ARE 213 Problem Set 1A

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 09/25/2020

Section 1

1. *Before getting started with the data work, first consider the table from Snow (1855) reproduced in the lecture notes ("Snow's Table IX"). The table reports only means.

(a) Develop an approximate 95% confidence interval for "Deaths per 10,000 Houses" for Southwark and Vaxhall customers. Develop another 95% CI for the same quantity for Lambeth. Do the confidence intervals overlap?

Note that that we're estimating p for a binomial distribution since deaths per 10,000 houses is the same as deaths per person (of course, scaled by persons per 10,000 households). Are we really dealing with a binomial distribution? Probably not, but it might not be a bad approximation if we think contaminated water is distributed randomly across space-time (so one person's probability exposure and subsequent death is the same and independent of another person's). Also, not everyone is equally susceptible to the virus (some have a higher p than others), but our estimate of p can be interpreted as an average p.

There are various ways to construct a confidence interval for an estimated binomial distribution. We use three different methods, all of which provide very similar estimates. The confidence intervals do not overlap.

```
# Southwark and Vauxhall
binom.confint(1263, 40046, method=c("asymptotic", "wilson", "agresti-coull"), type="central")
```

```
##          method    x     n       mean      lower      upper
## 1 agresti-coull 1263 40046 0.03153873 0.02987085 0.03329648
## 2    asymptotic 1263 40046 0.03153873 0.02982701 0.03325045
## 3        wilson 1263 40046 0.03153873 0.02987144 0.03329589
```

```
# Lambeth
binom.confint(98, 26107, method=c("asymptotic", "wilson", "agresti-coull"), type="central")
```

```
##          method  x     n        mean       lower       upper
## 1 agresti-coull 98 26107 0.003753783 0.003077893 0.004575688
## 2    asymptotic 98 26107 0.003753783 0.003011981 0.004495584
## 3        wilson 98 26107 0.003753783 0.003081460 0.004572122
```

(b) Discuss either formally or intuitively the critical assumption that underlies your confidence intervals. Give a 2 or 3 sentence quote from Snow's description (reproduced in Freedman (1991)) that supports this assumption.

To be confident that it is the choice of water company that is causing the difference in p and not some other factor, we need to be sure that there are not systematic differences between those who get their water from Southwark and Vauxhall and those who get it from Lambeth. John Snow argues that the two groups of people are comparable: "both rich and poor, both large houses and small" etc. In that case, we are reasonably certain that the difference in water company is what causes the difference in mortality risk.

Section 2

We now move to some analysis of real data. The data portions of Problem Sets 1a and 1b are based heavily on the paper Almond, Chay, and Lee (2005), and problem sets from Ken Chay and John DiNardo based on some of the data used in the paper. The goal of this assignment is to examine the research question: what is the causal effect of maternal smoking during pregnancy on infant birthweight and other infant health outcomes. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match. The data in Stata format can be downloaded from the bCourses website. There should be 48 variables in the data and, after you are finished with the cleaning steps desribed below, 114,610 observations.

The data here are "real" and quite imperfect, which will help simulate the unpleasantness of real world data work. Unlike the real world where you will confront this bleak situation largely alone, I will provide you with some hints for working your way through the raw data. You can download part of the codebook for the data to help you figure out the relevant variables.

2. The first order of business is to go through the code book, decide on the relevant variables, and process the data. This involves several steps:

(a) Fix missing values. In the data set several variables take on a value of, say, 9999 if missing. We have already checked for missing observations for about 2/3 of the variables. The remaining variables need to be checked and are the last 15 in the variables list (i.e. from 'cardiac' to 'wgain'). Refer to the codebook for missing value codes. Produce an analysis data set that drops any observations with missing values.

```
# According to the codebook, for the following medical risk factor variables, 8 corresponds to
# "Factor not on certificate" and 9 corresponds to "Factor not classifiable": cardiac, lung,
# diabetes, herpes, chyper, phyper, pre4000, preterm

med_risk_factors <- c('cardiac', 'lung', 'diabetes', 'herpes', 'chyper', 'phyper', 'pre4000', 'preterm')

for (var in med_risk_factors){
  mom_dt[var] <- replace(mom_dt[var], which(mom_dt[var] == 8, arr.ind = TRUE), NA)
  mom_dt[var] <- replace(mom_dt[var], which(mom_dt[var] == 9, arr.ind = TRUE), NA)
}

# Below, arr.ind = TRUE returns the indices at which the row equals a certain value

# According to the codebook, for tobacco, 9 corresponds to "Unknown or not stated"
mom_dt$tobacco <- replace(mom_dt$tobacco, which(mom_dt$tobacco == 9, arr.ind = TRUE), NA)

# According to the codebook, for cigar, 99 corresponds to "Unknown or not stated"
mom_dt$cigar <- replace(mom_dt$cigar, which(mom_dt$cigar == 99, arr.ind = TRUE), NA)

# According to the codebook, for cigar6, 6 corresponds to "Unknown or not stated"
mom_dt$cigar6 <- replace(mom_dt$cigar6, which(mom_dt$cigar6 == 6, arr.ind = TRUE), NA)

# According to the codebook, for alcohol, 9 corresponds to "Unknown or not stated"
mom_dt$alcohol <- replace(mom_dt$alcohol, which(mom_dt$alcohol == 9, arr.ind = TRUE), NA)

# According to the codebook, for drink, 99 corresponds to "Unknown or not stated"
mom_dt$drink <- replace(mom_dt$drink, which(mom_dt$drink == 99, arr.ind = TRUE), NA)

# According to the codebook, for drink5, 5 corresponds to "Unknown or not stated"
mom_dt$drink5 <- replace(mom_dt$drink5, which(mom_dt$drink5 == 5, arr.ind = TRUE), NA)

# According to the codebook, for wgain (assuming that's wtgain in codebook),
# 99 corresponds to "Unknown or not stated"
mom_dt$wgain <- replace(mom_dt$wgain, which(mom_dt$wgain == 99, arr.ind = TRUE), NA)

# Get rows with any missing value into one DT; remove all the rows with any missing value for main DT
```

```r
miss_dt <- mom_dt[!(complete.cases(mom_dt)),]
mom_dt <- na.omit(mom_dt)

# Now mom_dt contains 114,610 observations instead of the original 120,461
```

(b) If this were a real research project you would want to consider other approaches to missing data besides termination with extreme prejudice. What observations do you have to drop because of missing data? Might this affect your results? Do the data appear to be missing completely at random? How might you assess whether the data appear to be missing at random?

We know from the last problem set that if the data are missing at random then dropping them should not affect our results of the effect of smoking on birth weight. However, if the missing data is correlated with the treatment (smoking) or the outcome (birth weight) then it could bias our results. From the table below, it does appear that there are differences in the missing and nonmissing data. For example, the mothers in the missing data are younger, less educated, less likely to be married, have more previous children, received less prenatal care, have a shorter time since the last birth, have a lower gestation period, and have lower birth weight. As discussed in the last problem set, we could formally assess whether the data is missing at random by regressing an indicator for the missing variable on the treatment.

In the table, we calculate the mean and standard deviation of the variables for the missing and nonmissing data. We also calculate the t-stat:

$$t = \frac{Mean_m - Mean_{nm}}{\frac{Var_m}{n_m} + \frac{Var_{nm}}{n_{nm}}}$$

the degrees of freedom:

$$\frac{\frac{Var_m}{n_m} + \frac{Var_{nm}}{n_{nm}}}{\frac{(\frac{Var_m}{n_m})^2}{n_m - 1} + \frac{(\frac{Var_{nm}}{n_{nm}})^2}{n_{nm} - 1}}$$

