CS&SS/Stat 564: Assignment 1

Jeffrey Arnold

4/10/2017

Fork this repository, edit this file with your solutions, and turn in this assignment via pull request.

For help with Markdown and R markdown

- RStudio R markdown site
- R Markdown Cheatsheet
- R Markdown Reference Guide
- Math in R Markdown

This problem set will require loading the following packages:

```
# library()
library("rethinking")
library("boot")
```

You will (probably, but not necessarily) need **boot** for bootstrapping.

1 Statistical Rethinking, Ch. 2

Complete these problems from

1.1 2M1

1.2 2M2

1.3 2H1

1.4 2H2

1.5 2H3

1.6 2H4

2 Statistical Rethinking, Ch 3.

2.1 3H1

2.2 3H2

2.3 3H3

2.4 3H4

2.5 3H5

3 The German Tank Problem

The "German Tank Problem" is so named because it was firstly, or at least famously, used to estimate the total number of German tanks in WWII from the serial numbers of tanks they had destroyed. The general problem is to estimate the size of a population given a sequentially numbered sampled: given that you observe a sample with sequential numbers $\{12, 17, 33, 35, 50\}$, how large is the population from which that sample was drawn?[1] More recently, Gill and Spirling (2015), use this methodology to estimate the total number of US diplomatic cables, and the proportion leaked, from the Wikileaks dump of US diplomatic cables in 2011.

Can you think of, or have you come across another problem, in your own resaearch interests in which this method could be used?

We want to estimate the size of a finite population (N), given sequentially numbered sample of size $n \leq N$ sampled with equal and independent probabilities and without replacement from that population. Let X be the maximum value of that sample, $X = \max(X_1, \ldots, X_n)$. What we want to estimate is N given that we have observed a maximum value of $X = x$. By Bayes rule,

$$\Pr(N|x) = \frac{\Pr(x|N)\Pr(N)}{\Pr(x)}, \text{for } x \leq N < \infty$$

where $Pr(x)$ is the normalizing constant, $\Pr(N)$ the prior distribution of the total number of cables, adn $\Pr(m|N)$ is the probablity that the maximum numbered cable in the sample is m given that N cables were sent.

[1] In that example, it was 100, and the sample was generated via `sort(sample(1:100, 5))`. The other important assumption is that within the population, all obervations are sampled with equal probability and without replacement.

The likelihood, the probability of observing a sample maximum of $x = X$ given n, is

$$\Pr(x|N) = \begin{cases} \frac{\binom{x-1}{n-1}}{\binom{N}{n}} & \text{if } n \leq x \leq N \\ 0 & \text{otherwise} \end{cases},$$

and a log likelihood of

$$\log \Pr(x|N) = \begin{cases} \log\left[\binom{x-1}{n-1}\right] - \log\left[\binom{N}{n}\right] & \text{if } n \leq x \leq N, \\ -\infty & \text{otherwise} \end{cases}.$$

Note that you should always calculate likelihoods on the log-scale given that these probabilities can get too small to represent with floating point numbers.[2] Binomial coefficient and factorials should also always be calculated on the log scale, which is why R provides the function *lchoose*, since binomial coefficients quickly become larger than floating point accuracy The following R function calculates that log-likelihood,

```
maxint_loglik <- function(x, n, N) {
  ifelse(n <= x & x <= N, lchoose(x - 1, n - 1) - lchoose(N - 1, n), -Inf)
}
```

For these examples use the following generated data set:

```
# set.seed(35489)
# n <- 10
# N <- 100
# smpl <- sample.int(N, size = n)
smpl <- c(6, 17, 49, 75, 46, 71, 26, 66, 28, 74)
smpl_max <- max(smpl)
```

3.1 Frequentist Estimators

1. Show that the maximum likelihood estimator of the population size $\hat{N}_{\text{MLE}}$ is x (the maximum integer in the sample). This does not need to be a formal proof. You can show this by calculating the likelihood over a reasonable range of values and finding the maximum.
2. Goodman (1954) provides a minimum variance unbiased estimator of N,

   $$\hat{N}_{\text{Goodman}} = \frac{n+1}{n}x - 1.$$

   What is the minimum unbiased variance estimator for this sample?
3. Calculate 95% confidence intervals for both of these estimators using a simple bootstrap. **Do the confidence intervals make sense?**

