Categorical Data Analysis: Outline of Solutions to HW1

1. (a)

$$P(Y_1 = 1) = \frac{13}{52}$$

(b)

$$P(Y_2 = 1|Y_1 = 1) = \frac{12}{51}$$
$$P(Y_2 = 1|Y_1 = 0) = \frac{13}{51}$$

(c)

$$\begin{aligned} P(Y_2 = 1) &= P(Y_2 = 1|Y_1 = 1)P(Y_1 = 1) + P(Y_2 = 1|Y_1 = 0)P(Y_1 = 0) \\ &= \frac{12}{51} \times \frac{13}{52} + \frac{13}{51} \times \frac{39}{52} \\ &= \frac{1}{4} \\ &= \frac{13}{52} \end{aligned}$$

(d)

(e)

$$P(Y_2 = 1|Y_1 = 1) = \frac{12}{51} \neq \frac{13}{52} = P(Y_2 = 1)$$

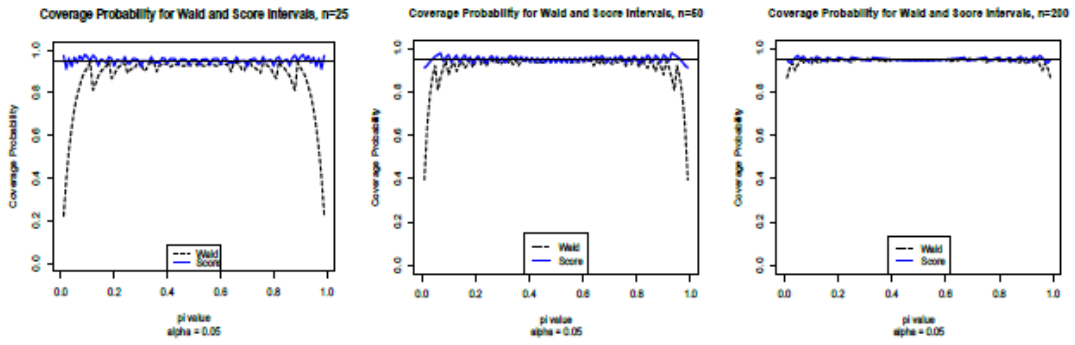
For part (f): (read carefully the help file for dhyper!)

```
> dhyper(3,4,48,5)
[1] 0.001736079

> dbinom(3,size=5,prob=4/52)
[1] 0.003878339
```

2. Here is a solution that computes the exact coverage for n=25 at many values of pi. However, setting n <- 15 and pi.vec <- c(0.05, 0.1, 0.25, 0.5) will give the solution for part a on the HW.

