QBS 120 - Problem Set 5

Rob Frost

1. Point and interval estimation for the first gene in the classic Golub microarray gene expression data set. Assume the observed values of the first gene in the Golub data can be modeled as i.i.d $\mathcal{N}(\mu, \sigma^2)$. The Golub data can be found in the *multtest* R package, which can be downloaded from the Bioconductor repository (see http://www.bioconductor.org/packages/release/bioc/html/multtest.html for details and download instructions).

   The normalized expression values for the first gene can be found using the following R code:

```
> library(multtest)
> data(golub)
> gene1.values = golub[1,]
> gene1.values[1:5]

[1] -1.45769 -1.39420 -1.42779 -1.40715 -1.42668
```

   (a) What is the MLE of μ ($\hat{\mu}_{mle}$)?
   (b) Is your estimate from part a) unbiased? Justify.
   (c) Is your estimate from part a) consistent? Justify.
   (d) If you cannot assume the data are normally distributed, is your estimate from a) for $E[X] = \mu$ still valid?
   (e) What is the MLE of σ^2 ($\hat{\sigma}^2_{mle}$)?
   (f) Is your estimate from part e) unbiased? Justify.
   (g) Is your estimate from part e) consistent? Justify.
   (h) If you cannot assume the data are normally distributed, is your estimate from e) for $Var(X) = \sigma^2$ still valid?
   (i) What is the distribution of $\hat{\mu}_{mle}$? If you cannot assume the data are normally distributed, does this sampling distribution still hold?
   (j) What is the MSE of the $\hat{\mu}_{mle}$ estimate?
   (k) Use the distribution of $\hat{\mu}_{mle}$ to compute a 95% CI for μ. Assume that the $\sigma^2 = \hat{\sigma}^2_{mle}$.

2. (Based on Rice 8.3) One of the earliest applications of the Poisson distribution was made by Student (1907) in studying errors made in counting yeast cells or blood corpuscles with a haemacytometer. In this study, yeast cells were killed and mixed with water and gelatin; the mixture was then spread on a glass and allowed to cool. Four different concentrations were used. Counts were made on 400 squares, and the data are summarized in the data.frame below. In this data.frame, each of the "concen.*" columns records the number of squares

associated with that concentration for which the number of counted cells equals the value in the "cells" column.

```
> yeast.counts = data.frame(cells=0:12,
+          concen.1 = c(213,128,37,18,3,1,0,0,0,0,0,0,0),
+          concen.2 = c(103,143,98,42,8,4,2,0,0,0,0,0,0),
+          concen.3 = c(75,103,121,54,30,13,2,1,0,1,0,0,0),
+          concen.4 = c(0,20,43,53,86,70,54,37,18,10,5,2,2))
```

(a) Compute the MLE estimate of the parameter λ for each of the four sets of data.

(b) Approximate the theoretical standard error of the $\hat{\lambda}$ values computed for Problem 1 part a). Do not use simulation.

(c) For the $\hat{\lambda}$ values compute for Problem 1 part a), estimate the standard error using the parametric bootstrap. How do these values compare to the approximate theoretical values? Do these results match you expectations?

(d) Find an approximate 95% confidence interval for each estimate.

(e) Compare observed and expected counts.

3. (Based on Rice 8.9) How would you respond to the following argument? This talk of sampling distributions is ridiculus! Consider Example A of Section 8.4. The experimenter found the mean of the number of fibers to be 24.9. How can this be a "random variable" with an associated "probability distribution" when it's just a number? The author of this book is guilty of deliberate mystification!

4. (Based on Rice 8.13) In Example D of Section 8.4, the MOM estimate was found to be $\hat{\alpha} = 3\bar{X}$. In this problem, you will consider the sampling distribution of $\hat{\alpha}$.

   (a) Show that $E[\hat{\alpha}] = \alpha$, i.e., the estimate is unbiased.

   (b) Show that $Var(\hat{\alpha}) = (3 - \alpha^2)/n$. Hint: What is $Var(\bar{X})$?

   (c) Use the CLT to deduce that a normal approximation to the sampling distribution of $\hat{\alpha}$. According to this approximation, if $n = 20$ and $\alpha = 1$, what is the $P(\hat{\alpha} > 0.5)$? Define in terms of $\Phi()$, the CDF for the standard normal.

5. (Based on Rice 8.58) For a population in Hardy-Weinberg equilibrium, alleles occur with the following frequencies:

   - AA: $(1 - \theta)^2$
   - Aa: $2\theta(1 - \theta)$
   - aa: θ^2

   For a specific sample of 190 people, the haptoglobin types occurred as follows:

   $$X_1 : Hp1 - 1 : 10$$
   $$X_2 : Hp1 - 2 : 68$$
   $$X_3 : Hp2 - 2 : 112$$

   Assume the haptoglobin genotype for this population is in Hardy-Weinberg equilibrium.

(a) Find the mle of θ.

(b) Find the asymptotic variance of the mle.

(c) Find an approximate 99% confidence interval for θ.

(d) Use the parametric bootstrap to estimate the sampling distribution of the MLE of θ. Plot this distribution along with the asymptotic distribution. How does the shape of the bootstrap sampling distribution compare to the asymptotic distribution?

(e) Use the bootstrap sampling distribution to estimate the variance of the MLE of θ. How does the bootstrap variance compare with the asymptotic variance?

(f) Compute the 99% CI for θ using the bootstrap percentile approach. How does this compare with the CI computed in part c)?