTERN AUSTRALIA

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.***

Example 1: Variability or Uncertainty

THE UNIVERSITY OF WESTERN AUSTRALIA

- Consider a machine that makes steel rods for use in optical storage devices. The specification for the diameter of the rods is 0.45 ±0.02 cm.
- During the last hour, the machine has made 1000 rods.
- The quality engineer wants to know approximately how many of these rods meet the specification. He does not have time to measure all 1000 rods. So he draws a random sample of 50 rods, measures them, and finds that 46 of them [ie (46/50)=92%] meet the diameter specification.
- How do the diameters of the rods vary? The variation between the rods might be related to:
  - actual differences between rods due to different machines or operators
  - measurement error.

Example 2: Variability or Uncertainty

Consider a population of bandicoots in a two square kilometre area of rural Australia.

How do **the weights** of the bandicoots **vary**?

List all the possible **sources of variation** that can affect a bandicoot's **weight**.

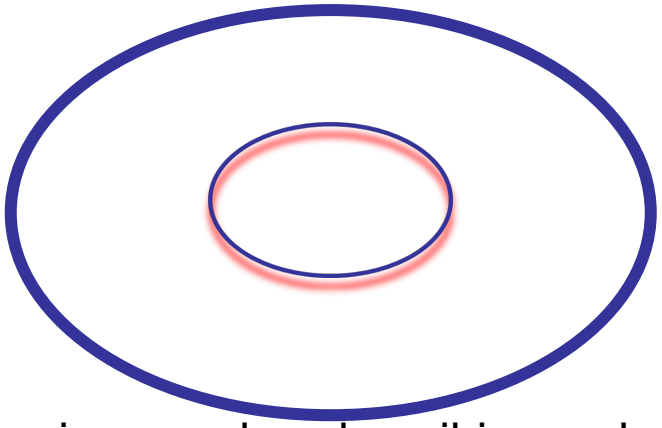
It might help to consider those factors that might

- cause bandicoots in this area to weigh more or less than bandicoots in the rest of Australia;
- cause bandicoot A to have a different weight compared to bandicoot B;
- cause variations in the weight of the same bandicoot.

Population vs Sample

Population

- **Population:** The entire group of individuals in which we are interested but can't usually assess directly.
- Example: All humans, all working-age people in WA, all tertiary students in Western Australia



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

Sample

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

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

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 3

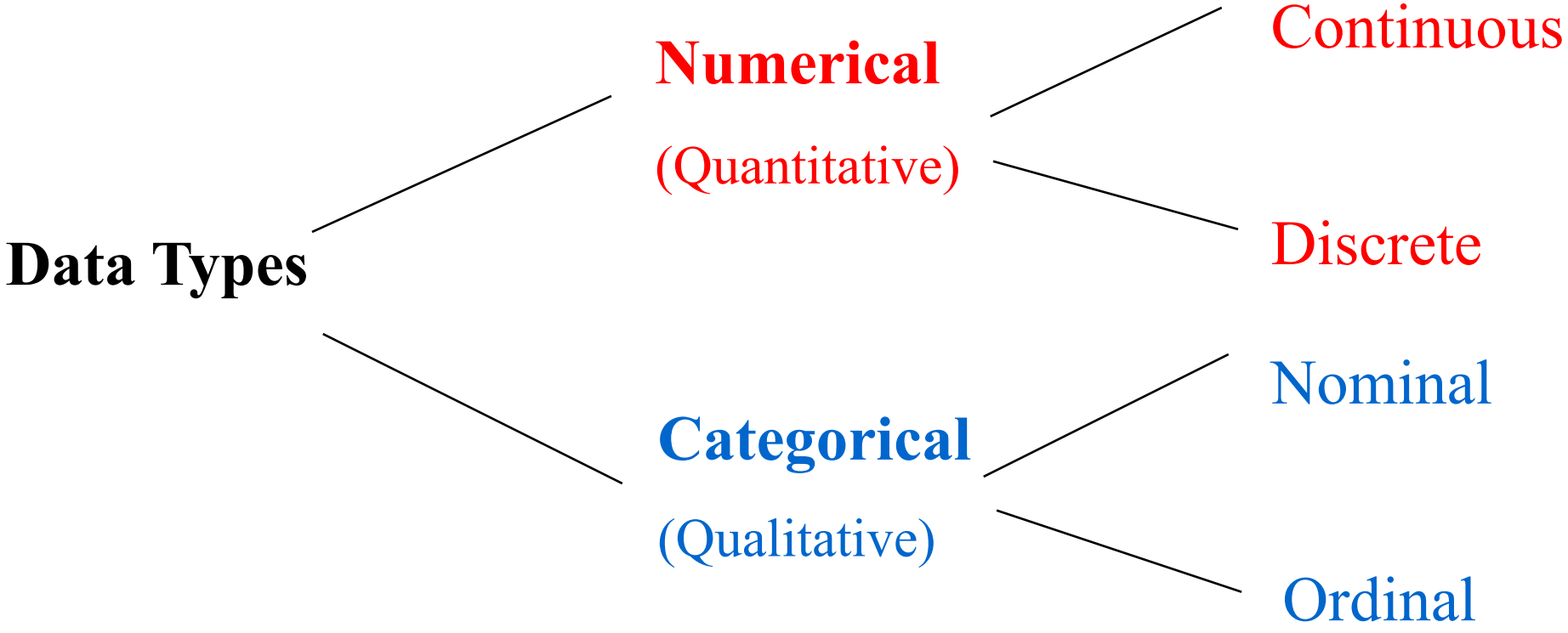
A study is conducted about the average age of a *random sample* of **50 students** from all students at the University of Western Australia.

- The age of each student is an *individual observation* or item of the data.
- The ages of 50 students together is a *sample* of observations.
- The ages of all students at Flinders University is the *population* of observations.
- The procedure of collecting the sample of students from the population is called the *sampling procedure*.
- The age of the students is the *variable* of interest in this study.

"In statistics usually we study the sample rather than the whole population."

Data Types

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



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: 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: 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 **ord**er)
    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 4:

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 5:

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

Number of Facebook/Twitter/LinkedIn users at UWA

In-class Exercise 2

1. Data on number of Facebook users by country

Type 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?

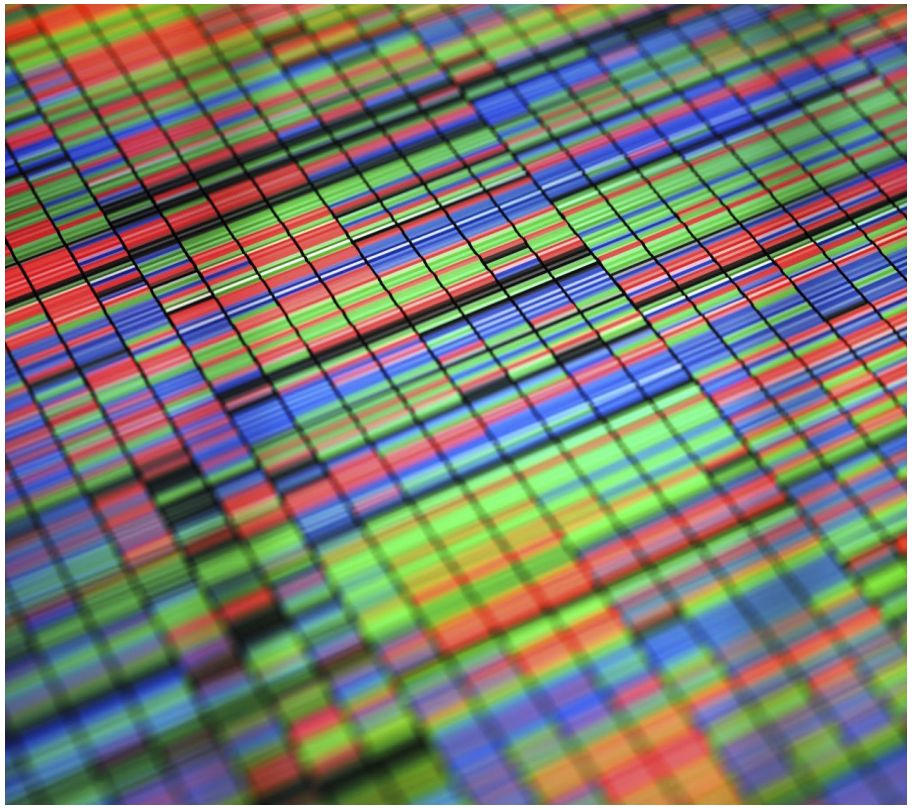
Categorical – ordinal (ordered)

Aim 2 Graphical Summaries (Moore et al Chapter 1.1)

Always, always, always,
always graph your data!

