Random variables and Probability distributions

A random variable is a variable whose value depends on the outcome of a random event/experiment.

For example, the score on the roll of a die, the height of a randomly selected individual from a given population, the income of a randomly selected individual, the number of cars passing a given point in an hour, etc. Random variables may be discrete or continuous.

Associated with any random variable is its probability distribution that allows us to calculate the probability that the variable will take different values or ranges of values.

Probability distributions for discrete variables

The probability distribution of a discrete r.v. X is given by a *probability function* f(X), which gives the probabilities for each possible value of X, and the range of possible values.

$f(X)$ = (some function) for $X = X_1, X_2, \ldots, X_n$
$f(X)=0$ for all other values of X

Where $f(X_i) = P(X=X_i)$

or (if, for example, X can take any positive integer value)

$f(X) = F(n)$ for $X=n$, a positive integer, where F is some function
$f(X) = 0$ for other values of X

This function must satisfy $f(X)>=0$ for all values of X, and $\sum f(X) |$all values of $X = 1$.

E.g., let X depend on the toss of a fair coin, with X=1 if the coin lands heads, 0 if tails. Then

$f(X) = 0.5$, for $X=0,1$
$f(X) = 0$ otherwise

Example 2: we toss a fair coin until it comes up heads. Let X be the number of times we toss the coin. It is fairly easy to show that

$f(X)=0.5^n$ for $X=n$, for $n=1,2,3,\ldots.$

f(X)=0 otherwise

(Since there is a probability of 0.5 of getting heads first time, a probability of 0.5*0.5=0.25 of getting tails first, then heads second time, probability 0.5*0.5*0.5 of getting T,T,H, etc.)

<u>Expected values</u>

Expected values are descriptive measures indicating characteristic properties of probability distributions. The expected value of a r.v. can be seen as the 'average' value – not in the sense of the average of a sample, but the 'theoretical' average we'd expect if we repeated the experiment a large number of times. (The average of a theoretical model, rather than a set of observations).

The <u>Expected Value</u> of a discrete r.v. X is defined as follows

Let S be the set of possible values that X can take

Then E(X) = $\sum_{X_i \in S} X_i f(X_i)$

E.g. suppose X is the score on the roll of a fair die, so f(X)=1/6 for X=1,2,3,4,5,6, 0 otherwise; then

E(X)=1*(1/6)+2*(1/6)+…+6*(1/6) = 21/6 = 3.5

E.g. 2: Let f(X) = 0.5^n for X=n, n=1,2,3,…, f(X)=0 otherwise; then

E(X) = $\Sigma \sum_{n=1}^{\infty} n.0.5^n$

Let E(X) = $1*0.5 + 2*0.5^2 + 3*0.5^3 + \ldots$

Then 0.5E(X) = $\quad 1*0.5^2 + 2*0.5^3 + \ldots$

So E(X)-0.5E(X) = $0.5+0.5^2+0.5^3 + \ldots.$

But we know from previous work (intro maths) that the r.h.s. of this equation is equal to 1.

So E(X)-0.5E(X) = 0.5E(X) = 1

Hence E(X)=2.

Other expected values can be readily defined (and are highly valuable in statistical analysis). E.g. $E(X^2) = \sum_{X_i \in S} X_i^2 f(X_i)$

Expected value operations

a) For a constant **a**, E(a) = $\sum_{X_i \in S} X_i f(X_i) = af(a)$ = a as f(X)=1 for X=a, 0 otherwise. So E(a) = a.
b) For a r.v. X and a constant a, E(aX) = $\sum_{X_i \in S} aX_i f(X_i) = a \sum_{X_i \in S} X_i f(X_i)$ = aE(X)
c) For two functions of X, g(X) and h(X), E[g(X)+h(X)] = $\sum_{X_i \in S}$ $[g(X_i)+h(X_i)]f(X_i)$ = $\sum_{X_i \in S}$ $g(X_i)f(X_i)$ + $\sum_{X_i \in S}$ $h(X_i)f(X_i)$ = E[g(X)] + E[h(X)]
d) For two r.v.s X and Y, E(X+Y) = E(X) + E(Y) (not so immediately easy to prove).

Variance of X. By analogy of the definition of the variance for an observed frequency distribution, we have

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

That is, Var(X) is the average value of the squared deviation of X from its mean.

