A Tutorial on Statistical Distributions
(for Economics Part I Paper 3)

Corrections to Dr Ian Rudy (http://www.robinson.cam.ac.uk/iar1/contact.html) please.

This document is intended to guide you on the main statistical distributions you will meet on the course, and which distribution you should use in which situation. It assumes you already understand the basic idea of discrete and continuous distributions, and the basic general idea of a confidence interval and a hypothesis test.

1. The Key Results

1.1 Discrete (Specifically, Binary) Variables

If individuals in an infinitely large population can be regarded as either having (with probability *p*) some characteristic of interest, or not having it, then the number of individuals in a sample of size *n* who have the characteristic will follow a Binomial distribution. If *n* is sufficiently large to obey $np(1\text{-}p) > 10$ then we can use the Normal approximation to the Binomial, with mean *np* and variance $np(1\text{-}p)$.

If we wish to calculate a confidence interval for p, based on a sample of size *n* in which a fraction $\hat{p}$ of individuals have the characteristic of interest, then if $n\hat{p}(1-\hat{p}) > 10$ we should use $\hat{p} \pm Z\sqrt{\hat{p}(1-\hat{p})/n}$ where *Z* is a value from the Normal distribution.

If we wish to compare two samples to test a null hypothesis that they come from the same population, and $n_1\hat{p}_1(1-\hat{p}_1) > 10$ and $n_2\hat{p}_2(1-\hat{p}_2) > 10$, then we should compare the following statistic:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with a suitable critical value from a Normal distribution. Here, $\hat{p}_i$ is the fraction of individuals in sample *i* (*i*=1 or 2) who have the characteristic, n_i is the size of sample *i*, and:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

1.2 Continuous Variables

Let us say that we take a sample of values x_i (*i*=1 to *n*) from an infinitely large population, where each x_i can take any value, or any value in a wide range. Then:

1.2.1 If we seek to use the sample mean $\bar{x}$ to estimate the population mean μ, or to do a hypothesis test that μ has some specific value:

1.2.1.1 If we know that individuals in the population follow a Normal distribution and we have been given the variance of individuals in the population (σ^2) precisely then we should use a Normal distribution.

1.2.1.2 If the sample size *n* is sufficiently large (say, 30 or more) and the distribution of individuals in the population is not extreme in respect of spikiness or outliers then we should use a Normal distribution.

1.2.1.3 If *n* is not large, but we know that individuals in the population follow a Normal distribution, we should estimate σ^2 from the sample, and use a t-distribution with *n*-1 degrees of freedom.

1.2.2 If we seek to estimate the variance of individuals in the population (σ^2) using the quantity $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$, or to do a hypothesis test that σ^2 has some specific value, and we know that individuals in the population follow a Normal distribution, then we should use a χ^2 distribution with *n*-1 degrees of freedom.

1.2.3 If we seek to test whether two different samples come from populations with the same variance, and we know that individuals in the population follow a Normal distribution, we should use the F-distribution to determine if $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$ (put the larger value on the top) is sufficiently greater than 1. For a two-tailed test with a significance level of α, the critical value of F is that which has an area $\alpha/2$ in the right hand tail.

1.2.4 If we seek to test whether two different samples come from populations with the same mean, and we know that individuals in the populations follow Normal distribution *with the same variance*, then we should compare the following statistic:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

with a suitable critical value from a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Here, $\bar{x}_i$ is the mean for sample *i* (*i*=1 or 2), n_i is the size of sample *i*, and:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

2. What Lies Behind the Key Results

This section discusses the theory behind the key results above, taking the distributions one by one.

