Introductory Statistics with Excel

Andrew Parnell

University College Dublin

Class 6 - Linear regression and control charts

Learning outcomes

In this class we will cover

- Correlation and bivariate data
- Linear regression models
- Simple control charts

Relationships between variables

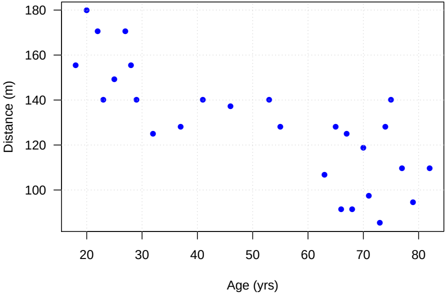
It is very rare for variables to change in **isolation**. Most often, multiple variables change **simultaneously**, and we are interested in understanding their **relationship** and possibly **predicting** one or more variables from the others.

In this class we will learn some methods for **quantifying** the relationships between quantitative and qualitative variables.

We have already met some graphical methods for performing some of these tasks, including **scatter plots**, **box plots** and **histograms**.

Relationships between quantitative variables

Example: some data measure the age of 30 drivers together with the maximum distance at which they could read a signpost. A scatter plot of the data is shown below



This is an example of a **negative statistical relationship**; in general as age goes up the maximum readability distance goes down

Looking for patterns in scatter plots

Some questions to ask:

- What is the **average pattern**? Is it a linear relationship (a straight line) or is it curved?
- What is the direction of the pattern? Is it a **positive** or **negative** relationship?
- How much do the individual points **vary** from the average pattern?
- Are there any **unusual** data points (outliers)?

Linear and non-linear relationships

Not all relationships between quantitative variables are linear. Consider the following examples:

- The relationship between height and age
- The relationship between car age and selling price

There are two analyses we often perform with pairs of quantitative variables. The first is to calculate the strength of the linear relationship between the two variables, known as the **correlation**. The second is to try and predict one of the variables from the other, known as **linear regression**

Correlation

Calculating the correlation coefficient

The Pearson **correlation coefficient** (often written as r) measures the strength and direction of a **linear** relationship by a number between 1 and -1.

- A number close to $r = 1$ or $r = -1$ indicates that the relationship (as seen in a scatter plot) is almost perfectly linear
- A number close to $r = 0$ indicates that there is no linear relationship between the two variables
- If the value is positive ($r > 0$) the two variables tend to increase together. If it is negative ($r < 0$) then as one variable increases the other decreases

Calculating the correlation coefficient

If we label one variable as x and the other as y so that the observations are in pairs $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$, the correlation coefficient can be calculated as:

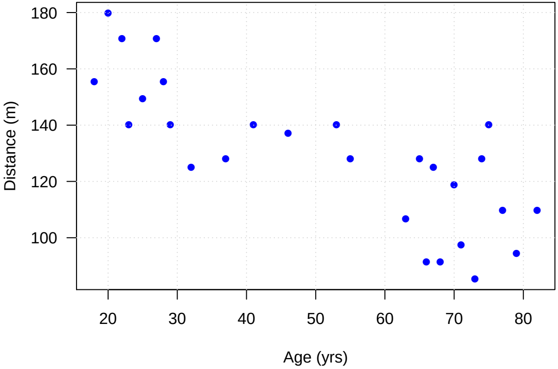
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, and s_x and s_y are the standard deviations of the x and y variables respectively.

Note that the terms inside the brackets are the **standardised data points** which we met in lecture 3 when we met the normal distribution

Some examples

The age/distance data set we met at the start of this lecture has a correlation value of $r = -0.795$



Some important notes about correlation

- If you look at the formula for calculating the correlation coefficient you can see how a positive or negative value of r can be obtained. For example, if the standardised x-value is above the x mean and standardised y-value is above the y mean then you are multiplying two positive values together. Similarly, if they are both below their respective means you are multiple negative values together
- Note that just because two variables are correlated, it does not mean that one is caused by the other. There may be a **confounding** variable
- There are different ways to calculate the r value, and different versions of it for qualitative or ordinal variables

