Solution

Chapter 8 and 9 Review – Applications with Whale Velocities (with Theoretical Practice Along the Way)

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. Some graphs and basic descriptive statistics can be found on the next page (along with the R commands I used to generate them).

First assume that the velocities can be modeled as Exp($\theta$), where $\theta$ is unknown.

a. Find the likelihood function for the 210 observations. (You can use n=210 if you want).

$f(y|\theta) = \frac{1}{\theta} e^{-y/\theta}, \; y > 0$ $\quad f_n(\underset{\sim}{y}|\theta) = \frac{1}{\theta^n} e^{-\sum y_i/\theta}$

$$L(\theta) = \frac{1}{\theta^n} e^{-\sum y_i/\theta}$$

b. Identify a sufficient statistic for $\theta$.

$$L(\theta) = \underbrace{\frac{1}{\theta^n} e^{-T/\theta}}_{g(T,\theta)} \cdot \underbrace{1}_{h(\underset{\sim}{y})}$$

$T = \sum y_i$ is suff for $\theta$ by FC.

c. Find the MLE for $\theta$ (formula and value using data).

$$\ell(\theta) = -n \log \theta - T/\theta$$

$$\ell'(\theta) = \frac{-n}{\theta} + \frac{T}{\theta^2} = 0$$

$\hat{\theta} = \bar{Y} = .606299$

$$\frac{T}{\theta^2} = \frac{n}{\theta} \qquad T = n\theta$$

$$\hat{\theta} = \frac{T}{n} = \frac{\sum y_i}{n} = \bar{Y}$$

d. Is the MLE minimal sufficient for $\theta$? Why or why not?

MLE is a 1-1 fn of $T = \sum Y_i$ which is suff, so $\hat{\theta}$ MLE is suff. then b/c it is an MLE, it is minimal sufficient.

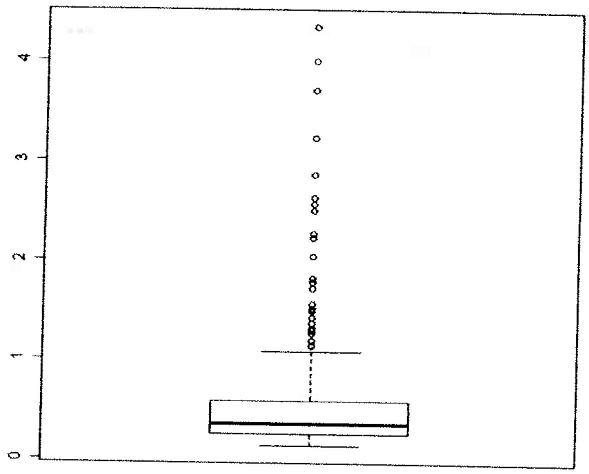
## Whale Velocity Summary Information and Basic Graphs with R Commands

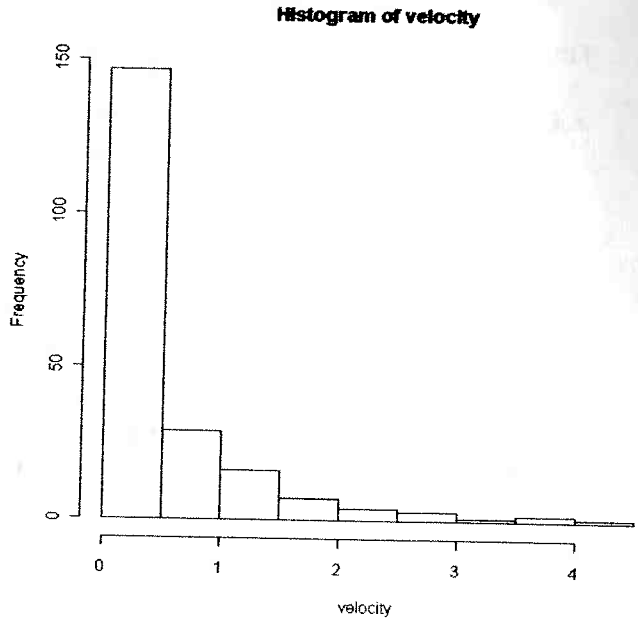
```
Velocity = read.table("C:/Documents and Settings/awagaman/My Documents/Math 30/Spring 2011/Handouts/whalevelocity.txt", header=TRUE)  %Reads in the data%

attach(Velocity)  %Lets you work with variable names directly from the data%

hist(velocity) %Makes a histogram%

boxplot(velocity) %Makes a boxplot%
```




```
summary(velocity) %computes basic descriptive statistics%
  Min. 1st Qu.  Median    Mean  3rd Qu.   Max.
0.1337  0.2525  0.3540  0.6063  0.5908  4.3480

mean(velocity)  %computes the sample mean with more precision%
0.606299

sd(velocity)    %computes the sample standard deviation%
0.6793837

length(velocity)   %computes number of observations in this variable%
210


sum(velocity)
127.3228

sum(log(velocity))
-177.4602
```

