Inference

Last week we considered sets of experimental data from two different groups performing the same experiment: timing a ball falling a fixed distance to estimate the acceleration due to gravity.

Attempt	1	2	3	4	5	6
Group 1 Time (s)	1.11	1.18	1.02	1.09	1.10	1.13
Group 2 Time (s)	1.12	1.10	1.11	1.10	1.11	1.10

Review: What are the best estimate and its error, for each of these data sets?

Group 1's result should be reported as:

Group 2's result should be reported as:

Now we want to consider the following critical question: how can we use these values to make scientific claims, or draw specific conclusions? Typically, we might want to know if our results match other people's, or if they match some theoretical value – that is,

1. do these experiments agree with one another?; or
2. does either of these experiments agree with some separate prediction?

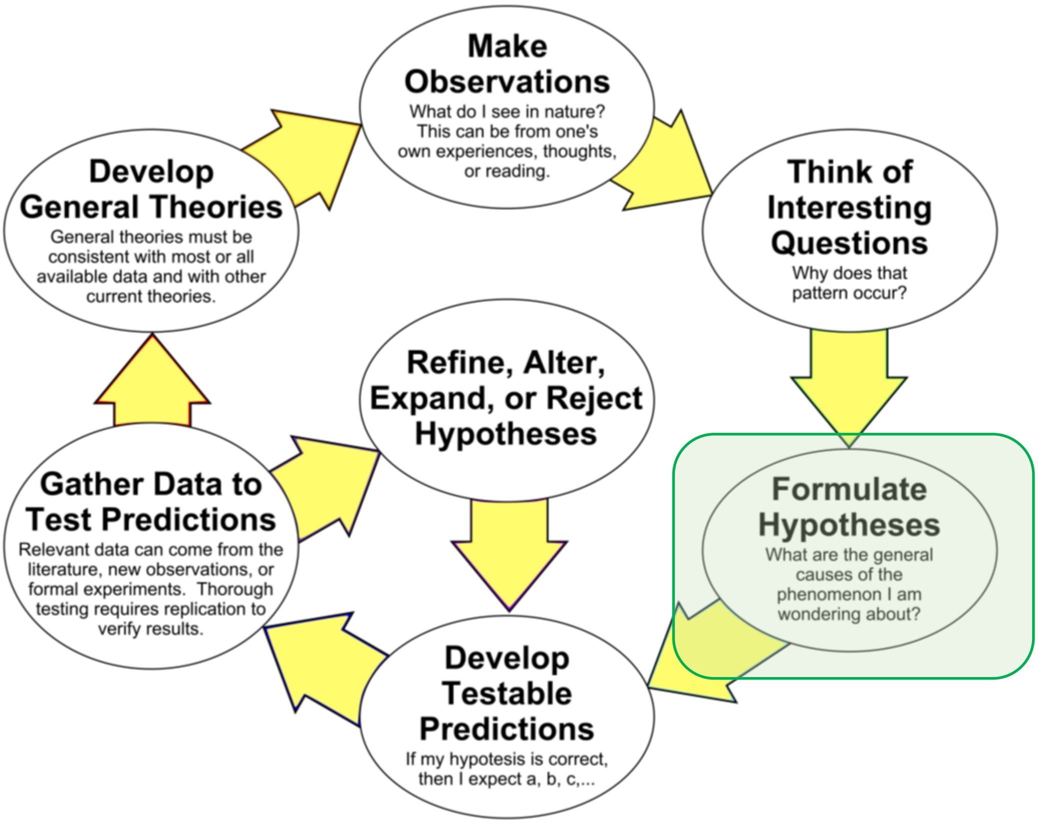
This process of drawing conclusions from our data, based on a statistical analysis, is known as *(statistical) inference*. We will look at the how statistical inference relates to the scientific method, and how we can answer the two above questions for the falling-ball data.

Hypotheses

In statistics, we frame these questions as a *hypothesis*, which we formulate as one of the steps in the scientific method. We hypothesise a particular outcome

- *The experimental results agree with each other*
- *The experimental results agree with the theoretical prediction*

and we use statistical methods to decide whether to *accept* or *reject the hypothesis*. We use the best estimate, its error, and the number of experiments to make this decision.



