Homework 3

Biology 697

9/21/2012

1 Problems from Whitlock & Schluter

Complete problems 10-12 on pg. 95. Use R where possible. Data sets (so you don't have to type things in) are available at `http://www.zoology.ubc.ca/~whitlock/ABD/teaching/datasets.html`.

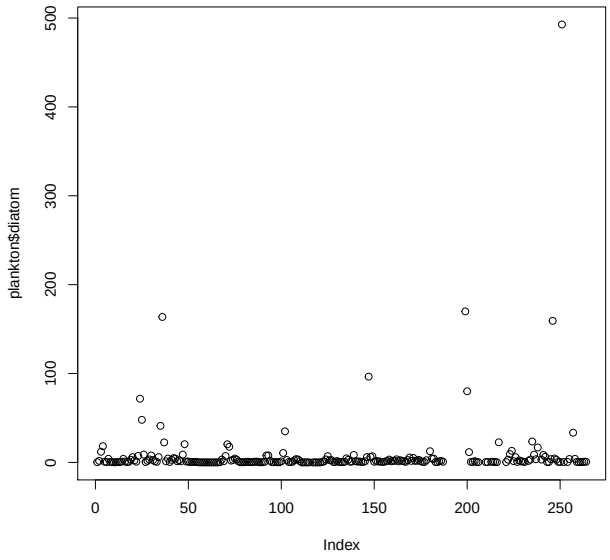
2 R and Plotting

N.B. There is more than one way to code correct answers for any of these. Have fun, and do them as efficiently as possible to the best of your abilities. And feel free to spice them up and/or make improvements as you see fit.

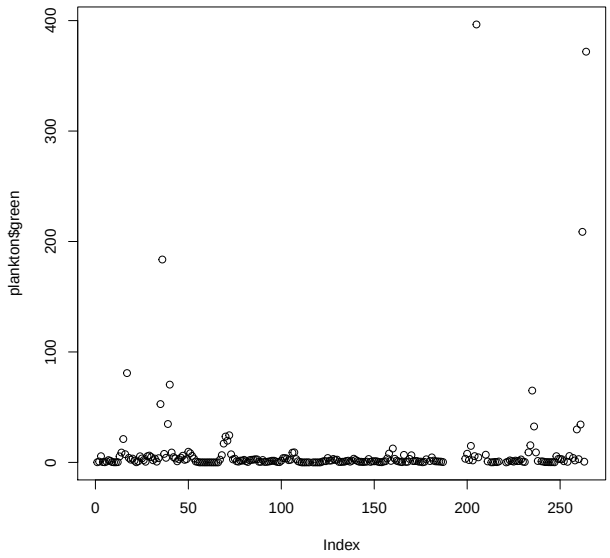
2.1 Loading and Cleaning

We'll be working with the Lake Baikal Plankton Data. To learn more about it, the instrumentation, etc., see `http://knb.ecoinformatics.org/knb/metacat/nceas.290/nceas`. Load it in. Screen it for any obvious outliers. Should they be eliminated? Why or why not?

```
plankton<-read.csv("../lectures/data/hampton.5.1-Baikal_74_97_moAvg_plankton.csv",
                   skip=1, na.strings=c("NA",  " NA", ".", " "))

#after visual inspection, filter outliers
#but check, before using data, if these are
#real...
plot(plankton$diatom)
```




```
plot(plankton$green)
```



```
plankton<-plankton[which(plankton$diatom<200),]
plankton<-plankton[which(plankton$green<100),]


##########################2.2a
```

2.2 Error of Estimates other than the mean

As we discussed in class, the re-sampling based approach to assessing error in parameter estimates can be incredibly simple and powerful. In particular, it can be quite powerful in the case of variables that have asymmetric confidence intervals. To estimate asymmetric confidence intervals, one re-samples their data as usual to calculate a test statistic, but then looks at the quantiles or percentiles of the test statistic to determine the range of values in which 95% of their sample estimates fall.
Let's look at how this works for medians.

(a) Calculate the naive bootstrapped standard error and 95% confidence intervals for the median of the values of `diatom` in the data. Use 5000 bootstrapped replicates.

(b) Compare this naive estimate to the percentile confidence intervals. Take a look at the arguments for the function `quantile`. Are they different? Why or why not?