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 (n observations) from a Gamma distribution where both $\alpha$ and $\beta$ are unknown?

$$f(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta}$$

$$L(\alpha,\beta) = f_n(y) = \frac{1}{(\Gamma(\alpha)\beta^{\alpha})^n} (\prod y_i)^{\alpha-1} e^{-\Sigma y_i/\beta}$$

$$l(\alpha,\beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha-1)\log(\Sigma y_i) - \Sigma y_i \log \beta$$

f. Does it look like the MLEs for $\alpha$ and $\beta$ are easily computable?

No, not easily computable $\rightarrow$ $\Gamma$ fn and dependence on $\alpha, \beta$ for each other.

The good news is that the computer can calculate the MLEs for us (or method of moments estimators if you want also) based off the data. In R, you need to first load a library called MASS and then use the fit distribution function (fitdistr) appropriately. This is what happens when you try to fit an exponential and a gamma distribution for the whale velocities:

```
library(MASS)
fitdistr(velocity,"exponential")
   rate
 1.6493511
(0.1138160)

fitdistr(velocity,"gamma")
   shape       rate
 1.5969630   2.6339528
(0.1425361) (0.2756002))
```

You get parameter estimates and standard errors (these are in the parentheses).

Important Notes! Notes on the exponential distribution in R specify that rate = $1/\theta$, because of their provided density function. Notes on the gamma distribution in R specify that shape = $\alpha$ and rate = $1/\beta$, because of their provided density function.

g. Does the MLE provided by the computer match your MLE from the exponential distribution?

$$\frac{1}{\hat{\theta}} = 1.6493511 \Rightarrow \hat{\theta} = .6063 = \bar{Y} \checkmark$$

yes it matches

h. Do the Gamma distribution estimates appear "consistent" with the exponential estimates?

$$\text{Exp}(\theta) = \text{Gamma}(1, \theta) \overset{?}{\approx} \text{Gamma}(\alpha, \beta)$$

$$\alpha = 1.5969630 \quad \beta = \frac{1}{2.6339528} = .379657524615$$

$$\alpha\beta = \bar{Y} = .606299$$ so "somewhat" consistent

$$\alpha \neq 1$$

i. Let's verify that the log-likelihood does appear to be maximized for the MLE estimates from the Gamma.

```
n=210
a=1.5969630
b=1/2.6339528;b        %to show the beta value%
 0.3796575
loglik=-n*a*log(b)-n*log(gamma(a))+(a-1)*sum(log(velocity))-(1/b)*sum(velocity)
loglik
 -92.78208
```

Here's a table with the loglik values for some choices of $\alpha$ and $\beta$.

| a\b | .35 | .3796575 | .38 | .40 |
|---|---|---|---|---|
| 1.55 | -94.92616 | -92.98399 | -92.97523 | -92.9182 |
| 1.5969630 | -93.9221 | $\boxed{-92.78208 \text{ Max}}$ | -92.78222 | -93.23105 |
| 1.6 | -93.87106 | -92.78292 | -92.78363 | -93.26518 |
| 1.63 | -93.45555 | -92.87983 | -92.88622 | -93.69091 |

Does it appear that the MLEs are maximizing the log-likelihood?

yes. all around the log-lik values are more negative.

j. We want to pick between the exponential and gamma distributions now. Let's sample from exponential and gamma distributions based on the MLE estimates. We can then make histograms and boxplots for these simulated samples (next page). Which histogram more closely resembles the whale velocity data? What about boxplots? Which distribution would you prefer to use to model the whale velocities? Are there any issues you see that might make you consider other distributions to use as models?