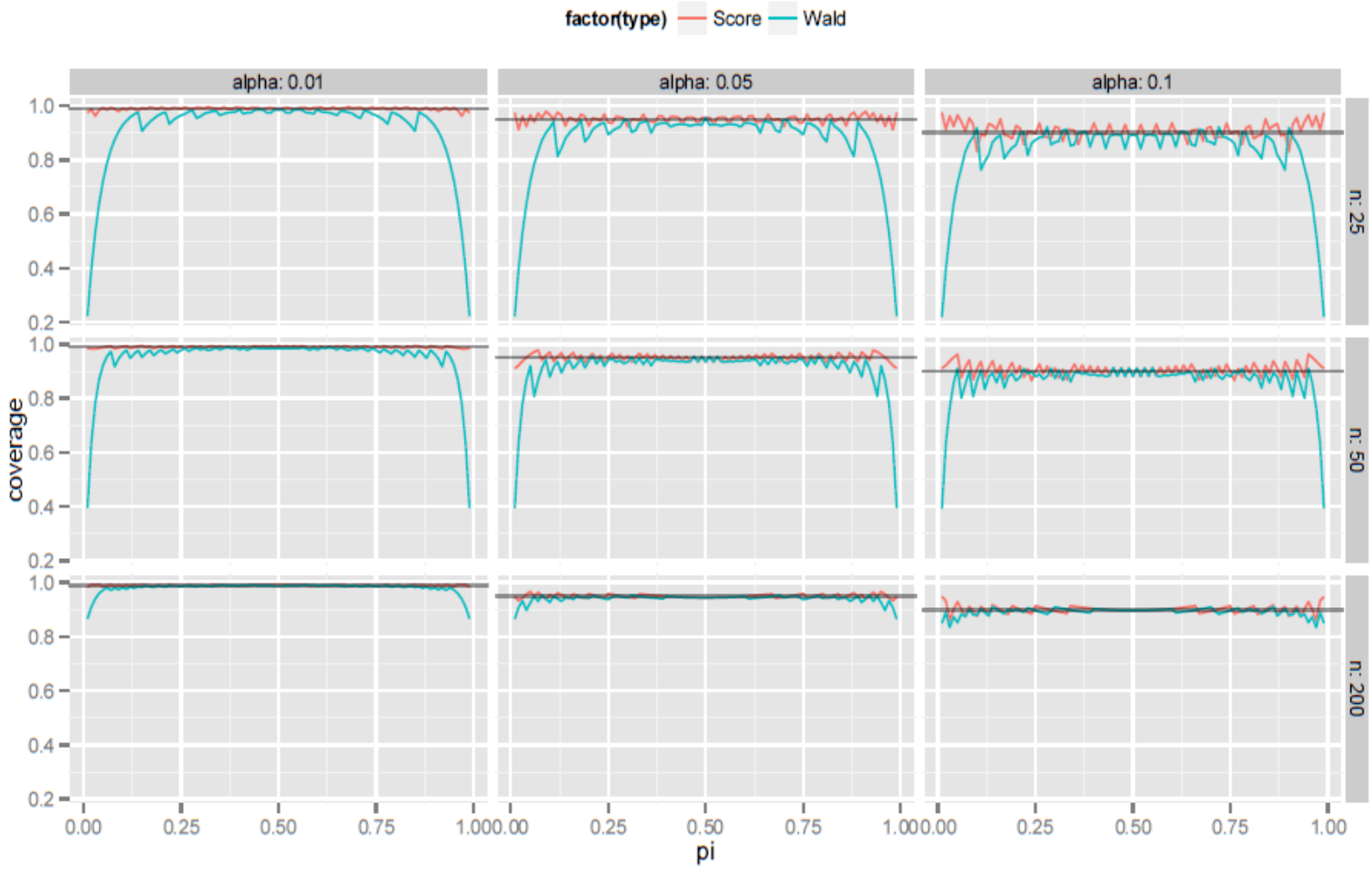
```
# can adjust n and alpha values as required
n <- 25
alpha <- 0.05
```

```
x.vec <- 0:n
c <- qnorm(1-alpha/2)
wald.ci <- sapply(x.vec, function(x){
pi.hat <- x/n
sd <- sqrt(pi.hat*(1-pi.hat)/n)
return(c(pi.hat - c*sd, pi.hat + c*sd))
}) # first row is the lower bounds, second row is the upper bounds
score.ci <- sapply(x.vec, function(x){
return(prop.test(x,n,correct=FALSE,conf.level=1-alpha)$conf.int)
}) # first row is the lower bounds, second row is the upper bounds

pi.vec <- seq(0.01,0.99,0.01)
coverage <- sapply(pi.vec, function(pi){
probs <- dbinom(x.vec, n, pi)
covers.wald <- wald.ci[1,] <= pi & pi <= wald.ci[2,]
covers.score <- score.ci[1,] <= pi & pi <= score.ci[2,]
return(c( sum(probs*covers.wald), sum(probs*covers.score) ))
}) # first row gives the Wald coverage, second row gives the score

plot(coverage[1,] ~ pi.vec,
type="l",
lty="dashed",
main=paste("Coverage Probability for Wald and Score Intervals, n=",n,sep=""),
sub = paste("alpha = ", alpha, sep=""),
xlab="pi value", ylab="Coverage Probability",
ylim = c(0,1)
)
lines( coverage[2,] ~ pi.vec, col="blue")
legend("bottom",c("Wald", "Score"),col=c("black", "blue"),lty=c(2,1))
abline(h=1-alpha)
```



