## Introduction to Probability and Statistics

This course will give you a flavour of:

- Data analysis
- Probability theory
- Random variables
- Statistical modeling

This course will let you become more familiar with mathematical equations and notation. It will also help you start to think of certain problems in a more mathematical way.
We will cover a range of topics:

- Data types
- Basic statistics
- Basic graphs
- Probability rules
- Conditional probability
- Discrete and continuous random variables
- Basic calculus

Lectures:

| | | |
|---|---|---|
| 1600-1800 | Wednesday 5th October | Taviton (16) 432 |
| 1600-1800 | Thursday 6th October | Chadwick 2.18 |

Lecturer: Dean Markwick, dean.markwick.15@ucl.ac.uk

These lecture notes and the problem sheet have kindly been provided by Sam Livingstone and updated by James Pitkin.

# 1 Data Analysis

A simple goal of any Statistics (data analysis, analytics, data science) task is to use some available data (observations, sample) to learn something about the world. We may be trying to answer specific questions, or just describe and understand what is happening. We might need to build a complex mathematical model, or we may need to simply organise a data set and plot a graph. It is *vital* to remember that **the object of interest is the things we learn, not the methods that were used**. Don't fall in love with your analysis! Do what is necessary to get some interesting answers, and focus on these.

## 1.1 Types of Data

Data come in many shapes and sizes, particularly in the modern world. Some examples are:

(1) Time in between earthquakes in a region (any positive real number)

(2) Tweets with the word "danger" in a single day (text/words)

(3) Patient satisfaction survey scores (number between 1 and 10, *ordinal*)

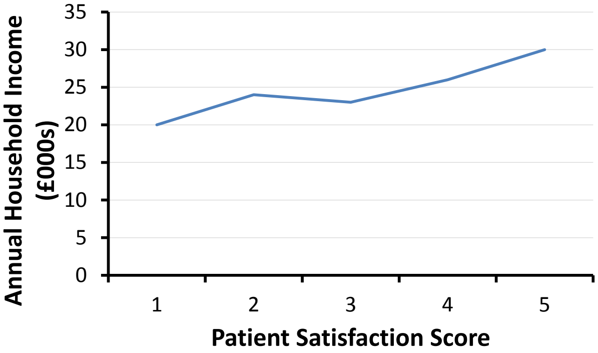
(4) Number of drowning deaths in a region per year, by nationality (1=European, 2=North American..., *categorical*)

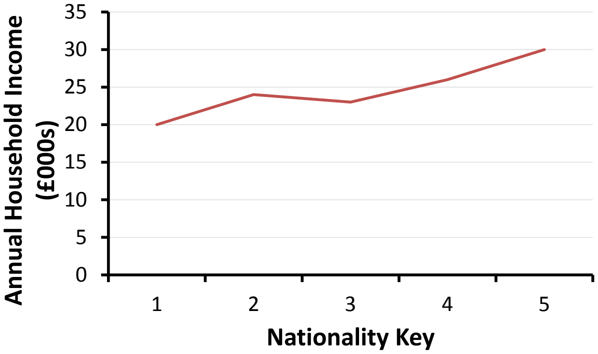
(5) Outcomes of a disease treatment (1=Recovery, 0=Mortality, *binary*)

(6) Outcome of a coin toss (1=Heads, 0=Tails, *binary*)

(7) Head shots of convicted criminals (images)

Notice two things: In (3) and (4) the data you receive will look the same, but they will mean different things. One of these graphs makes sense but the other does not:





In (5) and (6) the data will look the same, *and* the same techniques can be used to analyse them (so understanding coin flips can be useful!).

As in the graphs, we can sometimes use data to understand whether two things are **associated** with each other. We may then try to quantify this association (inference) or use one thing to make a guess for the other (prediction).