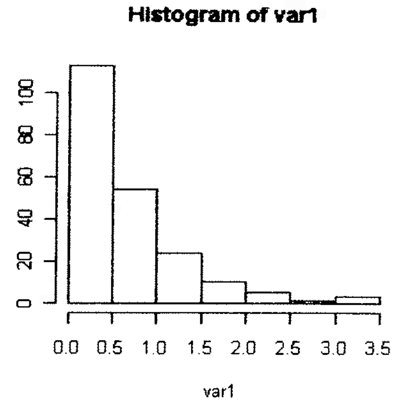
Overall, the Gamma plots look better.
However, still not quite getting spike low or that many outliers.

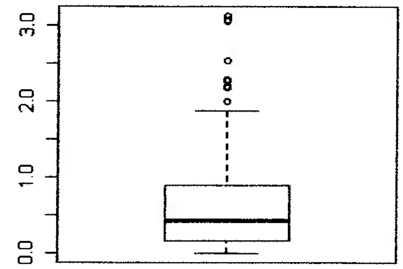
Here are the R commands used to generate these simulated samples. Note that I did simulated samples from each distribution twice, and I choose to simulate with n=210 to be consistent with the data.

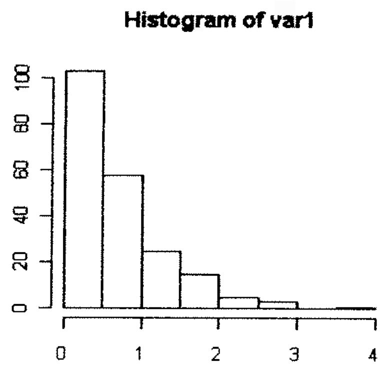
```
var1=rexp(210, 1.6493511)
var2=rgamma(210, 1.5969630,2.6339528)
```

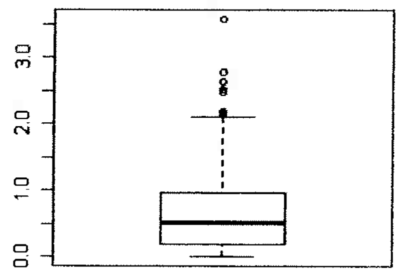
Graphs were generated via the hist and boxplot commands applied to var1 and var2.

From Exponential:

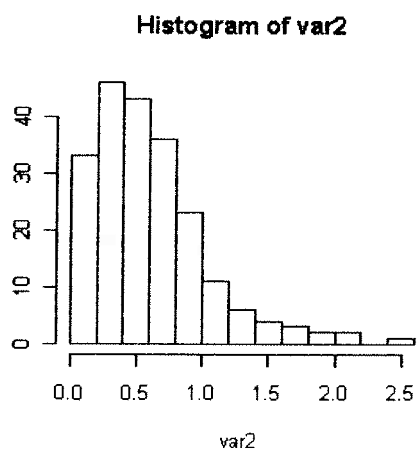


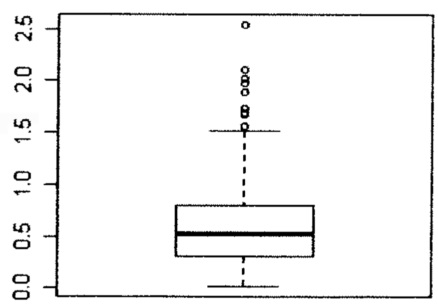



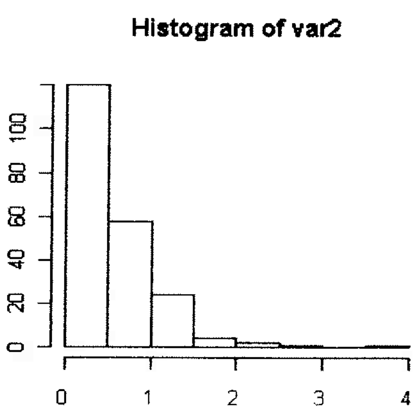


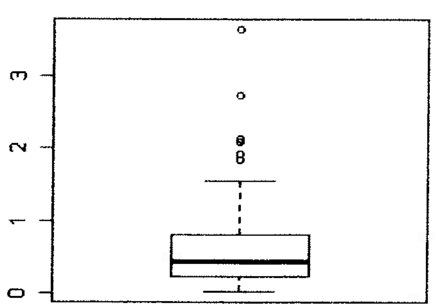



From Gamma:





Briefly, a bit more theory practice. Assume you are considering a random sample of n observations from an exponential distribution with unknown $\beta$, as we had at the start of this activity, in the context of the whale velocities. (Note, I swapped from $\theta$ to $\beta$)

k. Find unbiased estimators for $\beta$ - average velocity of the whales, $\beta^2$ - variance of the whale velocities, and $\beta(1-\beta)$.

$E(\bar{X}) = \beta$ b/c of properties of exponentials

$E(s^2) = \beta^2$ b/c $s^2$ is unbiased for $\sigma^2 = \beta^2$ here.

$\beta(1-\beta) = \beta - \beta^2 \Rightarrow E(\underbrace{\bar{X} - s^2}_{\text{unbiased for } \beta - \beta^2}) = E(\bar{X}) - E(s^2) = \beta - \beta^2 = \beta(1-\beta)$

l. In general, could you conclude that $1/\bar{X}$ is unbiased for $1/\mu$? To think about this more simply, think about just a single observation, is 1/X unbiased for $1/\mu$ (you can/should write out the expectation to look at it) for a general distribution with pdf f(x) and mean $\mu$?

No. $E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$ in general.

$$E\left(\frac{1}{X}\right) = \int_x \frac{1}{x} f(x)\,dx \neq \frac{1}{\int x f(x)\,dx} = \frac{1}{E(X)} = \frac{1}{\mu}$$

m. Show that $\frac{2}{\beta}\sum_{i=1}^{n} X_i$ has a chi-squared distribution with 2n degrees of freedom (hint: think about results we know for sums of Gamma RVs.), and that therefore it is a pivot for $\beta$.

$X \sim Exp(\beta)$ $\quad Y = \frac{2}{\beta} X$ $\quad M_Y(t) = \left(1 - \beta\left(\frac{2}{\beta}\right)t\right)^{-1}$

$= (1-2t)^{-1}$

$\sim Exp(2)$

$\Rightarrow \frac{2}{\beta}\sum X_i \sim Gamma(n, 2)$

$\Rightarrow \chi^2(2n)$

No dependence on $\beta$.

**n. Based on the CLT, we know that $\bar{X}$ is approx. normal, and we can standardize to make a standard normal random variable, $Z = \frac{\bar{X}-\beta}{\sqrt{(\beta^2/n)}}$. What distribution would Y (see below) have? Reduce the expression filling in what Z would be. Does it look like this quantity might be useful for making CIs for $\beta$? (In other words, is Y also a pivot?)**

$$Y = \frac{Z}{\left(\frac{2\sum X_i}{\beta(2n)}\right)^{(1/2)}} \Rightarrow \frac{N(0,1)}{\left(\frac{\chi^2(2n)}{2n}\right)^{1/2}} \sim t(2n)$$