We can use expected value operations to get an alternative expression for Var(X):

$Var(X) = E\{[X-E(X)]^2\}=E\{X^2 - 2E(X)X + [E(X)]^2\} = E(X^2) – E[2E(X)X] + E\{[E(X)]^2\}$

$= E(X^2) – 2E(X)E(X) + [E(X)]^2$ (since E[X] is a constant)

$= E(X^2) – 2[E(X)]^2 + [E(X)]^2$

$\mathbf{= E(X^2) – [E(X)]^2}$

That is, Var(X) is the "mean of the square minus the square of the mean".

E.g. let X be the score on the roll of a fair die. So f(X) = 1/6 for X =1,2,3,4,5,6, f(X)=0 otherwise. We know that E(X)=3.5.

Now $E(X^2) = 1^2*f(1) + 2^2*f(2) + \ldots + 6^2*f(6)$

$= (1/6)*(1+4+9+16+25+36) = 91/6.$

Linear function of X. If Y = a +bX where a and b are constants, we have

$E(Y) = E(a+bX) = E(a) + E(bX) = a + bE(X)$

$Var(Y) = E(Y^2) – [E(Y)]^2 = E(a^2 + 2abX + b^2X^2) – [a+bE(X)]^2$

$= (a^2 + 2abE(X) + b^2E(X^2) – [a^2 + 2abE(X) + b^2E(X)^2]$

$= b^2E(X^2) - b^2E(X)^2 = b^2[E(X^2) – E(X)^2]$

$= b^2Var(X)$

Note that the constant disappears when calculating the variance (as constants have no variance!), while the linear multiple of X is *squared*.

Exercise: If Y = a – bX, show that E(Y) = a – bE(X), and Var(Y) = b^2Var(X).

Probability distributions for continuous random variables

When dealing with a continuous r.v., it is not generally meaningful to talk of the probability of attaining any particular value. It is like asking what is the probability that a golf ball will land on a particular blade of grass. Instead, for a continuous r.v. X, we define a probability density function (pdf) f(X), which gives the *relative probability* of different values. We can meaningfully talk only of the actual probability of achieving certain *ranges* of values. For example, if X is the height of a random individual, then we don't talk of the probability of someone being *exactly* 5'8", but we can meaningfully talk of the probability of X lying between 5'7.5" and 5'8.5".

The probability density function f(X) is a function that assigns a non-negative value to each real number. If X can take only values between say, a and b, then f(X) will take the form

$f(X) = $ (some function) $a<X<b$
$f(X) = 0$ otherwise

The probability of any given range of values of X is given by the <u>area under the curve</u> of f(X) for that range of values. That is

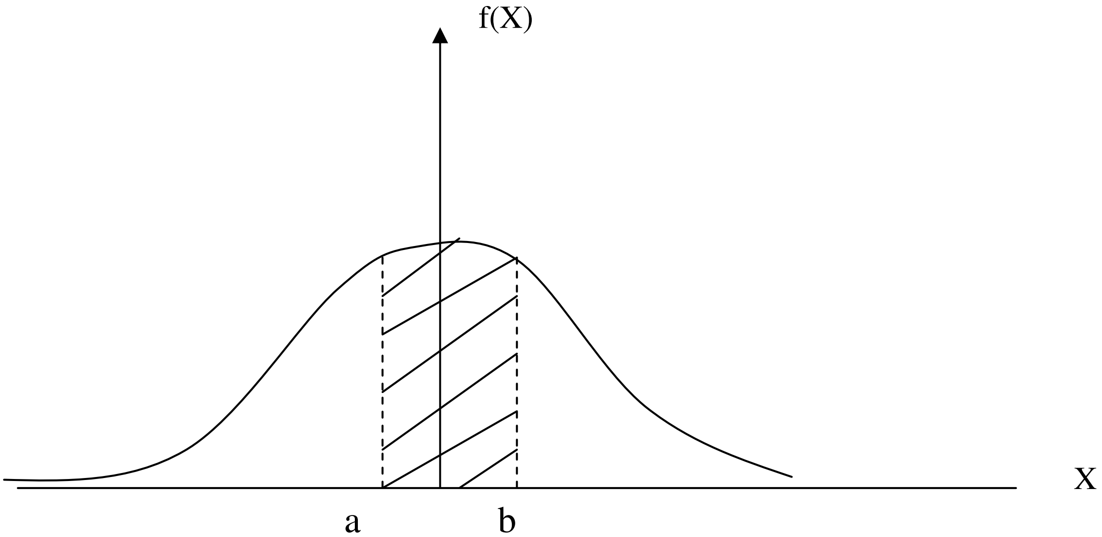
$$P(c<X<d) = \int_{X=c}^{X=d} f(X)\,dX$$

