Chapter 4: An Introduction to Probability and Statistics

4.1 Probability

The simplest kinds of probabilities to understand are reflected in everyday ideas like these:

(i) if you toss a coin, the probability that it will turn up heads is $1/2$ (sometimes we might say 50% but our probabilities will be fractions between 0 and 1);

(ii) if you roll a die, the probability that side 1 (one dot) will turn up is $1/6$;

(iii) if you own one raffle ticket in a raffle where 8000 tickets were sold, the probability that your ticket will win (or be drawn first) is $1/8000$.

All of these examples are based on a fairness assumption and are to some extent idealizations of the real world. This subject is relatively easy in theory, but becomes much more tricky if you want to find a theory that accurately reflects some real world situation. Our goal is to stick to simple situations where the theory can be used directly.

In all 3 examples above there is a *random experiment* involved, where the result is not entirely predictable. Even "predictable" scientific experiments are rarely entirely predictable and will usually have some randomness (caused by extraneous effects on the experimental apparatus or inaccuracies in measurements or other such factors). Thus scientific experiments are frequently treated as random experiments also.

For our theoretical framework for these random experiments, we want to concentrate on a few key ideas.

A All the possible outcomes of the experiment (in one *trial*). We think mathematically of the *set* of all possible outcomes and we refer to this by the technical term *the sample space* for the experiment. We may denote the sample space by S often.

B Some interesting sets of outcomes, or subsets of the sample space. These are called *events*. So an event is a subset $E \subset S$.

An example might be that in a game with dice, we want to toss a die and get an odd number. So we would be interested in the event $E = \{1, 3, 5\}$ in the sample space for rolling a die, which we take to be $S = \{1, 2, 3, 4, 5, 6\}$.

We sometimes use the term *simple event* for the events with just one element. Thus in the dice case the simple events are $\{1\}, \{2\}, \ldots, \{6\}$.

C The third term to explain is *probability*. We will deal now only with the case of a finite sample space and return later to the case of an infinite sample space.

Infinite sample spaces are possible if the experiment involves measuring a numerical quantity or counting something where there is no special limit on the number. Take for example the experiment of measuring the time (in seconds, say) taken for 100 oscillations of a given pendulum. In principle the result could be any positive number and each mathematically precise number will have zero probability. (There may be a positive probability of a measurement between 200.0 and 200.1 but there will be zero probability of getting exactly 200.00... as the measurement.)

A counting example with a sample space $\{0, 1, 2, 3, \ldots\}$ might be recording the number of gamma rays hitting a specific detector in 1 second. If the gamma rays are from some radioactive source there will be a certain probability for each number $0, 1, 2, \ldots$.

In the case where the sample space $S = \{s_1, s_2, \ldots, s_n\}$ is finite, we suppose there is a probability p_i attached to each outcome s_i in such a way that each p_i is in the range $0 \leq p_i \leq 1$ and the sum of all the probabilities is 1. (So $\sum_{i=1}^{n} p_i = 1$.)

Then we compute the probability of any event $E \subset S$ as the sum of the probabilities for the outcomes in E. So, if $E = \{s_1, s_5, s_7\}$, then the probability of E is $p_1 + p_5 + p_7$.

We write $P(E)$ for the probability of an event E and we sometimes write $P(s_i)$ for p_i, the probability of the outcome s_i.

If we take the example of the die, we are using $S = \{1, 2, 3, 4, 5, 6\}$ and each outcome has probability $1/6$. So, for the event $E = \{1, 3, 5\}$ we get probability $P(E) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$.

That was an example where each of the possible outcomes is equally probable, and in examples of this type calculating probabilities for events comes down to counting. You count the total number of outcomes and the reciprocal of that is the probability of each outcome. To find the probability of an event you count the number of outcomes in the event and multiply by that reciprocal (or equivalently divide by the total number of outcomes). We will **not** be dealing with examples of this type except as simple illustrations. Instead we will deal with the general theory and that deals with situations where the individual outcomes are not necessarily equally probable. A weighted coin or die are simple examples of this type.

You can easily see that for an infinite sample space we cannot assign the same positive probability to each of the infinitely many outcomes in such a way that they add to 1. Or, we cannot work out sensible probabilities by just dividing by infinity.

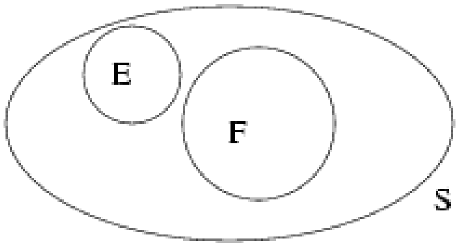
In general we get a probability $P(E)$ for each event $E \subset S$ in such a way that the following rules hold:

(i) $0 \leq P(E) \leq 1$ for each $E \subset S$;

(ii) $P(\emptyset) = 0$ and $P(S) = 1$;

(iii) if $E \subset S$ and $F \subset S$ are two events with $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$. (We call E and F *mutually exclusive* events if $E \cap F = \emptyset$.)