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

Charts for types of variables
CATEGORICAL

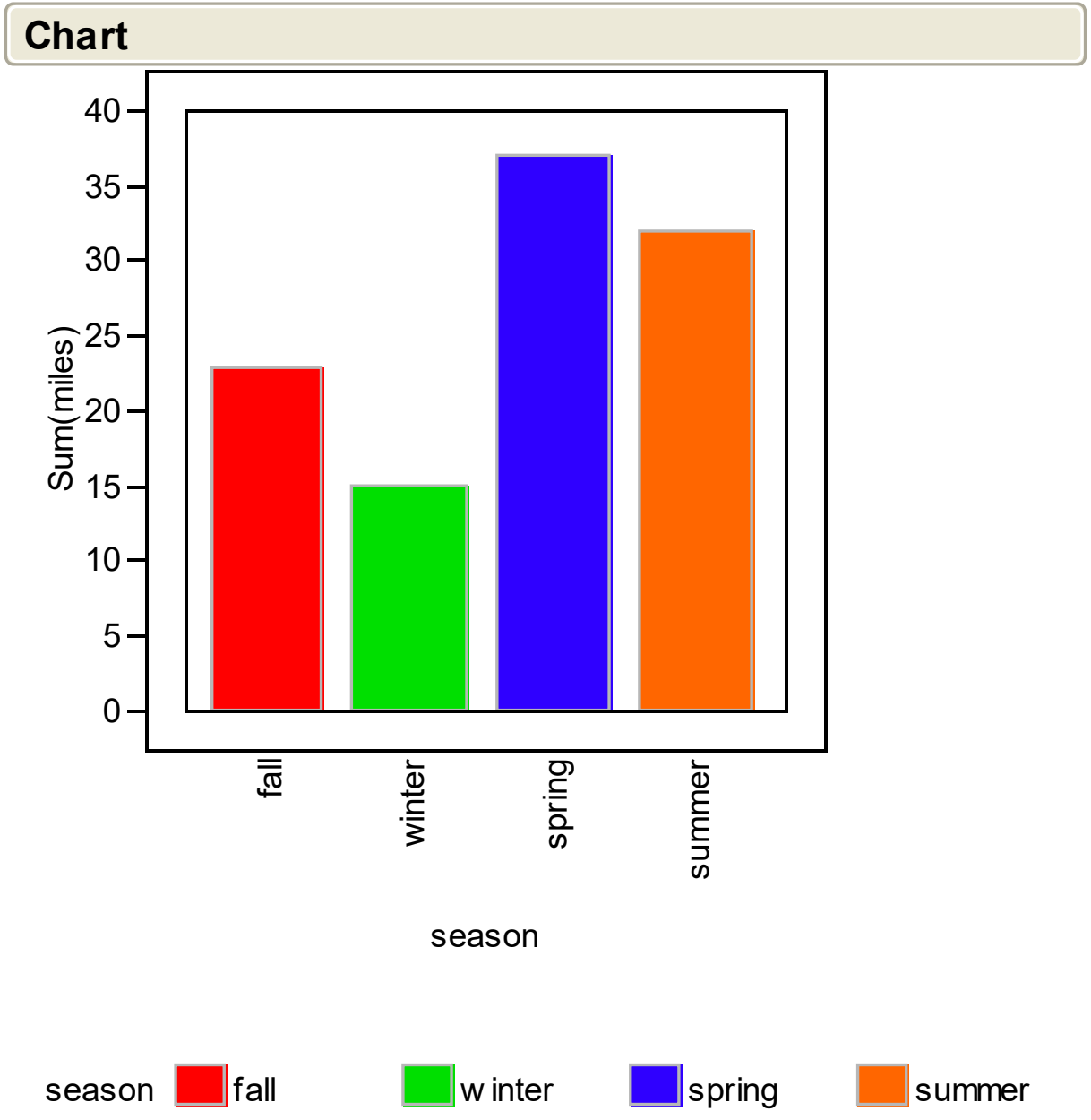


Ordinal variable

- Bar chart

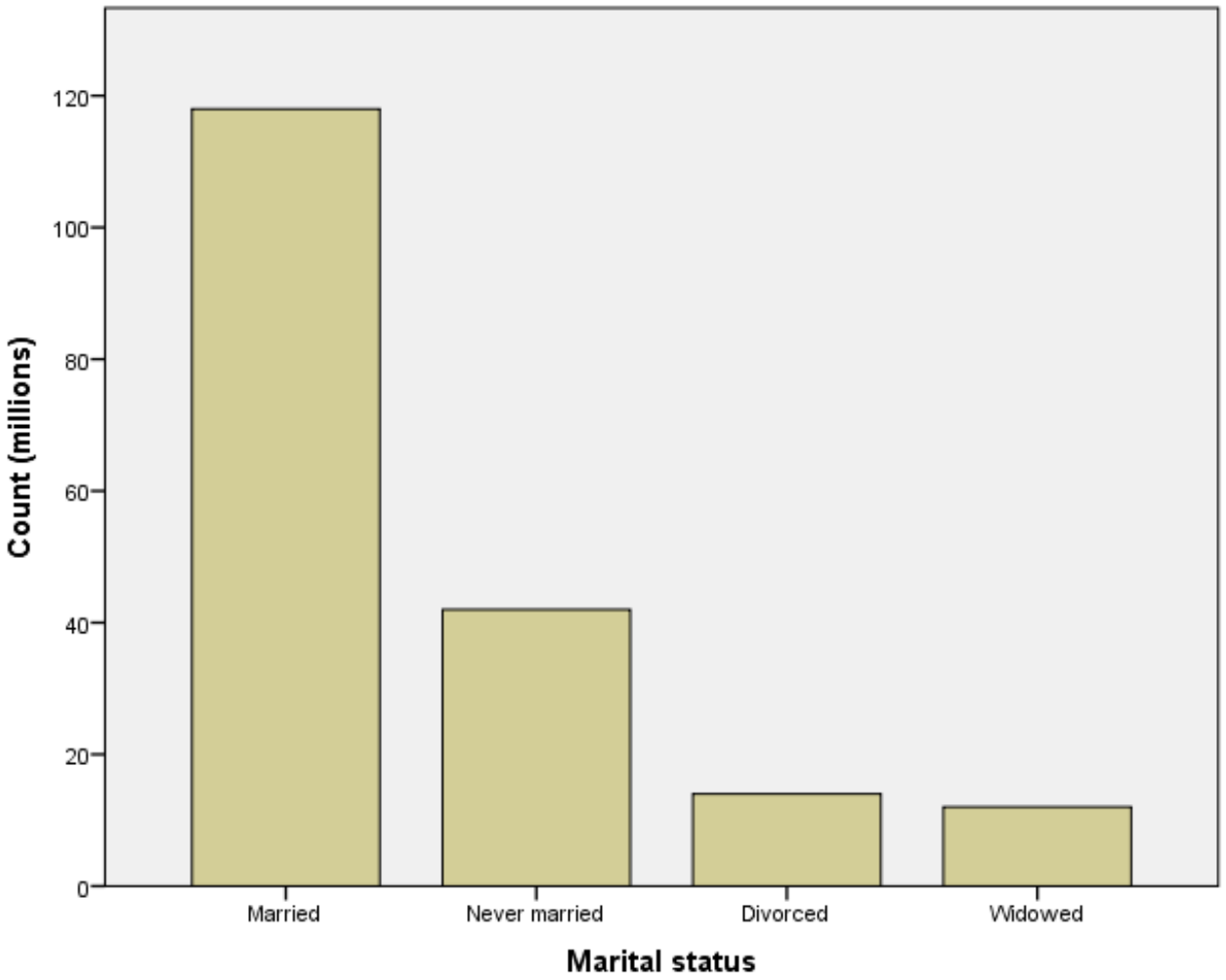
Nominal variable

- Pareto chart
- Pie chart



THE UNIVERSITY OF WESTERN AUSTRALIA

Bar chart (categorical – ordinal)



Pareto chart
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.)

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

Example 6 (Moore et al 2017): Top 10 causes of death in the United States 2006

THE UNIVERSITY OF WESTERN AUSTRALIA

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.

