Qualifying Exam 2019

Exam #7

6/1/2019

Question 1

a) Number of meals involving fish as a positive test

```
# epiR package for calculating sensitivity and specificity using contingency tables
sensspec0 <- epi.tests(ctable0)
sensspec1 <- epi.tests(ctable1)
sensspec2 <- epi.tests(ctable2)
sensspec3 <- epi.tests(ctable3)
sensspec4 <- epi.tests(ctable4)
sensspec7 <- epi.tests(ctable7)
sensspec14 <- epi.tests(ctable14)
sensspec21 <- epi.tests(ctable21)
```

	Sensitivity	Specificity
>=0	100	0.0
>=1	100	8.0
>=2	100	19.2
>=3	100	28.0
>=4	90	28.8
>=7	70	36.8
>=14	30	89.6
>=21	30	93.6

b) Appropriate thresholds

Sensitivity refers to the true positive rate, or the probability that a test will rule in disease correctly. Specificity indicates the true negative rate, or the probability that a test will correctly rule out disease. Therefore, the probability of a false negative is 100 - sensitivity and the the false positive rate is 100 - specificity.

i. True positives

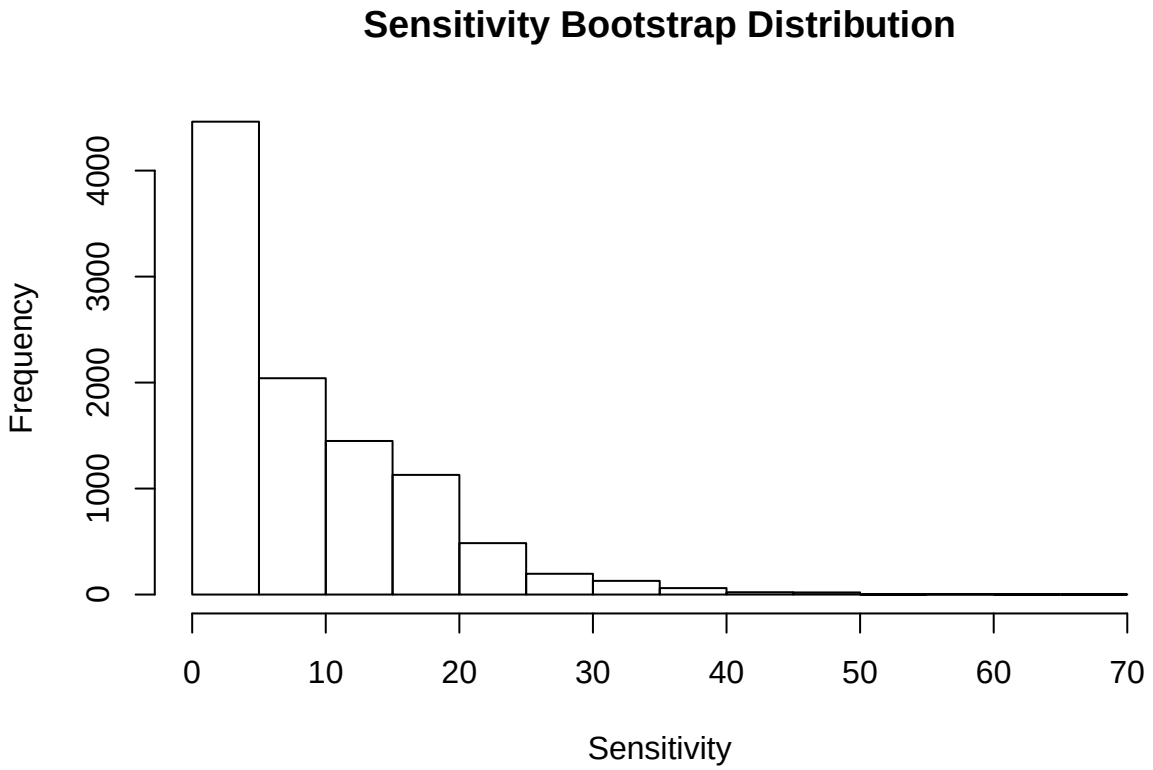
If we want to maximize true positives while minimizing false positives, the optimal threshold is the one with the highest sensitivity and lowest 100 - specificity. A threshold of >= 3 meals per week including fish would provide a 100% true positive rate and a 72% false negative rate.

