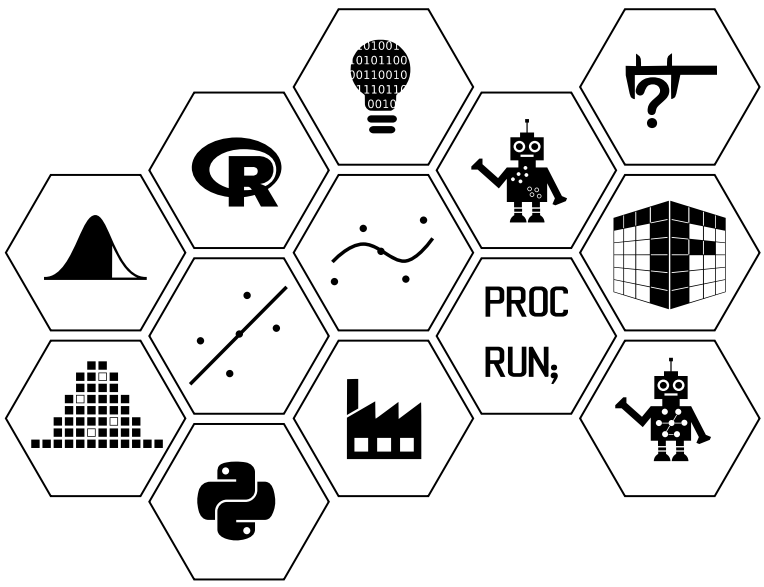
Learning from Data/Data Science Foundations

Week 2: Visualising and summarising data



DATA ANALYTICS
GLASGOW

Visualising and Summarising Data

In week 1, we introduced the ideas of different data sources, structures and data types; approaches to collecting and collating data and an introduction to approaches for extracting/summarising data. Once we have obtained our data, the next important step is to summarise and visualise the data in an appropriate form as part of quality assuring the data, and to get a subjective answer to our questions of interest.

Summaries and plots of your data can help you to identify an appropriate probability model for your data and can also be used to identify features in your data that it might be useful to account for within a statistical model. Identifying features about the nature of your data, or understanding complex features of your data will help to indicate the appropriate methods for analysis.

It's also important to carefully consider the context of your data, the question of interest and the type of data that have been recorded. It might be that important factors about your data cannot be identified through summaries and plots, but have arisen naturally through the data sourcing and collection process.

Week 2 learning material aims

The material in week 2 covers:

- summarising and visualising data;
- identifying data types from the data context;
- quality assuring data;
- exploring relationships in data;
- identifying data features.

Summarising Data

Once we have sourced data, collected it and collated it we refer to all the individual elements of data referring to one particular characteristic (variable) as a *dataset*.

Numerical Data

When all of the individual elements of a dataset are numeric then we are interested in how the numeric values of the data are *distributed* along a number line. We therefore refer to the *distribution* of the data. The *distribution* of a *dataset* can be summarised in terms of location, spread and shape. This enables us to summarise the main properties of the data and we will use these properties in later weeks of the course in order to consider appropriate probability distributions that the sample of data may have arisen from.

- **Location** - this is typically the centre of the distribution (mean/median)
- **Spread** - this is the variation/range of values covered by the data (range/standard deviation/inter-quartile range)
- **Shape** - this is the shape of the distribution

The **median** is found by arranging the data numeric values in ascending order and identifying a value that divides the dataset into two groups with, as near as possible, the same number of observations in each group. If the number of observations is an odd number, the median is defined to be the middle value; if the number of observations is even, the median is the average of the two middle values.

It is conventional to denote the (unordered) data as $x_1, x_2, \ldots, x_n$ (where n denotes the number of observations in the set of data). When the data are arranged into ascending order, it is conventional to denote the values by

$$x(1), x(2), \ldots, x(n).$$

The observation $x(i)$ is said to have depth i, so the smallest observation has depth 1 and the largest observation has depth n. The median is conventionally found at depth $\frac{1}{2}(n+1)$.

The **range** (maximum value - minimum value) is not a good measure of spread, because it is greatly affected by outliers (extreme values in the data distribution).

An alternative, and more *robust* measure of spread is obtained from the **quartiles**. Quartiles are values which, along with the median, break the dataset into four subsets of (nearly) equal size. Approximately $\frac{1}{4}$ of the observations will lie below the lower quartile (LQ) and about $\frac{3}{4}$ of them will lie below the upper quartile (UQ). There are many different methods to compute quartiles but typically the LQ is found at depth $\frac{1}{4}(n+1)$ and the UQ is found at depth $\frac{3}{4}(n+1)$.

The **inter-quartile range** (IQR) is a robust measure of spread, defined by: **IQR = UQ - LQ**.

A five number summary is often used to describe data and this includes:

- The minimum value of the data
- The lower quartile (LQ)
- The median

- The upper quartile (UQ)
- The maximum value of the data

For a set of data, $x_1, \ldots, x_n$, (with number of observations n) we denote the **(arithmetic) mean** by:

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

The mean is greatly influenced by *outliers*. Removing large observations and re-calculating the mean can change the result. However, the median is likely to remain unchanged. Therefore, we say that the median is a robust measure of location.

In general, the mean is only a useful measure of the location of a dataset when the data are reasonably symmetric around their median (in which case, the mean and the median are approximately equal). This suggests that the median is the better measure of location to use by default.

Another measure of location, which can be useful when the data are measured on a discrete scale, is the **mode** or most common value. The mode, too, is insensitive to outliers.

Linked to the mean is a measure of spread known as the **variance** (σ^2).

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

The quantity σ^2 is referred to as the **population** variance with **population** mean μ. However, as noted earlier when we were discussing ideas of collecting data, it is often not possible to collect information on all members of the population and we collect data from a sample of the population. When using data from a **sample** of the whole **population**, the population variance σ^2 and population mean μ are unknown. Therefore, in practice, it is the sample variance (s^2) that is computed for this reason, using the sample mean $\bar{x}$:

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

The standard deviation ($s = \sqrt{s^2}$) is usually preferred to the variance since it is a measure of the spread which has the same units as the original data. The idea behind the standard deviation is to measure the typical distance (or deviance) between an individual observation and the mean. No measure based on the sum of the deviances is suitable, since that is always 0.

A measure could be based on the sum of the absolute values of the deviances, but for historical and mathematical reasons it is preferred to base the variance on the sum of the squared deviances.

Basically, the variance is the average of the squared deviances, but the divisor $(n-1)$ is used instead of n.

The factor $n-1$ is called the *degrees of freedom*. The sum of the deviations around the mean is known to be zero (i.e. try to show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$). If we know the value for the first $(n-1)$ deviations, the last one is known. There are only $n-1$ independent pieces of information in this estimate of variance. Dividing by $n-1$ accounts for the fact that our information on the variance (or standard deviation) is based on an estimate of the mean and not the true population mean. (We'll learn more about this in week 5 onwards). As the sample size n increases towards including all members of the population $s^2 \to \sigma^2$.

The variance and standard deviation are very sensitive to outliers (even more so than the mean). In general, the IQR is a much more robust measure of spread than the standard deviation, and is generally to be preferred to it.

When these quantities e.g. mean, variance, median, inter-quartile range etc are computed on a sample of data (rather than on data from all members of our population of interest) we refer to them as **statistics** or **summary statistics**, and we refer to the size of the data set n as the **sample size**. We'll return to this terminology and related formal definitions in week 3.

Task 1

It can be very helpful for manipulating expressions in probability models in later sections of the course to re-write the sample variance in alternative forms. Prove that the formula for the sample variance can be expressed in the following way:

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\}.$$

A selection of `R` commands to display simple data summaries are included in the table below.

Description	`R` commands
To produce the five number summary and mean	`summary()`
mean	`mean()`
variance	`var()`
standard deviation	`sd()`
range of dataset	`range()`
median	`median()`
quantiles, e.g. quartile 25%	`quantile()`
tabulating categorical data	`table()`
proportions for categorical data	`prop.table()`

For **categorical data**, proportions and percentages are used to summarise the data.

Plotting Data

Obviously using numerical summaries as above gives a flavour of some of the important aspects of a dataset but to visualise either the data itself, or indeed these numerical summaries, drawing a picture of the data is much more effective.