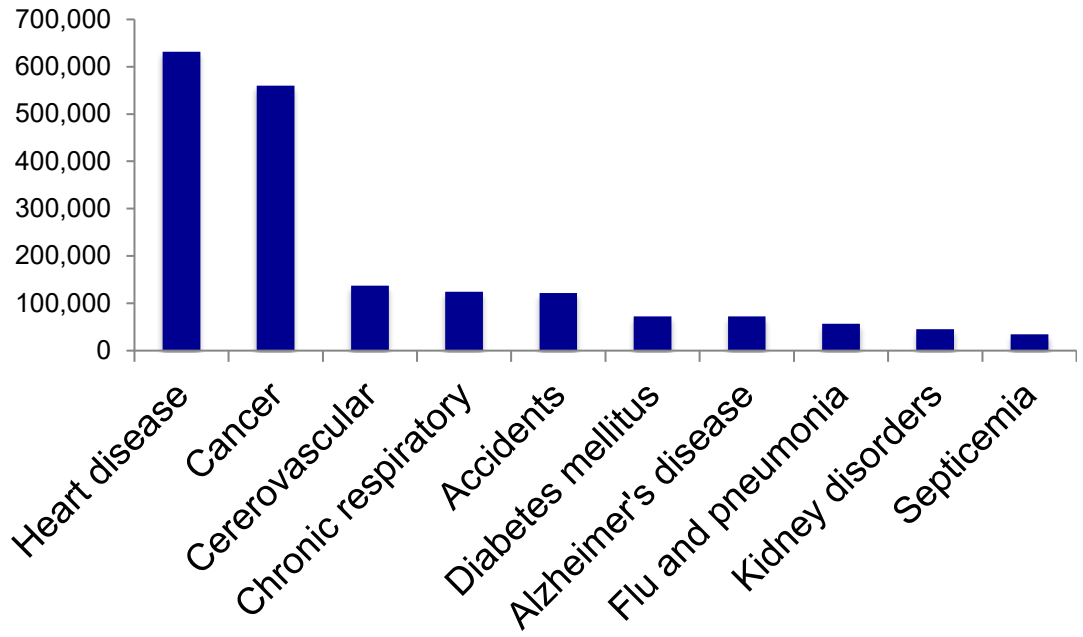
Rank	Causes of death	Counts	% of top 10s	% of total deaths
1	Heart disease	631,636	34%	*26%*
2	Cancer	559,888	30%	*23%*
3	Cerebrovascular	137,119	7%	*6%*
4	Chronic respiratory	124,583	7%	*5%*
5	Accidents	121,599	7%	*5%*
6	Diabetes mellitus	72,449	4%	*3%*
7	Alzheimer's disease	72,432	4%	*3%*
8	Flu and pneumonia	56,326	3%	*2%*
9	Kidney disorders	45,344	2%	*2%*
10	Septicemia	34,234	2%	*1%*
	All other causes	*570,654*		*24%*

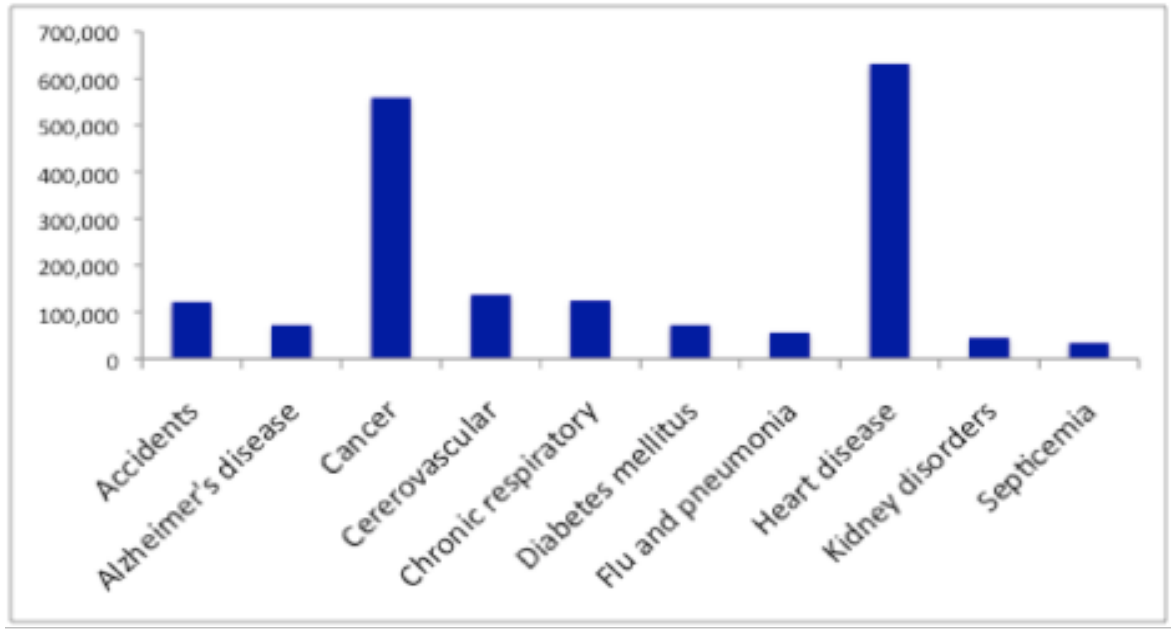
Pareto charts

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





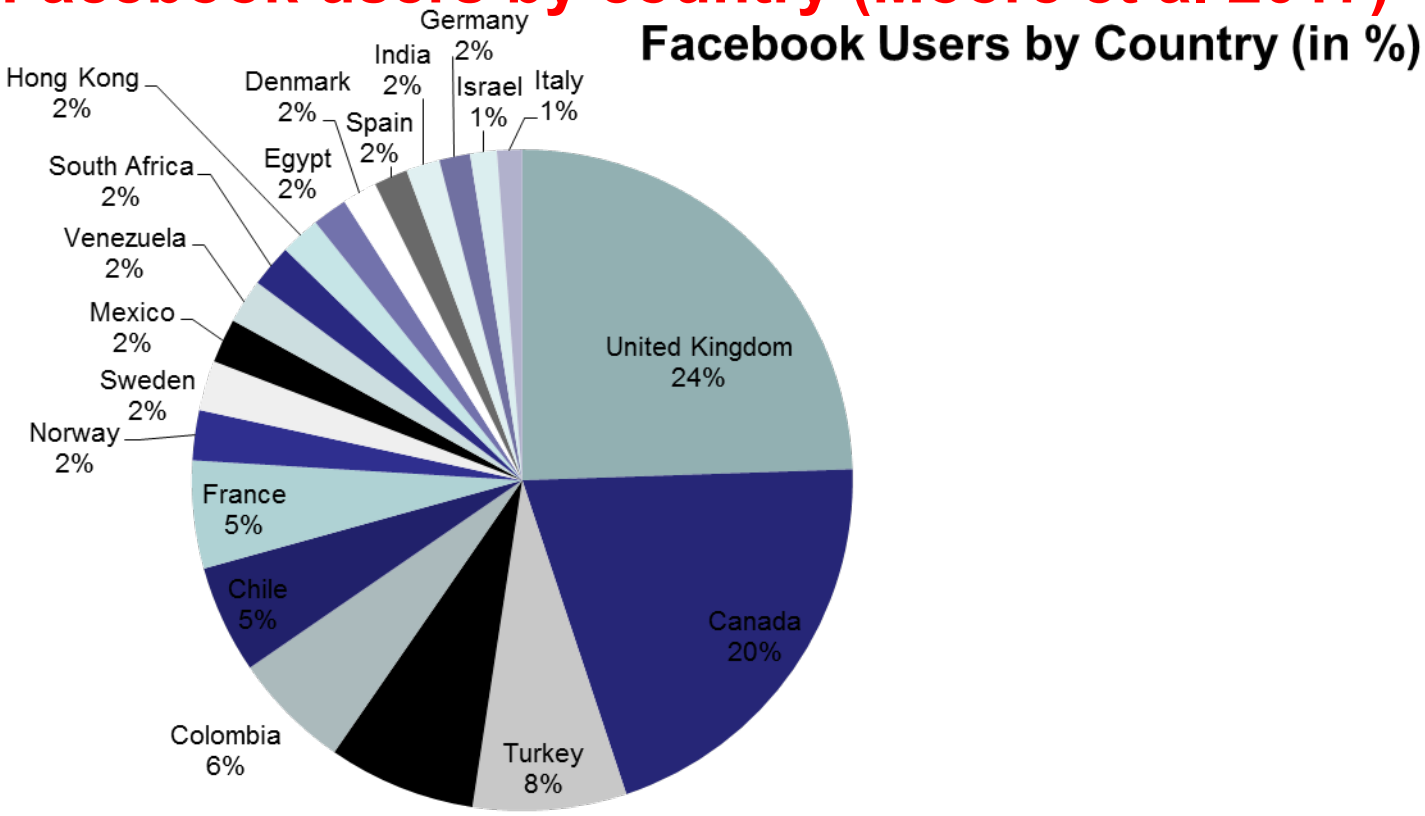
Sorted alphabetically
→ Much less useful