A Venn diagram may help to imagine what this third property says.



When formulated in this way, the idea of a probability works for the case of infinite sample spaces, which we come to soon.

4.2 Theoretical Means

The term *mean* (also called *expectation* sometimes) is another word for average, in this context a long-run average.

If you roll a die 5000 times you would expect that each of the 6 numbers on the die will show up about $5000/6$ times. Of course this would not happen exactly (5000 is not divisible by 6, but even if it was we would not expect each number to show up exactly as often as every other, only roughly as often). So if we were to write down the 5000 numbers that showed up and tot them up we would get roughly

$$\frac{5000}{6}\times 1+\frac{5000}{6}\times 2+\frac{5000}{6}\times 3+\frac{5000}{6}\times 4+\frac{5000}{6}\times 5+\frac{5000}{6}\times 6$$

So if we take the average number that turned up (the result of the tot divided by the total number 5000) we get

$$\begin{aligned}\text{average} &= \frac{1}{5000}\left(\frac{5000}{6}\times 1+\frac{5000}{6}\times 2+\frac{5000}{6}\times 3+\frac{5000}{6}\times 4+\frac{5000}{6}\times 5+\frac{5000}{6}\times 6\right)\\ &= \frac{1}{6}\times 1+\frac{1}{6}\times 2+\frac{1}{6}\times 3+\frac{1}{6}\times 4+\frac{1}{6}\times 5+\frac{1}{6}\times 6\\ &= \frac{1}{6}\left(\frac{6\times 7}{2}\right)=\frac{1}{6}(21)=\frac{7}{2}\end{aligned}$$

This mean of $\frac{7}{2}$ is obviously not a number that will ever show up on the die. It is the long run average of the numbers that show up. If we actually rolled the die 5000 times and did the tot of the numbers we got and divided that by 5000 we would not be likely to get exactly $\frac{7}{2}$, but we should get something close to that. If we rolled the die even more than 5000 times (10000 or 100000 times, say) we could expect our average to come out closer to $\frac{7}{2}$.

Looking beyond this particular example to a more general experiment, we can realize that we can only average numerical quantities. If we tossed a coin 5000 times we would get a list of heads and tails as our 5000 outcomes and it would not make any sense to average them. In the raffle example, we could rerun the draw lots of times and average the ticket numbers that show up, but this would not make a lot of sense.

With the coin experiment, suppose we had a game that involved you winning 50c if heads came up and losing 75c if tails came up. Then it would make sense to figure out your average winnings on each toss of the coin. Similarly, with the raffle, suppose there was just one prize of €1000 and each ticket cost 25c. That means that if your ticket wins then you end up gaining €999.75 and if any other ticket wins your gain is €-0.25. Figuring out your average (or mean) gain in the long run gives you an idea what to expect.

This idea of a numerical quantity associated with the result of a random experiment (eg the amount you win, which is something that depends on the numbers on your tickets plus the result of the draw) is a basic one and we have a technical term for it. A real-valued function $X : S \to \mathbb{R}$ on the sample space S is called a *random variable*. The *mean* of such a random variable is

$$p_1X(s_1) + p_2X(s_2) + \cdots + p_nX(s_n) = P(s_1)X(s_1) + P(s_2)X(s_2) + \cdots + P(s_n)X(s_n)$$

We can see that this is a simple formula (multiply the probability of each outcome by the value of the random variable if that is the outcome and add them up) and it can be motivated in the same way as we did the calculation above with rolling the die that lead to the long run average $\frac{7}{2}$.

If we did our experiment a large number N times, we would expect that each outcome s_i should happen about p_iN times (the proportion dictated by the probability). If we wrote down the values $X(s)$ for the random variable each time and totted them up, we should get roughly

$$p_1NX(s_1) + p_2NX(s_2) + \cdots + p_nNX(s_n)$$

and dividing by N to get an average, we would find that the average should be about the mean.

4.2.1 Definition. If we have a random experiment with sample space $S = \{s_1, s_2, \ldots, s_n\}$ and a random variable $X : S \to \mathbb{R}$, then the *mean* of X is

$$\text{mean} = \mu = P(s_1)X(s_1) + P(s_2)X(s_2) + \cdots + P(s_n)X(s_n) = \sum_{i=1}^{n} P(s_i)X(s_i).$$

The *variance* of the random variable X is

$$\sigma^2 = P(s_1)(X(s_1) - \mu)^2 + P(s_2)(X(s_2) - \mu)^2 + \cdots + P(s_n)(X(s_n) - \mu)^2$$

The square root σ of the variance is called the *standard deviation* of the random variable.

4.2.2 Remark. The variance is the mean square deviation of the random variable from its mean and the variance is large if the values of the random variable are often far away from the mean (often means often in the long run or with high probability). The standard deviation is the root mean square deviation and is easier to think about because it is in the same units as the quantity X. It has to do with the amount of scatter or spread in the values of X. If σ is rather small, then there is a good chance that the value of X will be near the mean, but if σ is very big that is not so.