Note that since the total probability of all possible values must be 1, we must have that

$$\int_{-\infty}^{\infty} f(X)\,dX = 1$$

And in particular if X can only take values between a and b, then

$$\int_{a}^{b} f(X)\,dX = 1.$$



The above graph shows an example pdf, f(X). The total area between the curve and the X-axis, stretching to infinity in each direction, must total one. The shaded area shows P(a<X<b).

<u>Expected values</u> are defined analogously to the discrete case:

$$E(X) = \int_{-\infty}^{\infty} Xf(X)dX$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(X)f(X)dX$$

The rules for manipulating expectations are the same as those for discrete variables given in section 3.2. Likewise, we define

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

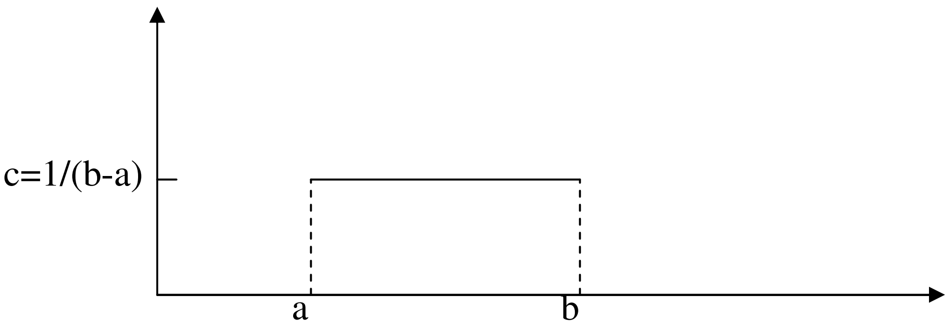
<u>Example</u>; The rectangular distribution is given by

$f(X) = c$, $a<X<b$
$f(X) = 0$ otherwise

That is, all values between a and b are equally likely, and no other value is possible.

Now the total area under the curve of f(X) here is a rectangle of height c and width (b-a), so the area is c(b-a). We know this area must total 1, and hence

$c = 1/(b-a)$.



We may now calculate E(X)

$$E(X) = \int_{a}^{b} \frac{X}{b-a} dX$$

We have not done integration. The result is $E(X) = (b^2-a^2)/2(b-a) = (b+a)/2$, that is, the average of a and b, or the midpoint of the range of values X can take.

Now $E(X^2) = \int_a^b \frac{X^2}{b-a}dX = \frac{1}{b-a}\int_a^b X^2 dX$

$= (b^3-a^3)/3(b-a) = (b^2+ab+a^2)/3$

Hence $\mathrm{Var}(X) = (b^2+ab+a^2)/3 - [(b+a)/2]^2 = (b-a)^2/12$

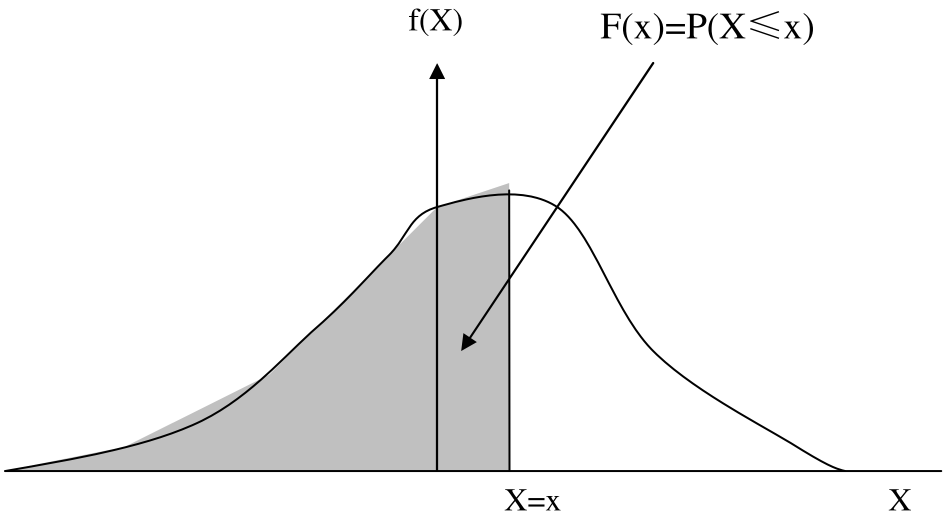
<u>Cumulative distributions</u>

For a discrete or continuous random variable X, we are often interested in the *cumulative* probability distribution for X, that is the probability that X will be *less than or equal to* any given value.