Pie charts

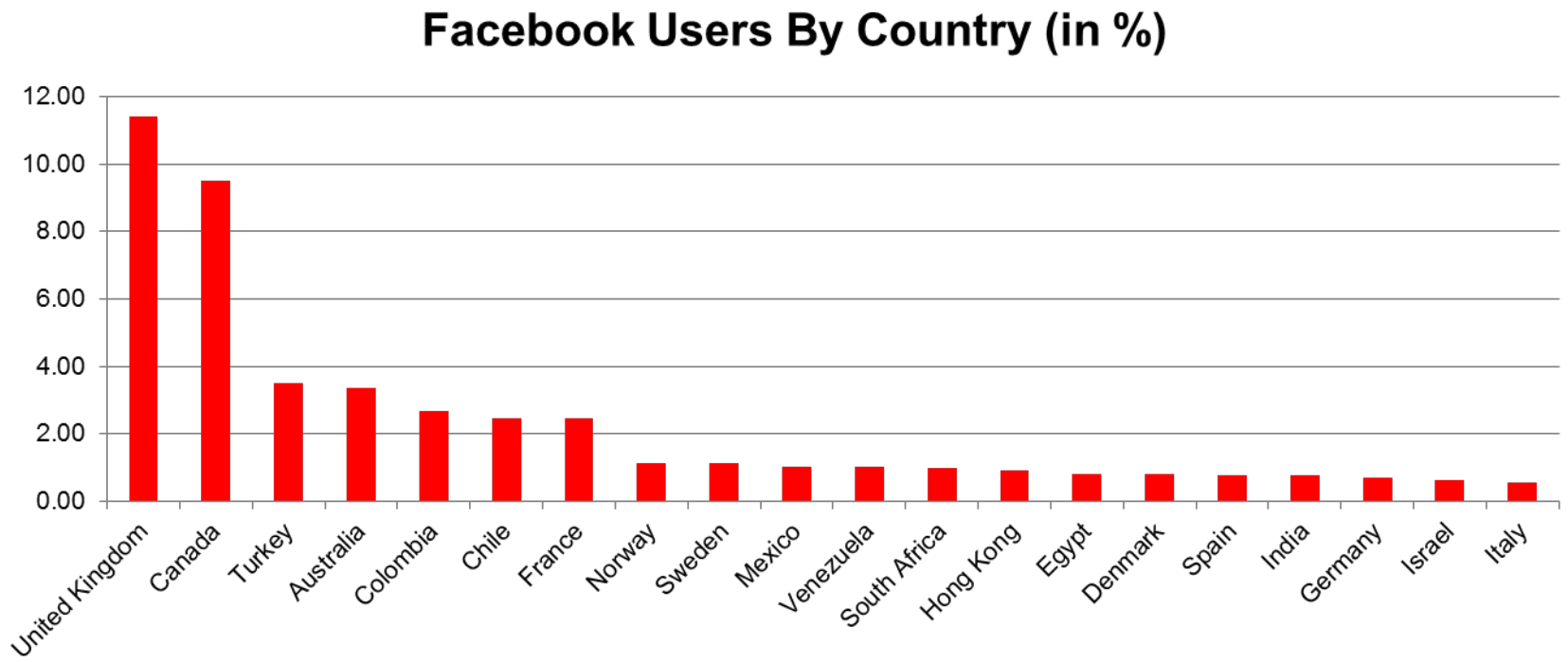
Each slice represents a piece of one whole.

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)



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.

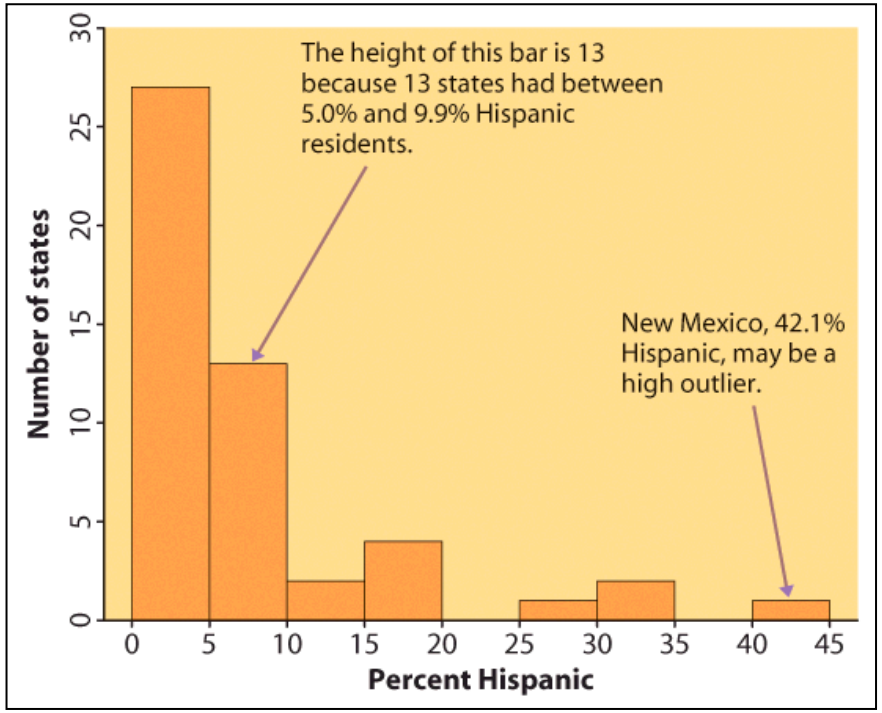
- **Boxplot**

Provides 5 number summary

Histograms

The range of values that a variable can take is divided into equal size intervals.

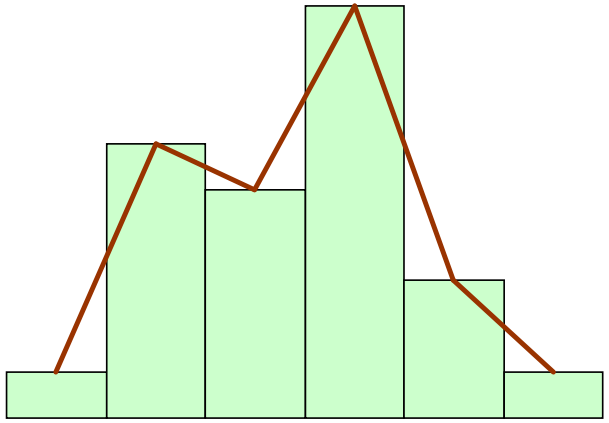
The histogram shows the number of individual data points that fall in each interval.



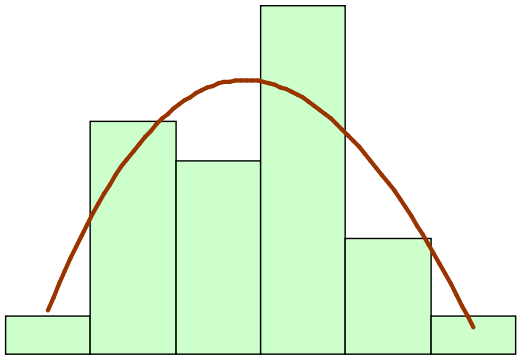
- The first column represents all states with a Hispanic percent in their population between 0% and 4.99%. The height of the column shows how many states (27) have a percent in this range.
- The last column represents all states with a Hispanic percent in their population between 40% and 44.99%. There is only one such state: New Mexico, at 42.1% Hispanics.

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 each column
→ too detailed



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

How to create a histogram

1. Divide the possible values into classes or intervals or bins of equal widths.
2. Count how many observations fall into each interval/bin. Instead of counts, one may also use percents.
3. Draw a picture representing the distribution—each bar height is equal to the number (or percent) of observations in its interval.

rule of thumb: start with 5 to 10 Bins

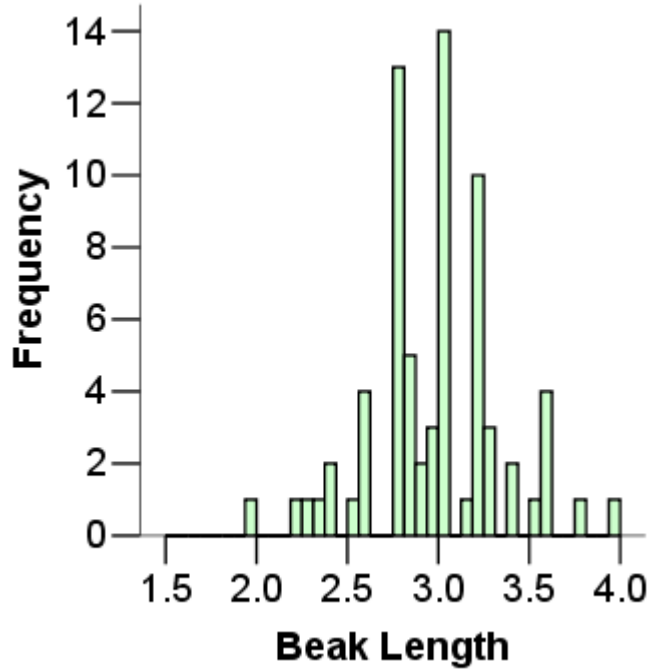
Look at the distribution and refine your bins (There isn't a unique or "perfect" solution)

