

PHAR1811

Data Analysis

GARTH TARR
SEMESTER 1, 2013



THE UNIVERSITY OF
SYDNEY

Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location


Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Housekeeping


Contact Details

- Email:  `garth.tarr@sydney.edu.au`
- Room: 806 Carslaw Building
- Consultation: by appointment (email to arrange a time)

Weekly workshops

- Week 4: Tutorial
- Week 5: Computer lab
- Week 6: Quiz

Materials

-  `sydney.edu.au/science/maths/u/gartht/PHAR1811`

Calculator



You need to bring a (non-programmable) calculator with you to all lectures, tutorials, labs and quizzes!

Overview

1. Why do you need to know statistics?
2. Summarising data sets numerically
 - Measures of location
 - Measures of variation
 - Min, max, quartiles
3. Summarising data sets graphically
 - Histograms
 - Frequency tables
 - Stem and leaf plots
 - Boxplots
4. Summarising bivariate data sets
 - Scatter plots
 - Correlation

Where is statistics used?

Sample size calculations

Each time you perform an experiment you use, on average, 15 grams of a particular chemical. This chemical is **very** expensive and takes three months to be imported from America. You need to run 10 successful experiments for your Honours thesis.

Question: How much of the drug should you order?

Hypothesis testing

The standard treatment for Rheumatoid arthritis had a measurable improvement for 63% of people. A new drug trialled has been trialled on 100 people. 68 people recorded a measurable improvement.

Question: Is the new drug better than the standard treatment?

Where is statistics used?

Quality control

Pfizer manufactures 10,000 Sildenafil citrate¹ tablets per hour. A random inspection of 500 tablets shows that 6 are of poor quality. The acceptable poor quality rating is 1% or lower.

Question: Is the production process working to the required standard?

Describing data

You have blood pressure measurements for 10,000 individuals in an Excel worksheet.

Question: How do you present this data in a meaningful way?

¹Commonly known as Viagra

Where is statistics used?

Business

You own a pharmacy. You think you sell more boxes of tissues in winter than in summer. You have monthly sales data for tissues for the past three years.

Question: How many tissues should you order each month?

Medicine

In the past, doctors did not wash their hands (or their surgical instruments) between patients. Florence Nightingale observed that less patients died in wards where nurses washed their hands.

Question: Does washing your hands save lives?

Where is statistics used?


Predictive Modelling

Certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate current flu activity around the world in near real-time.

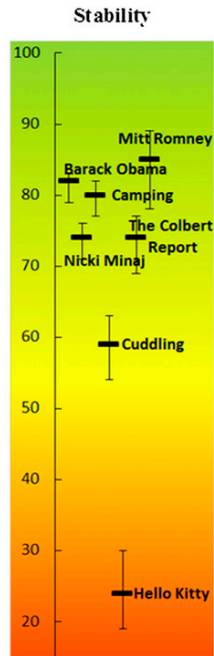


Where is statistics used?

Predictive Modelling

By looking at what pages you  Like on Facebook researchers can automatically and accurately predict a range of personal attributes.

Source: Proceedings of the National Academy of Sciences



What's involved in a statistical study?

Statistics exist because of **randomness**, i.e. variability due to chance. To deal with this, statistical studies generally involve:

1. **Experimental Design:** How to collect data to best target the population of interest and minimise bias.
 2. **Descriptive Statistics:** Present the data in meaningful ways to look for patterns after the data are collected. Calculate summary measures of location and variability to describe patterns.
 3. **Modelling:** Fit an appropriate probability model to account for the patterns and variability in the data.
 4. **Inference:** What does the observed sample tell us about the “target population”? How reliable is that information?
- Over the next two weeks we will consider step 2.

Descriptive Statistics

Numerical Summary

- Measures of Location
 - Arithmetic mean
 - Median
 - Mode
- Measures of Scale/Spread
 - Range
 - Standard deviation
 - Median absolute deviation
 - Interquartile range
- Other
 - Min and max
 - Quartiles
 - Correlation coefficient

Graphical Summary

- Categorical Data
 - Pie chart
 - Bar plot
 - Frequency table
 - Dot plot
- Univariate Data
 - Stem and leaf plot
 - Boxplot
 - Histogram
- Bivariate Data
 - Scatter plot

Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location

Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Population vs Sample

Definition (Population)

The entire set of individuals (or data points) with the characteristic that is of interest for a study.

Example (Population)

The set of all enrolled students across the University of Sydney.

Definition (Sample)

A subset of the population of interest that is collected during the course of a study.

Example (Sample)

A set of 1000 University of Sydney students interviewed while walking across City Road on Monday morning.

Why bother with samples?

Wouldn't it be better to work with populations?

- Populations are typically very large.
- It can be prohibitively expensive to survey a population.
- It can take a long time to survey a population.
- A sample represents a subset of manageable size.

Question

Can you think of any examples where the population is surveyed?

Samples need to be representative of the population.

- The best way to ensure a representative sample is to use “random sampling.”

Question

If we wanted to know how many University of Sydney students like attending lectures, should we just sample this class?

Variables

Definition (Variable)

A **variable** is a value or characteristic that can differ between individuals/observations.

Example

A study might include **variables** such as gender, blood pressure, placebo, quantity of drug prescribed, age, time since diagnosis, ...

Definition (Quantitative)

Quantitative variables are numerical values whose magnitude provides some meaning.

Definition (Qualitative)

Qualitative variables are used to distinguish between categories.

Variables

Qualitative/Categorical

- Marital Status
- Ethnicity
- Gender

Quantitative/Numerical

Discrete

- Number of children
- Score in a multiple choice quiz
- Number of attempts at a driving test

Continuous

- Height of an individual
- Weight of an individual
- Time since infection

Qualitative or Quantitative?

1. Blood pressure reading from a patient.
2. Recording a patient's doctor.

Arterial Blood Pressure – discrete or continuous?

- When you get your blood pressure measured often you'll end up with a reading along the lines of 120 over 80 mmHg.

Parameters of Interest

Definition (Parameter)

A **parameter** is a numerical measure (or fact) that describes a characteristic of a population.

Parameters are usually:

- unknown
- denoted in the literature by a Greek letter
 - μ (read: mu) for the population mean
 - σ (read: sigma) for the population standard deviation

Example

The mean prescription amount of Methylphenidate², μ , across all doctors in Sydney.

²Commonly known as Ritalin

Statistic

Definition (Statistic)

A **statistic** is a numerical measure that describes a characteristic of a sample.

Statistics are:

- Functions of the data – i.e. they are calculated using the observations you collect.
 - \bar{x} (read: “x bar”) for the sample mean
 - s for the sample standard deviation
- Used to **estimate** population parameters.

Example

The sample mean prescription amount of Methylphenidate, \bar{x} , prescribed by doctors in the Wentworth Building Health Centre.

General notation for writing observations

- For a general sample of size n we write the **observations** as $\{x_1, x_2, \dots, x_n\}$.
- In other words, the i th observation is denoted x_i for $i = 1, 2, \dots, n$.