If X is a discrete r.v. with probability distribution f(X), taking only integer values 0,1,2,3,…, we may define the cumulative probability function $F(x) = P(X \leqslant x)$ by $F(x) = \sum_{n=0}^{x} f(n)$. It is easy to see that $f(x)=F(x)-F(x-1)$.

For a continuous r.v. X with pdf f(X), we may define the cumulative pdf $F(x)=P(X \leqslant x)$ by $F(x) = \int_{X=-\infty}^{X=x} f(X)dX$

It is easy to show that $dF/dx = f(x)$. (see diagram)



Joint (bivariate) probability distributions

These distributions arise when we consider the values taken by two random variables arising from the same event(s).

The joint probability distribution for two r.v.s X and Y is some function f(X,Y), where $f(X_i,Y_j) = P(X=X_i \text{ and } Y=Y_j)$.

Hence, we must have f(X,Y)>=0 for all values of X and Y, and the sum of f(X,Y) over all possible values must total one.

For two *continuous* r.v.s X and Y, we can define the joint probability density function f(X,Y), a non-negative function such that the *double integral* of f(X,Y) over all values of X and Y is equal to one.

$$\int_{X=-\infty}^{X=\infty} \int_{Y=-\infty}^{Y=\infty} f(X,Y)\,dY\,dX = 1$$

Example. Let X = no. of people in a house and Y = no. of rooms in a house, for a group of households. The joint probabilities are shown in a table:

Y/X	**1**	**2**	**3**	**4**	*Total*
1	.12	.08	.05	0	*.25*
2	.1	.15	.15	.1	*.5*
3	.05	.05	.1	.05	*.25*

Total	*.27*	*.28*	*.3*	*.15*	**1.00**

Note that the probabilities sum to 1.

Example: Let X be the height of a randomly selected individual from a given group, and Y be that individual's weight. We will not attempt to define a pdf here.

Marginal probability distributions are obtained by ignoring one variable, and looking at the total probability (or pdf) for the other. In the above table, the marginal distribution for X is given by the numbers in the Total row at the bottom, while the marginal distribution for Y is given by the figures in the right-hand column.

For continuous variables, the marginal pdf for X, $f_X(X)$ is given by

$$\int_{Y=-\infty}^{Y=\infty} f(X,Y)dY$$

That is, for each possible value of X, we look at the graph of f(X,Y) against Y, and take the total area under this curve.

Similarly, the marginal pdf for Y, $f_Y(Y) = \int_{X=-\infty}^{X=\infty} f(X,Y)dX$

Example

Let X be a random variable taking values between a and b, and Y a r.v. taking values between c and d. Let f(X,Y) = 1/(b-a)(d-c) for $a<X<b$ and $c<Y<d$, f(X,Y)=0 otherwise. (A uniform distribution, with all possible values equally likely).

Then the marginal distribution for X, $f_X(X)$ =

$$\int_{Y=c}^{d} \frac{1}{(b-a)(d-c)}dY = \frac{d-c}{(b-a)(d-c)} = \frac{1}{b-a}$$

Similarly, we can easily show that $f_Y(Y)$ = 1/(d-c). This is as it should be, as it means the total probability for the marginal distribution of X adds up to (b-a)/(b-a)=1, and the same for the marginal distribution of Y.

Independence and dependence

R.v.s X and Y are said to be independent if f(X,Y) = f(X)f(Y) for all values of X and Y. (This definition applies to both discrete and continuous variables).

In the above table, the discrete variables X and Y are not independent. For example, f(1,1)=0.12, but f(1)f(1)=0.27*0.25 = 0.0675. (In other words, the combination of one room and one person is more likely than if they were independent – there being one room increases the probability of there being only one person.)

In the example of height and weight of a given individual, we would not expect these two (continuous) variables to be independent. We would expect the two variables to tend to be high or low together. For example, we would expect f(high weight, high height) to be greater than f(high weight)f(high height), but f(low weight, high height) to be less than f(low weight)f(high height).

