MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #7 (Estimation)

Last modified: March 12, 2014

Reading from Chihara and Hesterberg

- Chapter 6

Optional Reading from Haigh

- The key steps in the almost-proof of the Central Limit Theorem are on pages 110-112.

Proof of the Week

- Suppose that $X_1, X_2, \cdots X_n$ are independent random variables, all with expectation 0, variance 1, and the same moment generating function $M(t)$.

  Define $Z_n = \frac{1}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n)$,
  and call its moment generating function $H_n(t)$.

  Using the fact that if $\lim_{n\to\infty} nr_n = 0$ then

$$\lim_{n\to\infty}(1 + \frac{x}{n} + r_n)^n = e^x$$

  prove that

$$\lim_{n\to\infty} H_n(t) = e^{\frac{t^2}{2}}$$

  This is a special case of the proof on page 110 of Haigh)

R scripts

- Script 7A-MLE estimates.R

  Topic 1 - Maximum likelihood estimate of a parameter
  Topic 2 - using the mle() function to estimate one or more parameters
  Topic 3 - estimating two parameters to fit some real-world data

- Script 7B-EstimationMoments.R

  Topic 1 - choosing one parameter to make the theoretical mean match the sample mean
  Topic 2 - using two sample moments to estimate two parameters

- Script 7C-BiasEfficiency.R

  Topic 1 - looking for an efficient estimator
  Topic 2 - an efficient estimator of the parameter p in a Bernoulli distribution

- Script 7D-Consistent.R

  Topic 1 - the sample mean is a consistent estimator of the expectation.
  Topic 2 - an example of an estimator that is not consistent.
  Topic 3 - the sample mean can be a consistent estimator even when the variance is infinite.

- Script 7P-Proof7.R

  Topic 1 - A famous limit that you met in a calculus course
  Topic 2 - the mgf for the sum of many discrete random variables
  Topic 3 - the mgf for the sum of many continuous random variables

Mathematical notes

1. Prove that if $x_1, x_2, \cdots x_n$ are an independent random sample from a normal distribution with unknown parameters μ and σ, the maximum-likelihood estimators of the parameters are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

2. Suppose that $x_1, x_2, \cdots x_n$ are a random sample from a uniform distribution Unif$[0, \beta]$ with unknown parameter β. Show that the maximum likelihood estimator of β is $\max x_i$ and that this estimator can be made unbiased by multiplying it by $(n+1)/n$.

3. Proof of the week

Suppose that $X_1, X_2, \cdots X_n$ are independent random variables, all with expectation 0, variance 1, and the same moment generating function $M(t)$.

Define $Z_n = \frac{1}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n)$,
and call its moment generating function $H_n(t)$.

Using the fact that if $\lim_{n\to\infty} nr_n = 0$ then

$$\lim_{n\to\infty}(1 + \frac{x}{n} + r_n)^n = e^x$$

prove that

$$\lim_{n\to\infty} H_n(t) = e^{\frac{t^2}{2}},$$

the moment-generating function of the standard normal distribution.

(This is a special case of the proof on page 110 of Haigh)

4. (If time permits). This is a special case of the sequence of theorems on pages 158 and 159.

   Suppose that $X_1, X_2, \cdots, X_n$ are independent random variables from a distribution with mean μ and variance σ^2. Let $\overline{X_n}$ denote the sample mean. We have shown that $E[\overline{X_n}] = \mu,\ \mathrm{Var}[\overline{X_n}] = \sigma^2/n$.

   Show that the sample means are a consistent sequence of estimators for μ, in the sense that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| \geq \epsilon) = 0.$$

5. Mean-square error when estimating a proportion

This covers the example on page 156 and its generalizations (examples in script 7C and section problem 2)

There are n trials, each with an unknown probability p of success. We artifically add t trials with s successes, usually because we have reason to suspect that the parameter p is close to s/t. So our estimator is

$$\hat{p} = \frac{X+s}{n+t},$$

where X is a binomial random variable with paramenters n and p, while s, n, and t are just constants. The mean-square error is

$$MSE = E[(\hat{p}-p)^2] = E[(\frac{X+s}{n+t} - p)^2] = \frac{1}{(n+t)^2}E[(X+s-np-tp)^2].$$

$$MSE = \frac{1}{(n+t)^2}E[((X-np)+(s-tp))^2]$$

By linearity,

$$MSE = \frac{1}{(n+t)^2}(E[(X-np)^2] + 2E[(X-np)(s-tp)] + E[(s-tp)^2])$$

The first term is the variance of a binomial distribution. The middle term is zero because $E[X] = np$. The last term is a constant. So

$$MSE = \frac{np(1-p)+(s-tp)^2}{(n+t)^2}.$$

This agrees with the equation in the middle of page 156 if we set $t = 2, s = 1$.

Section problems

1. The NFL recently moved kickoffs from the 30-yard line to the 35-yard line, and they may change the rule again. To investigate the effects of another change, the rules committee is sponsoring some practice games in which the position of the ball at kickoff is a random variable X whose distribution is Unif$[\alpha, \beta]$. The parameters α and β are a well-guarded secret, but your spies have piloted the MetLife blimp over the practice field and oberved kickoffs from the following yard lines:
   (These data were generated by R using `runif(10, alpha, beta)`.)

   $X_1 = 34.5, X_2 = 32.0, X_3 = 36.8, X_4 = 31.5, X_5 = 37.2, X_6 = 38.5, X_7 = 29.1, X_8 = 35.6, X_9 = 36.6, X_{10} = 36.2$

   Different students can try two different ways of estimating α and β

   (a) Using the method of moments, as in problem 14 on page 162. See page 398 for the variance of the uniform distribution.

