Appendix D: The Binomial Distribution

In Chapter 9 we used the Chi Square value to calculate probabilities, but we did not discuss how these probabilities are calculated from the Chi Square value. In this appendix we will cover how these calculations are derived.

The table of the values of Chi Square is based on the properties of a normal distribution curve. The figures refer to the probability of drawing a sample at any given point under this curve. To see how these figures are derived, we must understand the principles of the *binomial distribution*.

Take a simple dichotomous variable – whether or not respondents had smoked in the last week. We saw in Table 8.2 (in Chapter 8) that 158 respondents had smoked while 169 had not. The proportion of respondents who smoke is therefore 48 per cent.

Now suppose that it were actually true that 48 per cent of all the adults in Brighton had smoked at some point in the last week, and suppose we took many, many samples of 327 individuals from this population. Some of these samples would come up with fewer than 158 smokers and some would come up with more, but most of our samples would be fairly close to the figure of 158 and only a few would be a long way off it. We would end up with a distribution looking very much like a normal distribution, the difference being that it will be in the form of a histogram rather than a smooth curve (because we have

only whole numbers). This is the binomial distribution – the distribution of sample proportions.

Because it is a normal curve, we can work out the probability of our sample proportion (48%) being similar to the actual proportion of smokers in the population. The logic is similar to that involved in calculating the Standard Error of the mean, but this time we calculate the *Standard Error of sample proportions*, SE(*p*).

We know that 95 per cent of sample proportions will fall within 2 standard errors either side of the mid-point; thus we can say with 95 per cent certainty that the proportion of smokers in the target population is within plus or minus 2 SE(*p*)s of our sample mid-point of 158.

What we are interested in, however, is the association between *two* variables – in our case, respondents who smoke, and respondents with friends who smoke. Obviously, therefore, we have to compare the smoking behaviour of the friends of smokers with that of the friends of non-smokers. From Table 8.2 we can see that:

- among the 158 smokers in our sample, there are 111 who report that half or more of their friends smoke (a proportion of 111/158 = 0.70);

- among the 169 non-smokers, there are 37 who report that half or more of their friends smoke (a proportion of 37/169=0.22).

Obviously, these proportions are different, but the question is whether they are sufficiently different to convince us that they reflect real differences in the population from which these sub-samples were drawn.

To judge this, we calculate how many times out of a hundred we would be likely to draw two sub-samples from the population with a difference as large as 0.48 in the proportions reporting that most of their friends did or did not smoke. This is what the look-up table tells us. The calculation of the Chi Square statistic gives a total for the differences of proportions in the sample, and this figure can then be checked against the figures in the look-up table to see how many times a total difference of this size would occur in every hundred random samples given no difference in the population.