### 1.2 Summarising Data

In this section we work with two data sets, showing the number of earthquakes per year over a ten year period in two different areas

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Area 1 | 1 | 10 | 7 | 8 | 4 | 2 | 7 | 5 | 11 | 6 |
| Area 2 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |

Summarising data simply means reducing the size. It's easier to look at fewer numbers to get a feel for what is going on than an entire data set.

#### *1.2.1 Averages*

A simple one number summary is the average, most commonly referring to the **mean**. "The average number of earthquakes per year in Area 1 is 6". We know instinctively how to calculate this:

$$\frac{1}{10}(1+10+7+8+4+2+7+5+11+6)=6$$

In general we notate the mean as '$\bar{x}$' or 'x bar', and write:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}(x_1+x_2+x_3+...+x_n) \tag{1}$$

This is just the same calculation as above, where $n=10$ is the size of the data set. The symbol $\Sigma_{i=1}^{n}x_i$ means 'the sum for $i=1$ to $n$ of $x_i$', or 'the sum of $x_1$, $x_2$, ... and $x_n$'.

Sometimes other measures of *average* are either preferred (when the data contain a few observations which are much larger or smaller than the rest) or needed (when the data are not numbers), such as the median or mode. See [**?**] for a good introduction here.

#### *1.2.2 Measures of Spread*

An average alone does not always tell us everything of interest. In the two data sets under study, in both cases $\bar{x}=6$, but they are clearly not equal. Some measure of the level of *variation* in the data is usually a sensible next step. We might think about how much each observation $x_i$ *deviates* from the average, i.e.

$$x_i-\bar{x}.$$

One seemingly sensible measure of spread might be to find how much an observation deviates from the mean *on average*. But we can see that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})=\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\bar{x}=\bar{x}-\bar{x}=0.$$

Two other options are to consider the *squared* and *absolute* deviations from the mean:

$$|x_i-\bar{x}|,\quad\text{and}\quad(x_i-\bar{x})^2,$$

giving two different measures:

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

The first of these is known as the *mean absolute deviation*, the second is called the *variance*.[1] Although both are reasonable measures of spread in a dataset, we prefer the second typically for theoretical reasons [**?**].

Variance is, however, in *squared* units. In the two data sets under study $s^2$ will be in 'squared earthquakes per year' or (earthquakes per year)$^2$, which doesn't mean much! So we typically take the (positive) square root of the resulting quantity, giving us a metric which is in the units we are interested in:

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

We call this the *standard deviation* (note that square rooting before the summation would give us the mean absolute deviation).

Another measure of spread is the *range* of the data, taken as the largest value minus the smallest. This has obvious limitations (as it only depends on two values), but can still be a useful quantity depending on the regularity of your data.

### *1.2.3 Interpreting measures of spread*

If you calculate the standard deviations for the two data sets you will see that they differ. It is intuitive that typically larger values mean that two randomly selected observations from the data set are likely to be further apart from one another. But some useful rules of thumb are:

- Typically around 65-70% of the data will lie within one standard deviation of the mean, i.e. in the range $(\bar{x}-s,\bar{x}+s)$
- Similarly around 95% of the data will lie within two standard deviations of the mean, i.e. in the range $(\bar{x}-2s,\bar{x}+2s)$

These rules will certainly will not always hold (in fact in the data set for Area 2 we see that they don't!) They are actually based on what would happen if your data was approximately *Normally distributed*, something to be discussed in a later course (a good starting point is [**?**]).

## 1.3 Visualising Data

Often the most informative thing to do with a data set is visualise it somehow. There are many creative ways to do this. Some innovative examples can be found on the website *Information is beautiful*,[2] and in many other places. We only give some simple suggestions here. Some good points to remember are:

[1] Sometimes you will see this with an $n-1$ in the denominator. There is no single best choice, so use whichever is more memorable. I realise this is a bit vague, so see [?] for more detail.

