STA841 HW1

Dipesh Gautam

September 17, 2015

1 Problem 1

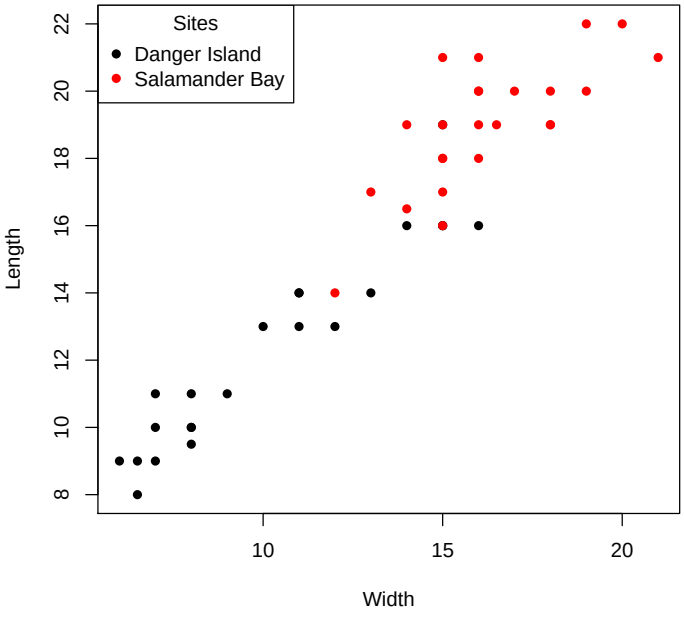

Figure 1:

I've plotted the given information from "jellyfish.dat" in Fig. 1 where color indicate the location where the jellyfish was caught. As we can see from the plot, length and width for jellysfish seem to be linearly related. And also that the jellyfish found in Danger Island tend to be smaller on both length and width than the ones found in Salamander Bay.

2 Problem 2

2.1

The response variable is the number of hours of sleep each person gets. Treatment is the explanatory variable. Since we're trying to model the effect of different treatments on insomnia patients and administering the different treatments on the same patients on different nights, patient Id will be the block factor in the study.

2.2

One of the concerns I have about the designs not addressed in the data is how far apart were the nights when each patient received the treatment. And if they were in back to back nights, what (if any) effects the previous night's treatment had on the next night?

2.3

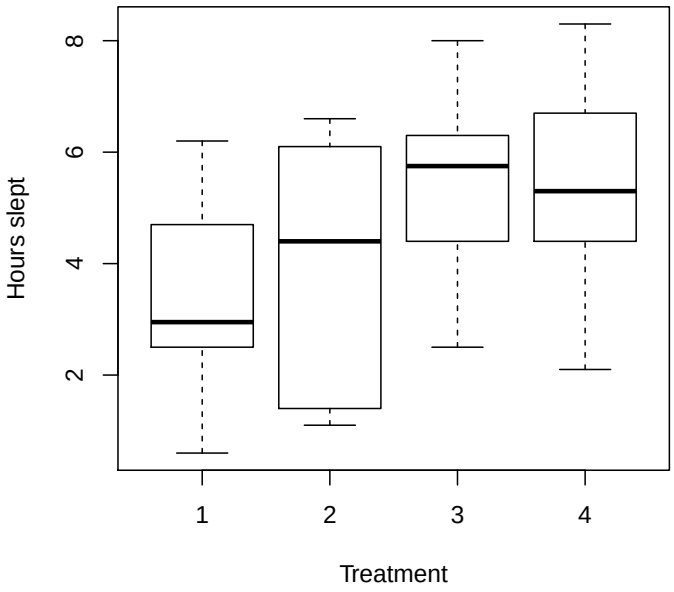

Figure 2: Plot summarizing the data

In the Fig. 2, we have a box plot to explore the difference in treatment effects over all patients. and we can see that overall treatment C and D seem to work better than the first two. Treatment B has seems to have really high variance.

Fig. 3 shows number of hours of sleep each person gets. We can see that each person is different and gets more a less sleep regardless of treatment. E.g. patient 10 gets a lot of sleep regardless of which treatment he gets whereas patient 1 gets less sleep in general regardless of treatment. We can also see the effect of treatment described by last plot where treatment C and D seemed to work better in general. For most people, it's either treatment C or treatment D that provides the most hours of sleep.

