CALIFORNIA STATE UNIVERSITY, HAYWARD
STATISTICS DEPARTMENT

Statistics 6601-01 Advanced Statistical Computing
Fall 2002

Midterm - Solution

1. **(30 points)**

(a) What is the sampling distribution of $\bar{x}$ for a random sample of size n from a $N(\mu, \sigma^2)$ distribution?
Ans: By the Central Limit Theorem, for a sample from a normal population, $N(\mu, \sigma^2)$,

$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1) \tag{1}$$

as $n \to \infty$. Or as a shorthand notation

$$\bar{x} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2}$$

The most important issue here is that $\bar{x}$ has a specific Normal distribution with mean μ, so $\bar{x}$ is unbiased, and it has a standard error that is divided by n. The division by n gives the increase in accuracy of the *statistic* as a estimator for the *parameter* μ.

(b) Define the term *coverage probability* when using the $100(1-2\alpha)\%$ one-sample t confidence interval for the population mean μ.

Ans: The coverage probability of the one-sample t confidence interval can be explained as follows: In repeated sampling from a normal population, $N(\mu, \sigma^2)$, approximately $100(1-2\alpha)\%$ of the confidence intervals

$$\bar{x} \pm t_{n-1}(\alpha)\frac{s}{\sqrt{n}} \tag{3}$$

will cover the (unknown, fixed) parameter μ and approximately $100(2\alpha)\%$ of the confidence intervals will not cover the parameter μ. In probability notation, for a random sample from the population, $X_1, X_2, ..., X_n$,

$$P\left(\bar{X} - t_{n-1}(\alpha)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1}(\alpha)\frac{S}{\sqrt{n}}\right) = (1 - 2\alpha) \tag{4}$$

where $\bar{X}$ and S are random variable since before an actual sample is collected these statistics are random variables.

(c) Explain how you would interpret the *coverage probability* when using a $100(1-2\alpha)\%$ Bootstrap confidence interval for the population mean μ.

Ans: The coverage probability of a one-sample Bootstrap confidence interval (the theory works for any type of Bootstrap CI) is essentially the same as for the classical one-sample t confidence interval. In repeated sampling from the population of interest approximately $100(1-2\alpha)\%$ of the bootstrap confidence intervals will cover the (unknown, fixed) parameter μ and approximately $100(2\alpha)\%$ of the confidence intervals will not cover the parameter μ. For each sample from the population B Bootstrap resamples are taken from the sample and a single bootstrap confidence interval is computed. In $R = 2000$ resamples are taken for each of the samples $B = 1000$ resamples are taken from the sample to produce the bootstrap sampling distribution and bootstrap confidence interval. This is why the coverage probability simulation program runs so slowly in comparison to the classical coverage probability simulation program.

2. **(40 points)** Suppose $x_1, x_2, ..., x_n$ are the ages of a sample of $n = 60$ chief executive officers (CEOs) of successful small businesses. Assume the population of chief executive officers of successful small businesses is large. Estimate the population variance σ^2 of the distribution of the ages of chief executive officers using

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{5}$$

(a) What statistic would you use to estimate the population variance?

Ans: The sample variance s^2, the unbiased estimate of σ^2. The purpose of this question is to make sure you realize that a *sample statistic* estimated a *population parameter*.

(b) Write the formula you would use to compute the standard classical confidence interval for σ^2. Show using a probability argument that it has the appropriate coverage probability.

Ans: See pages 69-71 in Lunneburg. Or see Ott. Since

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{6}$$

we can derive a $100(1-2\alpha)\%$ confidence interval for σ^2 having the correct coverage probability.

(c) Describe how you would produce the bootstrap estimate of σ^2. Also, give the formula you would use to estimate the bias of the estimator.

Ans: If we assume that the sample size is a small portion of the population of CEO's of small companies then we can use the nonparametric bootstrap where we resample with replacement from the original sample.

To compute a bootstrap confidence interval for σ^2 we would resample with replacement from the original sample of $n = 60$ of CEO's B times and compute s^2 B times. A histogram of these B bootstrap sample variances $t^* = s^{2^*}$ would give an estimate the the sampling distribution of s^2. The mean of these B s^{2^*} values would give the bootstrap estimate of the parameter σ^2, which would be the center of the bootstrap sampling distribution,

$$\frac{1}{B} \sum_{j=1}^{B} {s^2}^*_j \tag{7}$$

and the standard deviation of the B s^{2^*} values would give the bootstrap estimate of the standard error of the bootstrap sampling distribution of s^2.

$$\frac{1}{B-1} \sum_{j=1}^{B} \left({s^2}^*_j - \frac{1}{B} \sum_{j=1}^{B} {s^2}^*_j \right)^2 \tag{8}$$

To estimate the bias

$$\frac{1}{B} \sum_{j=1}^{B} {s^2}^*_j - s^2 \tag{9}$$

where s^2 is the sample standard deviation from the original sample of $n = 60$.