[2] To be found at `www.informationisbeautiful.net`

- Is the purpose of the graphical display to communicate (a) specific message(s) or simply to generate interest?
- If the former, then is that message clear?
- In either case, what conclusions (if any) do you want someone to draw from the visualisation?
- If you are making a chart, make the axes labels and titles (if there are any) in large font, so that they are readable

It is very easy to get things wrong, and spend lots of time making a pretty graph that does not say anything. Don't make that mistake! Some great references are [**?**, **?**, **?**], and there are several pieces of software available which are easy to use to create effective graphical displays (e.g. *Tableau*®, *Qlikview*®, *JMP*®).

### *1.3.1 Statistical Graphics*

Two very useful types of chart from a statistical perspective are histograms and scatter plots.

A histogram is a graphical display of the *relative frequency distribution* of a set of data. We divide the range of the data into 'bins' and count the number of data points that lie in each. We then divide each count by the total (relative frequency). A rectangle is plotted for each bin, and the height is (relative frequency)/(bin width). This procedure is a lot simpler when viewed as a table:

| **FTSE End of Day Price** | **frequency** | **relative frequency** | **bin width** | $\frac{\textbf{relative frequency}}{\textbf{bin width}}$ |
|---|---|---|---|---|
| 2000-2500 | 127 | 0.07 | 500 | 0.07/500 |
| 2500-3000 | 470 | 0.25 | 500 | 0.25/500 |
| 3000-3500 | 487 | 0.26 | 500 | 0.26/500 |
| 3500-4000 | 315 | 0.17 | 500 | 0.17/500 |
| 4000-4500 | 127 | 0.07 | 500 | 0.07/500 |
| 4500-5000 | 120 | 0.06 | 500 | 0.06/500 |
| 5000-5500 | 78 | 0.04 | 500 | 0.04/500 |
| 5500-6000 | 111 | 0.06 | 500 | 0.06/500 |
| 6000-6500 | 25 | 0.01 | 500 | 0.01/500 |

Note that dividing by bin width only makes a difference if the bin widths are different. Otherwise it just re-scales the data. The term relative frequency could also be interpreted as a *probability*, as

In the twenty-first century creating histograms is much easier than it used to be...

```
hist(EuStockMarkets[,4], freq=F)
```

Some examples are given in Figure 1.

Scatter plots are useful for understanding how two different things (variables) are associated. Again these days they are much easier to draw...

```
plot(EuStockMarkets[,3],EuStockMarkets[,4])
```

