ST117 E2

Homework Lab Group 003 Pod E

This submission was created by:

1. Name and WARWICK ID: QINLING SI 5614637 Question A
2. Name and WARWICK ID: TOM O'CONNELL 5628105 Question A
3. Name and WARWICK ID: ZHIJIAN LIN 5655296 Question B
4. Name and WARWICK ID: DANIEL GUO 5645242 Question B

Question A

For X, Y,U, and V be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ the covariance and the correlation are defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

1.Theory

1.(a) Compute Cov(X, a) and show that Cov(aX + b, Y) = a Cov(X, Y).

according to the definition:

$$Cov(X,a) = \mathbb{E}[(X - \mathbb{E}[X])(a - \mathbb{E}[a])]$$

since a is a constant,then

$$\mathbb{E}[a] = a, (a - E[a]) = a - a = 0$$

Therefore:

$$Cov(X,a) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot 0] = \mathbb{E}[0] = 0$$

proof:

according to the definition:

$$Cov(aX + b, Y) = \mathbb{E}[((aX + b) - \mathbb{E}[aX + b]) \cdot (Y - \mathbb{E}[Y])]$$

since $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$

$$LHS = \mathbb{E}[((aX + b) - \mathbb{E}[aX + b]) \cdot (Y - \mathbb{E}[Y])$$

$$LHS = \mathbb{E}[(aX + b - a\mathbb{E}[X] - b)(Y - \mathbb{E}[Y])] = \mathbb{E}[(aX - a\mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$$LHS = \mathbb{E}[a(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = a\mathbb{E}[(X - \mathbb{E}[x])(Y - \mathbb{E}[Y])]$$

$$LHS = aCov(X, Y) = RHS$$

Therefore,$Cov(aX + b, Y) = aCov(X, Y)$

end of the proof

1.(b)Show that Cov(aX + bY,U) = a Cov(X,U) + b Cov(Y,U).

proof:

according to the definition:

$$Cov(aX + bY, U) = \mathbb{E}[((aX + bY) - \mathbb{E}[aX + bY]) \cdot (U - \mathbb{E}[U])]$$

since $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$

then

$$Left\ Hand\ Side = \mathbb{E}[((aX + bY) - \mathbb{E}[aX + bY]) \cdot (U - \mathbb{E}[U])] = \mathbb{E}[(aX + bY - a\mathbb{E}[X] - b\mathbb{E}[Y])(U - \mathbb{E}[U])]$$

$$= \mathbb{E}[((aX - a\mathbb{E}[X]) + (bY - b\mathbb{E}[Y])) \cdot (U - \mathbb{E}[U])] = \mathbb{E}[a(X - \mathbb{E}[X])(U - \mathbb{E}[U])] + \mathbb{E}[b(Y - \mathbb{E}[Y])(U - \mathbb{E}[U])]$$

$$= a\mathbb{E}[(X - \mathbb{E}[X])(U - \mathbb{E}[U])] + b\mathbb{E}[(Y - \mathbb{E}[Y])(U - \mathbb{E}[U])] = aCov(X, U) + bCov(Y, U) = Right\ Hand\ Side$$

Therefore,$Cov(aX + bY, U) = aCov(X, U) + bCov(Y, U)$

end of the proof

1.(c)Derive a similar formula for Cov(X, aU + bV) without repeating a version of the above calculation but by instead naming and using a general property

according to 1(b),we have $Cov(aX + bY, U) = aCov(X, U) + bCov(Y, U)$

since covariance has the property of symmetry,then

$$Cov(X, aU + bV) = Cov(aU + bV, X)$$

Therefore, according to 1(b):$Cov(X, aU + bV) = aCov(U, X) + bCov(V, X)$

1.(d)Show that the covariance of two independent random variables is 0 but that the converse statement is not generally true. (Note that for the second part a counterexample is enough.)

if X and Y are independent, then

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

so

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X] \cdot \mathbb{E}[Y]]$$

$$= \mathbb{E}[X \cdot Y] - 2\mathbb{E}[Y] \cdot \mathbb{E}[X] + \mathbb{E}[X] \cdot \mathbb{E}[Y] = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = 0$$

therefore, covariance of two independent random variables is 0

converse statement: If Cov(X,Y)=0, then X and Y are independent

counterexample:

let X follows the standard normal distribution and $Y = X^2$

then

$$\mathbb{E}[X] = 1, Var(X) = 0$$

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = Var(x) - (\mathbb{E}[X])^2 = 1$$