Usually, we hypothesise the outcome that we might "reasonably" expect. This hypothesis is a statement, summing up the situation we want to test. For example

- *Hypothesis: The experimental results match the theory*
- *Hypothesis: The two experimental results match one another*
- *Hypothesis: There is nothing unusual going on*

In modern science we usually choose our hypothesis to be the result that is most consistent with our current level of knowledge. This may be considered a consequence of Occam's (or Ockham's) razor, a principle attributed to the 14th century logician and Franciscan friar William of Ockham. In a context useful for science, it states something like,

"when you have two competing theories that make exactly the same predictions, the simpler one is the better."

In other words, we choose our hypothesis to be the result that would be least surprising. Such a hypothesis is given a special, technical name: it is called the ***Null Hypothesis.*** The word "null" suggests nothing unusual is going on – no alarms and no surprises! The symbol H_0 is often used as shorthand for the Null Hypothesis, for a given experiment.

What would we give as the Null Hypothesis, if we considered the following questions?

- *I wonder if left-handed people more intelligent than right-handed people?*

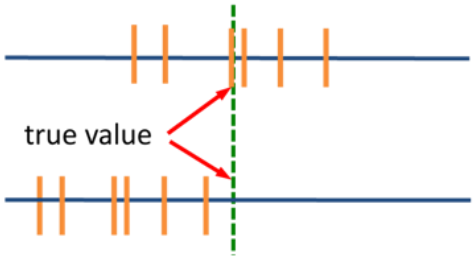
- *I wonder if my dogs are more popular than cats?*

Confidence and Significance

But what precisely does it mean to say that "*my experimental results match the theory*"? I might be able to use theory to predict a single value, but when I do my experiment, I might get a different value each time (depending on my experimental set-up), because of the random errors that creep into my results. How can I compare this set of values to any single value?

We make the following *ansatz*[1]:

- We assume there is a *true value* that my experiment is trying to determine (μ_T), which is the value I would get if I could do the experiment without any random errors. Unless we have a theory that has already been tested, we might not know what this true value μ_T is, but we hope that by doing our experiment well, the values we obtain lie close to it.
- If we do many experiments and the random errors are truly random, the experimental values we get should be distributed randomly about the true value μ_T (with some above and some below). If that were the case, the true value μ_T would be the average you would get from doing many, many experiments. This is why the average of our data is the best estimate (μ). The error for the best esimate (SE) gives us a sense of how far away the true value μ_T might be from the best estimate μ.
- After repeating my experiment, I obtain the best estimate for my experimental outcome (μ) and an error for the best esimate (SE).
- The problem with this is that we could be been unlucky, and got consistently wrong values, e.g. the random error in each experiment causes the value to be too low (see figure to the right). The more experiments we do, the more unlikely this outcome is, but we can never rule it out entirely as a possibility.



To get around this, we talk about ***confidence*** in the outcome. We express how likely it is that the true value μ_T matches our experiment, by expressing our ***percentage confidence that the μ_T lies within a particular range of values about μ.***

[1] An ansatz is an initial educated guess or assumption that allows you to solve a (usually mathematical) problem. Borrowed from German, "beginning, initial statement, estimate"

For example, for the case of the ball drop experiment we would say that:

The true value lies in the range 1.1 ± 0.4 seconds (95% CI).

Here, "95% CI" means a *95% confidence interval*. We are saying that we have 95% confidence that the true value lies in that range, in the sense that

- If we were to repeat the experiment many times …
- … and calculate the 95% confidence interval (as described in a few pages' time)
- … the true value would lie within this interval 95% of the time.

If the true value lies within the confidence interval 95% of the time, it must lie outside of this interval 5% of the time. This gives us an equivalent, alternative way of expressing this idea: we say that:

The true value lies in the range 1.1 ± 0.4 seconds, to **a** ***statistical significance of 5% (or 0.05)***

Constructing Confidence Intervals

We can construct our confidence in a couple of ways

1. Choose the interval size, and work out the level of confidence associated with it; or
2. Choose a level of confidence, and work out the corresponding interval size.

It is more common to take the second approach, choosing a high level of confidence. In modern scientific work, it is common to choose a confidence of 95%, 99%, or 99.9%.

What are the corresponding levels of statistical significance, for the 95%, 99%, or 99.9% CI?

To work out whether my data match an expected value (e.g. from theory), we use the *two-sided Student t-test*:

1. For my experiment, which I repeated n times
   - I calculate the best estimate μ and standard error $SE = s/\sqrt{n}$
   - I determine the expected value μ_0
2. I work out the t-factor t_f required for my chosen CI (from a table or software)
   - This will depend on the level of confidence (eg 95%) and number of experiments n
3. I check whether the expected value μ_0 lies in the range between $\mu \pm t_f \times SE$
   - If it does, my expected value agrees with my experiment (to 95% confidence) ☺
   - If it doesn't, my expected value disagrees (to a 5% level of statistical significance) ☹

Science History Sidebar!