Example (Observe the heights of 5 individuals)

| Name | i | x_i |
|------|-----|----------------------|
| Jack | 1 | $x_1 = 175\text{cm}$ |
| Jill | 2 | $x_2 = 163\text{cm}$ |
| Xiao | 3 | $x_3 = 182\text{cm}$ |
| Jim | 4 | $x_4 = 171\text{cm}$ |
| Jane | 5 | $x_5 = 159\text{cm}$ |

Sigma Notation

Definition (Sigma Notation)

We write the sum of n observations as:

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_{n-1} + x_n.$$

- The symbol, \sum , is the greek letter, capital sigma, hence the name “Sigma notation.”
- $\sum_{i=1}^n$ is read as “the sum from $i = 1$ to n .”
- You can use it to sum anything, not just observations:

$$\sum_{i=1}^3 1 = 1 + 1 + 1 = 3 \quad \text{or} \quad \sum_{i=1}^4 a = a + a + a + a = 4a.$$

Sigma Notation

Your turn...

$$\sum_{i=1}^3 (x_i + a) =$$

=

=

Sigma Notation

Example (Observe the heights of 5 individuals)

| Name | i | x_i |
|------|-----|----------------------|
| Jack | 1 | $x_1 = 175\text{cm}$ |
| Jill | 2 | $x_2 = 163\text{cm}$ |
| Xiao | 3 | $x_3 = 182\text{cm}$ |
| Jim | 4 | $x_4 = 171\text{cm}$ |
| Jane | 5 | $x_5 = 159\text{cm}$ |

The **sum** of these observations is:

$$\begin{aligned}\sum_{i=1}^5 x_i &= x_1 + x_2 + x_3 + x_4 + x_5 \\ &= 175 + 163 + 182 + 171 + 159 \\ &= 850.\end{aligned}$$

Sigma Notation

Example (Observe the heights of 5 individuals)

| Name | i | x_i |
|------|-----|----------------------|
| Jack | 1 | $x_1 = 175\text{cm}$ |
| Jill | 2 | $x_2 = 163\text{cm}$ |
| Xiao | 3 | $x_3 = 182\text{cm}$ |
| Jim | 4 | $x_4 = 171\text{cm}$ |
| Jane | 5 | $x_5 = 159\text{cm}$ |

The **sum of squares** of these observations is:

$$\begin{aligned}\sum_{i=1}^5 x_i^2 &= x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \\ &= 175^2 + 163^2 + 182^2 + 171^2 + 159^2 \\ &= 144,840.\end{aligned}$$

Your turn with Sigma Notation...

Consider the following data (and let a be any constant):

| i | 1 | 2 | 3 | 4 |
|-------|----|---|----|----|
| x_i | -1 | 1 | 5 | 10 |
| y_i | 5 | 4 | 10 | 12 |

1. $\frac{1}{4} \sum_{i=1}^4 x_i =$

2. $\sum_{i=1}^4 ay_i =$

3. $\sum_{i=1}^4 x_i^2 =$

4. $\sum_{i=1}^4 x_i y_i =$

Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location

Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Measure of Location: Mean

Definition (Sample Mean: \bar{x})

The sample mean, denoted \bar{x} , of n observations, $\{x_1, x_2, \dots, x_n\}$ is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

i.e. add together all the observations and divide by the sample size.

Example

Consider the following ages, y_i , of 12 individuals when first diagnosed with asthma, $\{8, 5, 4, 10, 12, 5, 25, 7, 6, 10, 11, 5\}$. The sample mean of these observations is:

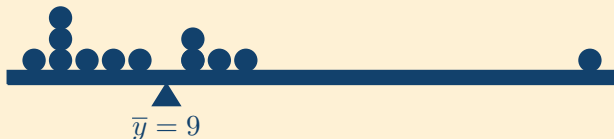
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{12} \times (8 + 5 + \dots + 11 + 5) = \frac{1}{12} \times 108 = 9.$$

The average age of diagnosis in this sample is 9.

Sample mean as the “centre of mass”

Example

Again, consider the ages of 12 individuals when first diagnosed with asthma, $\{8, 5, 4, 10, 12, 5, 25, 7, 6, 10, 11, 5\}$.



Properties of the Mean

Location equivariant

Consider $y_i = x_i + c$, for $i = 1, 2, \dots, n$ then,

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n (x_i + c) \\ &= \frac{1}{n} \left[\sum_{i=1}^n x_i + \sum_{i=1}^n c \right] \\ &= \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} nc \\ &= \bar{x} + c.\end{aligned}$$

If we “shift” a variable by adding a constant to each observation and then find the sample mean then it is the same as adding a constant to the the mean of the original variable.

Properties of the Mean

Scale equivariant

Consider $y_i = c x_i$, for $i = 1, 2, \dots, n$ then,

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n c x_i \\ &= \frac{1}{n} [c x_1 + c x_2 + \dots + c x_n] \\ &= \frac{1}{n} c \sum_{i=1}^n x_i \\ &= c \bar{x}.\end{aligned}$$

If we “rescale” a variable by multiplying each observation by a constant and find the sample mean then it is the same as multiplying the sample mean of the original variable by the same constant.

Ordered Data

Definition

An ordered set of numerical data $\{x_1, x_2, \dots, x_n\}$ is denoted $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ where

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Example (Ordered asthma data)

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-----------|---|---|---|----|----|---|----|----|----|----|----|----|
| x_i | 8 | 5 | 4 | 10 | 12 | 5 | 25 | 7 | 6 | 10 | 11 | 5 |
| $x_{(i)}$ | 4 | 5 | 5 | 5 | 6 | 7 | 8 | 10 | 10 | 11 | 12 | 25 |

So $x_{(1)} = 4$, $x_{(2)} = 5$, $x_{(3)} = 5$, $x_{(4)} = 5$, \dots , $x_{(12)} = 25$.

Measure of Location: Median

Definition (Median)

The **median**, denoted \tilde{x} , is in the middle of the data set in the sense that **at least** 50% of the data values are below \tilde{x} and **at least** 50% of the data are above \tilde{x} .

In a sample of n observations, the **median** is defined as:

- The $k = (n + 1)/2^{\text{th}}$ largest observation if n is **odd**:

$$\tilde{x} = x_{(k)}.$$

- The average of the $n/2^{\text{th}}$ and $(n/2 + 1)^{\text{th}}$ largest observations if n is **even**:

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Finding the median

Example (Median of the asthma data)

1. Order the data:

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-----------|---|---|---|----|----|---|----|----|----|----|----|----|
| x_i | 8 | 5 | 4 | 10 | 12 | 5 | 25 | 7 | 6 | 10 | 11 | 5 |
| $x_{(i)}$ | 4 | 5 | 5 | 5 | 6 | 7 | 8 | 10 | 10 | 11 | 12 | 25 |



2. Find the middle:

$$\tilde{x} = 7.5$$

- We have an even number of observations so we find the average of the $n/2 = 6^{\text{th}}$ and $(n/2 + 1) = 7^{\text{th}}$ observations:

$$\tilde{x} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{7 + 8}{2} = 7.5.$$

3. Check your answer:

- Is at least 50% of the data ≥ 7.5 ?
- Is at least 50% of the data ≤ 7.5 ?

Measure of Location: Mode

Definition (Mode)

The mode, is the most frequently occurring value.

- If all the data values are unique, then the mode is not defined.
- If two values are both share the highest frequency. The sample is said to be 'bimodal'.
- Similarly you can have 'trimodal' and 'multimodal' samples.

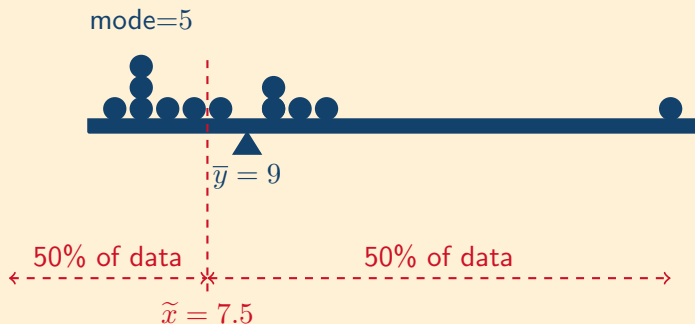
Example (Mode of the asthma data)

- The data are: {8, 5, 4, 10, 12, 5, 25, 7, 6, 10, 11, 5}.
- The mode is: 5.

