Stat 511 HW#7 Spring 2009 (not to be collected)

1. An R function that will create bootstrap samples (taken from page 399 of Efron and Tibshirani) is

```
> bootstrap<-function(x,nboot,theta)
+ {data<-matrix(sample(x,size=length(x)*nboot,replace=T),nrow=nboot)
+ return(apply(data,1,theta))
+ }
```

(A fancier version that will compute BCa and ABC intervals can be had by downloading the R contributed package "bootstrap" that has been written to accompany the Efron and Tibshirani book.) Load the MASS package.

Below are times to failure (in millions of cycles) for high speed turbine engine bearings made from two different compounds (taken from "Analysis of Single classification Experiments Based on Censored Samples from the Two-Parameter Weibull Distribution: by J.I. McCool, *Journal of Statistical Planning and Inference*, 1979). (These are pretty small samples to be trying out a bootstrap technology, but for sake of exercise we will ignore this fact.)

Compound 1	3.03	5.53	5.60	9.30	9.92	12.51	12.95	15.21	16.04	16.84
Compound 2	3.19	4.26	4.47	4.53	4.67	4.69	5.78	6.79	9.37	12.75

Enter these two sets of values into R as

```
> compound1<-c(3.03,5.53,5.60,9.30,9.92,12.51,12.95,15.21,16.04,16.84)
> compound2<-c(3.19,4.26,4.47,4.53,4.67,4.69,5.78,6.79,9.37,12.75)
```

a) Make normal plots for the two samples as

```
> qqnorm(compound1)
> qqnorm(compound2)
```

Do you expect "constant variance/normal distribution" ordinary statistical methods to be reliable in the analysis of these data? Why?

b) Make normal plots of the log-lifetimes here as

```
> qqnorm(log(compound1))
> qqnorm(log(compound2))
```

Are you any more optimistic about the reliability of "constant variance/normal distribution" ordinary statistical methods in the analysis of the log lifetimes?

c) Consider inference for the median of a distribution F that one might assume to be generating the transplant Compound 2 lifetimes. The sample median Compound 2 lifetime can be had by typing

```
> median(compound2)
```

As a way of assessing the precision of the sample median (as an estimator of the median of F, you may generate, say, $B = 10,000$ bootstrap sample medians by typing

```
> B<-10000
> comp2boot.non<-bootstrap(compound2,B,"median")
```

Based on these values, what number might you report as a standard error for the sample median?

b) Use the bootstrapped sample medians to produce a 95% (unadjusted) percentile bootstrap confidence interval for the median of F. Some R code for doing this is

```
> kl<-floor((B+1)*.025)
> ku<-B+1-kl
> sortcomp2boot.non<-sort(comp2boot.non)
> sortcomp2boot.non[kl]
> sortcomp2boot.non[ku]
```

c) A parametric bootstrap approach here might be to use a Weibull distribution to describe transplant survival times. It's easy to use maximum likelihood to fit a Weibull to these data by typing

```
> fit2<-fitdistr(compound2,"weibull")
> fit2
```

Then a set of 10,000 simulated sample medians from the fitted distribution can be simulated by typing

```
> Wboot<-function(samp,nboot,theta,shape,scale)
+ {data<-scale*matrix(rweibull(samp*nboot,shape),nrow=nboot)
+ return(apply(data,1,theta))
+ }
> comp2boot.Wei<-Wboot(10,B,"median",fit2$estimate[1],fit2$estimate[2])
```

Use these to produce a parametric bootstrap standard error for the sample median and a parametric bootstrap 95% (unadjusted) percentile bootstrap confidence interval for the median of F. How do these compare to what you found above?

d) As Koheler's Assignment #9 from Spring 2002 points out, large sample theory says that if f is the density function for F and $\theta = F^{-1}(.5)$ (the population median) the standard deviation of the sample median is approximately

$$\frac{1}{2f(\theta)\sqrt{n}}$$

How does a (Weibull) parametric bootstrap standard error for the sample median Compound 2 lifetime compare to a plug-in version of the above (where the sample median is used in place of θ and the fitted Weibull density is used for f)?

