April 5th

The Chi-square test

The chi-squared significance test for goodness of fit

Let $Y_1, Y_2, \ldots, Y_k$ be the observed cell counts in a table that arise from random sampling. Suppose their joint distribution is described by the multinomial model with probabilities p_1, p_2, ..., p_k. A significance test of

$$H_0 : p_1 = \pi_1, \cdots, p_k = \pi_k, \quad H_A : p_i \neq \pi_i \text{ for at least } i$$

can be performed with the χ^2 statistic. The π_i are specified probabilities. Under H_0 the sampling distribution is asymptotically the chi-squared distribution with $k-1$ degrees of freedom. This is a good approximation, provided that the expected cell counts are all five or more. Large values of the statistic support the alternative.

This test is implemented by the `chisq.test()` function. The function is called with

```
chisq.test(x, p=...)
```

The data is given in tabulated form in `x`; the null hypothesis is specified with the argument `p=` as a vector of probabilities. The default is a uniform probability assumption. This should be given as a named argument, as it is not the second position in the list of arguments. The alternative hypothesis is not specified, as it does not change. A warning message will be returned if any category has fewer than five expected counts.

The chi-squared test for independence of two categorical variables

Let $Y_{ij}, i = 1,\ldots,n_r, j = 1,\ldots,n_c$ be the cell frequencies in a two-way contingency table for which the multinomial model applies. A significance test of

$$H_0 : \text{the two variables are independent}$$

$$H_A : \text{the two variables are not independent}$$

can be performed using the chi-squared test statistic (9.2). Under the null hypothesis, this statistic has sampling distribution that is approximated by the chi-squared distribution with $(n_r - 1)(n_c - 1)$ degrees of freedom. The p-value is computed using $\mathsf{P}(\chi^2 \geq \text{ observed value } | H_0)$.

In R this test is performed by the `chisq.test()` function. If the data is summarized in a table or a matrix in the variable `x` the usage is

```
chisq.test(x)
```

If the data is unsummarized and is stored in two variables `x` and `y` where the ith entries match up, then the function can be used as

```
chisq.test(x,y).
```

Alternatively, the data could be summarized first using `table()`, as in `chisq.test(table(x,y))`.

For each usage, the null and alternative hypotheses are not specified, as they are the same each time the test is used.

The argument `simulate.p.value=TRUE` will return a p-value estimated using a Monte Carlo simulation. This is used if the expected counts in some cells are too small to use the chi-squared distribution to approximate the sampling distribution of χ^2.