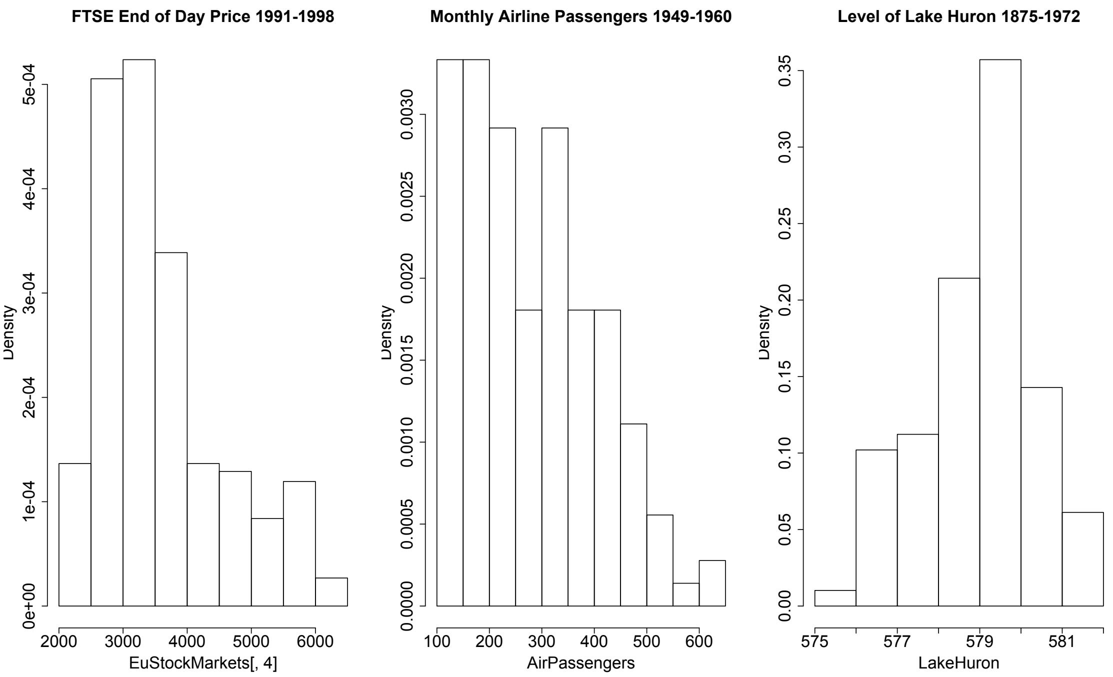


Figure 1: Some example histograms.

You will learn more about scatter plots in your follow on course.

## 2 Basic Probability

Probability theory is a strand of Mathematics. It just so happens that the insights drawn from it can help us better understand real world data. Ideas from probability can give us an indication of how precise a particular guess based on some data is (e.g. what will the average daily rainfall be next year in London?). They can also tell us how much data we need to collect in order to make a guess with a certain level of confidence. In addition we can test whether certain questions about our data are likely to be true or false (hypothesis testing). In essence, having some kind of understanding of the randomness involved in the *data generating process* allows us to make conclusions that go beyond simply what is in the data set itself.

Much of basic probability involves giving a formal mathematical language to things that are intuitively obvious. When situations get more complicated, however, and intuition fails us, we can still use our formal language to draw conclusions to our conclusions. So probability can seem a bit dull at first, but later on you will (hopefully) see the power of it.

### 2.1 Basic Rules

We call something that could happen an outcome, and a set of outcomes an event. The set of all possible outcomes is called the sample space $S$. We write $P(A)$ to mean 'the probability of the event $A$'.

***Example.*** *When rolling a fair six-sided die,* $S = \{1, 2, 3, 4, 5, 6\}$*, possible outcomes are 1, 2, 3 etc. and* $A = \{1, 2\}$ *is the event that the value of the die is 2 or less. So* $P(A) = 1/3$*.*

With these definitions, probability is based on three simple rules:

1. For any event $A$, $P(A) \geq 0$ (a probability must be positive)
2. $P(S) = 1$ (the probability of something happening is one - something must happen!)
3. If $A$ and $B$ cannot both happen, then $P(\text{at least one of } A \text{ or } B \text{ happen.}) = P(A) + P(B)$

The last rule may seem confusing, but again we use it implicitly all the time. In the die roll example, we know that the die cannot land on both 1 and 2 at the same time, and that the probability of each is 1/6, so we naturally say that the probability of the dice roll being two or less is 1/3.

***Example.*** *The events that a fair coin lands heads up and tails up cannot both happen. So the probability that at least one of them happens is* $1/2 + 1/2 = 1$*, as expected.*

Many questions in probability can be much more easily understood using Venn diagrams.

### 2.2 Conditional Probability

We are often interested in understanding how knowledge of one thing can help us make better guesses for another. Examples are:

- Can knowledge of a person's school academic performance help predict whether they will do well at university?
- Does understanding the maximum heights of tidal waves over the last one hundred years enable us to better guess what the largest might be over the next one hundred, and hence build appropriate flood defences?
- Will knowing how much money each Premier League football team spent in the transfer market help predict in what league position they will finish?

Mathematically we write 'the probability of the event $A$ given that we know the event $B$ has occurred' as

$$P(A|B).$$

The above statement is known as a *conditional probability.* We can calculate it using the Probability of both events occurring and the probability of $B$, as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2}$$

This formula is best understood through a Venn diagram.

