Gianni Spiga
Sta 144
Drake
4 April 2023

Homework 1

$\boxed{1.)}$ A3-1.) a Let $Y = g(x)$ s.t. $P(Y=y) = \sum_{g(x)=y} P(X=x)$

$$E(Y) = \sum_{Y} y \, P(Y=Y)$$

$$= \sum_{Y} y \sum_{g(x)} P(X=x)$$

$$= \sum_{Y} \sum_{g(x)} y \, P(X=x)$$

$$= \sum_{x} g(x) \cdot P(X=x) \quad \leftarrow \text{LOTUS}$$

2.) $E(aX+b) = \sum_{x} (aX+b) \, P(X=x)$

$$= \sum_{x} aX \cdot P(X=x) + bP(X=x)$$

$$= a \sum_{x} x \cdot P(X=x) + b \sum_{x} (X=x)$$

$$= a \cdot E(x) + b$$

3.) $E(XY) = \sum_{x} \sum_{Y} xy \, \underbrace{P(X=x, Y=y)}_{\text{joint distribution}}$

$$= \sum_{x} \sum_{y} xy \, P(X=x) \cdot P(Y=y) \quad \leftarrow \text{by indep.}$$

$$= \sum_{x} x \, P(X=x) \sum_{Y} y P(Y=y)$$

$$= E(X) \cdot E(Y)$$

4.) $Cov(X,Y) = E([X-E(X)][Y-E(Y)])$

$$= E\{XY - X \cdot E(Y) - Y \cdot E(X) + E(X) \cdot E(Y)\}$$

$$= E(XY) - E(X)E(Y) - E(X)\cdot E(Y) + E(X)E(Y)$$
$$= E(XY) - E(X)E(Y) \blacksquare$$

5.) $Cov(\sum_i a_iX_i + b_i, \sum_j c_jY_j + d_j) = E[(\sum a_iX_i + b_i)(\sum c_jY_j + d_j)] - E(\sum a_iX_i + b_i)\cdot E(\sum c_jY_j + d_j)$

$$= E(\sum_i\sum_j a_iX_ic_jY_j + a_iX_id_j + b_ic_jY_j + b_id_j) - \ldots$$
$$\ldots - E(\sum a_iX_i + b_i)\cdot E(\sum c_jY_j + d_j)$$
$$= \sum_i\sum_j (a_ic_j E(X_iY_j) + a_id_j E(X_i) + b_ic_j E(Y_j) + b_id_j)$$
$$- \sum\sum [a_iE(X_i) + b_i][c_jE(Y_j) + d_j]$$
$$= \sum_i\sum_j a_ic_j E(X_iY_j) + \cancel{a_id_jE(X_i)} + \cancel{b_ic_jE(Y_j)} + \cancel{b_id_j} - a_ic_jE(X_i)E(Y_j)$$
$$- \cancel{a_id_jE(X_i)} - \cancel{b_ic_jE(Y_j)} - \cancel{b_id_j}$$
$$= \sum_i\sum_j a_ic_j E(X_i, Y_j) - a_ic_j E(X_i)\cdot E(Y_j)$$
$$= \sum_i\sum_j a_ic_j Cov(X_i, Y_j) \blacksquare$$

6.) $Var(X) = E[(X - E(X))^2]$

$$= E[(X^2 - X\cdot E(X) - X\cdot E(X) + E(X)^2]$$
$$= E(X^2 - 2\cdot X\cdot E(X) + E(X)^2]$$
$$= E(X^2) - 2E(X)^2 + E(X)^2$$
$$= E(X^2) - E(X)^2 \blacksquare$$

7.) $Var(X+Y) = E((X+Y)^2) - [E(X+Y)]^2$

$$= E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2$$
$$= E(X^2) + 2E(XY) + E(Y^2) - [E(X)^2 + 2E(X)E(Y) + E(Y)^2]$$
$$= [E(X^2) - E(X)^2] + [E(Y^2) - E(Y)^2] + 2E(XY) - 2E(X)E(Y)$$
$$= Var(X) + Var(Y) - 2Cov(X,Y) \blacksquare$$

8.) From Sheldon Ross "First Course in Probability"

$$0 \le Var\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) = 2 \pm 2Corr(X,Y)$$

$2 + 2Corr(X,Y) \ge 0$	$2 - 2Corr(X,Y) \ge 0$
$2Corr(X,Y) \ge -2$	$-2Corr(X,Y) \ge -2$
$Corr(X,Y) \ge -1$	$Corr(X,Y) \le 1$

$\blacksquare$

Homework 1

Gianni Spiga

2023-04-11

Contents

Problem 1

Please see written file uploaded.