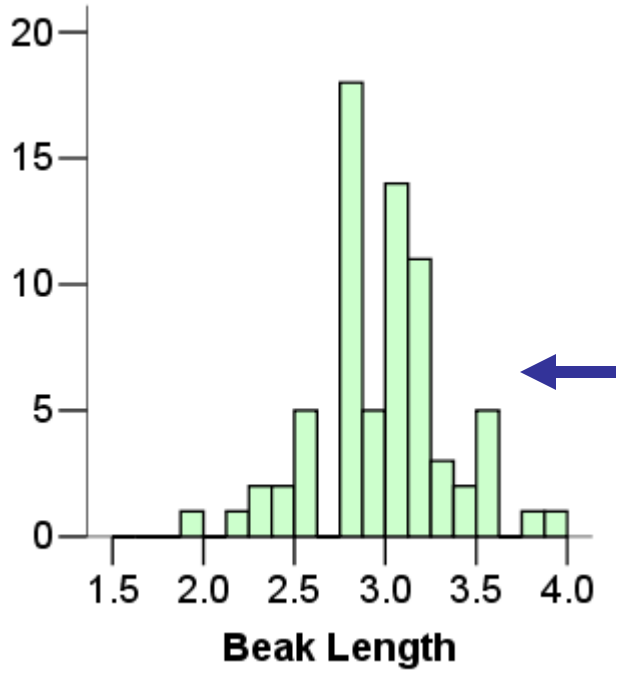
It is an iterative process – try and try again. What bin size should you use?

Not too many bins with either 0 or 1 counts

Not overly summarized that you lose all the information

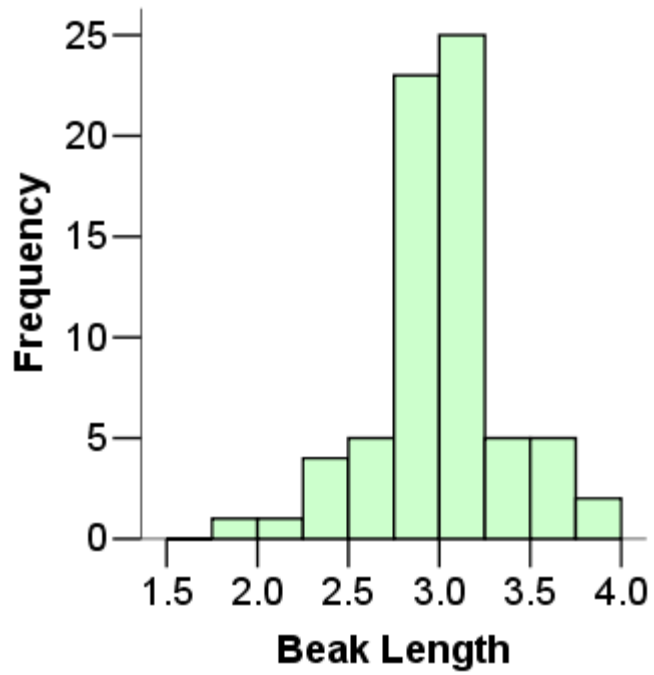
Not so detailed that it is no longer summary

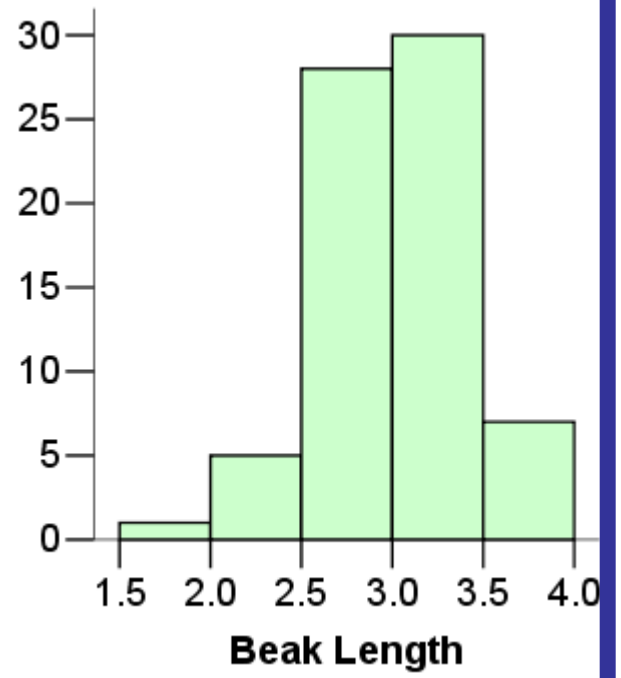




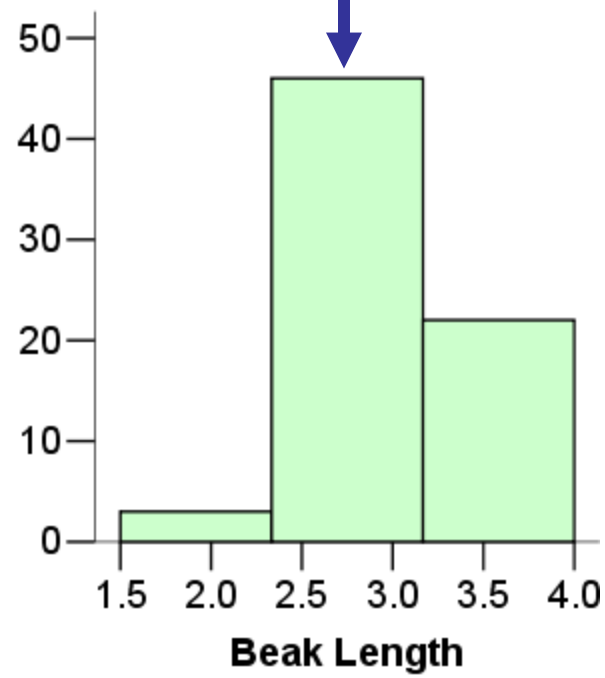
Same data set

Not summarized enough





Too summarized



Example 8. IQ data

Moore et al 2017 Chapter 1

TABLE 1.3

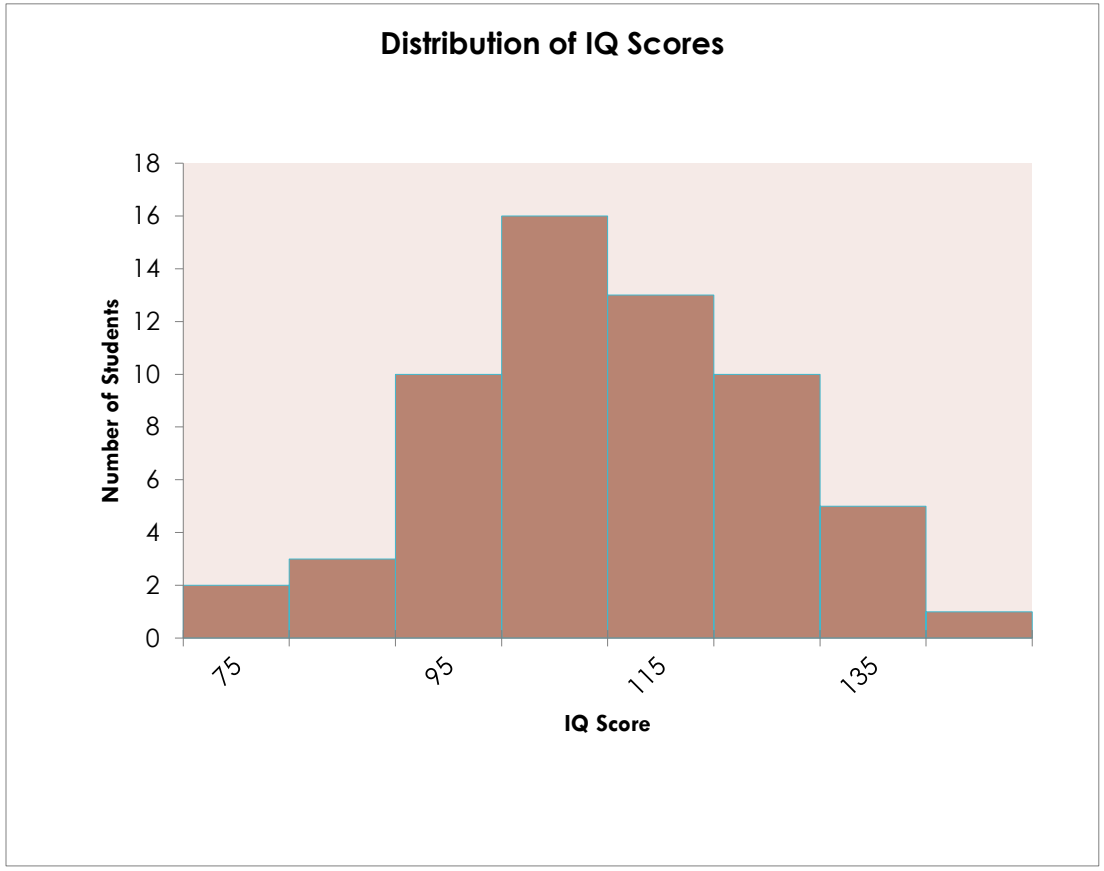
IQ test scores for 60 randomly chosen fifth-grade students

