Introduction to data analysis (CMM020)

- Statement for Audio and Video Learning Resources
- Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.
- If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.

Introduction to Data Analysis

Ref: D. Diez, M. Cetinkaya-Rundel, C. Barr, *OpenIntro Statistics* (4th Edition), OpenIntro, 2019 [available for download at www.openintro.org/book/os/, accessed 27/01/2020]

Content

- Data
- Probability basics
- Statistics
  - Descriptive statistics for univariate distributions

Data

- Nominal
  - values are symbolic, e.g. desk, table, bed, wardrobe
  - no relation between nominal values
  - Boolean attributes are a special case
    - 0 and 1 or True and False
  - also called categorical, enumerated or discrete
- Ordinal
  - values are ordered, e.g. small, medium, large, x-large
  - small < medium < large < x-large
  - but difference between 2 values is not meaningful

… data

- Interval – Discrete quantitative data
  - Quantities are ordered
  - measured in fixed equal units
    - E.g. years 2001, 2002, 2003, 2004
  - difference between values meaningful: 2005 - 2004
    - but sum or product is not always meaningful: 2005 + 2004
- Ratio – Continuous quantitative data
  - Depth of the North sea in different areas …
  - 1.233m, 100.785m, …

Nominal vs. Ordinal

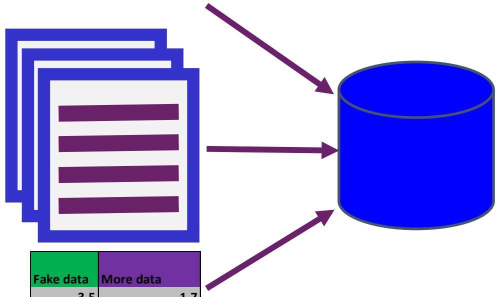
- Questionnaire responses as nominal
  - Strongly Disagree
  - Disagree
  - Neither agree nor disagree
  - Agree
  - Strongly agree

- Questionnaire responses as ordinal
  - Strongly Disagree < Disagree < Neither agree nor disagree
  - Neither agree nor disagree < Agree < Strongly agree

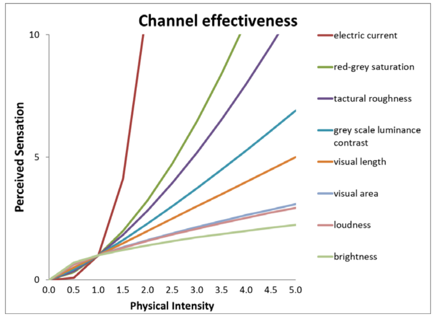
Data Processes

- Collection
  - From one or more sources
  - In one or more formats
  - Probably over a period of time
- Preparation
  - Get the right data into the right format for the desired analysis
- Analysis and visualisation
  - Make sense of the data

	Electric	Red-grey saturation	Tactual roughness	Grey scale luminance contrast
n=	3.5	1.7	1.5	1.2
0	0	0	0	0
0.5	0.088388	0.307786103	0.353553391	0.435275282
1	1	1	1	1
1.5	4.133514	1.99230186	1.837117307	1.626707657
2	11.31371	3.249009585	2.828427125	2.29739671
2.5	24.70529	4.747861206	3.952847075	3.002811085
3	46.76537	6.47300784	5.196152423	3.737192819
3.5	80.21178	8.412317884	6.547900427	4.49657305
4	128	10.55606329	8	5.278031643
4.5	193.3053	12.89618556	9.545941546	6.079320173
5	279.5085	15.42584657	11.18033989	6.898648307



Fake data	More data
3.5	1.7
0	0
0.088388	0.307786103
1	1
4.133514	1.99230186
11.31371	3.249009585
24.70529	4.747861206
46.76537	6.47300784
80.21178	8.412317884
128	10.55606329



Data collection

- From one or more sources
- In one or more formats
- Probably over a period of time
- May use both internal and external sources
  - E.g. sales data (internal to company)
    weather data (external, e.g. from met office)
- Collection itself may need to be planned.

Data preparation

- Preparing data for analysis and visualisation is difficult and demanding
  - Data may need assembling, integrating, aggregating and cleaning
    - if data set is huge, a sample may be used
    - Sample must be representative
- Various types of data may be used
  - Nominal and numeric are most common

… data preparation

- **Wrangling** - transforming data into another format to make it more suitable and valuable for a task
- **Cleansing** (cleaning) - detecting and correcting errors in the data.
- **Scraping** - automatic extraction of data from a data source.
- **Integration** - combining data from several disparate sources into a (useful) dataset.
    - Often challenging due to differences in practices

      E.g. measures by week vs. measures by month

Missing data

- Missing data may be unknown, unrecorded, irrelevant
- Causes
  - Equipment faults
  - Difficult to acquire (e.g. age, income)
  - Measurement is not possible
- The fact that a value is missing may be informative
  - e.g. missing test in medical examination
  - BUT this is NOT usually the case
- Represented in R as *NA*

… missing data

- What action is taken when data is missing?
  - Depends on the data
    - Ignore record
    - Data imputation
      - Replace by average, median, mode (see below)
      - Use machine learning techniques to derive likely value.
        - E.g. same value as most “similar” record

Inaccurate / incorrect values

- **Noise / outliers**
- Errors and omissions which do not affect original purpose of data collection
  - E.g. age of bank customers not important
  - E.g. Customers IDs not important
- Typographical errors in nominal attributes
  - e.g. Pepsi vs Pepsi-cola
- Deliberate errors
  - People may lie about their mental health history
  - Companies may downplay their impact on the environment
- Duplicates
  - Analysis may be very sensitive to this
- Outlier – sudden vibration causes sensor misreading.

Data analysis

- Make sense of data
  - Data exploration
    - Initial examination of data
    - Usually with target questions in mind
  - Data analysis and visualisation

… data analysis

- **Inference**
  - Using sample data to draw conclusions about a population.
- **Prediction**
  - Forecast - future values from past data.
  - Estimate - unseen values from seen values.
- **Causal Analysis (cause – effect)**
  - Does x cause y?
  - Hard!
  - E.g. in London people using umbrellas lived longer
    - Umbrella → longer life? …
    - It not because people use umbrellas, it is because they are wealthier than others
- Present the results – often visualisations
  - To a variety of audiences.
- Visualisation may be used for all tasks

Analysis and Visualisation

- Although taught separately, they are part of the same process
- Visualisation guides analysis and helps ascertain whether results of analysis make sense.
- Visualisation and summary statistics very useful for
  - Exploratory Data Analysis
  - Presentation of Results

Data Analysis and Probability

- What is the most likely explanation of the data?
- If something appears in the data – how likely is it to be real, rather than coincidence?
- Questions like these lead to discussion of **probability**
  - A quantitative expression of how likely some event is.
- Consequently, any statistical analysis involves probability

Probability

- P(A) is the probability of an event A
  - If A is certain to occur, $\boldsymbol{P(A) = 1}$
  - If A cannot occur, $\boldsymbol{P(A) = 0}$
  - Otherwise, $\boldsymbol{0 \leq P(A) \leq 1}$
  - Often converted to percentages, from 0% to 100%
- Variations of the Probability Concept
  - **Classical Probability** involves counting the ways in which an event can occur (often cards, dice, selecting *m* from *n*,…)
  - **Empirical Probability** is estimated from experimental evidence (e.g. 3 successes in 5 trials [3/5] equals 0.6)
  - **Subjective Probability** is an individual's estimate (guessing!)