4.2.3 Example. Suppose we take the example of the die, sample space $S = \{1, 2, 3, 4, 5, 6\}$, each outcome has probability $1/6$ and the random variable $X : S \to \mathbb{R}$ which has the 6 values $X(1) = 1$, $X(2) = -1$, $X(3) = 2$, $X(4) = 2$, $X(5) = -1$ and $X(6) = 1$. Then the mean is

$$\begin{aligned} \mu &= P(1)X(1) + P(2)X(2) + \cdots + P(6)X(6) \\ &= \frac{1}{6}(1) + \frac{1}{6}(-1) + \frac{1}{6}(2) + \frac{1}{6}(2) + \frac{1}{6}(-1) + \frac{1}{6}(1) \\ &= \frac{2}{3} \end{aligned}$$

and the variance is

$$\begin{aligned} \sigma^2 &= P(1)(X(1) - \mu)^2 + P(2)(X(2) - \mu)^2 + \cdots + P(6)(X(6) - \mu)^2 \\ &= \frac{1}{6}((1 - 2/3)^2 + (-1 - 2/3)^2 + (2 - 2/3)^2 + (2 - 2/3)^2 + \\ &\quad (-1 - 2/3)^2 + (1 - 2/3)^2) \\ &= 14/9 \cong 1.555... \end{aligned}$$

Thus $\sigma = \sqrt{14/9} \cong 1.247219$ is the standard deviation in this example.

4.3 Sample means and variances

Suppose we perform a real experiment several times and get measurements $X_1, X_2, \ldots, X_n$. We believe there is a sample space and some probabilities behind this experiment, but we are not sure how to work out the probabilities explicitly. We can try to work things out form the data at our disposal. The *sample mean* is

$$\text{sample mean} = \text{average} = m = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It gives us an estimate of what the theoretical mean would be if we could work out the appropriate probabilities.

The *sample variance* is not quite an average. It is

$$\frac{1}{n-1}((X_1 - m)^2 + (X_2 - m)^2 + \ldots + (X_n - m)^2))$$

The $n - 1$ gives a better idea of what the real theoretical variance is than you would get if you replaced the $\frac{1}{n-1}$ factor by $\frac{1}{n}$. (We are not in a position to explain why that is so. However, if n is big there is not such a big difference between $\frac{1}{n-1}$ and $\frac{1}{n}$. If n is small enough for the difference to matter much, then there probably is not enough data to be able to draw conclusions.)

4.4 Conditional probabilities

Suppose we have some (partial) information about the result of a random experiment. Specifically, suppose we know that the outcome is in a subset A of the sample space S (or that the event A has occurred). What effect should that have on the probabilities?

With conditional probabilities we assume that the relative likelihood of the outcomes within A remain as they were before we had any information, but have to be scaled up to give a total probability of 1.

The *conditional probability* of an event B given that event A has occurred is defined to be

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

4.4.1 Example. For example, suppose we have a biased die where the 6 possible outcomes $S = \{1, 2, 3, 4, 5, 6\}$ have probabilities

$$\frac{1}{12}, \frac{1}{12}, \frac{3}{12}, \frac{3}{12}, \frac{2}{12}, \frac{2}{12}$$

(in that order) and we see that at least 2 dots are visible after it has been rolled. That means we know that $A = \{2, 3, 4, 5, 6\}$ has occurred. As

$$P(A) = \frac{1}{12} + \frac{3}{12} + \frac{3}{12} + \frac{2}{12} + \frac{2}{12} = \frac{11}{12}$$

we then reassign probabilities to the remaining possible outcomes by dividing by $11/12$. That will leave probabilities

$$\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{2}{11}, \frac{2}{11}$$

for the outcomes $2, 3, 4, 5, 6$. If we compute $P(B|A)$ for $B = \{1, 2, 3\}$ we get the revised probability for $B \cap A = \{2, 3\}$ (since we know 1 has not happened). In summary, in this example,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{12} + \frac{3}{12}}{\frac{11}{12}} = \frac{4}{11}$$

An important concept is the idea of *independent events*, which means events $A, B \subset S$ with $P(B|A) = P(B)$. This is the same as

$$P(A \cap B) = P(A)P(B)$$

To get an example, imagine we have 20 balls in a hat of which 10 are blue and 10 are red. Suppose half (5) of the red ones and 5 of the blue have a white dot on them and the others have no dot. If a ball is drawn out at random (so each ball has the same probability $1/20$ of being drawn), you should be easily able to check that the events $A =$ a red ball and $B =$ a ball with a dot are independent.

4.5 The binomial distributions

Suppose we have a coin with probability p of turning up heads and probability $q = 1 - p$ of turning up tails (here $0 \leq p \leq 1$). Our experiment will now be to toss the coin a certain number n of times and record the **number of times heads shows up**. To the outcome will be a count between 0 and n, or the sample space will be $S = \{0, 1, 2, \ldots, n\}$. What probabilities should we assign to the points in this sample space?