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X^3] = 0$$

so $Cov(X,Y) = 0 - 1 * 0 = 0$

but X,Y are not independent

Therefore,the converse statement is not generally true

1.(e)Let X1, X2, . . . , Xn be random variables on a probability space. Derive a formula that expresses the variance of their sum as the sum of their variance plus a suitable sum of covariances.

since we have $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

then

$$Var(X_1, X_2, \cdots, X_n) = Var(\sum_{i=1}^{n} X_i) = Cov(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i) = \sum_{j=1}^{n}\sum_{i=1}^{n} Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{1 \le i < j \le n}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n} Cov(X_i) + 2\sum_{1 \le i < j \le n}^{n} Cov(X_i, X_j)$$

Therefore, $Var(X_1, X_2, \cdots, X_n) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Cov(X_i) + 2\sum_{1 \le i < j \le n}^{n} Cov(X_i, X_j)$

1.(f) We need to show that $-1 \leq \rho(X,Y) \leq 1$

proof:

since we have

$$\rho(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

and according to Cauchy-Schwarz inquality:

$$|\mathrm{Cov}(X,Y)| \leq \sqrt{Var(X) \cdot Var(Y)}$$

so

$$|\mathrm{Cov}(X,Y)| \leq SD(X) \cdot SD(Y)$$

then

$$-SD(X) \cdot SD(Y) \leq \mathrm{Cov}(X,Y) \leq SD(X) \cdot SD(Y)$$

divide $SD(X) \cdot SD(Y)$ on both sides:

$$-1 \leq \frac{Cov(X,Y)}{SD(X)SD(Y)} \leq 1$$

therefore $-1 \leq \rho(X,Y) \leq 1$

end of the proof

2. Examples

2.(a)In n rolls of a fair die let X and Y be the numbers of 1's and 2's observed, respectively. Calculate Cov(X, Y) and $\rho(X,Y)$

since

$X \sim binomial(n, p_1)$ whiere $p_1 = \frac{1}{6}$

$Y \sim binomial(n, p_2)$ whiere $p_2 = \frac{1}{6}$

then,

$$\mathbb{E}[X] = np_1 = \frac{1}{6}n$$

$$\mathbb{E}[Y] = np_2 = \frac{1}{6}n$$

$$Var(X) = np_1(1-p_1) = \frac{1}{6}n \cdot \frac{5}{6} = \frac{5}{36}n$$

$$Var(Y) = np_2(1-p_2) = \frac{1}{6}n \cdot \frac{5}{6} = \frac{5}{36}n$$

$$\mathbb{E}[XY] = p_1 \cdot n \cdot p_2 \cdot (n-1) = \frac{1}{36}(n^2 - n)$$

then,

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = \frac{1}{36}n^2 - \frac{1}{36}n - \frac{1}{36}n^2 = -\frac{1}{36}n$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-\frac{1}{36}n}{\sqrt{\frac{5}{36}n \cdot \frac{5}{36}n}} = \frac{-\frac{1}{36}n}{\frac{5}{36}n} = -\frac{1}{5}$$

Therefore, $Cov(X,Y) = -\frac{1}{36}n$ and $\rho(X,Y) = -\frac{1}{5}$

2.(b)Let X and Y be independent and uniformly distributed on [0, 1], Z = max(X, Y), and W = min(X, Y). Calculate Cov(Z, W) and $\rho(Z,W)$

since $X, Y \sim Uniform(o,1)$ and they are independent

so,$f_{Z,W}(z,w) = 2$ when $0 \leq w \leq z \leq 1$

then,

$$f_Z(z) = 2z \ where(0 \leq z \leq 1)$$

$$f_W(w) = 2(1-z) \ where(0 \leq w \leq 1)$$

$$\mathbb{E}[Z] = \int_0^1 (z \cdot 2z)\, dz = 2\int_0^1 (z^2)\, dz = \left[\frac{2}{3}z^3\right]_0^1 = \frac{2}{3}$$

$$\mathbb{E}[W] = \int_0^1 (2w \cdot (1-w))\, dw = 2\int_0^1 (w - w^2)\, dw = \frac{1}{3}$$

