Understanding Data and Statistical Design (60117)

Chapter 1

Introduction to Statistics

Subject Coordinator: Stephen Woodcock
Lecture notes: Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Chapter outline

Topics:

- data and variables
  - numerical
  - categorical
- exploring data
  - charts
  - statistics
- probability and RVs
  - discrete case
  - continuous case
  - population parameters
- scientific experimentation
  - glossary
  - as a process
  - introductory example

Data and variables

In this module we introduce some basic ideas and techniques used to describe and explore data.

We will do so with reference to data collected from a study of 92 individuals.

The data was generated by recording various attributes of these individuals and their resting pulse rates.

The participants were then divided (randomly) into two groups.

The first group then went for a short run while the second group remained at rest.

After the first group had finished their run, all 92 individuals had their pulse rates recorded for a second time.

Data and variables

The attributes of the participants and the two recorded pulse rates are what we call **variables**.

The variables from the study are summarised in the table below.

Name	**Type**	**Description**
height	numerical	height (in)
weight	numerical	weight (lb)
sex	categorical	"female", "male"
smokes	categorical	regular smoker: "no", "yes"
activity	categorical	activity level: "low", "medium", "high"
ran	categorical	run group: "no", "yes"
*pulse*1	numerical	pulse rate recording 1 (bpm)
*pulse*2	numerical	pulse rate recording 2 (bpm)

If we want to be more precise, we can consider the data collected from the experiment to be observations of **random variables (RVs)**.

Data and variables

The data from this study is in chapter1.csv (available on Canvas), an excerpt of which is copied below.

	height	weight	sex	smokes	activity	ran	pulse1	pulse2
1	66.00	140	male	no	medium	yes	64	88
2	72.00	145	male	no	medium	yes	58	70
3	73.50	160	male	yes	high	yes	62	76
4	73.00	190	male	yes	low	yes	66	78
5	69.00	155	male	no	medium	yes	64	80
6	73.00	165	male	no	low	yes	74	84
7	72.00	150	male	no	high	yes	84	84
8	74.00	190	male	no	medium	yes	68	72
9	72.00	195	male	no	medium	yes	62	75
10	71.00	138	male	no	medium	yes	76	118

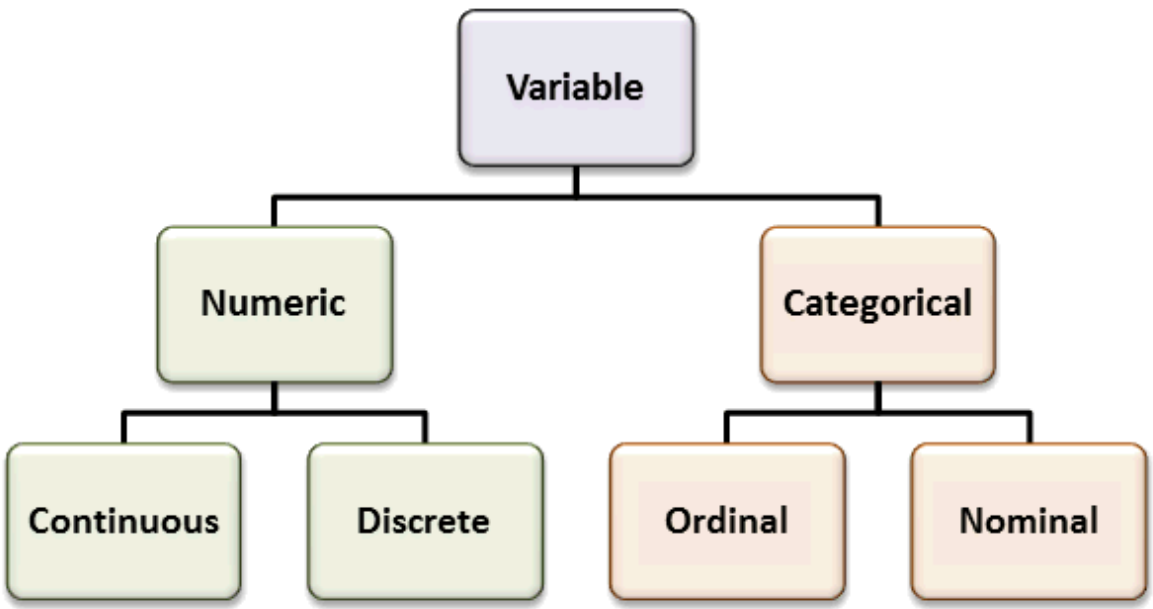
To refer to elements in the data sample we use subscript notation, e.g.

$$height_2 = 72.00 \quad \text{or} \quad ran_9 = \text{“yes”}.$$

Data and variables

Notice that we have classified the variables into two types, **numerical** and **categorical**, which is important as these take different types of values.

We can also go further in our classification of variables, as the diagram below shows.



Source: www.abs.gov.au

Data and variables – numerical

These variables take values that are numbers.

We can further classify numerical variables as either **discrete** or **continuous**.

Discrete numerical variables take values in sets that are called "countable". Examples include

- all whole numbers between 0 and 10
- the set of all whole numbers (called "integers").

Continuous numerical variables take values in sets that are called "uncountable". Examples include

- all numbers between 0 and 10
- the set of all numbers (called "reals" or the "real line").

The discrete/continuous classification is important as these two types of numerical variables have very different properties.

Data and variables – categorical

These variables take values that are not numbers, called the **categories**, **levels** or **states** of the variable.

Nominal categorical variables are those where the sates have no intrinsic order. Examples include:

- postcode
- whether a person exercises or not.

Ordinal categorical variables are those where the sates have an intrinsic order. Examples include

- level of education
- defence force rank.

Note that sometimes it is useful to code the states of categorical variables as numbers, but this does not makes such variables numerical (an example is the postcode variable above).

Exploring data

We often have to deal with large amounts of data, sometimes many thousands or even millions of data points (billions or more in the case of "big data").

This means it is virtually impossible to analyse all data points and have any hope of extracting useful information.

In statistics we instead make use of various tools to describe characteristics of the data.

These tools are either **graphical (charts)** or **numerical (statistics)**.

Charts can be used to explore all variables.