[In fact the same formulae will apply to any random experiment with two outcomes repeated n times (provided the n times are independent). Instead of 'heads' and 'tails' we usually refer to 'successes' and 'failures', where the term "success" refers to the ones we count and "failure" to those we don't count. We need to divide the results into a class called "success' that happens with some probability p and then that class will fail to happen wit probability $1 - p$. Repeat n times and count.]

The idea of independent events comes in here because we assume in our analysis that each of the n times we toss the coin is independent. Thus a heads to start with does not make it any more or less likely that the second toss will show heads. We can then analyze that the probability of heads ($= H$, say for short) on the first toss should be p, whereas the probability of $T =$ tails should be q. And that is true each time we toss the coin. So the probability of the first 3 tosses showing up HTH in that order is

$$P(H)P(T)P(H) = pqp = p^2q$$

Now if $n = 3$, this is not the same as the probability of counting exactly 2 heads because there are two other ways to get that: THH and HHT. Each by themselves have probability p^2q but the probability of exactly 2 heads in 3 tosses is $3p^2q$. For $n = 3$, the outcomes $S = \{0, 1, 2, 3\}$ (numbers of heads) have probabilities

$$q^3, 3q^2p, 3qp^2, p^3$$

in that order. These add up to 1 because by the formula for the cube of a sum

$$q^3 + 3q^2p + 3qp^2 + p^3 = (q + p)^3 = (1 - p + p)^3 = 1$$

To explain the general case, we rely on the 'binomial theorem' which explains what the n^{th} power of a sum is:

4.5.1 Theorem (Binomial theorem)**.** *For any $n = 1, 2, \ldots$ and any two numbers p and q the expansion of $(p + q)^n$ works out as*

$$(p + q)^n = p^n + \binom{n}{1}p^{n-1}q + \binom{n}{2}p^{n-2}q^2 + \cdots + \binom{n}{n-1}pq^{n-1} + q^n$$

where the coefficients are given by

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

(and are known as binomial coefficients).

Using summation notation the formula can be rewritten

$$(p + q)^n = \sum_{i=0}^{n} \binom{n}{i} p^{n-i} q^i$$

or

$$(p+q)^n = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i}$$

Proof. This can be proved by induction on n.

We would need to recall the notion of a *factorial*. Recall that

$$n! = n(n-1)\cdots(2)(1)$$

is the product of the numbers $1, 2, \ldots, n$ (so that for instance $4! = 4(3)(2)(1) = 24$ and is 4 times $3! = 6$). However zero factorial is not covered by the general rule but we define $0! = 1$.

With these conventions

$$\binom{n}{0} = \frac{n!}{0!n!} = \frac{n!}{(1)n!} = 1$$

and

$$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = \frac{n!}{n!(1)} = 1.$$

For $n = 1$, the binomial theorem just says

$$(p+q)^1 = \binom{1}{0}p^1 + \binom{1}{1}q^1 = 1p + 1q$$

and that is true. In the summation version $\sum_{i=0}^{1} p^i q^{1-i}$ there is another convention that q^0 is to be interpreted as 1 and also $p^0 1$. (This is the right rule when $q \neq 0$ or $p \neq 0$ but normally 0^0 is **not** defined.)

An obvious thing from the definition of the binomial coefficients is the property that

$$\binom{n}{i} = \binom{n}{n-i} \qquad (0 \leq i \leq n)$$

In order to make the proof by induction work, one usually relies on the less obvious fact that

$$\binom{n+1}{i+1} = \binom{n}{i} + \binom{n}{i+1} \qquad (0 \leq i < n).$$

This can be verified by fairly simple manipulations

$$\begin{aligned}
\binom{n}{i} + \binom{n}{i+1} &= \frac{n!}{i!(n-i)!} + \frac{n!}{(i+1)!(n-i-1)!} \\
&= \frac{n!}{i!(n-i-1)!}\left(\frac{1}{n-i} + \frac{1}{i+1}\right) \\
&\qquad \text{using } (n-i)! = (n-i)(n-i-1)! \text{ and} (i+1)! = (i+1)i! \\
&= \frac{n!}{i!(n-i-1)!}\left(\frac{i+1+n-i}{(n-i)(i+1)}\right) \\
&= \frac{n!(n+1)}{(i+1)i!(n-i)(n-i-1)!} \\
&= \frac{(n+1)!}{(i+1)!(n-i)!} = \binom{n+1}{i+1}
\end{aligned}$$

This property is just what you need for the induction step, where you multiply the formula for $(p+q)^n$ by $p+q$ and try to reconcile that with the formula for $(p+q)^{n+1}$. □

