# Methods Homework 3

*Tim Vigers*

*September 23, 2018*

## A

### 1.

```r
# Read in the data.
ozone <- read.csv("C:/Users/timbv/Documents/School/UC Denver/Biostatistics/Biostatistical Methods 1/Home
# Divide the number of good days by the total number of days.
good.prob <- length(which(ozone$AQI.Category == "Good")) /
  length(ozone$AQI.Category)
# Print probability of good ozone levels.
good.prob
```

```
## [1] 0.5298013
```

### 2.

```r
# Add the probability of 5 good days, 6 good days, and 7 good days.
dbinom(5,7,good.prob) + dbinom(6,7,good.prob) + dbinom(7,7,good.prob)
```

```
## [1] 0.2783011
```

### 3.

```r
# Find the normal interval using P(x-0.5 < X < x+0.5).
np <- 7 * good.prob
sigma <- sqrt(np * (1-good.prob))
pnorm(0.5/sigma)-pnorm(-0.5/sigma)
```

```
## [1] 0.295043
```

### 4.

I don't think it makes sense to use the binomial distribution for calculating days with "good" ozone status, because the binomial distribution assumes independent trials (like flipping a coin). Air quality from one day to the next can't be independent, because the air quality one day affects the air quality the next. Like with temperature, you can't really have wild swings from one day to the next because it's a continuous process.

# B.

## 1.

```
# Read in the procedure cost data.
proc.cost <- read.csv("C:/Users/timbv/Documents/School/UC Denver/Biostatistics/Biostatistical Methods 1/
# Add a column reducing cost to two factors.
proc.cost$Cost.factor <- 0
proc.cost$Cost.factor[proc.cost$Cost > 0] <- 1
# Make a new table without full cost, for frequency count.
temp <- proc.cost[,c(1,3)]
# Frequency table.
cost.count <- table(temp,exclude = "Cost")
# Format the frequency table.
dimnames(cost.count)$`Cost.factor` <- c("Zero","Non-Zero")
print(cost.count)
```

```
##          Cost.factor
## Procedure Zero Non-Zero
##         1   48       72
##         2   15       65
```

## 2.

```
# Calculate the proportion of non-zero costs for each procedure.
p1 <- cost.count[1,2] / sum(cost.count[1,])
p2 <- cost.count[2,2] / sum(cost.count[2,])
p1
```

```
## [1] 0.6
```

```
p2
```

```
## [1] 0.8125
```

```
# Calculate the mean non-zero cost for each procedure.
m1 <-
  mean(proc.cost$Cost[which(proc.cost$Procedure == 1 &
                              proc.cost$Cost.factor == 1)])
m2 <-
  mean(proc.cost$Cost[which(proc.cost$Procedure == 2 &
                              proc.cost$Cost.factor == 1)])
m1
```

```
## [1] 2.155417
```

```
m2
```

```
## [1] 1.085077
```

```
# Calculate the variance of non-zero cost for each procedure.
v1 <-
  var(proc.cost$Cost[which(proc.cost$Procedure == 1 &
                             proc.cost$Cost.factor == 1)])
v2 <-
```

```
  var(proc.cost$Cost[which(proc.cost$Procedure == 2 &
                              proc.cost$Cost.factor == 1)])
v1
```

```
## [1] 1.262825
```

```
v2
```

```
## [1] 1.58376
```

### 3.

Expected value of Y, assuming R and Z are independent:

$$E[RZ] = E[R] * E[Z] = pm$$

Variance of Y, assuming R and Z are independent:

$$V[RZ] = E[(RZ)^2] - E[RZ]^2 = E[R^2]E[Z^2] - E[R]^2E[Z]^2 =$$

$$(Var(R) + E[R]^2)(Var(Z) + E[Z]^2) - E[R]^2E[Z]^2 =$$

$$Var(R)Var(Z) + Var(R)E[Z]^2 + Var(Z)E[R]^2 + E[R]^2E[Z]^2 - E[R]^2E[Z]^2 =$$

$$Var(R)Var(Z) + Var(R)E[Z]^2 + Var(Z)E[R]^2 = Var(R)v + Var(R)m^2 + vE[R]^2$$

$$Var(R) = E[R^2] - E[R]^2 = p - p^2 = p(1-p) \text{ so:}$$

$$V[RZ] = p(1-p)v + p(1-p)m^2 + vp^2 =$$

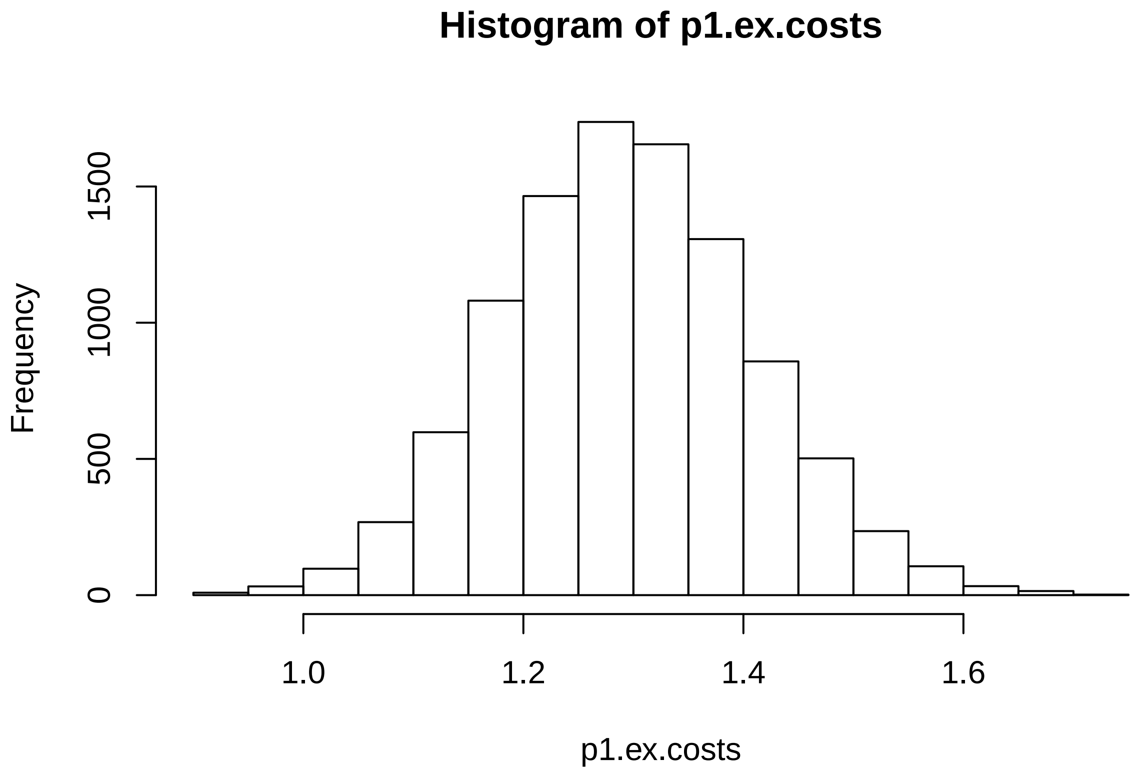
$$p(1-p)(v + m^2) + vp^2$$

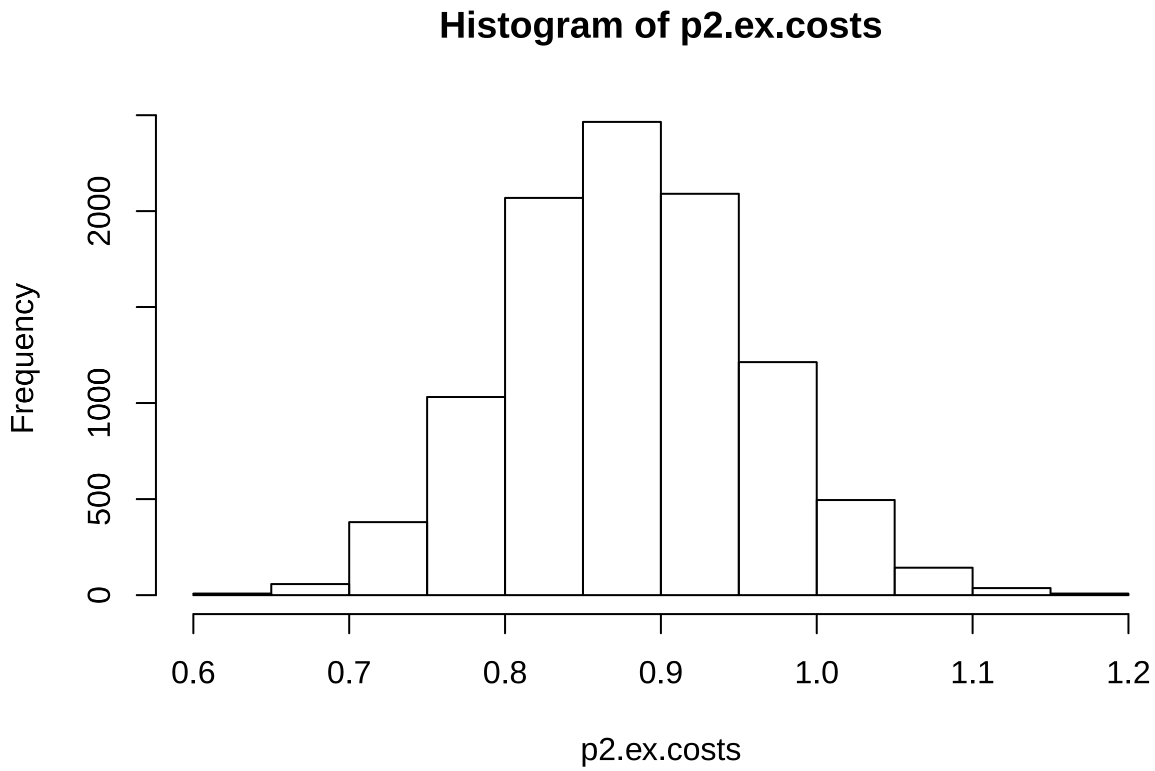
### 5.

```
# Set seed, determine number of simulations and sample size.
set.seed(1017)
number_of_sims <- 10000
n1 <- 120
n2 <- 200
p1.ex.costs <- rep(-9, number_of_sims)
p2.ex.costs <- rep(-9, number_of_sims)
# Plug in calculated values.
shape1 <- m1^2/v1
scale1 <- v1/m1
shape2 <- m2^2/v2
scale2 <- v2/m2
# Loop through simulations.
for (i in 1:number_of_sims) {
  R1 <- rbinom(n1,1,p1)
  Z1 <- rgamma(n1, shape = shape1, scale = scale1)
  p1.ex.costs[i] <- mean(R1) * mean(Z1)

  R2 <- rbinom(n2,1,p2)
  Z2 <- rgamma(n2, shape = shape2, scale = scale2)
  p2.ex.costs[i] <- mean(R2) * mean(Z2)
}
```

```r
# Because of the CTL, the simulated sample means will be normally distributed.
p1.sim.cost <- qnorm(0.8,mean(p1.ex.costs),sd(p1.ex.costs)) * 120
p2.sim.cost <- qnorm(0.8,mean(p2.ex.costs),sd(p2.ex.costs)) * 200
total.sim.cost <- p1.sim.cost + p2.sim.cost

as.numeric(total.sim.cost)
```

```
## [1] 356.3308
```


Histogram of p1.ex.costs
Frequency
0
500
1000
1500
1.0
1.2
1.4
1.6
p1.ex.costs


Histogram of p2.ex.costs
Frequency
0
500
1000
2000
0.6
0.7
0.8
0.9
1.0
1.1
1.2
p2.ex.costs