(c) Look at the `bootstrap` function in the bootstrap package. Can you use it to get the 95% CIs in two lines of code?

```
n.sims<-5000
diatomMedian<-rep(NA, n.sims)
for(i in 1:n.sims){
  diatomMedian[i]<-median(sample(plankton$diatom,
                                 size=nrow(plankton), replace=TRUE), na.rm=T)
}

diatomCIn<-2*sd(diatomMedian)
c(median(plankton$diatom) - diatomCIn,
  median(plankton$diatom)+diatomCIn)

## [1] 0.7673 1.3782


##########################2.2b
quantile(diatomMedian, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 0.8164 1.3910
```

```
############################2.2c
```

```
diatomMedian2<-bootstrap(plankton$diatom, 5000, median)
quantile(diatomMedian2$thetastar, c(0.025, 0.975))
```

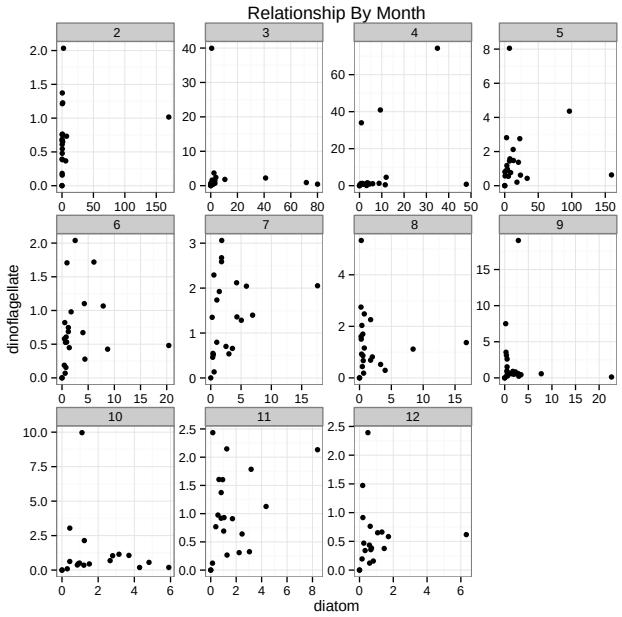
```
##   2.5%  97.5%
## 0.8164 1.3910
```

```
############################2.3
```

2.3 Faceting and Loops

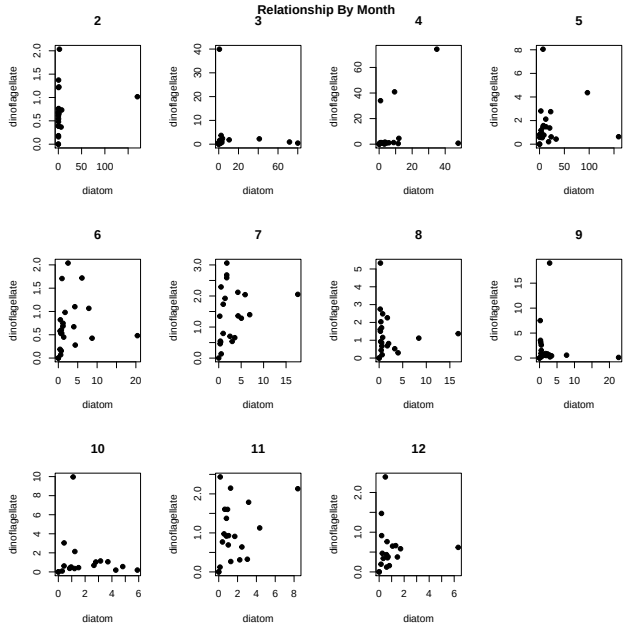
One of the really interesting ways to look at the relationships in this data is to split them by month. This lets us see trends within months so that we can directly compare processes between years. For example, we can look at the Diatom-Dinoflagellate relationship as follows.

```
  qplot(diatom, dinoflagellate, data=plankton) +
    facet_wrap(~Month, scale="free") +
  theme_bw() +
  ggtitle("Relationship By Month")
```



Reproduce this plot using both ggplot2 and the basic R graphing package. The former is straightforward. The latter should look something like:

```
#set the # of panels
par(mfrow=c(3,4))

#loop over months, plotting on panel per month
for(a.month in unique(plankton$Month)){
  plot(dinoflagellate ~ diatom,
       data=plankton[which(plankton$Month == a.month),],
       pch=19, main=a.month)
}
title("Relationship By Month", outer=TRUE, line=-1.5)
```