4.5.2 Remark. The binomial coefficient $\binom{n}{i}$ is often read as "n choose i' because it give the number of ways of choosing i object from n, where the order is not important (or the number of subsets of size i in a set of size n). The explanation is that if you think of picking out the i things one at a time, then there are n choices for the first one, but then only $n-1$ left and so $n-1$ choices for the second. In total then there are $n(n-1)$ choices for the first two (if you keep track of the order in which you are picking them out). In total then to pick out i (keeping track of the order they are picked out in) there are

$$n(n-1)\cdots(n-i+1)$$

ways to do it. However if you don't care about the order, there are a total of

$$i(i-1)\cdots(2)(1)$$

ways to reorganize i elements. (To see that use the same sort of argument of picking them out one by one in order.) So each collection of i elements that is counted by the number $n(n-1)\cdots(n-i+1)$, each one counted $i! = i(i-1)\cdots(2)(1)$ times. So the number of different ways to pick out i elements from n where the order is not important is

$$\begin{aligned}\frac{n(n-1)\cdots(n-i+1)}{i(i-1)\cdots(2)(1)} &= \frac{n(n-1)\cdots(n-i+1)}{i!}\\ &= \frac{n(n-1)\cdots(n-i+1)(n-i)\cdots(2)(1)}{i!(n-i)(n-i-1)\cdots(2)(1)}\\ &= \frac{n!}{i!(n-i)!} = \binom{n}{i}\end{aligned}$$

This explains the 'n choose i' terminology.

Another fact that is very well known is the relationship between the binomial coefficients and 'Pascal's triangle'. That is

$$\begin{array}{ccccccccccc} &&&&&1&&&&& \\ &&&&1&&1&&&& \\ &&&1&&2&&1&&& \\ &&1&&3&&3&&1&& \\ &1&&4&&6&&4&&1& \\ 1&&5&&10&&10&&5&&1 \end{array}$$

where there are 1's along the edges and each other number is the sum of the two just above it. That property is exactly the

$$\binom{n+1}{i+1} = \binom{n}{i} + \binom{n}{i+1} \qquad (0 \le i < n).$$

property.

We return now to the probability question under discussion (named because of its relationship with the binomial theorem). We are considering a repetition n independent times of a random experiment where the probability of 'success' on each trial is p (and the probability of 'failure' each time is $q = 1 - p$) and we are counting the number of 'successes' in the n trials. The possible counts are $0, 1, 2, \ldots, n$.

In general (for any n) the appropriate probabilities in $S = \{0, 1, 2, \ldots, n\}$ are given by

$$P(i) = \binom{n}{i} p^i q^{n-i}$$

We could check using the binomial theorem that this is a valid assignment of probabilities (they are ≥ 0 and add up to 1). A counting argument is needed to relate these probabilities to the probabilities we mentioned for the number of heads.

4.5.3 Theorem (Properties of the binomial distributions)**.** *The binomial distribution (for the number of 'successes' in n independent trials where the probability of success is p on each trial) has mean*

$$\mu = np$$

and variance

$$\sigma^2 = npq = np(1 - p)$$

Proof. We will not verify (or prove) these but the formula for the mean is

$$\mu = \sum_{i=0}^{n} P(i)i = \sum_{i=0}^{n} i \binom{n}{i} p^i q^{n-i} = \sum_{i=1}^{n} i \frac{n!}{i!(n-i)!} p^i q^{n-i}$$

in this case and it is not so hard to simplify this to get np. The variance is

$$\sum_{i=0}^{n} P(i)(i - \mu)^2$$

which is slightly more tricky to simplify to npq. □

4.5.4 Example. A fair die is rolled 4 times. Find the probability of obtaining at least 3 sixes.

This exactly fits the scenario for the binomial distribution with $n = 4$ independent trials of the experiment of rolling the die, if we regard 'success' as rolling a 6. Then $p = 1/6$ and the probability we want is

$$P(3) + P(4) = \binom{4}{3} p^3 q^{4-3} + \binom{4}{4} p^4 q^{4-4} = 4\left(\frac{1}{6}\right)^3 \frac{5}{6} + \left(\frac{1}{6}\right)^4 = \frac{21}{6^4} = \frac{7}{432}$$

4.6 The Poisson distribution

This can be obtained as a limiting case of the binomial distributions where $n \to \infty$ but p is adjusted so that $\mu = np =$ constant.

The idea is that there are experiments where things are likely to happen fairly rarely. One classical story is about servicemen in the Prussian cavalry being fatally kicked by horses. Statistics from 10 different corps were collected over 20 years and there were 122 deaths. So that gives the average number of deaths in one corps in one year as $\frac{122}{200} = 0.61$. If you imagine dividing the year into seconds, there would be a very small chance of any fatality each second, but over a long period there is a noticeable probability of a fatality. The Poisson distribution was found to fit the observed data in terms of the numbers of year-corps in which there were $0, 1, 2, \ldots$ fatalities.

The sample space in this case is $S = \{0, 1, 2, \ldots\}$ (which is infinite) and

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

The number μ is a parameter in the Poisson distribution, which means there are many Poisson distributions — one for each choice of $\mu > 0$.

Using our knowledge of power series we can check that this is a valid assignment of probabilities (that is that they are ≥ 0 and sum to 1).

$$\sum_{n=0}^{\infty} P(n) = \sum_{n=0}^{\infty} \frac{\mu^n}{n!} e^{-\mu} = e^{-\mu} \sum_{n=0}^{\infty} \frac{\mu^n}{n!} = e^{-\mu} e^{\mu} = 1$$

It was observed in 1910 that the Poisson distribution provides a good model for the (random) number of alpha particles emitted per second in a radioactive process.