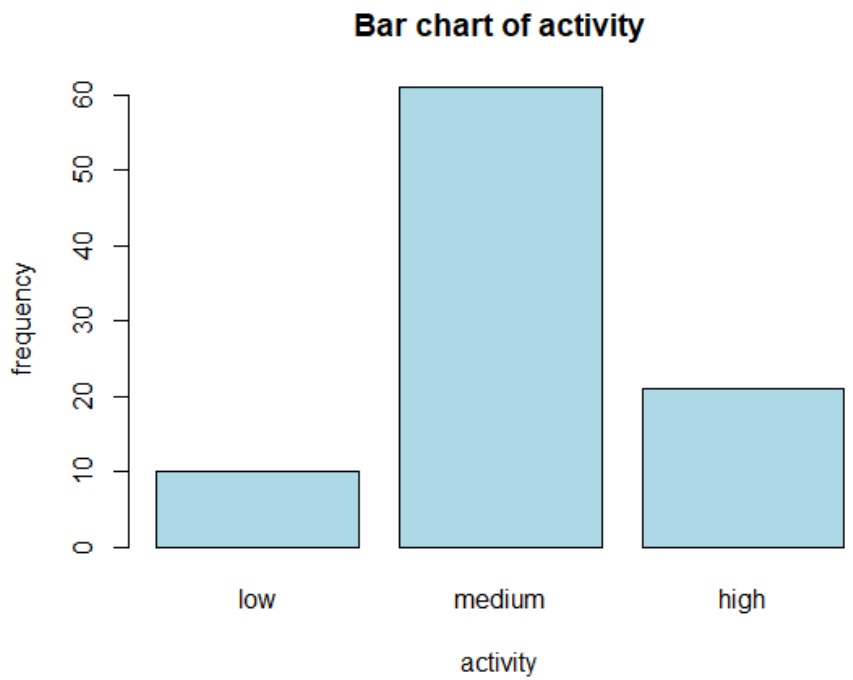
Statistics are used only on numerical variables.

Exploring data – charts

Bar chart

A bar chart is used for categorical variables and counts the number of occurrences (frequencies) of each state of the variable.

The chart below is of the *activity* variable and has been constructed using R (see code file chapter1.R on Canvas).



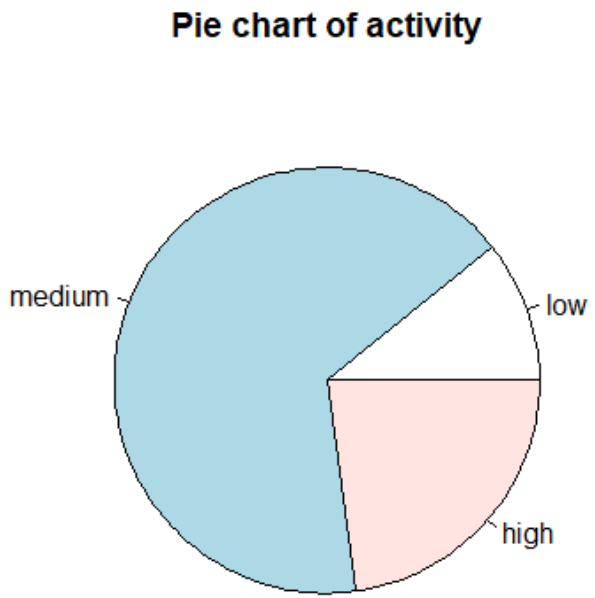
The frequencies can be read off the vertical axis.

Exploring data – charts

Pie chart

An alternative to a bar chart is a pie chart, which displays the frequencies as segments of a circle.

The chart below is again of the *activity* variable.



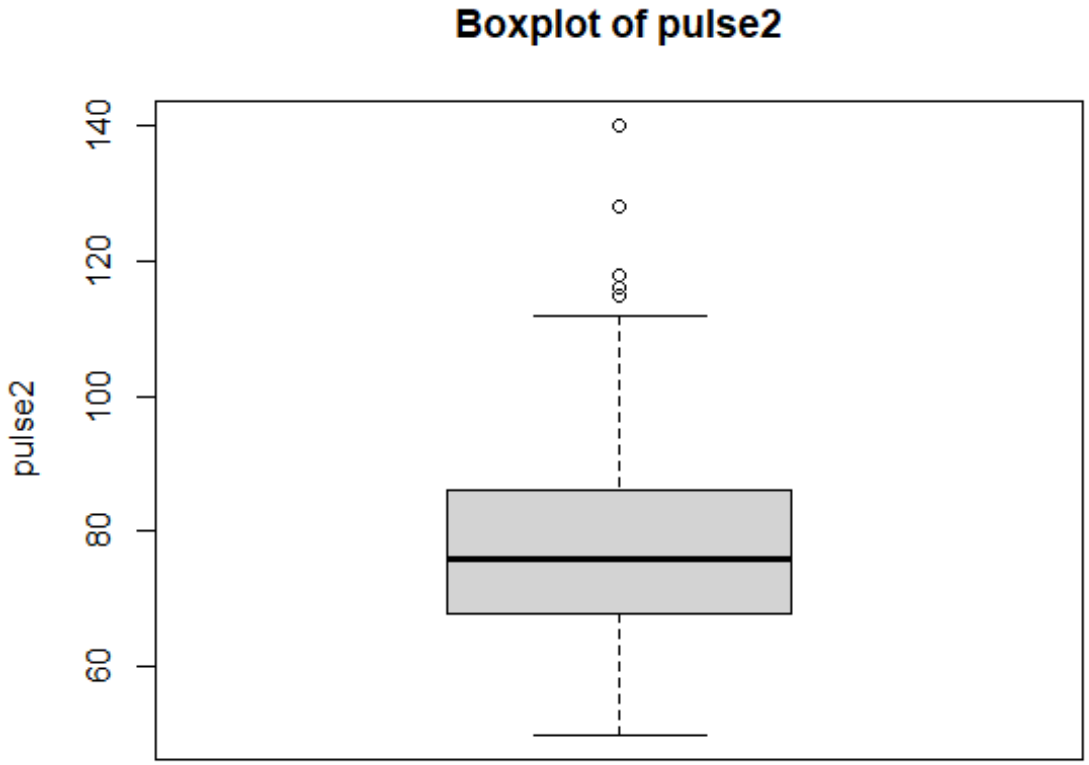
Bar and pie charts becomes harder to interpret when the number of states of the variable becomes large (which is why they are seldom used for numerical variables).