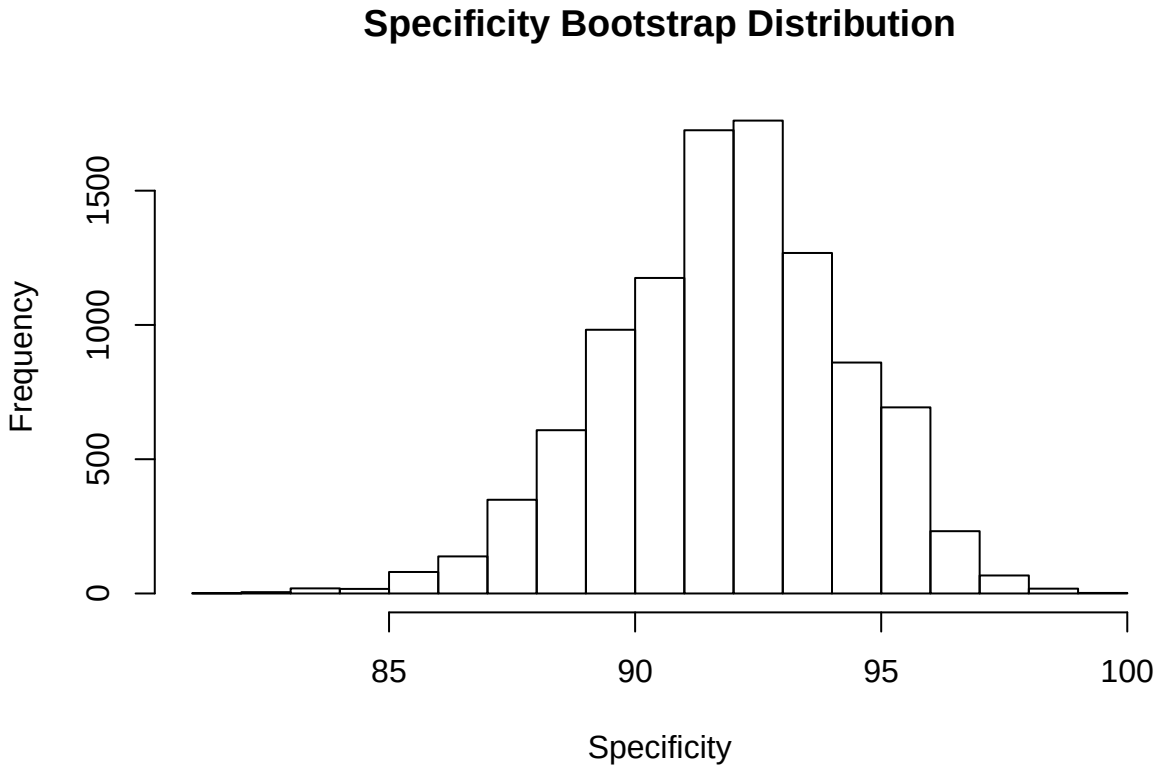
ii. True negatives

Maximizing true negatives first and then true positives requires choosing the test with highest specificity and highest sensitivity. In this case a threshold of >= 21 meals including fish per week would provide a true negative detection rate of 93.6% and a true positive rate of 30%.

c) Bootstrap sampling for >= 21 meals threshold

i. Plots





ii. Mean, SE, and Bias From Bootstrap Distributions

TALK ABOUT THESE RESULTS

	Mean	Standard Error	Bias
Sensitivity	8.16521	0.0922186	-21.83479
Specificity	91.87575	0.0240551	-1.72425

iii. 90% Bootstrap and Normal Percentile Confidence Intervals

	Normal Percentile	Coverage	Bootstrap CI
Sensitivity	-7%, 23.34%	0%, 6.71%	0%, 25%
Specificity	87.92%, 95.83%	5.58%, 4.9%	87.8%, 95.8%

The bootstrap distribution for sensitivity is not at all normal. The 90% confidence interval for this distribution using normal percentiles is (-7.00%, 23.34%), which does not make sense as sensitivity cannot be negative. Also, none of the bootstrap values were below the lower limit (again, because this is impossible), when we'd expect that 5% would be for a normal distribution. So in this case it would probably be better to use the bootstrap confidence interval (0%, 25%).