Probability rules

- Complementary Events
  - Assume A is an event and $\bar{A}$ (not A) is event A not happening

$$\boldsymbol{P(A) + P(\bar{A}) = 1}$$

- General Rule for Combination of Events
  - For events A and B

$$\boldsymbol{P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B) \leq 1}$$

- Mutually Exclusive Events
  - If events A and B cannot occur together then

$$\boldsymbol{P(A\ and\ B) = 0}$$
$$\boldsymbol{P(A\ or\ B) = P(A) + P(B) \leq 1}$$

Conditional Probability

- Conditional Probability
  - **The probability of A given** that B has happened is $P(A|B)$
  - E.g. Probability of a hypothesis being true, given the evidence
    - $P(A|B) = P(A \text{ and } B)/P(B)$
    - $P(A \text{ and } B) = P(A|B) * P(B) = P(B|A) * P(A)$

- Independent Events
  - The probability of A is unaffected by the success or failure of B
    - Observing B tells us nothing about A
  - Then A and B are **independent** events
    - $P(A|B) = P(A|\bar{B}) = P(A)$

$\bar{B}$ means "not B", i.e. B is false.

Conditional Probability Example

- A Geologist estimates that the probabilities of striking oil at locations A and B are
  - $P(A) = 0.6$
  - $P(B) = 0.5$
- If the events were independent, then
  - $P(A \text{ and } B) = 0.6 * 0.5 = 0.3$
- But the geologist estimates
  - $P(A \text{ and } B) = 0.4$
- A strike at one location increases the probability of a strike at the other
  - $P(A|B) = P(A \text{ and } B)/P(B) = 0.4/0.5 = 0.8$
  - $P(B|A) = P(A \text{ and } B)/P(A) = 0.4/0.6 = 0.667$

Probability Distributions

- Take an experiment or trial with exactly n possible (mutually exclusive) outcomes
- Then the probabilities of these outcomes sum to one
- The set of probabilities is a discrete probability distribution
- Important standard examples:
  - Binomial (total successes in repeated trials, n=2)
  - [Poisson]
  - Uniform – all probabilities equal to 1/n

Continuous Probability Distributions

- A variable x can take any value in an interval
- $p(x)$ is a probability density function describing the relative probability of values of x
- $P(a < x < b)$ is given by the area under the graph of $p(x)$ between a and b
  - The integral, if you've done calculus
- The total area under the graph is one
- Important Standard Examples
  - Normal (Gaussian or Bell Curve)
  - Uniform over a finite interval e.g. [0,1]

Descriptive Statistics

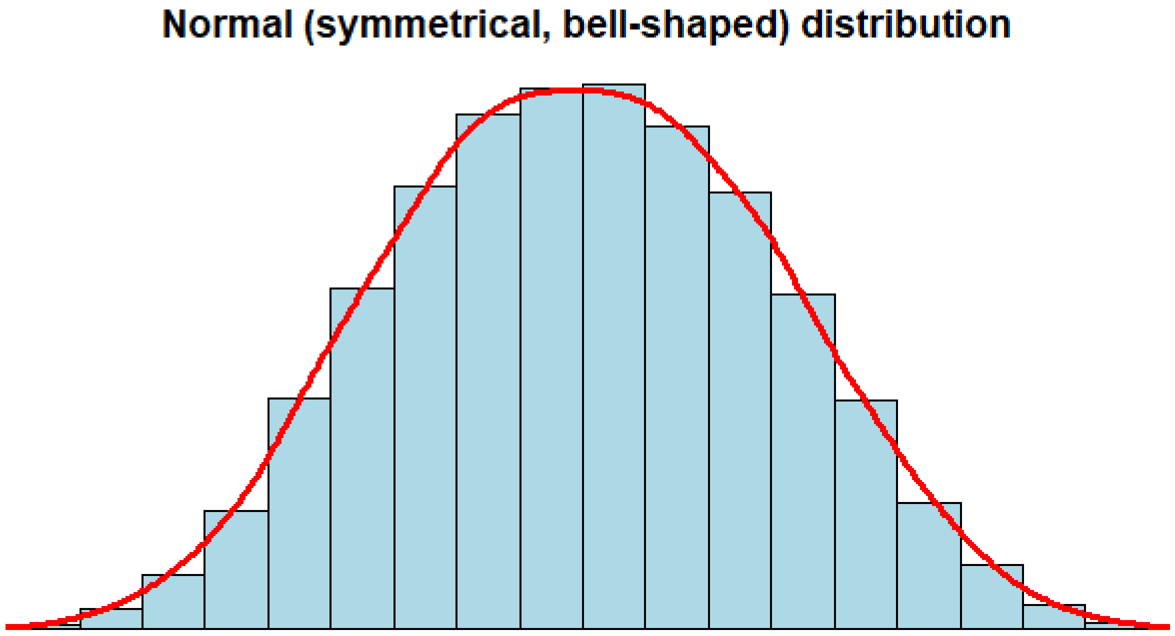
- Visualise Data
  - Tables
  - Charts
- Summarise Data
  - Measures of Location (Averages)
  - Measures of Dispersion (Spread)
  - And many more…
- Main Uses
  - Exploratory Data Analysis
  - Presenting Results

Univariate Distributions

- To examine a univariate (1-D or single variable) distribution:
  - Tabulate the **frequency or relative frequency** of each distinct value
  - Construct a (relative) frequency bar chart or a histogram
- Calculate summary statistics to indicate typical values and spread of values