Problem 2

a.)

```
samp <- c(15, 34, 35, 36, 11, 17, 36, 15)

sampWReplace <- expand.grid(samp, samp, samp)
# We will 8^3 possibilities, 512

# Now we find all possible, *unique*, sums
Tsums <- rowSums(sampWReplace)

unique(Tsums)
```

```
##  [1]  45  64  65  66  41  47  83  84  85  60  86  61  67  87  62  68  37  43  49
## [20] 102 103 104  79 105  80 106  81  56 107  82  88  57  63  69 108  89  58  70
## [39]  33  39  51
```

b.)

```
# sampling distribution of T
table(Tsums) / length(Tsums)
```

```
## Tsums
##          33          37          39          41          43          45
## 0.001953125 0.011718750 0.005859375 0.023437500 0.023437500 0.021484375
##          47          49          51          56          57          58
## 0.023437500 0.011718750 0.001953125 0.005859375 0.005859375 0.011718750
##          60          61          62          63          64          65
## 0.023437500 0.023437500 0.058593750 0.011718750 0.046875000 0.023437500
##          66          67          68          69          70          79
## 0.070312500 0.023437500 0.052734375 0.005859375 0.011718750 0.005859375
##          80          81          82          83          84          85
## 0.011718750 0.029296875 0.023437500 0.035156250 0.023437500 0.064453125
##          86          87          88          89         102         103
## 0.058593750 0.076171875 0.023437500 0.023437500 0.001953125 0.005859375
##         104         105         106         107         108
## 0.017578125 0.025390625 0.035156250 0.023437500 0.015625000
```

c.)

To find the correlation, we would first be interested in calculating the covariance, specifically:

$$cov(z_i, z_j) = E(z_i, z_j) - E(z_i)E(z_j)$$

However, in drawing with replacement, we know that each draw is independent, thus we have:

$$cov(z_i, z_j) = E(z_i)E(z_j) - E(z_i)E(z_j) = 0$$

Thus, with covariance 0, we can soundly conclude that independent draws have correlation 0.

d.)

```r
sampWOReplace <- t(combn(samp, m = 3))
TsumsNoRep <- rowSums(sampWOReplace)

# sampling distribution of T w/o replacement
table(TsumsNoRep) / length(TsumsNoRep)
```

```
## TsumsNoRep
##         41         43         47         60         61         62         63
## 0.01785714 0.03571429 0.01785714 0.03571429 0.03571429 0.08928571 0.01785714
##         64         65         66         67         68         80         81
## 0.05357143 0.01785714 0.07142857 0.03571429 0.07142857 0.01785714 0.03571429
##         82         83         84         85         86         87         88
## 0.03571429 0.01785714 0.03571429 0.07142857 0.08928571 0.07142857 0.03571429
##         89        105        106        107
## 0.01785714 0.03571429 0.01785714 0.01785714
```

e.)

We once again have:

$$cov(z_i, z_j) = E(z_i, z_j) - E(z_i)E(z_j)$$

However, these draws are not independent, thus the joint expectation is not separable. However, since our variables are binary, there is only one instance when the covariance is not zero, and that is when both z_i and z_j are both 1.

$$\begin{aligned}
cov(z_i, z_j) &= P(z_i = 1, z_j = 1) - P(z_i = 1)P(z_j = 1) \\
&= P(z_j = 1 | z_i = 1)P(z_i = 1) - P(z_i = 1)P(z_j = 1) \\
&= \frac{n-1}{N-1}\frac{n}{N} - \frac{n^2}{N^2} \\
&= \frac{n}{N}(\frac{n-1}{N-1} - \frac{n}{N}) \\
&= \frac{3}{8}(\frac{2}{7} - \frac{3}{8}) \\
&= -0.0335
\end{aligned}$$

$$
\begin{aligned}
Var(z_i) &= \frac{3}{8} * \frac{5}{8} \\
&= \frac{15}{64} \\
Sd(z_i) &= \sqrt{\frac{15}{64}} \\
corr(z_i, z_j) &= \frac{\frac{3}{8}(\frac{2}{7} - \frac{3}{8})}{\frac{15}{64}} \\
&= -0.1429
\end{aligned}
$$

f.)

```
# Expected value with replacement
sum(as.numeric(attr(table(Tsums), "dimnames")$Tsums) * (table(Tsums) / length(Tsums)))
```

```
## [1] 74.625
```

```
# Or you can just do this
mean(Tsums)
```

```
## [1] 74.625
```

```
var(Tsums)
```

```
## [1] 331.726
```

```
# For without replacement
mean(TsumsNoRep)
```

```
## [1] 74.625
```