Linear regression

Linear regression

Suppose now we want to predict one variable from another. As in Class 2, we call the variable we want to predict the **response variable** and the remaining variable the **explanatory variable**. This is a common problem in science:

- We might want to predict whether someone has a disease based on the results of a medical test that they take
- We might want to predict how big a product's sales will be based on the amount spent on advertising it

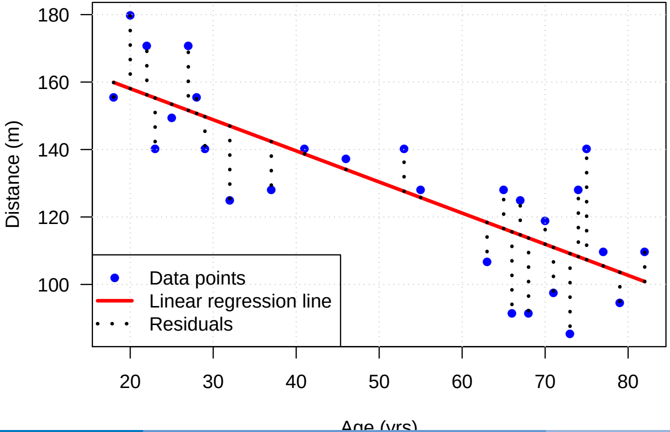
If we are happy to assume a linear relationship we create a **statistical model**:

$$\text{response} = a + b \times \text{explanatory} + \text{residual}$$

where a is the **intercept** and b is the **slope**, both to be estimated from the data

Fitting the regression equation

One way of estimating the values of the slope and intercept is via **least squares** where we try to minimise the squared differences between the estimated values from the model (i.e. ignoring the residual component) and the true values of y.



Some notes about linear regression

- The vertical distances between the data points and the fitted line are called **residuals** as they represent the leftover variation in the data points beyond that of the fitted linear regression model
- Often the square of the correlation coefficient (r^2) is reported along with the slope and the intercept. This is often used to illustrate the **proportion of variation 'explained' by the linear regression**.
- We should be very careful about **extrapolating** our line beyond the x range of the data.
- The name **regression** comes from Francis Galton who used it to derive the relationships between the heights of parents and children. He originally called the method 'reversion'

Measuring relationships between categorical variables

We have already met methods for presenting categorical variables with two categories, such as **cross-tabulations** and **bar charts**.

Example: deaths on the Titanic

Class	1st	2nd	3rd	Crew
Survived	203	118	178	212
Died	122	167	528	673
Total	325	285	706	885

What can we say about the relationship between the class of passenger and the survival rate?

Risk and relative risk

The first thing we can compute is the **risk** (or **baseline risk**) of being in an undesirable category:

$$\text{Risk} = \frac{\text{Number in category}}{\text{Total number in group}}$$

For example, the risk of a 1st class passenger dying is $\frac{122}{325} \approx 0.375$ Next, we could calculate the **relative risk** which is the ratio of risks in two different groups:

$$\text{Relative Risk} = \frac{\text{Risk in category 1}}{\text{Risk in category 2}}$$

For example, the relative risk of a 3rd class passenger compared to a first class passenger is $\frac{0.748}{0.375} \approx 2$. Thus you were approximately **twice as likely to die** if you were in 3rd class compared to 1st!

Odds and Odds ratios

The **odds** of an event compare the chance that it does not happen with the chance that it does. We are used to seeing odds written as 'a to b', for example in horse racing you might see a horse given as 3 to 1 to win the race.

We can transform odds into probabilities by calculating $\frac{b}{a+b}$ in the odds. Thus a 3 to 1 horse has probability 0.25 of winning the race. Some notes:

- Bookmakers usually publish the odds **against** an event rather than the events **for** an event. This can make things quite confusing
- The odds given by bookmakers, when transformed into probabilities, don't usually add to 1. The difference is their profit margin!

The **odds ratio** is calculated by dividing the odds in two different categories.

