1 WYNTKA… Statistics

Statistics is one subject area in which many otherwise well-prepared students seem to have little knowledge of before they start studying engineering – and this is a pity, since it is one of the most useful and important topics in mathematics for modern life. It's a continuing puzzle to me why it is not more widely taught, when other less useful subjects fill up the maths syllabus in schools.

Anyway, this chapter will try and cover the basics, at least those basics that will be required for a study of the rest of this book. Before I start, I should say that I won't attempt to provide strict definitions and rigorous mathematical derivations of the results, I'm only aiming to provide an indication of where the key results come from, and why they are useful. Anyone who wants more details is recommended to consult a statistics textbook.

Symbols Used in this Chapter:

x, y	Random variables[1]
x_i, y_i	The i^{th} sample of a series of values of the random variables x and y
X, Y	Fixed, known, well-defined possible values of the random variables x and y
$p_x(x)$	A continuous probability density function
$q_x(x)$	A discrete probability density function
$P_x(x)$	A continuous cumulative distribution function
$Q_x(x)$	A discrete cumulative distribution function
N	The number of samples taken of values with a certain statistical distribution

Just about all modern communications systems are designed on a statistical basis, which is another way of saying they don't always work. Perhaps that's a bit unfair: they 'work' in the sense that they meet their design criteria (or they should), it's just that no engineer specifying or designing a communications system expects all attempts to use the communications system to be successful. They know that's impossible, or at the very least, hopelessly uneconomical.

End users? Well, they're different, they usually do expect communication systems to always work and get upset if, for example, they get a message telling them "all lines are engaged, please try again later", or their web page takes several minutes to load, or they can't get service from their mobile phone in some small village in a mountainous region.

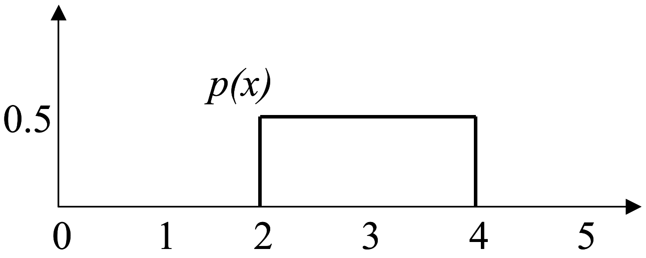
To provide enough lines for everyone to always get through to any call centre on the first attempt; to provide enough capacity in the Internet to ensure all web-pages could download within a few seconds; to provide enough base stations with sufficient power to offer mobile phone service everywhere in the country: this could only be done with enormous investment, which would increase everyone's bill. And then the end users would really complain. Like it or not, we're going to have to analyse systems on a statistical basis, so that the operators can decide how much they want to invest, and what service levels they want to offer, and design and specify systems accordingly. This requires a basic knowledge of statistics.

[1] I'm using the term 'random variable' here to represent a quantity which can take any of a range of values, and varies (usually with time), in such a way that I cannot exactly predict what value it is going to take at any specified time in the future. I can, however, describe the variable in terms of how likely it is that it has any particular value at any particular time. Since this includes things like the amplitude of an audio speech signal, it seems a little odd to call them 'random variables' (I hope most of what I say isn't 'random', although some of my students may disagree). However, other terms such as 'statistical variable' or 'stochastic variable' sound a little less friendly and more confusing, so I've decided to stick with 'random variables' here.

1.1 Probability Distributions

The key point about a random variable is that unlike a normal variable, which has a fixed, if sometimes unknown value, random variables don't have a unique, fixed value. It's like the difference between asking how far it is from my house to my office (a fixed value, even if you might not know what it is, at least it doesn't change) and asking how far I walk each day (there is no single fixed answer to that question, the distance changes every day according to which route I take). We have to describe random variables in terms of probability distributions: how likely it is that they have a certain value.

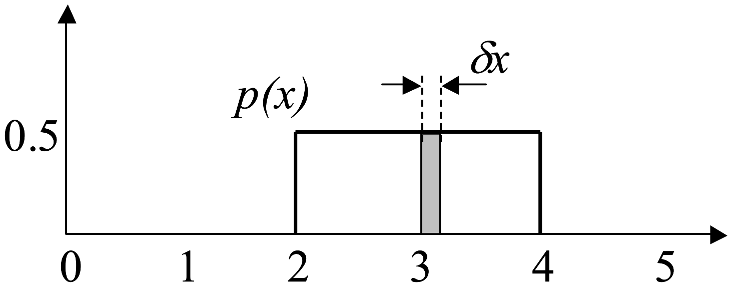
For example: consider a random variable that can take any value between two and four. Suppose it is equally likely to take any value in that range. We can represent the probability distribution of this variable as a function, *p(x)*, as shown in the figure below. This is known as a *probability density function*.