(d) Describe how you would use the Bootstrap sampling distribution of σ^2 to compute a empirical Bootstrap confidence interval for σ^2.

Ans: Take the lower α percentage point and upper α percentage point from the bootstrap sampling distribution.

3. **(40 points)** A soda company takes a sample $n = 100$ cans of soda from its assembly line every day to check to see if the line is within the manufacturing specifications of no more than 10% and no less than 5% underweight cans of soda on average per day. These limits were determined to maximize profits and minimize the likelihood of complaints by customers.

   The quality control manager has been using the the standard one-sample z confidence interval to do the calculation of the confidence interval for p, the population proportion of defective (under weight) cans.

$$\hat{p} \pm z(\alpha)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{10}$$

   Suppose today's random sample contains 5 under weight cans of soda.

   (a) Why might the standard one-sample z confidence interval not be accurate near zero?

   Ans: Since $\hat{p} = 5/100 = 0.05$ is so close to zero. The one-sample z confidence interval works best when $\hat{p}$ is in the middle of 0 and 1, that is when it is close to 0.50. Also, note that the standard checks, $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$ are barely met.

   (b) Describe the shape of the bootstrap distribution of the sample proportion $\hat{p}$ in Fig. 1.

   Ans: The distribution looks skewed to the right. So if we use the normal distribution we might get confidence intervals that are incorrect since the upper tail of the sampling distribution of $\hat{p}$ might not be normal.

   (c) Use the Minitab and S-Plus output below to compare the results of computing a classical 95% confidence interval for p and the 95% bootstrap confidence interval for p. Why might a bootstrap 95% confidence interval for p be more accurate? What do these results say about the *robustness* of the one-sample z confidence interval for p?

   Ans: Classical 95% one-sample z confidence interval for p. (0.007284, 0.092716). Empirical 95% Bootstrap confidence interval for p. (0.02, 0.09025) BCa 95% Bootstrap confidence interval for p. (0.01592, 0.09) The bootstrap confidence intervals might be more accurate since they are taking into consideration the non-normality of the sampling distribution of the sample statistics $\hat{p}$. However, the classical z confidence interval for p gives results that are very close to the Boostrap results and the conclusion reached are the same. So the classical method is quite robust to violations in the assumptions.

   Also note that the classical z confidence interval is based on the same sample size as the bootstrap confidence interval calculation. The B in the Bootstrap is like n going to infinity. The bootstrap is NOT creating more data out of thin air! The sample size n is the same for both methods.

   (d) Is the batch acceptable or unacceptable?

   Ans: The null hypothesis is

$$H_0 : p \in [0.05, 0.10]. \tag{11}$$

   Since the confidence interval overlap the null hypothesis we fail to reject the null hypothesis. Therefore, the batch of soda is acceptable.

4. **(40 points)** Suppose $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ is a random sample from the joint normal distribution $N(\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho)$. The population correlation between X and Y is defined as

$$\rho = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X])^2]}\sqrt{E[(Y - E[Y])^2]}} \tag{12}$$

and recall that the sample correlation coefficient is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}. \tag{13}$$

(a) Why is the sample correlation coefficient r considered to be a plug-in estimator of the population correlation ρ?

Ans: Since the formula for r is based on the formula for ρ.

(b) Can the sampling distribution of r be normal? Yes or No. Explain.

Ans: No. This is because r can only take values between o and 1.

(c) Describe how you would produce the nonparametric bootstrap estimate of ρ. (Hint: Sample pairs of observations.) Describe how you would compute the 95% empirical bootstrap confidence interval for ρ from the bootstrap distribution of r.

Ans: From the original sample, re-sample pair (x_i, y_i) n times and recalculate r B time. A histogram of the values of r give the bootstrap sampling distribution of r. From the bootstrap sampling distribution of r one can select the appropriate percentiles to get an the empirical bootstrap confidence interval for ρ.

(d) Fisher's z-transformation is useful for transforming the sample correlation coefficient r, when sampling from a bivariate normal distribution, to a standard normal distribution

$$\mathbf{t}(r) = \left(\frac{1}{2}\right)\left\{log\left[\frac{1+r}{1-r}\right] - log\left[\frac{1+\rho}{1-\rho}\right]\right\}/\sqrt{1/(n-3)} \rightarrow N(0,1) \tag{14}$$

as $n \rightarrow \infty$. Describe how you would create a bootstrap t confidence interval for ρ. Use the following S-Plus output to help guide your answer.

Ans: When re-sampling as described above compute $\mathbf{t}(\mathbf{r})$ many times and back-transform to produce the Bootstrap t confidence interval for ρ.