Here is a solution that uses the ggplot2 package for plotting the exact coverage at various alpha levels:

```
# need to install and load the ggplot2 package first:
install.packages("ggplot2")
require(ggplot2)
settings.mat <- expand.grid(n=c(25,50,200),alpha=c(0.05, 0.10, 0.01))
pi.vec <- seq(0.01,0.99,0.01)
plot.data <- apply(settings.mat, 1, function(settings){
n <- as.numeric(settings[1])
alpha <- as.numeric(settings[2])
x.vec <- 0:n
    c <- qnorm(1-alpha/2)
wald.ci <- sapply(x.vec, function(x){
pi.hat <- x/n
sd <- sqrt(pi.hat*(1-pi.hat)/n)
return(c(pi.hat - c*sd, pi.hat + c*sd))
}) # first row is the lower bounds, second row is the upper bounds
score.ci <- sapply(x.vec, function(x){
return(prop.test(x,n,correct=FALSE,conf.level=1-alpha)$conf.int)
}) # first row is the lower bounds, second row is the upper bounds
coverage <- sapply(pi.vec, function(pi){
probs <- dbinom(x.vec, n, pi)
covers.wald <- wald.ci[1,] <= pi & pi <= wald.ci[2,]
covers.score <- score.ci[1,] <= pi & pi <= score.ci[2,]
return(c( sum(probs*covers.wald), sum(probs*covers.score) ))
}) # first row gives the Wald coverage, second row gives the score
cov.wald <- coverage[1,]
cov.score <- coverage[2,]
data.temp1 <- data.frame(pi = pi.vec, coverage = cov.wald)
data.temp1$type="Wald"
data.temp2 <- data.frame(pi = pi.vec, coverage = cov.score)
data.temp2$type="Score"
data <- rbind(data.temp1, data.temp2)
data$n <- n
data$alpha <- alpha
return(data)
})

plot.data <- do.call(rbind, plot.data)
plot.data$n <- factor(plot.data$n)
plot.data$alpha <- factor(plot.data$alpha)
hline.data <- data.frame(alpha=levels(plot.data$alpha),
hl=1-as.numeric(levels(plot.data$alpha)))
qplot(x=pi,y=coverage,color=factor(type), geom="line", data=plot.data) +
  facet_grid(n~alpha,labeller=label_both) +
  geom_hline(aes(yintercept=hl), data=hline.data, alpha = 0.4) +
    theme(legend.position = "top")
```


factor(type)
Score
Wald
alpha: 0.01
alpha: 0.05
alpha: 0.1
n: 25
n: 50
n: 200
coverage
pi
1.0
0.8
0.6
0.4
0.2
0.00
0.25
0.50
0.75
1.00

3. Let

$$Y_{ij} = \begin{cases} 1 & \text{when the } i\text{th trial results in group } j \\ 0 & \text{when the } i\text{th trial does not result in group } j \end{cases}$$

Then, we have $P(Y_{ij} = 1) = \pi_j$ and $P(Y_{ij} = 0) = 1 - \pi_j$. In other words, $Y_{ij} \sim$ Bernoulli(π_j). Then,

$$\begin{aligned} \text{Var}(N_j) = & \text{Var}\left(\sum_{i=1}^{N} Y_{ij}\right) \\ &= \sum_{i=1}^{N} \text{Var}(Y_{ij}) \quad \text{(independence)} \\ &= \sum_{i=1}^{N} \pi_j(1-\pi_j) \quad \text{(variance of a Bernoulli)} \\ &= n\pi_j(1-\pi_j) \end{aligned}$$