BMI 713: Computational Statistics for Biomedical Sciences

Assignment 5

October 7, 2010 (due October 14)

1. Sampling Distribution of Sample Proportions

(a). Suppose we have a random sample of size 100 from a Binomial distribution with the population proportion of 0.3. What are the expected mean $E(\hat{p})$ and variance $Var(\hat{p})$ of the sample proportion $\hat{p}$? What is the sampling distribution of the sample proportion $\hat{p}$?

(b). Generate 1000 samples of size 100 from a Binomial distribution with population proportion of 0.3, calculate the sample proportion for each sample, and save them in the vector `sample.p`.

(c). Calculate the mean and variance of the 1000 sample proportions, and compare to the results in (a).

(d). Plot the density of the sample proportions, and compare it to the curve of Normal distribution with mean `E.p` and variance `V.p`, where `E.p` is the expected mean $E(\hat{p})$ and `V.p` is the variance $Var(\hat{p})$ from (a), using the following commands:

```
plot(density(sample.p), main="Density Plot of Sample Proportions", xlab="Sample
Proportions", xlim=c(0.1, 0.5), col="red")
curve(dnorm(x, mean=E.p, sd=sqrt(V.p)), from=0.1, to=0.5, add=TRUE, col="green")
```

2. Hypothesis Testing

Alcohol and breast cancer: The following are partial results from a *case-control* study involving a sample of *cases* (women with breast cancer) and a sample of *controls* (demographically similar women without breast cancer). (Data from L. Rosenberg, *et. al.*, A Case-Control Study of Alcoholic Beverage Consumption and Breast Cancer, *American Journal of Epidemiology* 131 (1990): 6-14). Are the occurrences of women breast cancer related to drinking habits?

	Breast Cancer	
	Cases	Controls
Fewer than 4 drinks per week	330	658
4 or more drinks per week	204	386

(a). State the null hypothesis and the alternative hypothesis.

(b). Compute the sample proportions of women with breast cancer among women who have fewer than 4 drinks per week and who have 4 or more drinks per week, $\hat{p}_1$ and $\hat{p}_2$ respectively.

(c). How large should the sample sizes be for the adequacy of the normal approximation?

(d). What test do you need to perform for the hypothesis testing? What is the value of the test statistic? What is the *p-value*? What conclusion can you make?

3. Confidence Interval

(a). Given a sample of size n from a population with proportion p (unknown), and the estimated sample proportion $\hat{p}$ (known), write a function `CI(n, p.hat, a)` in R to calculate the confidence interval (CI) for p, at the significance level of α. The function should return a vector of two elements: the first is the lower bound of the CI, and the second is the upper bound of the CI. Note: for this problem, please do not use the function `prop.test` or `binom.test`.

(b). Generate 1000 samples of size 50 from a Binomial distribution with proportion of 0.75, and use the `CI` function in (a) to derive the 95% confidence interval of the population proportion p for each sample, based on the estimated sample proportion $\hat{p}$ (pretending that p is unknown). How many times out of the 1000 simulations do the confidence intervals contain the true population proportion p (i.e., 0.75)? How would you interpret confidence interval based on this problem?