$$Y = \frac{\frac{\bar{X}-\beta}{\beta/\sqrt{n}}}{\frac{(\sum X_i)^{1/2}}{\sqrt{\beta}\sqrt{n}}} = \frac{\sqrt{n}\sqrt{\beta}}{(\sum X_i)^{1/2}} \cdot \frac{\bar{X}-\beta}{\beta/\sqrt{n}} = \frac{n(\bar{X}-\beta)}{\beta^{1/2}(\sum X_i)^{1/2}}$$

can also sub $\frac{\sum X_i}{n}$ as $\bar{X}$ and rearrange

No, not very useful

**o. Noting that the variable in n. might be hard (or at least not very appealing) to use to make CIs, try using just Z, which is approximately standard normal. Give a formula for a 90 percent CI for $\beta$. Note the .95 quantile from the normal distribution is 1.645.**

$$Z = \frac{\bar{X}-\beta}{\beta/\sqrt{n}} = \frac{\sqrt{n}\,\bar{X} - \sqrt{n}\,\beta}{\beta} = \sqrt{n}\left(\frac{\bar{X}}{\beta}-1\right) \sim N(0,1)$$

$$P\left(-1.645 \le \sqrt{n}\left(\frac{\bar{X}}{\beta}-1\right) \le 1.645\right) = .90$$

$$P\left(-\frac{1.645}{\sqrt{n}} \le \frac{\bar{X}}{\beta} - 1 \le \frac{1.645}{\sqrt{n}}\right) = .90$$

$$P\left(\frac{-1.645+\sqrt{n}}{\sqrt{n}} \le \frac{\bar{X}}{\beta} \le \frac{1.645+\sqrt{n}}{\sqrt{n}}\right) = .90$$

$$P\left(\frac{\sqrt{n}}{\sqrt{n}-1.645} \ge \frac{\beta}{\bar{X}} \ge \frac{\sqrt{n}}{1.645+\sqrt{n}}\right) = .90$$

$$\left(\frac{\bar{X}\sqrt{n}}{\sqrt{n}-1.645} \ge \beta \ge \frac{(\sqrt{n})\,\bar{X}}{1.645+\sqrt{n}}\right) \Rightarrow (.5445, .68345)$$

$\bar{X} = .6063$

$n = 210$ $\quad \sqrt{n} = 14.49$

Solutions

# Chapter 8 and 9 Review: A Bit More Practice

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

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

and 0, otherwise.

a. Find the likelihood function for the $n$ observations.
b. Identify a sufficient statistic for $\theta$.
c. Find the MLE for $\theta$.
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?
g. How would you check to see if the MLE is consistent for $\theta$?
h. Would it be appropriate to compute the relative efficiency of the estimators in c. and f. with the information you have about those estimators right now?

a. $L(\theta) = f_n(\underline{x}|\theta) = (\theta+1)^n (\prod x_i)^{\theta}$, $0 \le x_i \le 1 \quad \forall i=1,\dots,n$
$0$, o.w.

b. By FC, $(\prod x_i)$ is sufficient. $T = \prod x_i$

c. $\ell(\theta) = n \log(\theta+1) + \theta \log(\prod x_i) = n\log(\theta+1) + \theta \sum (\log X_i)$

$\ell'(\theta) = \frac{n}{\theta+1} + \sum(\log X_i) = 0 \Rightarrow \frac{n}{\theta+1} = -\sum \log X_i$

$\frac{-n}{\sum \log X_i} = \theta + 1 \Rightarrow \hat{\theta} = \frac{-n}{\sum \log X_i} - 1$ or $\frac{-n}{\log(\prod x_i)} - 1$

d. MLE is a 1:1 fn of T which is suff. so it is also suff. and hence is min. suff.

e. $f(x|\theta) = a(\theta)b(x)\exp[c(\theta)d(x)]$
$\Rightarrow a(\theta) = \theta+1 \quad b(x) = 1 \quad c(\theta) = \theta \quad d(x) = \log x$

f. $T = \sum d(x) = \sum \log X_i$ is suff.



g. Check unbiased and if so see if $\lim_{n\to\infty}$ Var is 0. Getting pdf would NOT be fun.

h. We'd need to know they are unbiased before computing eff.