Measures of location

Example

Again, consider the ages of 12 individuals when first diagnosed with asthma, $\{8, 5, 4, 10, 12, 5, 25, 7, 6, 10, 11, 5\}$.



Robustness

What happens to our location estimates if one (or more) of our observations is really large?

Example (Asthma data with $x_7 = 52$ instead of $x_7 = 25$)

The data are now: $\{8, 5, 4, 10, 12, 5, 52, 7, 6, 10, 11, 5\}$.

- The **median** stays the same (check for yourself):

$$\tilde{x} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{7 + 8}{2} = 7.5.$$

- The **mode** stays the same: 5.
- The **sample mean** is now:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{12} \times 135 = 11.25,$$

when before it was 9.

Which measure of location is best?

- The mean is more sensitive to extreme observations than the median. For this reason, the median is said to be more **robust** than the mean.
- The mean is easier to compute and easier to deal with theoretically than the median.
- For **skewed** data, the median is usually preferable.
- For **symmetric** data, the mean is usually preferable since it is less variable between samples.

Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location

Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Measures of Scale/Spread/Dispersion

Another quantity of interest to describe a dataset or population is the variability.

Example

Systolic Blood Pressure³ (SBP) measurements for 6 males and 6 females.

- Male Sample $\{x_i\} = \{132, 139, 156, 133, 141, 133\}$.
- Female Sample $\{y_i\} = \{132, 109, 121, 120, 126, 85\}$.

Question: Is one dataset more “spread apart” than the other?

³maximum blood pressure over the course of a heartbeat

Range

Definition (Range)

The range of a sample of size n is the maximum value minus the minimum value:

$$\text{Range} = x_{(n)} - x_{(1)}.$$

Example (Systolic Blood Pressure (SBP) Data)

1. For each data set, find the maximum and the minimum.
 - Males $\{x_i\} = \{132, 139, 156, 133, 141, 133\}$
 - Females $\{y_i\} = \{132, 109, 121, 120, 126, 85\}$
2. Subtract the minimum from the maximum.
 - Male range = $156 - 132 = 24$
 - Female range =
3. In this sample, female SBP appears to be more variable than male SBP.

Population variance and standard deviation

Definition (Population variance)

The **population variance**, σ^2 (read: sigma squared), is the average of the squared deviations of each observation from the mean:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \text{where} \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- N is the population size
- μ (read: mu) is the population mean.

Definition (Population standard deviation)

The **population standard deviation**, σ (read: sigma), is the square root of the **population variance**:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Sample variance and standard deviation

In practice we rarely observe all N individuals in the population. Instead we take a sample of size n .

Definition (Sample variance and standard deviation)

- The **sample variance**, denoted s^2 , is defined as:

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

- The **sample standard deviation**, denoted s , is the square root of the **sample variance**:

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

Population vs sample variance

Population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Sample variance

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

- The denominator for the sample variance is $n - 1$.
- As the sample size, n increases,

$$\frac{1}{n-1} \approx \frac{1}{n}$$

(try it yourself, use $n = 5$ and then $n = 500$)

- In large samples, the difference between $1/n$ and $1/(n-1)$ is negligible.
- In the sample case, we are still pretty much finding the average of the squared deviations from the mean.

Calculation formula for sample variance

- To make calculation easier, we can rewrite the formula for the sample variance as:

$$s^2 = \frac{1}{n-1} S_{xx} = \frac{1}{n-1} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \quad (1)$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]. \quad (2)$$

- In equation (1) we first need to find n and \bar{x} . We then need to calculate the squared deviation from the mean for each observation, $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$ and then add them all together.
- In equation (2) we still need to find n and \bar{x} but then we only need to calculate $\sum_{i=1}^n x_i^2$.

Sample variance and standard deviation

Example (Systolic Blood Pressure (SBP) Data)

- Male Sample $\{x_i\} = \{132, 139, 156, 133, 141, 133\}$.

- Sample size: $n = 6$.

- Sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} \times 834 = 139$.

- Sum of squares:

$$\sum_{i=1}^n x_i^2 = 132^2 + 139^2 + 156^2 + 133^2 + 141^2 + 133^2 = 116,340.$$

- Sample variance,

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \\ &= \frac{1}{6-1} [116,340 - 6 \times 139^2] = 82.8. \end{aligned}$$

- Sample standard deviation: $s = \sqrt{s^2} = \sqrt{82.8} \approx 9.1$.

Sample variance and standard deviation

Your turn with Systolic Blood Pressure (SBP) Data

- Female Sample $\{y_i\} = \{132, 109, 121, 120, 126, 85\}$.
- Sample size: $n =$
- Sample mean: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i =$
- Sum of squares: $\sum_{i=1}^n y_i^2 =$
- Sample variance,

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right]$$

=

- Sample standard deviation: $s = \sqrt{s^2} =$
- It appears that SBP is more variable than SBP.

Properties of the sample standard deviation

Location invariant

Consider $y_i = x_i + c$, then,

$$s_y = s_x,$$

i.e. if you add a constant to all observations, the standard deviation does not change.

Scale equivariant

Consider $y_i = cx_i$, then,

$$s_y = cs_x,$$

i.e. if you multiply all observations by a constant, the standard deviation is also multiplied by the same constant amount.

Interquartile Range

Definition (Quartiles)

The quartiles are the values that divide the **ordered** dataset into four (approximately) equal parts.

- The 1st quartile, denoted Q_1 , is a value that is greater than or equal to at least 25% of the data and less than or equal to at least 75% of the data.
- Q_2 is the median, \tilde{x} .
- $Q_3 = 3^{\text{rd}}$ quartile is a value such that at least 75% of the data are $\leq Q_3$ and at least 25% of the data are $\geq Q_3$.

Definition (Interquartile range)

The Interquartile Range (IQR) is the distance of the third quartile from the first quartile: $\text{IQR} = Q_3 - Q_1$.

Finding Quartiles

Two cases:

1. If $n/4$ is a whole number then

- $Q_1 = \frac{1}{2}(x_{(k)} + x_{(k+1)})$ where $k = \frac{n}{4}$.
- $Q_3 = \frac{1}{2}(x_{(k)} + x_{(k+1)})$ where $k = \frac{3n}{4}$.

Example (Quartiles of the asthma data)

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-----------|---|---|---|-----------|----|---|----|----|----|--------------|----|----|
| x_i | 8 | 5 | 4 | 10 | 12 | 5 | 25 | 7 | 6 | 10 | 11 | 5 |
| $x_{(i)}$ | 4 | 5 | 5 | 5 | 6 | 7 | 8 | 10 | 10 | 11 | 12 | 25 |
| | | | | ↑ | | | | | | ↑ | | |
| | | | | $Q_1 = 5$ | | | | | | $Q_3 = 10.5$ | | |

The Interquartile Range is therefore:

$$\text{IQR} = Q_3 - Q_1 = 10.5 - 5 = 5.5$$

Finding Quartiles

2. If $n/4$ is not a whole number then

- To find Q_1 you round $n/4$ up to the next whole integer and use that order statistic.
- To find Q_3 you round $3n/4$ up to the next whole integer and use that order statistic.

Example (Consider a data set with $n = 9$ observations)

- To find Q_1 : $n/4 = 2.25$ and we round up to the next integer: $Q_1 = x_{(3)}$. The third biggest observation.
- To find Q_3 : $3n/4 = 6.75$ and we round up to the next integer: $Q_3 = x_{(7)}$. The seventh biggest observation.

Important!

You need to divide the ordered data set into 4 parts, if n is not a multiple of four, **round up** to the next whole number.