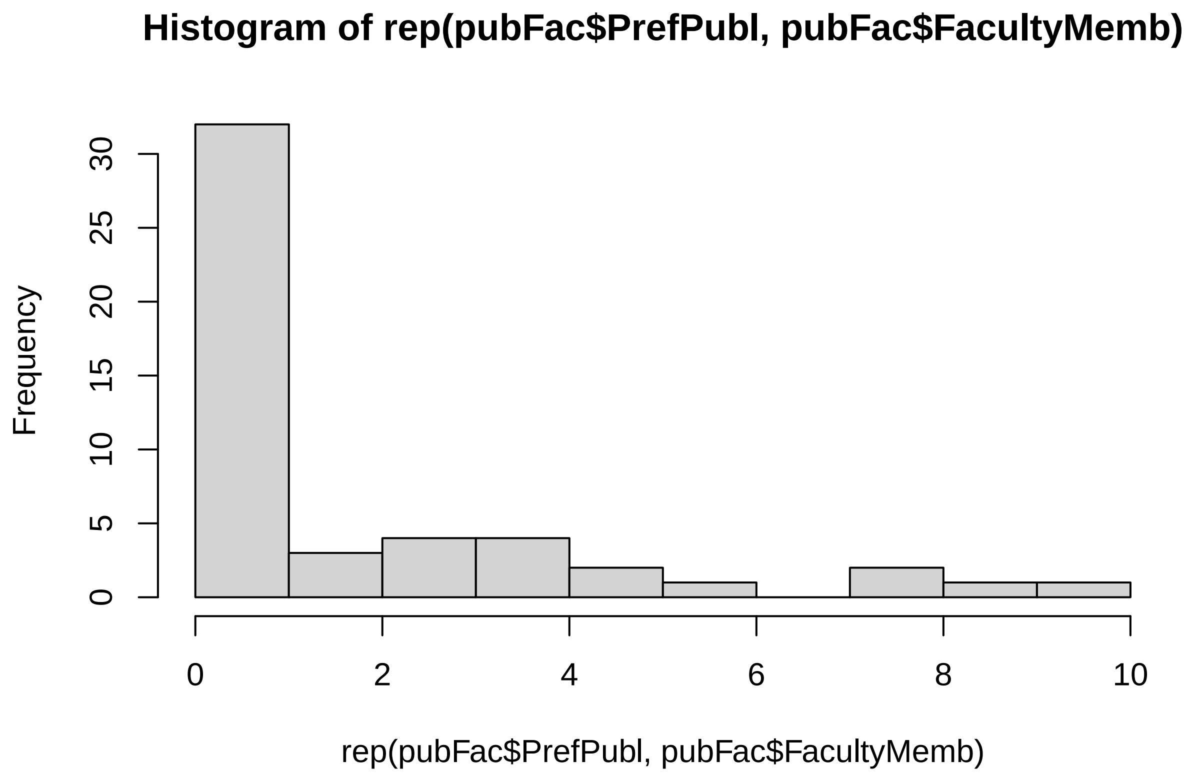
```
var(TsumsNoRep)
```

```
## [1] 240.7841
```

Problem 3

a.)

```
pubFac <- data.frame("PrefPubl" = rep(0:10),
                     "FacultyMemb" = c(28, 4, 3, 4, 4, 2, 1 , 0, 2, 1, 1))
hist(rep(pubFac$PrefPubl,pubFac$FacultyMemb), breaks = 10)
```



We can see that the data is highly right skewed, where 28 faculty members had 0 refereed publications.

b.)

```
long_dat <- rep(pubFac$PrefPubl,pubFac$FacultyMemb)
mean(long_dat)
```

```
## [1] 1.78
```

```
sqrt((var(long_dat) / 50) * (1 - 50/807))
```

```
## [1] 0.3674151
```

c.)

No, given the skewness of the data, we would need a much larger sample for any justification of normality.

d.)