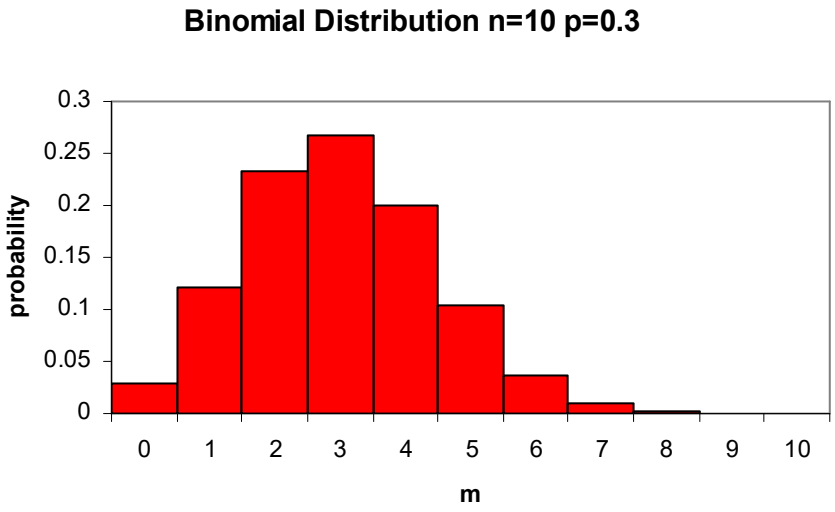
2.1 Binomial Distribution

The Binomial distribution is a discrete distribution that arises when we carry out a series of trials, each of which has only two possible outcomes. The classic example would be throwing up a coin, which can come down either heads or tails. We often denote the two outcomes as "success" and "failure", though it is often arbitrary as to which is which. If the probability of a success in a single trial is p then the probability of m successes in n trials is:

$$ {}^{n}C_{m}\, p^{m} (1-p)^{n-m} $$

where:

$$ {}^{n}C_{m} = \binom{n}{m} = \frac{n!}{m!(n-m)!} $$



The mean of the Binomial distribution is np.
The variance of the Binomial distribution is $np(1-p)$.

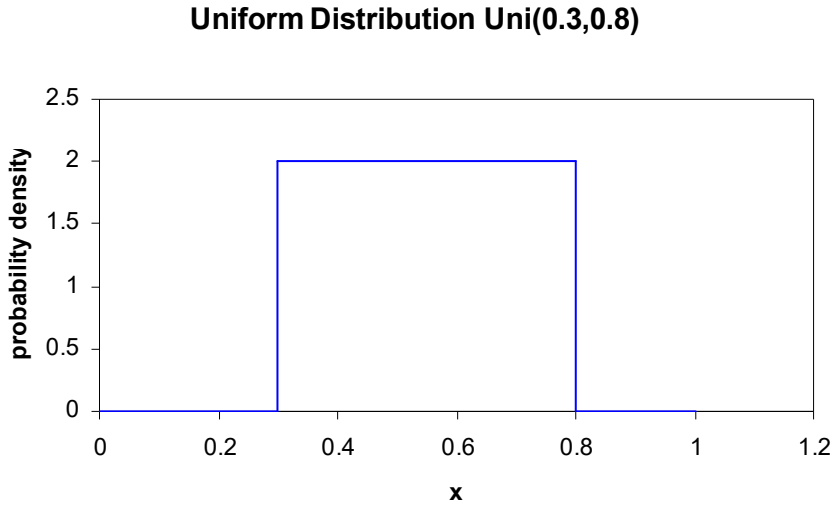
A special case of the Binomial distribution occurs when there is only one trial, so n=1. This is known as the Bernoulli distribution. It is a discrete distribution, with the only two possible values of m being 0 and 1. By implication, a Binomial distribution is just the sum of n random variables from a Bernoulli distribution.

As n tends to infinity, the Binomial distribution becomes more and more like a Normal distribution (see below), and we often use a Normal distribution to do calculations, to save having to use the formula above for several values of m. The Formulae Sheet quotes a rule for when the Normal approximation to the Binomial is

valid, namely $np(1-p) > 10$. Other textbooks and lecturers may have different versions of this rule; it all boils down to how precise one wants to be.

2.2 Uniform Distribution

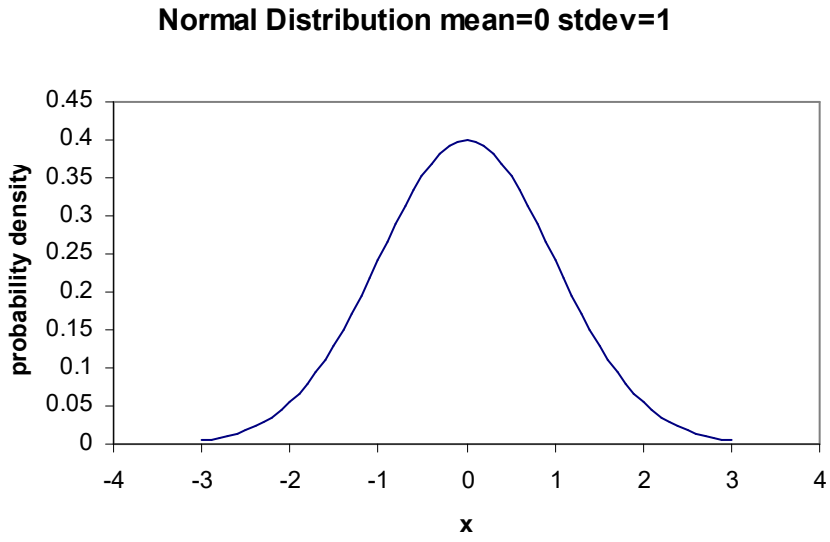
The Uniform distribution is a continuous distribution. The probability density is a constant for all values of the random variable within a certain range:



The Uniform distribution is most often used to model a variable whose value we do not know, other than knowing it is within a certain range of values. It is in effect a statement that we have no idea of the precise value of the variable within this range. One particular use is to describe the possible values of a probability (such as the probability it will rain tomorrow) when we have absolutely no idea if it will rain: we would use a Uniform distribution between 0 and 1.

2.3 Normal Distribution

The Normal distribution arises when we take n independent random values from almost any other distribution, and sum them: if we were to do this repeatedly, the distribution of the sums will follow a Normal distribution, as n tends to infinity. The principle is known (curiously) as the *Central Limit Theorem*. You will meet the Normal distribution more often than any other distribution, because many of the random variables one meets in the world can be regarded sums of other, independent, random values. Because the mean of n random values is just their sum divided by n, the mean of a set of independent, random values from almost any distribution also follows a Normal distribution. In particular, the mean of a random sample of size n from almost any distribution will follow a Normal distribution fairly closely as long as n is 30 or more. (Whether you use the number 30 or some other number depends on how fussy and precise you wish to be.) The only exceptions are where the original distribution is extreme in respect of spikiness or outliers.



The Normal distribution with mean μ and variance σ^2 is often denoted by N(μ,σ^2).

[Aside: a bit of jargon you will often come across in this context is i.i.d. This stands for independent, identically distributed, and means a set of random variables which are independent and from the same distribution. A very important example is a random sample, which meets these criteria.]

We can sometimes use the Normal distribution to calculate a confidence interval for the population mean, or to do a hypothesis test involving the population mean. If the variance of individuals in a population is σ^2, then the variance of the mean of samples of size n about the population mean can be shown to be σ^2 / n. Hence if individuals in the population follow a Normal distribution and we know σ, then a 95% confidence interval for the population mean, based on a single sample of size n is:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

If we wish to do a hypothesis test that the population mean has some particular value μ, then we might calculate the quantity:

$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}}$$

and compare it with some critical value from the Normal distribution.

But both of these procedures require us to know σ^2, the variance of the individuals in the population, and we do not usually know this quantity precisely. σ^2 is sometimes given in examination questions, but in real life, usually the only way to estimate it is from a finite sample. We denote the estimate of σ^2 based on a sample as $\hat{\sigma}^2$, calculated using the formula:

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

[Aside: many books and lecturers denote this quantity by s^2, though this has the slight downside that a few other books and lecturers use this symbol for the formula above with n on the bottom, rather than n-1.]

If the finite sample is large (most people would say $n \geq 30$, though this is arbitrary and ultimately depends on how precise you wish to be) then $\hat{\sigma}^2$ is a fairly precise estimate of σ^2, and the formulae above can be used with the Normal distribution to calculate confidence intervals and to do hypothesis tests. And in fact we can even relax the requirement that the individuals in the population follow a Normal distribution in this case, because the Central Limit Theorem says that the means of samples will be approximately Normal, as long as the distribution of individuals in the population is not extreme in respect of spikiness or outliers.

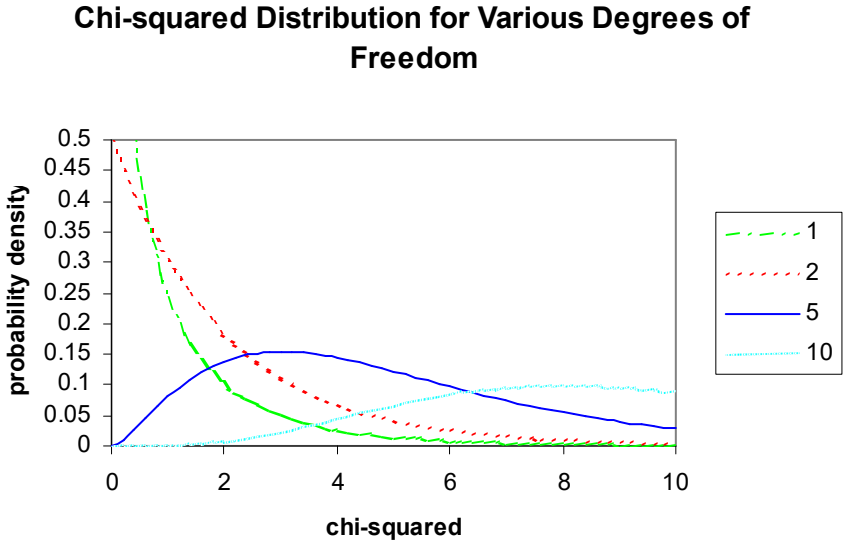
If the finite sample is not large (so $n < 30$), we need to allow for the error in $\hat{\sigma}^2$. To do this, we use the t-distribution instead of the Normal distribution. Before describing the t-distribution, we need to describe another distribution, the χ^2 distribution.