Exploring data – charts

Boxplot

When a numerical variable can take many values, it is useful to get an idea of its **distribution**, which we can do by looking at the **shape of the sample data**.

One way is with a **boxplot**, shown here for the variable *pulse*2.



Exploring data – charts

The **line in box** identifies the **sample median** which separates the lower and upper halves of observations for that variable. This gives us an idea as to the **location** or **centre** of the sample data.

The **box height** is called the **sample interquartile range (IQR)** and covers (approximately) the middle half of observations for that variable. This gives us an idea as to the **scale** or **spread** of the sample data.

The **whiskers** and the box cover the range of the data excluding **outliers**, which are marked with circles.

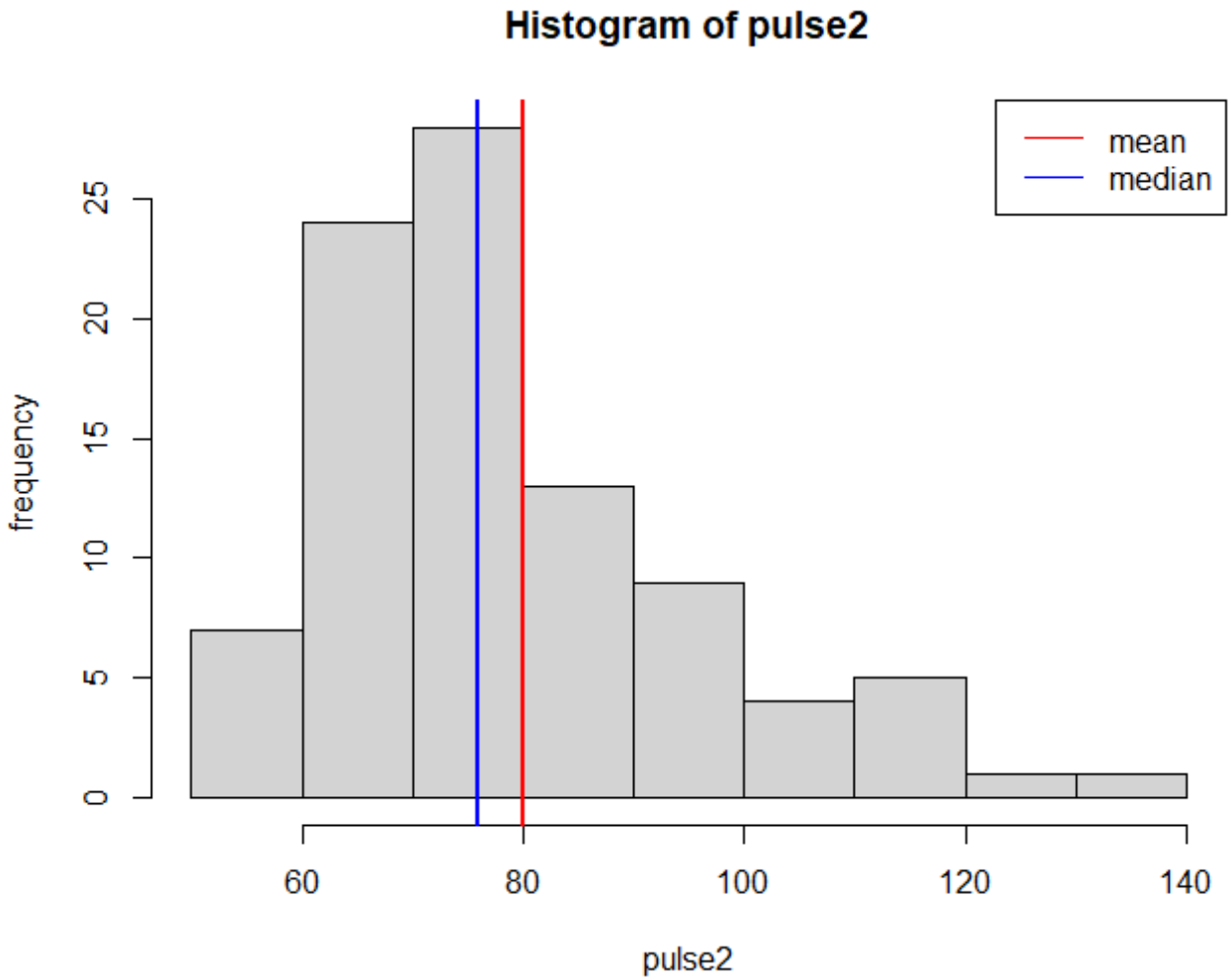
The boxplot can also be used to assess the **symmetry** of the sample data.

For *pulse*2 we see the sample data is not symmetric, as the upper tail is much longer than the lower tail.

Exploring data – charts

Histogram

An alternative to a boxplot is a histogram, shown below for *pulse*2.



Exploring data – charts

A histogram counts the frequencies of observations occurring in ranges called **bins**.

The red line represents the **sample mean** or **sample average** of observations for the variable.

The blue line represents the **sample median**.

