3.5 : Goodness of Fit for Discrete Distributions

We can illustrate some of the important ideas of hypothesis testing by considering how we test the goodness of fit of data to **discrete** distributions. We do this using the χ^2 statistic.

Suppose we carry out n observations and obtain as our results k different discrete outcomes, $E_1, ..., E_k$ which occur with frequencies $o_1, ..., o_k$ ('o' for 'observed'). An example of such observations might be the number of meteors observed on n different nights, or the number of photons counted in n different pixels of a CCD.

Consider the null hypothesis that the observed outcomes are a sample from some model discrete distribution (e.g. a Poisson distribution). Suppose, under this null hypothesis, that the k outcomes, $E_1, ..., E_k$, are expected to occur with frequencies $e_1, ..., e_k$ ('e' for 'expected'). We can test our null hypothesis by comparing the observed and expected frequencies and determining if they differ significantly. We construct the following χ^2 test statistic.

$$\chi^2 \quad = \quad \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where $\sum o_i = \sum e_i = n$. Under the null hypothesis this test statistic has approximately a χ^2 pdf with $\nu = k - 1 - m$ degrees of freedom. Here m denotes the number of parameters (possibly zero) of the model discrete distribution which one needs to estimate before one can compute the expected frequencies, and ν is reduced by one further degree of freedom because of the constraint that $\sum e_i = n$. In other words, once we have computed the first $k - 1$ expected frequencies, the k^{th} value is uniquely determined by the sample size n.

This χ^2 goodness of fit test need not be restricted *only* to discrete random variables, since we can effectively produce discrete data from a sample drawn from a continuous pdf by binning the data. Indeed, as we remarked in Section 2.2.7 the Central Limit Theorem will ensure that such binned data are approximately normally distributed, which means that the sum of their squares will be approximately distributed as a χ^2 random variable. The approximation to a χ^2 pdf is very good provided $e_i \geq 10$, and is reasonable for $5 \leq e_i \leq 10$.

3.5.1 : Example 1

A list of 1000 'random' digits – integers from 0 to 9 – are generated by a computer. Can this list of digits be regarded as uniformly distributed?

Suppose the integers appear in the list with the following frequencies:-

r	0	1	2	3	4	5	6	7	8	9
o_r	106	88	97	101	92	103	96	112	114	91

Let our NH be that the digits are drawn from a uniform distribution. This means that each digit is expected to occur with equal frequency – i.e. $e_r = 100$, for all r. Thus:-

$$\chi^2 \quad = \quad \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \quad = \quad 7.00$$

Suppose we adopt a 5% level of significance. The number of degrees of freedom, $\nu = 9$; hence the critical value of $\chi^2 = 16.9$ for a one-tailed test. Thus, at the 5% significance level we **accept** the NH that the digits are uniformly distributed.

3.5.2 : Example 2

A coin is tossed 200 times, and 115 heads and 85 tails are recorded. Test the null hypothesis that the coin is fair, using a 5% level of significance.

Under the NH of a fair coin we have $e_1 = e_2 = 100$. Thus:-

$$\chi^2 \quad = \quad \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \quad = \quad 4.5$$

Here, the number of degrees of freedom, $\nu = 1$, for which we have a critical value of $\chi^2 = 3.84$. Hence we **reject** the NH at the 5% significance level – i.e. the coin is *not* fair.

3.5.3 : Example 3

The table below shows the number of nights during a 50 night observing run when r hours of observing time were 'clouded out'. Fit a Poisson distribution to these data for the pdf of r and determine if the fit is acceptable at the 5% significance level.

r	0	1	2	3	4	> 4
No. of nights	21	18	7	3	1	0

Of course one might ask whether a Poisson distribution is a sensible model for the pdf of r since a Poisson RV is defined for any non-negative integer, whereas r is clearly at most 12 hours. However, as we saw in Section 1.3.2, the shape of the Poisson pdf is sensitive to the value of the mean, μ, and in particular for small values of μ the value of the pdf will be negligible for all but the first few integers, and so we neglect all larger integers as

possible outcomes. Hence, in fitting a Poisson model we also need to estimate the value of μ. We take as our estimator of μ the **sample mean**, i.e.