Expected values for joint distributions

We can define expected values for combinations of two r.v.s X and Y, e.g. E(X+Y), E(XY), etc. For example, for a discrete distribution, suppose X can take values X_i in some set S, and Y can take values Y_j in some set T, then

E(XY) = $\sum_{X_i \in S} \sum_{Y_j \in T} X_i Y_j f(X_i, Y_j)$

For a continuous distribution,

E(XY) = $\int_{X=-\infty}^{\infty} \int_{Y=-\infty}^{\infty} XYf(X,Y)dYdX$

For E(X+Y), replace XY in the above equations with X+Y, etc.

We quote several important results without proof:

E(X+Y) = E(X) + E(Y)
E(X-Y) = E(X) – E(Y)

For *independent* variables, we have similar results for variances:

Var(X+Y) = Var(X) + Var(Y)
Var(X-Y) = Var(X) + Var(Y) *Why are these both positive?*

For *dependent* variables, the result is not so simple. We need measures of to what extent X and Y depend on each other. Later we will define measures such as the correlation coefficient. For now, we will define the covariance of X and Y,

Cov(X,Y) = E{[X-E(X)][Y-E(Y)]} = E(XY) – E(X)E(Y)

Going back to variances, we obtain the results

Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)
Var(X-Y) = Var(X) + Var(Y) – 2Cov(X,Y)

Let us suppose that X and Y tend to go together, that is tend to be both high or both low. Then we will usually get X-E(X) and Y-E(Y) to be either both positive or both negative, so their product will on average be positive, giving a positive Covariance. If X and Y tend to go in opposite directions, then we will usually get one of X-E(X) and Y-E(Y) to be positive, and the other negative, so that their product is negative, giving a negative Covariance. If X and Y are independent, then [X-E(X)] and [Y-E(Y)] are equally likely to be positive and negative in either combination, so the Covariance will be zero. In fact we can define independence in this way.

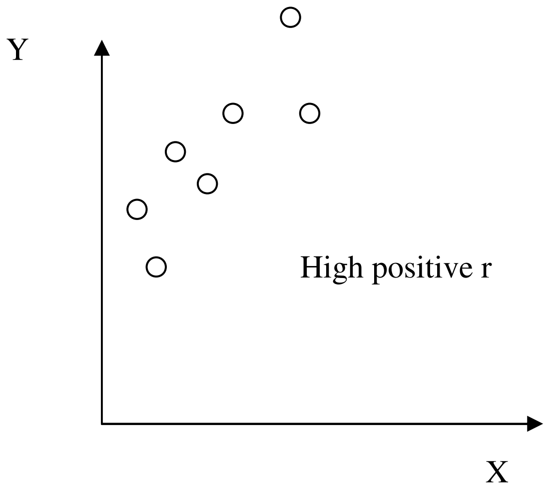
Correlation coefficient

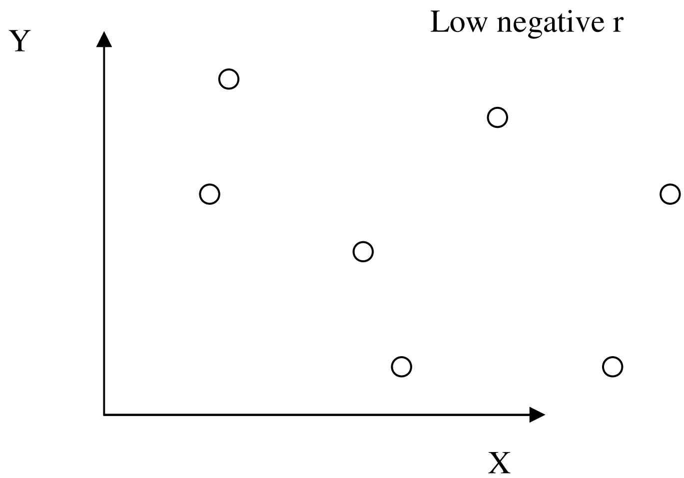
For two random variables X and Y, the correlation coefficient, r, is defined as

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

The correlation coefficient always lies between –1 and 1. If r=1, this means perfect positive correlation – it means that Y has an exact positive linear relationship with X. If r=-1, we have perfect negative correlation.