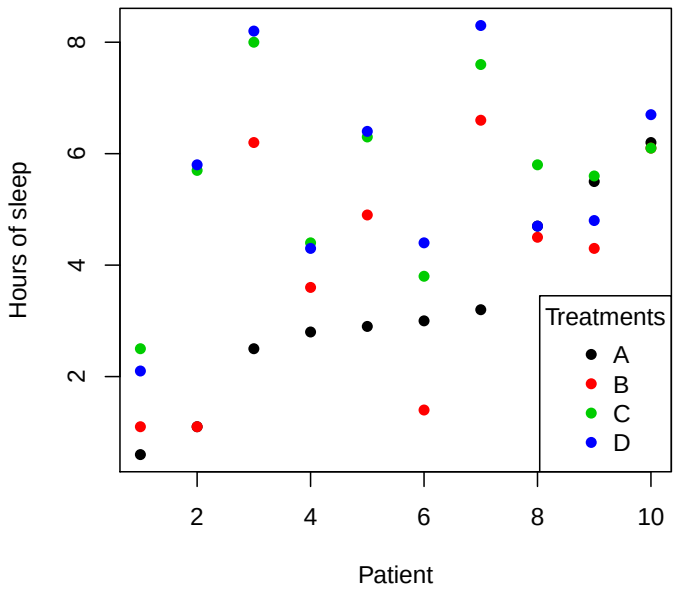

Figure 3: Scatterplot represenation of given data

2.4

For this problem we're going to try to fit a linear model with as:
Hours $\sim$ Treatment + Patient

We need independence between treatments and also between treatments and individuals.

2.5

```
Call:
lm(formula = Hours ~ Treatment + factor(Patient), data = sleep)

Residuals:
   Min     1Q Median     3Q    Max
-2.380 -0.475  0.190  0.640  1.795

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.2300     0.6606   0.348 0.730431
TreatmentB          0.7300     0.5183   1.409 0.170371
TreatmentC          2.3300     0.5183   4.496 0.000118 ***
TreatmentD          2.3200     0.5183   4.477 0.000124 ***
factor(Patient)2    1.8500     0.8194   2.258 0.032259 *
factor(Patient)3    4.6500     0.8194   5.675 5.01e-06 ***
factor(Patient)4    2.2000     0.8194   2.685 0.012252 *
factor(Patient)5    3.5500     0.8194   4.332 0.000183 ***
factor(Patient)6    1.5750     0.8194   1.922 0.065208 .
factor(Patient)7    4.8500     0.8194   5.919 2.62e-06 ***
factor(Patient)8    3.3500     0.8194   4.088 0.000350 ***
factor(Patient)9    3.4750     0.8194   4.241 0.000233 ***
factor(Patient)10   4.7000     0.8194   5.736 4.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.159 on 27 degrees of freedom
Multiple R-squared:  0.7842,        Adjusted R-squared:  0.6883
F-statistic: 8.177 on 12 and 27 DF,  p-value: 3.339e-06
```

We can also run two way anova model to check the statistical significance of the two factor variables.

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Treatment        3  41.08  13.694  10.197 0.000116 ***
factor(Patient)  9  90.70  10.078   7.504 2.02e-05 ***
Residuals       27  36.26   1.343
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results reflect what we expected from part 3. We see statistical significance in almost all of them and alsotreatment C and treatment D seem to be working better. The patients underlying response to the treatments is also reflected as predicted by the plots.

2.6

If we try to estimate the treatment x patient interaction based on this data set, we'll end up with more covariates than the number of observations. So, the estimation is not possible using the simple model described above.

3 Problem 3

The plot provided in lecture 3 is reproduced in Fig. 4. The code is included in appendix.

