Workshop 6: Likelihood

In this workshop we will use likelihood methods to estimate parameters and test hypotheses. Likelihood methods are especially useful when modeling data having a probability distribution other than the normal distribution (e.g., binomial, exponential, etc).

Maximum likelihood estimate

To estimate a parameter, we treat the data as given and vary the parameter to find that value for which the probability of obtaining the data is highest. This value is the *maximum likelihood estimate* of the parameter. The likelihood function is also used to obtain a *likelihood-based confidence interval* for the parameter. This confidence interval is a large-sample approximation, and may be inaccurate for small sample size, depending on the probability distribution of the data.

Log-likelihood ratio test

The log-likelihood ratio test can be used to compare the fits of two nested models to the same data. The "full" model fits the data using the maximum likelihood estimates for the parameter(s) of interest (for example, a proportion p). The "reduced" model constrains the parameter values to represent a null hypothesis (for example, that $p = 0.5$, or that p is equal between two treatments). The G statistic is calculated as twice the difference between the log-likelihoods of the two models ("full" minus "reduced"):

```
G <- 2 *(loglikefull - loglikereduced)
```

G is also known as the deviance. Under the null hypothesis, G has an approximate χ^2 distribution with degrees of freedom equal to the difference between the "full" and "reduced" models in the number of parameters estimated from data. We'll work through an example below.

Warmup

We'll start by getting familiar with the commands in R to calculate probabilities.

1. The probability of heads in a coin toss is 0:

4. To get closer to this value, repeat steps (1) to (3) using a narrower range of values for p

Left-handed owers

Individuals of most plant species are hermaphrodites (with both male and female sexual organs) and are, therefore, prone to inbreeding of the worst sort: having sex with themselves. The mud plantain, *Heteranthera multi ora*, has a simple mechanism to avoid such \sel ng." The style de ects to the left in some individuals and to the right in others. The anther is on the opposite side. Bees visiting a left-handed plant are dusted with pollen on their right side, which then is deposited on the styles of only right-handed plants visited later. To investigate the genetics of this variation, Jesson and Barrett (2002, *Proc. Roy. Soc. Lond., Ser. B, Biol. Sci.* 269: 1835-1839) crossed pure strains of left- and right-handed owers, yielding only right-handed F_1 o spring, which were then crossed with one another. The expectation under a simple model of inheritance would be that their F_2 o spring should consist of left- and right-handed individuals in a 1:3 ratio (i.e., 1=4 of the plants should be left-handed). The data from this cross can be downloaded here.

1. Calculate the log-likelihood curve and the maximum-likelihood estimate of the proportion of left-handed owers (to the nearest hundredth). Use the data in the data frame, rather than summaries of the frequencies of left- and right-handed owers, to calculate the likelihoods. In otherwords, use the vector of 'left' and 'right' to calculate the likelihood, rather than the numbers 6 and 21 (which correspond to the number of 'left' and 'right', respectively. Practice with this approach will help you later below.

2. Calculate the likelihood-based 95% con dence interval of the population proportion.

3. We can compare the ts of two models to these same data, to test the null hypothesis that the proportion of left-handed owers in the cross is 1=4 (i.e., the proportion predicted by the simplest genetic model for ower handedness). To begin, obtain the log-likelihood corresponding to the maximum likelihood estimate of the proportion of left-handed owers. This represents the t of the \full" model to the data. This model estimated one parameter from the data (p, estimated using maximum likelihood).

4. Now obtain the log-likelihood of the value for p speci ed by the null hypothesis. This represents the t of the \reduced" model to the data. This reduced model estimated zero parameters from the data (instead, p was speci ed by the null hypothesis).

5. Calculate the G statistic for the log-likelihood ratio test (see above for formula). To obtain a P-value for the test, calculate the tail probability from the ^{2} distribution as follows,

   ```
   1 - pchisq(G, df)
   ```

   where `df` is the degrees of freedom, calculated as the di erence between the two models in the number of parameters estimated from the data.

6. Optional: How similar is the result from your log-likelihood ratio test to that from an ordinary ^{2} goodness of t test? Analyze the same data using the chisq.test command in R and comment on the outcome.

Voyaging voles

The movement or dispersal distance of organisms is often modeled using the geometric distribution, assuming steps are discrete rather than continuous. For example, M. Sandell, J. Agrell, S. Erlinge, and J. Nelson (1991, *Oecologia*, 86: 153-158) measured the distance separating the locations where individual voles, *Microtus agrestis*, were first trapped and the locations they were caught in a subsequent trapping period, in units of the number of home ranges. The data for 145 male and female voles are here. The variable \dispersal" indicates the distance moved (number of home ranges) from the location of first capture. If the geometric model is adequate, then

$$\begin{aligned} Pr[X = 0 \text{ home ranges moved}] &= p \\ Pr[X = 1 \text{ home ranges moved}] &= (1 - p)p \\ Pr[X = 2 \text{ home ranges moved}] &= (1 - p)^2 p \end{aligned} \tag{1}$$

and so on. p is the probability of success (capture) at each distance from the location from the first capture.

1. Using the appropriate commands in R, calculate the maximum-likelihood estimate of p, the parameter of the geometric distribution. Use the data from both sexes combined.
2. Use the appropriate R command and the maximum likelihood estimate of p to calculate the predicted fraction of voles dispersing 0, 1, 2, 3, 4, and 5 home ranges.
3. Use the result in (2) to calculate the expected number of voles (out of 145) dispersing 0-5 home ranges, assuming a geometric distribution. How does this compare with the observed frequencies?

Life of bees

Life spans of individuals in a population are often approximated by an exponential distribution. To estimate the mortality rate of foraging honey bees, P. K. Visscher and R. Dukas (1997. Insectes Sociaux 44: 1-5) recorded the entire foraging life span of 33 individual worker bees in a local bee population in a natural setting. The 33 life spans (in hours) are here.

1. Plot the frequency distribution of lifespans of the 33 bees. Choose the option to display probability density instead of raw frequency. Does the distribution of lifespans resemble an exponential distribution (make sure to try different bin widths of the histogram)?
2. Use the exponential approximation and maximum likelihood to estimate the hourly mortality rate of bees.
3. Using the maximum likelihood estimate, calculate the probability density for the ex-