2.19 gives the formula $SE(\hat{p}) = \sqrt{(1 - \frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}}$.

```r
phat <- 28 / 50

sePhat <-
  sqrt(((807 - 50) / (807 - 1)) * ((phat) * (1 - phat) / (50-1)))

# 95% CI
c(phat - 1.96 * sePhat, phat + 1.96 * sePhat)
```

```
## [1] 0.4253027 0.6946973
```

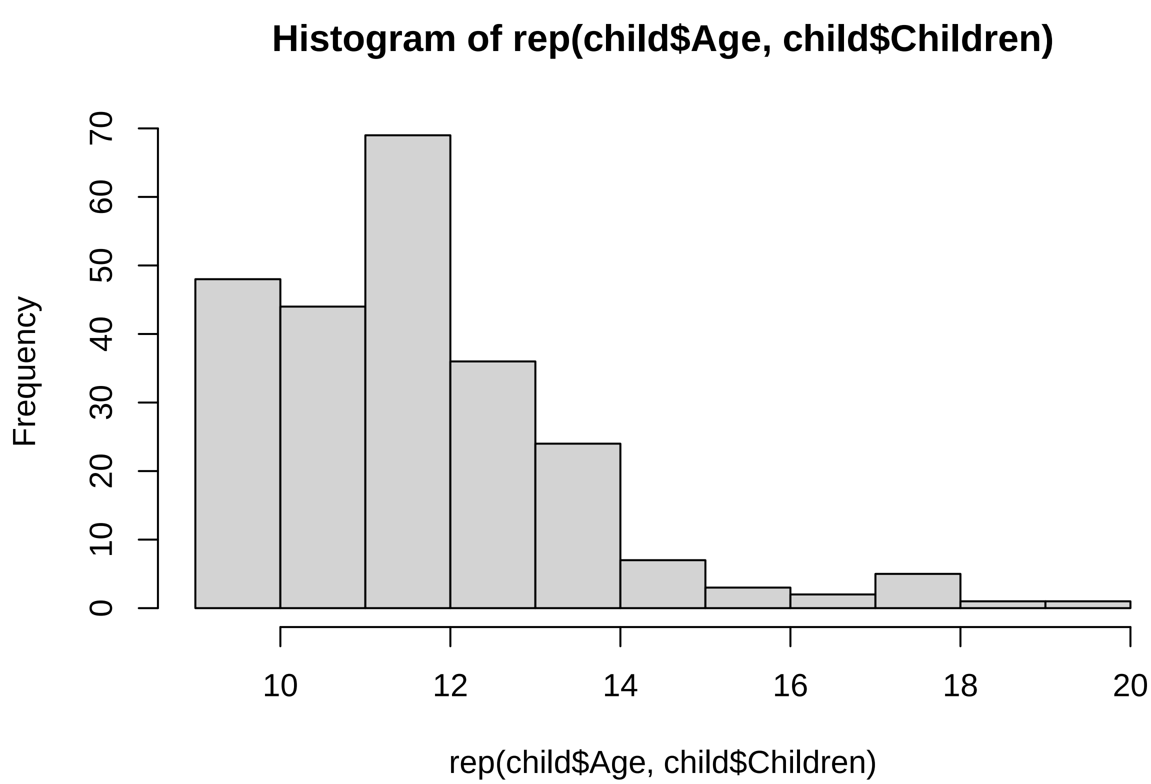
Problem 4

Problem 11, page 63 2nd ed

a.)

```r
child <- data.frame("Age" = rep(9:20),
                     "Children" = c(13, 35, 44, 69, 36, 24, 7 , 3, 2, 5, 1, 1))

hist(rep(child$Age,child$Children), breaks = 10)
```



The shape is not normally distributed, there is noticeable right skew in the data. However, since this data is not as heavily skewed as it could be and our sample size is large, we can deduce the sample mean would be approximately normal.

b.)

```
# Mean
mean.age <- mean(rep(child$Age,child$Children))
mean.age
```

```
## [1] 12.07917
```

```
#SE (ignore FPC, since we do not know population size)
se.age <- sqrt((var(rep(child$Age,child$Children)) / 240))
se.age
```

```
## [1] 0.1242478
```

```
# 95% CI for mean
c(mean.age - 1.96 * se.age, mean.age + 1.96 * se.age)
```

```
## [1] 11.83564 12.32269
```

c.)

```
# From page 47 (pdf 60) 2nd ed
n_desire <- (1.96)^2 * var(rep(child$Age,child$Children)) / (0.5^2)
ceiling(n_desire)
```

```
## [1] 57
```