Rather than stare at a spreadsheet of numbers, appropriate graphs or diagrams:

- may throw up an unsuspected view of the data such as a pattern in it or what seem odd observations;
- allow one to give a subjective answer or judgement on the specific questions posed;
- provide one with a means of assessing at least parts of any probability model assumed;
- aid in the presentation of results and conclusions.

Some of the most common graphical methods you will already have met in the `R` **programming/Statistical computing** course and some of their basic uses related to the data types described here are given in the table below.

Plot	Basic Use
Dotplot	to summarise discrete data
Histogram	to summarise continuous data
Bar chart	to summarise categorical data
Stem and leaf diagram	to summarise continuous data
Boxplot	to summarise continuous data
Scatterplot	to compare two continuous variables

You will already have used a selection of these graphical methods in the `R` **programming/Statistical computing course**. However, we will consider their use in **Learning from Data** in more detail below and it is useful to make a few additional comments about a couple of these.

A **histogram** is a plot of the number (or proportion) of observations in a particular range (or *bin*) of the variable of interest against the central value of that *bin*. The choice of the width of each *bin* can often make large differences to the apparent *shape* of the sample so *bin-width* has to be carefully chosen.

A **boxplot** (also known as a box-and-whisker diagram or plot) is a line plot of the five-number summary of the sample made up by the

- median
- quartiles (lower and upper)
- and extremes (minimum and maximum).

(An aside: In `R`, *hinges* are used instead of the sample quartiles, although the values will be very similar. See `R` help files or documentation for more information on this.)

There is a mathematical method to highlight *outliers* (extreme/unusual observations in the data) and to determine *fences*, upper and lower thresholds from which to check for outliers.

After determining the lower and upper quartiles and the interquartile range as outlined above, then the *fences* are often determined using the following formula (although different methods can be used - see `R` documentation for more details):

$$\text{Lower Fence} = \text{LQ} - 1.5(\text{IQR})$$

$$\text{Upper Fence} = \text{UQ} + 1.5(\text{IQR}),$$

where LQ and UQ are the lower and upper quartiles, respectively. The Lower Fence is the *lower limit* and the Upper Fence is the *upper limit* of the data. Anything below the Lower Fence or above the Upper Fence can be considered an outlier and will be identified on a boxplot by a symbol, typically a circle in `R`.

This video illustrates approaches to summarising and displaying data providing explanation on specific features of summary statistics, histograms and the boxplot:

Video

Summarising and displaying data

Duration 10:41

Examples

The following examples will be used to illustrate some of the methods of displaying and summarising data. The datasets are very simple, and small here. However, they are useful for illustration purposes to fully explore the ideas.

Example 1

Absence from Work

In a recent year, the 15 employees in one department of a financial institution recorded the following numbers of days absence due to illness:

```
library(ggplot2)

absence <- data.frame("absences"=c(0,   0,   0,   0,   0,   0,   1,   1,
1,
                                   1,   1,   2,   4,   10,   39),
                      "ID"=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15))

ggplot(absence) +
    geom_point(aes(x = absences, y=ID))+
  theme(axis.title.y=element_blank(),axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  xlab("Number of days absence")
```

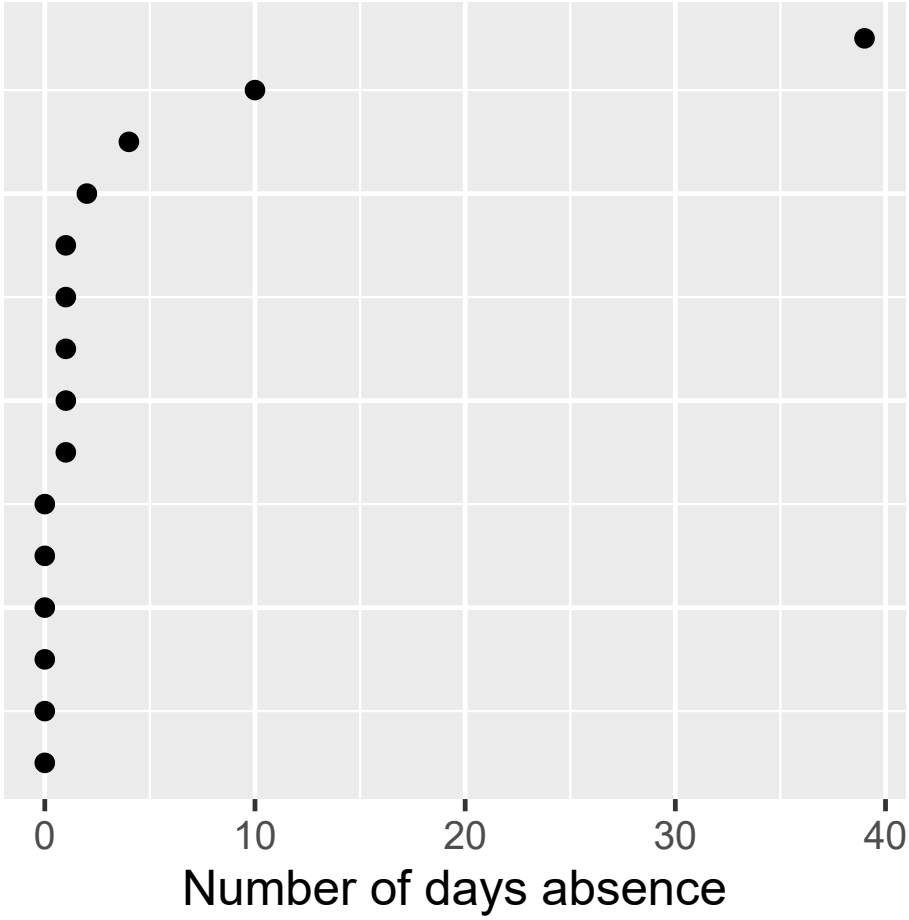


Figure 1

A dotplot is ideal for displaying discrete data.

The (horizontal) axis of the dotplot emphasises that the data have been measured on a numerical scale, whilst the separation between groups of points indicates that the data are discrete.

- **location**: Median = x(8) = 1 day, so typically these employees had just one days absence.
- **spread**: IQR = x(12) - x(4) = 2 - 0 = 2 days, so the spread of (the majority of) the data is small.
- **shape**: there is a pronounced right skew, with possible outliers at 10 and 39 days.

The sample mean for this dataset is $\bar{x} = 4$ days which is greater than all but two of the observations. The sample mean is greatly influenced by outliers.

Omitting the largest observation, x(15) = 39 and re-calculating the mean gives 1.5 days, which is much smaller than before.

On the other hand, re-calculating the median gives an unchanged value of 1 day.

Example 2

Cars

Here we have a dataset containing average prices for various types of cars in the United States in 2002.

Question: What is the *average car price* in the U.S.?

Sample: Each of a sample of 48 cars (of all different types i.e. small, sport, large, compact, etc.) had their average price recorded in 2002.

These continuous data are displayed in the stem and leaf plot, histogram and boxplot below.

The data [1] are contained in the RData object for week 2, which can be found at: RData

```
library(gridExtra)

boxplot<-ggplot(cars, aes(x="", y=Price))+
  geom_boxplot()+
  xlab("Price ($ Thousand)")+
  ylab("Price ($ Thousand)")

hist<-ggplot(cars, aes(x=Price))+
  geom_histogram(binwidth=5, fill="white", color="black")+
  xlab("Price ($ Thousand)")
```

```
grid.arrange(hist, boxplot, ncol=2)
```

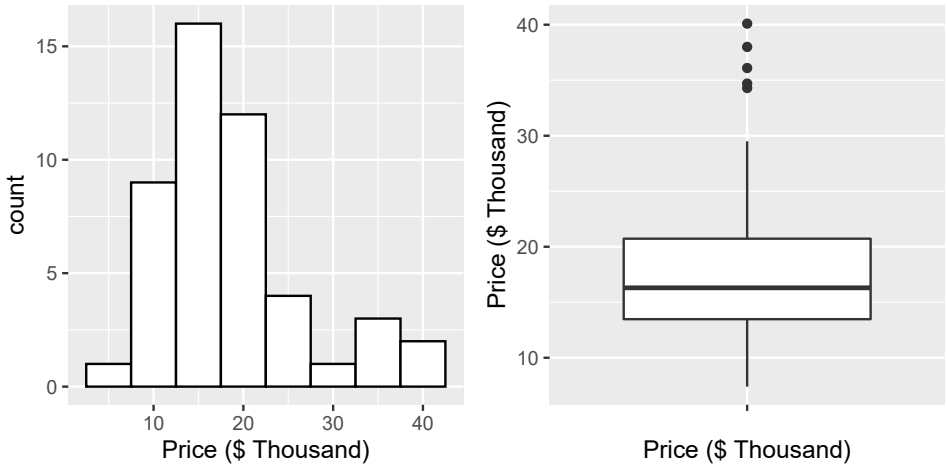


Figure 2

The boxplot on the right highlights that there are five large values in the dataset and in fact these have been flagged in `R` as outliers and displayed as circles.