Note that the area under the graph is one. That's always true for all probability density functions – no exceptions. Question: what is the probability that a sample taken at random from this distribution has a value of five? Answer: zero. That's obvious, I hope. Now: what's the probability that another sample value has a value of three? The answer is also zero.

This might seem surprising at first sight – the function has a value of 0.5 when the value is three – but that's not what the probability density function is about. The probability density function is a *density*. Only when integrated over a range of possible values does the probability density function become a real probability.

For example, consider a range of values of x between 3 and $3+\delta x$. In the figure below, the shaded area has an area of $p(x)\delta x$, and this is the probability that any random value has a value between 3 and $3+\delta x$.



If the quantity x had units of volts, then the probability density function $p(x)$ would have units of volts^{-1}, so that the product $p(x)dx$ was dimensionless (as all probabilities are).

In the more general case, the probability that any particular value lies between two numbers a and b is:

$$probability\ \left(a \le x < b\right) = \int_a^b p(x)\ dx$$

It's very easy to show that the area under the whole probability density curve must always be one. It's equivalent to asking: "what's the probability that the value of a random variable is between minus infinity and infinity?" The answer is one. It's got to have some value, after all.

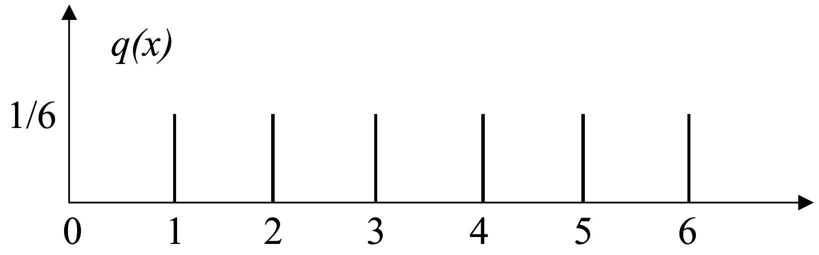
This is also why the probability of the value of a random sample of the distribution being exactly three is zero:

$$probability\ (x = 3) = \int_3^3 p(x)\ dx = 0$$

At least that's true for continuous distributions – distributions in which the random variable can have any value within a certain range: for example the length of a phone call can have any value between zero seconds and infinity.

1.1.1 Discrete Distributions

There's another important type of distribution as well: a discrete distribution. In this case, the possible values of the random variable are restricted to certain discrete values, often integers. For example: what's the probability that a single roll of a die gives three? Answer: one-sixth (at least assuming the die is evenly weighted). The probability density function can be shown like this:



although it's important to note that in this case, these vertical lines are impulses of infinite height, but area 1/6 (so that the total area under the probability density function is still one).

Unlike with a continuous distribution, the probability density function $q(x)$ for a discrete distribution is the probability that the statistical value is exactly equal to x. We now have:

$$probability\ (a \le x \le b) = \sum_{i=a}^{b} q(i)$$

1.1.2 Cumulative Probability Distributions

Quite often, it's useful to plot the integral from minus infinity to some value x of the probability distribution. This results in a function known as the cumulative distribution function (CDF), which is given for continuous and discrete distributions by:

$$P(x) = \int_{-\infty}^{x} p(y)\,dy \qquad Q(x) = \sum_{i=-\infty}^{x} q(i)$$

In other words, *P(x)* is the probability that the statistical variable has some value equal to or less than *x*. It's obvious, I hope, that $P(\infty)=Q(\infty)=1$ for all probability distributions.

1.2 Averages: Modes, Medians, Means and rms Values

Since we can't work out a single value for a random variable (because they keep changing), and it's often unwieldy to have to talk about the entire shape of the probability distribution function whenever we want to describe the sort of values we might expect a statistical variable to take, it's useful to be able to define a few simple parameters that can characterise the distributions. The first of these is the average value – although 'average' isn't a good word to use in statistics, since it doesn't have a well-defined meaning. Instead, there are four common values, all of which are occasionally used to describe the average value in communications[2].

The *mode* is the most likely value. The probability of exceeding the *median* value is 50%. The *mean* is the value you would get by taking a very large number of samples of the random variable, adding them all up, then dividing by the number of samples. The *rms* (*root mean square*) value is similar to the mean, except you square the values before adding them, then divide by the number of samples, and finally take the square root of the result[3] [4].