Problem 5

Problem 16, page 65 in 2nd ed

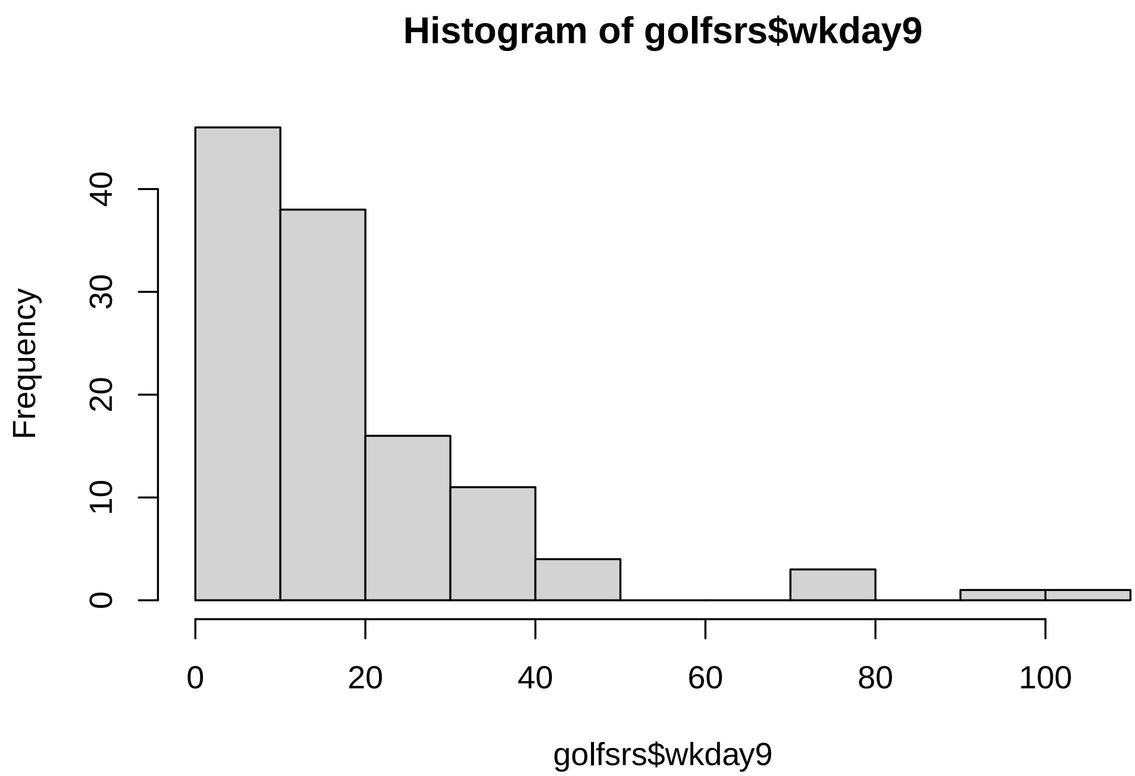
a.)

```
golfsrs <- read.csv("~/Github/MStats/STA144/Homework/HW1/golfsrs.csv")
head(golfsrs)
```

```
##      RN state holes type yearblt wkday18 wkday9 wkend18 wkend9 backtee rating
## 1  5491    RI    18 priv    1923      25     25      35     25    6453   71.8
## 2 10276    VT    18 semi    1972      40     24      45     24    6549   71.1
```

```
## 3  6025    MN     9  pub    1939      NA      10      NA      10    3058   69.2
## 4  9739    GA    18 semi    1991      37      37      45      37    6766   72.2
## 5  3463    CA    18  pub    1970      17      10      20      10    6706   71.4
## 6  5883    MN    18  pub    1996      16      12      18      12    7002   73.5
##   par cart18 cart9 caddy pro
## 1  69   15.0   7.5     y   y
## 2  72   30.0  18.0     n   y
## 3  35   16.5  11.0     n   n
## 4  72    0.0   0.0     n   y
## 5  72   22.0  15.0     n   y
## 6  72   10.0   7.0     n   y
```

```
hist(golfsrs$wkday9)
```



The data is right strongly right skewed with tails.

b.)

```
mean(golfsrs$wkday9)
```

```
## [1] 20.15333
```

```r
sqrt(var(golfsrs$wkday9) / 120 * (1 - 120 / 14938))
```

```
## [1] 1.629866
```

Problem 6

Problem 22 2nd ed

a.)