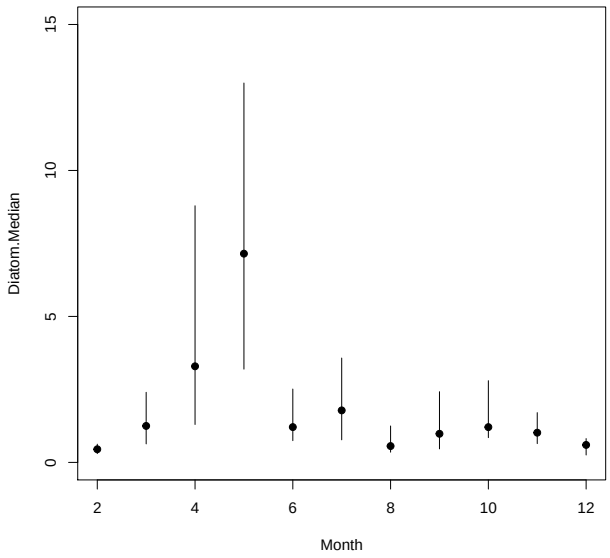

```
#now be nice and reset plotting conditions
par(mfrow=c(1,1))


############################2.4
```

2.4 Representing Variation

Often in plots we want to show an estimate and the variation around that estimate. Boxplots do this for a whole sample, but what if we want to see means and the variation around the means? Reproduce the following two plots. These are plots of the median diatom abundance in different months and the bootstrapped 95% confidence interval around the medians. Produce the plot using both the base R graphics package and in ggplot2. You'll need to look at some additional plotting functions to get those error lines in the base graphics package. Likewise, you'll need to play with some additional geoms for ggplot2. Feel free to spice up your graphs beyond what I have presented here.

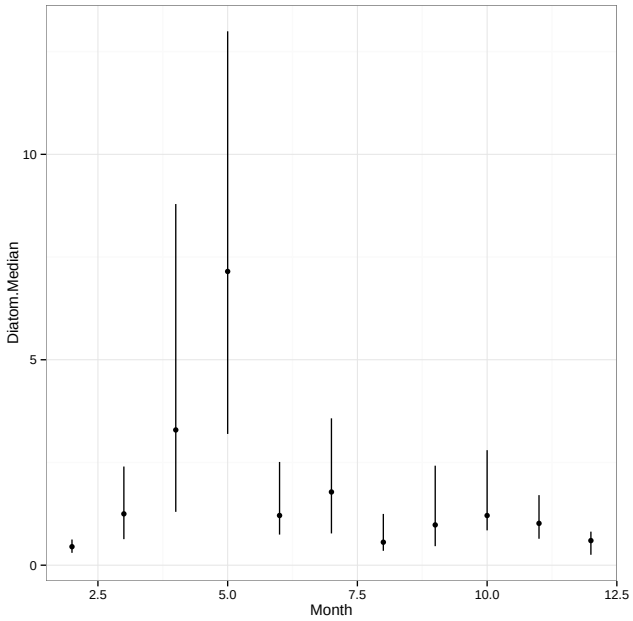
Base graphics package:

```
#1) Create a new data frame that will have the information for plotting
#   as we need one row per month
newPlankton<-data.frame(Month=unique(plankton$Month))

#2) For loop to calculate the aggregated properties
for (i in 1:nrow(newPlankton)) {
  #3) Get the monthly data set
  shortDF<-plankton[which(plankton$Month==newPlankton$Month[i]),]

  #4) bootstrapped CIs
  shortDiatomMedian<-bootstrap(shortDF$diatom, 5000, median)

  newPlankton$Diatom.Median[i]<-median(shortDF$diatom)

  #5) Extract the monthly CIs
  newPlankton$Diatom.lowerCI[i]<-quantile(shortDiatomMedian$thetastar, 0.025)
  newPlankton$Diatom.upperCI[i]<-quantile(shortDiatomMedian$thetastar, 0.975)

}

#6) plot for points, segments for error bars
plot(Diatom.Median ~ Month, data=newPlankton, pch=19, ylim=c(0,15))
segments(newPlankton$Month, newPlankton$Diatom.lowerCI,
         newPlankton$Month, newPlankton$Diatom.upperCI)
```



Ggplot2:

```
#7) the ggplot2 way uses geom_point and geom_linerange
#although geom_pointrange would also work
  ggplot(data=newPlankton, aes(x=Month, y=Diatom.Median,
                               ymin=Diatom.lowerCI, ymax=Diatom.upperCI)) +
  geom_point() +
  geom_linerange() +
  theme_bw()
```


Diatom.Median
10
5
0
2.5
5.0
7.5
10.0
12.5
Month