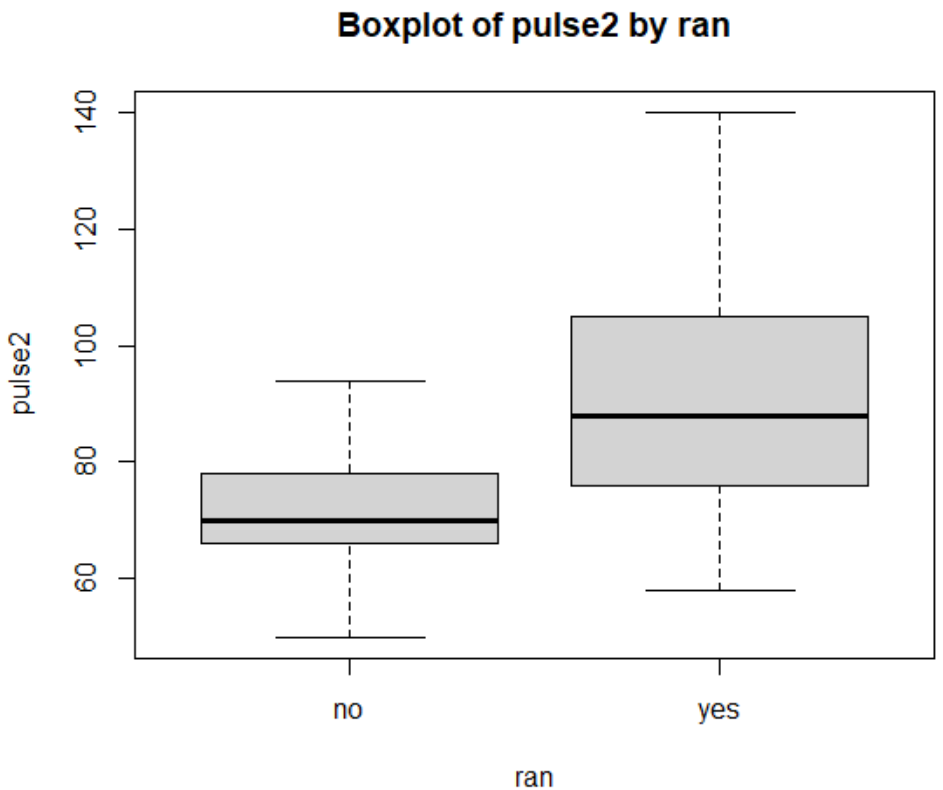
The histogram can also be used to assess the **symmetry** of the sample data.

For *pulse*2 we see the sample data is not symmetric as the upper (or right) tail is much longer than the lower (or left) tail.

Exploring data – charts

Grouped boxplot

Sometimes it is useful to construct boxplots of a numerical variable for each state of a categorical variable, shown below of *pulse*2 for each level of *activity*.

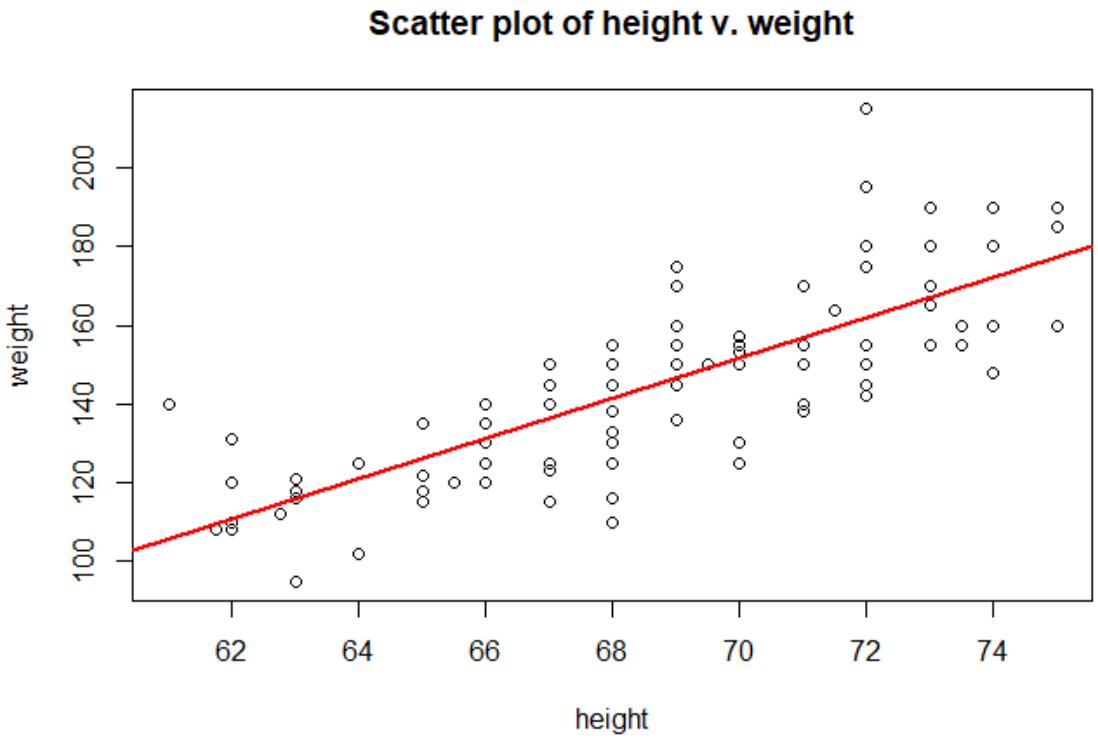


The effect of running on the second pulse rate reading is clear, with different centres and scales of the two samples of data.

Exploring data – charts

Scatter plot

A scatter plot can be used to assess the relationship between two numerical variables, shown below for *height* and *weight* (with fitted regression line superimposed over the top).



We see a positive relationship between the variables, which makes sense – the taller the individual the heavier (on average).

Exploring data – statistics

Just as graphical tools can be used to assess the centre, scale and symmetry of sample data, so too can numerical tools.

These numerical tools come in the form of **statistics**, some of which we have already encountered in the example charts above.

We will introduce some of the these for the variables *pulse*2, *height* and *weight*.

Exploring data – statistics

Centre/location statistics
These statistics are used to describe the centre or location of the sample data.

The **sample mean** is the most commonly used, which for *pulse2* is denoted and calculated as

$$\begin{aligned}\overline{pulse2} &= \frac{1}{92}\sum_{i=1}^{92} pulse2_i \\ &= \frac{pulse2_1 + pulse2_2 + \cdots + pulse2_{92}}{92} \\ &= \frac{88 + 70 + \cdots + 76}{92} = \frac{7360}{92} = 80.\end{aligned}$$

That is, we sum all the observations and divide by the sample size.

Exploring data – statistics

The **sample median** is another commonly used centre/location statistic, which is the value such that above and below are an equal number of observations.

If we want to find the median of the sample of *pulse*2, we first need to sort the observations.

Because the sample size 92 is an even number, we then take the average of the middle two sorted values.

That is, we calculate the median as

$$m = \frac{pulse2_{46}^{(\text{sort})} + pulse2_{47}^{(\text{sort})}}{2} = \frac{76 + 76}{2} = 76$$

where $pulse2_i^{(\text{sort})}$ is the i-th sorted value of *pulse*2 sample.

Were the sample size an odd number, the median would be the middle sorted value.

Exploring data – statistics

Scale/spread statistics
These statistics are used to describe the scale or spread of the sample data.