Example of bootstrapping confidence intervals for the maximum likelihood estimator.

```
# Number of simulations
nsims <- 2000
# For clarity write a function for the estimator
estimator <- function(x) max(x)
# initialize a vector to save results of the bootstrapping
results <- vector("numeric", nsims)
# repeat `nsims` times:
for (i in seq_len(nsims)) {
  # resample the sample
  newsmpl <- sample(smpl, size = length(smpl), replace = TRUE)
```

[2] If you are not familiar with the term "floating point", see Computerphile (2014), Burns (2011 ch. 1), and Goldberg (1991).

```
  # calculate estimate and save to the
  results[i] <- estimator(newsmpl)
}
# A 95% confidence interval is
quantile(results, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    66    75
```

Now, you can edit the code above and replace `estimator` with a with the Goodman estimator:

```
estimator <- function(x) {
  n <- length(x)
   (n + 1) / n * (max(x) - 1)
}
```

3.1.1 Bayesian Posterior: Proper Uniform Prior and Grid Estimation

We now turn to Bayesian estimation of the population maximum.[3]

The posterior probability of the the population maximum given the sample is $p(N|x,n)$ is

$$p(N|x) = \frac{p(x|N)p(N)}{p(x)}$$

The likelihood $p(x|N)$ is the same as the MLE estimator,

$$\Pr(x|N) = \begin{cases} \frac{\binom{x-1}{n-1}}{\binom{N}{n}} & \text{if } n \leq x \leq N \\ 0 & \text{otherwise} \end{cases},$$

The marginal probabiliry of the data, is the sum of $p(x|N)p(N)$ for all values of N which have non-zero probability in the prior,

$$p(x) = \sum_{m \in \{N : p(N) \neq 0\}} p(x|m)p(x).$$

The first prior we will consider is a proper uniform uniform distribution with a minimum of 0 and a maximum of $N > K$,

$$N \sim U(0, K),$$

which has the probability mass function of

$$p(N) = \begin{cases} \frac{1}{K} & \text{if } N \in \{0, K\} \\ N & \text{otherwise} \end{cases}$$

In this example use $K = 400$.

```
K <- 400
```

In R, you can calculate the probability mass function of the uniform distribution with the function with `dunif`. Even though the pmf calculation for this function is trivial, using it will make your code more readable because it will more clearly expresses your intent than `1 / K`.[4]

We will calculate the Bayesian posterior distribution by grid estimation as described in *Rethinking Statistics*.

[3] See Höhle (2006) for various Bayesian estimators of this problem.

[4] See R for Data Science for a discussion of how code is for humans to read.

1. Suppose that the prior probability for N is distributed uniform between 0 and some maximum value $K \geq N$. $K = 400$.
2. Compute the prior probablity at each value
3. Compute the likelihood at each value
4. Compute the unstandardized and standardized likelihood at each value
5. On the same plot, plot the probability mass functions of the likelihood, prior, and posterior distributions.
6. Calculate the maximum a posterior, mean, and median estimators.
7. Calculate the 95% central credible interval. How does it differ from the frequentist confidence intervals, both in its values and in interpretation.

References

Burns, Patrick. 2011. *The R Inferno*. http://www.burns-stat.com/pages/Tutor/R_inferno.pdf.

Computerphile. 2014. "Floating Point Numbers - Computerphile: Floating Point Numbers - Computerphile." January 22. https://www.youtube.com/watch?v=PZRI1IfStY0.

Gill, Michael, and Arthur Spirling. 2015. "Estimating the Severity of the Wikileaks U.S. Diplomatic Cables Disclosure." *Political Analysis* 23 (02). Cambridge University Press (CUP): 299–305. doi:10.1093/pan/mpv005.

Goldberg, David. 1991. "What Every Computer Scientist Should Know About Floating-Point Arithmetic." *ACM Computing Surveys*. doi:10.1145/103162.103163.

Goodman, Leo A. 1954. "Some Practical Techniques in Serial Number Analysis." *Journal of the American Statistical Association* 49 (265). Informa UK Limited: 97–112. doi:10.1080/01621459.1954.10501218.

Höhle, Held. 2006. "Bayesian Estimation of the Size of a Population." Technical report 499. Sonderforschungsbereich 386. https://epub.ub.uni-muenchen.de/2094/1/paper_499.pdf.