22S:138
Bayesian Statistics

Intro to One-Parameter Models: Learning about a Proportion

Lecture 3
Aug. 30, 2004

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Example:

- You read in last Monday's newspaper that one of the Iowa regents wants to raise tuition at the 3 Iowa universities by 10%.
- You want to send the regents some arguments against this idea.
- To support your argument, you would like to tell the regents what proportion of current UI students are likely to quit school if tuition is raised by that much.
- Your research question is: what is the unknown *population parameter* p – the proportion in the entire population of UI students who would be likely to quit school.

- You do not have the time or resources to locate and interview all 28000+ students, so you cannot evaluate p exactly.
- You pick a simple random sample of $n = 50$ students from the student directory and ask each of them whether he/she would be likely to quit school if tuition were raised by 10%.
- You wish to use your sample data to estimate the population proportion p, and to determine the amount of uncertainty in your estimate.

The binomial distribution

- For each of the 50 students in your sample, define a "Bernoulli" random variable indicating whether they say that would quit school (yes or no).
  - "Bernoulli" or "binary" random variable – random variable that can assume one of only two values
  - one value is arbitrarily called a "success," the other a "failure"
  - we'll call a "yes" answer a success
- The unknown population proportion p is also the probability that a randomly-selected student from this population would answer "yes."

- Because we know nothing about the people in your sample except that they were in the student directory, it is reasonable to assume that they all have the same probability of saying yes to your questions – namely p.
  - If we knew more about the students, this assumption would not be reasonable.
- We also will assume, because you drew a simple random sample, that the responses from the individual students are *independent*.
  - This would not be reasonable if you had chosen 25 pairs of roommates, sets of siblings, etc.

- Define a random variable Y as the count of the number of successes in your sample.
- Y meets the definition of a "binomial random variable" – it is the count of the number of successes in n independent Bernoulli trials, all with the same success probability

$$Y \sim Binomial(n, p)$$

- What are the possible values of Y?
- If we knew p, we could use the binomial probability mass function to compute the probability of obtaining any one of the possible values of Y in our sample.

$$p(y|p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n$$

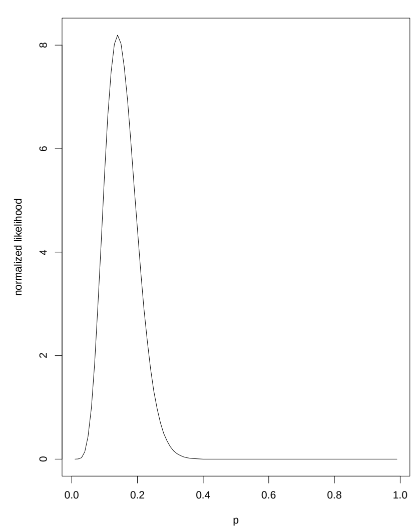
The likelihood function

- But we don't know p.
- Instead, after you interview the 50 students, we know that $y = 7$, and we want to estimate p.
- In this case, we may change perspective and regard the sampling distribution as a function of the unknown parameter p. When regarded in this way, the sampling distribution is called the "likelihood function."

$$L(p) = \binom{50}{7} p^7 (1-p)^{43}, \quad 0 < p < 1$$

- We could compute this likelihood for different values of p. Intuitively, values of p that give larger likelihood evaluations are more consistent with the observed data.

The likelihood function for a binomial sample with 7 successes in 50 trials



Frequentist approach to estimating p: Maximum likelihood estimation

- Frequentist approach to estimating p: find the value of p for which the likelihood function attains its maximum
  - This is the value that makes the observed data most likely.
  - This value of p is called the "maximum likelihood estimate" or MLE.
- Usually preferable to maximize the natural log of the likelihood function.
  - log transformation is monotonic, so maximizing the log gives the same answer as maximizing the original function
  - original likelihood is usually a product, so log is a sum – easier to differentiate!
  - curvature of log likelihood is related to sample variance of the MLE

$$l(p) = log\binom{n}{y} + y\ log(p) + (n-y)log(1-p), \quad 0 < p < 1$$

- Take first derivative of log likelihood with respect to p, set equal to 0, and solve for p.

$$\frac{dl(p)}{dp} = \frac{y}{p} - \frac{n-y}{1-p} = 0$$
$$\hat{p} = \frac{y}{n}$$

- In example, MLE is

$$\hat{p} = \frac{7}{50} = .14$$

Sampling distribution of the MLE

- The sampling distribution of an estimator is the distribution of the values of that statistic calculated from all possible samples of size n drawn from the population of interest.
  - (or, if the population is infinite or only theoretical, the distribution of values obtained in the limit under repeated sampling)
- binomial setting: sampling distribution of $\hat{p}$ is the probability distribution of the values of this estimator if repeated binomial samples are taken with sample size n and fixed success probability p

- asymptotic approximation to sampling distribution of any MLE when n is "large"
  - generically, if $\hat{\theta}$ is the MLE for a population parameter θ, then

    $$\hat{\theta} \simeq N\left(\theta, -l^{(2)}(\hat{\theta})^{-1}\right)$$

    where
    * $l^{(2)}(\hat{\theta})$ is the second derivative of the log-likelihood evaluated at the MLE.
    * and $N(a,b)$ is the normal distribution with mean a and variance b
- for binomial proportion p,

$$\hat{p} \simeq N\left(p,\ \frac{\hat{p}(1-\hat{p})}{n}\right)$$

Frequentist confidence interval

- The estimated standard deviation of an estimator is called its *standard error*.
- The approximate standard error of $\hat{p}$ is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- a level C confidence interval for population proportion p can be calculated using asymptotic sampling distribution of $\hat{p}$

  $$\left(\hat{p} - z^*_{1-C/2}\ se(\hat{p}),\ \hat{p} + z^*_{1-C/2}\ se(\hat{p})\right)$$

  where $z^*_{1-C/2}$ is the 1 - C/2 quantile of the standard normal distribution.