The nearer r is to 1 or –1, the closer the relationship between X and Y. The closer to 0, the weaker the relationship. (See graphs below)





It should be noted that correlation between two variables does not necessarily mean causation between them.

The Binomial distribution

The binomial distribution is probably the most important probability distribution for discrete variables. It arises as follows.

We perform an experiment (trial) that can have two results: success (S) or failure (F). We repeat the trial n times under exactly the same conditions. The results of the trials are independent of each other, and in each case, the probability of success is P. Such trials are called Bernoulli trials.

Let X be the number of successes from n trials. Then X has the binomial distribution with parameters (n,P). The binomial distribution is given by:

f(X) = ${}^{n}C_{X}\,P^{X}\,(1-P)^{n-X}$, for X = 0,1,2,…,n, f(X) = 0 otherwise.

Where ${}^{n}C_{X}$ is the expression for the number of combinations of X from n, that is the number of ways of choosing X objects out of n, where order is irrelevant. ${}^{n}C_{X}$ = n!/(X!(n-X)!)

Why? Well, we can obtain X successes from n trials in ${}^{n}C_{X}$ different ways, each of which is a mutually exclusive sequence of successes and failures. Therefore the probability of getting one of them is equal to the sum of their individual probabilities. What is the probability of any given sequence of successes and failures? E.g., if n=5, what is the probability of getting S, S, F, S, F? Since the trials are independent, we may multiply together the individual probabilities of the result of each trial, so P(SSFSF) = P(S)P(S)P(F)P(S)P(F) = P*P*(1-P)*P*(1-P) = $P^3(1-P)^2$

In general, the probability of any sequence with X successes and n-X failures is $P^X(1-P)^{n-X}$, and there are ${}^{n}C_{X}$ of them, giving the distribution for f(X) shown above.

It can be shown that E(X) = nP, Var(X) = nP(1-P), and $\sum_{X=0}^{n} f(X) = 1$.

Example Suppose we toss a fair coin 5 times and let X be the number of heads. (Successes). Then X has binomial distribution with n=5, P=0.5. We write X~B(5,0.5)

We can calculate f(X) for X=0,1,…,5. So, f(0) = ${}^{5}C_{0}(0.5)^{0}(1-0.5)^{5-0}$

= [5!/(5!*0!)]*0.5^5 = 1/32 (since 0! Is defined to be equal to 1 – the number of ways of arranging 0 objects.)

f(1) = ${}^{5}C_{1}(0.5)^{1}(0.5)^{5-1}$ = [5!/(4!*1!)]*0.5^5 = 5/32

Similarly, f(2)=${}^{5}C_{2}$/32 = [(5*4)/(2*1)]/32 = 10/32
Also f(3)=10/32
f(4)=5/32
f(5) = 1/32.

It is not hard to show that if P=0.5, then f(X)=f(n-X) – since getting X successes is the same as getting n-X failures, and successes and failures are equally likely.

The binomial distribution is positively skewed when P<0.5, and negatively skewed when P>0.5. It is symmetrical when P=0.5. When n is large, the binomial distribution approximates to the normal distribution (see below), irrespective of the value of P.

Tables of the individual or cumulative probabilities of the binomial distribution are included in most text books and collections of statistical tables.

In an actual series of Bernoulli trials, we define the sample proportion to be p=X/n. Thus p is itself a random variable. (It is discrete, even though it takes non-integer values, since it can only take fractions denominated by n). We have

$E(p)=E(X/n)=(1/n)E(X)=nP/n=P$.

So *on average* the sample proportion will be equal to the actual success probability;

$Var(p) = Var(X/n) = Var(X)/n^2 = nP(1-P)/n^2 = P(1-P)/n$

Hence, the standard deviation of p, SD(p)= $\sqrt{\dfrac{P(1-P)}{n}}$

E.g., if P=0.5, then E(p) = 0.5, and Var(p) = 0.25/n, SD(p)= $\dfrac{0.5}{\sqrt{n}}$.

Thus, the variance (and SD) of p diminishes as n increases, in other words, the more trials we conduct, the closer the sample proportion is likely to be to the actual probability.

The normal distribution

The normal distribution is of fundamental importance in statistical analysis. Many continuous variables are distributed normally (e.g. height, weight, very often test scores). It approximates some observed distributions, and it arises frequently in sampling problems.