The **sample variance** is most commonly used, which for *pulse*2 is denoted and calculated as

$$\begin{aligned} s^2 &= \frac{1}{92-1}\sum_{i=1}^{92}(pulse2_i - \overline{pulse2})^2 \\ &= \frac{(pulse2_1 - \overline{pulse2})^2 + \cdots + (pulse2_{92} - \overline{pulse2})^2}{91} \\ &= \frac{(88-80)^2 + \cdots + (76-80)^2}{91} = \frac{26590}{91} = 292.1978. \end{aligned}$$

The **sample standard deviation** $s = 17.09379$ is the square root of the sample variance and has the same units as *pulse*2 (i.e. bpm).

Exploring data – statistics

Association/relationship statistics
These statistics are used to describe the strength of association or relationship between two variables.

The **sample covariance** a common measure of linear association, which for *height* and *weight* is calculated as

$$\begin{aligned} q &= \frac{1}{92-1}\sum_{i=1}^{92}(height_i - \overline{height})(weight_i - \overline{weight}) \\ &= \frac{(66-68.72)(140-145.15)+\cdots+(61.75-68.72)(108-145.15)}{91} \\ &= \frac{6204.457}{91} = 68.18084. \end{aligned}$$

The positive value tells us that there is a positive (linear) association between *height* and *weight*, which agrees with what we saw in the scatter plot earlier.

Exploring data – statistics

The **(Pearson) sample correlation** is often preferred over the sample covariance as the former does not depend on the scale of the data.

For *height* and *weight* this is calculated as

$$r = \frac{q}{s_h s_w} = \frac{68.18084}{3.659291 * 23.7394} = 0.7848664$$

where s_h and s_w are sample standard deviations of *height* and *weight* respectively.

That is, to calculate sample correlation we take the sample covariance and divide by the product of the sample standard deviations.

Correlation is a number between -1 and 1 where

- -1 represents perfect negative (linear) association
- 0 represents no linear association
- 1 represents perfect positive (linear) association.

The value $r = 0.7848664$ indicates the linear relationship between *height* and *weight* is medium-strong.

Exploring data – statistics

A summary of various statistics can be generated using R.

Below are location and scale sample statistics for *pulse*2.

	stat	value
1	sample size	92.00000
2	mean	80.00000
3	median	76.00000
4	variance	292.19780
5	standard deviation	17.09379
6	IQR	17.00000

Below are scale and association sample statistics for *height* and *weight*.

	stat	value
1	standard deviation height	3.6592907
2	standard deviation weight	23.7393978
3	covariance height and weight	68.1808409
4	correlation height and weight	0.7848664

Probability and RVs

When we collect sample data, it is not the sample that is of primary interest to us – it is the population from which the sample was collected that we wish to learn about.

The example sample data we have been considering is of 92 individuals, but we wish to apply what we can learn about this sample to the much larger population of individuals from whom the 92 were selected.

We call this type of analysis **statistical inference**.

For many of the statistical tools we use in our analysis, assumptions need to be made about the nature of the sample data that we have collected.

More precisely, we often **need to assume a sample of data is from a particular distribution**.

In this section we look at some basics of **probability theory that underpins statistical analysis**.

Probability and RVs – discrete case

Let X be a discrete RV.

Associated with such a RV is its **probability mass function (PMF)**

$$p_X(x) = \text{Prob}(X = x)$$

with the properties

$$\sum_x p_X(x) = 1 \quad \text{and} \quad 0 \leq p_X(x) \leq 1.$$

Associated with any RV is the **cumulative distribution function (CDF)**, defined in the discrete case as

$$F_X(x) = \text{Prob}(X \leq x) = \sum_{u \leq x} p_X(u),$$

which is the sum of probabilities up to and including $X = x$.

The PMF and CDF of X define the **distribution** of X.

Probability and RVs – discrete case

Binomial distribution
A binomial RV X is denoted $X \sim Bin(n, p)$, where n and p are called the **parameters** of the distribution.

Such a RV can be used to model a very simple experiment consisting of a sequence of n trials, where each trial has probability of success p and where the outcome of each trial is **independent** of the outcomes of other trials.

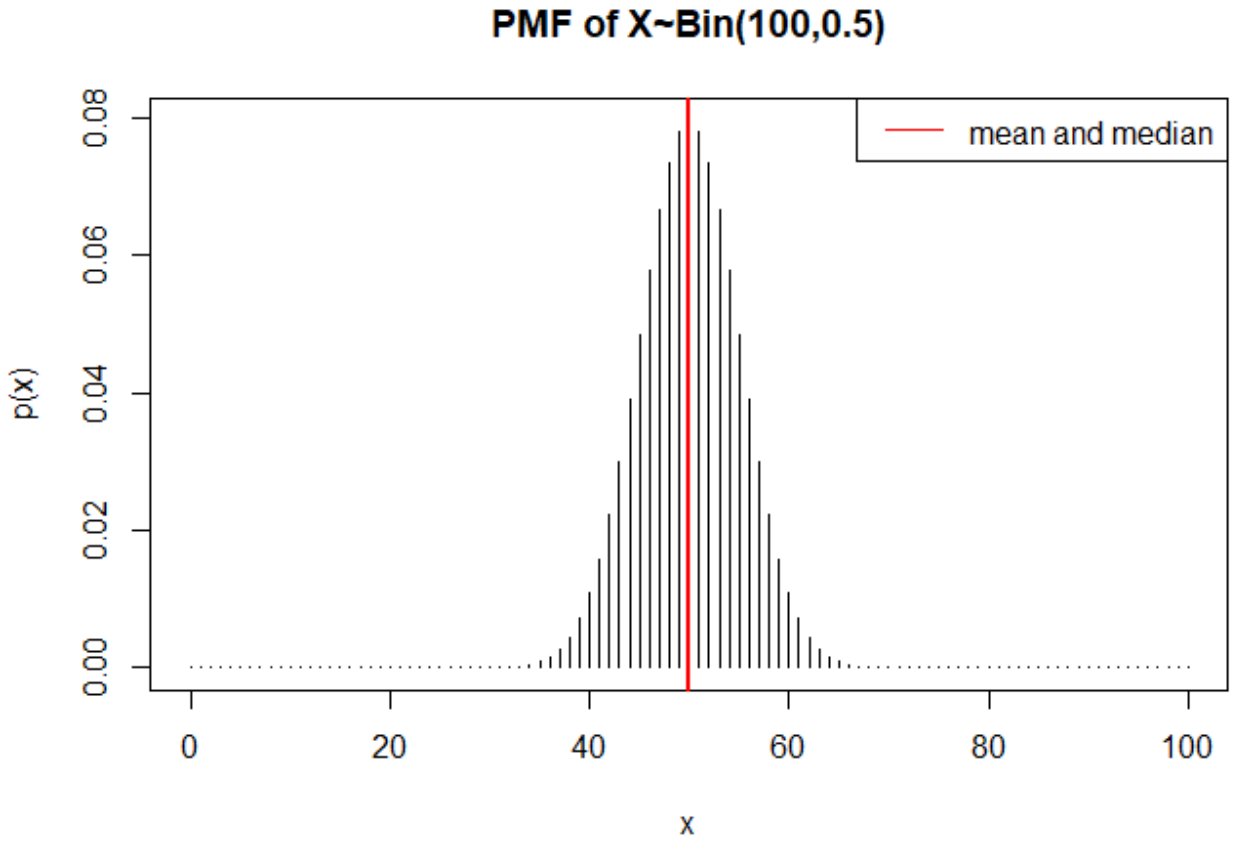
The RV X represents the number of successful outcomes from the n trials, so X can take the values $\{0, 1, 2, \ldots, n\}$.