Computing summary statistics in `R` (using the `R` command `summary`) for the data provides the following:

```
summary(cars$Price)
```

```
R Console
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
7.40   13.47   16.30   18.57   20.73   40.10
```

These summary statistics provide the minimum value (Min.), the lower quartile (1st Qu.), the median, the mean, the upper quartile (3rd Qu.) and the maximum value (Max) respectively. *Note, that the quartiles are computed in slightly different ways depending on the statistical package used and you will sometimes get slightly different answers when computing these by hand.*

Therefore, for the car price data we can say that:

- **location**: The median of the data is 16,300 with a mean of 18,570.
- **spread**: The IQR is 20.73-13.47 = 7.26 and the standard deviation can be calculated to be 7.82 (*this quantity is not provided automatically in the above summary but has been computed with the command* `sd` *in* `R`).

- **shape**: The distribution of the data is skewed to the right (illustrated in the stem and leaf plot and histogram) with five outliers at the right tail of the distribution.

Example 3

Occupational Stress

As part of a study of health and working conditions, a random sample of 180 Scottish dentists completed a general health questionnaire. One of the questions was *How often do you feel under stress at work?*

Responses to this question are tabulated below.

Stress level	Never	Occasionally	Most of the time	Some of the time	Total
Dentist Response	24	108	45	3	180
	13.3%	60%	25%	1.7%	100%

It is often helpful to add sample percentages to a frequency (%) table, as illustrated. These ordinal data may be displayed effectively on a bar chart. This display has just one axis, a frequency scale on the vertical axis (Never (N), Occassionally (O), Most (M), Some (S)). Since the individual observations are not numeric, there can be no horizontal axis. There is also a space between consecutive bars, for a similar reason.

```
stress <- data.frame("dentists"=c(13.3, 60, 25 ,1.7),
                     "group"=c("N", "O", "M", "S"))

ggplot(stress, aes(x=group, y=dentists))+
  geom_bar(stat="identity", fill="grey", color="black")
```

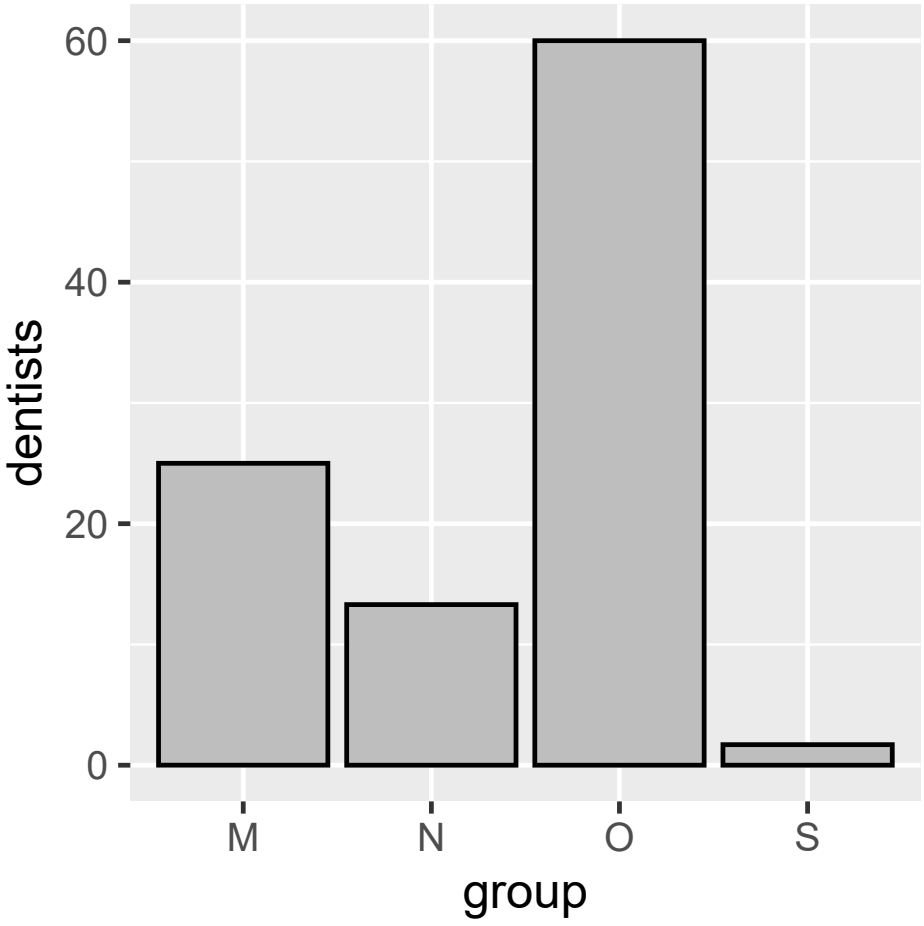


Figure 3

The data highlight that the majority of the dentists in the sample feel under stress occasionally at work.

The same summaries and displays can be used for nominal data.

Example 4

The Olympics

These data represent the men's gold 100m sprint times (in seconds) from the Olympic games in 1896 to 2016.

The data [2] are contained in the RData object for week 2, which can be found at: RData

```
gold <- olympic[olympic$Medal == "Gold", ]

ggplot(gold, aes(x=Year, y=Time))+
  geom_point()+
```

```
  xlab("Year")+
  ylab("Time (seconds)")
```

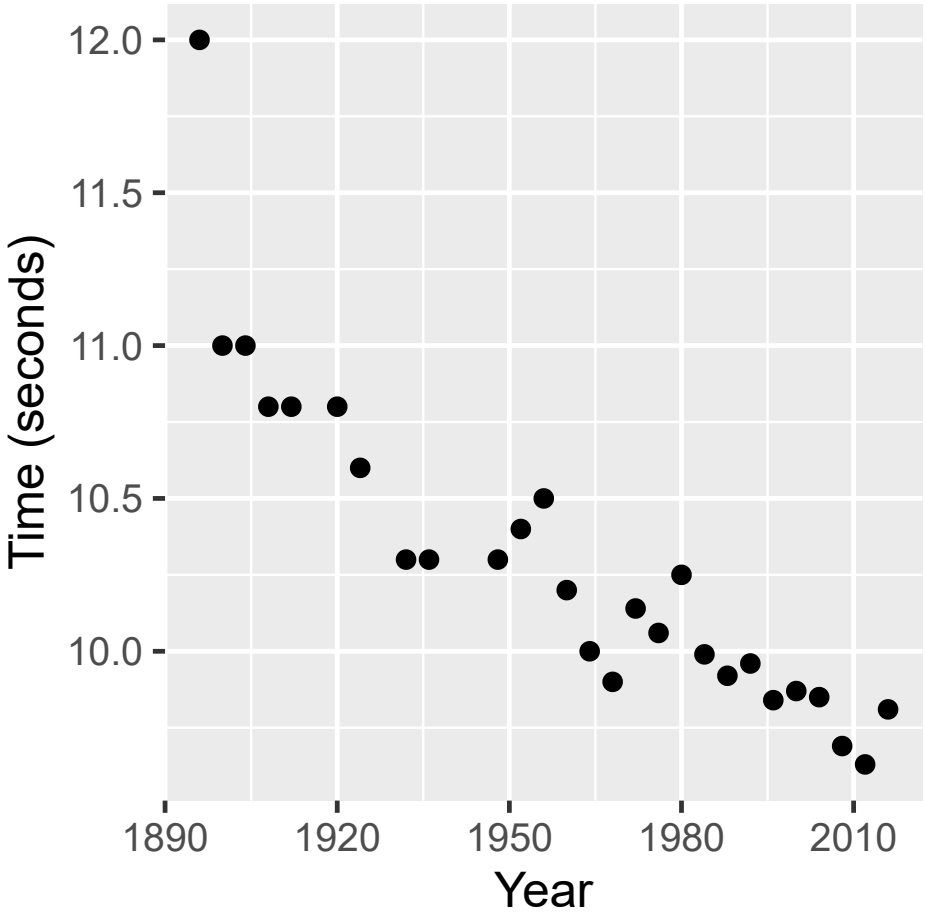


Figure 4

The figure above shows that the relationship between time and year is a decreasing one. At first glance, a straight line relationship looks sufficient. However, given some thought, would that make sense?