IQR of the SBP data

Example (Female Sample $\{y_i\}$: $\{132, 109, 121, 120, 126, 85\}$)

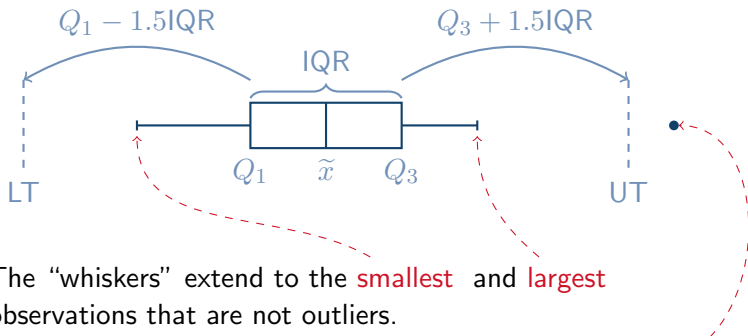
1. Order the data: $\{y_{(i)}\}$: $\{85, 109, 120, 121, 126, 132\}$
2. $n/4 = 1.5$, round up: $Q_1 = x_{(2)} = 109$.
3. $3n/4 = 4.5$, round up: $Q_3 = x_{(5)} = 126$.
4. $\text{IQR} = Q_3 - Q_1 = 17$.

Your turn: Male Sample $\{x_i\}$: $\{132, 139, 156, 133, 141, 133\}$

1. Order the data:
2. $n/4 = 1.5$, round up: $Q_1 =$
3. $3n/4 = 4.5$, round up: $Q_3 =$
4. $\text{IQR} = Q_3 - Q_1 =$

Quartiles and boxplots

Boxplots graphically highlight Q_1 , \tilde{x} , Q_3 and outliers.



- The “whiskers” extend to the **smallest** and **largest** observations that are not outliers.
- Outliers (above UT and below LT) are identified using **dots**.

Median Absolute Deviation (MAD)

Definition (Median Absolute Deviation)

The **median absolute deviation** (MAD) is defined to be the median of the absolute deviations from the median.

$$\text{MAD} = \text{Median of } \{ |x_i - \tilde{x}| \}.$$

Recall the variance is (approximately, when n is large) the average of the squared deviations from the mean:

$$s^2 \approx \text{Mean of } \{ (x_i - \bar{x})^2 \}.$$

Median Absolute Deviation (MAD)

Example (Male SBP data)

Male SBP data: $\{x_i\} = \{132, 139, 156, 133, 141, 133\}$.

1. Order the data: $\{x_{(i)}\} = \{132, 133, 133, 139, 141, 156\}$.

2. Median: $\tilde{x} = \frac{133 + 139}{2} = 136$.

3. Calculate the deviations from the median.

$$\begin{aligned} &\{132 - \tilde{x}, 139 - \tilde{x}, 156 - \tilde{x}, 133 - \tilde{x}, 141 - \tilde{x}, 133 - \tilde{x}\} \\ &= \{-4, 3, 20, -3, 5, -3\}. \end{aligned}$$

4. Take the absolute values, $\{4, 3, 20, 3, 5, 3\}$.

5. Calculate the median by **ordering** and finding the middle:

$$\text{Median } \{3, 3, 3, 4, 5, 20\} = 3.5.$$

6. $\text{MAD} = 3.5$.

Median Absolute Deviation (MAD)

Your turn: Female SBP data

Female Sample $\{y_i\} = \{132, 109, 121, 120, 126, 85\}$.

1. Order the data: $\{y_{(i)}\} =$
2. Median: $\tilde{x} =$
3. Calculate the deviations from the median.
4. Take the absolute values,
5. Calculate the median by ordering and finding the middle:
6. MAD =

Which measure of scale is appropriate?

- This question is very similar to the mean vs median argument for measure of location.
- The standard deviation is easier to deal with theoretically and more efficient for symmetric datasets.
- The IQR and MAD are more 'robust' measures of spread, i.e. less sensitive to extreme observations.

Example (SBP measurements)

Question: Is one dataset more "spread apart" than the other?

| | Males | Females |
|--------------------|-------|---------|
| Range | 25.0 | 47.0 |
| Standard deviation | 9.1 | 16.8 |
| IQR | 8.0 | 17.0 |
| MAD | 3.5 | 8.5 |

Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location

Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Summarising categorical data

Definition (Categorical data)

If the observations in a sample fall into categories the frequencies are called **categorical data**.

Example

Calls to 000 for an ambulance are categorised as follows:

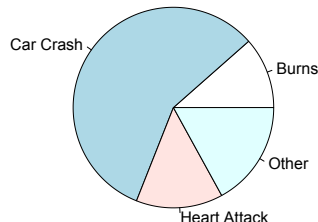
| Reason | Burns | Car Crash | Heart attack | Other |
|------------------|-------|-----------|--------------|-------|
| Frequency | 23 | 115 | 28 | 34 |

- We can use pie charts, bar plots or dot charts.

Pie chart

- Pie charts are a **very bad** way of displaying information.
- The eye is good at judging linear measures (in a bar chart) but bad at judging relative areas (in a pie chart).
- Cleveland (1985): "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements."⁴

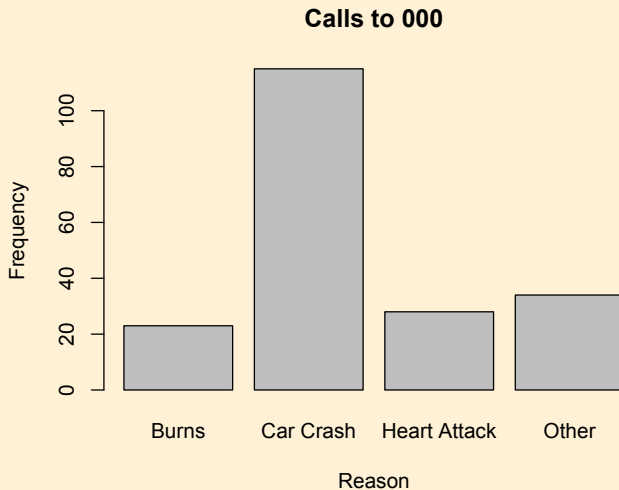
This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by **perceptual psychologists**.



⁴Cleveland, W. S. (1985) *The elements of graphing data*. Wadsworth: Monterey, CA, USA.

Bar plot

Example

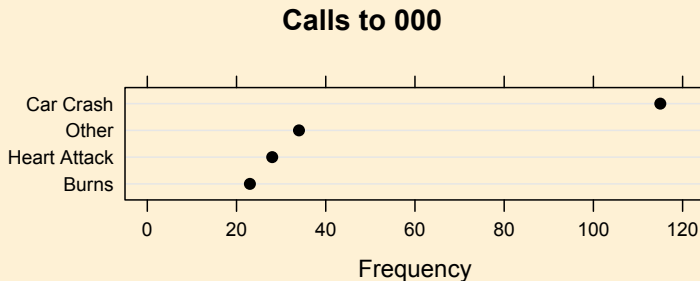


Dot chart

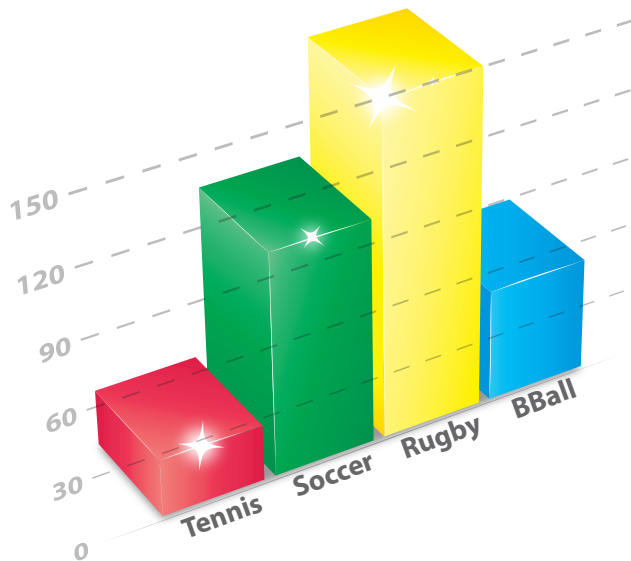
A dot chart is an alternative to a barplot or pie chart.