$$\mathbb{E}[ZW] = \int_0^1 \int_0^z (2zw)\, dw\, dz$$

where $\int_0^z (2zw)\, dw = \left[zw^2\right]_0^z = z^3$

so

$$\mathbb{E}[ZW] = \int_0^1 (z^3)\, dz = \left[\frac{1}{4}z^4\right]_0^1 = \frac{1}{4}$$

$$Cov(Z,W) = \mathbb{E}[ZW] - \mathbb{E}[Z]\mathbb{E}[W] = \frac{1}{4} - \frac{1}{3} * \frac{2}{3} = \frac{1}{36}$$

$$\mathbb{E}[Z^2] = \int_0^1 (z^2 \cdot 2z)\, dz = \left[\frac{2}{4}z^3\right]_0^1 = \frac{1}{2}$$

$$\mathbb{E}[W^2] = \int_0^1 (w^2 \cdot 2(1-w))\, w = \left[\frac{2}{3}w^2 - \frac{2}{4}w^3\right]_0^1 = \frac{1}{6}$$

$$Var(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$$

$$Var(W) = \mathbb{E}[W^2] - (\mathbb{E}[W])^2 = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{18}$$

$$\rho(Z,W) = \frac{Cov(Z,W)}{SD(Z)SD(W)} = \frac{Cov(Z,W)}{\sqrt{Var(Z)Var(W)}} = \frac{\frac{1}{36}}{\sqrt{\frac{1}{18} \cdot \frac{1}{18}}} = \frac{\frac{1}{36}}{\frac{1}{18}} = \frac{1}{2}$$

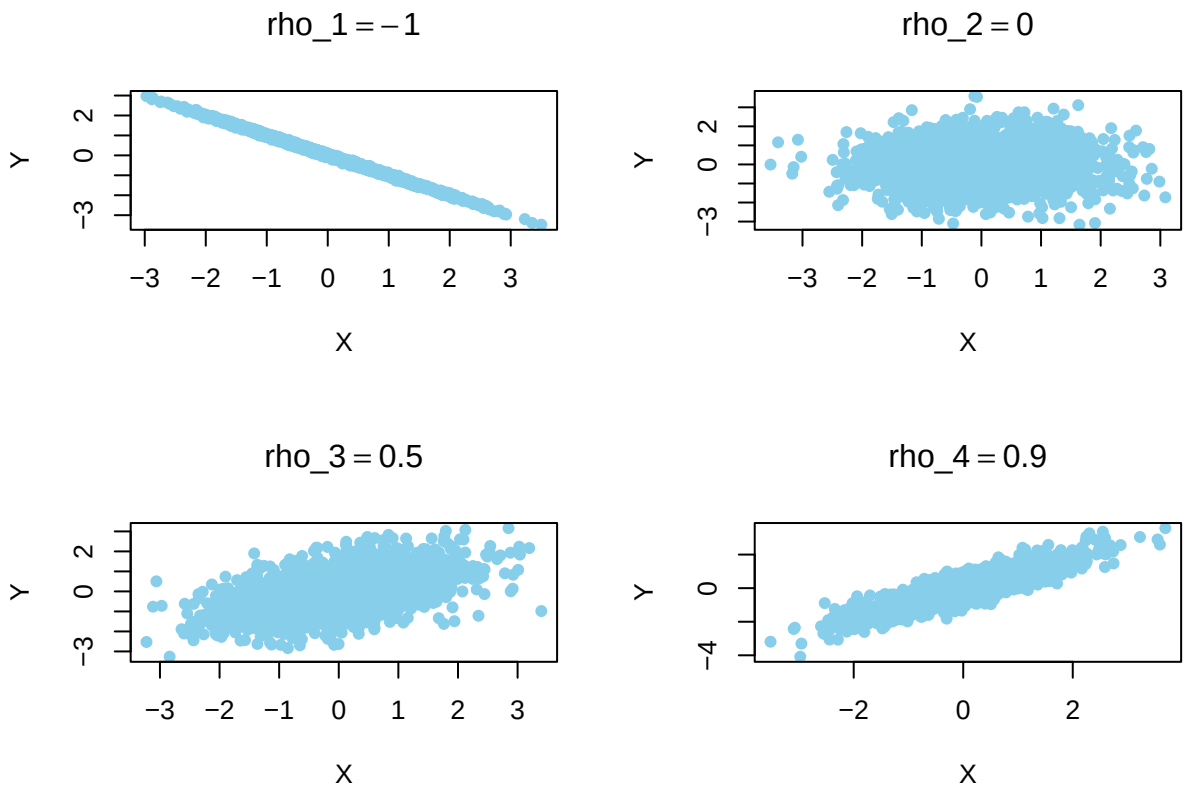
Therefore, $Cov(Z,W) = \frac{1}{36}$ and $\rho(Z,W) = \frac{1}{2}$