A summary slide on risk and odds

Suppose you have the following table:

	Response Variable		
Explanatory variable	Response 1	Response 2	Total
Category of interest	A_1	A_2	T_A
Baseline category	B_1	B_2	T_B

Now:

- The **risk** (of response 1) for the category of interest is A_1/T_A
- The **odds** (of response 1 to response 2) is A_1 to A_2
- The **relative risk** is $\frac{A_1/T_A}{B_1/T_B}$
- The **odds ratio** is $\frac{A_1/A_2}{B_1/B_2}$

Common mistakes with risk

A recent study of 6000 women claims that drinking 5 cups of coffee per day reduces women's risk of breast cancer by 57%[a].

[a] http://www.dailymail.co.uk/health/article-1385763/Five-cups-coffee-day-protect-breast-cancer.html

Should you start drinking a bucket of coffee a day?

- We should always get the sample size and the **baseline risk**. It might be that such a small proportion of people suffer from breast cancer that a 57% change doesn't make much difference
- Remember all the other ways that statistical studies can fool us: confounding variables, study limitations, etc
- Revisit lecture 1: disaster in the skies

Control charts

Quality Control

- Variability is inevitable in any manufacturing process.
  e.g. Two screws made by the same machine and from the same raw materials will show some variation in length and diameter.
- As in the previous statistical models we have met, variability may be partitioned into **residual variation** and **assignable cause**.
- Residual variation is that which is left over, i.e. beyond simple explanation.
- Assignable causes could be worn machine parts, poorly trained operators or the production environment.
- If the variability in a production process is only due to residual variation then the process is said to be **in control**.

Control Charts

- Control charts consist of a **central line** and upper and lower **control limits**.
- The observed values are plotted and if any fall outside the control limits, the process is said to be **out of control**.
- Control limits and the central line can be estimated from data collected when the process is known to be in control.

Important

Control limits are appropriate only for analysing past data (data used in their calculation). They can be used for future data only when the process is in control and/or when the limits have been modified.

Example

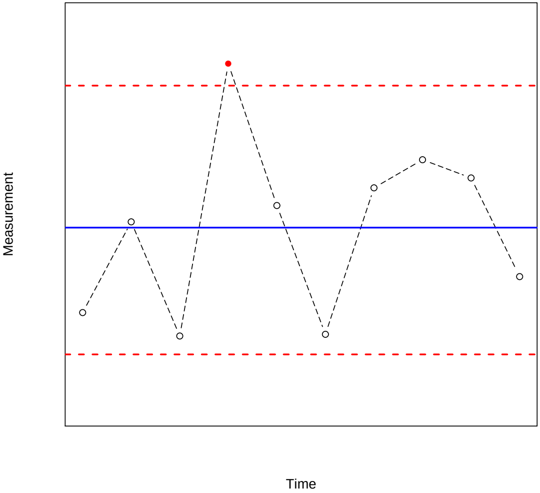


Chart Setup

- When control charts are first applied the process may not be in control.
- Even if it is in control, the behaviour of the process will not yet be fully understood.
- Data needs to be collected from the process and any assignable causes of variation removed. This is the **chart setup** stage.
- After the process has been operating in control for some time, a large amount of data is available and control charts may be used to monitor the process.
- We will focus on the process monitoring stage here.

Benefits of Control Charts

- **Improve productivity:** Scrap and rework are reduced by monitoring a process with control charts.
- **Defect prevention:** An in control process, producing goods consistently correctly, reduces the amount of defective produce.
- **Prevent unnecessary process adjustments:** If processes are only checked periodically, operators often overreact to random error. This is not the case if the process is monitored continuously.
- **Diagnostic information:** An experienced engineer can often use a control chart (in tandem with other information) to diagnose a process problem.
- **Process capability:** A control chart monitors important process parameters and their stability over time. This information can be used to assess if the process is fit for purpose or not.