The bootstrap distribution for specificity appears to be much closer to normal than for sensitivity. The 90% normal percentile confidence interval is (87.92%, 95.83%), which matches the bootstrap confidence interval very closely (87.80%, 95.80%). Also, approximately 5% percent of the bootstrap values were in each tail, which is what we would expect from a normal distribution.

d. 90% Confidence Intervals Using Exact and Asymptotic Methods

	Clopper-Pearson	Simple Asymptotic
Sensitivity	8.73%, 60.66%	6.16%, 53.84%
Specificity	88.75%, 96.78%	90%, 97.2%

The Clopper-Pearson CI for sensitivity is (8.73%,60.66%). The simple asymptotic CI for sensitivity is (6.16%,53.84%). The Clopper-Pearson CI for specificity is (88.75%, 96.78%). The simple asymptotic CI for specificity is (90.00%, 97.20%).

In general, the normal approximation works best for large sample sizes. Although as a general rule of thumb the Central Limit Theorem applies to sample sizes over 30, the boostrap distribution of sensitivity was not normally distributed so I would use the exact confidence interval in this case.

Confidence intervals are essentially the range for a parameter that is consistent with the data. So based on the exact confidence intervals, if we were to repeat this experiment many times, sensitivity for this test would be between 8.73% and 60.66% in 90% of those experiments.

e. Linear Regression

i. Model Equation

$$\hat{MeHg} = \hat{\beta}_0 + \hat{\beta}_1 X_{\text{fisherman}} + \hat{\beta}_2 X_{\text{fish meals per week}} + \hat{\beta}_3 X_{\text{fish parts=1}} + \hat{\beta}_4 X_{\text{fish parts=2}} + \hat{\beta}_5 X_{\text{fish parts=3}}$$

In the model above, $\hat{\beta}_1$ is the estimate for the effect of being a fisherman on mercury levels. $\hat{\beta}_2$ is the estimated effect of the number of fish meals per week on mercury levels. In this model we are treating the number of fish meals per week as continuous. $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$ are the estimated effect of eating muscle tissue only, muscle tissue and sometimes the whole fish, or the whole fish (respectively). $\hat{\beta}_0$, the intercept, is the average mercury level for someone who is not a fisherman, eats 0 fish meals per week, and do not consume any fish parts.

ii. Results

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.9040000	0.8420135	1.0736170	0.2849986
Fisherman = Yes	0.2464962	0.7417227	0.3323293	0.7401801
Fish Meals per Week	0.0964710	0.0568348	1.6973942	0.0920335
Fish Part = Muscle	3.0608992	1.0782386	2.8387957	0.0052624
Fish Part = Muscle and Whole	1.6757469	1.0186581	1.6450533	0.1023934
Fish Part = Whole	3.0091672	1.3660412	2.2028378	0.0293821

iii. Summary

On average, being a fisherman increases mercury levels by 0.246 (95% CI: -1.221,1.714), but this relationship is not statistically significant (p = 0.740).

f. Fishermen Who Eat 4 Meals of Whole Fish Each Week

i. Average

$$\mathbf{a} = \begin{pmatrix} 1 & 1 & 4 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} 0.904 & 0.246 & 0.096 & 3.061 & 1.676 & 3.009 \end{pmatrix}$$

$$\hat{Y} = \mathbf{a}^{\mathrm{T}}\boldsymbol{\beta} = 4.543$$

$$\text{CI} = \hat{Y} \pm t_{\frac{\alpha}{2}} \sqrt{(MSE)\mathbf{a}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{a}}$$

On average, fishermen who eat 4 meals of whole fish each week will have a mercury level of 4.546 (95% CI: 3.071,6.019).