We call two events $A$ and $B$ *independent* if knowledge of one doesn't influence the probability of the other. In this case

$$P(A|B) = P(A).$$

***Example.*** *The events that one fair coin lands heads up and another also does are independent*

***Example.*** *The events that one fair die takes value greater than or equal to 4 and that a different die takes value 2 or less are independent*

***Example.*** *The events that that there will be more males than females in a building and that the building is a gentlemen's club are not independent.*

Note form the last example that $P(A|B) \neq P(B|A)$ in general. If there are more males than females in a building then the probability that the building is a gentlemen's club is still quite low, as there aren't many gentlemen's clubs around. But If the building *is* a gentlemen's club then the probability that there are more males than females inside is very high!

### 2.3 Notation Cheat Sheet

Here is a quick guide to help you get used to some of the notation. The important thing to remember is mathematicians are not cleverer than you, we are just lazier! Notation is not designed to confuse, just so that we can write things down without as much effort. It can take a bit of getting used to though.

| **Notation** | **Pronunciation** | **Meaning** |
|---|---|---|
| $A$ | The event $A$ | A collection of things that could happen |
| $S$ | The sample space | The collection of all possible things that could happen |
| $P(A)$ | The probability of $A$ | |
| $A \subset S$ | $A$ is a subset of $S$ | Each possible outcome in $A$ will also be in $S$ |
| $A \cup B$ | $A$ or $B$ | The collection of things that are in at least one of $A$ and $B$ |
| $A \cap B$ | $A$ and $B$ | The collection of things that are in both $A$ and $B$ |
| $A\|B$ | $A$ given $B$ | |
| $B\|A$ | $B$ given $A$ | |

## 3 Random Variables

Random variables allow us to bring probability from statements about abstract events into the world of numbers. Many very different phenomena can also give rise to the same random variables, allowing us to develop a very general framework for modelling.

A random variable can be thought of in many ways. A simple definition is that it is a number that varies randomly! Another way to think about it is as the numerical outcome of an experiment *before* the experiment has happened. Another way to describe it is as a map from the sample space $S$ into the set of numbers. We seek to make this idea more concrete through examples.

Essentially, the key point is that it is a lot easier if all the things we are trying to guess are numbers. So random variables are a way to bring everything into the world of numbers.

### 3.1 Discrete Random Variables

These can only take a finite (or countable) number of values. The simplest possible example is a binary random variable:

$$X = \begin{cases} 1 & \text{if coin lands head up.} \\ 0 & \text{otherwise.} \end{cases}$$

We have simply defined a number which takes the value 1 if a coin lands head up and a 0 otherwise. So the ‘experiment’ here is the coin flip, and the random variable $X$ represents the outcome *before* we flip the coin. After we flip the coin $X$ will either be 1 or 0, but beforehand we cannot be sure.

We can define another random variable $Y$ to be the number of car crashes on the M25 motorway this year. We don’t know how many there will be yet as the year is not over, so we can’t exactly say what value $Y$ will take, hence we call it a random variable.

After the experiment has been done, we refer to the outcome as a single *realisation* of the random variable. So if we flipped a coin three times and got three heads, then 1,1 and 1 would be three realisations of the random variable that takes the value 1 if the coin lands heads.

Note that a disease with a 50% mortality rate could be treated exactly the same as a coin flip in this framework. We would simply define a random variable that takes the value 1 if someone survives and a 0 otherwise.

### *3.1.1 Probability Distributions*

A probability distribution is just a table that summarises the possible values that a random variable can take, along with the probability that it takes each of these values. So for a fair coin we have

| $x$ | 0 | 1 |
|---|---|---|
| $P(X=x)$ | 1/2 | 1/2 |

We usually use upper case letters for random variables and lower case letters for the possible values they can take. So $P(X = x)$ means 'the probability that the random variable $X$ takes the value $x$.' The table above tells us that $P(X = 0) = 1/2$, and similarly for $P(X = 1)$.