The mean of a Poisson distribution is

$$\begin{aligned} \sum_{n=0}^{\infty} nP(n) &= \sum_{n=1}^{\infty} n \frac{\mu^n}{n!} e^{-\mu} \\ &= \sum_{n=1}^{\infty} \frac{\mu^n}{(n-1)!} e^{-\mu} \\ &= \mu e^{-\mu} \sum_{n=1}^{\infty} \frac{\mu^{n-1}}{(n-1)!} \\ &= \mu e^{-\mu} e^{\mu} \\ &= \mu \end{aligned}$$

and so there is no confusion in using the symbol μ for the parameter in the Poisson distribution.

The variance also turns out to be $\sigma^2 = \mu$.

4.6.1 Example. If the number of alpha particles detected per second by a particular detector obeys a Poisson distribution with mean $\mu = 0.4$, what is the probability that at most 2 particles are detected in a given second?

The answer is $P(0) + P(1) + P(2)$ where $P(n)$ is given by

$$P(n) = \frac{(0.4)^n}{n!} e^{-0.4}$$

In other words

$$e^{-0.4} + (0.4)e^{-0.4} + \frac{(0.4)^2}{2} e^{-0.4} = 0.99207$$

4.7 Continuous probability densities

We now move on to look into the question of what happens when we have infinite sample spaces where each individual outcome has zero probability. We have seen one type of example already (experiments resulting in numerical measurements). For another example, consider a factory that fills 1 litre cartons of milk. Each carton produced will have somewhere near 1 litre of milk in it, but there is no chance of getting exactly 1 litre of milk into the carton in a mathematically precise sense of infinite precision. You might get 1.0 litres within 0.01 litres, but you cannot expect exactly 1 litre. Due to inherent inaccuracies in the machines, we can regard the amount of milk that goes into each carton as the value of a random variable with a continuous range of possible values and where each individual value will have probability zero.

We work with a *probability density function*, which is a function $f(x)$ with the characteristic property that the probability that we will get a value in the range $[x, x + dx)$ is $f(x)\,dx$ (when dx is very small or infinitesimally small). In summary

$$P([x, x + dx)) = f(x)\,dx$$

From the probability density function we can work out the probability of a result in a given range $[a, b]$ by integration.

$$P([a, b]) = \int_a^b f(x)\,dx$$

Since we want our probabilities to be always between 0 and 1 and we want the total probability to be 1, we need our probability density function to be always nonnegative and have integral 1 over the entire range of values. Thus any function with the two properties

(i) $f \colon \mathbb{R} \to [0, \infty)$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (that is an improper integral)

can be a probability density function

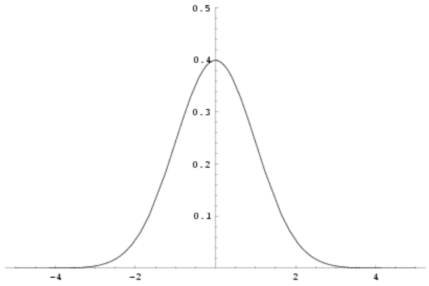
4.8 Normal probability density

One of the types of probability density functions that is most often used in practice is the *normal probability density* function. Actually there is a whole lot of different ones. There are two parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ that we get to choose to suit our problem and the normal density with mean μ and standard deviation σ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

A good case to consider is the case $\mu = 0$ and $\sigma = 1$, which is called the *standard normal density* function. It is

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$



4.9 Probability distribution functions

We use probability density functions to work out probabilities

$$P([a,b]) = \int_a^b f(x)\,dx$$

and we can work these out if we know the values of

$$F(b) = P((-\infty,b]) = \int_{-\infty}^b f(x)\,dx$$

because

$$P((-\infty,b]) = P((-\infty,a)) + P([a,b])$$

and so

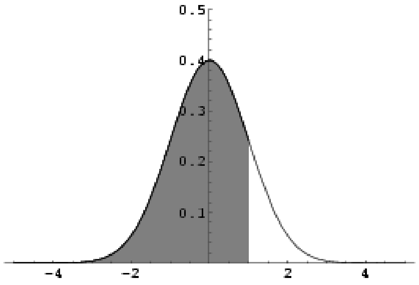
$$P([a,b]) = P((-\infty,b]) - P((-\infty,a)) = F(b) - F(a)$$

The *probability distribution function* associated with a probability density $f(x)$ is the function

$$F(x) = \int_{-\infty}^x f(t)\,dt$$

In the case of the standard normal, these integrals cannot be worked out explicitly except by using numerical integration and the values of the standard normal distribution function are tabulated in the log tables (look at pages 36–37).

Here is a picture for the standard normal distribution $F(1)$ as the area under the curve corresponding to the standard normal density function.