ii. Individual

$$\hat{Y} = \mathbf{a}^{\mathrm{T}}\boldsymbol{\beta} = 4.543$$

$$\text{CI} = \hat{Y} \pm t_{\frac{\alpha}{2}} \sqrt{(MSE)(1 + \mathbf{a}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{a})}$$

An individual fisherman who eats 4 meals of whole fish each week will have a mercury level of 4.546 (95% CI: -0.011,9.102).

iii. Prediction Interval vs. Confidence interval

The confidence interval above gives us information about the average mercury level for fishermen who eat 4 meals of whole fish each week in the current sample. However, the prediction interval refers to the mercury level for a theoretical new study participant. So because we are trying to make inference about a broader population, we need to account for some uncertainty in our estimators, which results in a wider interval.

Question 2

a.

$$log(Y_i^*) \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$Yi = 1 \text{ when } Y_i^* > l = 0.001$$

$$Yi = 1 \text{ when } log(Y_i^*) > log(l)$$

$$P(log(Y_i^*) > log(l)) = 1 - P(log(Y_i^*) \leq log(l))$$

$$= 1 - P(\frac{log(Y_i^*) - \beta_0 - \beta_1 X_i}{\sigma} \leq \frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma}) = 1 - P(Z \leq \frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma})$$

$$= 1 - \Phi(\frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma})$$

$$\text{Set } P(log(Y_i^*) > log(l)) = \theta_i$$

$$Y_i \sim \text{Bernoulli}(\theta_i)$$

In order for this model to work, $\frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma}$ must be between 0 and 1, so $0 \leq log(l) - (\beta_0 + \beta_1 X_i) \leq \sigma$, which means $log(l) \geq (\beta_0 + \beta_1 X_i)$.

Next calculate the log likelihood of Yi:

$$L(Y_i|\theta) = \prod_{i=1}^{n} \theta_i^{Y_i} (1 - \theta_i)^{1 - Y_i}$$

$$logL(Y_i|\theta) = \sum_{i=1}^{n} Y_i log(\theta_i) + (1 - Y_i) log(1 - \theta_i)$$

$$= \sum_{i=1}^{n} Y_i log(1 - \Phi(\frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma})) + (1 - Y_i) log(\Phi(\frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma}))$$

This is a probit regression model, so we can use glm() to calculate the coefficient estimates and their standard error:

```
probit_mod <-
  glm(Yi ~ factor(location), family = binomial(link = "probit"), data = pcbs)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7215223	0.2365641	3.050007	0.0022884
factor(location)Niagara	1.1679877	0.4933294	2.367561	0.0179058

b.

Because $log(Y_i^*)$ is a standard normal distribution with a location and scale shift, we can rewrite the distribution as $log(Y_i^*) \sim \frac{1}{\sigma}\phi(\frac{log(Y_i^*)-(\beta_0+\beta_1 X_i)}{\sigma})$. Next we define an indicator variable:

$$I(Y_i) = \begin{cases} 0, & \text{if } log(Y_i^*) \leq log(l) \\ 1, & \text{if } log(Y_i^*) > log(l) \end{cases}$$

The resulting likelihood is similar to above, but includes the normal PDF of $log(Y_i^*)$:

$$L(log(Y_i^*)|\theta) = \prod_{i=1}^{n}(\frac{1}{\sigma}\phi(\frac{log(Y_i^*) - (\beta_0 + \beta_1 X_i)}{\sigma})^{I(Y_i)}(\Phi(\frac{log(l) - \beta_0 - \beta_1 X_i}{\sigma}))^{1-I(Y_i)}$$

So for observations where $log(Y_i^*) > log(l)$, the contribution to the likelihood function is the PDF of $log(Y_i^*)$. For observations that are below the detection threshold, the contribution to the likelihood is just the probability of being below the threshold, as defined above. This is a type I Tobit model, which can be fit using the VGAM package:

```
tobit <- AER::tobit(contam.lev ~ factor(location), left = 0.001, data = pcbs)
summary(tobit)$coefficients
```