In mathematical notation:

$$mean=\lim_{N\to\infty}\left(\frac{\sum_{i=1}^{N}x_i}{N}\right)\qquad \frac{1}{2}=\int_{-\infty}^{median}p(x)\,dx\qquad rms=\lim_{N\to\infty}\left(\sqrt{\frac{\sum_{i=1}^{N}x_i^2}{N}}\right)$$

While it's obvious how to calculate the median value from the probability distribution[5], it might not be quite so obvious for the mean and rms.

Consider a discrete distribution: the probability of any sample of the random variable having a value of *x* is *q*(*x*), so if there are a very large number of samples *N* of the statistical variable taken, then the number of samples with a value equal to *x* is going to tend towards *N q(x)*[6]; and

[2] Actually there are more than four types of average: but these four are the most common in communications engineering, and of these four the mean is the most common, and the mode the least common.

[3] This means that the rms value is always positive, so it can have values very different from the mean, median and mode. Consider a statistical variable that always had values very close to –1. The mean, median and mode would all be close to –1: whereas the rms value would be around +1.

[4] The rms value is useful since the average power in a signal is proportional to the mean value of the square of the signal: so the rms value is proportional to the square root of the average power.

[5] Well, perhaps not always obvious. The formula given is fine for well-behaved continuous distributions (by which I mean distributions with no discontinuities), but suppose you have a discrete distribution in which values are equally likely to be any integer from one to six (i.e. the result of a die throw). What's the median? The probability of having a value lower than, e.g. 3.1 is 50%, so that works; but so does any other number between three and four. In these circumstances it is conventional to take a value half-way between the highest and lowest possible medians. In this case, that means 3.5.

[6] Mathematicians like to do lengthy proofs of statements like this. Engineers usually just say "yes, that's sort of obvious" and move on.

each of these contributes a value of *x* to the sum. This is true for all possible values of *x*, so we can write:

$$\sum_{i=1}^{N} x_i \to \sum_{x=-\infty}^{\infty} x\,N\,q(x)$$

$$mean = \lim_{N\to\infty}\left(\frac{\sum\limits_{x=-\infty}^{\infty} x\big(N\,q(x)\big)}{N}\right) = \sum_{x=-\infty}^{\infty} x\,q(x)$$

For continuous distributions exactly the same logic applies: for a very large number of samples *N*, the probability of any sample lying between *x* and *x+dx* is *p(x) dx*, so the number of samples in this range is *N p(x) dx*, each of which contributes a value *x* to the sum, so we can write:

$$\sum_{i=1}^{N} x_i \to \sum_{x=-\infty}^{\infty} x\,N\,q(x) \to \int_{x=-\infty}^{\infty} x\,N\,p\big(x\big)\,dx$$

$$mean = \lim_{N\to\infty}\left(\frac{\int\limits_{x=-\infty}^{\infty} x\big(N\,p(x)\big)dx}{N}\right) = \int_{x=-\infty}^{\infty} x\,p(x)\,dx$$

Calculating the rms follows exactly similar reasoning, and gives (for discrete and continuous distributions respectively):

$$rms = \sqrt{\sum_{x=-\infty}^{\infty} x^2\,q(x)} \qquad rms = \sqrt{\int_{x=-\infty}^{\infty} x^2\,p(x)\,dx}$$

1.2.1 Expected Values

“Expected value” or “expectation” is a term that implies the mean of a random variable. For example, the expectation value of *x* is the mean value of *x*, and the expectation value of x^2 is the square of the rms value of *x* (in other words, the mean of x^2). The expectation value of a quantity is notated by writing a line above the quantity, or by using the function E( ), which takes a large number of arguments, and returns their mean value:

$$E(x) = \overline{x} = mean(x) \qquad E(x^2) = \overline{x^2} = \big(rms(x)\big)^2$$

1.2.2 A Continuous Example – The Uniform Distribution

Consider the first example distribution, the uniform distribution where the statistical variable *x* was equally likely to take any value between two and four.

The mean:

$$mean = \int_{x=-\infty}^{\infty} x\,p(x)\,dx = \int_{x=2}^{4} x\frac{1}{2}dx = \frac{1}{2}\left[\frac{x^2}{2}\right]_2^4 = 3$$

The rms:

$$rms = \sqrt{\int_{x=-\infty}^{\infty} x^2\,p(x)\,dx} = \sqrt{\int_{x=2}^{4} x^2\frac{1}{2}dx} = \sqrt{\frac{1}{2}\left[\frac{x^3}{3}\right]_2^4} = \sqrt{\frac{28}{3}} = 3.055...$$

The median:

$$\frac{1}{2} = \int_{-\infty}^{median} p(x)\,dx = \int_{2}^{median}\frac{1}{2}dx = \left[\frac{x}{2}\right]_2^{median}$$

$$\frac{1}{2} = \frac{median}{2} - \frac{2}{2}$$

$$median = 3$$

although you don’t need to do that last calculation, it should be obvious from the shape of the probability distribution what value has 50% of all samples of the statistical variable below it.