For example, this RV can model a sequence of $n = 100$ coin tosses where the probability of obtaining a head on each coin toss is $p = 0.5$, i.e. $X \sim Bin(100, 0.5)$.

We can then use this distribution to calculate, for example, the probability of obtaining 50 heads from 100 coin tosses, which turns out to be $P(X = 50) = 0.07958924$.

Probability and RVs – discrete case

Below is a plot of the PMF of the RV $X \sim Bin(100, 0.05)$.



Note that the distribution of this RV is symmetric and the **population mean** and **population median** are equal.

Probability and RVs – continuous case

Let Y be a continuous RV.

If one can write

$$\text{Prob}(a \leq Y \leq b) = \int_a^b f_Y(y)\,\mathrm{d}y,$$

then f_Y is called the **probability density function (PDF)** of the RV Y, which possess the properties

$$\int_{-\infty}^{\infty} f_Y(y)\,\mathrm{d}y = 1 \quad \text{and} \quad f_Y(y) \geq 0.$$

In this case the **cumulative distribution function (CDF)** is defined as

$$F_Y(y) = \text{Prob}(Y \leq y) = \int_{-\infty}^{y} f_Y(u)du,$$

which is the area between f_Y and the horizontal axis up to and including $Y = y$.

The PDF and CDF of Y define the distribution of Y.

Probability and RVs – continuous case

Normal/Gaussian distribution
A normal RV Y is denoted $Y \sim N(\mu, \sigma)$, where μ and σ are called the **parameters** of the distribution.

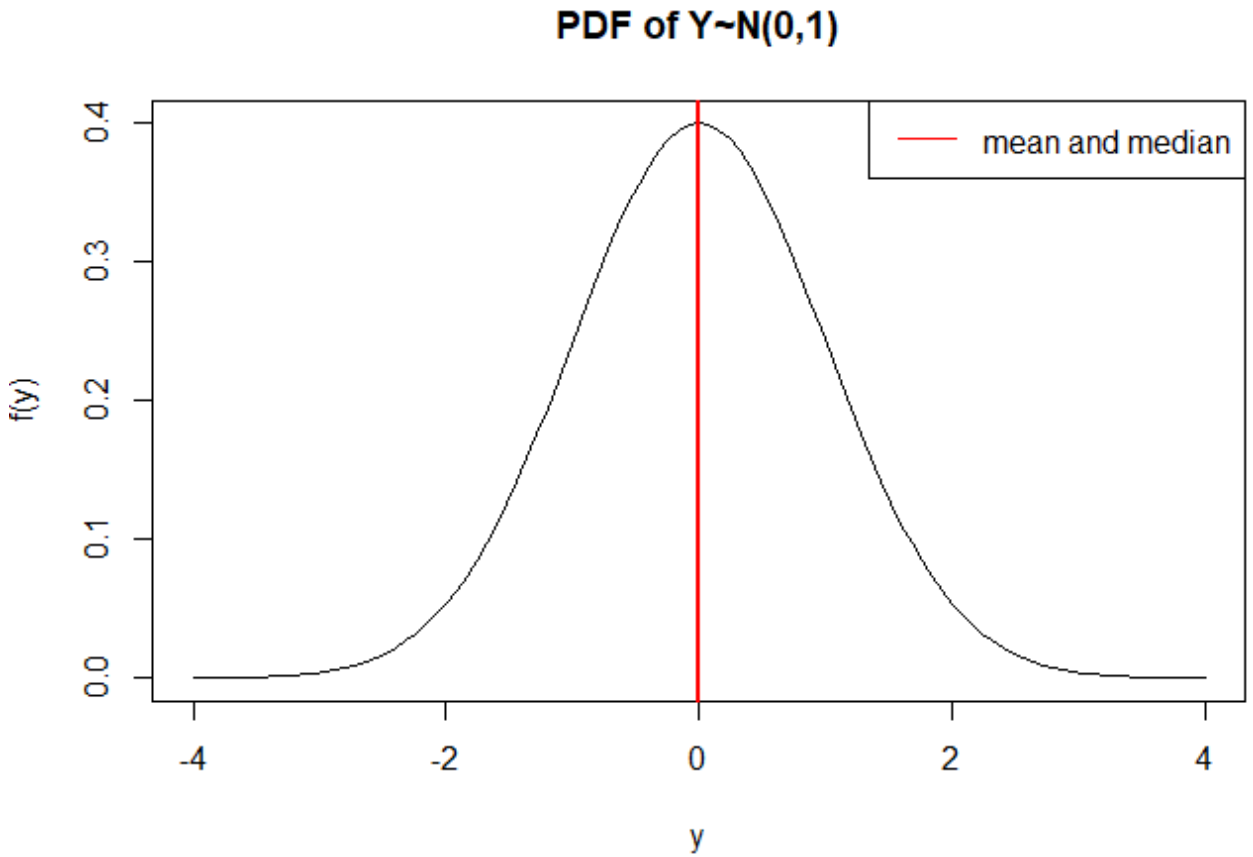
The parameters μ and σ are respectively the **population mean** and **population standard deviation** of the distribution.