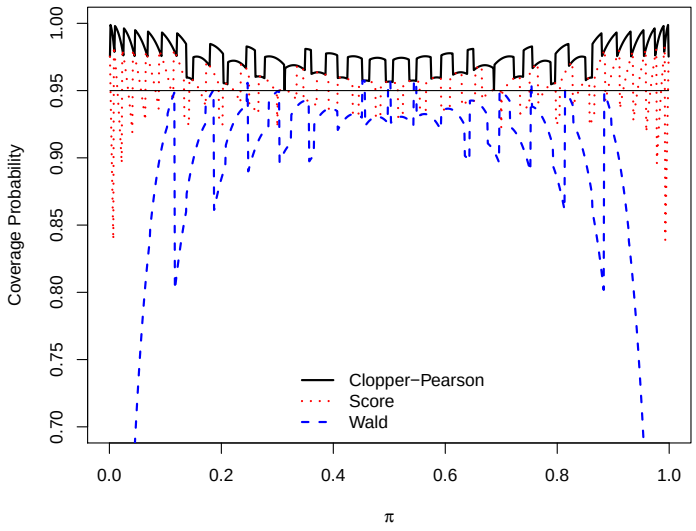

Figure 4: Exact coverage propability for three CIs

4 Appendix

Codes used:
Problem 1

```
> jellyfish=read.table("jellyfish.dat")
> plot(jellyfish$Width,jellyfish$Length, type="p", pch=16, col=jellyfish$Site,
+      xlab="Width", ylab="Length")
> legend("topleft", c("Danger Island", "Salamander Bay"),
+        col=1:length(unique(jellyfish$Site)),pch=16,title="Sites")
```

Problem 2

```
> sleep=read.table("sleep.dat", header=T)
> sleep$Treat=as.numeric(sleep$Treatment)
> #2.3
> boxplot(sleep$Hours~sleep$Treat, xlab="Treatment", ylab="Hours slept")
> plot( sleep$Patient, sleep$Hours, col=sleep$Treat, pch=16,
+       xlab="Patient", ylab="Hours of sleep")
> legend("bottomright", c("A", "B", "C", "D"),col=1:length(unique(sleep$Treat)),
+        pch=16, title="Treatments")
> fit<-lm(Hours~Treatment + factor(Patient), data=sleep)
> summary(fit)
> fit1<-aov(Hours~Treatment + factor(Patient), data=sleep)
> summary(fit1)
>
>
```

Problem 3

```
> ##clopper-pearson
> N=999
> n=25
> step=.1/(N+1)
> prob=seq(step,N*step,step)
> coverageCP=rep(0,N)
> for (pi in prob){
+   cov=0
+   for (x in 0:n){
+     int=rep(0,2)
+     int[1] = qbeta(.025,x,n-x+1)
+     int[2] = qbeta(.975,x+1,n-x)
+     if (pi>=int[1]&pi<=int[2]){
+       cov=cov+dbinom(x,n, pi)
+     }
+   }
+   coverageCP[pi*(N+1)]=cov
```

```
+ }
> #wald
> coverageWald=rep(0,N)
> for (pi in prob){
+   cov=0
+   for (x in 0:n){
+     int=rep(0,2)
+     thetaHat=x/n
+     int[1] = thetaHat-1.96*sqrt(thetaHat*(1-thetaHat)/25)
+     int[2] = thetaHat+1.96*sqrt(thetaHat*(1-thetaHat)/25)
+     if (pi>=int[1]&pi<=int[2]){
+       cov=cov+dbinom(x,n, pi)
+     }
+   }
+   coverageWald[pi*(N+1)]=cov
+ }
> ##Score
> coverageScore=rep(0,N)
> for (pi in prob){
+   cov=0
+   for (x in 0:n){
+     int=rep(0,2)
+     thetaHat=x/n
+     z=1.96
+     int[1] = 1/(1+z^2/n)*(thetaHat+z^2/(2*n)-z*sqrt(thetaHat*(1-thetaHat)/n+z^2/(4*n^2)))
+     int[2] = 1/(1+z^2/n)*(thetaHat+z^2/(2*n)+z*sqrt(thetaHat*(1-thetaHat)/n+z^2/(4*n^2)))
+     if (pi>=int[1]&pi<=int[2]){
+       cov=cov+dbinom(x,n, pi)
+     }
+   }
+   coverageScore[pi*(N+1)]=cov
+ }
> plot(prob,coverageCP,type="l",col="black", ylim=c(.7,1),
+      ylab= "Coverage Probability", xlab=expression(pi), lwd=2)
> lines(prob,coverageScore, col="red", lty=3, lwd=2)
> lines(prob,coverageWald,col="blue", lty=2, lwd=2)
> lines(prob,rep(.95,N))
> legend("center", c("Clopper-Pearson", "Score", "Wald"),
+        lty=c(1,3,2),lwd=c(2,2,2), col=c("black", "red", "blue"))
```