Normal distribution

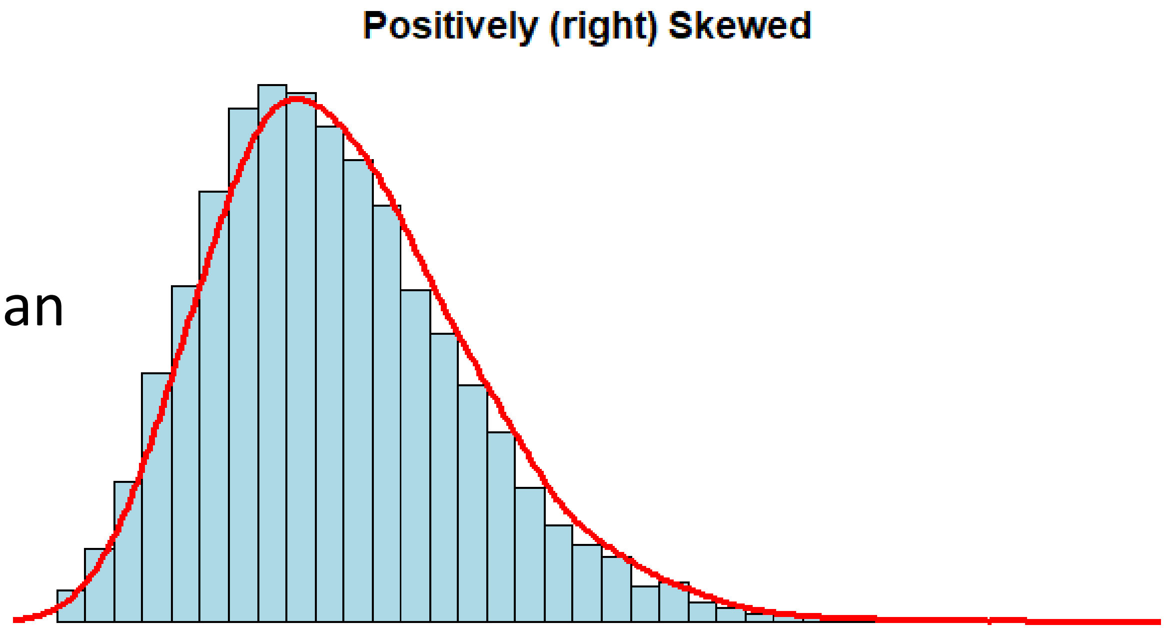
- Normal distribution
  - Gaussian
  - Bell shaped
  - mean=median
- The height indicates the number of values falling in that range.
- Most frequent value(s) – top of bell



Skewed distribution - right

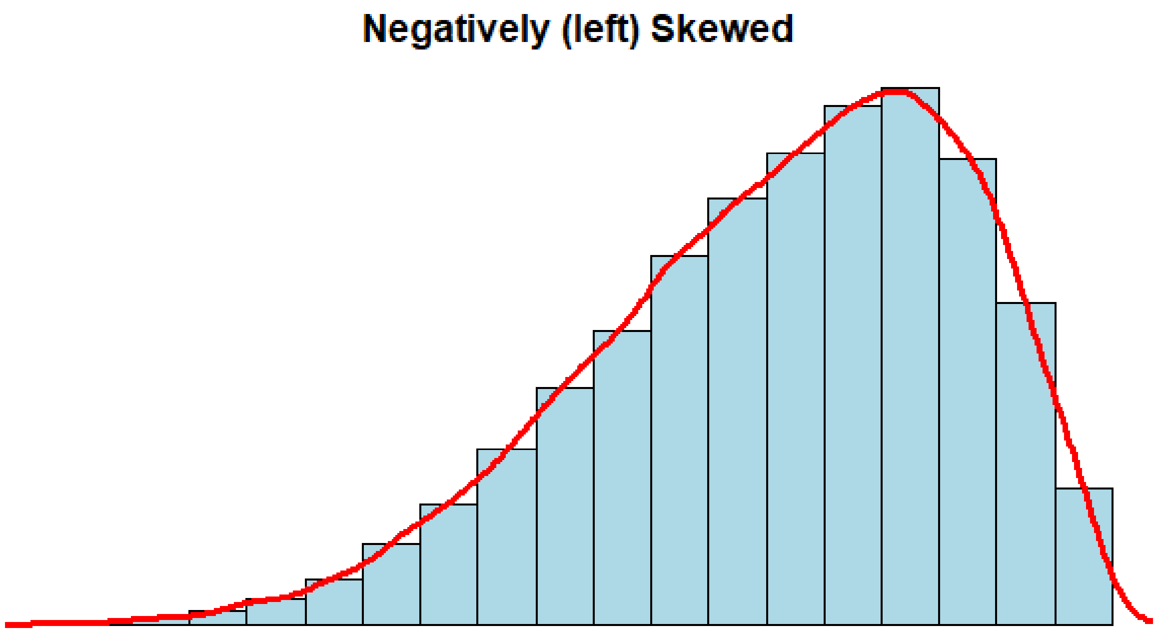
Right (Positive) Skew

- RH tail is longer or fatter
- Top half of data is further from median than bottom half
- Mean to the right of median



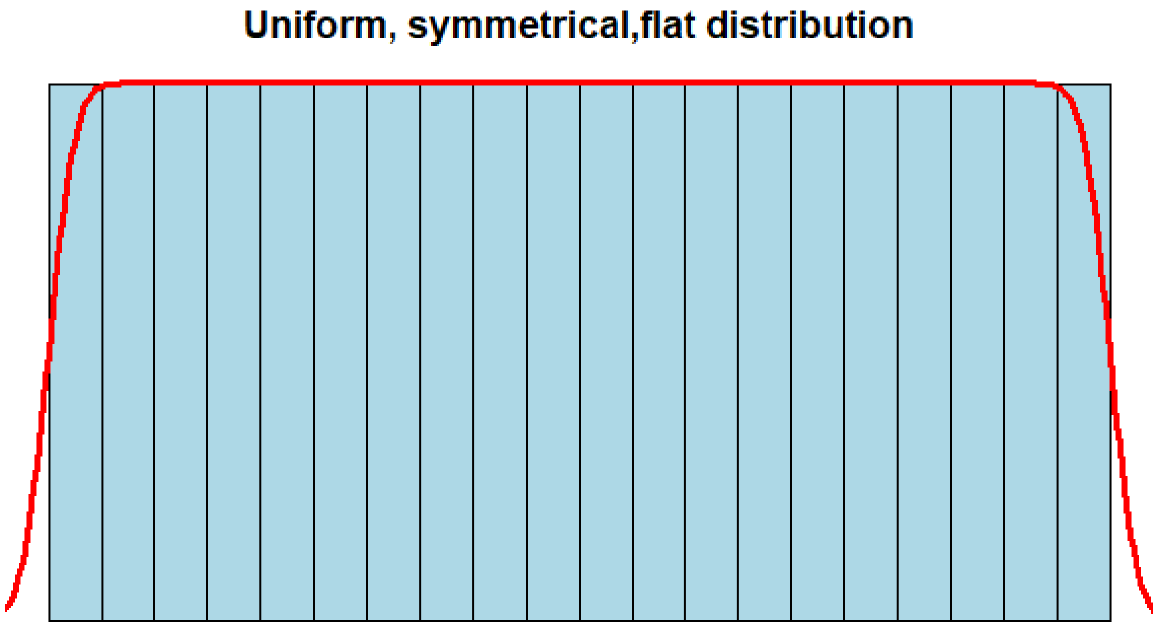
Skewed distribution - left

- Left (negative) skew
  - LH tail is longer or fatter
  - Bottom half of data is further from median than top half
  - Mean to the left of median



Uniform, symmetrical distribution

All values have similar frequency



Statistical Models

- Statistical procedures are usually based on a statistical model of the data and the underlying population.
- These models have probability distributions as inputs and outputs.
- Typically, random variables take values drawn from given probability distributions
- The normal distribution is particularly important.