2.4 χ^2 Distribution

The χ^2 (pronounced "kai-squared" and often spelt chi-squared) distribution arises when we take a sum of squares of random values from a Normal distribution. The formal definition is:

$$\chi^2 = \sum_{i=1}^{k} Z_i^2$$

where the Z_i are a random values from a Normal distribution with mean 0 and standard deviation 1. The variable k is known as the *degrees of freedom*. Here are some graphs of the χ^2 distribution for various degrees of freedom:



The main situation in which you will need to use the χ^2 distribution is when you are working with a random sample of values from a Normal distribution, and wish to calculate a confidence interval for the variance, or to do hypothesis tests involving the variance. To see why this is, it is useful to study the quantity:

$$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

where the x_i are a set of n random variables from a Normal distribution. It can be shown that this quantity follows a χ^2 distribution with n-1 degrees of freedom. You can see that the quantity *might* follow a χ^2 distribution, in that it is the sum of squares of random variables $\frac{x_i-\bar{x}}{\sigma}$, which look somewhat like N(0,1) Normal variables. You might wonder why the degrees of freedom is only n-1 rather than n: ultimately that is because we are not adding squares of $\frac{x_i-\mu}{\sigma}$, but rather $\frac{x_i-\bar{x}}{\sigma}$, and that slightly reduces the scope the individual terms in the sum have to vary.

The quantity above can be rewritten as:

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

ie

$$\frac{(n-1)}{\sigma^2}\hat{\sigma}^2$$

Therefore we know $\frac{(n-1)}{\sigma^2}\hat{\sigma}^2$ follows a χ^2 distribution with n-1 degrees of freedom.

To find a 95% confidence interval for the population variance of individuals (σ^2), we would argue as follows:

$$\text{Prob}\left[\frac{(n-1)}{\sigma^2}\hat{\sigma}^2 > U\right] = 0.025$$

where U is the value of χ^2 corresponding to an area 0.025 in the right hand (ie upper) tail.

$$\text{Hence Prob}\left[\sigma^2 < \frac{(n-1)}{U}\hat{\sigma}^2\right] = 0.025$$

$$\text{and Prob}\left[\frac{(n-1)}{\sigma^2}\hat{\sigma}^2 < L\right] = 0.025$$

where L is the value of χ^2 corresponding to an area 0.025 in the left hand (ie lower) tail.

$$\text{Hence Prob}\left[\sigma^2 > \frac{(n-1)}{L}\hat{\sigma}^2\right] = 0.025$$

And hence overall:

$$\text{Prob}\left[\frac{(n-1)}{U}\hat{\sigma}^2 < \sigma^2 < \frac{(n-1)}{L}\hat{\sigma}^2\right] = 0.95$$

In finding U and L, we use a χ^2 distribution with n-1 degrees of freedom.

2.5 t-Distribution

It was noted in Section 2.3 above that if we wish to do a hypothesis test that the population mean has some particular value μ, then we might calculate the quantity:

$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}}$$

but that in practice, if we do not know σ^2, and have to use $\hat{\sigma}^2$ (estimated from a sample) instead, then the resulting quantity

$$\frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$$

may not follow a Normal distribution sufficiently precisely. The t-distribution (sometimes known as the Student t-distribution) is designed to handle this situation. We can analyse how the quantity

$$\frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$$

behaves, We first rewrite it as follows:

$$\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{(n-1)\sigma^2}}}$$

ie:

$$\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{\sigma^2}\Big/(n-1)}}$$

which, if the $\bar{x}$ is the mean of a sample of values from a Normal distribution, can be symbolically written as:

$$\frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

and that is the definition of the t-distribution with *n*-1 degrees of freedom.