   (b) Using the minimum and maximum to get unbiased estimates of α and β.

   Jonathan knows the values of α and β that I used to generate the data and will reveal them once estimates have been announced on Piazza.

2. NFL kickers make 40-yard field goals about 80% of the time.

   You are running a tryout camp where prospective kickers attempt 10 40-yard field goals and you count the number of successsful attempts, x. You want to estimate the probability p of success on such a field goal from the results of this tryout, and you are considering two different estimators:

   - $\hat{p}_1 = x/10$
   - $\hat{p}_2 = (x + 4)/15$

   Using the approach of script 7C, Topic 2, find the range of values of p for which the second estimator has a smaller mean-squared error than the first.

3. The lifetime of a Mortal Cyberpet is a random variable X that has an exponential distribution with a scale parameter T (weeks) that specifies the expected lifetime. Feeding a cyberpet costs 1 dollar per week, charged to Mom or Dad's Mastercard or Visa.

   In a certain fifth-grade class each of the 30 students acquires exactly r cyberpets. After all these pets have finally died, parents report to the school principal the total amount spent on feeding their kid's cyberpets, a random variable which has a gamma distribution with shape r, scale T. The results are in the file `Cyberpets.csv` on the Class 7 page.

   No one thought to tell the principal the values of T and r. Estimate these parameters for her in two different ways.

   (a) by using the method of moments (see page 398). The "rate" parameter for the gamma distribution is $\lambda = 1/T$.

   (b) by using the `mle()` function with the sum of the log of !dgamma()! To get a reasonable initial estimate, note that the expected feeding cost equals rT, then make a guess about r.

   (c) by trying integer values of r between 5 and 30, in each case choosing T so that rT equals the average food bill, and determining what r maximizes the product of the density functions for the observed data (or the sum of their logarithms).

   Again, Jonathan knows the values of r and T that I used to generate the data and will reveal them once some estimates have been announced on Piazza.

Homework assignment - due on April 2, a week after midterm projects
This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 7 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. Exercise 17 on page 162 – Weibull model for earthquakes.
   Please do this problem by modifying the analysis done in the authors' script (URL is on page146, and there is also a copy on the "Textbook Scripts" page of the course Web site.) You can check your answer using the mle() function. Also study the derivation of equations (6.12) and (6.13), for which there was insufficient time during class. The Weibull distribution might be an interesting theme for a final project, since it seems to model complex natural phenomena surprisingly well.

   Include the graphs that are requested, but do not take the time to replicate the goodness-of-fit analysis, since that shows up in the next problem.

2. Exercise 16 on page 162. So far, all you know about the gamma distribution is that it arises as the sampling distribution for the exponential distribution, in which case the "shape" parameter must be an integer. However, the formulas on page 398 are valid even when r is nor an integer. The goodness-of-fit analysis can be modeled on the wind speed analysis

3. (This is equivalent to problem 20 in the textbook)

   Karl is applying for membership in the Stats Guild. He is required to submit the results of five different tests, each of which yields a score in the interval $[0, 1]$. Since he just guesses, his score on each attempt at a test is a random variable with the distribution Unif[0,1].

   Karl may not be very bright, but he is rich, and so he pays to take each test θ times, submitting only the best score for each. Here are his results, pasted in from a simulation that I did in R.

   $X_1 = 0.855, X_2 = 0.891, X_3 = 0.913, X_4 = 0.989, X_5 = 0.943.$

   The pdf for the maximum of θ samples from Unif[0,1] is

   $f(x;\theta) = \theta x^{\theta-1}, 0 \leq x \leq 1, \theta > 0$

   (a) Find the MLE of θ if it is not required to be an integer.

   (b) Find the method of moments estimate of θ.

   (c) Find the MLE estimate of θ, requiring it to be an integer. Just have R crank out the likelihood of the given results for $\theta = 1, 2, 3, \cdots, N$ choosing N large enough that you are sure that you have found the maximum.

After the homework has been submitted, I will post the short script that I used to generate the data for this problem. There is no guarantee that the MLE estimate of the parameter is equal to the value that I used.

When I Googled "taking SAT multiple times," I found no suggestion that colleges might use this sort of analysis to compensate for the fact that some students submit only the best of n SAT scores. There might be an interesting term project here!

4. Carry out a simulation in R to confirm the results of parts (a) and (d) of Exercise 37 on page 164. By doing the sampling N times, you can get an excellent estimate of the expectation of each of the two estimates.

5. The Pareto distribution with shape 1 and scale s has density function

$$f(x) = \frac{s}{(x+s)^2}, x \geq 0.$$

It is supported by the "actuar" package, which you probably installed earlier. If not, remove the # from the first line below. Then

```
#install.packages("actuar")
library(actuar)
x <- rpareto(10, 1 , 2); x
```

will draw a sample of size 10 from a Pareto distribution with shape 1 and scale 2.

(a) Investigate the behavior of sample means for samples of various sizes drawn from this distribution, and do a plot like the one for the Cauchy distribution in script 7D to show that the sample means are not a consistent estimator of anything.

(b) Make a histogram of sample medians for samples of size 10 from the Pareto distribution with shape 1, scale 2.

(c) Invent a consistent estimator of the parameter s and show that it works by creating a graphic like figure 6.7 in the textbook, where you display the result of taking the median of larger and larger samples.

Note: to do this problem you are going to have to work out the distribution function for the Pareto distrbution, attempt to calculate its expectation, and figure out its median. An R script is not a good medium for showing this sort of mathematics, but at least state your answers.