and the p-value which is the probability that the t-statistic achieves the value calculated.

```r
# Compare missing to non missing
compare_dt <- data.table("Variable" = c("Mother age", "Mother educ", "Marital status", "Prenatal adequacy",
                                        "Number living child", "# dead/living child",
                                        "Total birth/terminations", "Birth order", "Month prenatal began",
                                        "Number prenatal visits", "Time since last birth", "Father age",
                                        "Father educ", "Gestation", "Child sex",
                                        "Birth weight", "Number born", "One min Apgar", "Five min Apgar",
                                        "Anemia", "Cardiac disease", "Lung disease", "Diabetes",
                                        "Herpes", "Chron. hypertension", "Preg. hypertension",
                                        "Previous heavy birth", "Previous preterm", "Tobacco use",
                                        "Number cigarettes", "Alcohol use", "Number drinks", "Weight gain"),
                         "Miss means" = round(as.numeric(lapply(miss_dt[, c(6, 9:18, 20, 22,
                                                                            25:30, 33:43, 45:46, 48)],
                                                                            mean, na.rm=TRUE)), 3),
                         "Miss sd" = round(as.numeric(lapply(miss_dt[, c(6, 9:18, 20, 22, 25:30, 33:43,
                                                                         45:46, 48)],
                                                                         sd, na.rm=TRUE)), 3),
                         "Nonmiss means" = round(as.numeric(lapply(mom_dt[, c(6, 9:18, 20, 22, 25:30,
                                                                              33:43, 45:46, 48)],
                                                                              mean)), 3),
                         "Nonmiss sd" = round(as.numeric(lapply(mom_dt[, c(6, 9:18, 20, 22,
                                                                           25:30, 33:43, 45:46, 48)],
                                                                           sd)), 3))

# add difference and t stat
compare_dt[, "Difference" := `Miss means` - `Nonmiss means`]
```

```
compare_dt[, "t-stat" := Difference / sqrt((`Miss sd`)^2/nrow(miss_dt)) + ((`Nonmiss sd`)^2/nrow(mom_dt)))]
compare_dt[, "DF" := round(((((`Miss sd`)^2/nrow(miss_dt)) + ((`Nonmiss sd`)^2/nrow(mom_dt)))^2) /
             ((((`Miss sd`)^2/nrow(miss_dt))^2/(nrow(miss_dt) - 1)) +
             (((`Nonmiss sd`)^2/nrow(mom_dt))^2/(nrow(mom_dt) - 1))), 0)]
compare_dt[, "p-Value" := round(2*pt(-abs(`t-stat`), DF), 3)]
```

```
print(xtable(compare_dt, caption = 'Difference in Means Missing v Nonmissing', digits = 2),
      include.rownames = FALSE, size = "small", comment = FALSE)
```

Variable	Miss means	Miss sd	Nonmiss means	Nonmiss sd	Difference	t-stat	DF	p-Value
Mother age	27.05	5.97	27.76	5.70	-0.71	-8.84	6406.00	0.00
Mother educ	12.51	2.26	13.21	2.27	-0.70	-23.16	6466.00	0.00
Marital status	1.44	0.50	1.25	0.43	0.19	27.99	6316.00	0.00
Prenatal adequacy	1.63	0.79	1.30	0.55	0.33	31.59	6137.00	0.00
Number living child	1.24	1.43	0.97	1.15	0.27	14.17	6241.00	0.00
# dead/living child	2.27	1.47	1.99	1.17	0.28	14.40	6235.00	0.00
Total birth/terminations	2.81	1.87	2.42	1.52	0.39	15.72	6250.00	0.00
Birth order	2.78	1.74	2.41	1.46	0.37	15.95	6275.00	0.00
Month prenatal began	2.80	1.92	2.50	1.33	0.30	11.81	6138.00	0.00
Number prenatal visits	9.32	4.90	11.15	3.52	-1.84	-28.31	6163.00	0.00
Time since last birth	315.97	355.26	350.41	362.32	-34.44	-7.23	6487.00	0.00
Father age	29.61	7.04	30.06	6.41	-0.46	-4.84	6355.00	0.00
Father educ	12.67	2.29	13.28	2.33	-0.60	-19.62	6480.00	0.00
Gestation	38.53	3.42	39.15	2.44	-0.62	-13.77	6159.00	0.00
Child sex	1.49	0.50	1.49	0.50	0.00	0.00	6462.00	1.00
Birth weight	3191.90	716.95	3373.29	585.17	-181.39	-19.03	6254.00	0.00
Number born	1.04	0.21	1.03	0.17	0.01	4.26	6259.00	0.00
One min Apgar	7.91	1.57	8.12	1.26	-0.21	-10.16	6240.00	0.00
Five min Apgar	8.88	1.03	9.01	0.71	-0.13	-9.47	6135.00	0.00
Anemia	1.99	0.12	1.99	0.10	-0.00	-2.55	6278.00	0.01
Cardiac disease	1.99	0.09	1.99	0.08	-0.00	-0.86	6406.00	0.39
Lung disease	1.99	0.10	1.99	0.09	-0.00	-1.56	6327.00	0.12
Diabetes	1.97	0.16	1.97	0.16	0.00	0.00	6462.00	1.00
Herpes	1.99	0.10	1.99	0.08	-0.00	-2.35	6251.00	0.02
Chron. hypertension	1.99	0.10	1.99	0.09	-0.00	-0.77	6340.00	0.44
Preg. hypertension	1.97	0.16	1.97	0.17	0.00	2.32	6559.00	0.02
Previous heavy birth	1.99	0.10	1.99	0.12	0.00	2.18	6690.00	0.03
Previous preterm	1.98	0.15	1.99	0.12	-0.01	-5.35	6201.00	0.00
Tobacco use	1.57	0.49	1.84	0.37	-0.27	-41.46	6181.00	0.00
Number cigarettes	3.94	7.42	1.91	5.30	2.03	20.70	6158.00	0.00
Alcohol use	1.63	0.48	1.99	0.10	-0.36	-56.95	5875.00	0.00
Number drinks	0.16	1.47	0.03	0.62	0.13	6.64	5957.00	0.00
Weight gain	30.79	13.14	30.36	11.88	0.43	2.45	6348.00	0.01