So in doing hypothesis tests where we do not know σ^2, and have to estimate it from a sample value ($\hat{\sigma}^2$), we compare the quantity

$$\frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}}$$

with a critical value from a t-distribution rather than a Normal distribution. However, if *n* is large (again, "30 or more" is a common rule), then we can actually revert to using the Normal distribution, because then $\hat{\sigma}^2$ is a sufficiently precise estimator of σ^2.

A similar analysis can be used to show that we should also use the t-distribution to calculate a confidence interval for the population mean. So for a single sample of size *n*, the 95% confidence interval would be

$$\bar{x} \pm t\frac{\sigma}{\sqrt{n}}$$

where *t* is the value from the t-distribution with *n*-1 degrees of freedom, and which is only exceeded in magnitude 5% of the time.

2.6 F-distribution

The F-distribution is used to compare two sample variances, usually to test whether they come from populations with the same variance. Rather than analyse the

difference between the two sample variances, we analyse their ratio. Let us say the two samples are of sizes n_1 and n_2, and give estimates of their respective population variances as $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. We will call the hypothesised variance of both populations from which they both come σ^2. We need to assume that individuals in the two populations follow Normal distributions. We can write the ratio of the sample variances as:

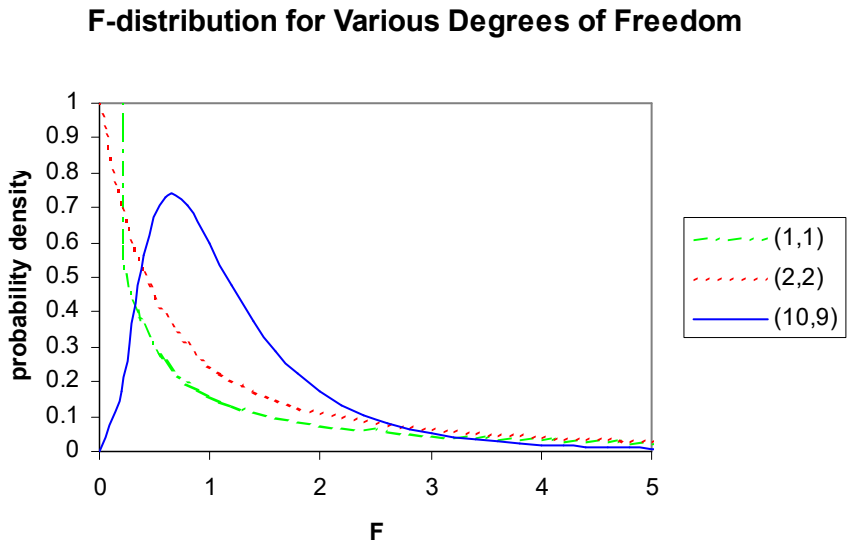
$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\left[\frac{(n_1-1)\hat{\sigma}_1^2}{\sigma^2}\right] \Big/ (n_1-1)}{\left[\frac{(n_2-1)\hat{\sigma}_2^2}{\sigma^2}\right] \Big/ (n_2-1)}$$

which can be symbolically written as:

$$\frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)}$$

where χ_i^2 is a χ^2 distribution with $n_i - 1$ degrees of freedom.

This is the definition of an F-distribution with (n_1-1, n_2-1) degrees of freedom.



To carry out a two tailed hypothesis test (at a significance level of, say, 5%) that the two samples come from populations with the same variances, we need to see if $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ is in the 2.5% region of the right hand tail or the 2.5% region of the left hand tail of an F-distribution with (n_1-1, n_2-1) degrees of freedom. The easiest way to do this is to arbitrarily label the samples so that $\hat{\sigma}_1^2$ is always bigger than $\hat{\sigma}_2^2$. Then all we have to

do is to compare $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ with the 2.5% critical value from the right hand tail of an F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

You will in fact find that tables of the F-distribution usually provide information about *only* the *right* hand tail. Apart from the fact that the procedure of the previous paragraph means we do not need to look at the left hand tail, there is another reason: these tables can be used to obtain information about the left had tail, as follows. The probability that $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ is less than (say) k is the same as the probability that $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$ is more than $1/k$. So if we wanted to find the critical value of F with (m,n) degrees of freedom, beyond which there is only a 2.5% area in the left hand tail, we would take the inverse of the critical value of the F distribution with (n,m) degrees of freedom such there is only a 2.5% area in the right hand tail.