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 0.5. If you flip a coin 10 times, what is the probability of obtaining exactly 3 heads and 7 tails? Hint: use the `dbinom` command.

2. The fraction of human babies born who are boys is about 0.512. If 20 newborn babies are randomly sampled, what is the probability that exactly 10 are boys?

3. Plot the entire probability distribution for the number of boys in families having six children. Assume the probability that any one child is a boy is 0.512.

4. If mortality is independent of age, then the probability of surviving X years after birth, and then dying in the $(X + 1)^{st}$ year, will follow a geometric distribution. X is any integer from 0 to ∞. If the probability of dying in any given year is 0.1, what fraction of individuals are expected to survive 10 years and then die in their 11th year?

5. Refer to the previous question. If the probability of death in any give year is 0.1, what fraction of individuals die before they reach their 6th birthday?

6. In an environment where prey are randomly distributed, the search time between discovered prey items will follow an exponential distribution. Imagine an environment in which the mean search time between prey items is 0.5 hours. What is the probability density corresponding to a search time of 2 hours?

7. Refer to the previous problem. Create a line plot of the exponential probability curve over most the range of possible values for search time between items (e.g., over the range of values between 0 and 5 hours).

Illegal tender

Jenkins, A. J. (2001, *Forensic Sci. Int.* 121: 189-193) collected 50 US 1-dollar bills and tested them for drug contamination. 46 of the bills were found to contain trace amounts of cocaine (subsequent studies have corroborated this astonishing finding – can you think of a reason?). Let's use these data to estimate the proportion of contaminated bills in the US dollar population. For the purposes of this exercise, assume that the sample of bills was a random sample (in fact, 10 bills were sampled from each of 5 cities).

1. Generate a vector that includes a range of possible values for the population proportion p, from 0.01 to 0.99 in increments of 0.01.

2. Given the dollar bill data above, calculate the log-likelihood of each value for p. Hint: use the `dbinom` command.

3. Create a line plot of the log-likelihood against the range of values for p. What is the resulting curve called? Can you see approximately the value of p corresponding to the highest point of the curve? What is this value called?

4. To get closer to this value, repeat steps (1) to (3) using a narrower range of values for p surrounding the highest point in the curve.

5. Use your results to determine the maximum likelihood estimate of the proportion of US 1-dollar bills contaminated with cocaine.

6. Provide a likelihood-based 95% confidence interval for the population proportion.

7. Optional: try constructing the 95% confidence interval using the `uniroot` command to solve for the root of the likelihood function (shifted, as we did for the example in lecture).

Counting elephants

Counting elephants is more challenging than you might think, at least when they live in dense forest and feed at night. Eggert et al. (2003, *Mol. Ecol.* 12: 1389-1402) used mark-recapture methods to estimate the total number of forest elephants inhabiting Kakum National Park in Ghana without having to see a single one. They spent about two weeks in the park collecting elephant dung, from which they extracted elephant DNA. Using five genetic markers, they generated a unique DNA fingerprint for every elephant encountered in this way. Over the first seven days of collecting they identified 27 elephant individuals. Refer to these 27 elephants as **marked**. Over the next eight days they sampled 74 individuals, of which 15 had been previously marked. Refer to these 15 elephants as **recaptured**. We would like to use these numbers to estimate the total number of elephants in the park.
If we can make the following assumptions,

- There were no births, deaths, immigrants, or emigrants while the study was being carried out, and

- The dung of every elephant, marked or unmarked, regardless of its output, had an equal chance of being sampled, and selection of each individual was independent of the others.

then the number of recaptured (i.e., previously marked) individuals X in the second sample should have a hypergeometric distribution with parameters k (the size of the second sample of individuals), m (total number of marked individuals in the population), and n (total number of unmarked individuals in the population).

1. Using the `dhyper` command inR (for the hypergeometric distribution), calculate the maximum likelihood estimate for the total number of elephants in the park. Note that the total number is $n + m$, where n is the unknown parameter to be estimated. Note also that only integer values for n are allowed, and that n cannot be smaller than $k - X$, the observed number of unmarked individuals in the second sample.

2. Calculate a likelihood-based 95% confidence interval for the total number of elephants.

Left-handed flowers

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 multiflora*, has a simple mechanism to avoid such "selfing."

The style deflects 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 flowers, yielding only right-handed F_1 offspring, which were then crossed with one another. The expectation under a simple model of inheritance would be that their F_2 offspring 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 flowers (to the nearest hundredth). Use the data in the data frame, rather than summaries of the frequencies of left- and right-handed flowers, 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% confidence interval of the population proportion.

3. We can compare the fits of two models to these same data, to test the null hypothesis that the proportion of left-handed flowers in the cross is 1/4 (i.e., the proportion predicted by the simplest genetic model for flower handedness). To begin, obtain the log-likelihood corresponding to the maximum likelihood estimate of the proportion of left-handed flowers. This represents the fit 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 specified by the null hypothesis. This represents the fit of the "reduced" model to the data. This reduced model estimated zero parameters from the data (instead, p was specified 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 χ^2 distribution as follows,

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

where df is the degrees of freedom, calculated as the difference 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 χ^2 goodness of fit 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

$$Pr[X = 0 \text{ home ranges moved}] = p$$
$$Pr[X = 1 \text{ home ranges moved}] = (1 - p)p \qquad (1)$$
$$Pr[X = 2 \text{ home ranges moved}] = (1 - p)^2 p$$

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 exponential distribution across a range of values for lifespan. Draw this relationship between probability and lifespan on top of the frequency distribution you plotted in (1). Comment on the fit between the data and the exponential distribution you fitted. Is the exponential distribution a good fit to these data?

4. Assume (for the purposes of this exercise) that an exponential probability model is reasonable. Calculate the likelihood-based 95% confidence interval for the mortality rate.

5. Calculate the maximum likelihood estimate for the mean lifespan, with approximate 95% likelihood based confidence interval.