- Doesn't take up as much space as a barplot.
- No distractions, simple and clear.
- Much easier to compare categories than a pie chart.

Example



Chartjunk has no place in science!



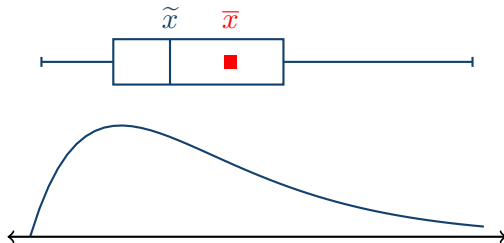
Univariate data and distributions

Definition (Distribution)

The term **distribution** is used to describe how the data is organised across the real line.

1. Left skewed, negative skew or long left tail.
2. Symmetric
3. Right skewed, positive skew or long right tail.
4. Bimodal, trimodal and multimodal.

Right skew

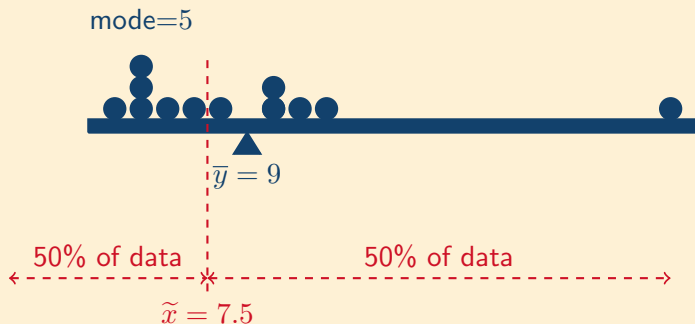


- For right (positive) skew distributions the tail is to the right (positive) end of the axis.
- In right skew distributions, $\tilde{x} < \bar{x}$, i.e. **Median** < **Mean**.
- Examples **might** include:
 - blood pressure readings,
 - defect counts in quality control,
 - ground surface temperature readings.

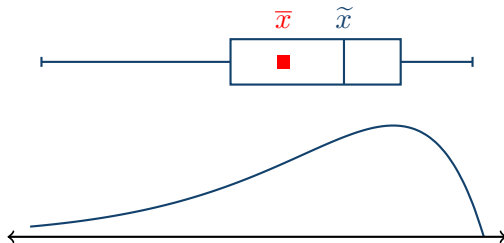
Skew and measures of location

Example

Recall, the ages of 12 individuals when first diagnosed with asthma, $\{8, 5, 4, 10, 12, 5, 25, 7, 6, 10, 11, 5\}$.

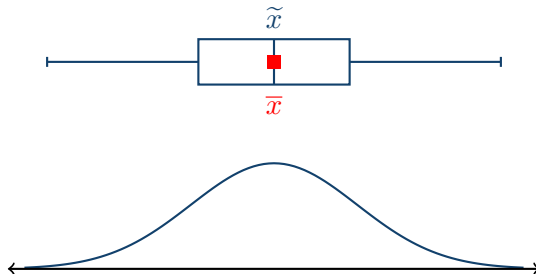


Left skew



- For left (negative) skew distributions the tail is to the left (negative) end of the axis.
- In left skew distributions, $\tilde{x} > \bar{x}$, i.e. **Median** > **Mean**. (Why?)
- Examples **might** include:
 - student marks in a quiz,
 - Mathematics Extension 2 HSC marks,
 - relative humidity readings.

Symmetric

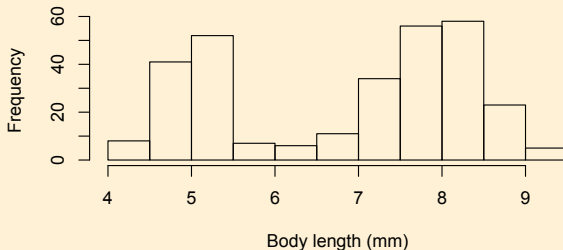


- For symmetric distributions there is no skew.
- In symmetric distributions, $\tilde{x} \approx \bar{x}$, i.e. **Median** \approx **Mean**.
- Examples **might** include:
 - height data,
 - IQ results,
 - birth weight of full term babies.

Multimodal distributions

- If there is one 'bump' in the distribution shape, then the data is said to be unimodal.
- If there are multiple 'bumps' then the data is said to be multimodal (bimodal for two bumps).

Example (Body lengths of 300 weaver ant workers)



The **bimodal** distribution arises due to existence of two distinct classes of workers: major workers and minor workers.

Frequency tables

Definition (Frequency table)

A **frequency table** is able to summarise a large dataset to help identify patterns while retaining all the information in the data set. For each unique value (or class) in the frequency table the frequency of occurrences is recorded.

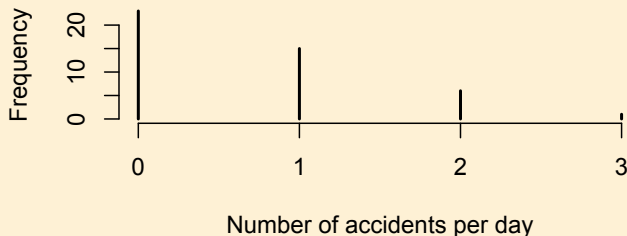
Frequency tables

Example (Discrete data)

Consider the dataset for the number of car accidents on a particular 'black spot' over the course of 45 days.

| | | | | |
|----------------------|----|----|---|---|
| Outcome (x_i): | 0 | 1 | 2 | 3 |
| Frequency (f_i): | 23 | 15 | 6 | 1 |

You could also plot this (discrete) data in an **ordinate diagram**:



Calculations using frequency tables

If data is already summarised in a frequency table, we can still calculate descriptive statistics such as the mean and variance.

- Assume a set of n observations take k unique values, $\{x_1, x_2, \dots, x_k\}$ with respective frequency counts $\{f_1, f_2, \dots, f_k\}$:

| | | | | | |
|------------------------|-------|-------|-------|---------|-------|
| Outcome (x_j): | x_1 | x_2 | x_3 | \dots | x_k |
| Frequencies (f_j): | f_1 | f_2 | f_3 | \dots | f_k |

- The mean is given by

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k f_j x_j$$

where $n = \sum_{j=1}^k f_j$ (the sum of the frequencies).