$$\hat{\mu} \quad = \quad \frac{21 \times 0 \, + \, 18 \times 1 \, + \, +7 \times 2 \, + \, 3 \times 3 \, + \, 1 \times 4}{50} \quad = \quad 0.90$$

Substituting this value into the Poisson pdf we can compute the *expected* outcomes, $e_r = 50\, p(r; \hat{\mu})$, where

$$p(0; 0.90) = 0.4066 \qquad p(1; 0.90) = 0.3659 \qquad p(2; 0.90) = 0.1647$$
$$p(3; 0.90) = 0.0494 \qquad p(4; 0.90) = 0.0111 \qquad p(5; 0.90) = 3.3 \times 10^{-5}$$

If we consider only five outcomes, i.e. $r \leq 4$, since the value of the pdf is negligible for $r > 4$, then the number of degrees of freedom, $\nu = 3$ (remember that we had to estimate the mean, μ). The value of the test statistic is $\chi^2 = 0.68$, which is smaller than the critical value. Hence we **accept** the NH at the 5% level – i.e. the data are well fitted by a Poisson distribution.

3.5.4 : The Binomial Distribution

In Section 3.5.3 we could have fitted the data with another discrete model – the **binomial distribution**. Suppose there are a total of n hours in each observing night (e.g. $n = 8$ or $n = 12$). Let θ denote the probability of any single hour being 'clouded out'. The binomial distribution gives the probability of getting r out of n hours clouded out ($r = 0, 1, ..., n$), viz:-

$$p(r; \theta) = \frac{n!}{r!(n-r)!} \theta^r (1-\theta)^{n-r}$$

$p(r; \theta)$ is the binomial pdf. It is quite straightforward to show (see handout) that the binomial distribution has mean, $E(r) = n\theta$ and variance, $\mathrm{var}(r) = n\theta(1-\theta)$.

As in Section 3.5.3, we have to estimate a single parameter – in this case θ (assuming that the number of observing hours, n, is known) – in fitting the data to a binomial model. We do this by equating the sample mean, $\hat{\mu}$, with the expected value of r, i.e. $n\theta$. We can then construct a χ^2 statistic exactly as in 3.5.3. (remembering to reduce the number of degrees of freedom by one because we need to estimate θ).

3.6 : The Kolmogorov-Smirnov Test

Suppose we want to test the hypothesis that a sample of data is drawn from the underlying population with some given pdf. We could do this by binning the data and comparing with the model pdf using the χ^2 test statistic. This approach might be suitable, for example,

for comparing the number counts of photons in the pixels (i.e. the bins) of a CCD array with a bivariate normal model for the 'point spread function' of the telescope optics, where the centre of the bivariate normal defines the position of a star.

For small samples this does not work well, however, as we cannot bin the data finely enough to usefully constrain the underlying pdf – particularly if our pdf is *multivariate*, as in the case of the bivariate normal example above, and requires several parameters to define it.

A more useful approach in this situation is to compare the sample **cumulative distribution function** with a theoretical model. We can do this using the **Kolmogorov-Smirnov** (KS) test statistic.

Let $\{x_1, ..., x_n\}$ be an iid random sample from the unknown population. Suppose the $\{x_i\}$ have been arranged in ascending order. The sample cdf, $S_n(x)$, of X is defined as:-

$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n} & x_i \leq x < x_{i+1}, \qquad 1 \leq i \leq n-1 \\ 1 & x \geq x_n \end{cases}$$

i.e. $S_n(x)$ is a step function which increments by $1/n$ at each sampled value of x.