Such a RV is a fundamental statistical object whose importance stems from the **central limit theorem**, which is one of the most important theorems in probability and statistics.

Many of the statistical tools we consider in this subject rely on the assumption that the sample data is from a normal distribution.

Probability and RVs – continuous case

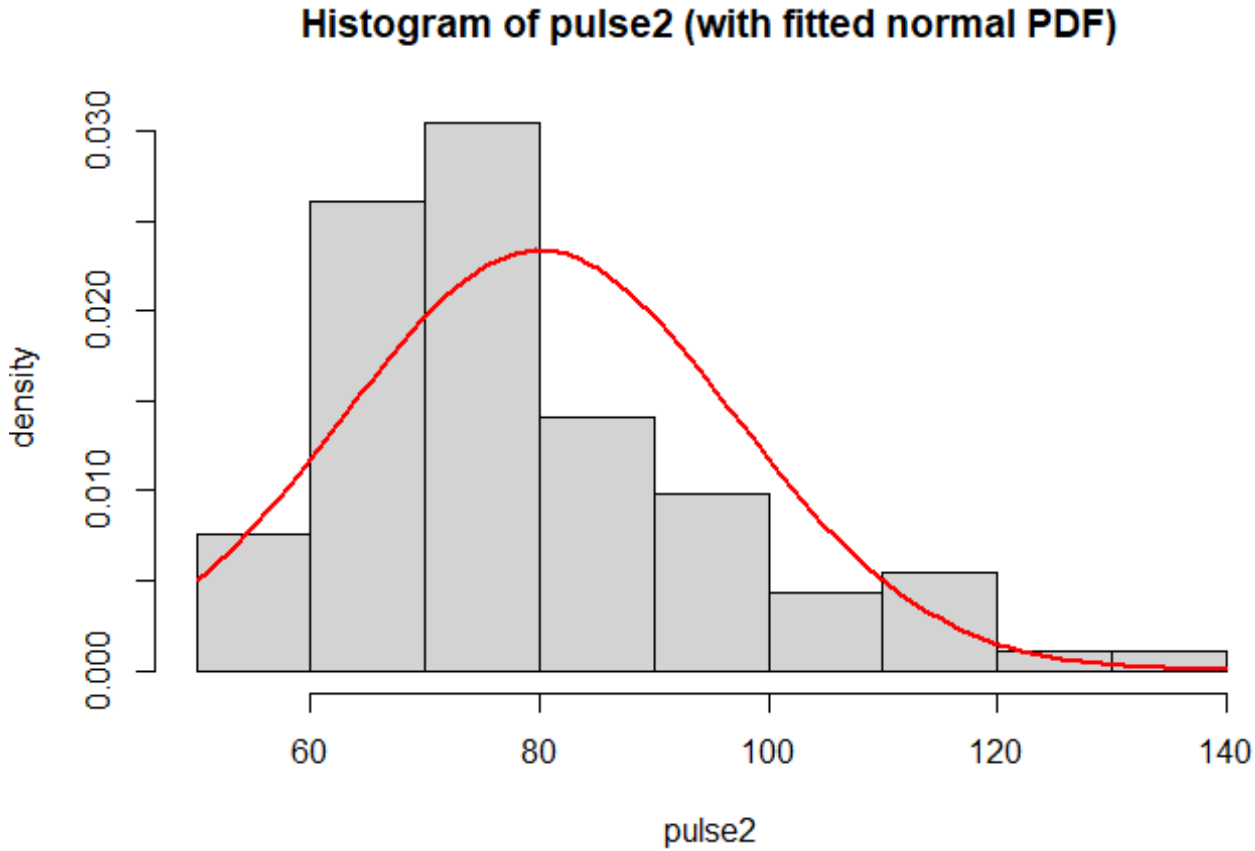
Below is a plot of the PDF of the RV $Y \sim N(0, 1)$.



Note that the distribution of this RV is symmetric and the **population mean** and **population median** are equal.

Probability and RVs – continuous case

Below is a histogram of *pulse2* data with fitted normal PDF in red.



Comparing the histogram to the normal density, we can conclude it is unlikely that the *pulse2* sample is from a normal distribution, as the histogram and density look too dissimilar.

Probability and RVs – population parameters

We have already encountered **statistics**, which can be used to describe the location and scale of **sample data**.

These statistics are estimates of population parameters.

The **population parameters** are used in a similar way to describe the location and scale of **distributions**.

We always consider a sample of data to be from a population distribution, although we may not know what distribution that is.

The population parameters are defined using the PMF or PDF of the RV.

Probability and RVs – population parameters

Centre/location parameters
These parameters are used to describe the centre or location of the distribution of a RV.

The **population mean** or **population average** or **expected value** of the discrete RV X is defined as

$$\mu_X = \mathbb{E}[X] = \sum_x x p_X(x)$$

and the continuous RV Y as

$$\mu_Y = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y.$$

For a sample of data drawn from a population, the sample mean is an estimate of the expected value.

The **population median** is another example of a location parameter, which can be estimated with the sample median.

Probability and RVs – population parameters

Scale/spread parameters
These parameters are used to describe the scale or spread of the RV (or population distribution).

The **population variance** of the discrete RV X is defined as

$$\sigma_X^2 = \text{var}(X) = \mathbb{E}[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 p_X(x)$$

and the continuous RV Y as

$$\sigma_Y^2 = \text{var}(Y) = \mathbb{E}[(Y - \mu_Y)^2] = \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y)\,\mathrm{d}y.$$

For a sample of data drawn from a population, the sample variance is an estimate of the population variance.

The **population standard deviation** is the square root of the population variance and is an estimate of the sample standard deviation.

Probability and RVs – population parameters

Association/relationship parameters
These parameters are used to describe the strength of association or relationship between two RVs.

The **population covariance** is a common measure of linear association, which for RVs X_1 and X_2 (discrete or continuous) is defined as

$$\text{covar}(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)],$$

where μ_1 and μ_2 are the expected values of the RVs X_1 and X_2 respectively.