We need to show that

$$CV(\hat{p}) = \sqrt{\frac{N-n}{N-1}\frac{1-p}{np}}$$

From the definition in 2.13:

$$\begin{aligned}
CV(\bar{y}) &= \sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}\bar{y}_U} \\
CV(\hat{p}) &= \sqrt{\left(1-\frac{n}{N}\right)}\frac{\sqrt{\left(\frac{N}{N-1}\right)p(1-p)}}{\sqrt{n}p}\text{(We substitue for } \hat{p}\text{)} \\
&= \sqrt{\left(1-\frac{n}{N}\right)\left(\frac{N}{N-1}\right)\frac{p(1-p)}{np^2}} \\
&= \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{N}{N-1}\right)\frac{(1-p)}{np}} \\
&= \sqrt{\left(\frac{N-n}{N-1}\right)\frac{(1-p)}{np}}
\end{aligned}$$

If the sample size $n = 1$, we would have the $CV(\hat{p}) = \sqrt{\frac{(1-p)}{p}}$.

$$
\begin{aligned}
n = 1 &= \frac{z^2_{\alpha/2} S^2}{(r\bar{y}_U)^2 + \frac{z^2_{\alpha/2} S^2}{N}} \\
&= \frac{z^2_{\alpha/2} \frac{N}{N-1} p(1-p)}{(rp)^2 + \frac{z^2_{\alpha/2} p(1-p)}{N-1}} \\
&= \frac{z^2_{\alpha/2} \frac{N}{N-1} \frac{p(1-p)}{p^2}}{\frac{(rp)^2}{p^2} + \frac{z^2_{\alpha/2} p(1-p)}{(N-1)p^2}} \\
&= \frac{z^2_{\alpha/2} \frac{N}{N-1} CV^2(\hat{p})}{\frac{(rp)^2}{p^2} + \frac{z^2_{\alpha/2} CV^2(\hat{p})}{(N-1)}} \\
\frac{(rp)^2}{p^2} + \frac{z^2_{\alpha/2} CV^2(\hat{p})}{(N-1)} &= z^2_{\alpha/2} \frac{N}{N-1} CV^2(\hat{p}) \\
\frac{(rp)^2}{p^2} &= z^2_{\alpha/2} \frac{N}{N-1} CV^2(\hat{p}) - \frac{z^2_{\alpha/2} CV^2(\hat{p})}{(N-1)} \\
\frac{(rp)^2}{p^2} &= \frac{z^2_{\alpha/2} CV^2(\hat{p})}{N-1} (N-1) \\
r^2 &= z^2_{\alpha/2} CV^2(\hat{p}) \\
CV(\hat{p}) &= \frac{r}{z_{\alpha/2}}
\end{aligned}
$$

b.)

Below is a data frame wit the neccessary sample sizes for the fixed and relative margin of error for each corresponding value of p. When using a fixed margin of error, the values of n are small for small values of p. However, when using a relative margin of error, we need an extremely large sample size. By making the necessary MOE relative to the value of p, the true margin of error is much smaller.

```r
p.vec <- c(0.001, 0.005, 0.01, 0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95, 0.99, 0.995, 0.999)
# Fixed MoE
FMOE <- sapply(p.vec, function(p) {1.96^2*p*(1-p) / 0.03^2})

# Relative MOE
RMOE <- sapply(p.vec, function(p) {1.96^2*p*(1-p) / (0.03*p)^2})

MOE.df <- data.frame("p" = p.vec, "Fixed"= FMOE, "Relative" = RMOE)
MOE.df
```

```
##        p       Fixed     Relative
## 1  0.001    4.264176 4.264176e+06
## 2  0.005   21.235511 8.494204e+05
## 3  0.010   42.257600 4.225760e+05
## 4  0.050  202.751111 8.110044e+04
## 5  0.100  384.160000 3.841600e+04
## 6  0.300  896.373333 9.959704e+03
## 7  0.500 1067.111111 4.268444e+03
## 8  0.700  896.373333 1.829333e+03
## 9  0.900  384.160000 4.742716e+02
## 10 0.950  202.751111 2.246550e+02
```

```
## 11 0.990   42.257600 4.311560e+01
## 12 0.995   21.235511 2.144947e+01
## 13 0.999    4.264176 4.272717e+00
```