What is more, if we start with *any* probability distribution for a random variable X (within certain conditions), and take the *average* value of X over a large number of repeated trials, then the distribution of this average value approximates a normal distribution as n gets large. This is true, for example, of the binomial distribution, as mentioned above. This makes the normal distribution extremely important.

The normal distribution is defined by the probability density function:

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right), \text{ for } -\infty<X\infty.$$

(Exp(x) is the same as e^x).

Here, μ and σ are the parameters of the distribution. We write $X \sim N(\mu,\sigma^2)$.

The normal distribution has the famous "bell curve" shape. It can be shown that

(a) E(X) = μ i.e. μ is the arithmetic mean
(b) Var(X) = σ^2, i.e. σ is the standard deviation.
(c) It is symmetric, with the mean μ also the median and the mode of the distribution.
(d) Since the distribution is symmetrical, f(μ+a)=f(μ-a) for any constant a.

Area under the normal curve

We know $P(a<X<b)=\int_a^b f(X)dX$, the area under the curve between a and b. Unfortunately, this integral is not easy to find, that is, there is no simple function whose differential is f(X). Fortunately, there are standard tables of the cumulative probability density function of the standard normal distribution – the normal distribution with μ=0 and σ^2=1.

Also fortunately, the cumulative distribution of any normal distribution can easily be calculated in terms of the standard normal distribution. What we must do is to express the value of a normal variable in terms of the *number of standard deviations from the mean.*

That is, if $X \sim N(\mu,\sigma^2)$, then it can be shown that

$P(X \leqslant X_0) = P(z \leqslant (X_0-\mu)/\sigma)$ where z is a standard normal variable, $z \sim N(0,1)$.

For example, suppose $X \sim N(10,4)$. (That is μ=10, σ^2=4). Then

$P(X \leqslant 14) = P(z \leqslant (14-10)/2) = P(z \leqslant 2)$ where z is a standard normal variable. Thus, we can look up $P(Z \leqslant 2)$ in a table of the standard normal distribution, and we have our answer.

Some additional simple rules will help us:

i) $P(a<X<b) = P(X<b) - P(X<a)$ (This applies to any distribution)
ii) $P(X>a) = 1-P(X<a)$
iii) If $Z_0<0$, then $P(z<Z_0) = 1-P(z>Z_0) = 1-P(z<-Z_0)$ (see graph on whiteboard).

For example, let $X \sim N(10,2^2)$. What is $P(7<X<11)$?

$P(7<X<11) = P(X<11) – P(X<7)$

$= P(z<\frac{11-10}{2}) – P(z<\frac{7-10}{2}) = P(z<0.5) – P(z<-1.5)$, where $z \sim N(0,1)$

$= P(z<0.5) – (1-P(z<1.5))$ by rule 3

$= P(z<0.5) + P(z<1.5) – 1$.

We can now look these last two probabilities up from a table of the cumulative standard normal distribution.

A couple of useful facts

$P(-1<z<1) \approx 0.68$

$P(-1.96<z<1.96) \approx 0.95$

Where $z \sim N(0,1)$

In other words, roughly 95% of observations from a normal distribution are within two standard deviations of the mean.

Linear transformation

If $X \sim N(\mu,\sigma^2)$ and $Y = a+bX$ where a and b are constants, it can be shown that $Y \sim N(a+b\mu,b^2\sigma^2)$.

Reproductive property

This is another property of the normal distribution which is important in sampling theory. It states that:

If X_1~$N(\mu_1,\sigma^2{}_1)$ and X_2~$N(\mu 2,\sigma^2{}_2)$

And if X_1 and X_2 are independent, then

(X_1+X_2)~$N(\mu_1+\mu_2,\sigma_1{}^2+\sigma_2{}^2)$ and

(X_1-X_2)~$N(\mu_1-\mu_2,\sigma_1{}^2+\sigma_2{}^2)$.

This is important because it means that if a random quantity is made up from adding together a lot of independent different factors, each of which are themselves normally distributed, then the result will also be normally distributed.

Example

Suppose we know that male weekly earnings are normally distributed with mean £300 per week and variance 100^2, and female earnings have mean £240 and variance 80^2. What will be the probability distribution of the *difference* between a randomly selected man and a randomly selected woman?

By the above property, the distribution will be normal, with mean £60 (in the man's favour) and variance $100^2 + 80^2 \approx 128^2$. In other words, on average the man's income will be £60 higher, as we'd expect, but the high standard deviation of £128 means that quite often the woman's income will be higher.