3. Simulations

3.(a)Use R to create four different point clouds (scatter plots) showing correlations of approximately (i.e., within ± 0.05) of $\rho_1 = -1$, $\rho_2 = 0$, $\rho_3 = 0.5$, and $\rho = 0.9$, respectively.

```r
par(mfrow = c(2, 2))
set.seed(5114)
n<-2000
X1 <- rnorm(n)
Y1 <- -X1 + rnorm(n, sd = 0.05)
X2 <- rnorm(n)
Y2 <- rnorm(n)
X3 <- rnorm(n)
Y3 <- 0.5 * X3 + sqrt(1 - 0.5^2) * rnorm(n)
X4 <- rnorm(n)
Y4 <- 0.9 * X4 + sqrt(1 - 0.9^2) * rnorm(n)

plot(X1, Y1, main = expression(rho_1 == -1), xlab = "X", ylab = "Y", col = "skyblue", pch = 16)
plot(X2, Y2, main = expression(rho_2 == 0), xlab = "X", ylab = "Y", col = "skyblue", pch = 16)
plot(X3, Y3, main = expression(rho_3 == 0.5), xlab = "X", ylab = "Y", col = "skyblue", pch = 16)
plot(X4, Y4, main = expression(rho_4 == 0.9), xlab = "X", ylab = "Y", col = "skyblue", pch = 16)
```