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

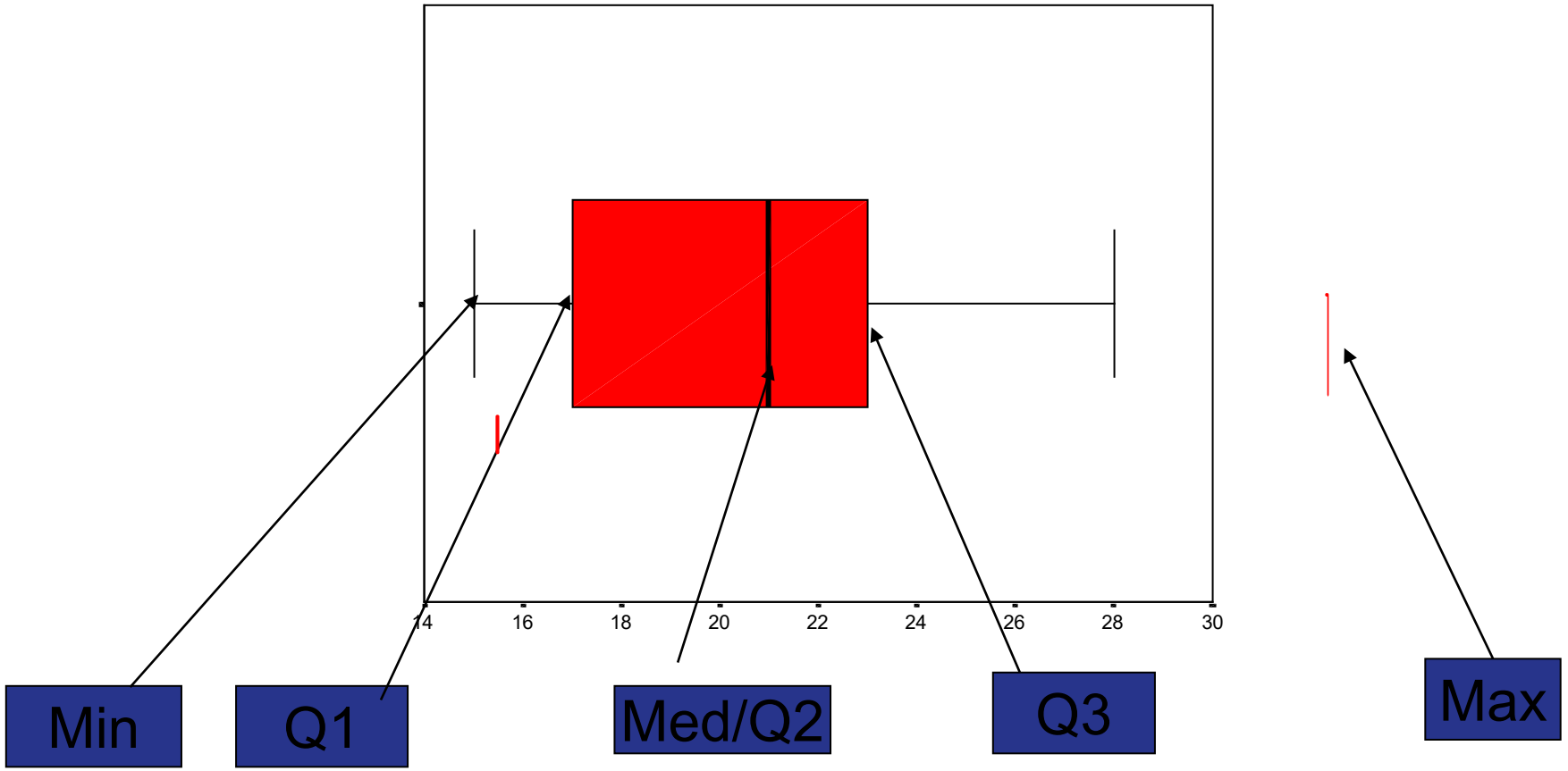
Histograms: IQ data



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

Box plot

Provides 5-number summary



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.

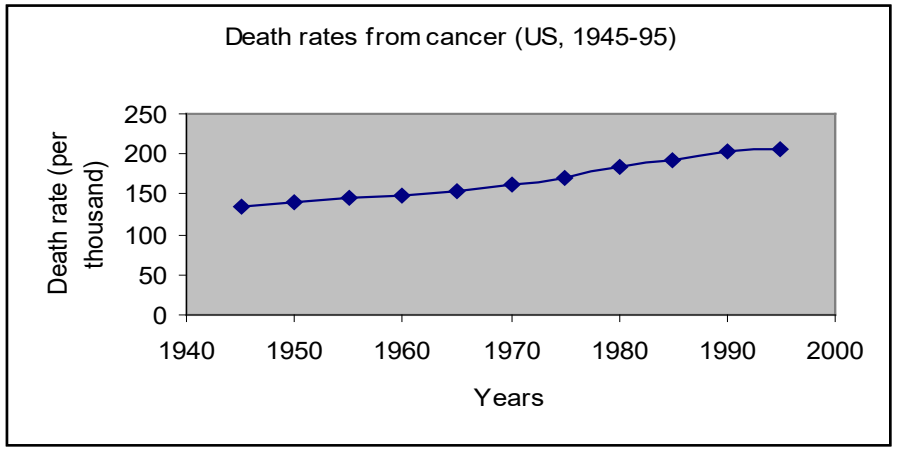
- Do they start at zero?
- are they linear?

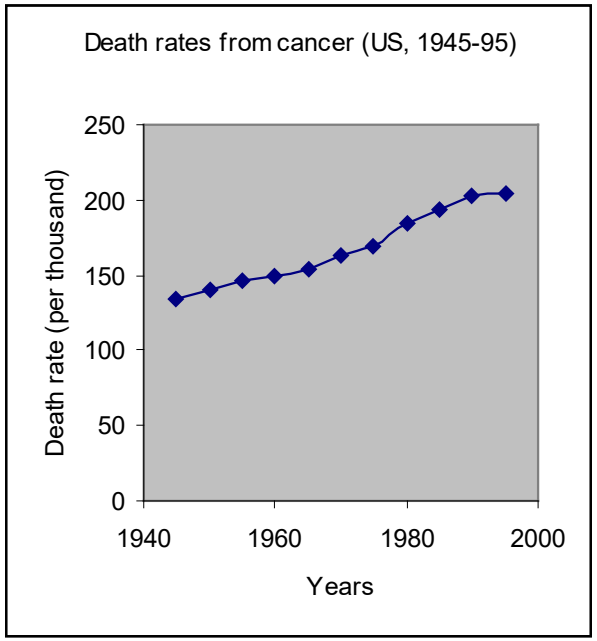
Is there misleading chart junk, effects or perspective?

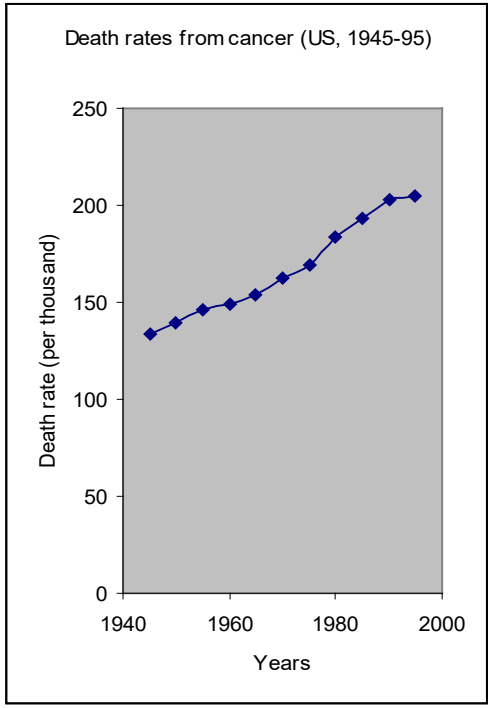
Is the graphical message relevant?

