3.6: General Hypothesis Tests

The χ^2 goodness of fit tests which we introduced in the previous section were an example of a **hypothesis test**. In this section we now consider hypothesis tests more generally.

3.6.1: Simple Hypothesis Tests

A **simple hypothesis test** is one where we test a **null hypothesis**, denoted by H_1 (say), against an **alternative hypothesis**, denoted by H_2 – i.e. the test consists of only **two** competing hypotheses. We construct a **test statistic**, t, and based on the value of t observed for our real data we make one of the following two decisions:-

1. accept H_1, and reject H_2
2. accept H_2, and reject H_1

To carry out the hypothesis test we choose the **critical region** for the test statistic, t. This is the set of values of t for which we will choose to **reject** the null hypothesis and accept the alternative hypothesis. The region for which we accept the null hypothesis is known as the **acceptance region**. Note that we must choose the critical region and acceptance region ourselves. For example we might choose the critical region as the set of values of t for which $t > 0$.

3.6.2: Level of Significance

The **level of significance** of a hypothesis test is the maximum probability of incurring a type I error which we are willing to risk when making our decision. In practice a level of significance of 5% or 1% is common. If a level of significance of 5% is adopted, for example, then we choose our critical region so that the probability of rejecting the null hypothesis when it is *true* is no more than 0.05.

If the test statistic *is* found to lie in the critical region then we say that the null hypothesis is rejected at the 5% level, or equivalently that our rejection of the null hypothesis is **significant** at the 5% level. This means that, **if** the null hypothesis **is** true, and we were to repeat our experiment or observation a large number of times, then we would expect to obtain – by chance – a value of the test statistic which lies

in the critical region (thus leading us to reject the NH) in no more than 5% of the repeated trials. In other words, we expect our rejection of the null hypothesis to be the *wrong* decision in no more than 5 times out of every 100 experiments.

3.6.3: 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 again 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 $5 \leq e_i \leq 10$.

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.

Example 2

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.

3.6.4: 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.

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 e.g. from numerical recipes algorithms.

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.

3.6.5: Hypothesis Tests on the Sample Correlation Coefficient

The final type of hypothesis test which we consider is associated with testing whether two variables are statistically independent, which we can do by considering the value of the **sample correlation coefficient**. In Section 3.1 we defined the covariance of two RVs, x and y, as

$$\mathrm{cov}(x, y) \quad = \quad E[(x - \mu_{\mathrm{x}})(y - \mu_{\mathrm{y}})]$$

and the correlation coefficient, ρ, as

$$\rho \quad = \quad \frac{\mathrm{cov}(x, y)}{\sigma_{\mathrm{x}} \sigma_{\mathrm{y}}}$$

We estimate ρ by the **sample correlation coefficient**, $\hat{\rho}$, defined by:-

$$\hat{\rho} \quad = \quad \frac{\sum(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sqrt{[\sum(x_i - \hat{\mu}_x)^2]\,[\sum(y_i - \hat{\mu}_y)^2]}}$$

where, as usual, $\hat{\mu}_x$ and $\hat{\mu}_y$ denote the sample means of x and y respectively, and all sums are over $1, ..., n$, for sample size, n. $\hat{\rho}$ is also often denoted by r, and is referred to as 'Pearson's correlation coefficient'.

If x and y *do* have a bivariate normal pdf, then ρ corresponds precisely to the parameter defined in Section 3.1. To test hypotheses about ρ we need to know the sampling distribution of $\hat{\rho}$. We consider two special cases, both of which are when x and y have a bivariate normal pdf.

(i): $\rho = 0$ (i.e. x and y are independent)

If $\rho = 0$, then the statistic

$$t \quad = \quad \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$$

has a *student's t distribution*, with $\nu = n - 2$ degrees of freedom. Hence, we can use t to test the hypothesis that x and y are independent.

(ii): $\rho = \rho_0 \neq 0$

In this case, then **for large samples**, the statistic

$$z \quad = \quad \frac{1}{2}\log_e\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$

has an approximately normal pdf with mean, μ_z and variance σ_z^2 given by

$$\mu_z \; = \; \frac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) \qquad \sigma_z^2 \; = \; \frac{1}{n-3}$$