Commonly Encountered Datasets

- **One-dimensional distribution** (univariate – one column)
  - E.g. Average daily time online – one piece of data per user
- **Two-dimensional distribution** (bivariate – two columns)
  - Number of accesses to website vs. total time spent on it
- **Time Series Data** (one of the dimensions is time)
  - Accesses to a website throughout the year
- **Rectangular Multivariate Data**
  - Several measurements on each class member
  - Can be viewed as a table or treated as a matrix

Univariate (one variable) Distributions

- To examine a univariate (1-D or single variable) distribution:
  - Tabulate the **frequency or relative frequency** of each distinct value
  - Construct a (relative) frequency bar chart or a histogram
- Calculate summary statistics to indicate typical values and spread of values
- Measures divided into
  - Measures of location
  - Measures of dispersion

Measures of Location

- Measures of central tendency
  - Typical value
    - But what is typical?
  - Mean (average) $\bar{x}$
    - Add up data and divide by number of items
  - Median
    - Middle value of ordered data
  - Mode
    - Most common data value
- Quartiles Q1, Q2, Q3
  - Divide data into 4 equally sized groups

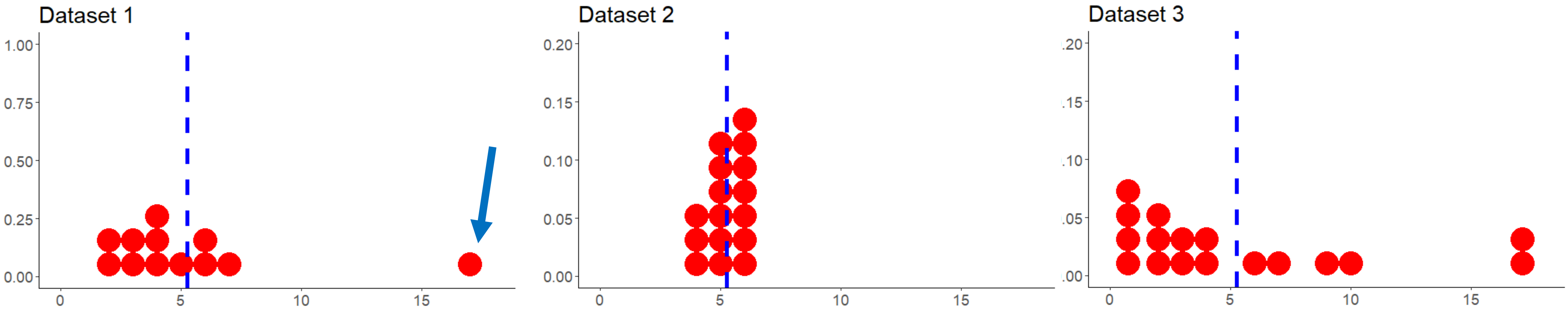
Mean (μ or $\bar{x}$)

- Sum of values divided by number of values

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

Same mean, different picture

- Dataset 1: 2, 2, 3, 4, 5, 4, 6, 7, 3, 17, 4, 6 → mean 5.25
- Dataset 2: 4, 4, 5, 5, 6, 6, 5, 5, 5, 6, 4, 5, 6, 6, 6, 6 → mean 5.25
- Dataset 3: 0.5,0.5,1,1,2,2,3,4,10,2 ,7,3,17,4,6,9, 17.25 → mean 5.25

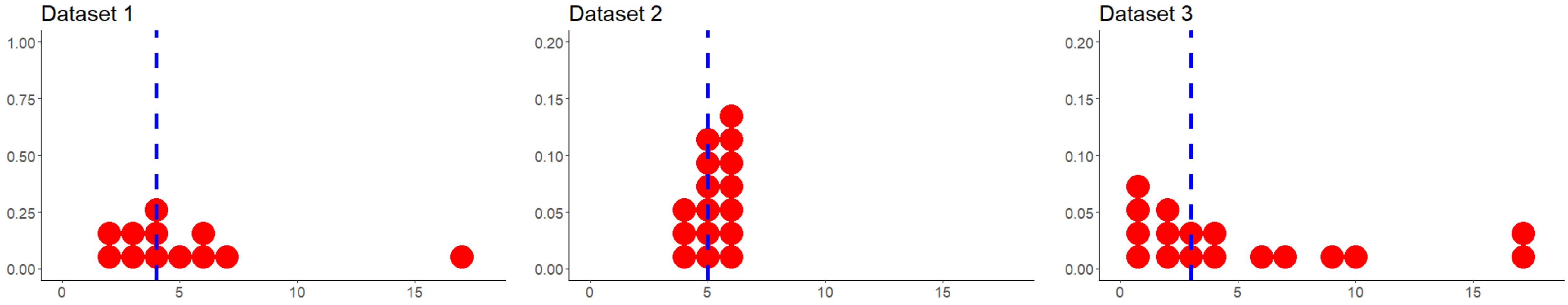


Median (η)

- Median = “(N+1)/2”th value
  - N=7, median is value at position(7+1)/2, i.e. 4th value
  - N=10, median is 5.5th value,
    - so average 5th and 6th values
  - Data must be in order
  - For a frequency distribution, we need to count cumulative frequency up to middle number
- Can also be defined for ordinal data, but not nominal data

… Median example

- Dataset 1: 2, 2, 3, 4, 5, 4, 6, 7, 3, 17, 4, 6 → median 4
- Dataset 2: 4, 4, 5, 5, 6, 6, 5, 5, 5, 6, 4, 5, 6, 6, 6, 6 → median 5
- Dataset 3: 0.5,0.5,1,1,2,2,3,4,10,2 ,7,3,17,4,6,9, 17.25 → median 3



Mode

- Most frequent value
- Disadvantages
  - May be several values of equal frequency
  - May not be near middle of distribution
  - In many applications, mode is not useful
- However, does apply to all data types, including nominal
- E.g. Browser users – frequency counts

chrome	edge	firefox	other	safari	samsung Int.
173	9	16	19	54	9

Mode is chrome, as it is the most frequent value.