Our fastest time was in the 2012 London games where Bolt ran 9.63 seconds. If we were to predict a future olympic 100m sprint time, a straight line relationship would give a faster time than this. Then, the games after that a faster time again, and so on. Therefore, it should be clear that a linear relationship is not likely to be the most appropriate for these data. This example illustrates the an important skill needed in learning from data: to think about the context of the relationship between your variables.

A scatterplot can be used to explore the relationship between two continuous variables. Statistical analysis can be more straight forward in the situation where the relationship between two variables is linear. However, this cannot be assumed and the nature of the relationship should be checked using a scatterplot. Approaches for fitting such relationships will be the focus of the **Predictive Modelling** course.

Task 2

Emerging market countries often face high levels of inflation. The data below are inflation figures (%) for 19 emerging markets.

```
inflation <- data.frame("infla"=c(6.5, 14.0, 13.5, 18.0, 14.5,
9.0, 18.0,
                                  42.0, 7.5, 6.0, 25.0, 12.0,
52.0, 20.0,
                                  16.0, 15.0, 11.5, 2.5, 2.0))
```

Use appropriate plots and summary statistics to comment on the location, spread and shape of the distribution of the data.

Supplement 1

Additional material for the course is available at **Supplementary Material**. Specifically,

Data visualisation

The `R` plots displayed above are quite basic in their construction. See the associated material on data visualisation at **DataViz** for some examples on how to develop visualisations in more detail.

Histograms and boxplots

See the material at the link **HistBox** for an example of the difference changing the bin-width can make in a histogram, and examples of boxplots for data distributions that have one or more modes.

Features identified from a plot of the data

Example 5

Non-linear relationships, non-constant variance and transformations

Shot Putt Data

Here we have a dataset containing measurements from 28 females. It is of interest to explore the relationship between Shot Putt distance (in metres) and the personal best bench power clean weight lift (in kilograms).

The data [3] are contained in the RData object for week 2, which can be found at: RData

A scatterplot of the data is shown in the plot below on the left.

```
putt1<-ggplot(shotputt, aes(x=shot.putt, y=power.clean))+
  geom_point()+
  xlab("Shot Putt (m)")+
  ylab("Power Clean Weight Lift (kg)")

putt2<-ggplot(shotputt, aes(x=shot.putt, y=log(power.clean)))+
  geom_point()+
  xlab("Shot Putt (m)")+
  ylab("log(Power Clean Weight Lift (kg))")

grid.arrange(putt1, putt2, ncol=2)
```

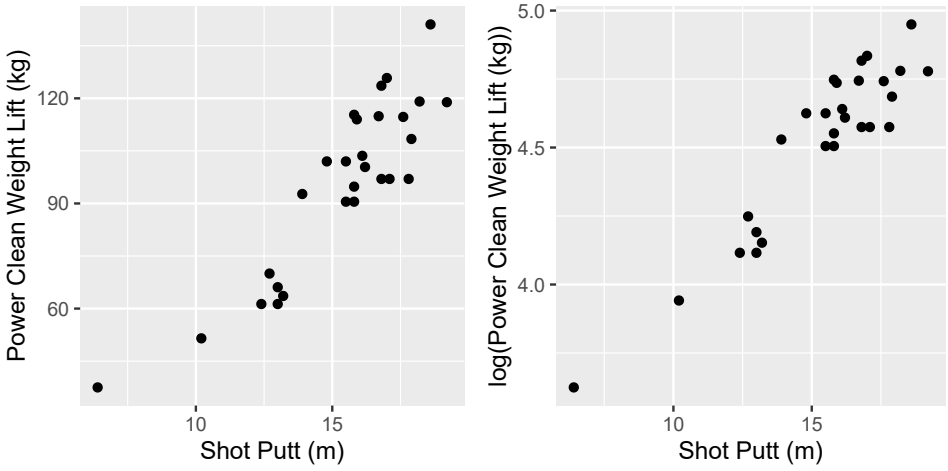


Figure 6

From the plot on the left, we can see that this is **not a linear** relationship, there is a gentle curve to the data. There is also some evidence in the plot that possibly the variance of the data increases as the values increase. As you would have noted in your Probability courses often probability distributions make assumptions about the mean/variance relationship. For example, the normal distribution is a symmetric distribution which assumes a constant variance term σ^2. However, in real data it is common to find examples where the variance increases as the mean increases, or where the distribution of a variable is highly skewed (skewness) or peaked (kurtosis). In such circumstances, it might be possible to find an appropriate **transformation** of the data (e.g. natural log, or square root) in order to transform the distribution to be more symmetric and to stabilise the variance. If a transformation is not appropriate then it might be that a different distribution should be considered or a robust form of inference may be required based on e.g. the median of the data.

In the shot putt example above, the relationship is clearly non-linear in the plot on the left, and the variability appears to change a little as the values on the x-axis increase (i.e. there is some evidence of the spread between values increasing over the x-axis (non-constant variance)). If we take a natural log transformation of the variable on the y-axis, the relationship becomes much more linear, as can be seen in the plot on the right (and there is some evidence that the variability appears to be a little more stable over the x-axis). These ideas are explored again in the second video for this week on page 15.

In addition to identifying the nature of relationships, and investigating if any possible transformation of the data may simplify the model building process, an effective plot of your data may also help you to **quality assure** your data e.g. identify mistakes in the data, or important features that should be taken into account in your statistical analysis.

Example 6

Extremes, Changepoints and Missing data

Sulphur Dioxide

Consider the example below for sulphur dioxide in the air that has been recorded at a particular monitoring station in Europe weekly for approximately 20 years.

The data are contained in the RData object for week 2, which can be found at: **RData**

```
s02$decimalyear <- s02$Years+s02$Weeks/52
#this creates a decimal year variable to have a time variable that is
continuous
ggplot(s02, aes(x=decimalyear, y=ln.SO2.))+
  geom_point()+
  xlab("Year")+
  ylab("log(SO2)")
```

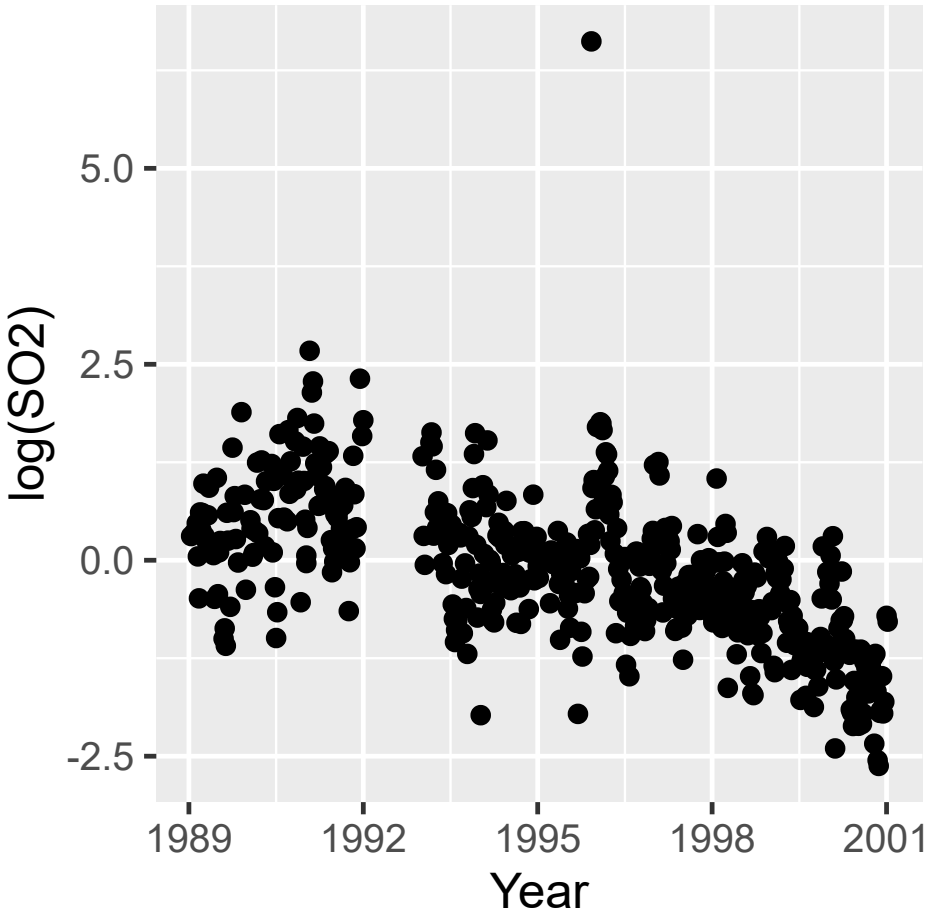


Figure 7

In this dataset it is the natural log of the SO_2 that is of interest (the natural log transformed the data to have closer to constant variability over time). The plot helps us to easily identify a period of **missing**