Student's t-test and t-distribution was a product of a scientific paper published by a statistician working at the Guinness brewery in Ireland. At the time, some employers preferred their staff publish under pseudonyms – of which "Student" was William Sealy Gosset's!

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Let's check whether our fall-time experimental results match the theoretical estimate $\mu_0 = 1.1$**, to 95% confidence** (establishing confidence intervals)**.**

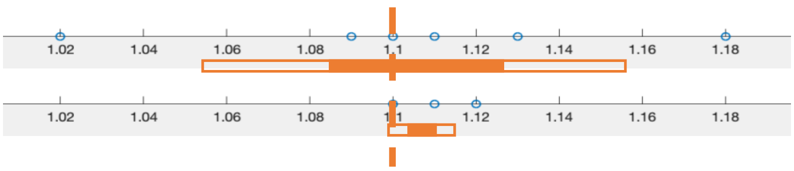
Group 1: $n = 6,\ \mu = 1.105,\quad s = 0.052,\quad SE = 0.021$

Group 2: $n = 6,\ \mu = 1.1067,\ s = 0.0082, SE = 0.0033$

Let's check whether our fall-time experimental results match the theoretical estimate $\mu_0 = 1.1$, to 95% confidence (using t-statistic)**.**

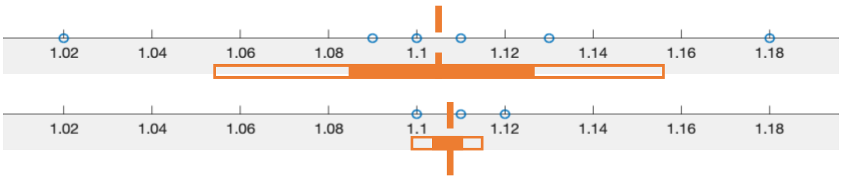
Group 1: $n = 6,\ \mu = 1.105, \quad s = 0.052, \quad SE = 0.021$

Group 2: $n = 6,\ \mu = 1.1067,\ s = 0.0082, SE = 0.0033$



To work out whether two sets of data match each other, we use the *Welch t-test*[2] (named after Welch):

1. For each experiment, repeated several times
   - I calculate the best estimate μ and standard error $SE = s/\sqrt{n}$
   - These values $(n_1,\ s_1,\ \mu_1,\ SE_1)$ will be different to $(n_2,\ s_2,\ \mu_2,\ SE_2)$
2. I calculate the t-statistic $t = \frac{|\mu_1-\mu_2|}{\sqrt{(SE_1)^2+(SE_2)^2}}$
3. I calculate the degrees of freedom $df = \frac{\left((SE_1)^2+(SE_2)^2\right)^2}{\frac{(SE_1)^4}{n_1-1}+\frac{(SE_2)^4}{n_2-1}}$ (!!)
4. I work out up the t-factor t_f required for the chosen CI and *df* (from a table or software)
5. If $t < t_f$, the results agree; if $t > t_f$, they do not agree (within statistical significance)



Let's check whether our two sets of experimental results agree, to 95% confidence.

Group 1: $n = 6,\ \mu = 1.105,\quad s = 0.052,\quad SE = 0.021$

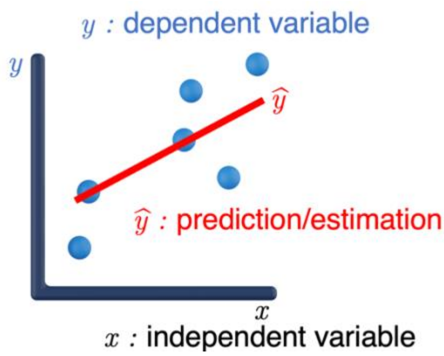
Group 2: $n = 6,\ \mu = 1.1067,\ s = 0.0082, SE = 0.0033$

[2] Equivalent to ANOVA for two means, for those who've done some stats before

Hypothesis testing: Conclusions

- We can't say absolutely whether …
- We use statistics to infer …
- For hypotheses based on results from repeated experiments
  - comparing one experiment with theory – use …
  - comparing two experiments – use …
- Are they the only hypothesis tests?