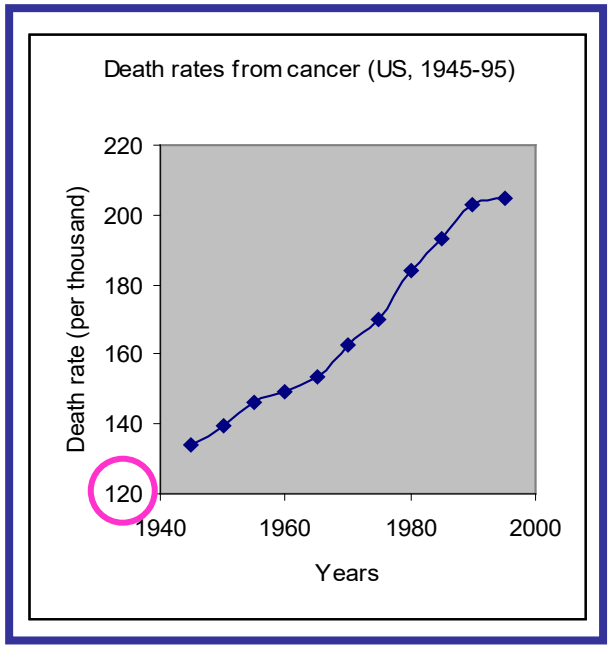
Scales matter

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

Summary 1: Types of variables

THE UNIVERSITY OF WESTERN AUSTRALIA

Numerical

Continuous

(A continuum of values is possible)

e.g. Height (m)

1.86
2.34
1.57

e.g. Age (years)

15
24
35

Discrete

(Can only take on discrete values)

e.g. Count of how many

Google's daily visits (in millions)

7
2

Categorical

Ordinal

(Categories that have order)

e.g. SIZE

Small
Medium
Large

e.g. SIZE

<5cm
5-10cm
>10cm

Nominal

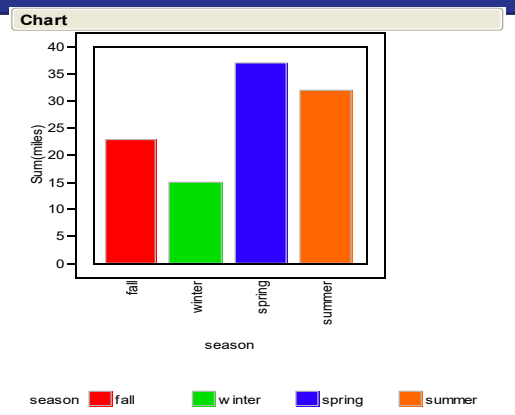
(Categories with no order)

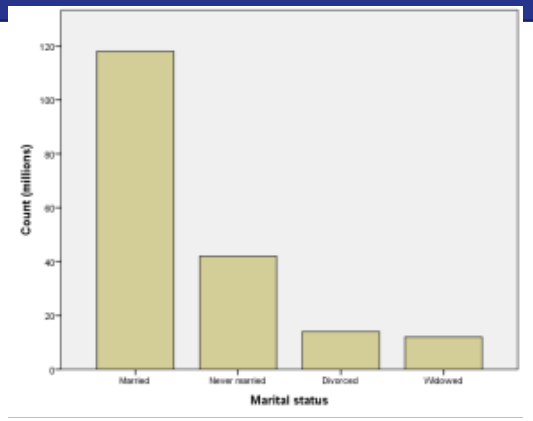
e.g. social media type

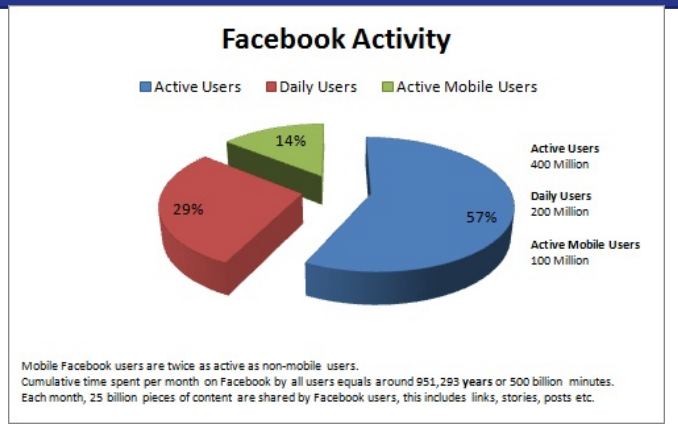
Facebook
Twitter
Myspace
Bebo

Summary 2: Graphs

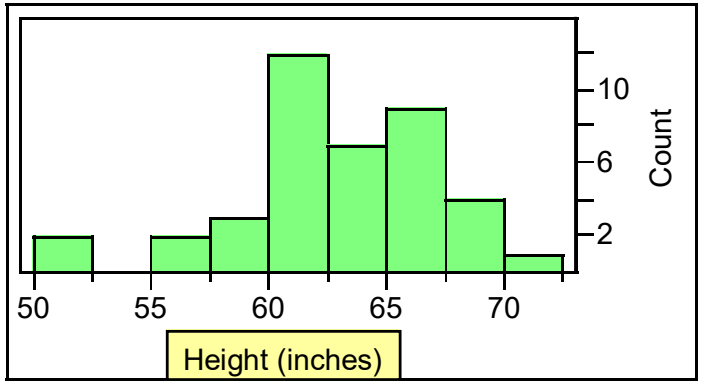
- bar charts (ordinal), pareto or pie charts (nominal)

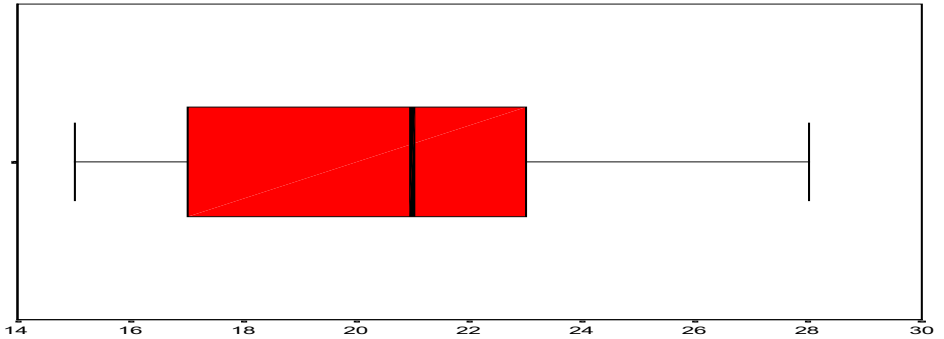






- histograms, box plots (numerical)





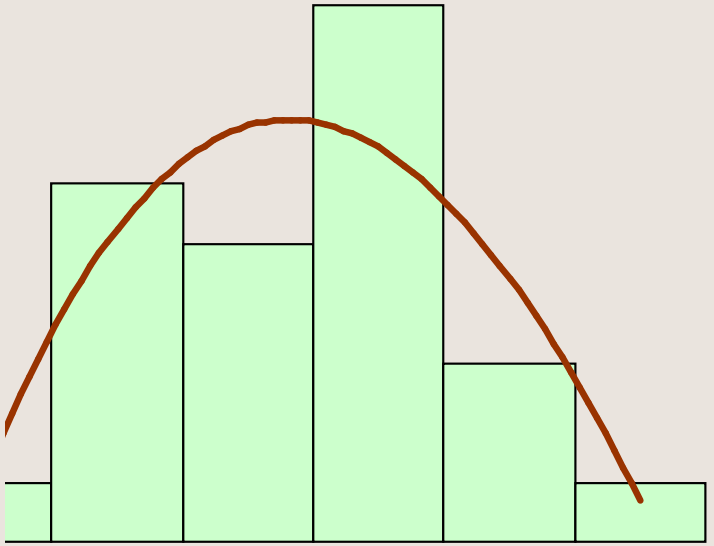
THE BIG PICTURE
Which statistical modeling?

Type of objective	Type of data	Statistical method/model
RELATIONSHIPS	2 numerical (numerical explanatory, numerical response)	Linear regression
COMPARISONS	1 categorical (2 sub-categories) – 1 sample	z-test for a proportion
	Numerical – 1 sample from 1 population	One sample t-test
	Numerical – 1 sample paired	Paired t-test
	Numerical – 2 samples from 2 independent populations	Two sample t-test

Aim 3 Describing Distributions – 3S

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

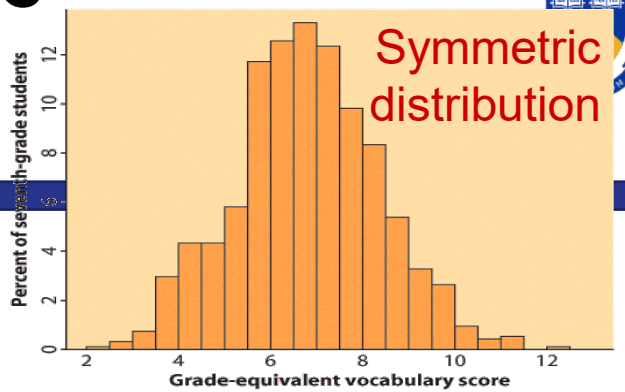
We can describe the overall pattern of a histogram by its Shape, (S)center, and Spread (3S).