Table 1: Difference in Means Missing v Nonmissing

(c) Produce a summary table describing the final analysis data set.

We create a summary table similar to the table in b, but this time we compare the means of smokers vs non-smokers. To see means/standard deviations for the entire dataset, refer to the nonmissing data columns of part b. We will discuss the differences between the groups in 3b.

```
setDT(mom_dt)
# Summary table: broken out by smoker/non smoker
summary_dt <- data.table("Variable" = c("Mother age", "Mother educ", "Marital status", "Prenatal adequacy",
                                        "Number living child", "# dead/living child",
                                        "Total birth/terminations", "Birth order", "Month prenatal began",
                                        "Number prenatal visits", "Time since last birth", "Father age",
                                        "Father educ", "Gestation", "Child sex",
```

```
                                    "Birth weight", "Number born", "One min Apgar", "Five min Apgar",
                                    "Anemia", "Cardiac disease", "Lung disease", "Diabetes",
                                    "Herpes", "Chron. hypertension", "Preg. hypertension",
                                    "Previous heavy birth", "Previous preterm",
                                    "Number cigarettes", "Alcohol use", "Number drinks", "Weight gain"),
                          "Smoker mean" = round(as.numeric(lapply(mom_dt[tobacco == 1,
                                                                         c(6, 9:18, 20, 22, 25:30, 33:41, 43,
                                                                         45:46, 48)],
                                                                         mean)), 3),
                          "Smoker sd" = round(as.numeric(lapply(mom_dt[tobacco == 1,
                                                                       c(6, 9:18, 20, 22, 25:30,
                                                                       33:41, 43, 45:46, 48)],
                                                                       sd)), 3),
                          "Nonsmoker mean" = round(as.numeric(lapply(mom_dt[tobacco == 2,
                                                                            c(6, 9:18, 20, 22,
                                                                            25:30, 33:41, 43, 45:46, 48)],
                                                                            mean)), 3),
                          "Nonsmoker sd" = round(as.numeric(lapply(mom_dt[tobacco == 2, c(6, 9:18, 20, 22,
                                                                            25:30, 33:41, 43, 45:46, 48)],
                                                                            sd)), 3))

# add difference and t stat
summary_dt[, "Diff" := `Smoker mean` - `Nonsmoker mean`]
summary_dt[, "t-stat" := Diff / sqrt(((`Smoker sd`)^2/nrow(mom_dt[tobacco==1])) +
                                       ((`Nonsmoker sd`)^2/nrow(mom_dt[tobacco==2])))]
summary_dt[, "DF" := round(((((`Smoker sd`)^2/nrow(mom_dt[tobacco==1])) +
                               ((`Nonsmoker sd`)^2/nrow(mom_dt[tobacco==2])))^2) /
             ((((`Smoker sd`)^2/nrow(mom_dt[tobacco==1]))^2/(nrow(mom_dt[tobacco==1]) - 1)) +
             (((`Nonsmoker sd`)^2/nrow(mom_dt[tobacco==2]))^2/(nrow(mom_dt[tobacco==2]) - 1))), 0)]
summary_dt[, "p-Value" := round(2*pt(-abs(`t-stat`), DF), 3)]
```

```
print(xtable(summary_dt, caption = 'Difference in Means Smoker v Nonsmoker', digits = 2),
      include.rownames = FALSE, size = "small", comment = FALSE)
```