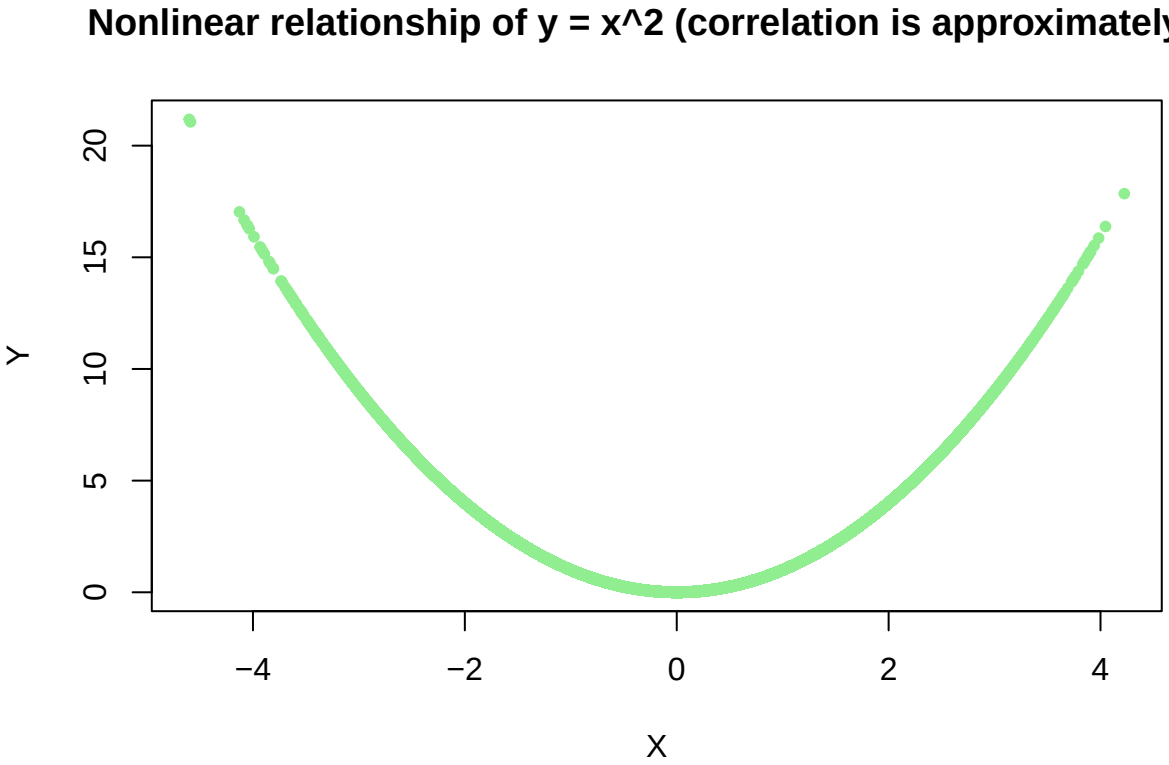


3.(b) Use R to create a point cloud that corresponds to a polynomial relationship between the X and Y but the correlation is approximately 0. Why is this possible even though X and Y are not independent?

```
set.seed(5114)
n<-200000
x <- rnorm(n)
y <- x^2 #Because the counterexample I used earlier is the same example
correlation<-cor(x,y)
cat("the coorelation is:",round(correlation, 4),"\n")
```

```
## the coorelation is: -9e-04
```

```
plot(x,y,
     main = "Nonlinear relationship of y = x^2 (correlation is approximately 0)",
     xlab = "X",
     ylab = "Y",
     col = "lightgreen",
     pch=20)
```



Why is this possible even though X and Y are not independent?

answer:This phenomenon occurs because Pearson correlation measures only linear relationships, while X and Y have a nonlinear relationship.

Question B

1. Theory

1.(a)

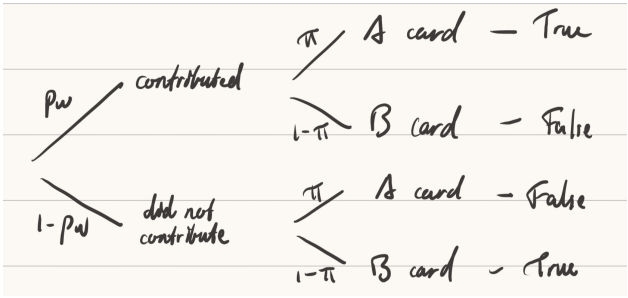


Figure 1: Tree Diagram of the survey

p_w is the proportion of ST117 students who contributed to their pod's E1 submission.

$1 - p_w$ is the proportion of which who did not contribute.

Let π be the proportion of A cards in the box.

Let $1 - \pi$ be the proportion of B cards in the box.

Students who contributed will answer "True" if A is pulled and "False" if B is pulled.

Students who did not contributed will answer "True" if B is pulled and "False" if A is pulled.

The tree diagram is shown in figure 1

Therefore, the probability r that a student says "True" for their card is:

$$r = p_w\pi + (1 - p_w) \times (1 - \pi)$$

1.(b)

Rearrange the formula derived from question 1:

$$r = p_w\pi + (1 - p_w) \times (1 - \pi)$$

Expand the brackets:

$$r = p_w\pi + 1 - p_w - \pi + p_w\pi$$

Rearrange:

$$r = 2p_w\pi - p_w + (1 - \pi)$$

Therefore:

$$r = (2\pi - 1)p_w + (1 - \pi)$$

1.(c)

Rearrange the formula derived from question 2:

$$r = (2\pi - 1)p_w + (1 - \pi)$$

$$r - (1 - \pi) = (2\pi - 1)p_w$$

$$p_w = \frac{r + \pi - 1}{2\pi - 1}$$

Under the condition when $\pi \neq \frac{1}{2}$, as it would be undefined.

Also, assuming that the fraction of $\frac{r+\pi-1}{2\pi-1}$ is positive.

1.(d)

We have $p_w = \frac{r+\pi-1}{2\pi-1}$

Since r is the probability that a randomly sampled student confirms the sentence in their envelope and R is the proportion of the sample taken by the SSLC rep that actually confirm the sentence in their envelope.