Control Chart for Means: $\bar{x}$-chart

- Rather than one measurement at each time point, take a number of measurements, n, at each time point and calculate their sample mean $\bar{x}$.
- By the central limit theorem the sample mean should vary about the true process mean μ with standard deviation σ, and should fall within the interval $\mu \pm 3\frac{\sigma}{\sqrt{n}}$ with a very high probability (> 0.99) like a wide confidence interval
- The limits of this interval will be the upper control limit (UCL) and lower control limit (LCL).

Control Chart for Means: $\bar{x}$-chart

- As we almost never know the true value of μ or σ these must be estimated from some collected data.
- Let the time points be denoted by $i = 1, \ldots, k$. Recall that n measurements are taken at each time point.
- The mean at each time point is denoted by $\bar{x}_i = \sum_j x_{ij}/n$.
- The center line is estimated by the grand mean:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

Control Chart for Means: $\bar{x}$-chart

- It is standard in control charts to use R_i, the range at each time point i, to estimate σ.

  $$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{\sum_{i=1}^{k} R_i/k}{d_2}$$

  where d_2 is a constant depending on n. (Values of d_2 for differing n are at the back of these slides)

- Thus:

  $$3\frac{\sigma}{\sqrt{n}} \approx 3\frac{\hat{\sigma}}{\sqrt{n}} = \frac{3(\bar{R}/d_2)}{\sqrt{n}} = \frac{3}{d_2\sqrt{n}}\bar{R} = A_2\bar{R}$$

  where $A_2 = \frac{3}{d_2\sqrt{n}}$

Control Chart for Means: $\bar{x}$-chart

Summary

- Center line : $\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$
- UCL: $\bar{\bar{x}} + A_2\bar{R}$
- LCL: $\bar{\bar{x}} - A_2\bar{R}$

where

- $k =$ number of samples, each of size n.
- $\bar{x}_i =$ mean of i^{th} sample.
- $R_i =$ range of i^{th} sample.
- $\bar{R} = \sum_{i=1}^{k} R_i/k$

Interpretation

Out of control: One or more of the sample points fall outside the control limits. This indicates there may be some problem with the production process and should be investigated. **Trends** in the process mean may also indicate the presence of some assignable variation.

In control: All sample points fall within the control limits. There may still be problems with the process but it is most likely better to leave the process alone than to look for problems that may not exist.

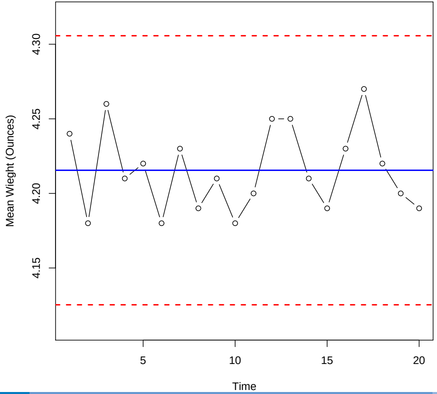
Example:

The following are the sample means and ranges of 20 samples, each consisting of the weight of 5 castings in ounces. Construct an $\bar{x}$-chart from these data.

$\bar{x}$	R	$\bar{x}$	R
4.24	0.09	4.20	0.21
4.18	0.12	4.25	0.20
4.26	0.14	4.25	0.17
4.21	0.24	4.21	0.07
4.22	0.15	4.19	0.16
4.18	0.28	4.23	0.16
4.23	0.06	4.27	0.19
4.19	0.15	4.22	0.20
4.21	0.09	4.20	0.12
4.18	0.15	4.19	0.16

Example

- Center line: $\bar{\bar{x}} = 4.216$
- UCL: $\bar{\bar{x}} + A_2\bar{R} = 4.216 + (0.58)(0.156) = 4.306$
- LCL: $\bar{\bar{x}} - A_2\bar{R} = 4.216 - (0.58)(0.156) = 4.125$



Control Chart for Process Variation: R-Chart

- Quality control engineers also need to monitor the variability of a process.
- Large variability in a process will produce non-uniform, inferior products.
- Changes in the variability of a process can also be indicative of assignable variation.
- The variability of a process can be monitored using an R-chart.