Raw data example

- 3, 7, 5, 2, 4, 5, 3, 8, 3, 7
  - N = 10
- Mean

$$\bar{x} = \frac{3+7+2+5+4+5+3+8+3+7}{10} = \frac{47}{10} = 4.7$$

- Median
  - Order data: 2, 3, 3, 3, **4, 5**, 5, 7, 7, 8
  - Take value at position (10+1)/2 = 5.5
  - So average 5^{th} and 6^{th} values
    - 5.5^{th} value = (4+5)/2 = 4.5
- Mode is 3

Frequency Distribution – one-way table

Values	x_i	5	10	15	20
Frequency counts	f_i	13	8	6	9

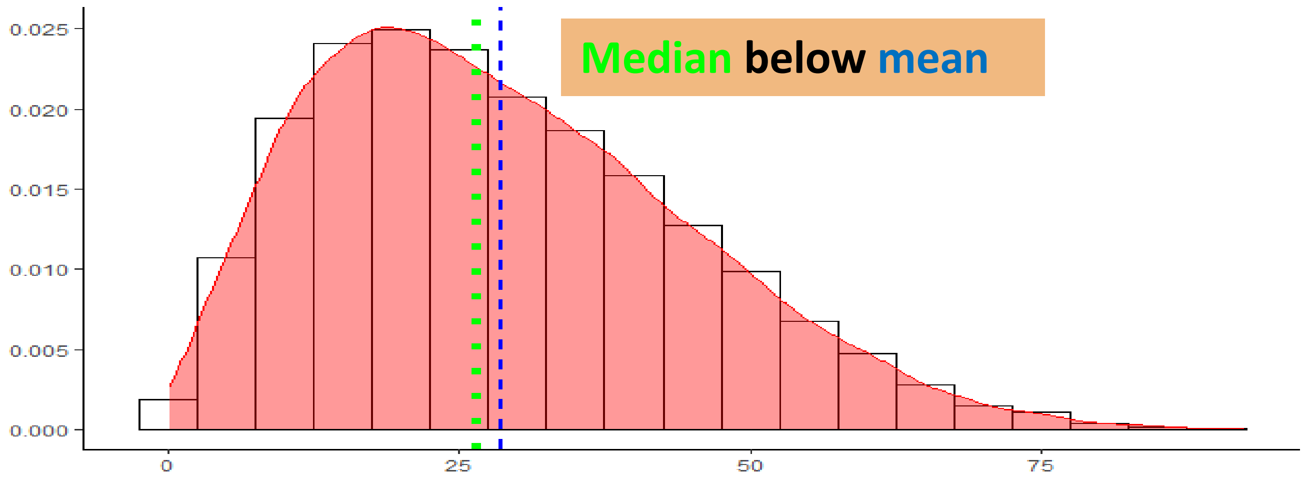
Value 20 appears 9 times

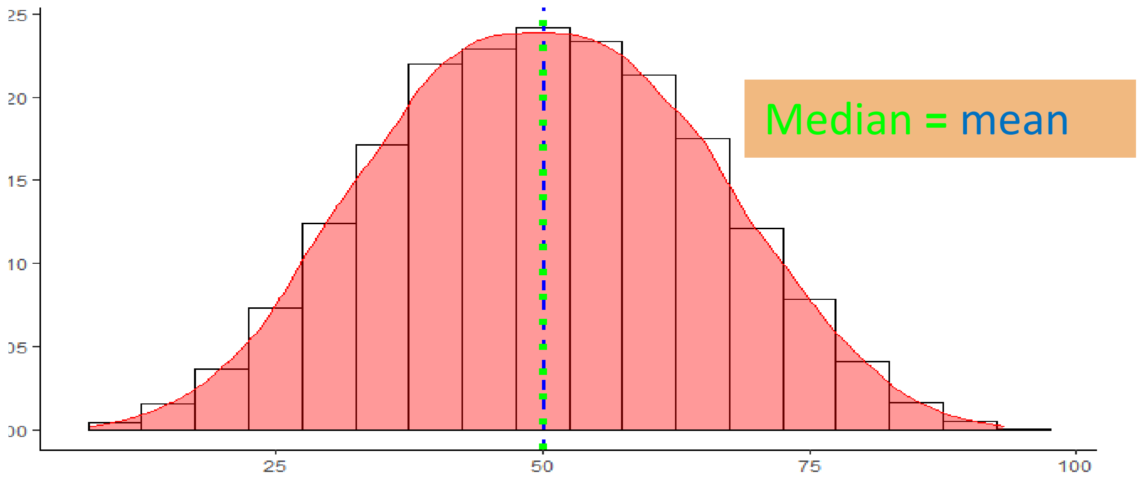
13+8+6+9 = 36 data values

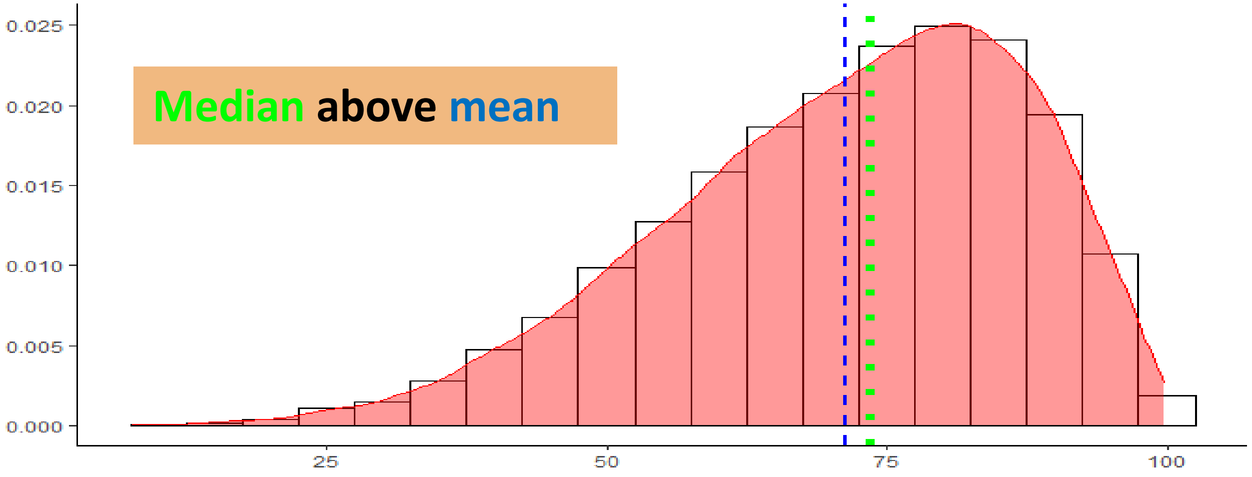
$$\bar{x} = \frac{13\times 5 + 8\times 10 + 6\times 15 + 9\times 20}{13+8+6+9} = \frac{415}{36} = 11.53$$

- Median = (37/2)th value = 18.5th value
  - 18th and 19th are both 10, so median is 10
- Mode is 5

Distributions vs mean and median values







Measures of Dispersion

- Range
- Inter-quartile range (IQR)
- Standard deviation and variance
- 5 point summary - boxplots

Quartiles

- The quartile provide information about the location of the data
  - Q1 is the median of bottom half of data (25% data is equal or below Q1)
  - Q2 is the median for the data (50% of data is equal or below Q2)
  - Q3 is the median of top half (75% data is equal or below Q3)
  - [There are some differences in the way these are calculated depending on method, leading to different results]
- Q1, Q2, Q3 Divide data in four
  - Confusingly, the quarters are sometimes called quartiles
- Used with minimum and maximum value to provide 5-point description of data
  - Min, Q1, Q2, Q3, Max
  - See below boxplots

