317 pset 1

In collaboration with Matthew Jacob, Elliot Britton, Megan McQueen, Kim Dao, Abbyy Steckel and Alan George

Valerie Nguyen

2/23/2021

Problem 1

(a)

i.

This is FALSE. We can see the counter example as follows:

Let $A_1, A_2, A_3 \sim N(0, 1)$ and they are independent random variables so that

$$X = A_1 + A_2$$

$$Y = A_1 + A_3$$

$$Z = A_1 - A_3$$

From this, since A_1, A_2, A_3 are all independent, the covariances between them will be 0, and we can write the covariances between X, Y, Z as follows:

$$Cov(X, Y) = Cov(A_1 + A_2, A_1 + A_3) = Cov(A_1, A_1) + Cov(A_2, A_1) + Cov(A_1, A_3) + Cov(A_2, A_3)$$

$$= var(A_1)$$

Similarly we have:

$$Cov(X, Z) = Cov(A_1 + A_2, A_1 - A_3) = var(A_1)$$

$$Cov(Z, Y) = Cov(A_1 - A_3, A_1 + A_3) = Cov(A_1, A_1) - Cov(A_3, A_3) + Cov(A_1, A_3) - Cov(A_3, A_1)$$

$$= var(A_1) - var(A_3)$$

- If $Cov(X, Y), Cov(X, Z) > 0$, then that means $var(A_1) > 0$.
- If $Cov(Z, Y) \geq 0$ then $var(A_1) \geq var(A_3)$, but $var(A_3)$ can always be greater than $var(A_1)$ which would make $Cov(Z, Y) < 0$, making the given statement untrue.

ii.

We can use a property of the correlation between two rv's $-1 \leq \rho(X, Y) \leq 1$:

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

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

$$-\sigma_X\sigma_Y \leq Cov(X,Y) \leq \sigma_X\sigma_Y$$

Therefore $-\sigma_X\sigma_Y \leq Cov(X,Y)$. #### iii. This is FALSE, we can consider the counter example below:

```
X <- c(-2,-4,-6)
Y <- c(4,6,8)

(cov_xy <- cov(X,Y))
```

```
## [1] -4
```

```
(cor_xy <- cor(X,Y))
```

```
## [1] -1
```

```
cor_xy <= cov_xy
```

```
## [1] FALSE
```

(b)

i.

We have $X \sim Unif[0,1]$:

- $\mathbb{E}(X) = \frac{1}{2}$
- $f_X(x) = 1$ for $0 \leq x \leq 1$

Since $Y \sim Unif[0,X]$, we have:

- $f_{Y|X}(y|x) = \frac{1}{x}$ for $0 \leq y \leq x$

The joint density of X & Y:

- $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = \frac{1}{x} * 1 = \frac{1}{x}$ for $0 \leq x \leq 1; 0 \leq y \leq x$

We can find the pdf of Y: * $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_y^1 \frac{1}{x}dx = ln(x)|_y^1 = -ln(y)$ for $0 \leq y \leq 1$.

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_0^1 y(-ln(y))dy$$

$$\mathbb{E}(Y) = \frac{1}{4}$$

ii. Find $Cov(X,Y)$ * First we find $\mathbb{E}(XY) = \int\int xyf_{X,Y}(x,y)dxdy$:

$$\mathbb{E}(XY) = \int_0^1 \int_0^x xy\frac{1}{x}dydx$$

$$= \int_0^1 \frac{y^2}{2}|_0^x dx$$

$$= \int_0^1 \frac{x^2}{2} dx$$

$$= \frac{1}{2}$$

We have

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{2} - \frac{1}{2} * \frac{1}{4} = \frac{1}{24}$$

iii. Find $Cor(X, Y)$ We have $Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y))}}$.

Find $Var(X)$:

- $\mathbb{E}(X^2) = \int_0^1 x^2 * 1 dx = \frac{x^3}{3}|_0^1 = \frac{1}{3}$
- $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{3} - \frac{1}{2}^2 = \frac{1}{12}$

Find $Var(Y)$: * $\mathbb{E}(Y^2) = \int_0^1 y^2 * (-ln(y)) dy = \frac{y^3}{9}|_0^1 = \frac{1}{9}$

- $Var(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \frac{1}{9} - \frac{1}{4}^2 = \frac{7}{144}$

We know that $Cov(X, Y) = \frac{1}{24}$. Therefore:

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y))}}$$

$$Cor(X, Y) = \frac{\frac{1}{24}}{\sqrt{\frac{1}{12} * \frac{7}{144}}} = \sqrt{\frac{3}{7}}$$