Therefore, R is the unbiased estimator of r and by substituting R into p_w, we have:

$$\hat{p_w} = \frac{R + \pi - 1}{2\pi - 1}$$

And:

$$\hat{p_w} = \frac{\hat{R} + \pi - 1}{2\pi - 1}$$

To check unbiased-ness, we need $\mathbb{E}[\hat{p_w}] = p_w$:

$$\hat{p_w} = \frac{\hat{R} + \pi - 1}{2\pi - 1}$$

$$\mathbb{E}[\hat{p_w}] = \mathbb{E}[\frac{R+\pi-1}{2\pi-1}] = \frac{\mathbb{E}[R]+\pi-1}{2\pi-1}$$

Since R is an unbiased estimator of r, $\mathbb{E}[R] = r$:

$$\mathbb{E}[\hat{p_w}] = \frac{r+\pi-1}{2\pi-1} = p_w$$

Thus, proving that it is unbiased, so that the unbiased estimator is:

$$\hat{p_w} = \frac{\hat{R}+\pi-1}{2\pi-1}$$

1.(e)

Since R is the proportion of students in the sample by the SSLC rep, who confirm the sentence on their card, it can be computed as:

$$R = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_i is the Bernoulli Random Variable which returns 1 if the student ith student confirms the card, and returns 0 otherwise.

Each X_i has the probability of r of returning 1, so by the expected value and variance formula:

$\mathbb{E}[X_i] = r \times 1 + (1-r) \times 0 = r$ and $Var(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = r - r^2 = r(1-r)$

Therefore:

$$Var(R) = \frac{Var(X_i)}{n} = \frac{r(1-r)}{n}$$

and

$$Var(\hat{p_w}) = Var(\frac{R+\pi-1}{2\pi-1}) = Var(\frac{R}{2\pi-1} + \frac{\pi-1}{2\pi-1})$$

By the formula that $Var(aX+b) = a^2 Var(X)$

$$Var(\hat{p_w}) = (\frac{1}{2\pi-1})^2 Var(R) = (\frac{1}{2\pi-1})^2 \frac{r(1-r)}{n} = \frac{r(1-r)}{n(2\pi-1)^2}$$

From part (b) we have $r = (2\pi-1)p_w + (1-\pi)$, substitute:

$$Var(\hat{p_w}) = \frac{((2\pi-1)p_w + (1-\pi))(1-((2\pi-1)p_w + (1-\pi)))}{n(2\pi-1)^2}$$

Followed by simplifying:

$$Var(\hat{p_w}) = \frac{((2\pi-1)p_w + (1-\pi))(1-(2\pi-1)p_w - (1-\pi))}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{(2\pi-1)p_w - (2\pi-1)^2 p_w^2 - (2\pi-1)p_w(1-\pi) + (1-\pi) - (1-\pi)(2\pi-1)p_w - (1-\pi)^2}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{(2\pi-1)p_w - (2\pi-1)^2 p_w^2 - 2(2\pi-1)p_w(1-\pi) + (1-\pi)(1-1+\pi)}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{(2\pi-1)p_w(1-(2\pi-1)p_w - 2(1-\pi)) + \pi(1-\pi)}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{(2\pi-1)p_w((2\pi-1)-(2\pi-1)p_w)) + \pi(1-\pi)}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{(2\pi-1)^2 p_w(1-p_w)) + \pi(1-\pi)}{n(2\pi-1)^2}$$

$$Var(\hat{p_w}) = \frac{p_w(1-p_w))}{n} + \frac{\pi(1-\pi)}{n(2\pi-1)^2}$$

Check the dependence on π:

First term $\frac{p_w(1-p_w))}{n}$ is independent on π,

Second term $\frac{\pi(1-\pi)}{n(2\pi-1)^2}$ is dependent on π.

Let $n = 10$ and $p_w = 0.5$

When $\pi = 0.1$, $Var(\hat{p_w}) = 0.03906$. When $\pi = 0.9$, $Var(\hat{p_w}) = 0.03906$

When $\pi = 0.3$, $Var(\hat{p_w}) = 0.15625$. When $\pi = 0.7$, $Var(\hat{p_w}) = 0.15625$

When $\pi = 0.45$, $Var(\hat{p_w}) = 2.5$. When $\pi = 0.55$, $Var(\hat{p_w}) = 2.5$

From this function while keeping n and p_w constant, $\lim_{\pi \to 0.5} Var(\hat{p_w}) = \infty$

Therefore, the variance increases sharply as $\pi \to 0.5$, from both sides, due to the second term of $Var(\hat{p_w})$, $\frac{\pi(1-\pi)}{n(2\pi-1)^2}$. The closer π is to 0 or 1, the smaller the variance, $Var(\hat{p_w})$ will be.

1.(f)

This survey assumes the honesty from the cohort: students may lie about their contribution with the fear of being identified despite said to be anonymous.