You will see that the tables only work for $x > 0$, but there is a symmetry to the picture that tells you that (for the standard normal case) $F(0) = P((-\infty, 0]) = 1/2$ and

$$F(-x) = P((-\infty, -x]) = P([x, \infty)) = 1 - P((-\infty, x]) = 1 - F(x).$$

From these rules plus the tables you can figure out all the values.

4.10 Mean and variance for continuous distributions

We will not go into this in any detail, but you can define the mean for a continuous random variable with density $f(x)$ to be

$$\text{mean} = \mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

(if the integral converges). You can also define the variance to be

$$\text{variance} = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

(again only when the integral converges).

One fortunate thing is that the mean and variance for a normal density with parameters μ and σ do turn out to be mean $= \mu$ and variance $= \sigma^2$. We will not check this out, but it is not so hard to do it. You can see that it would be confusing to call the parameters μ and σ if this did not work out.

4.10.1 Proposition (Relating normals to standard normals)**.** *Say*

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\, dt$$

is the normal distribution function with mean μ and variance σ and

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$$

is the standard normal distribution. One can show by a simple change of variables $u = (t-\mu)/\sigma$ that there is a relationship

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

between the two distribution functions.

It might be clearer to write $\Phi_{\mu.\sigma}$ in place of F and then $\Phi_{0,1}$ rather than Φ. Then the above can be rewritten

$$\Phi_{\mu,\sigma}(x) = \Phi_{0,1}\left(\frac{x-\mu}{\sigma}\right)$$

4.10.2 Remark. In this way we can relate all normal distribution functions to the standard normal that is tabulated in the log tables.

4.10.3 Example. Suppose a production line is producing cans of mineral water where the volume of water in each can produced can be thought of as (approximately) obeying a normal distribution with mean 500ml and standard deviation 0.5ml. What percentage of the cans will have more than 499ml in them?
Solution: We have

$$P(> 499 \text{ in a can}) = 1 - P(< 499) = 1 - \Phi_{\mu,\sigma}(499)$$

where $\Phi_{\mu,\sigma}(x)$ is the normal distribution function with mean $\mu = 500$ and standard deviation $\sigma = 0.5$. From the previous idea of relating normals to standard normals, we can say

$$1 - \Phi_{\mu,\sigma}(499) = 1 - \Phi_{0,1}\left(\frac{499-\mu}{\sigma}\right) = 1 - \Phi_{0,1}(\frac{499-500}{0.5}) = 1 - \Phi_{0,1}(-2).$$

From the symmetry properties of the standard normal distribution, we then have

$$\begin{aligned} 1 - \Phi_{\mu,\sigma}(499) &= 1 - \Phi_{0,1}(-2) \\ &= 1 - P(\text{standard normal} < -2) \\ &= P(\text{standard normal} > -2) \\ &= P(\text{standard normal} < 2) \end{aligned}$$

and from the tables this is 0.9772.

This is the proportion of cans that will have more than 499ml. Expressing the answer as a percentage we get 97.72%.

4.11 Confidence intervals

In a scientific context it is not that unusual to state results of measurements with an error that is intended to be interpreted in a probabilistic way. In engineering, on the other hand, it is more usual to give error bounds or tolerances. So an engineer might say that a rivet is 5mm in diameter with an error or ± 0.1mm, meaning that there is a certainty that the diameter is between $5 - .1 = 4.9$ and 5.1mm. However, in scientific data, the measurement analysis is likely to be more based on a claim that a large error is improbable (but not ruled out). For a scientist the statement that a result is 5 ± 0.1 (in mm, say) means that the probability distribution has a mean of 5 and a standard deviation of 0.1.

According the a theorem called the 'Central Limit Theorem' most things behave like a normal distribution in the long run. More precisely, suppose you sample a random variable n independent times, or think of n different random variables

$$X_1, X_2, \ldots, X_n$$

which are independent and 'identically distributed'. By identically distributed, the probability that X_i lies in any range $a \leq X_i \leq b$ is that same for all $i = 1, 2, \ldots, n$. By independent we mean that

$$P(a_1 \leq X_1 \leq b_1 \text{ and } a_2 \leq X_2 \leq b_2) = P(a_1 \leq X_1 \leq b_1)P(a_2 \leq X_2 \leq b_2)$$

and generalisations of that to more that two of the random variables.

It follows from them being identically distributed that all X_i have the same mean (which we call μ) and the same variance σ^2. It is a short step from that to say that

$$X_1 + X_2 + \cdots + X_n$$

has mean $n\mu$ and a bit trickier to show that it has variance $n\sigma^2$. So

$$Z = \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}}$$

will have mean 0 and variance (or standard deviation) 1.

The central limit theorem says that if n is large, then the probability that Z is in any given range (like $a \leq Z \leq b$) is close to the probability that a standard normal is in the same range.