Problem 2

(a) Show that τ is unbiased

```
y_ic <- c(0,-1,2,4,0,1,0,1)
y_it <- c(3,1,3,4,2,5,3,0)

(ATE <- mean(y_it) - mean(y_ic))
```

```
## [1] 1.75
```

We find here that ATE is 1.75. This value is also equal to the difference between the expected value of the treated subset of Y_i and the expected value of the control subset of Y_i.

$$\hat{\tau} = E(treated) - E(control)$$

Now, when we take a look at each of these expected values, we see that the mean of 3 randomly chosen treated units should be an unbiasted estimator of the treated group:

$$E(treated) = E(\frac{1}{3}\sum_{i=1}^{3}(treated\ unit))$$

Similarly, we would have:

$$E(control) = E(\frac{1}{5}\sum_{i=1}^{5}(control\ unit))$$

Thereby $\hat{\tau} = E(treated) - E(control)$ should be an unbiased estimate of $\bar{\tau}$

$$E(\hat{\tau}) = E(\frac{1}{3}\sum_{i=1}^{3}(treated\ unit)) - E(\frac{1}{5}\sum_{i=1}^{5}(control\ unit))$$

```
set.seed(571)
nit <- 10000
hat_tau <- rep(0, nit)

for (i in 1:nit) {
  it <- sample(c(1:8), 3, replace = F)
  hat_tau[i] <- mean(y_it[it]) - mean(y_ic[-it])
}
mean(hat_tau)
```

```
## [1] 1.76238
```

We find that the average of 10000 simulations of $\hat{\tau}$ is 1.762, really close to the ATE, therefore $\hat{\tau}$ is unbiased.

(b) Single run

```
t <- sample(c(1:8), 3, replace = F)
(hat_se <- sqrt(var(y_it[t])/3 + var(y_ic[-t])/(8-3)))
```

```
## [1] 0.6879922
```

(c) Estimate SE of $\hat{\tau}$

```
#standard error of simulated hat_tau
sd(hat_tau)
```

```
## [1] 0.9978877
```

```
#using the formula from (b), run 10000 simulations to calculate the actual value of the
 SE of hat_tau and average over these simulations

nit <- 10000
hat_tau_se <- rep(0, nit)

for (i in 1:nit) {
  it <- sample(c(1:8), 3, replace = F)
  hat_tau_se[i] <- sqrt(var(y_it[it])/3 + var(y_ic[-it])/(8-3))
}
mean(hat_tau_se)
```

```
## [1] 1.115617
```

From the simulation, I see that the formula generates a higher value (by 0.2) on average for the SE of $\hat{\tau}$ than what we found from $\hat{\tau}$ from previous simulation. It seems that the standard formula might be a little biased as it produces a lower SE and makes the result seem more precise.

(d)

```
tau_m1 <- median(y_it - y_ic)
tau_m2 <- median(y_it) - median(y_ic)
c(tau_m1, tau_m2)
```

```
## [1] 2.0 2.5
```

We see that $median(Y_{it} - Y_{ic}) \neq median(Y_{it}) - median(Y_{ic})$. Now we run 10000 simulations to see if our estimator could be an unbiased estimator of either median equations:

```
nit <- 10000
tau_m <- rep(0, nit)

for (i in 1:nit) {
  it <- sample(c(1:8), 3, replace = F)
  tau_m[i] <- median(y_it[it]) - median(y_ic[-it])
}
mean(tau_m)
```

```
## [1] 2.2142
```

From the simulations, we see that the average value for τ_M is 2.2293 which is right in between the result of two median equations (which are 2 and 2.5)

Problem 3

(a)

10000 simulations, generate potential outcomes, approximate the variance of $\hat{\tau}$

```
library(MASS)
set.seed(517)
N <- 12
m <- 4
y <- mvrnorm(N, c(2,0), matrix(c(1, 0.25, 0.25, 2), ncol=2))
?mvrnorm()
y_control <- y[ ,2]
y_treated <- y[ ,1]

nit <- 10000
path3a1 <- rep(0, nit)
for (i in 1:nit) {
  it <- sample(c(1:12), 4, replace = F)
  path3a1[i] <- mean(y_treated[it]) - mean(y_control[-it])
}

var(path3a1)
```

```
## [1] 0.4960607
```

Now we compare this value we got with the formula used in class

```
tau <- mean(y_treated) - mean(y_control)
Stc_sqrd <- sum((y_treated - y_control - tau)^2) / (length(y_treated) - 1)

nit <- 10000
path3a2 <- rep(0, nit)
for (i in 1:nit) {
  it <- sample(c(1:12), 4, replace = F)
  path3a2[i] <- var(y_treated[it])/m + var(y_control[-it])/(N-m) - Stc_sqrd/N
}

mean(path3a2)
```

```
## [1] 0.501702
```

The variance of $\hat{\tau}$ approximated through the Monte-Carlo simulation is very close to the variance estimate using the formula in class. This means that our approximation does correspond to the formula result.