We call $P(X = x)$ a 'probability mass function' or 'pmf'. Sometimes it is easier to simply write the mass function rather than writing the entire distribution as a table. For the coin flip example we could write

$$P(X=x)=\begin{cases}1/2 & \text{for } x=0,1\\ 0 & \text{otherwise.}\end{cases}$$

If we consider the $Z$ to be the number of flips required before (and including) the first head. Then we cannot say for definite how many flips this will take, so the distribution table is not feasible to draw. But we can write the mass function.

$$P(Z=z)=\begin{cases}\left(\frac{1}{2}\right)^z & \text{for } z=1,2,...\\ 0 & \text{otherwise.}\end{cases}$$

We sometimes also use the *cumulative distribution function*, or 'cdf'. This is simply $P(X \leq x)$, or the probability that a random variable $X$ takes a value less than or equal to $x$. These are more commonly used for continuous random variables, which we discuss later. The relation between the two is

$$P(X \leq x) = \sum_{y \leq x} P(X = y).$$

### *3.1.2 Expectation*

In the same we that we like to summarise data in data analysis with simple one or two number summaries, we would like to do the same thing about a random variable. The 'expectation' or 'expected value' of a random variable $X$ is like an idealised average. We write $E[X]$. If we had several realisations $x_1, x_2, ...x_n$ of the random variable, the average $\bar{x}$ would likely be close to $E[X]$. In fact, the Law of Large Numbers states that as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to E[X].$$

This is why we think of expectation in this way.

We calculate the expected value of a random variable using the formula

$$E[X] = \sum_x xP(X = x)$$

This looks complicated, but it's really just a weighted average. We weight each possible value $X$ can take by how likely it is to occur, and add up the result.

Imagine that we toss a coin 10 times and record the outcomes as a 1 for heads and 0 for tails. So we are working with the random variable

$$X = \begin{cases} 1 & \text{if the coin lands heads,} \\ 0 & \text{if the coin lands tails.} \end{cases}$$

We obtain the data set

$$1, 1, 0, 0, 1, 0, 0, 1, 1, 0.$$

The average value of our random variable is

$$\frac{1}{10}(1+1+0+0+1+0+0+1+1+0),$$

which we could also write this as

$$\frac{1}{10}(0 \times 5 + 1 \times 5) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2},$$

which is the same as:

$$0 \times P(X=0) + 1 \times P(X=1) = \sum_x xP(X=x).$$

Now, clearly we won't get 5 heads and 5 tails every time we flip a coin 10 times, but that is why we call expectation an *idealised* average.

We can similarly define the *variance* of a random variable as

$$\mathrm{Var}[X] = E[(X - E[X])^2].$$

### 3.2 A note on continuous data and Calculus

Often, the data we deal with (or the things we would like to model) are not things that can take only a discrete number of possible values. Examples are the times in between volcanic eruptions, heights of waves or the geographical distances between seismic events. We can model these with *continuous* random variables.

Unfortunately the theory is a little more complicated here. Imagine picking any number between 0 and 1 at random. We denote the number you choose by $X$. But there are so many possible values $X$ could take that the probability you pick any single number is basically zero. So we prefer to talk about the probability that a continuous random variable is in a certain region, e.g. $P(X \leq 1/2)$. The theory has been developed around this idea.