Calculations using frequency tables

Example (Car accidents at a particular black spot)

| | | | | |
|----------------------|----|----|---|---|
| Outcome (x_j): | 0 | 1 | 2 | 3 |
| Frequency (f_j): | 23 | 15 | 6 | 1 |

- We know that $n = \sum_{j=1}^k f_j = 23 + 15 + 6 + 1 = 45$ days.
- Thus,

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{j=1}^k f_j x_j \\ &= \frac{1}{45} (23 \times 0 + 15 \times 1 + 6 \times 2 + 1 \times 3) \\ &= \frac{2}{3}.\end{aligned}$$

Frequency tables for continuous data

- Continuous data can be grouped into **interval classes** to simplify the dataset without losing too much information.
- If **interval classes** are used, there is a possible issue when a value lies on a boundary of an interval. Two options:
 1. Include the left boundary point but not the right boundary point; **OR**
 2. Include the right boundary point but not the left boundary point.
- If the table uses intervals, with interval means m_j , the sample mean can be approximated by

$$\bar{x} \approx \frac{1}{n} \sum_{j=1}^k f_j m_j.$$

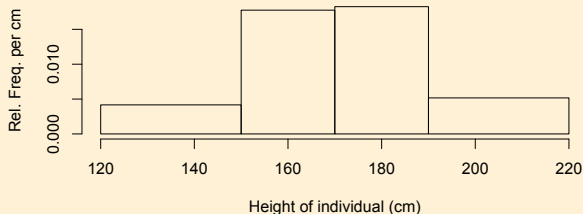
Frequency tables

Example (Continuous data)

Consider the heights (in cm) for $n = 200$ individuals,

| Interval Class | Class Mean (m_j) | Frequency (f_j) |
|----------------------|----------------------|---------------------|
| $120 < x_i \leq 150$ | 135 | 25 |
| $150 < x_i \leq 170$ | 160 | 71 |
| $170 < x_i \leq 190$ | 180 | 73 |
| $190 < x_i \leq 220$ | 205 | 31 |

We can plot this in a histogram:



Calculations using frequency tables

Your turn with continuous data

| Interval Class | Class Mean (m_j) | Frequency (f_j) |
|----------------------|----------------------|---------------------|
| $120 < x_i \leq 150$ | 135 | 25 |
| $150 < x_i \leq 170$ | 160 | 71 |
| $170 < x_i \leq 190$ | 180 | 73 |
| $190 < x_i \leq 220$ | 205 | 31 |
| | | <hr/> |
| | | $n =$ |

$$\bar{x} \approx \frac{1}{n} \sum_{j=1}^k f_j m_j$$

=

=

Histograms

Definition (Histogram)

A histogram is a diagram consisting of rectangles. The width of each rectangle is the width of its class. The area of each rectangle represents the relative frequency of observations in that class.

- The locations of the rectangles will correspond to the choice of class intervals along the x -axis.
- In Microsoft Excel, the class intervals are called bins.
- The **area** (not height!) of each rectangle represents the relative frequency of observations that occur within the interval. The total area sums to one.

Important!

Bar charts and histograms are not the same. The width of the rectangles in a bar chart does not matter (bars represent categories).

Histograms

- The choice of both the number and size of intervals can influence the shape (interpretation) of a histogram.
- The area of the rectangle corresponds to the relative frequency within each class.
- Care should be taken when intervals are not of equal length.

Example (Blood sugar readings)

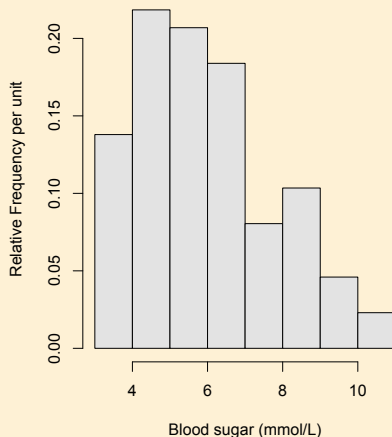
87 blood sugar readings (in millimoles/liter) of someone taken over Jan/Feb using the same machine. The readings vary across the time periods as a result of meals, physical activity, stress. . .

4.1 6.7 8.8 6.4 4.3 3.9 5.7 4.4 4.6 7.8 3.6 4.5 8.5 3.1 4.4 4.9 5.8 4.2 7.5 3.3 9.7 5.2
8.5 6.2 6.6 4.3 6.7 9.7 5.0 6.6 5.2 5.6 4.1 4.9 7.7 5.7 6.8 3.9 5.9 4.5 8.1 8.5 4.4 7.2
5.2 5.7 3.8 6.0 9.4 3.0 6.9 5.8 5.7 8.4 3.8 5.4 5.6 3.7 8.9 3.9 10.4 6.9 6.1 9.4 5.7 3.8
8.3 4.2 10.6 4.4 3.4 6.9 6.1 7.8 4.6 6.8 5.4 6.7 6.3 5.1 7.2 4.2 5.4 4.2 8.7 4.2 7.6

Histograms

Example (Blood sugar readings)

| Class interval | Relative freq. |
|------------------|----------------|
| $3 \leq x < 4$ | $12/87 = 0.14$ |
| $4 \leq x < 5$ | $19/87 = 0.22$ |
| $5 \leq x < 6$ | $18/87 = 0.21$ |
| $6 \leq x < 7$ | $16/87 = 0.18$ |
| $7 \leq x < 8$ | $7/87 = 0.08$ |
| $8 \leq x < 9$ | $9/87 = 0.10$ |
| $9 \leq x < 10$ | $4/87 = 0.05$ |
| $10 \leq x < 11$ | $2/87 = 0.02$ |
| | 1.00 |



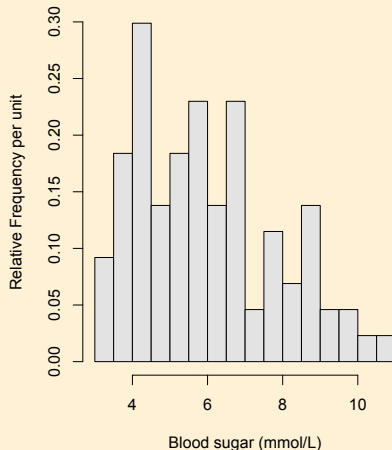
The relative frequencies sum to one, as does the area under the histogram.

Histograms

Example (Blood sugar readings)

| Class interval | Rel. Freq. |
|--------------------|---------------|
| $3 \leq x < 3.5$ | $4/87 = 0.05$ |
| $3.5 \leq x < 4$ | $8/87 = 0.09$ |
| \vdots | \vdots |
| $10 \leq x < 10.5$ | $1/87 = 0.01$ |
| $10.5 \leq x < 11$ | $1/87 = 0.01$ |
| | 1.00 |

Note that because the intervals are 0.5 units wide, we multiply the relative frequencies by 2 to get the height of each rectangle, so the total area sums to 1.

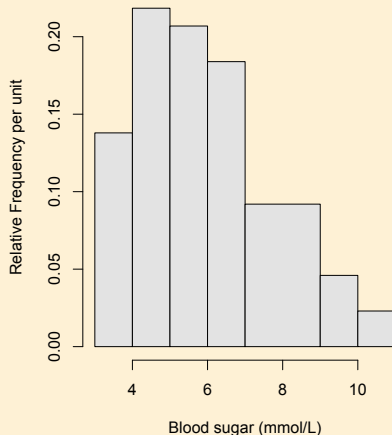


Histograms

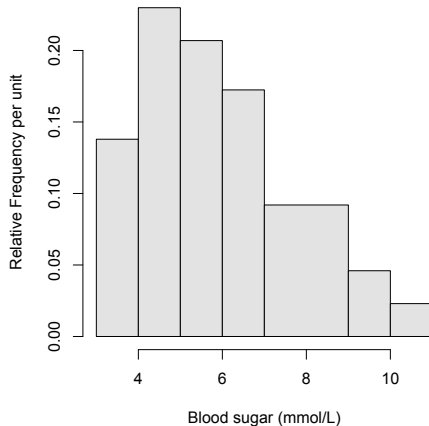
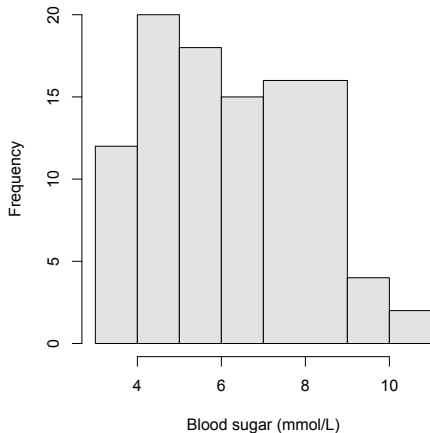
Example (Blood sugar readings)

| Class int. | Relative freq. |
|------------------|----------------|
| $3 \leq x < 4$ | $12/87 = 0.14$ |
| $4 \leq x < 5$ | $19/87 = 0.22$ |
| $5 \leq x < 6$ | $18/87 = 0.21$ |
| $6 \leq x < 7$ | $16/87 = 0.18$ |
| $7 \leq x < 9$ | $16/87 = 0.18$ |
| $9 \leq x < 10$ | $4/87 = 0.05$ |
| $10 \leq x < 11$ | $2/87 = 0.02$ |
| | 1.00 |

One interval is 2 units wide, we divide its relative frequency by 2 to get the relative freq. per unit.