For two samples of data drawn from two populations, the sample covariance is an estimate of the population covariance.

Probability and RVs – population parameters

The **(Pearson) population correlation** is often preferred over the population covariance as the former does not depend on the scale of the distribution.

We define this as

$$\rho = \text{corr}(X_1, X_2) = \frac{\text{covar}(X_1, X_2)}{\sigma_1 \sigma_2},$$

where σ_1 and σ_2 are the population standard deviations of the RVs X_1 and X_2 respectively.

Correlation is a number between -1 and 1 where

- -1 represents perfect negative linear association
- 0 represents no linear association
- 1 represents perfect positive linear association.

For two samples of data drawn from two populations, the sample correlation is an estimate of the population correlation.

Scientific experimentation

When we talk about **experiment design** we mean a statistical approach to design.

No matter what the field of research, all experiments have one thing in common – the use of statistics to analyse data and to draw conclusions.

If you understand statistics then you can analyse and interpret data from scientific experiments in many different disciplines.

Scientific experimentation

The statistical methodology may be very simple, such as reporting frequencies of experimental outcomes, or complex, such as multi-factor ANOVA and multiple regression.

Many of the more advanced statistical tools rely on certain assumptions and this requires the experiment be designed so that they are satisfied. For example, the common statistical requirement that experimental observations be independent is behind the use of **randomisation**.

Moreover, an understanding of statistics can inform the design process in ways that allow conclusions to be strengthened. We will see some examples of this later, such as the use of **blocking factors** to control sources of variation.

Scientific experimentation – glossary

Factor. A categorical variable taking certain **levels** that is manipulated by the experimenter, e.g. dosage of a drug. A factor is an independent variable, often called an **explanatory variable** or **predictor**.

Treatment. Combination of factor levels that is allocated to an **experimental unit**, e.g. a medical procedure.

Response. A measurement or observation of a **measurement unit**, e.g. blood pressure of an individual. A response is a dependent variable.

Randomisation. Treatments assigned randomly to experimental units (or vice versa). This is necessary for the common assumption of independent observations to be satisfied.

Replication. Experiment is often performed multiple times to reduce variability in results. Many statistical properties improve as sample size increases.

Scientific experimentation – as a process

The following outline of the process of scientific experimentation is adapted from Montgomery (1984). This is an older reference but a classic.

The text breaks the process down into seven sequential components.

1. **Definition of Research Problem**. Seemingly obvious, but essential to the planning process.
2. **Selection of Factors, Levels and Treatments**. The factors are the independent variables and the levels the values they take (factors may be fixed or random – we consider only fixed). The set of combinations of factor levels are the treatments. It is the effect of the treatments that are the subject of the research.
3. **Selection of Response variable.** The effect of treatments is measured via the response variable. Responses can be numerical and/or categorical, a property that determines the analytical tools available for the problem.

Scientific experimentation – as a process

4. **Experiment Design.** Of primary importance, this step determines whether research questions can be answered or not. Considerations here must include:
   - effect size – threshold value or change in value of response variable
   - risk tolerance – Type I and II error settings
   - determination of sample size – a function of effect size and error tolerances.

   Once these questions have been answered various template designs can be considered, e.g. completely randomised-block, Latin squares etc.
5. **Experiment Execution.** Where the rubber hits the road, the step where the treatments are administered and the data collected.

Scientific experimentation – as a process

6. **Data Analysis.** Statistical analysis of predictor and response variable data. Tools include:
   - descriptive statistics
   - graphical analysis
   - hypothesis tests
   - linear regression, logistic regression
   - etc.
7. **Conclusions and Recommendations.** Interpretation of results from data analysis, conclusions drawn as to effectiveness of treatments, recommendation for further action.

Our focus will be on Steps 4 and 6, Experiment Design and Data Analysis.

Scientific experimentation – introductory example

Let's illustrate the main ideas with a simple example.

The Scenario. Consider a new drug suspected of having an effect on blood pressure, and the problem of constructing an experiment to test whether this effect is statistically-significant.

We go through the seven steps from above.

1. **Definition of Research Problem**. We wish to test whether the population mean blood pressure of those administered the drug can be inferred to be different from those taking a placebo.
2. **Selection of Factors, Levels and Treatments**. The dosage of the drug is the **factor**, the levels are the different dosages at which the drug will be administered and the **treatments** correspond to the levels (treatments and levels are equivalent for single-factor experiments). A zero dosage level represents administration of a placebo.
3. **Selection of Response.** The response is the blood pressure of the individuals (**measurement units**) administered the drug.

Scientific experimentation – introductory example

4. **Experiment design**. Each treatment is randomly assigned to an **experimental unit** (groups of individuals), with the principle of replication followed so that each treatment is tested multiple times. Considerations:
   - effect size – difference between actual and hypothesised mean blood pressure we wish to detect
   - risk tolerance – Type I error or **significance level** is usually set as $\alpha = 0.05$ or $\alpha = 0.01$, Type II error is often set in compromise to sample size (and therefore cost).
   - determination of sample size – a function of effect size and error tolerances.

   We also need to consider potential extraneous sources of variability in the response variable:
   - are we testing on both males and females?
   - are there other characteristics of the experimental units that can influence the response (e.g. weight, general level of health etc.)?

   If yes, then blocking factors need to be considered (more when we look at two-way ANOVA).

Scientific experimentation – introductory example

5. **Experiment Execution**. Drug administered, blood pressure data collected.
6. **Data Analysis**. The tools employed depend on the number of levels, and also whether blocking factors have been used:
   - if no blocking factors have been employed and only two levels (dosages) administered, then a two sample independent T-test is appropriate (more next chapter)
   - if no blocking factors have been employed and more than two levels (dosages) administered, then one way ANOVA and associated F-test is appropriate (more in later chapters)
   - if blocking factors have been employed, then two way, three way etc. ANOVA and associated F-tests are appropriate (more in later chapters).
7. **Conclusions and Recommendations.** Depend on outcome of Step 6.

Question. How can this design be improved? Think about the use of the response variable and if this is adequate for the aims of the experiment.

References I

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27:17–21.

Montgomery, D. C. (1984). *Design and analysis of experiments*. 2 edition.

Peck, R., Olsen, C., and Devore, J. (2012). *Introduction to statistics & data analysis*. Brooks/Cole, 4th edition.