Functions



Correlation and regression are methods that relate two sets of numerical data. Correlation gave us a sense of whether the variation in one quantity could be explained, or predicted, through the variation of the other; then we looked at regression as a way of finding a line of best fit. This line of best fit provides a model for the relationship between the two sets of data – a *function* that estimates the response $\hat{y}$ for a given predictor x. We assumed the function relating these variables was linear: that is, that it took the form $y = mx + c$. However, there are an infinite variety of possible relationships that could arise between the variables: in practice, there is a core set of common functions that we need to be familiar with, in a scientific context. Now we will turn our attention to what functions are, looking at a range of important functions that you should know, and some important skills that we need when working with functions.

Examples of functions

What are some situations where linear functions arise?

Any type of function that is not a linear function (not of the form $y = mx + c$) is called a *non-linear function.*

What are some examples of nonlinear functions, and situations where they arise?

Important knowledge and skills when exploring relationships between data sets include

- familiarity with the more common nonlinear functions
  - what they look like when graphed, situations where they arise
- fitting nonlinear equations to data
- finding y for a given x, or vice versa
- finding important properties of the functions, such as maxima, minima, slopes, etc.

We will develop some of these skills in the coming weeks, learning about the relevant mathematical techniques and software.

What, exactly, *is* a function?

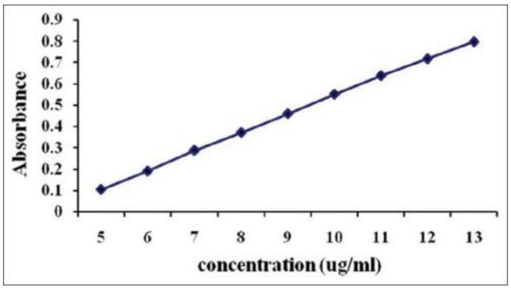
Functions express a relationship between two[3]* variables. They describe

- how a quantity evolves over time
- how a response depends on a predictor
- how an effect depends on a cause

Examples include:

Functions can be expressed in various ways

- As a graph



- In words
  - *Your final mark for this course is determine by …*
  - *To calculate the tax payable on your taxable income, …*
- As a formula

$$y = f(x) = sin(ax + b)$$

- As a procedure
  - *For natural number n, f(n) is the number of factors that divide into n.*

[3] functions can exist between 3 or more variables, but we won't consider them in this course

Until the end of the 19th century, the formula approach was the fundamental definition of a function – it had to be something that could be expressed in terms of a formula. Nowadays we take a more liberal approach, thinking of functions as "black boxes": for allowed input values x, the function f returns corresponding output values of $y = f(x)$. This allows us to include procedural definitions of functions, where we can't find a nice neat formula.

Function input and output

Function input (*abscissa*, *predictor*, *independent* variable):

- It's the **independent** variable, because we are free to choose its value.
- Generally called x if there is no preferred symbol
- Drawn on the **horizontal** axis
- The allowed input values are called the functions **domain** (where it lives)

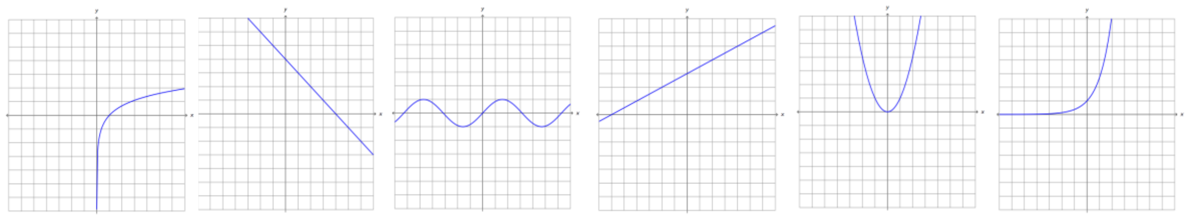
Function output (*ordinate*, *response*, *dependent* variable):

- It's the **dependent** variable: its value depends on the choice of x.
- Generally called y if there is no preferred symbol
- Drawn on the *vertical* axis
- The allowed output values are called the functions **range** (where it can reach)

If I have two sets of data, which do I choose which should be the independent variable x, and which should be the dependent variable y?

