Categorical Data Analysis (CDA). Categorical data analysis (CDA) is the analysis of data that are in the form of counts in various categories. We will deal primarily with observations of categories. Such observations represent various hair colors (black, brown, blonde, red), or various eye colors (brown, hazel, blue, green), and each cell contains a count of the number of people who fall in that particular category. Thus, we are interested in proportions of hair color or eye color. We emphasize that the data considered in this analysis are counts, rather than continuous measurements. Particularly when data are dichotomous we will make heavy use of binomial distributions and their normal approximation.

Binomial distribution. Suppose that n independent experiments, or trials, are performed, and that each experiment results in a "success" with probability p and a "failure" with probability $(1-p)$. The total number Y of successes is a binomial random variable. The probability that $Y = j$ can be found in the following way: Any particular sequence of j successes occurs with probability $p^j(1-p)^{n-j}$. The total number of such sequences is

$$\binom{n}{j} = \frac{n!}{j!(n-j)!}$$

since there are $\binom{n}{j}$ ways to assign j successes to n trials. Thus, the probability of any particular sequence times the number of such sequences becomes

$$P(Y = j) = \binom{n}{j} p^j (1-p)^{n-j}$$

Example: Apple scab disease. $n = 20$ apples of the variety "Summer red" were collected at random from an old apple tree. Y of these apples had signs of the black or grey-brown lesions associated with apple scab. Suppose that the probability of having those signs is $p = 0.25$. Then the exact distribution is given by the binomial distribution

$$P(Y = j) = \binom{20}{j} (0.25)^j (0.75)^{20-j}$$

The term $(0.25)^j$ corresponds to the probability of apple scab for each of the j apples, and the term $(0.75)^{20-j}$ corresponds to each of the $(20-j)$ apples without scab.

Hypothesis test. The number of apples with apple scab among 20 sampled apples is binomial, with size $n = 20$ and the probability p of finding apple scab, and observe $Y = 3$. Assume that we wish to test the hypothesis

$$H_0 : p = 0.25$$

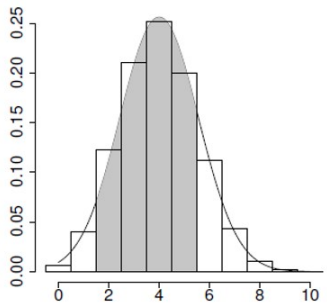
If H_0 is true then the probability of observing 3 apples with scab is $P(Y = 3) = 0.1339$. Thus, we do not reject the hypothesis that the proportion of apples with scab is 25%.

Normal approximation. The binomial distribution has mean np and variance $np(1-p)$ and the form is symmetric and resembles that of a normal distribution provided that p is not too close to zero or one. We can therefore try to approximate the binomial distribution with a normal distribution with the same mean and variance, $N(np, np(1-p))$. A frequently used rule of thumb is that the approximation is reasonable when $np > 5$ and $n(1-p) > 5$. We

can calculate probabilities for the binomial distribution using the standard normal cumulative distribution $\Phi(z)$ as follows:

$$P(Y \leq y) \approx \Phi\left(\frac{(y+0.5)-np}{\sqrt{np(1-p)}}\right)$$

Notice how we add 0.5 to Y in the numerator. This is because the normal distribution is a continuous distribution, whereas the binomial distribution is a discrete distribution.



If we wish to approximate the probability that a binomial variable Y results in values between 2 and 5, then we say that we get the best approximation if we compare that value to the interval $(1.5, 5.5)$ for the continuous distribution.

$$P(2 \leq Y \leq 5) \approx \Phi\left(\frac{5.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{1.5-np}{\sqrt{np(1-p)}}\right)$$

Example: Apple scab disease. The number of apples with apple scab among 20 sampled apples is binomial with size $n = 20$ and the probability $p = 0.25$ of finding apple scab, and observe $Y = 3$.

$$\begin{aligned} P(Y=3) &= P(Y \leq 3) - P(Y \leq 2) \\ &= \Phi\left(\frac{3.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{2.5-np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Then we get 0.1209, which in fact approximates the exact probability of 0.1339 reasonably well.

There is a single parameter p for a binomial distribution, and the obvious estimate for that is obtained by the observed number Y and then dividing it by the number n of sample. With size $n = 20$ we observe $Y = 7$. Our estimate of the proportion of apples infected with apple scab from this tree will be

$$\hat{p} = \frac{7}{20} = 0.35$$

The standard error (SE) of the population estimate becomes

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The SE for the above example will be 0.1067.

We can use the standard approach to construct a confidence interval for proportion estimate. The standard error (SE) is determined from proportion alone (there is no extra variance parameter in this model). Then we can use a quantile from the normal distribution. The $(1-\alpha)\%$ confidence interval for the proportion p becomes

$$\hat{p} \pm (z_{\alpha/2})\text{SE}(\hat{p})$$

where $z_{\alpha/2}$ denotes the critical value with level $\alpha/2$ for the standard normal distribution. The 95% confidence interval for the proportion of apples that are infected with scab will be

$$\begin{aligned} 0.35 \pm (1.96)\sqrt{\frac{(0.35)(0.65)}{20}} &= 0.35 \pm (1.96)(0.1067) \\ &= (0.1410, 0.5590) \end{aligned}$$

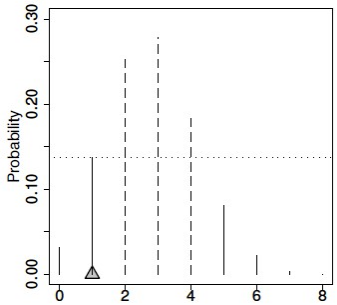
Exact test. We can make a formal test for

$$H_0 : p = p_0$$

Here we use the observed number Y_{obs} itself as our test statistic. Recall that the p-value is defined as the probability of observing something that is as extreme or more extreme, i.e., is less in accordance with the null hypothesis than our observation. Outcomes with probabilities less than $P(Y = Y_{obs})$ are more extreme, so we must add the probabilities of all possible outcomes and obtain the p-value by

$$\sum_{y:P(Y=y)\leq P(Y=Y_{obs})} P(Y = y)$$

Assume we wish to test the hypothesis that $H_0 : p = 0.35$, and we have observed the value $Y_{obs} = 1$ with size $n = 8$.



The dotted horizontal line is the probability corresponding to the observed value. The p-value of 0.2752 corresponds to the sum of the outcomes that are at least as "extreme" as our observation. The solid vertical lines correspond to those outcomes and probabilities.