Chapter 6 Review with Applications for Math 30: Estimators and Sufficiency

1. Theory Practice.

Consider n observations sampled from a distribution with pdf given by:

$$f(x|\theta) = (\theta + 1)x^{\theta}, 0 \leq x \leq 1,$$

and 0, otherwise.

a. Find the likelihood function for the n observations.

b. Identify a sufficient statistic for θ.

c. Find the MLE for θ.

d. Is the MLE minimal sufficient? Why or why not?

e. Rexpress the pdf so that it can be identified as a member of the exponential family of distributions (not exponential, just in the family).

f. Based on the distribution being in the exponential family, what other statistic can be shown to be sufficient?

2. Application.

The data set whalevelocity (available online if you want) contains 210 whale velocities - time in hours that it took a whale to travel 1 kilometer. The velocities were computed based on paired distance measures at known times for the same whale. The attached sheet contains a histogram, boxplot, and summary statistics from the data.

a. First assume that the velocities can be modeled as Exp(θ), where θ is unknown. What is the MLE of θ (formula and value using data)?

b. Now assume that you had previous information that the value of θ was near 2. What distribution could be used to model θ that would be in a conjugate family and incorporate the information that θ is near 2? (many possible answers for specific hyperparameters for the prior)

c. Provide the hyperparameters for the posterior distribution of θ given the prior that you chose and compute the Baye's estimator under squared error loss using that posterior distribution.

d. If the original prior had been a Gamma($\alpha_0 = 24, \beta_0 = 12$), what would the Baye's estimator under SEL have been? Is that similar to your estimator in c.?

e. Now assume that an exponential model for the velocities was not appropriate to start with. Perhaps a Gamma distribution is more appropriate. What is the log-likelihood for data from a Gamma distribution where both α and β are unknown?

f. Does it look like the MLEs for α and β are easily computable?

The computer program R is a tool that many statisticians use. It can perform hypothesis tests and generate confidence intervals, perform many more advanced statistical procedures, and also functions as a basic calculator, etc. R can also simulate from many different distributions and also "fit" distributions to data - estimating the parameters of those distributions by MLE methods. Many other programs (Matlab) also fit distributions to data. In R (which is free software), the function *fitdistr* requires the MASS library, but can be run as follows for data contained in *variable*:

library(MASS)
fitdistr(variable, "exponential")
or
fitdistr(variable, "gamma")

Notes on the gamma distribution in R specify that shape = α and rate = β, because in R, scale = $1/\beta$ in the provided density function.

Fit exponential and gamma distributions to the data (I will read in the data and verify syntax). Verify the computed log-likelihood is correct for the gamma distribution for the MLE estimates. Test a few other values for α and β and verify that the log-likelihood does appear to be maximized for the MLE estimates.

Sample from exponential and gamma distributions based on the MLE estimates. Make histograms and boxplots for your samples. Which histogram more closely resembles the whale velocity data? What about boxplots? Which distribution would you prefer to use to model the whale velocities?

Commands:
var1=rexp(1000,θ) generates 1000 observations from an Exp(θ) distribution and stores them in var1
var2=rgamma(1000,α, β) generates 1000 observations from a Gamma(α, β) distribution and stores them in var2
hist(var1) and *boxplot(var1)* are example histogram and boxplot creation commands