e) If one models the Compound 1 survival times as iid from second distribution, G, it might be of interest to estimate the difference in medians $F^{-1}(.5) - G^{-1}(.5)$. Modify the above to create $B = 10,000$ (nonparametric) bootstrap sample medians from the Compound 1 cases. Then compute 10,000 differences between a bootstrap Compound 2 sample median and a bootstrap Compound 1 sample median. Based on these 10,000 differences, find 95% percentile confidence limits for the difference in underlying median lifetimes. Is this difference clearly non-zero?

2. Problem 3 of the Spring 2003 Stat 511 Final Exam is a bootstrap problem. Do the R computing required to get your own copy of the printout provided on that Exam. Then answer the Exam questions.

3. Problem 3 of the Spring 2004 Stat 511 Final Exam is a bootstrap problem. Do the R computing required to get your own copy of the printout provided on that Exam. Then answer the Exam questions.

4. The table below contains information on 23 (out of 24) pre-Challenger space shuttle flights. (On one flight, the solid rocket motors were lost at sea and so no data are available.) Provided are launch temperatures, t (in °F), and a 0-1 response, y, indicating whether there was post-launch evidence of a field joint primary O-ring incident. (O-ring failure was apparently responsible for the tragedy.) $y = 1$ indicates that at least one of the 6 primary O-rings showed evidence of erosion.

Temperature	O-ring Incident?
66	0
70	1
69	0
68	0
67	0
72	0
73	0
70	0
57	1
63	1
70	1
78	0

Temperature	O-ring Incident?
67	0
53	1
67	0
75	0
70	0
81	0
76	0
79	0
75	1
76	0
58	1

Treat the response variables, y_i, as Bernoulli distributed (binomial $(1, p_i)$) and independent launch to launch. Note that $\mu = \mathrm{E}y = p$ here. We'll model

$$h(\mu) = \log\left(\frac{\frac{\mu}{n}}{1-\frac{\mu}{n}}\right) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 t$$

This is a generalized linear model, with binomial (Bernoulli) response, and the "logit" link … otherwise known as a logistic regression model. We may use R's `glm` function to analyze these data. (You may learn about the function in the usual way, by typing `?glm` .) Load the `MASS` package.

Read in the data for this problem by typing

```
> temp<-c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,
76,79,75,76,58)
> incidents<-c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1)
```

And create and view a logical variable equivalent to the 0-1 response variable by typing

```
> indicate<-(incidents>0)
> indicate
```

a) We may then fit and summarize a generalized linear model here by typing

```
> shuttle.out<-glm(indicate~temp,family=binomial)
> summary(shuttle.out)
```

The logit is the default link for binomial responses, so we don't need to specify that in the function call.

Notice that the case $\beta_1 < 0$ is the case where low temperature launches are more dangerous than warm day launches. NASA managers ordered the launch after arguing that these and other data data showed no relationship between temperature and O-ring failure. Was their claim correct? Explain.

b) `glm` will provide estimated mean responses (and corresponding standard errors) for values of the explanatory variable(s) in the original data set. To see estimated means $\widehat{\mu_i} = \widehat{p_i} = \dfrac{1}{1+\exp\left(-\beta_0 - \beta_1 t_i\right)}$ and corresponding standard errors, type

```
> shuttle.fits<-predict.glm(shuttle.out,type="response",
se.fit=TRUE)
> shuttle.fits$fit
> shuttle.fits$se.fit
```

Plot estimated means versus t. Connect those with line segments to get a rough plot of the estimated relationship between t and p. Plot "2 standard error" bands around that response function as a rough indication of the precision with which the relationship between t and p could be known from the pre-Challenger data.