There is no unique value for the mode, since this distribution is equally likely to take any value over the whole range; there is no single value more likely than the other values. It could be said that any value between two and four is a mode of this distribution.

1.2.3 A Discrete Example – A Weighted Die

Suppose you have a die that is unfairly weighted. It is twice as likely to come up with a value of one than it is to come up with a value of 2,3,4,5 or 6.

The mean value is[7]:

$$mean = \sum_{i=1}^{6} p(i)\,i = \frac{2}{7}.1 + \frac{1}{7}.2 + \frac{1}{7}.3 + \frac{1}{7}.4 + \frac{1}{7}.5 + \frac{1}{7}.6$$

$$= \frac{22}{7} = 3.1429...$$

The mode is one: this distribution has a single most likely value. If the die were fairly weighted, so that every number was equally likely to appear, we’d have to say that 1,2,3,4,5 and 6 were all *modes* of the distribution.

[7] You might find it obvious that the probability of rolling a one is 2/7. If not: let the probability of rolling a 6 be *p*. Then the probability of rolling a 2,3,4 or 5 is also *p*, and the probability of rolling a one is *2p*. The sum of all terms in a discrete probability must be one, so *2p + p + p + p + p + p = 1*, and so *p = 1/7*, etc.

The median value is sometimes (although not here) problematic[8]. The probability that the value is less that three is 3/7, and the probability that the value is above three is also 3/7, so that makes three the median.

1.3 Variances and Standard Deviations

The most frequently-quoted parameter of a distribution is the mean, sometimes referred to as the "first moment" of the distribution[9]. After knowing the mean of a distribution, the second most useful thing to know is to have some idea of the shape of the distribution: in particular whether all values from the distribution are likely to be very close to the mean, or whether there is a large spread in possible values and values exist a long way from the mean. The variance is a single parameter that gives some information about the spread of the statistical variables from the mean. It's defined as:

$$\sigma^2 = \lim_{N\to\infty}\left(\frac{\sum_N \left(x_i - \overline{x}\right)^2}{N}\right)$$

In other words, it's the expectation (mean) value of the square of the difference between the statistical variable and the mean of the distribution: a measure of how likely it is that the variable is a long way from the mean[10]. Using similar arguments to above, the variance can be derived from the formulas:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\overline{x})^2\, p(x)\, dx \qquad \sigma^2 = \sum_{-\infty}^{\infty} (x-\overline{x})^2\, p(x)$$

for continuous and discrete distributions respectively. There's another way of working out the variance as well, which is often easier to do in practice, and that's to note that:

$$\sigma^2 = \overline{\left(x-\overline{x}\right)^2} = \overline{\left(x^2 - 2x\overline{x} + \left(\overline{x}\right)^2\right)}$$

[8] Note that, for a fair die (and other circumstances), it can be difficult to calculate a single unique value for the mode and the median. That's one of the reasons why the mean is often preferred: it's almost always possible to find a single unique value for the mean of any useful distribution.

[9] The square of the rms value is sometimes referred to as the "second moment". In general, the n^th^ moment of the distribution can be defined according to:

$$n^{th}\ moment\, of\ p(x) = \int_{-\infty}^{\infty} x^n\, p(x)\ dx$$

Third and higher moments are used in communications theory, particularly in signal identification algorithms.

[10] Note that in the special case where the mean is zero, the variance is equal to the second moment of the distribution: in other words it is proportional to the mean power in the signal (if the random variable is a signal, of course).

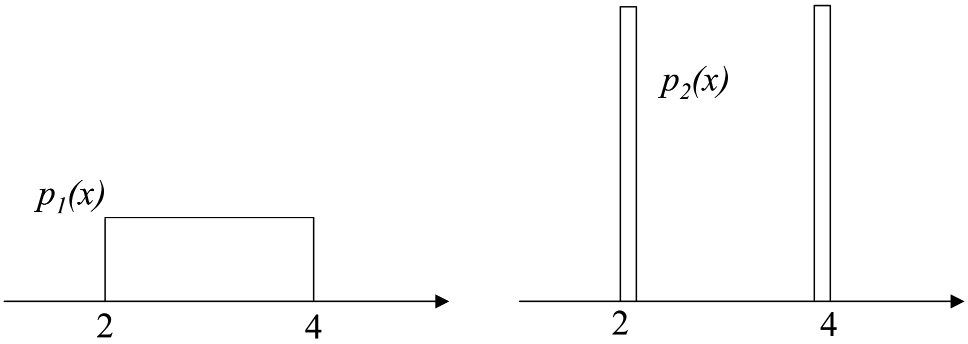
There is a theorem that states that the sum of two statistical variables has a mean equal to the sum of the means of the individual statistical variables[11]. An simple extension to this theorem can be used here, to show that the mean of the sum of these three terms (two of them statistical variables, the third (the square of the mean) is just a constant) is equal to the sum of the means of these three terms. Remembering that $\bar{x}$ is just a constant term, so that its expectation value is equal to itself, we can show that:

$$\sigma^2 = \overline{\left(x^2 - 2x\bar{x} + \left(\bar{x}\right)^2\right)} = \overline{x^2} - \overline{2x\bar{x}} + \overline{\left(\bar{x}\right)^2}$$

$$= \overline{x^2} - 2\bar{x}\,\bar{x} + \left(\bar{x}\right)^2 = \overline{x^2} - \left(\bar{x}\right)^2$$

This provides a simple way to calculate the variance: it's just the difference between the mean of the squares and the square of the mean. In other words, it's the second moment minus the square of the first moment of the distribution.

For example, consider two distributions, with the same mean, as shown in the figure below.



In the first distribution, all values in the distribution lie between two and four. We know from the previous example that the mean is three and the rms value is $\sqrt{28/3}$ =3.0551. In the second distribution, values can either lie between 2.0 and 2.1, or between 3.9 and 4.0, but nowhere else. Again, the mean is three, but now the rms value is[12]:

$$rms = \sqrt{\int_{-\infty}^{\infty} x^2\, p(x)\, dx} = \sqrt{\int_{2}^{2.1} x^2\, 5\; dx + \int_{3.9}^{4} x^2\, 5\; dx}$$

$$= \sqrt{5\left(\left[\frac{x^3}{3}\right]_2^{2.1} + \left[\frac{x^3}{3}\right]_{3.9}^{4}\right)} = 3.147...$$

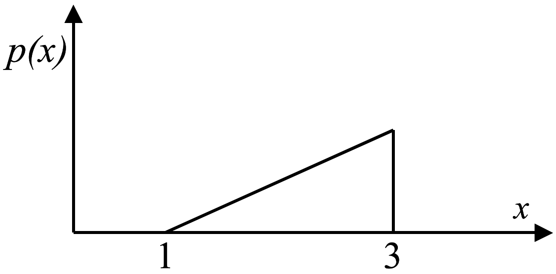
[11] See later for a proof of this.

[12] Note that *p(x)* has a maximum value of 5 here: the value of x must lie between 2.0 and 2.1; or between 3.9 and 4.0. That's a range of 0.2, and we need the integral of the probability distribution to be equal to 1, and since *p(x)* is a uniform probability distribution, that means *p(x)* = 1 / 0.2 = 5.

So the variance of the first distribution is ${}^{28}\!/_{3} - 3^2 = {}^{1}\!/_{3}$, and the variance of the second distribution is $3.147^2 - 3^2 = 0.904...$ and the standard deviations are $\sqrt{{}^{1}\!/_{3}} = 0.577...$ and 0.951… respectively. The more likely the values of a statistical variable are to be a long way from the mean, the larger the variance.

1.4 Problems

1) A continuous probability density function has the form:



a) What is the value of the probability density function when $x = 2$?

b) What is the mean value of this distribution? The mode? The median?

c) Derive the equation of the CDF (cumulative probability distribution) for this probability distribution, and plot the result.

2) Consider the weighted die described in section 1.2.3. What's the standard deviation of this distribution?

3) Another weighted die always comes up with either a one or a six, and never a two, three, four or five. What is the difference between the mean and median of this die, and the mean and median of a fairly-weighted die which is equally likely to end up with any of the six sides pointing up?

What is the difference in the standard deviation of the results from rolling these two dice?

4) Prove that for any probability distribution that is symmetrical about the median, the mean is equal to the median. What can be said about the mode in these cases?

5) Every day, suppose you leave your house, and start taking steps, each step being in a different, entirely random direction. After a very large number of steps, how far are you away from the house? That's a statistical variable, and it has a Rayleigh distribution[13], which has the form:

[13] It's interesting to note that the Rayleigh distribution has also been used with some success to describe how long it takes to complete the writing of a piece of software. This doesn't say a lot about the ability of software engineers to plan their work.

$$p(x) = \frac{2x}{r^2} \exp\left(-\frac{x^2}{r^2} \right)$$

where *r* is the rms (root mean square) value of the distribution. What's the median value? How about the mode?