Why use relative frequencies?



Stem and Leaf Plots

- The stem-and-leaf provides a simple and easy way to summarise data and display the shape.
- It retains all data information and shows their distribution.
- It is suitable for relatively small data sets.

Procedure:

1. Separate each data point into stem components and a leaf component. The leaf is in general the least significant figure (last digit).
2. Write all stem digits left of a vertical line.
3. Write all leaf digits right of the vertical line.
4. Re-arrange each leaf digit so each row is ordered – this makes it easier to find the median and quartiles!
5. State the scale, e.g. “the leaf is the 1st decimal place”.

Stem and leaf plots

Example (Blood sugar readings continued)

Single stem:

The decimal point is at the |

```

3 | 013467888999
4 | 1122222334444556699
5 | 012224446677777889
6 | 0112346677788999
7 | 2256788
8 | 134555789
9 | 4477
10 | 46

```

Double stem:

The decimal point is at the |

```

3 | 0134
3 | 67888999
4 | 1122222334444
4 | 556699
5 | 01222444
5 | 6677777889
6 | 011234
6 | 6677788999
7 | 22
7 | 56788
8 | 134
8 | 555789
9 | 44
9 | 77
10 | 4
10 | 6

```

Outliers

Definition (Outliers)

Outliers are measurements that lie outside the so called **upper threshold** (UT) and **lower threshold** (LT). These are defined (at least in this course),

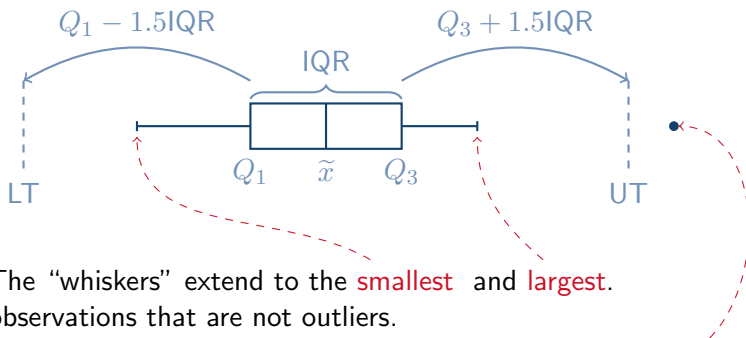
- $LT = Q_1 - 1.5 \times IQR.$
- $UT = Q_3 + 1.5 \times IQR.$

Important!

If an outlier is identified, it should be checked to see if it is genuine. If it was recorded in error (misprint, measurement error) then it should be corrected or excluded from any analysis.

Boxplot

- A boxplot is used to identify outliers and given an impression of the shape of the data.
- The **five number summary** is $\{\min, Q_1, \tilde{x}, Q_3, \max\}$.
- Boxplots graphically highlight Q_1 , \tilde{x} , Q_3 and outliers.



- The “whiskers” extend to the **smallest** and **largest** observations that are not outliers.
- Outliers (above UT and below LT) are identified using **dots**.

Example (Blood sugar readings, $n = 87$)

1. Calculate the five number summary.

\min = smallest observation: $x_{(1)} = 3.0$.

Q_1 = 22nd ordered observation: $x_{(22)} = 4.4$.

$Q_2 = \tilde{x}$ = 44th ordered observation: $x_{(44)} = 5.7$.

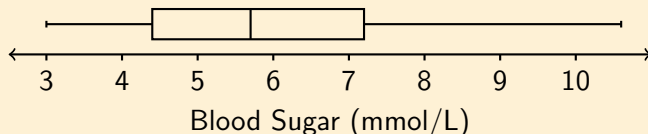
Q_3 = 66th ordered observation: $x_{(66)} = 7.2$.

\max = largest observation: $x_{(87)} = 10.6$.

2. Calculate

- $IQR = Q_3 - Q_1 = 7.2 - 4.4 = 2.8$.
- $LT = Q_1 - 1.5 \times IQR = 4.4 - 1.5 \times 2.8 = 0.2$.
- $UT = Q_3 + 1.5 \times IQR = 7.2 + 1.5 \times 2.8 = 11.4$.

3. The whiskers extend to the most extreme observations that still lie within the lower and upper thresholds.



Group Comparison

If data is split into several groups, boxplots can be used as a comparison tool.

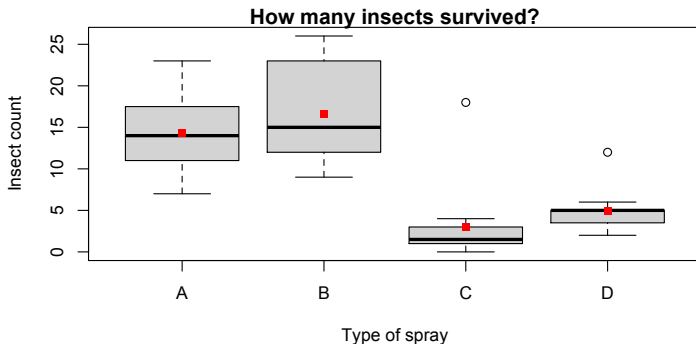
Example

Consider 534 individuals sampled from the US Current Population Survey of 1985.



Why is there more than one measure of variation?

| Type of spray | A | B | C | D |
|--------------------|------|-------------|-------------|------|
| MAD | 4.4 | 6.7 | 1.5 | 1.5 |
| Standard Deviation | 4.7 | 6.2 | 4.9 | 2.5 |
| Range | 16.0 | 17.0 | 18.0 | 10.0 |
| IQR | 4.8 | 10.0 | 2.0 | 1.2 |



Outline

Introduction and Motivation

Notation, Definitions and Sigma Notation

Measures of Location

Measures of Variation

Visualising Data Using Tables and Graphs

Analysing Bivariate Data

Bivariate Data

- A data point may sometimes be observed as a pair, i.e. two measurements are observed on each 'individual' in the study.

Example (Bivariate data)

A researcher is exploring whether there is a relationship between blood pressure and weight. For each person in the study, she observes the person's weight and their blood pressure.

- Usual convention is to write the pairs of observations as:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

- Another popular convention is to display in tabular format

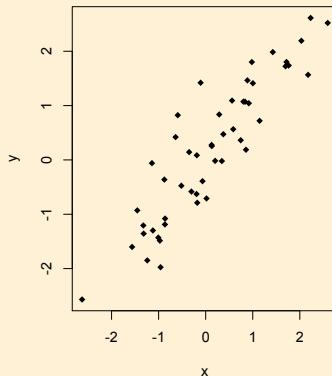
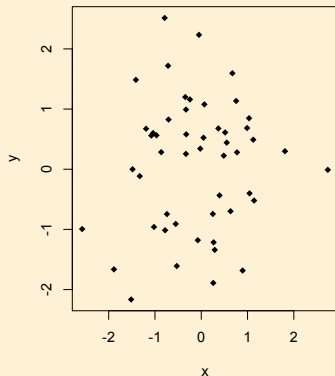
| Observation (i) | 1 | 2 | 3 | ... | n |
|----------------------|-------|-------|-------|-----|-------|
| Variable 1 (x_i) | x_1 | x_2 | x_3 | ... | x_n |
| Variable 2 (y_i) | y_1 | y_2 | y_3 | ... | y_n |

- We want to see if there is a **relationship** between x and y .