Quantiles

- **Quartiles** can be generalised to any number of divisions
- Dividing values are **Quantiles**
- **Percentiles** divide the data into 100 equal parts
- **Deciles** divide the data into 10 equal parts

- Balance of detail vs effective summary

Measures of Dispersion (spread)

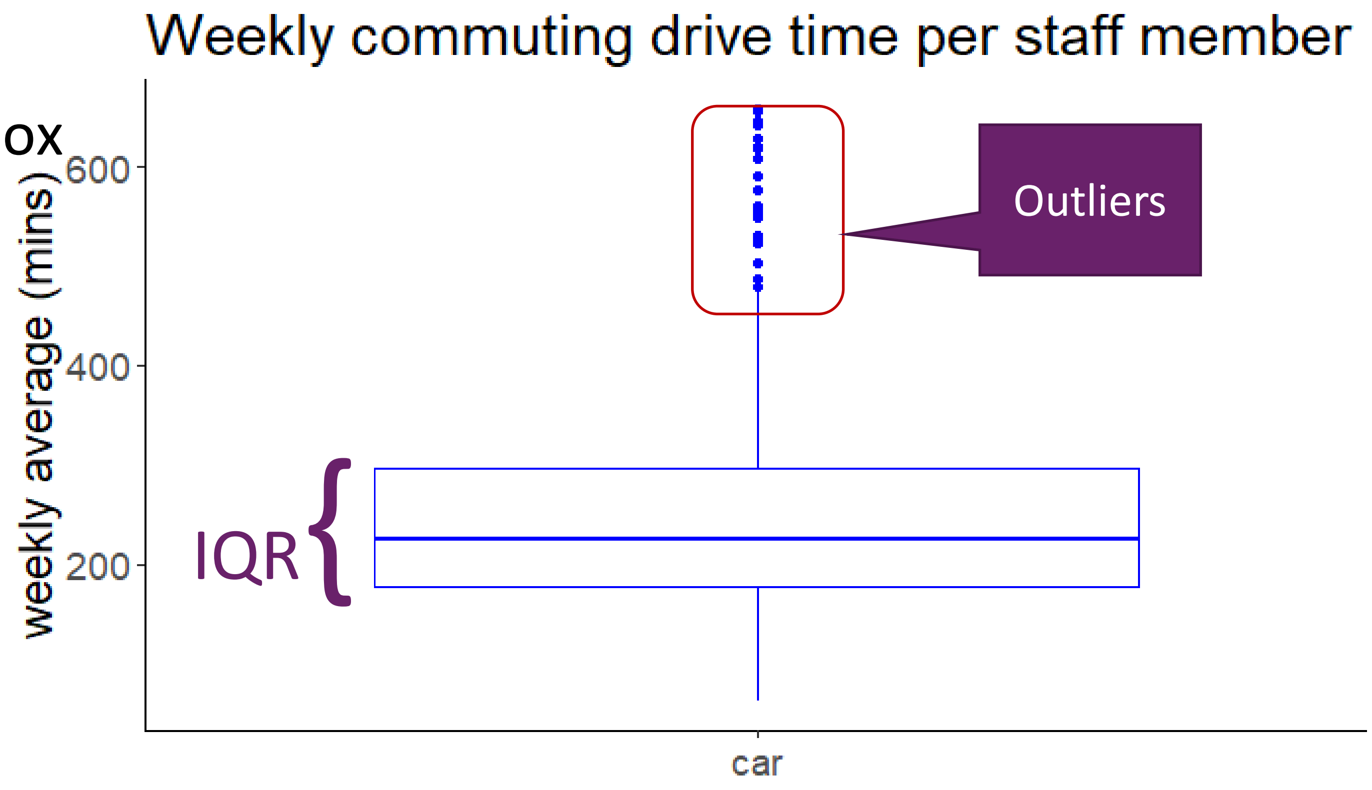
- **Range:** maximum value – minimum value
- **Inter-quartile range:** range of middle half, Q3 – Q1
  - Often paired with median
- **5- Number Summary {**Min, Q1, Q2, Q3, Max**}**
- **Standard deviation**
  - Pairs with mean
  - **σ** = population SD
  - **S** = sample SD