Let the model cdf be $P(x)$, corresponding to pdf $p(x)$, and let the null hypothesis be that our random sample is drawn from $p(x)$. The KS test statistic is

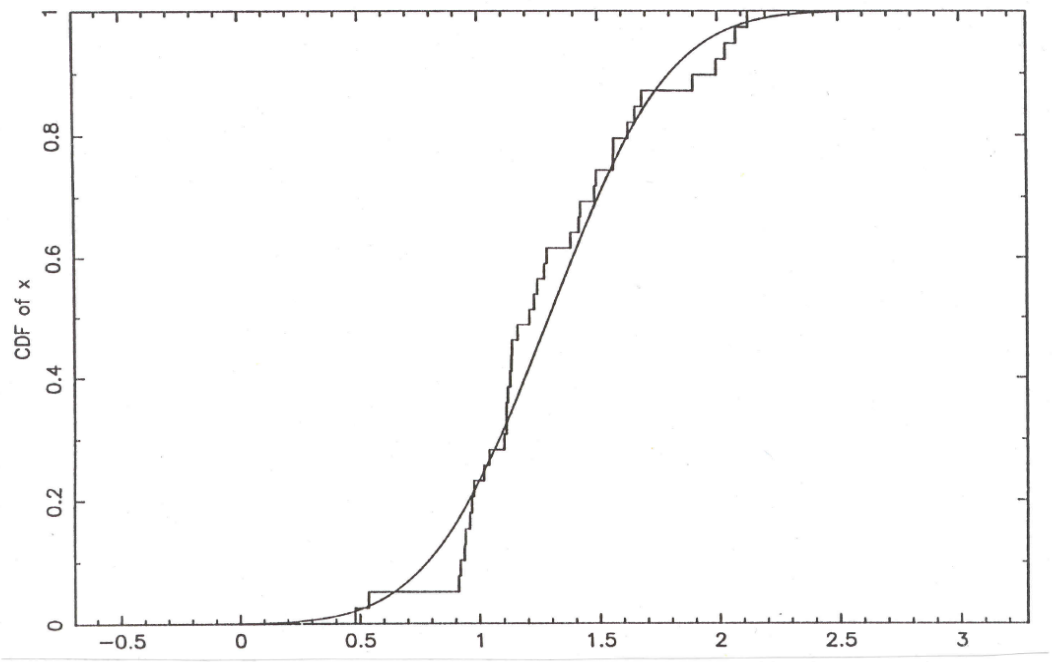
$$D_n \;=\; \max |P(x) - S_n(x)|$$

It is easy to show that D_n always occurs at one of the sampled values of x. The remarkable fact about the KS test is that the distribution of D_n under the null hypothesis is **independent of the functional form** of $P(x)$. In other words, whatever the form of the model cdf, $P(x)$, we can determine how likely it is that our *actual* sample data was drawn from the corresponding pdf. Critical values for the KS statistic are tabulated or can be obtained from numerical algorithms.

Figure 17 shows the KS test applied to the log period distribution in a sample of LMC Cepheids. Shown is the sample cdf of the 39 stars, together with the model cdf with which they are being compared: a normal distribution with mean and variance equal to the sample mean and variance of the real data. The observed value of the test statistic for these data, $D_{\rm obs} = 0.124$. Comparison with the critical values of the distribution show that $\mathrm{Prob}(D_n > D_{\rm obs}) = 0.562$. Thus, if the NH is true, there is a more than 50% chance that one would obtain as large, or indeed larger, a value of D_n for a randomly chosen

sample of 39 Cepheids drawn from the model normal pdf. This clearly suggests that we should accept the null hypothesis for these data – i.e. the distribution of log periods is adequately described by a normal pdf.

Figure 17: Example KS Test



There is also a two-sample version of the KS test, where one tests the null hypothesis that the two samples are drawn from the same underlying population. The test statistic is now

$$D_{m,n} = \max |S_m(x) - S_n(x)|$$

where S_m and S_n denote the sample CDFs of two samples of size m and n respectively. Again the distribution of $D_{m,n}$ under the null hypothesis is independent of the underlying pdf. This is especially useful, because it means that we can test whether two samples are drawn from the same underlying population **without having to assume anything about the form of that population**.

The KS test is an example of a **robust**, or **nonparametric**, test since one can apply the test with minimal assumption of a parametric form for the underlying pdf. The price for this robustness is that the **power** of the KS test is lower than other, parametric, tests. In other words there is a higher probability of accepting a false null hypothesis – that two samples *are* drawn from the same pdf – because we are making no assumptions about the parametric form of that pdf.