data around about 1992, and there is one **extreme** data point visible around 1996. At this point you would want to think about the reasons and viability of such data points.

- Is there a reason for the missing data?
- Are there data available that have not been included in the spreadsheet?
- Is the extreme value a 'real' data value?
- Is it truly an **extreme** or outlying data point?
- Is it a mistake or 'typo' in the spreadsheet?

It looks like there is a pattern in the data over time. It looks as though as time increases the overall $S0_2$ levels decrease. However,

- Is the pattern actually a non-linear decrease?
- What is happening in the early period of the data?

To answer these questions, you might need more information about the context of the data and the background to the environmental processes at that time. In this case there is actually a **changepoint** in the data around 1992. This could have been caused by a change in regulation or a change in the monitoring equipment. These are two good reasons to consider the dataset in two different sections i.e. pre and post 1992 since the data are arising from two slightly different underlying conditions. Is there also a changepoint around 1996?

What a plot doesn't show us.....

We've already discussed that we can explore the relationship between two continuous variables using a scatterplot. The scatterplot can be a useful tool to help us identify the nature of the relationship between variables e.g. linear, non-linear.

However, we need to be aware that identifying a relationship between two variables does not necessarily imply that a change in one variable causes a change in the other variable i.e. a relationship (or association) between two variables does not necessarily imply **causation**.

Here are two clearly ridiculous examples [4].

Example 7

Spurious relationships - association is not causation

Cheese and Bedsheets

The first example in the plot on the left below considers the per capita cheese consumption (in lbs) and number of people who died by becoming tangled in their bedsheets.

Divorce and Margarine

The second example in the plot on the right below is the divorce rate in Maine (per 1,000 people) and per capita consumption of margarine (in lbs).

```
cheese.bed <- data.frame("cheese" = c(29.8, 30.1, 30.5, 30.6, 31.3,
31.7, 32.6, 33.1, 32.7, 32.8),
                         "bedsheets" = c(327, 456, 509, 497, 596, 573,
661, 741, 809, 717))

cheese.bed.1<-ggplot(cheese.bed, aes(x=cheese, y=bedsheets))+
  geom_point()+
  xlab("cheese consumption(lbs)")+
  ylab("No. of people that died")

div.marg<-data.frame("divorce"=c(5,4.7,4.6,4.4,4.3,4.1,4.2,4.2,4.2,4.1),
                     "marg"=c(8.2, 7, 6.5, 5.3, 5.2, 4, 4.6, 4.5, 4.2,
3.7))

div.marg.1<-ggplot(div.marg, aes(x=divorce, y=marg))+
  geom_point()+
  xlab("Divorce rate (per 1,000 people)")+
  ylab("Margarine consumption(lbs)")
par(mfrow=c(1,1))

grid.arrange(cheese.bed.1, div.marg.1, ncol=2)
```

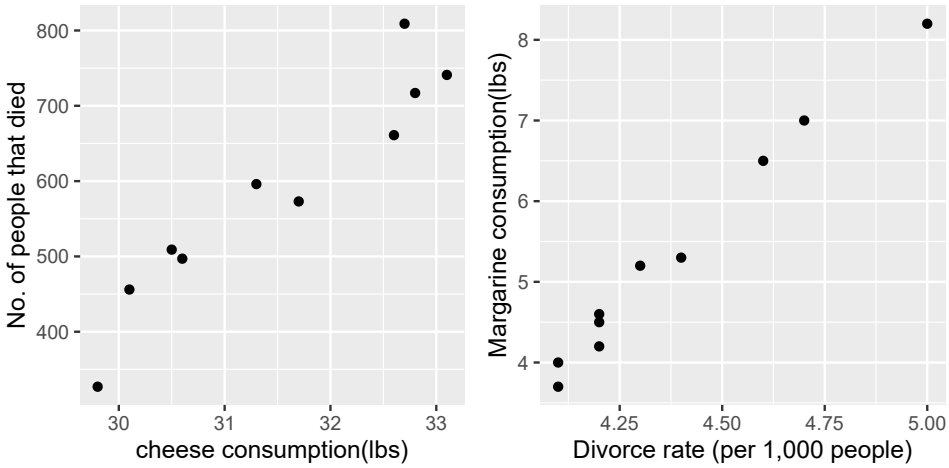


Figure 8

In both cases it looks from the plot like there is a relationship between the two variables. It clearly does not make sense in either case that this is a causal relationship between the variables, and hence you would not try to build a model to predict one variable from the other. These are great examples to illustrate that it's important to consider your question of interest before collating, and exploring your data. Just searching for relationships and patterns in data can come up with spurious links, which at best are misleading and at worst can be dangerous.

See the **Predictive Modelling** course for more information on this topic.

The context is important too.....

Within a dataset some data points may be more highly related than others.

When two pieces of information have been recorded on the same *individual*, the data are usually referred to as **paired data** since measurements on the same *individual* will be related. In this situation it can be of interest to compare the measurements and, when the data are continuous, this can be done using a scatterplot.

Example 8

Paired Data

Cats

This dataset is based on a dataset used for a study investigating whether or not there was a difference in the amount (ml) of water cats drink when the water is flowing or still. The data recorded is the average water drank by the cat over 4 days (for both flowing and still). We have 20 observations.

```
cat<-ggplot(cats, aes(x=still, y=flow))+
  geom_point()+
  xlab("Still Water, ml")+
  ylab("Flowing Water, ml")+
  geom_abline(intercept = 0, slope = 1)
      # this provides a line of equality y=x

cats$diffs <- cats$flow-cats$still

diff<-ggplot(cats, aes(x="", y=diffs))+
  geom_boxplot()+
  ylab("water drinking (flow-still)")+
```

```
  theme(axis.title.x=element_blank(),axis.text.x = element_blank(),
        axis.ticks.x = element_blank())+
  geom_hline(yintercept = 0, colour="red", lty=2)
      # this adds a horizontal line at 0

grid.arrange(cat, diff, ncol=2)
```

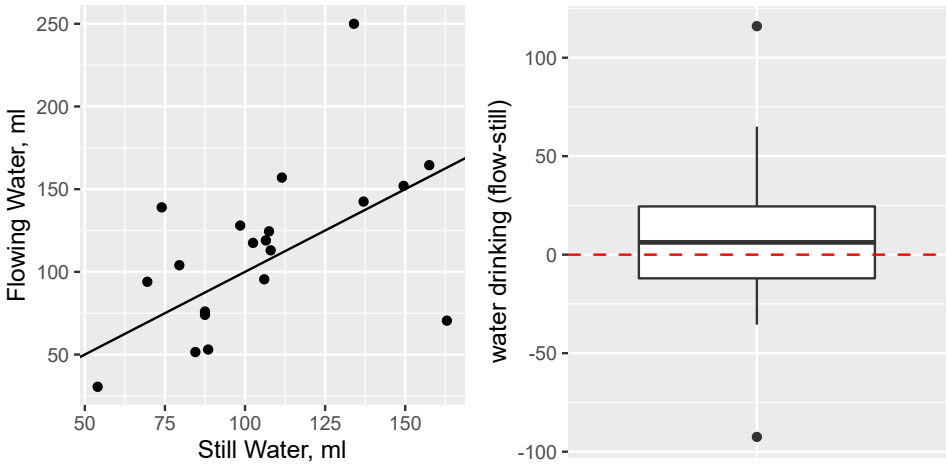


Figure 9

The data [5] are contained in the RData object for week 2, which can be found at: RData

As you can see from the plot on the left, slightly more data points lie above the line of equality. This indicates that these individual cats drank more water when the water was flowing compared to when it was still. Note here that the data for a single cat for flowing and still are not independent since both measurements were taken on the same cat.