```
##
## Test of coefficients:
##
##                            Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)               0.0034856  0.0278652   0.1251    0.9005
## factor(location)Niagara   0.1763509  0.0384882   4.5819 4.607e-06 ***
## Log(scale)               -1.8670597  0.0929449 -20.0878 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 3

```
cat_mod <- lmer(logvl~trt*week+(1|pid),data=hiv)
cat_mod <- lme(logvl~trt*week,random = ~1|pid,data=hiv)
summary(cat_mod)$tTable
```

```
##                    Value    Std.Error   DF   t-value      p-value
## (Intercept)  4.013867591 0.0772666570 2910 51.948250 0.000000e+00
## trt         -0.837790042 0.1087869774  222 -7.701198 4.417947e-13
## week         0.005365891 0.0008358772 2910  6.419473 1.591148e-10
## trt:week     0.001361022 0.0011768667 2910  1.156479 2.475802e-01
```

Appendix

Load libraries, import and format data

```
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
library(epiR)
library(ggplot2)
library(lme4)
library(nlme)
library(VGAM)
library(AER)
```

```
# Import datasets
fish <- read.csv("/Users/timvigers/Documents/School/Qualifying Exams/2019 MS/fishermen_mercury.csv")
pcbs <- read.csv("/Users/timvigers/Documents/School/Qualifying Exams/2019 MS/pcb_conc.csv")
hiv <- read.csv("/Users/timvigers/Documents/School/Qualifying Exams/2019 MS/hiv_comp_tall.csv",header =
colnames(hiv) <- c("pid","trt","week","logvl")
```

Question 1

Create dummy variables for fish data

```
# Create indicator variables and response
fish$meal0 <- ifelse(fish$fishmlwk >= 0,1,0)
fish$meal1 <- ifelse(fish$fishmlwk >= 1,1,0)
fish$meal2 <- ifelse(fish$fishmlwk >= 2,1,0)
fish$meal3 <- ifelse(fish$fishmlwk >= 3,1,0)
fish$meal4 <- ifelse(fish$fishmlwk >= 4,1,0)
fish$meal7 <- ifelse(fish$fishmlwk >= 7,1,0)
fish$meal14 <- ifelse(fish$fishmlwk >= 14,1,0)
fish$meal21 <- ifelse(fish$fishmlwk >= 21,1,0)
fish$response <- ifelse(fish$MeHg >= 8,1,0)
# Create contingency tables
ctable0 <- table(factor(fish$meal0,levels=1:0),factor(fish$response,levels=1:0))
ctable1 <- table(factor(fish$meal1,levels=1:0),factor(fish$response,levels=1:0))
ctable2 <- table(factor(fish$meal2,levels=1:0),factor(fish$response,levels=1:0))
ctable3 <- table(factor(fish$meal3,levels=1:0),factor(fish$response,levels=1:0))
ctable4 <- table(factor(fish$meal4,levels=1:0),factor(fish$response,levels=1:0))
ctable7 <- table(factor(fish$meal7,levels=1:0),factor(fish$response,levels=1:0))
ctable14 <- table(factor(fish$meal14,levels=1:0),factor(fish$response,levels=1:0))
ctable21 <- table(factor(fish$meal21,levels=1:0),factor(fish$response,levels=1:0))
# Make results table
sens_spec_results <- as.data.frame(matrix(ncol = 2,nrow = 8))
colnames(sens_spec_results) <- c("Sensitivity","Specificity")
rownames(sens_spec_results) <- c(">=0",">=1",">=2",">=3",">=4",">=7",">=14",">=21")
```