Back to quitting school example

- $\hat{p} = 0.14$
- $n = 50$
- $se(\hat{p}) = \sqrt{\frac{.14(1-.14)}{50}} = .049$
- for a 90% confidence interval, $z^* = 1.645$
- 90% confidence interval for population proportion p is

  $$(0.14 - 1.645 \times 0.049,\ 0.14 + 1.645 \times 0.049)$$
  $$(0.059, 0.221)$$

- Note: if we had obtained a *different* random sample of 50 students, we would have gotten not only a different $\hat{p}$ but also a different confidence interval.
- Interpretation of this frequentist confidence interval: if
  - we took many, many different random samples from this population

  - and applied this procedure for computing a confidence interval to each of the samples

  then

  - 90% of the resulting confidence intervals would include the true population proportion p
- We don't know whether the particular confidence interval from the sample we actually have is one of the 90% or one of the 10%.
- The frequentist cannot say that there is 90% probability that the true p is in this interval. The true p is some fixed number, even though we don't know what it is. That number is or is not in this interval.

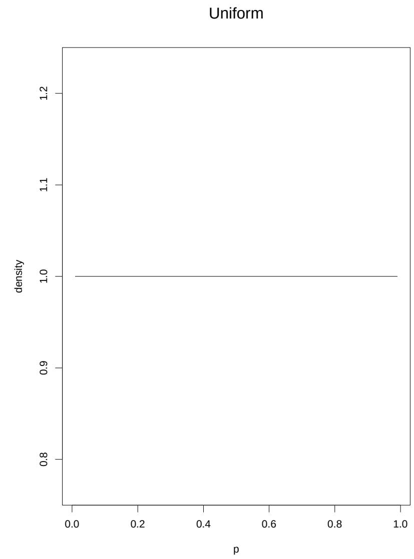
Bayesian inference regarding a proportion
Constructing a prior

- parameter of interest is still the unknown population proportion p
- p could take on any value in interval (0, 1)
- We need to assess our knowledge or belief about this unknown parameter *before* we observe the data from the survey.
- Because p can take on any of a continuum of values, we express this knowledge or belief most appropriately by means of a *probability density function*.
  - unlike our previous problems where there was a discrete set of models to which we assigned probabilities

Constructing a prior, continued

- person who has little or no knowledge about likely values of this proportion
  - might consider all values in (0, 1) equally plausible before seeing any data
  - *uniform* density on (0,1) describes this belief mathematically

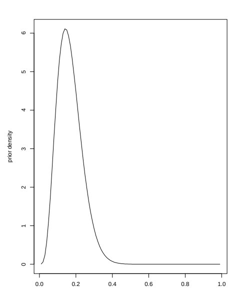
$$p \sim U(0, 1)$$
$$p(p) = 1, \quad 0 < p < 1$$

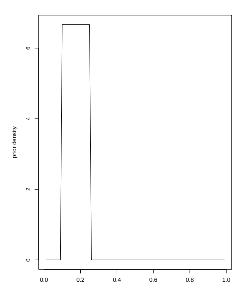


- called "vague" or "noninformative" prior

Other possible priors

- If a person has knowledge or belief regarding likely values of p, his/her prior will be informative
- examples: two different possible priors expressing the belief that the most likely values of p are between .1 and .25





- A histogram prior
- A discrete prior, such as

$$p(p) = \frac{1}{7}, \; p = .1, .125, .15, .175, .20, .225, .25$$

- An infinite number of other possibilities

Updating prior beliefs

- Bayes theorem for probability density functions

$$p(p|data) \propto p(p)\ L(p)$$

- Recall quitting school example and binomial likelihood:

$$L(p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad 0 < p < 1$$

- Combining the prior density and the likelihood to get the posterior density

$$p(p|data) \propto p(p)\, p^y (1-p)^{n-y}, \quad 0 < p < 1$$

- If the uniform prior, $p(p) = 1$ had been chosen, and there were 7 people who said they would quit out of 50 surveyed:

$$p(p|data) \propto p^7 (1-p)^{43}, \quad 0 < p < 1$$

- With the noninformative, uniform prior, the posterior density is proportional to the likelihood function!
  - But the Bayesian and frequentist interpretations are different.
  - The Bayesian says that the population parameter p can be treated as if it were random variable, and the posterior distribution is a probability distribution representing beliefs about its value.
  - The frequentist says that the same curve represents the probability of the sample result (7 successes in 50 trials) for different fixed values of the unknown population parameter p.