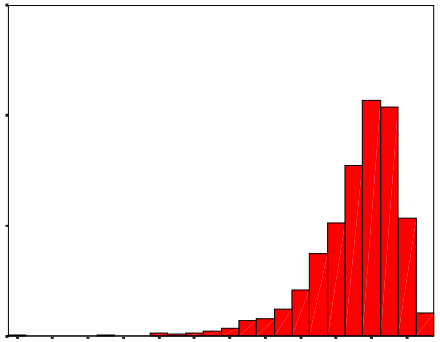
Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

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

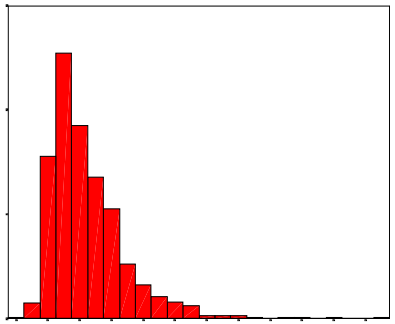


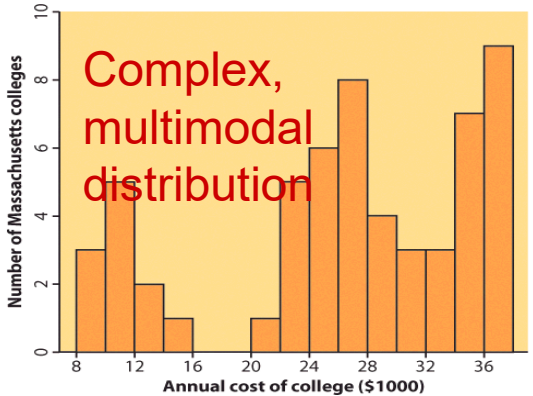
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side.
- It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.



Skewed to the left

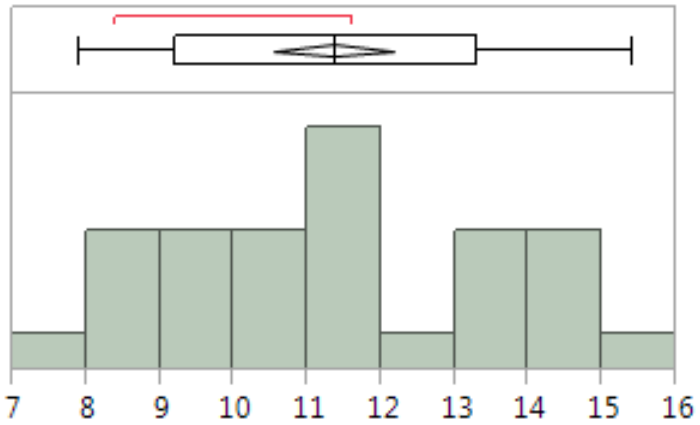
Skewed to the right



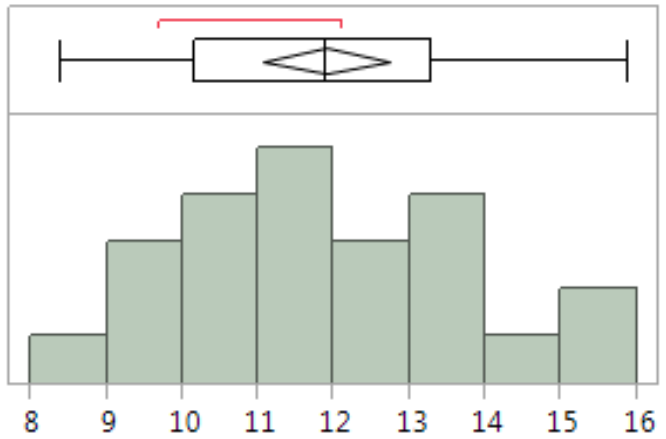


- Not all distributions have a simple overall shape, especially when there are few observations.

Example 9: SCE Rates



Control/Normal



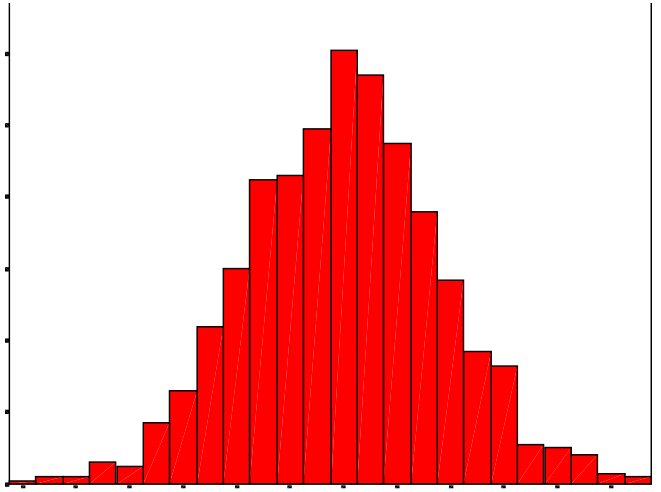
Smoker

- A measure of the extent of sister-chromatid exchange (SCE) tells us whether the cell is suffering stress.
- For each person, an SCE rate measured in number of exchanges per mitosis is obtained, and a potential problem exists when this is high.
- This study conducted by Seshadri et al (1982), who wanted to see the effects that drugs had on the SCE rates of mothers and their new babies.
- These drugs included nicotine, alcohol, hair dyes and more.
- In this experiment, the control group was 30 healthy mothers who were not exposed to any of the drugs in the investigation. They are referred to here as the "Normal Mothers."
- The SCE rates determined from blood samples of 23 "Smoker Mothers".

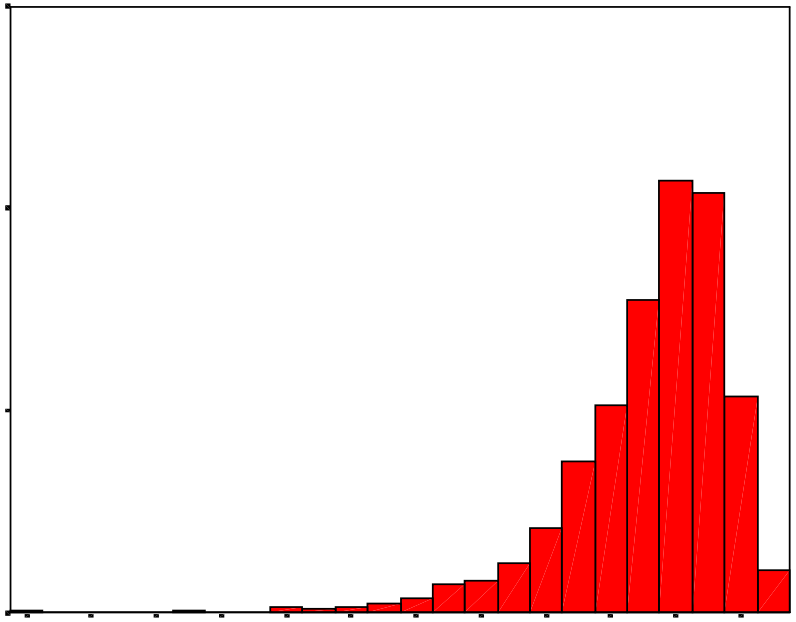
Distributional Shape

THE UNIVERSITY OF WESTERN AUSTRALIA

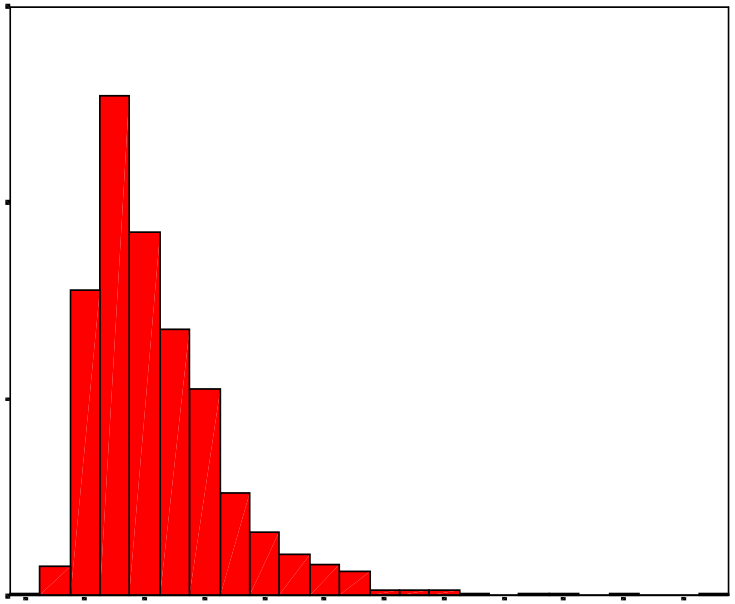
Symmetric



Skewed to the left



Skewed to the right



THE UNIVERSITY OF WESTERN AUSTRALIA

Aim 4

Introduction to R (Packages): tidyverse;ggplot2

Computer Lab 2