ROBERT GORDON UNIVERSITY ABERDEEN

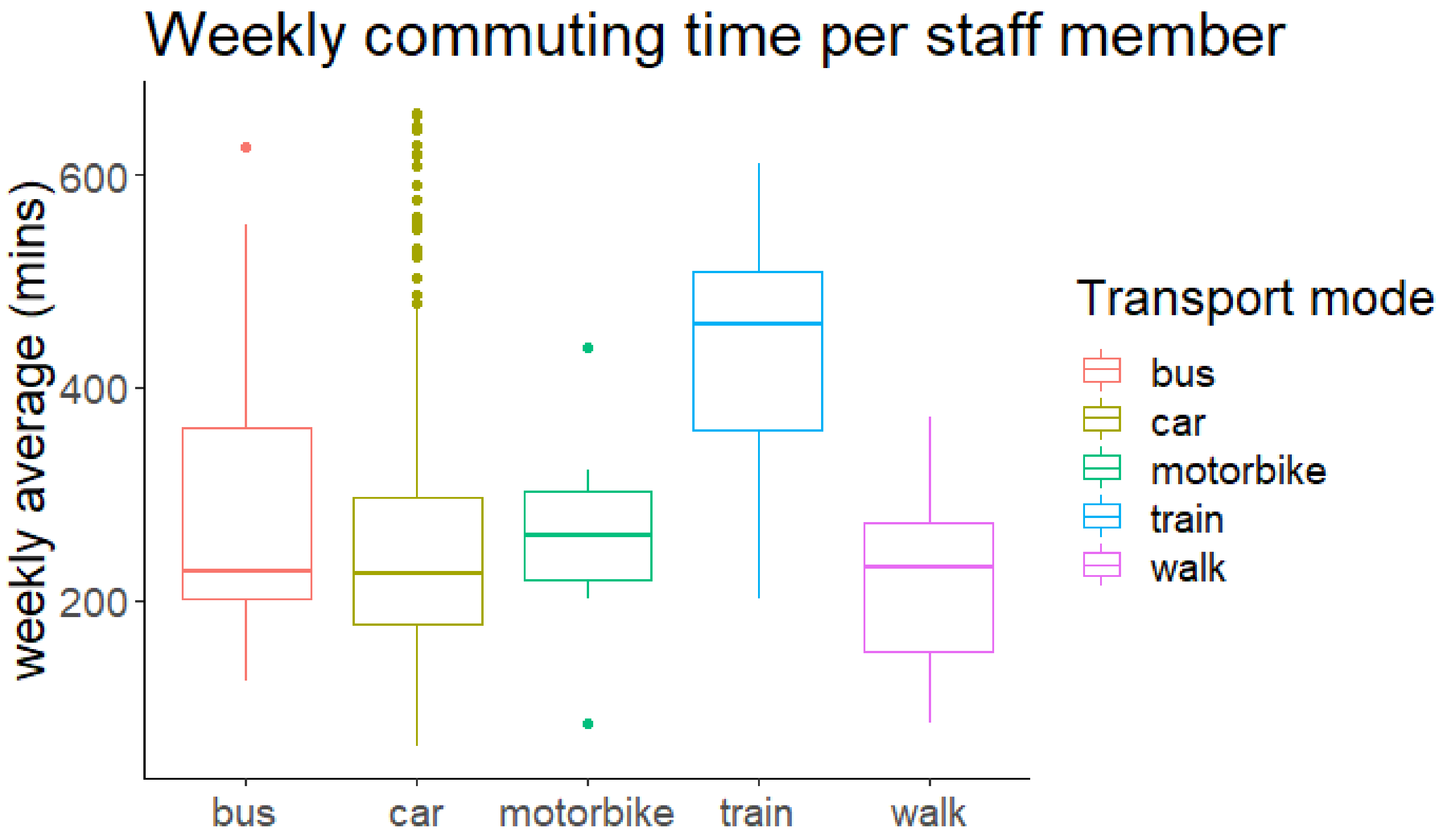
Boxplot

- **IQR** is difference between bottom and top of each box (blue area).
- **Q2 = Median** is horizontal bar in each box
- **Q1** is bottom of box
- **Q3** is top of box
- **Min** is bottom of bottom line (if no outliers)
- **Max** is top of top line (if no outliers)
- Outliers indicated by dots
- **5 point summary**

  {min, Q1,median, Q3, max}



Boxplot example



Outliers

- Unusually high/low value compared to the rest of the data in the dataset
  - Data incorrectly recorded (needs correction or removal)
  - Data should not have been included in dataset (does not belong - remove)
  - Data is correct and belongs in the dataset, but is unusual – skewness
    - Retain?
- Outliers
  - Below **lower inner fence**: Q1 – 1.5 * IQR
  - Above **upper inner fence**: Q3 + 1.5 * IQR

IQR and 5 number summary example

- Data
  - 3, 7, 5, 2, 4, 5, 3, 8, 3, 7
- Sorted data
  - 2, 3 ,3, 3, 4, 5, 5, 7, 7, 8
- Range = max – min = 8 - 2 = 6
- Median (Q2) = (4 + 5)/2 = 4.5
- Q1 = 3
- Q3 = 7
- IQR = 7 – 3 = 4
- 5-number summary {2,3,4.5,7,8}

Standard deviation

- Indicates spread of data
- Low value
  - Data is close to the mean
- Large value
  - Data is spread more
- Is a low SD value better than a high SD value?
  - Depends on the problem domain
  - Just because the data is more disperse it does not mean it is worse.
    - E.g. Time of the day at which each student starts using computer resources
    - Ideally start times are spread (up to a point) throughout the day so there is more resource availability per student -> higher SD

Standard Deviation: population and sample

Population standard deviation σ_N

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N}}$$

Sample standard deviation $s = \sigma_{N-1}$

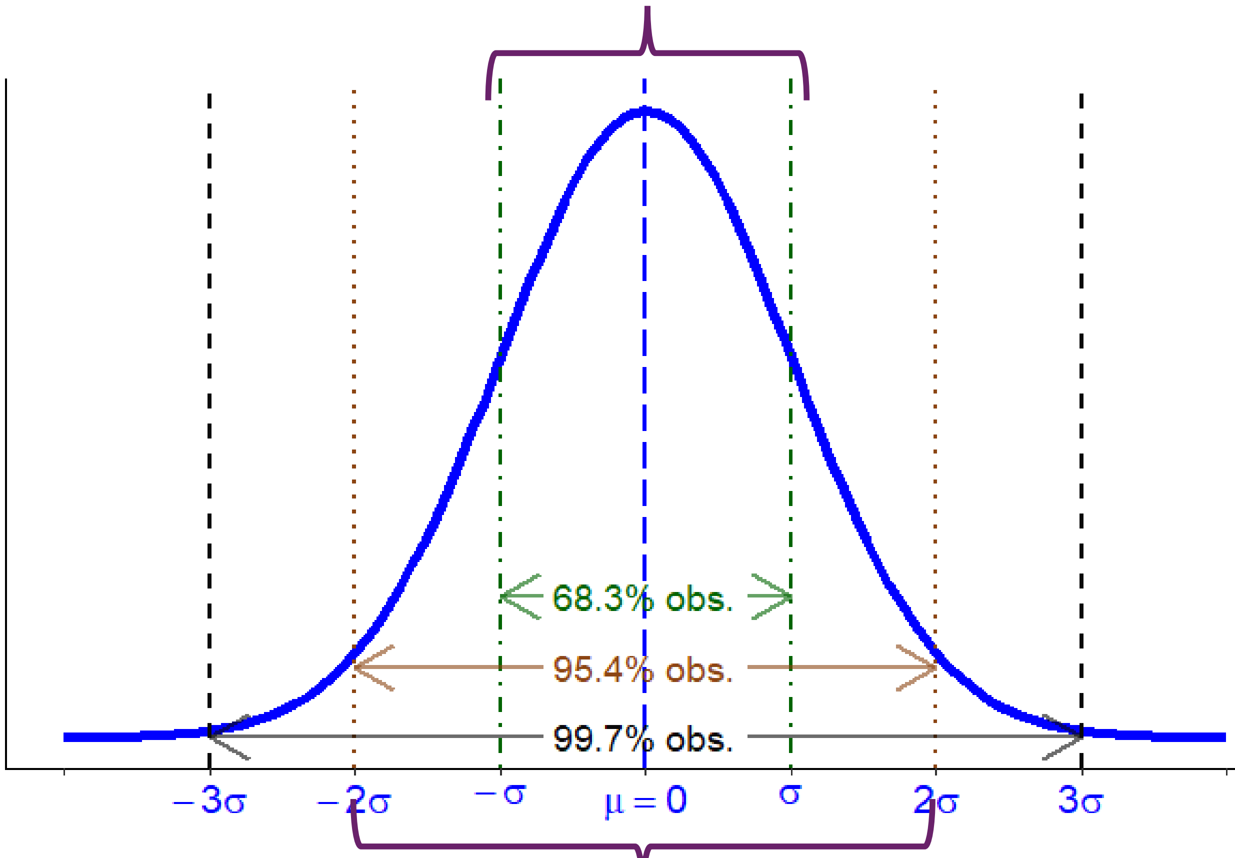
$$\sigma_{N-1} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N-1}}$$

- σ is for entire **population**
- s is for **sample** of larger population

Interpreting σ for a normal distribution

- For a normal distribution
  - **~68%** of observations will fall within 1 standard deviation of the mean
  - **~95%** of observations will fall within 2 standard deviations of the mean
  - almost **all** observations will fall within 3 standard deviations of the mean