Variable	Smoker mean	Smoker sd	Nonsmoker mean	Nonsmoker sd	Diff	t-stat	DF	p-Value
Mother age	26.17	5.61	28.06	5.67	-1.88	-41.57	25844.00	0.00
Mother educ	11.99	1.63	13.44	2.30	-1.46	-102.70	33713.00	0.00
Marital status	1.48	0.50	1.21	0.41	0.27	70.10	23024.00	0.00
Prenatal adequacy	1.41	0.63	1.27	0.53	0.14	27.42	23335.00	0.00
Number living child	1.15	1.22	0.93	1.13	0.22	22.54	24523.00	0.00
# dead/living child	2.18	1.26	1.95	1.15	0.23	23.04	24357.00	0.00
Total birth/terminations	2.74	1.67	2.36	1.48	0.38	29.04	24002.00	0.00
Birth order	2.73	1.60	2.35	1.42	0.38	29.95	24042.00	0.00
Month prenatal began	2.75	1.51	2.45	1.28	0.30	25.13	23486.00	0.00
Number prenatal visits	10.63	3.84	11.25	3.45	-0.63	-20.51	24176.00	0.00
Time since last birth	306.63	349.76	358.71	364.07	-52.08	-18.33	26329.00	0.00
Father age	28.96	6.65	30.27	6.34	-1.31	-24.60	24977.00	0.00
Father educ	12.13	1.67	13.49	2.37	-1.37	-93.98	33870.00	0.00
Gestation	39.05	2.71	39.17	2.39	-0.13	-5.87	23955.00	0.00
Child sex	1.48	0.50	1.49	0.50	-0.00	-0.99	25672.00	0.32
Birth weight	3171.14	572.08	3411.62	579.73	-240.48	-51.98	25884.00	0.00
Number born	1.02	0.15	1.03	0.18	-0.01	-4.78	28778.00	0.00
One min Apgar	8.10	1.27	8.12	1.26	-0.02	-1.67	25574.00	0.10
Five min Apgar	9.01	0.71	9.01	0.71	0.00	0.00	25672.00	1.00
Anemia	1.99	0.12	1.99	0.10	-0.01	-5.48	23242.00	0.00
Cardiac disease	1.99	0.08	1.99	0.08	0.00	1.57	26933.00	0.12
Lung disease	1.99	0.10	1.99	0.08	-0.00	-3.89	23358.00	0.00
Diabetes	1.97	0.16	1.97	0.16	0.00	0.00	25672.00	1.00
Herpes	1.99	0.08	1.99	0.08	-0.00	-1.54	24921.00	0.12
Chron. hypertension	1.99	0.08	1.99	0.09	0.00	1.51	27312.00	0.13
Preg. hypertension	1.98	0.14	1.97	0.18	0.01	10.92	30443.00	0.00
Previous heavy birth	1.99	0.09	1.98	0.12	0.01	10.22	32622.00	0.00
Previous preterm	1.98	0.15	1.99	0.11	-0.01	-10.83	21882.00	0.00
Number cigarettes	11.96	7.47	0.00	0.00	11.96	216.58	18265.00	0.00
Alcohol use	1.97	0.18	2.00	0.07	-0.03	-21.74	19278.00	0.00
Number drinks	0.14	1.44	0.01	0.25	0.12	11.73	18475.00	0.00
Weight gain	29.47	13.45	30.52	11.56	-1.05	-9.92	23646.00	0.00

Table 2: Difference in Means Smoker v Nonsmoker

3. The next part of the assignment is to try to estimate the "causal" effect of maternal smoking during pregnancy on infant birth weight. Let's start out using techniques that are familiar, and think about whether they are likely to work in this context. Answer the following questions.

(a) Compute the mean difference in APGAR scores (both five and one minute versions) as well as birthweight by smoking status.

```
# According to the codebook, omaps is the one minute APGAR score and fmaps is the five minute APGAR score
# Both are a score from 0-10
# dbrwt (assuming that corresponds to dbirwt in codebook) is birthweight in grams
# tobacco is 1: yes, tobacco use during pregnancy and 2: no tobacco use during pregnancy

smoker <- subset(mom_dt, mom_dt$tobacco == 1)
nonsmoker <- subset(mom_dt, mom_dt$tobacco == 2)

# Mean difference in one minute APGAR score by smoking status
mean_diff_1min_apgar <- mean(smoker$omaps) - mean(nonsmoker$omaps)
print(mean_diff_1min_apgar)
```

```
## [1] -0.01743508
```

```
# Mean difference in five minute APGAR score by smoking status
mean_diff_5min_apgar <- mean(smoker$fmaps) - mean(nonsmoker$fmaps)
print(mean_diff_5min_apgar)
```

```
## [1] -0.0001498085
```

```r
# Mean difference in birthweight by smoking status
mean_diff_birthweight <- mean(smoker$dbrwt) - mean(nonsmoker$dbrwt)
print(mean_diff_birthweight)
```

```
## [1] -240.4778
```