By the assumption that more student who did not contribute lie about their contribution, it could underestimate the proportion of p_w. By the assumption that more student who contributed lie about their contribution, it could overestimate the proportion of p_w.

The survey also assumes that there is no misinterpretation: student may confuse or misread the card's statement.

This would lead to higher uncertainties as: if contributed students misunderstood the card, it would underestimate the proportion of p_w; if non-contributors misunderstood the card, it would overestimate the proportion of p_w.

The estimator, $\hat{p_w} = \frac{\hat{R}+\pi-1}{2\pi-1}$, relied on truthful response. If the sampled data is not truthful, there would be bias in $\hat{p_w}$, such that the estimate would be unreliable

2. Simulations

```
#Used write.csv(grade_book, "grade_book.csv", row.names = FALSE)
#to save the dataframe as a csv on Exercise Set 1 Rmd file.
grade_book <- read.csv(
  "/Users/danielguo/Desktop/University/Year 1/ST117/Exercise Set1 1/grade_book.csv")
#Read the dataframe csv file

set.seed(1234567) # Set seed for consistency

# Set parameters
n_students <- 272 # number of students
# assuming one student dropped out as mentioned in exercise 1
```

```r
p_w <- 0.6 # Proportion of students who contributed
pi <- 0.9 # Proportion of Card A
grade_book$contribution <- rbinom(n_students, 1, p_w)  # 1 if contributed, 0 if not

head(grade_book) #print the first few rows of the dataframe
```

```
##   Pod_ID First_Name Last_Name Lab_Group Participated_A0 Participated_A1
## 1      1          A     Aliyya   Borunda           1              No             Yes
## 2      1          A      Jerod el-Hashim           1             Yes             Yes
## 3      1          A    Timothy   Sanchez           1             Yes             Yes
## 4      1          B     Morgan  Arellano           1              No             Yes
## 5      1          B  Jacqueline        Ky           1              No              No
## 6      1          B    Liberty    Torrez           1             Yes             Yes
##   Marks_A1 Participated_A2 Marks_A2 Marks_Q1 Marks_Q2 Homework_Pods E1_Score
## 1        1             Yes        1     0.60     0.79             A     0.62
## 2        1             Yes        1     0.51     0.85             A     0.62
## 3        1             Yes        1     0.72     0.80             A     0.62
## 4        1             Yes        1     0.32     0.94             B     0.79
## 5        0             Yes        1     0.49     0.92             B     0.79
## 6        1             Yes        1     0.45     0.36             B     0.79
##   E2_Score E3_Score Report_Pod Log_Participation Passed_Logs Mark_Draft
## 1     0.60     0.55      Pod 9                 5        0.83          1
## 2     0.60     0.55     Pod 13                 5        0.83          1
## 3     0.60     0.55     Pod 49                 4        0.67          1
## 4     0.53     0.98     Pod 10                 6        1.00          1
## 5     0.53     0.98      Pod 2                 6        1.00          1
## 6     0.53     0.98     Pod 21                 5        0.83          1
##   WR_Marks    MM contribution
## 1     0.79 71.85            1
## 2     0.49 63.00            0
## 3     0.87 75.20            0
## 4     0.84 83.80            1
## 5     0.76 79.10            0
## 6     0.89 76.95            1
```

```r
# Function to run the survey simulation
simulate_survey <- function(n) {
  pw_estimates <- numeric(100)  # Store estimates
  for (i in 1:100) {
    # Random sample of students
    sample_students <- grade_book[sample(1:n_students, n), ]

    # Students randomly pick a card (Type A with probability pi, Type B 1-pi)
    cards <- rbinom(n, 1, pi)  # 1 is card A, 0 is card B

    # Students response, depending on the card
    responses <- ifelse((sample_students$contribution == 1 & cards == 1) |
                        (sample_students$contribution == 0 & cards == 0), 1, 0)

    # proportion R
    R_hat <- mean(responses)

    # caclate estimated pw_hat with the formula from 1d
    pw_hat <- (R_hat - (1 - pi)) / (2 * pi - 1)
```

```r
    # Store result
    pw_estimates[i] <- pw_hat
  }

  return(pw_estimates)
}

# Run simulations for both sample sizes
pw_estimates_10 <- simulate_survey(10)
pw_estimates_40 <- simulate_survey(40)

# Empirical mean and variance
mean_10 <- mean(pw_estimates_10)
var_10 <- var(pw_estimates_10)
mean_40 <- mean(pw_estimates_40)
var_40 <- var(pw_estimates_40)

# The theoretical values:
theoretical_r <- (2 * pi - 1) * p_w + (1 - pi)
theoretical_var_10 <- (theoretical_r * (1 - theoretical_r)) / 10 / (2 * pi - 1)^2
theoretical_var_40 <- (theoretical_r * (1 - theoretical_r)) / 40 / (2 * pi - 1)^2

# Print the empirical and theoretical results
cat("Empirical Mean for n=10:", mean_10)
```

```
## Empirical Mean for n=10: 0.6425
```

```r
cat("Empirical Variance for n=10:", var_10)
```

```
## Empirical Variance for n=10: 0.04072601
```

```r
cat("Empirical Mean for n=40:", mean_40)
```

```
## Empirical Mean for n=40: 0.6346875
```

```r
cat("Empirical Variance for n=40:", var_40)
```

```
## Empirical Variance for n=40: 0.008161596
```