One SD from mean 68% data



Two SDs from mean 95% data

Interpretation of sample standard deviation

- It is possible that no observations fall within one standard deviation of the mean
  - In $[\bar{x} - \sigma, \ \bar{x} + \sigma]$

- At least **3/4** of observations (75%) fall within 2 standard deviations of the mean
  - In $[\bar{x} - 2\sigma, \ \bar{x} + 2\sigma]$

- At least **8/9** of observations (89%) fall within 3 standard deviations of the mean.
  - In $[\bar{x} - 3\sigma, \ \bar{x} + 3\sigma]$

Standard deviation example

- Data
  - 3, 7, 5, 2, 4, 5, 3, 8, 3, 7 with $\bar{x} = \mathbf{4.7}$

$$\sigma_N = \sqrt{\frac{(3-4.7)^2 + (7-4.7)^2 + \ldots + (7-4.7)^2}{10}}$$

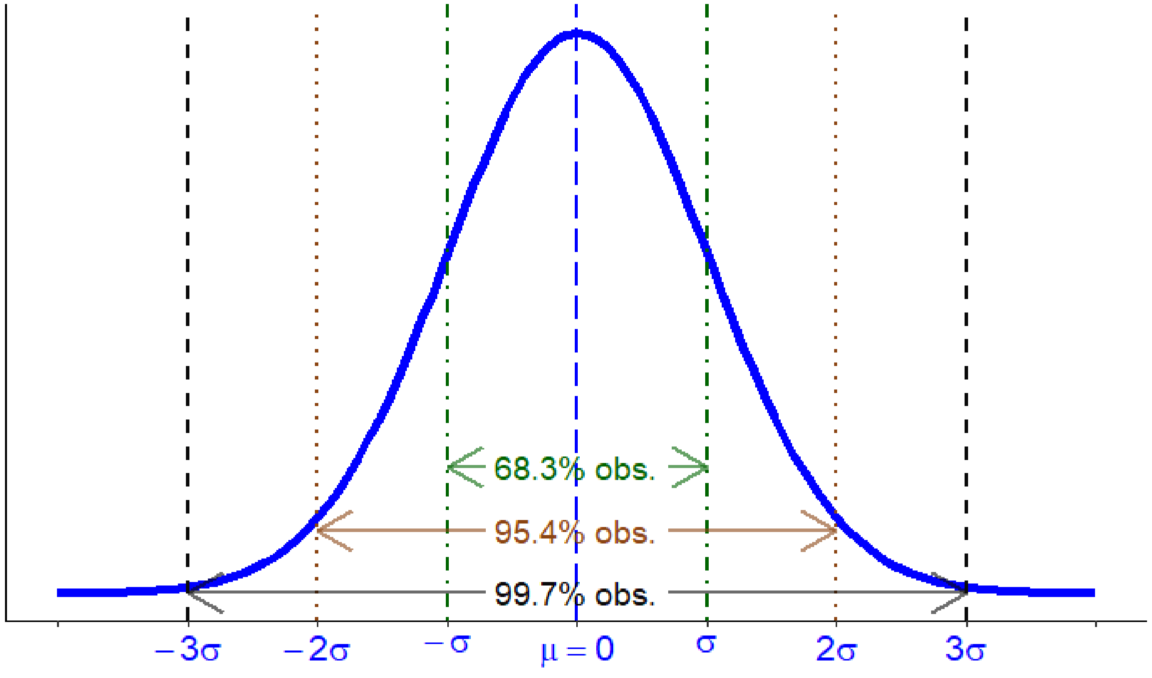
$$= \sqrt{\frac{38.1}{10}} = \sqrt{3.81} = 1.95$$

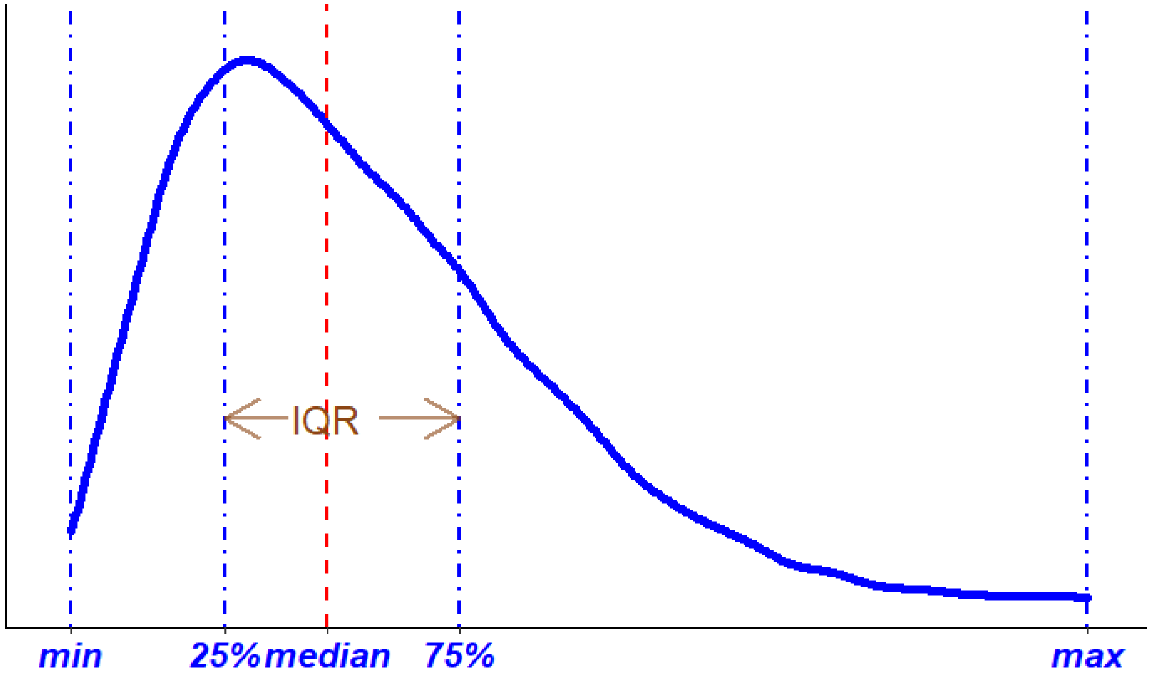
Variance vs Standard Deviation

- Variance = σ^2
- Sample Variance = s^2

What measures?

- For **normal distributions**
  - Summarise data using **mean** and **standard deviation**.
- For **skewed distributions**
  - Summarise data using **median** and **interquartile range** as they are more representative of data centre and spread.
  - But also use mean and standard deviation in addition.
  - Or five number.





Summary

- Data analysis depends on types of data
- Data cannot be analysed without 2 previous processes
  - Collection
  - Preparation
- Basics of probability and stats can tell important information about data
- Important measures are
  - Mean
  - Median
  - Mode
  - Range
  - Standard deviation
  - Interquartile range
  - 5-point summary {min, Q1, Median, Q3, max}