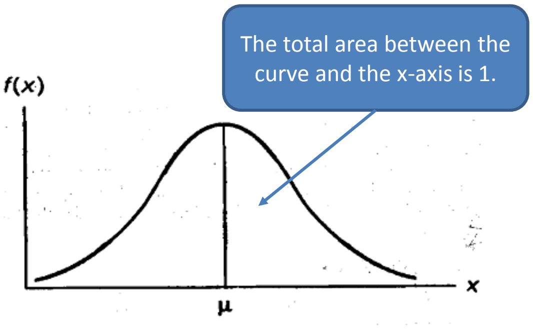
### 3.3 Continuous random variables

Technically, we call a continuous random variable one that can take *uncountably* many values. Countability is an abstract Mathematical concept, but a really nice introduction is given in the short (7 minute) TED talk ‘How big is infinity?’ by Dennis Wildfogel.[3] Basically the idea are there are so many possible values for the random variable that it is easier to say

$$P(a \leq X \leq b)$$

than to talk about the probability of $X$ being equal to any particular value (since the latter is basically zero).

We develop this theory through the concept of a ‘probability density function’. This is a function $f(x)$ which is always non-negative, and that if we graph it, the area underneath it is equal to 1.



The total area between the curve and the x-axis is 1.

A toy example which helps intuition is to think of a game show with a spinner (we do this on the board, I’ve left some space here in case you want to copy out some of the drawings).

[3] Available online at `https://www.youtube.com/watch?v=UPA3bwVVzGI`.

### 3.3.1 How do we calculate the area under a curve?

To calculate the area under a curve we use a tool called *integration*, which is one half of Calculus (the other is differentiation. Here we just give some intuition.[4]

We know how to calculate the area of a rectangle, so a simple solution is to just approximate the area of the curve by drawing rectangles underneath it and calculating the area of those. This gives us the sum

$$\text{Area} \approx \sum_{i=1}^{n} f(x_i) \triangle x.$$

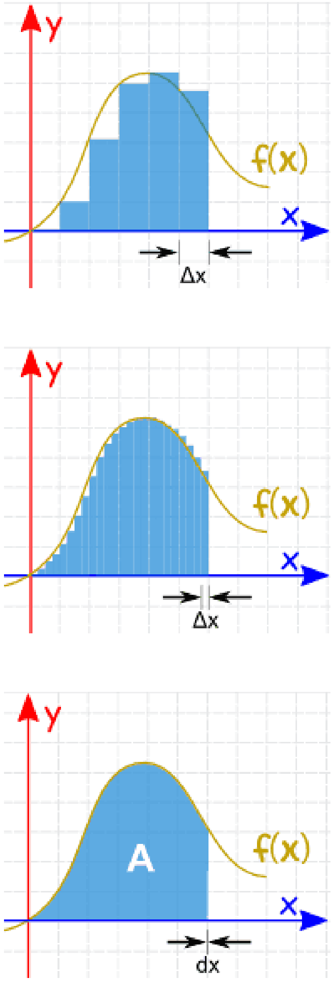
The clearly the more rectangles we use (or mathematically, as $n$ gets bigger) the approximation becomes more accurate. The 'integral' is basically the limiting process - what would happen if we took an infinite number of rectangles. We replace the $\triangle x$ with $dx$, which represents an infinitessimally small length, and calculate the area of an infinite number of these infinitely thin rectangles. The notation we use is

$$\text{Area } = \int_a^b f(x)dx$$

The notation $\int$ is just a big S, which stands for sum! In the case where $f(x)$ is a density for a random variable $X$ we can write

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

To actually calculate the integral we use several 'off the shelf' results. For example, if $f(x) = x^2$ then the integral is $(b^3 - a^3)/3$. A first course in Calculus basically involves memorising many of these 'off the shelf' results!



### 3.3.2 Expectation

The expectation of a continuous random variable follows the same 'weighted average' intuition as that used for the discrete case. Here we just have to replace sums with integrals. We think of

$$f(x)dx \approx P(X = x),$$

(note that this will be an extremely small number). We also replace $\sum$ with $\int$. So we have

$$E[X] = \int x f(x)dx,$$

[4] Figure taken from `http://www.mathsisfun.com/calculus/integration-introduction.html`.

just like in the discrete case

$$E[X] = \sum_x xP(X = x).$$

Hopefully this gives at least some sort of conceptual picture.