Therefore, for **paired data** it's common to consider the differences between the two measurements. This reduces the problem to a one-sample problem of independent differences in the amount of water drunk in this context. Here the differences (flow-still) for each cat have been plotted in the boxplot on the right. This also indicates that these individual cats drank more water when the water was flowing compared to when it was still since the median lies above the horizontal line at 0.

There can also be different relationships present between data points that we should consider. For example data that are collected over time or at locations close together in space may be more closely related than data that are further apart in time and space. Such data are often referred to as time series or spatial data and it can be important to take account of this feature.

An example could be the carbon dioxide data recorded from the Mauna Loa volcano in Hawaii. This dataset is already available in `R`.

Example 9

Temporal correlation

Volcanos

Mauna Loa Volcano, Hawaii, Atmospheric CO_2 Concentration (parts per million). The data are collected over time monthly from 1959 to 1997 and are available directly in `R`.

```
data(co2)
ts.plot(co2)
```

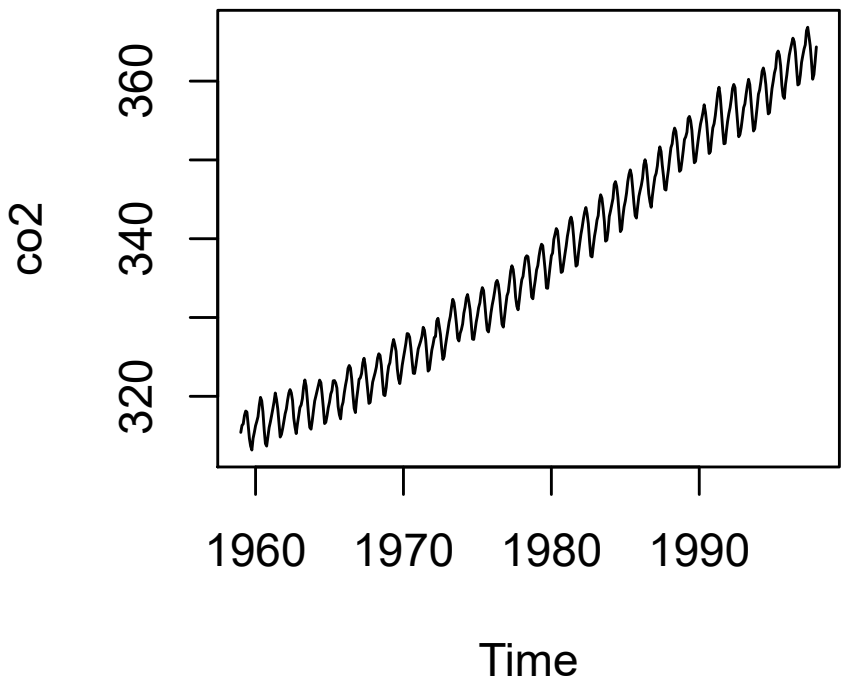


Figure 10

```
# The ts.plot() command is for data formatted as a time series object in
R;
# This automatically provides a plot of the data over time;
# A time index does not need to be created;
# The time index information is already contained in the R object.
```

It is clear in the plot that there is an increase in the $C0_2$ concentration over time and also that there are repeated patterns in the data. The repeated pattern is every 12 months, creating a **seasonal pattern** and illustrates a relationship between data points that are close together in time (**temporal correlation**). You will learn more about these type of data in the course on **Advanced Predictive Modelling**.

Examples of data that are related over space are:

- brain signals recorded at multiple electrodes;
- river water quality recorded at different monitoring stations on the same network;
- transport routes for main suppliers of companies.

This video provides a detailed discussion of the topic of summarising and describing data for an example on the water quality of a lake:

Video

Describing data

Duration 10:33

Supplement 2

Simpson's paradox: Consider the situation where we have data that have arisen from two groups, and we are interested in a relationship between two variables X_1 and X_2. Simpson's paradox occurs if we have the situation that when we look at the nature of the relationship by simply plotting X_1 vs X_2, we get a different relationship than if we investigate the nature of the relationships between X_1 and X_2 for each group separately.

Confounding variables: Additionally, a relationship between two variables can be masked by a third variable that has not been considered or a strong relationship between two variables can be the result of each variable having a strong relationship with a third, unexplored, variable.

See an example of Simpson's paradox here

Summary

Many of these data features will be introduced more fully with appropriate modelling approaches considered throughout different modules of this MSc programme. For now the main take home message is that it is very important to carefully consider your question of interest, the context of your data and the measurement process, and then to fully explore your data through appropriate data visualisations and summaries to: check and quality assure your data, explore patterns and relationships informally, and to inform the choice of probability model and the nature of relationship to be explored.

Task 3

The data for the task below are contained in the Rdata object for week 2, which can be found at: RData, under the data object, **dogs**.

Here we have a dataset [1] containing measurements of anxiety levels in labrador dogs before and after a treatment to lower stress levels when hearing an explosive in police dog training. Two measurements have been taken; the Salivary Cortisol and the Plasma Cortisol.
The lower the reading for both of these variables, the higher the stress levels of the dog. Is there a reduction in anxiety levels of the dogs after the stress-reducing treatment?

Data description:

- SC_pre = pre-test Salivary Cortisol mg/dL
- SC_post = post Salivary Cortisol mg/dL
- dSC_popr = SC_post - SC_pre
- PC_pre = pre-test Plasma Cortisol mg/dL
- PC_post = post test Plasma Cortisol mg/dL
- dPC_popr = PC_post - PC_pre

Produce appropriate summary statistics and plots for these data in order to explore the question of interest above?

Learning outcomes for week 2

By the end of week 2, you should be able to:

- describe data in terms of the location, spread and shape of its distribution;
- compute medians, quartiles, means and variances/standard deviations;
- provide examples of situations where it may be more appropriate to use a median compared to a mean;
- identify and use appropriate approaches to summarise and display data.
- interpret appropriate plots and summaries for different types of data;
- identify data features from summaries, plots and the context of the data to inform statistical modelling;

Review exercises, selected video solutions and written answers to all tasks/review exercises are provided overleaf.

Review exercises

Task 4

An experiment was carried out to compare the weight gains of rats fed on four different diets, distinguished by the amount of protein (low and high) and the source of the protein (beef or cereal). 40 rats were randomly allocated to the four diets and the table below shows the weight gain of each rat (grams). Compute the **median**, **lower** and **upper** quartiles for each diet, and decide informally from these numerical summaries whether one of the diets is better than the others for increasing the weight of the animals.

```
Beef (low)    90   76   90   64   86   51   72   90   95    78

        Beef (high)   73  102  118  104   81  107  100   87
117   111

        Cereal (low) 107   95   97   80   98   74   74   67
89    58

        Cereal (high) 98   74   56  111   95   88   82   77
86    92
```

Task 5

Prove that the formula for the sample variance can be expressed in the following way:

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}.$$

Task 6

Using the numerical summaries given below based on data from a sample, calculate the mean and standard deviation in each case.

Mean = 305/19 = 16.05

Variance = 1/18*(7712.5 - ((sqr(305)/19))) = 156.4692982

SD = 156.4692982 = 12.5

1. $n = 19$, $\sum_{i=1}^{n} x_i = 305.0$, $\sum_{i=1}^{n} x_i^2 = 7712.5$

2. $n = 30$, $\sum_{i=1}^{n} x_i = 10.976$, $\sum_{i=1}^{n} x_i^2 = 6.107$

3. $n = 40$, $\sum_{i=1}^{n} x_i = 312$, $\sum_{i=1}^{n} x_i^2 = 4850$

Task 7

The data in the plots overleaf are the supine systolic blood pressures (SBP) of a group of 15 patients with moderate essential hypertension (high blood pressure), immediately before and two hours after taking the drug captopril.