Format sensitivity and specificity results

```
# Format results
sens_spec_results[">=0","Specificity"] <-
  round(sensspec0$elements$specificity$est*100,1)
sens_spec_results[">=0","Sensitivity"] <-
  round(sensspec0$elements$sensitivity$est*100,1)
sens_spec_results[">=1","Specificity"] <-
  round(sensspec1$elements$specificity$est*100,1)
sens_spec_results[">=1","Sensitivity"] <-
  round(sensspec1$elements$sensitivity$est*100,1)
sens_spec_results[">=2","Specificity"] <-
  round(sensspec2$elements$specificity$est*100,1)
sens_spec_results[">=2","Sensitivity"] <-
  round(sensspec2$elements$sensitivity$est*100,1)
sens_spec_results[">=3","Specificity"] <-
  round(sensspec3$elements$specificity$est*100,1)
sens_spec_results[">=3","Sensitivity"] <-
  round(sensspec3$elements$sensitivity$est*100,1)
sens_spec_results[">=4","Specificity"] <-
  round(sensspec4$elements$specificity$est*100,1)
sens_spec_results[">=4","Sensitivity"] <-
  round(sensspec4$elements$sensitivity$est*100,1)
sens_spec_results[">=7","Specificity"] <-
  round(sensspec7$elements$specificity$est*100,1)
sens_spec_results[">=7","Sensitivity"] <-
  round(sensspec7$elements$sensitivity$est*100,1)
sens_spec_results[">=14","Specificity"] <-
  round(sensspec14$elements$specificity$est*100,1)
sens_spec_results[">=14","Sensitivity"] <-
  round(sensspec14$elements$sensitivity$est*100,1)
sens_spec_results[">=21","Specificity"] <-
  round(sensspec21$elements$specificity$est*100,1)
sens_spec_results[">=21","Sensitivity"] <-
  round(sensspec21$elements$sensitivity$est*100,1)
```

Bootstrap

```
# Vector for storing results
set.seed(1234)
B <- 10000
sens_results <- numeric(B)
spec_results <- numeric(B)
# Loop
for (i in 1:B) {
  meals <- sample(fish$fishmlwk,replace = T)
  meals <- ifelse(meals >= 21,1,0)
  response <- sample(fish$MeHg,replace = T)
  response <- ifelse(response >= 8,1,0)
  table <- table(factor(meals,levels=1:0),factor(response,levels=1:0))
  sens_results[i] <- (table[1,1]/sum(table[,1])) * 100
  spec_results[i] <- (table[2,2]/sum(table[,2])) * 100
}
```

Mean, SE, and Bias for Bootstrap Distributions

```
boot_results <- as.data.frame(matrix(ncol = 3,nrow = 2))
colnames(boot_results) <- c("Mean","Standard Error","Bias")
rownames(boot_results) <- c("Sensitivity","Specificity")
# Sensitivity
boot_results["Sensitivity","Mean"] <- mean(sens_results)
boot_results["Sensitivity","Standard Error"] <-
  sd(sens_results)/sqrt(length(sens_results))
boot_results["Sensitivity","Bias"] <-
  mean(sens_results) - sensspec21$elements$sensitivity$est*100
# Specificity
boot_results["Specificity","Mean"] <- mean(spec_results)
boot_results["Specificity","Standard Error"] <-
  sd(spec_results)/sqrt(length(spec_results))
boot_results["Specificity","Bias"] <-
  mean(spec_results) - sensspec21$elements$specificity$est*100
```

Bootstrap and Normal Percentile Confidence Intervals

```
# Sensitivity
# Normal percentiles
L <- mean(sens_results) - (1.645 * sd(sens_results))
U <- mean(sens_results) + (1.645 * sd(sens_results))
# Specificity
# Normal percentiles
Lc <- mean(spec_results) - (1.645 * sd(spec_results))
Uc <- mean(spec_results) + (1.645 * sd(spec_results))
# Results table
results <- as.data.frame(matrix(ncol = 3,nrow = 2))
rownames(results) <- c("Sensitivity","Specificity")
colnames(results) <- c("Normal Percentile","Coverage","Bootstrap CI")
L <- round(L,2)
U <- round(U,2)
Lc <- round(Lc,2)
Uc <- round(Uc,2)
results["Sensitivity",] <-
  c(paste0(L,"%, ",U,"%"),
    paste0(round(sum(sens_results < L)/B * 100,2),"%, ",
           round(sum(sens_results > U)/B * 100,2),"%"),
    paste0(paste(round(quantile(sens_results,c(0.05,0.95)),2),collapse = "%, "),"%"))
results["Specificity",] <-
  c(paste0(Lc,"%, ",Uc,"%"),
    paste0(round(sum(spec_results < Lc)/B * 100,2),"%, ",
           round(sum(spec_results > Uc)/B * 100,2),"%"),
    paste0(paste(round(quantile(spec_results,c(0.05,0.95)),2),collapse = "%, "),"%"))
```