(c)

Three sensible estimators for $\mathbb{V}(\hat{\tau})$

Let N = # units, N_t = # treated units, N_c = # control units, and W = vector of treatment assignments.

First we have $\mathbb{V}_{neyman}(\hat{\tau}) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t}$. This estimator is unbiased when the treatment effect $\tau_i = Y_i(1) - Y_i(0)$ hoolds constant for all i.

Next, we have $\mathbb{V}_{\rho_{ct}}(\hat{\tau}) = s_c^2 * \frac{N_t}{N*N_c} + s_t^2 * \frac{N_c}{N*N_t} + \rho_{ct} * s_c^2 * s_t^2 * \frac{2}{N}$. This estimator is unbiased if we know the correlation between $Y_i(0)$ and $Y_i(1)$.

Finallly, we have $\mathbb{V}_{const}(\hat{\tau}) = s^2 * (\frac{1}{N_c} + \frac{1}{N_t})$. This estimator is unbiased when the treatment effect $\tau_i = Y_i(1) - Y_i(0)$ hoolds constant for all i. In addition, $s^2 = s_c^2 = s_t^2$.

Problem 4

(a)

Since our assumption is that SUTVA holds. As a result, the outcome for one student would not have an effect on the outcome for another student. There would be a total of $2 * N$ potential outcomes, with each unit having 2 potential outcomes under either $t = 0$ or $t = 1$.

The effect of the virus on a student i:

$$\tau_i = (Y_i|T_i = 1) - (Y_i|T_i = 0)$$

Using the formula, we have the ATE:

$$\hat{\tau} = \frac{1}{m}\sum_{i=1}^{m} Y_i T_i + \frac{1}{N-m}\sum_{i=1}^{N-m} Y_i T_i$$

(b)

The fact that the virus is contagious does not necessarily violate SUTVA because it does not necessarily change the effect of the virus on any given student i. Although the virus spreads between students, the fact that one student has the virus does not change the expected effect of the virus on another student since this expected effect has already factored in the possibility oof catching the virus for this one student. In other words, if we look at the formula for effect of the virus on a student i, $(Y_i|T_i = 1) - (Y_i|T_i = 0)$ is not influenced by any given $T_j = 1$.

(c)

If more than half the class get the virus, the whole class's outcomes are affected. This definitely violates SUTVA as the potential outcome of a unit (one student) is immpacted by the potential outcomes of other units.

For the number of potential outcomes, we have two different scenarios to account for. If less than half the class gets the virus, SUTVA holds, so the total potential outcomes is still 2N. However, in the scenario that more than half get the virus, we would have an extra number of potential outcomes. This can be written down as follows:

- If N is even:

$$Total(potential\ outcomes) = 2N + \sum_{k=N/2}^{N} \binom{N}{k}$$

- If N is odd:

$$Total(potential\ outcomes) = 2N + \sum_{k=(N+1)/2}^{N} \binom{N}{k}$$

(d)

- Contagion can be a threat to SUTVA since the assignment might not be random when it comes to the effect of virus. One example is that immunocompromised students have higher likelihood of catching the virus in the first place, which violates the probability assignment property of SUTVA. This can make ATE biased.
- The conditions in (c) can pose a threat to the experiment if $m \geq \frac{N}{2}$, i.e. more than half the class was treated, all the students will suffer and their potential outcomes are influenced by the treatment. This violates SUTVA.

(e)

Contagion is a threat to SUTVA since if treatment is considered injecting the virus and getting infected through contagion is irrelevant. This is due to the fact that un-injected students who catch the virus from injected students will still count as the control group which poses a threat to the experiment design.

(f)

When only 6 students get infected, we're using the estimator ATE for the 6 treated units. Assuming that the matched students were randomly chosen, the estimatoor is still $\hat{\tau} = Y_i(1) - Y_i(0)$.

When only 6 students are healthy, we're using the estimator Average Treatment Effect for 6 control units.