By referring to the plots, describe the apparent effect of the drug on systolic blood pressure in this group of patients. A line of equality is added to the plot on the left, and a line at zero added to the plot on the right.

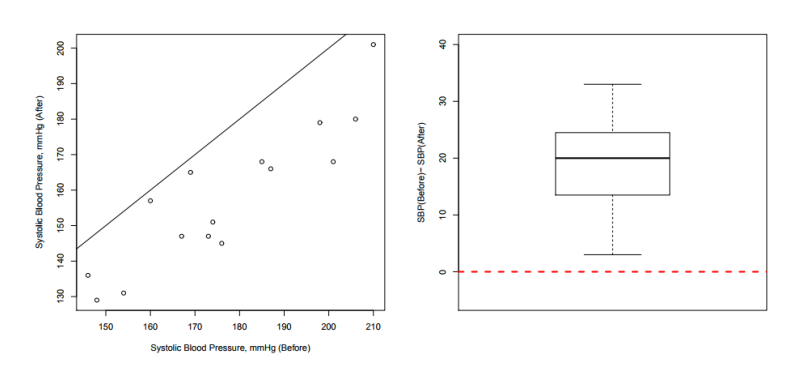


Figure 13

Answer 1

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\bar{x}+n\bar{x}^2\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\frac{\sum_{i=1}^{n}x_i}{n}+\frac{(\sum_{i=1}^{n}x_i)^2}{n}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2\frac{(\sum_{i=1}^{n}x_i)^2}{n}+\frac{(\sum_{i=1}^{n}x_i)^2}{n}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2\right\}
\end{aligned}
$$

Answer 2

Inflation in emerging market countries

```
stem(inflation$infla, scale=2)
```

R Console

```
The decimal point is 1 digit(s) to the right of the |

0 | 236789
1 | 224455688
2 | 05
3 |
4 | 2
5 | 2
```

```
summary(inflation$infla)
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.00    8.25   14.00   16.05   18.00   52.00
```

```
ggplot(inflation, aes(x="", y=infla))+
  geom_boxplot()+
  theme(axis.title.x=element_blank(),axis.text.x =
element_blank(),
        axis.ticks.x = element_blank())+
  ylab("% inflation")
```

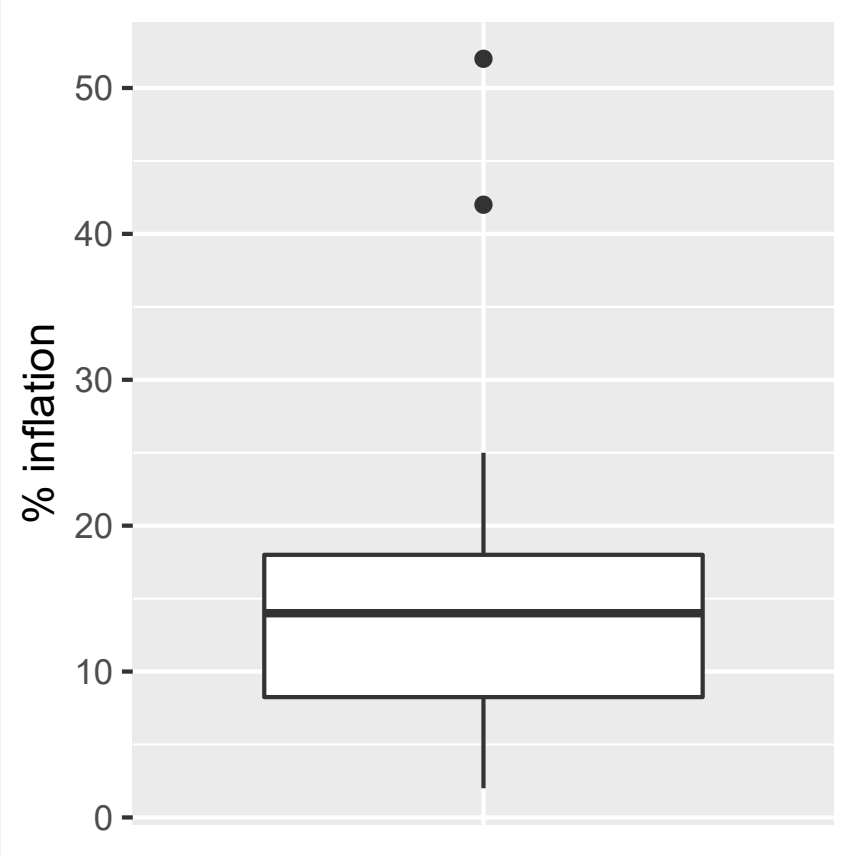


Figure 5

- location: The median of the data is 14 (%) and the mean is 16.05 (%). The difference between these indicates that the distribution of the data is not symmetric.
- spread: The IQR is 18-8.25 = 9.75
- shape: The distribution is right skewed with two `large' observations, outliers at 42 and 52 (%).

Answer 3

Dogs explosives:

```
dog1<-ggplot(dogs, aes(x=SC_pre, y=SC_post))+
  geom_point()+
  xlab("Salivary Cortisol Pre")+
  ylab("Salivary Cortisol Post")+
  geom_abline(intercept = 0, slope = 1)

dog2<-ggplot(dogs, aes(x="", y=dSC_popr))+
  geom_boxplot()+
  ylab("Difference in Salivary Cortisol (post - pre)")+
  theme(axis.title.x=element_blank(),axis.text.x =
element_blank(),
        axis.ticks.x = element_blank())

grid.arrange(dog1, dog2, ncol=2)
```

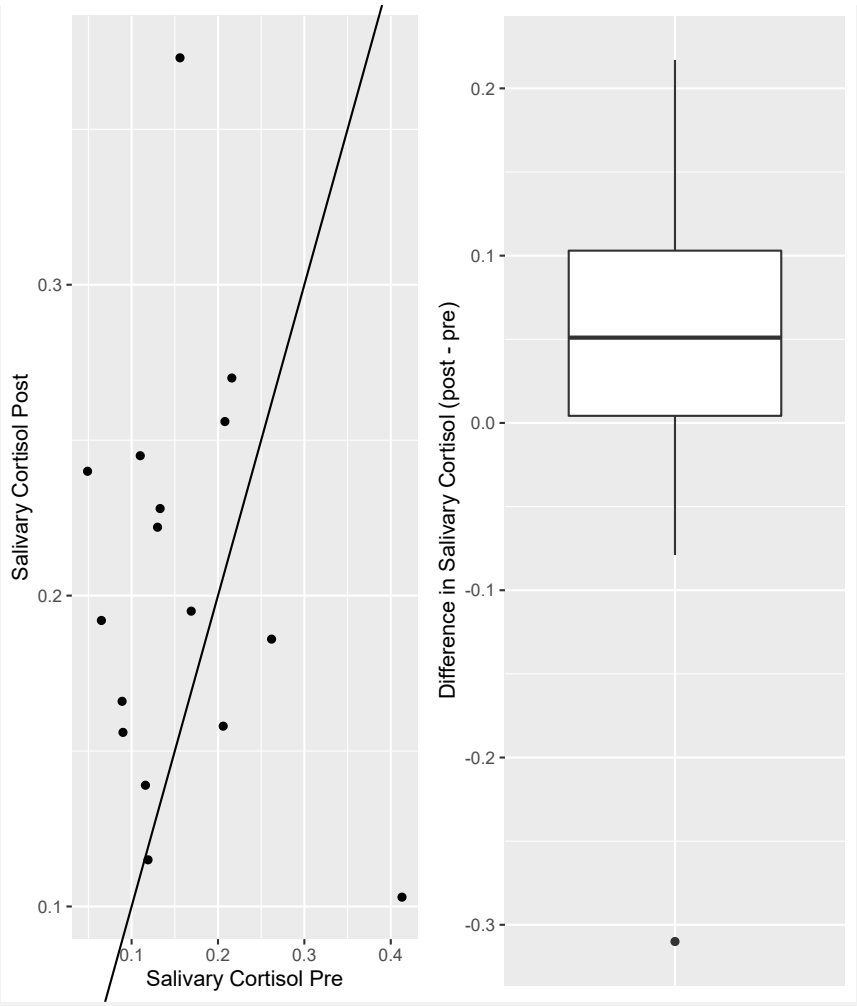


Figure 11

```
dog3<-ggplot(dogs, aes(x=PC_pre, y=PC_post))+
  geom_point()+
  xlab("Plasma Cortisol Pre")+
  ylab("Plasma Cortisol Post")+
  geom_abline(intercept = 0, slope = 1)

dog4<-ggplot(dogs, aes(x="", y=dPC_popr))+
  geom_boxplot()+
  ylab("Difference in Plasma Cortisol (post - pre)")+
  theme(axis.title.x=element_blank(),axis.text.x =
element_blank(),
        axis.ticks.x = element_blank())

grid.arrange(dog3, dog4, ncol=2)
```