So (more or less) no matter what the underlying probabilities are, Z will always be like a standard normal. For instance the probability

$$P(-1 < Z < 1) \cong P(-1 < Z_{0,1} < 1)$$

where $Z_{0,1}$ means a standard normal. We have

$$\begin{aligned} P(-1 < Z_{0,1} < 1) &= P(Z_{0,1} < 1) - P(Z_{0,1} < -1) \\ &= P(Z_{0,1} < 1) - P(Z_{0,1} > 1) \\ &= P(Z_{0,1} < 1) - (1 - P(Z_{0,1} < 1)) \\ &= 2P(Z_{0,1} < 1) - 1 \\ &= 2\Phi_{0,1}(1) - 1 = 2(0.8413) - 1 = .6826 \end{aligned}$$

If we extrapolate from the Central Limit Theorem to apply this with $n = 1$ or just assume that X is described by the normal distribution with mean μ and variance σ^2, then we could say that

$$P\left(-1 < \frac{X - \mu}{\sigma} < 1\right) \cong .68$$

and that is the same as

$$P(\mu - \sigma < X < \mu + \sigma) \cong 0.68$$

so that X has a 68% chance for being in the range $\mu \pm \sigma$.

If we do this with $\pi \pm 2\sigma$ instead we get 0.95 (or 95% chance) and with $\pi \pm 3\sigma$ we get 99.7%.

So this is the idea of confidence intervals (stated in terms of the normal distribution case).

4.12 Student's t-distribution

Suppose X is a normal random variable but we don't know the mean μ or the standard deviation σ. Then we can aim to discover μ by experiment.

We would like to be able to say something about how confident we can be about μ based on a sample.

What we can do is take n samples and say we get values

$$x_1, x_2, \ldots, x_n.$$

We can calculate the sample mean

$$m = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

and the sample variance — which we will denote by

$$s^2 = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2}{n - 1}$$

William Sealy Gosset (1876–1937) was a statistician working in the Guinness brewery in Dublin and he published under then pseudonym "student". He worked out that if

$$X_1, X_2, \ldots, X_n$$

are normal random variables which are independent and identically distributed with true mean μ, then the distribution of

$$\frac{m - \mu}{s/\sqrt{n}}$$

with

$$m = \frac{X_1 + X_2 + \cdots + X_n}{n}, \qquad s^2 = \frac{(X_1 - m)^2 + (X_2 - m)^2 + \cdots + (X_n - m)^2}{n - 1}$$

could be calculated. It is a symmetric distribution with mean 0 and is known as the t-distribution with $n - 1$ degrees of freedom. Let's call that distribution F (though it is not to be confused with the normal distribution and depends on the sample size n, or really on $n - 1$, the so-called 'number of degrees of freedom'). The values of F can be found in the tables (pages 40–41) but the numbers there are actually $1 - F$.

Say we want to give a confidence interval for the true value of μ. So we want to find a number k so that if μ is not likely to be outside the range $m \pm k$.

Step 1. Choose the confidence level γ we want (say $\gamma = .95$ or $\gamma = .99$)

Step 2. From the tables we choose c so that

$$F(c) - F(-c) = P(\text{range } [-c, c]) = \gamma$$

Because the t-distribution is symmetric $F(-c) = 1 - F(c)$ and so we want c so that

$$2F(c) - 1 = \gamma$$

or

$$F(c) = \frac{1+\gamma}{2}$$

or

$$1 - F(c) = \frac{1-\gamma}{2}$$

Step 3. Calculate the sample mean m and sample variance s^2.

Step 4. Then with $k = sc/\sqrt{n}$ we get probability γ that μ is really in the range $m - k$ to m_k.

4.12.1 Example. Suppose these are 6 samples

1.16	1.12	1.09	0.98	1.28	1.25

from a normal distribution with mean μ. Use Student's t-distribution to give a symmetric 99% confidence interval for μ.
Solution:

Step 1. We have $\gamma = 0.99$.

Step 2. Then

$$\frac{1+\gamma}{2} = \frac{1.99}{2} = 0.995$$

For F the student t-distribution with $6 - 1 = 5$ degrees of freedom we find from the tables that $F(4.032) = 0.995$ (or actually that $1 - F(4.032) = 0.005$).

Step 3.

$$m = \frac{1.16 + 1.12 + 1.09 + 0.98 + 1.28 + 1.25}{6} = 1.14667$$

$$s^2 = \frac{(1.16-m)^2 + (1.12-m)^2 + (1.09-m)^2 + (0.98-m)^2 + (1.28-m)^2 + (1.25-m)^2}{5}$$

works out as 0.0120667. So $s = \sqrt{0.0120667} = 0.109848$.

Step 4. Then with $k = sc/\sqrt{n} = (0.109848)(4.032)/\sqrt{6} = 0.180816$ we get probability 0.99 that μ is in the range $m \pm 0.180816 = 1.14667 \pm 0.180726$.

So in the interval $[0.965854, 1.32749]$ with confidence 0.99 (or 99%).

(If we went for 95% confidence instead the number 4.032 would change to 2.571 and k would become 0.115297. The interval would be $[1.03137, 1.26197]$.)

Richard M. Timoney March 25, 2014