90% Confidence Intervals Using Exact and Asymptotic Methods

```
# Sensitivity
# Clopper-Pearson
```

```
results <- as.data.frame(matrix(ncol = 2,nrow = 2))
rownames(results) <- c("Sensitivity","Specificity")
colnames(results) <- c("Clopper-Pearson","Simple Asymptotic")
n <- sum(ctable21[,1])
x <- ctable21[1,1]
L <- x/(x+((n-x+1)*qf(0.95,(2*(n-x+1)),2*x))) * 100
U <- (x+1)*qf(0.95,(2*(x+1)),2*(n-x))/((n-x)+(x+1)*qf(0.95,(2*(x+1)),2*(n-x))) * 100
results["Sensitivity",1] <- paste0(round(L,2),"%, ",round(U,2),"%")
# Simple asymptotic
n <- sum(ctable21[,1])
phat <- 0.3
L <- (phat - qnorm(0.95)*sqrt((phat*(1-phat))/n))*100
U <- (phat + qnorm(0.95)*sqrt((phat*(1-phat))/n))*100
results["Sensitivity",2] <- paste0(round(L,2),"%, ",round(U,2),"%")
# Specificity
# Clopper-Pearson
n <- sum(ctable21[,2])
x <- ctable21[2,2]
L <- x/(x+((n-x+1)*qf(0.95,(2*(n-x+1)),2*x))) * 100
U <- (x+1)*qf(0.95,(2*(x+1)),2*(n-x))/((n-x)+(x+1)*qf(0.95,(2*(x+1)),2*(n-x))) * 100
results["Specificity",1] <- paste0(round(L,2),"%, ",round(U,2),"%")
# Simple asymptotic
n <- sum(ctable21[,2])
phat <- 0.936
L <- (phat - qnorm(0.95)*sqrt((phat*(1-phat))/n))*100
U <- (phat + qnorm(0.95)*sqrt((phat*(1-phat))/n))*100
results["Specificity",2] <- paste0(round(L,2),"%, ",round(U,2),"%")
```

Linear Model

```
lin_mod <- lm(MeHg ~ factor(fisherman)+fishmlwk+factor(fishpart),data = fish)
results <- as.data.frame(summary(lin_mod)$coefficients)
rownames(results) <- c("Intercept","Fisherman = Yes",
                       "Fish Meals per Week","Fish Part = Muscle",
                       "Fish Part = Muscle and Whole",
                       "Fish Part = Whole")
```

Confidence and Prediction Intervals

Confidence

```
a <- matrix(c(1,1,4,0,0,1))
b <- as.numeric(summary(lin_mod)$coefficients[,1])
yhat <- t(a)%*%b
mse <- mean(lin_mod$residuals^2)
t <- qt(0.1/2,133,lower.tail = F)
x <- model.matrix(lin_mod)
L <- yhat - t*sqrt(mse*(t(a)%*%(solve((t(x)%*%x)))%*%a))
U <- yhat + t*sqrt(mse*(t(a)%*%(solve((t(x)%*%x)))%*%a))
```

Prediction

```r
a <- matrix(c(1,1,4,0,0,1))
b <- as.numeric(summary(lin_mod)$coefficients[,1])
yhat <- t(a)%*%b
mse <- mean(lin_mod$residuals^2)
t <- qt(0.1/2,133,lower.tail = F)
x <- model.matrix(lin_mod)
L <- yhat - t*sqrt(mse*(1+(t(a)%*%(solve((t(x)%*%x)))%*%a)))
U <- yhat + t*sqrt(mse*(1+(t(a)%*%(solve((t(x)%*%x)))%*%a)))
```

Question 2

Format data