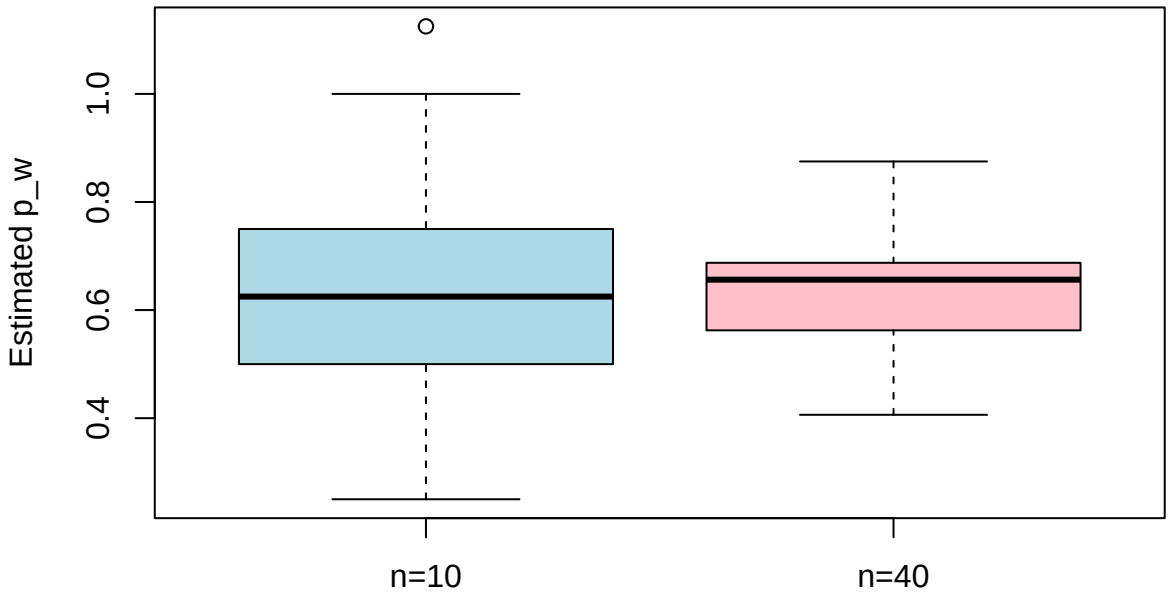
```r
cat("Theoretical Variance for n=10:", theoretical_var_10)
```

```
## Theoretical Variance for n=10: 0.0380625
```

```r
cat("Theoretical Variance for n=40:", theoretical_var_40)
```

```
## Theoretical Variance for n=40: 0.009515625
```

```r
boxplot(
  pw_estimates_10, pw_estimates_40,
  names = c("n=10", "n=40"),
  col = c("lightblue", "pink"),
  main = "Distribution of pw estimates for sample sizes of 10 and 40",
  ylab = "Estimated p_w"
)
```



Overall, the empirical data is similar to the theoretcal.

When $n = 10$, the estimates fluctuate more (wider spread in the boxplot).

When $n = 40$, the estimates are more concentrated around p_w.

$\hat{(}p_w)$ has higher variability for $n = 10$ than $n = 40$.

This aligns with: variance decreases as sample size increases. (law of large numbers)