Problem 5

(a)

```
villages <- read.csv(file = "villages.csv")
nas <- matrix(0, ncol = 2, nrow = 6)
for (i in 1:6) {
  nas[i,1] <- names(villages)[i+2]
  nas[i,2] <- length(which(is.na(villages[,i+2])))
}
nas
```

```
##      [,1]                    [,2]
## [1,] "pct.missing"           "90"
## [2,] "share.total.unskilled" "73"
## [3,] "head.edu"              "5"
## [4,] "mosques"               "2"
## [5,] "pct.poor"              "7"
## [6,] "total.budget"          "2"
```

The variables in the dataset do have missing values as reported above. However, the missing outcomes for some units should not raise any flags as long as units are selected at random for treatment. However, similar to the AIDS-treatment-group example in Imbens and Rubin's book, if these missing outcomes correspond to a pre-treatment covariate that we do not account for in the experiment, unit-level randomization might be influenced and treatment effect might be affected by these pre-treatment covariates.

(b)

```
#villages <- na.omit(villages)
cis <- matrix(NA, ncol = 3, nrow = 6)

for (i in 1:6) {
  y_treated <- villages[villages$treat.invite == 1, i+2]
  y_control <- villages[villages$treat.invite == 0, i+2]
  t_test <- t.test(y_treated, y_control)
  cis[i, 1] <- names(villages)[i+2]
  cis[i, 2] <- round(t_test$conf.int[1], 2)
  cis[i, 3] <- round(t_test$conf.int[2], 2)
}
cis
```

```
##      [,1]                    [,2]    [,3]
## [1,] "pct.missing"           "-0.09" "0.04"
## [2,] "share.total.unskilled" "-0.03" "0.02"
## [3,] "head.edu"              "-0.54" "0.41"
## [4,] "mosques"               "-0.21" "0.08"
## [5,] "pct.poor"              "-0.03" "0.05"
## [6,] "total.budget"          "-9.97" "6.45"
```

All of these confidence intervals include 0, which means randomization did indeed succeed in producing comparble treatment and control groups, despite the missing outcomes.

(c)

```
y_treated <- villages[villages$treat.invite == 1, "pct.missing"]
y_control <- villages[villages$treat.invite == 0, "pct.missing"]
(t_test <- t.test(y_treated, y_control))
```

```
##
##  Welch Two Sample t-test
##
## data:  y_treated and y_control
## t = -0.70443, df = 334.89, p-value = 0.4817
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08778496  0.04149022
## sample estimates:
## mean of x mean of y
## 0.2289582 0.2521056
```

We see that the 95% confidence interval is (-0.090, 0.040), while the p-value for this estimate is 0.4515.

(d)

We model the data by this linear model $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$

```
mod <- lm(villages$pct.missing ~ villages$treat.invite)

#summary on beta1
summary(mod)$coef[2,]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.02314737  0.03321720 -0.69684898  0.48623819
```

The p-value from t-test is 0.4515, while the p-value from the regression model is 0.4563

They are different because the variance assumptions are different in the two models. The linear model seem to assume the same variances, while the t-test doesn't.

(e)

```
newmod <- lm(villages$pct.missing ~ (villages$treat.invite + villages$share.total.unskil
led + villages$head.edu + villages$mosques + villages$pct.poor + villages$total.budget))

summary(newmod)
```

```
##
## Call:
## lm(formula = villages$pct.missing ~ (villages$treat.invite +
##     villages$share.total.unskilled + villages$head.edu + villages$mosques +
##     villages$pct.poor + villages$total.budget))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2705 -0.2143 -0.0190  0.1852  1.4386
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.3796376  0.0889853   4.266 2.41e-05 ***
## villages$treat.invite         -0.0262205  0.0331812  -0.790  0.42980
## villages$share.total.unskilled  0.0748300  0.1291188   0.580  0.56250
## villages$head.edu             -0.0053682  0.0058124  -0.924  0.35619
## villages$mosques              -0.0499372  0.0192213  -2.598  0.00967 **
## villages$pct.poor             -0.1203165  0.0749802  -1.605  0.10925
## villages$total.budget          0.0005061  0.0002908   1.740  0.08252 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3415 on 465 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.0301, Adjusted R-squared:  0.01758
## F-statistic: 2.405 on 6 and 465 DF,  p-value: 0.02675
```

We see here with the new regression that the treatment coef is still 0.429, which is not significant. The t-test also reports insignificant p-val of treatment effect. Therefore I fail to reject the null-hypothesis, and treatment had no effect.

(f)

It seems doubtful to me that the treatment had an effect. The t-test was not close to significant while the regression estimate was, because this could have been caused by adding on multiple covariates to the regression just as we did in (e), thus driving down the p-value of the model. In the context of an experiment, adding multiple covariates could be unreliable (and dishonest) since the covariates could potentially have a correlation with our causal relationship of interest.