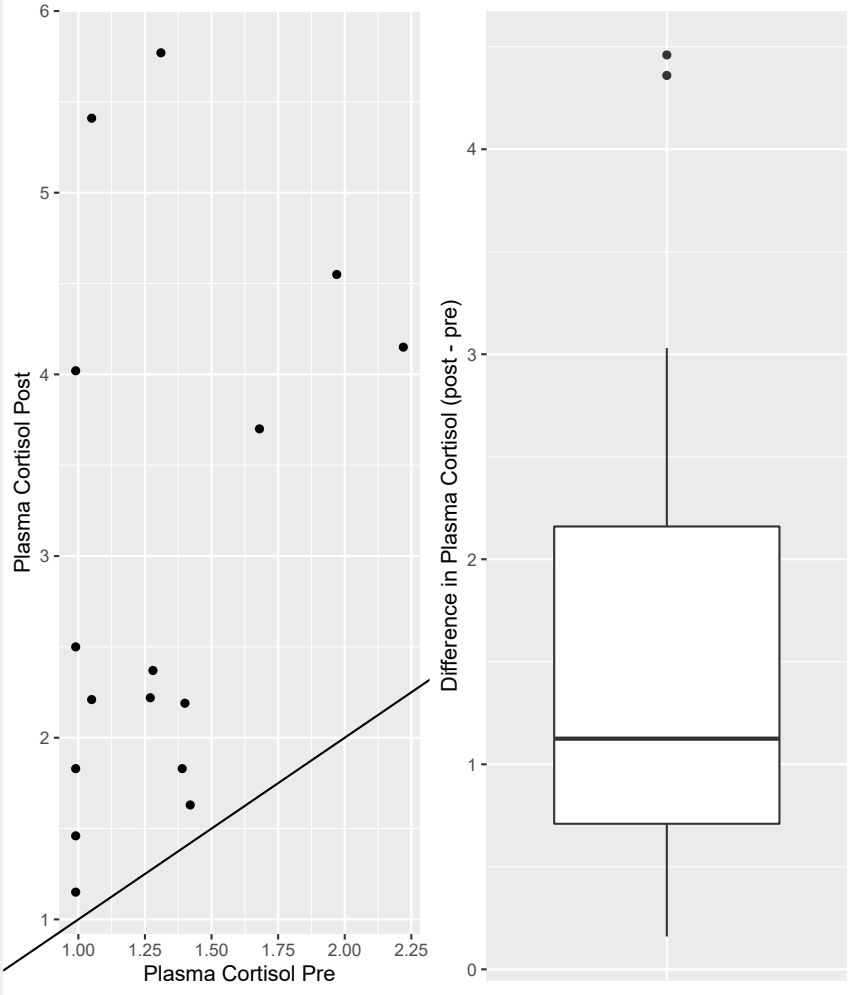


Figure 12

```
summary(dogs$SC_pre)
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0490  0.1050  0.1315  0.1582  0.2065  0.4130
```

```
sd(dogs$SC_pre)
```

R Console

```
[1] 0.08983186
```

```
summary(dogs$SC_post)
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1030  0.1575  0.1935  0.2028  0.2412  0.3730
```

```
sd(dogs$SC_post)
```

R Console

```
[1] 0.06744529
```

```
summary(dogs$PC_pre)
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.990   0.990   1.275   1.312   1.405   2.220
```

```
sd(dogs$PC_pre)
```

R Console

```
[1] 0.3720971
```

```
summary(dogs$PC_post)
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.150   1.830   2.295   2.937   4.053   5.770
```

```
sd(dogs$PC_post)
```

R Console

```
[1] 1.451825
```

After the informal impression of the data, it appears that there is a reduction in stress levels of dogs after the treatment. To see this we look at the scatterplot of Salivary Cortisol before and after the treatment and the boxplot of the difference in anxiety (a common approach since the data are paired measurements i.e. 2 from each dog). The scatterplot shows that the majority of values after treatment are higher than before, and the boxplot shows that most of the differences lie above zero - indicating a positive effect of the treatment. Similarly, when looking at the plots for the Plasma Cortisol, this is further emphasised. Lastly, we can look at the summary statistics and standard deviations which also indicate that there could be a difference. More formal statistical inference is needed to form a conclusion for the population of labrador dogs, and we'll start to introduce these ideas in week 3.

Answer 4

Numerical summaries:

```
Beef (low)    51   64   72    76   78   86   90   90   90   95

        Beef (high)   73   81   87   100  102  104  107  111
117  118

        Cereal (low)  58   67   74    74   80   89   95   97
98  107

        Cereal (high) 56   74   77    82   86   88   92   95
98  111
```

Diet	Median	LQ	UQ
B/lo	82	70	90
B/hi	103	85.5	112.5
C/lo	84.5	72.25	97.25
C/hi	87	76.25	95.75

(Note: the summary statistics here have been computed by hand.)

Summary statistics if computed in `R` (see the note on page 5 about differences here in the way `R` computes the quartiles)

Diet	Median	LQ	UQ
B/lo	82	73	90
B/hi	103	90.25	110
C/lo	84.5	74	96.5
C/hi	87	78.25	94.25

The high-protein beef diet seems most successful in increasing the average weight of the rats. The distribution of results seems to have similar spread for all four groups.

Answer 5

It is useful to practice the different formulations for the sample variance since one formulation may be easier to compute than another depending on the form that the data, or summaries of the data are provided.

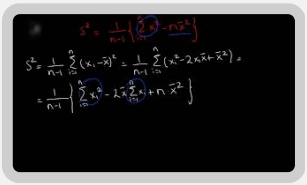
Later in the course, we will use such expressions and manipulations as part of model fitting and inference and so being able to move between these expressions will help to simplify results.

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\bar{x}+n\bar{x}^2\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2n\frac{\sum_{i=1}^{n}x_i}{n}\bar{x}+n\bar{x}^2\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\right\} \\
&= \frac{1}{n-1}\left\{\sum x_i^2-n\bar{x}^2\right\}
\end{aligned}
$$

Video

Video model answers

Duration 3:49



Answer 6

mean and variance:

Scenario 1

mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{305}{19} = 16.05$

variance,

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right\} = \frac{1}{18}\left\{7712.5 - \frac{(305.0)^2}{19}\right\} = 156.469$$

standard deviation: $\sqrt{156.469} = 12.51$

Scenario 2

mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{10.976}{30} = 0.3659$

variance:

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right\} = \frac{1}{29}\left\{6.107 - \frac{(10.976)^2}{30}\right\} = 0.07211$$

standard deviation: $\sqrt{0.07211} = 0.2685$

Scenario 3

mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{312}{40} = 7.8$

variance:

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right\} = \frac{1}{39}\left\{4850 - \frac{(312)^2}{40}\right\} = 61.959$$

standard deviation: $\sqrt{61.959} = 7.87$

Answer 7

Since all of the points are below the line of equality in the scatterplot and all of the differences in the boxplot are positive then it appears that captopril lowers SBP in this group of patients.

Footnotes

1. The original data are available from: http://vincentarelbundock.github.io/Rdatasets/datasets.html. ↩

2. The original data are available from: https://www.olympic.org/athletics/100m-men. ↩

3. The original data are available from: http://users.stat.ufl.edu/~winner/datasets.html) ↩

4. These examples come from the following website: http://www.tylervigen.com/spurious-correlations, which contains links to the data sources ↩

5. The original data for this example are based on data that are available at: http://users.stat.ufl.edu/~winner/datasets.html) ↩

6. The original data for this example are available at: http://users.stat.ufl.edu/~winner/datasets.html) ↩