Control Chart for Process Variation: R-Chart

Summary

- Center line : $\bar{R}$
- UCL: $D_4\bar{R}$
- LCL: $D_3\bar{R}$

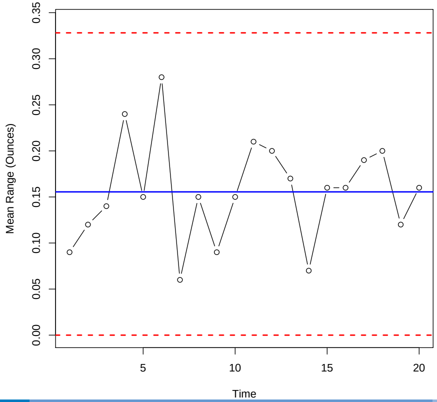
where D_3 and D_4 are constant related to n, the size of the individual samples. (Values of D_3 and D_4 for differing n are at the back of these slides)

Control Chart for Process Variation: R-Chart

- R-charts can be interpreted in the same way as $\bar{x}$ charts. If any points fall outside the control limits the process is said to be out of control.
- Trends in the variability of a process may also indicate the presence of some assignable variation.
    - Increasing/decreasing variability.
    - Large numbers of consecutive points above/below the center line.
- Only if the variability is in control (R-chart) should the quality measurement ($\bar{x}$-chart) be checked. Variabilty which is out of control may mask problems with the quality measurement.

Example

- Center line: $\bar{R} = 0.156$
- UCL: $D_4\bar{R} = (2.11)(0.156) = 0.328$
- LCL: $D_3\bar{R} = (0)(0.156) = 0$



Summary of class 6

- Correlation and linear regression very useful for quantifying the **relationship** between two **quantitative** variables
- Relative and absolute risk/odds useful for quantifying **categorical** relationships
- $\bar{x}$-bar control charts useful for quantifying **mean** problems, and R-charts useful for quantifying **variability** problems

Summary of course

- You now have had reasonable exposure to a good chunk of 20th century statistical methodology
- If you want to update yourself for the 21st century you need to:
  1. Learn R or Python
  2. Learn about Bayesian statistics (and hierarchical modelling)
  3. Take a course in machine learning

Control Chart Constants

n	A2	A3	d2	D3	D4	B3	B4
2	1.88	2.66	1.13	0.00	3.27	0.00	3.27
3	1.02	1.95	1.69	0.00	2.57	0.00	2.57
4	0.73	1.63	2.06	0.00	2.28	0.00	2.27
5	0.58	1.43	2.33	0.00	2.11	0.00	2.09
6	0.48	1.29	2.53	0.00	2.00	0.03	1.97
7	0.42	1.18	2.70	0.08	1.92	0.12	1.88
8	0.37	1.10	2.85	0.14	1.86	0.18	1.81
9	0.34	1.03	2.97	0.18	1.82	0.24	1.76
10	0.31	0.97	3.08	0.22	1.78	0.28	1.72
11	0.28	0.93	3.17	0.26	1.74	0.32	1.68
12	0.27	0.89	3.26	0.28	1.72	0.35	1.65
13	0.25	0.85	3.34	0.31	1.69	0.38	1.62
14	0.23	0.82	3.41	0.33	1.67	0.41	1.59
15	0.22	0.79	3.47	0.35	1.65	0.43	1.57
16	0.21	0.76	3.53	0.36	1.64	0.45	1.55
17	0.20	0.74	3.59	0.38	1.62	0.47	1.53
18	0.19	0.72	3.64	0.39	1.61	0.48	1.52
19	0.19	0.70	3.69	0.40	1.60	0.50	1.50
20	0.18	0.68	3.73	0.41	1.58	0.51	1.49
21	0.17	0.66	3.78	0.42	1.57	0.52	1.48
22	0.17	0.65	3.82	0.43	1.57	0.53	1.47
23	0.16	0.63	3.86	0.44	1.56	0.55	1.46
24	0.16	0.62	3.90	0.45	1.55	0.56	1.45
25	0.15	0.61	3.93	0.46	1.54	0.56	1.44