The temperature at Cape Canaveral for the last Challenger launch was 31° F. Of course, hind-sight is always perfect, but what does your analysis here say might have been expected in terms of O-ring performance at that temperature? You can get an estimated 31° F mean and corresponding standard error by typing

```
> predict.glm(shuttle.out,data.frame(temp=31),se.fit=TRUE,
type="response")
```

5. An engineering student group worked with the ISU press on a project aimed at reducing jams on a large collating machine. They ran the machine at 3 "Air Pressure" settings and 2 "Bar Tightness" conditions and observed

y = the number of machine jams experienced in

k seconds of machine run time

(Run time does not include the machine "down" time required to fix the jams.) Their results are below.

Air Pressure	Bar Tightness	y, Jams	k, Run Time
1 (low)	1 (tight)	27	295
2 (medium)	1	21	416
3 (high)	1	33	308
1	2 (loose)	15	474
2	2	6	540
3	2	11	498

Motivated perhaps by a model that says times between jams under a given machine set-up are independent and exponentially distributed, we will consider an analysis of these data based on a model that says the jam counts are independent Poisson variables. For

μ = the mean count at air pressure i and
bar tightness j

suppose that

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + \log k_{ij} \qquad (*)$$

Notice that this says

$$\mu_{ij} = k_{ij} \exp\left(\mu + \alpha_i + \beta_j\right)$$

(If waiting times between jams are independent exponential random variables, the mean number of jams in a period should be a multiple of the length of the period, hence the multiplication here by k_{ij} is completely sensible.) Notice that equation (*) is a special case ($\gamma = 1$) of the relationship

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma \log k_{ij}$$

which is in the form of a generalized linear model with link function $h(\mu) = \log\left(\mu\right)$. As it turns out, `glm` will fit a relationship like (*) for a Poisson mean that includes an "offset" term ($\log k_{ij}$ here). Enter the data for this problem and set things up by typing

```
> A<-c(1,2,3,1,2,3)
> B<-c(1,1,1,2,2,2)
> y<-c(27,21,33,15,6,11)
> k<-c(295,416,308,474,540,498)
> AA<-as.factor(A)
> BB<-as.factor(B)
> options(contrasts=c("contr.sum","contr.sum"))
```

a) Fit and view some summaries for the Poisson generalized linear model (with log link and offset) by typing

```
> collator.out<-glm(y~AA+BB,family=poisson,offset=log(k))
> summary(collator.out)
```

The log link is the default for Poisson observations, so one doesn't have to specify it in the function call. Does it appear that there are statistically detectable Air Pressure and Bar Tightness effects in these data? Explain. If one wants small numbers of jams, which levels of Air Pressure and Bar Tightness does one want?

b) Notice that estimated "per second jam rates" are given by

$$\exp\left(\mu + \alpha_i + \beta_j\right)$$

Give estimates of all 6 of these rates based on the fitted model.

c) One can get `R` to find estimated means corresponding to the 6 combinations of Air Pressure and Bar Tightness for the corresponding values of k. This can either be done on the scale of the observations or on the log scale. To see these first of these, type

```
> collator.fits<-predict.glm(collator.out,type="response",
se.fit=TRUE)
> collator.fits$fit
> collator.fits$se
```

How are the "fitted values" related to your values from b)?

To see estimated/fitted log means and standard errors for those, type

```
> lcollator.fits<-predict.glm(collator.out,se.fit=TRUE)
> lcollator.fits$fit
> lcollator.fits$se
```

5. Do Problem 2 of the Spring 2003 Stat 511 Final Exam.

6. Do Problem 2 of the Spring 2004 Stat 511 Final Exam.

7. Do Problem 5 of Chapter 11 of Faraway's *Extending the Linear Model with R*. You will need to install and load the `faraway` package in order to get to the data. Try out all of the smoothers discussed in Faraway's Chapter 11, and for each that has a choice of parameter(s) try more than one (set of) value(s).