MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #6 (Bootstrapping)

Last modified: March 5, 2014

Reading from Chihara and Hesterberg

- Chapter 7, Section 7.1.1 – an introduction to confidence intervals
- Chapter 5 – the bootstrap

Proof of the Week

- A normal random variable $X \sim N(\mu, \sigma^2)$ has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Prove the following:

- The moment generating function of X is

$$M(t) = e^{\mu t + \sigma^2 t^2/2}.$$

- $E[X] = \mu$.
- $\mathrm{Var}[X] = \sigma^2$.
- If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, and X_1 and X_2 are independent, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

(This is partially done on pp. 368-369)

R scripts

- 6A-ConfidenceInterval.R
  Topic 1 – sampling from a normal distribution
  Topic 2 – sampling from a gamma distribution

- 6B-BootstrapIntro.R
  Topic 1 - A bootstrap sampling distribution
  Topic 2 – trying a bootstrap where we know the population distribution
  Topic 3 – the bootstrap reveals skewness in the population

- 6C-BootstrapRealData.R
  Topic 1 - a case where the population distribution is both unknown and skewed
  Topic 2 - a two-sample bootstrap
  Topic 3 – Bootstrapping other statistics

- 6D-BootstrapCardio.R
  Topic 1 – doing a bootstrap when we have only a couple of proportions to work with

- 6P-Proof 6.R
  Topic 1 – density functions for normal distributions
  Topic 2 – moment generating function for a normal distribution

Mathematical notes

1. Confidence intervals as random variables

   Suppose that we are drawing samples of size n from a known population with mean μ and variance σ^2. The sample mean $\overline{X}$ is a random variable whose expectation is also μ. Let α be a smallish number, typically $\alpha = 0.05$. Then a "$1 - \alpha$ confidence interval" is specified by two random variables L and U with the property that

   $$P(L \geq \mu) = P(U \leq \mu) = \alpha/2.$$

   Thus the probability of the event $L < \mu < U$ is $1 - \alpha$ (typically 95%).

   Let q_1 and q_2 denote the $\alpha/2$ and $(1 - \alpha/2)$ quantiles for the sampling distribution of X.

   (a) Show that $U = \overline{X} + \mu - q_1$ and that $L = \overline{X} + \mu - q_2$. Explain why it is reasonable for U to depend on the "lower" quantile q_1 and vice versa.

   (b) Show that if the sampling distribution is symmetical about μ, then $U = \overline{X} + q_2 - \mu$ and that $L = \overline{X} - (\mu - q_1)$.

   (c) For the normal distribution $N(0, 1)$ and $\alpha = .05$, $\mu = 0$, $q_1 = -1.96$ and $q_2 = 1.96$. Show that if the central limit theorem applies, then $L = \overline{X} - 1.96\sigma/\sqrt{n}; U = \overline{X} + 1.96\sigma/\sqrt{n}$.

2. Proof of the week

A normal random variable $X \sim N(\mu, \sigma^2)$ has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Prove the following. For parts b-d use moment generating functions.

- The moment generating function of X is

$$M(t) = e^{\mu t + \sigma^2 t^2/2}.$$

- $E[X] = \mu$.
- $\text{Var}[X] = \sigma^2$.
- If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, and X_1 and X_2 are independent, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Section problems Scripts that solve these problems can be posted to the "Section problem scripts" topic box on the Class 6 page.

1. Exercise 8 on page 131. This is similar to example 5.2. If someone posts a script for parts (a) - (d), others can use it to do part (e).
2. Exercise 10 on page 131. A quick edit of part of script 5B will do most of what is needed.
3. Exercise 13 on page 132. One answer to part (c) is on page 402; it would be interesting to see others during section.

Homework assignment 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 6 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. "Mortal cyberpets," whose lifetime is specified by an exponential distribution, have become big business, and you have been hired by Consumer Reports to write a feature article on them. It turns out that many youngsters have been buying 20 of these cyberpets, keeping vital statistics on them, and then complaining when the mean lifetime for their sample is less than the expected lifetime advertised by the pets' creator.

   So you plan to publish a formula for a confidence interval that has a 95% probability of including the true expectation of the lifetime, along with a warning that there is stiil a 2.5% chance that the true expected lifetime will lie outside the interval on either side. Since the central limit theorem does not give a good approximation to the mean of a sample of just 20 exponential random variables, you plan to develop and test a formula for the endpoints of the confidence interval by using the exact sampling distribution (a gamma distribution with shape = 20), then see if you can replicate your results by using a bootstrap percentile confidence interval from a single sample.

   For simplicity, measure time in lunar months (1 month= 28 days) and take lambda = 1. Replicate the approach of script 6A, with the bootstrap then done as in script 6B. It is OK to paste and edit code from these scripts. Include a graphical display of 100 confidence intervals.

   My suspicion is that the bootstrap percentile confidence interval will pass the 95% test reasonably well but that when the confidence interval fails to include 1 month it is far more likely to miss on one side than on the other. We will try after spring break to deal with this problem!

2. Exercise 12 on page 132.

3. Exercise 17 on page 133. You can check your numerical answers against page 402.

4. Of the students who entered Harvard as freshmen in Fall 1997, 41 of 819 men and 23 of 756 women failed to graduate within 5 years. Conduct a bootstrap analysis, in the manner of Example 5.7, of the relative risk of failing to graduate within five years based on gender. Include a plot like figure 5.17 that illustrates the bootstrap percentile confidence interval for the relative risk.