- Any variable you have control over: x
- Any timestamp variable[4]: x
- Any variable showing a response: y
- Any variable with repeated values: y

Drawing the graph of a function



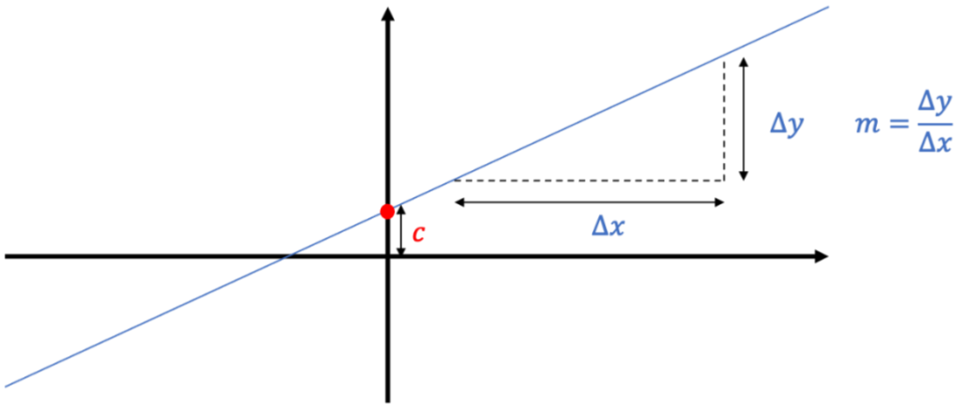
- The simplest approach is to remember the shapes of standard functions
- If you don't know/remember: use sample values of x to calculate f(x)

[4] Not a measurement of the time taken for something to occur, but a measurement of when (time and/or date) something has occurred, like the road toll data. We don't have control over time, but when a timestamp is one of the two data sets, we usually want to consider how the response is varying with time.

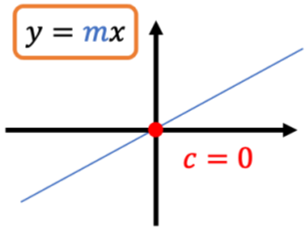
Linear functions

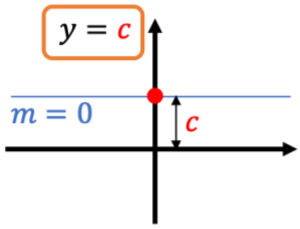
The straight line has the formula $y = mx + c$, where

- m is the slope
- c is the y-intercept



Special cases:



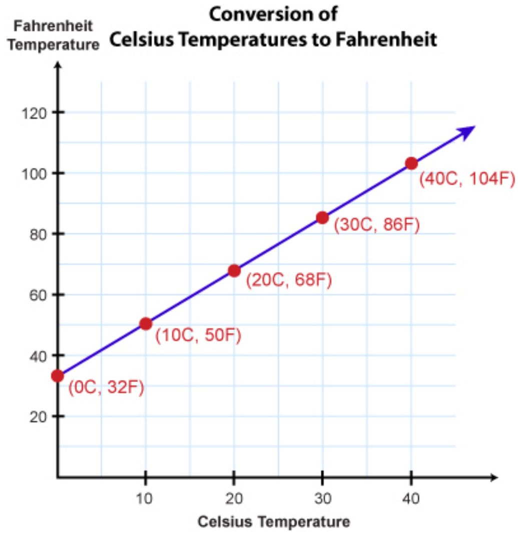


How do you plot $y = \frac{x}{2} + 3$?

Calculations involving $y = mx + c$

Example: converting between Celsius and Fahrenheit: $F = \frac{9}{5}C + 32$

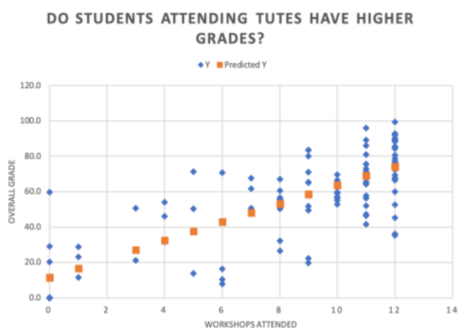
- Currently it's 30°C in Brisbane – what's that for my American friends?
- My recipe says to heat my oven to 350°F – what's that in Celsius?
- What would *Fahrenheit 451* be called in Celsius?



Example: using the model obtained from a linear regression

- What mark do students get attending 8 workshops?
- How many workshops would I need to attend to get a '5'?

$$M = 5.22\,W + 11.5\ (\pm 17)$$



Example: using a concentration-absorbance regression curve

- What concentration is a sample with Absorbance 0.40?

$$A = 0.0875\,W - 0.3375$$