Scatterplot

- A scatterplot is used to plot bivariate data.
- Useful tool to determine if there is an association or relationship between the two variables.

Example



Correlation

Definition (Correlation coefficient)

The measure of linear association of bivariate data is the **correlation coefficient**,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

- $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$
- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$
- $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2.$

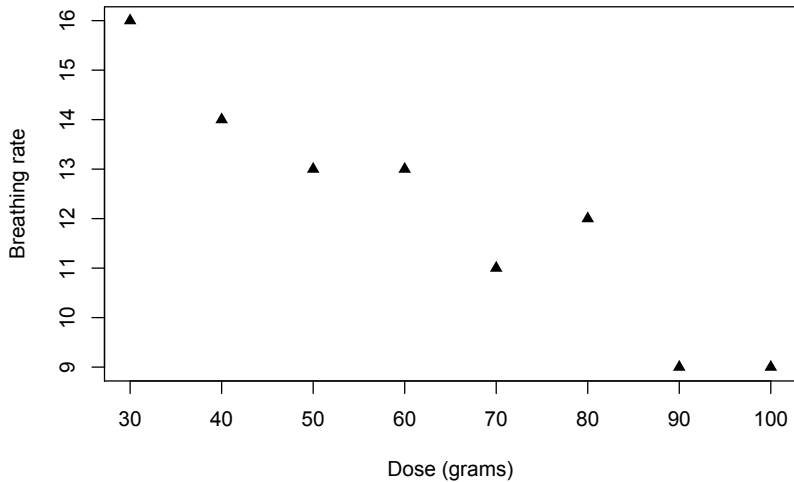
Haldane effect

John Scott Haldane (1860–1936) was a Scottish physiologist famous for intrepid self-experimenting which led to many important discoveries about the human body and the nature of gases.

- One of Haldane's specialties was the physiology of gas absorption and binding in humans.
- He frequently experimented on himself and his second wife.
- To assess carbon dioxide regulation of blood pH he ingested large quantities of sodium bicarbonate (NaHCO_3) to make his blood basic (as opposed to acidic).
- The results are as follows:

| | | | | | | | | |
|-------------------------|----|----|----|----|----|----|----|-----|
| Dose (grams) (x): | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| Breathing Rate (y): | 16 | 14 | 13 | 13 | 11 | 12 | 9 | 9 |

Haldane effect



Correlation of the Haldane data

We need to know:

- $n = 8$
- $\sum_{i=1}^8 x_i = 520$
- $\sum_{i=1}^8 y_i = 97$
- $\sum_{i=1}^8 x_i^2 = 38000$
- $\sum_{i=1}^8 y_i^2 = 1217$
- $\sum_{i=1}^8 x_i y_i = 5910$

Your turn...

- $\bar{x} = \frac{1}{8} \sum_{i=1}^8 x_i =$
- $\bar{y} = \frac{1}{8} \sum_{i=1}^8 y_i =$
- $S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 =$
- $S_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 =$
- $S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} =$
- $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} =$

Properties of r

- r is bounded by ± 1 , i.e. $-1 \leq r \leq 1$.
- r is not affected by a change of scale or origin. In other words, r is scale invariant and location invariant.
- r is symmetric in x and y . I.e. the correlation between x and y is the same as the correlation between y and x .
- r reflects the linear trend of the points
 - If $r > 0$ this means that, in general, y increases when x increases
 - If $r < 0$ this means that, in general, y decreases when x increases
- r^2 is the proportion of all variability in the y 's "explained by" a straight line fitted through the observations.
 - More on this in PHAR2821!

Caution: misinterpreting r

- A value close to ± 1 indicates that nearly all the variability in y can be explained by x .

Important!

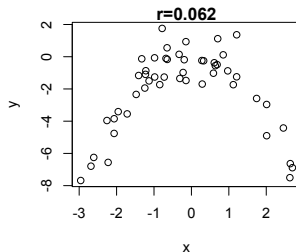
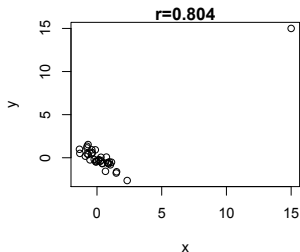
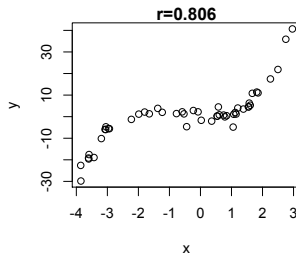
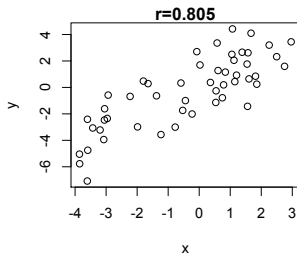
A value of r close to ± 1 is not necessarily an indication of causality!

Example (Ice cream and swim suits)

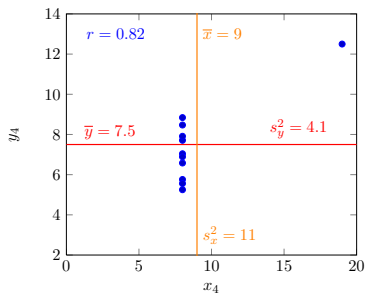
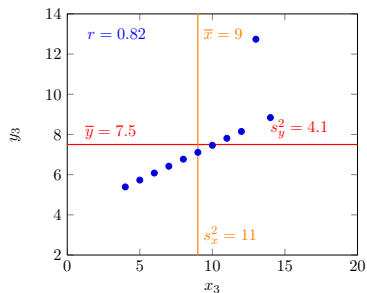
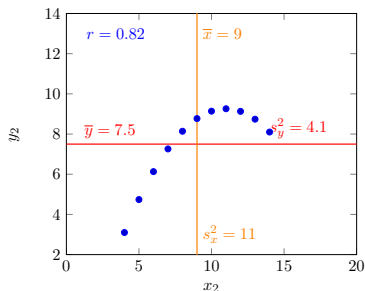
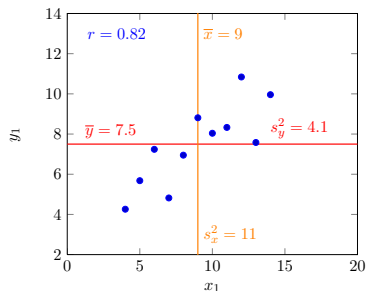
It has been observed that the correlation between ice cream sales and swim suit sales is 0.95. Does this mean that increasing sales of swim suits **causes** an increase in ice cream sales?

Caution: misinterpreting r

The correlation coefficient is a measure of **linear** association.



The importance of graphing your data



References



F.J. Anscombe.

Graphs in Statistical Analysis.

The American Statistician, 27(1):17–21, 1973.



M.C. Phipps and M.P. Quine.

A Primer Statistics.

Pearson Education Australia, 4th edition, 2001.



J.A. Rice.

Mathematical statistics and data analysis.

Duxbury Press, 1995.



H. Rosling.

The Joy of Stats – 200 Countries, 200 Years, 4 Minutes.

BBC Four, 2010.