Problem Set 1: Suggested Solution

Objectives

- To be familiar with simple mean comparison and causal reasoning.
- To be able to use Stata or R to generate basic summary statistics.

Immigrant Earnings (40 points)

Immigration is one of the most contentious policy debates in the US. Understanding the economic performance of immigrants helps to shape our stance on immigration policies.

A first simple fact about immigration is that on average immigrants earn less than natives. Using the 1-percent 2000 Public Use Microdata Sample of the US census, Table 1 presents the means of immigrants and natives' earnings (The dummy variable `Native` indicates whether a person is native or immigrant). The data refer to employed male workers who are twenty-five to sixty-four years old.

Table 1: Means of Immigrants and Natives' Earnings in 2000

Native	Mean	Std. Dev	Count
0	41158	52425	72203
1	50402	54023	491564

Moreover, immigrants who stay in US longer tend to earn more on average. Table 2 documents immigrants earnings by year of entry to the US.

Table 2: Mean of Immigrants' Earning in 2000, by Year of Entry

yr2us	Mean	Std. Dev.	Count
1970 and before	59420	67231	10399
1971 - 1980	46011	56497	16589
1981 - 1990	37323	46229	23668
1991 - 2000	32819	44261	21547

Questions

1. Earning Differentials between Immigrants and Natives

    1. Provide at least two interpretations to the mean comparisons displayed in Table 1. That is, what are the possible underlying causal mechanisms. (8 points)

- **There are many possibilities. For example, the skill distribution, perhaps measured by the distribution of years of schooling, are different for the two groups. Discriminations against immigrants in the US labor market can be another possibility.**

    2. Suppose you want to test one of your interpretations, what kinds of mean comparisons would you make? Be specific about outcome variables and "treatment" variables. (8 points)

- **To test whether native workers are more educated, simply perform a naive comparison with "treatment" being whether a person is a native, and "outcome" being education attainment.**

2. Immigrant Earnings and Time in the US

    1. Provide at least two interpretations to the mean comparisons displayed in Table 2. That is, what are the possible underlying causal mechanisms. (8 points)

- **Again there are many possible explanations. One explanation can be assimilation. That is, as immigrants spend more time in the US, they gradually pick up labor market skills that enable them to earn more. It's also possible that recent immigrants are more likely to be lower-skilled workers. Another possibility can be that lower earning immigrants gradually choose to leave the US and we only observe higher earning immigrants for earlier periods.**

    2. Suppose you want to test one of your interpretations, what kinds of mean comparisons would you make. Be specific about outcome variables and "treatment" variables. (8 points)

- **To test whether lower earning immigrants are more likely to leave the US, one can perform a naive comparison with "treatment" being an immigrant's income level, and "outcome" being whether an immigrant is leaving the US permanently.**

3. Are there any policy implications of your interpretations of the data? If so, briefly discuss these implications. If not, briefly discuss why your interpretations are not enough to give policy suggestions.

- **Even if we get the correct underlying causal story, it can still be very hard to provide public polices because often we need to specify social objectives. On issues such as immigrations, any posisble policy are likely to create benefit to some groups at a cost to certain other groups. For example, a less strict immigration policy that attracts lower-skilled workers might hurt lower-skilled natives. Thus, if the US social objective is to maximize the welfare of all natives, perhaps a more strict immigration policy is better, but if maximizing the welfare of all human beings is the objective, perhpas a less strict immigration policy is better.**

Selection Bias (20 points)

Based on the National Health Interview Survey 2009, Table 3 presents mean comparison for a health index of insured and uninsured wives. Similarly, Table 4 presents mean comparisons for three characteristic variables.

Table 3: Health of Insured and Uninsured Wives in the NHIS

Outcome	Some HI	No HI	Difference
Health Index	4.02 [0.92]	3.62 [1.01]	0.39 (0.04)

Table 4: Characteristics of Insured and Uninsured Wives in the NHIS

Characteristics	Some HI	No HI	Difference
Education	14.44	11.80	2.64

Characteristics	Some HI	No HI	Difference
			(0.11)
Employed	0.77	0.56	0.21
			(0.02)
Family Income	106,212	46,385	59,828
			(1,406)

1. Is the mean difference in health index statistically significant? How do you reach your decision? (5 points)

- **We're interested in knowing whether the difference in population mean is statistically significant, which means whether it's significantly different from zero. Here is a usual step for testing the difference:**
  - **Set up a null hypothesis: $H_0 : \Delta\mu = 0$. That is, the population "difference in mean", $\Delta\mu$, is assumed to be zero.**
  - **Given the null hypothesis, construct a t-value:**

$$t = \frac{\textbf{ΔHealth Index} - \Delta\mu}{SE(\textbf{ΔHealth Index})} = \frac{\textbf{ΔHealth Index} - 0}{SE(\textbf{ΔHealth Index})}.$$

    You can find the difference in sample average, ΔHealth Index, and standard error of difference in sample average, SE(ΔHealth Index), in the last column of Table 3.
  - **If the t-value is greater than 2, which means an unlikely event happens (based on the standard normal distribution of a large sample t-statistic), we claim the null hypothesis can be rejected. Since the null hypothesis is that the population "difference in mean", $\Delta\mu$, is zero, rejecting the null hypothesis means that the population "difference in mean", $\Delta\mu$, is significantly different from 0.**
- **Based on Table 3, the t-value is $0.39/0.04 = 9.75 > 2$; the null hypothesis can be rejected.**

2. Can we interpret the mean difference in health index as the causal impact of having health insurance? What is the "gap" between the simple mean difference and causal effect? Use potential outcome notations to make your answer concrete. (10 points)

- **No. Mean comparisons based on observed outcomes without a credible research design can hardly be interpretated as causal impacts, which are defined based on potential outcomes. The gap between the simple mean difference and causal effect can be derived as the following:**

$$\begin{aligned}
&E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
=&E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \\
=&E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] + \\
&E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]
\end{aligned}$$

- **In this case, the causal parameter of interest is the average causal effect on the treated, $E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]$, and the selection bias term is $E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$.**

3. Briefly describe some basic facts one can learn from Table 4. (5 points)

- **On average, women who have health insurance are more educated, more likely to be employed, and have a higher level of family income.**

The Rand HIE (40 points)

For this part of the questions, use Stata or R or any software you like. Download the dataset `rand.dta` from Blackboard or Econ300.com. The dataset contains the following variables:

Table 5: Variable Descriptions (`rand.dta`)

Variable Names	Descriptions
any_ins	= 1 if has any health insurance assigned; = 0 otherwise
female	= 1 if female; = 0 otherwise
blackhisp	= 1 if nonwhite; = 0 otherwise
educper	years of education
hosp	hospitalized last year
ghindx	pre-treatment outcome: general health index
cholest	pre-treatment outcome: cholesterol level (mg/dl)
ghindxx	post-treatment outcome: general health index
cholestx	post-treatment outcome: cholesterol level (mg/dl)

1. Generate basic summary statistics such as mean, standard deviation, and total number of observations for variables `female`, `blackhisp`, `educper`, `ghindx`, and `cholest`, separated by the `any_ins` indicator. Your table for each variables should look similar as Table 1 in this problem set. You can also combine all the information in just one table. (10 points)

```
rand = rio::import("http://econ300.com/rand.dta")
library(dplyr)

rand %>%
  group_by(any_ins) %>%
  summarise("mean(female)" = mean(female, na.rm=T),
            "sd(female)" = sd(female),
            "n(female)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(female) sd(female) n(female)
1       0    0.5599473         NA      2689
2       1    0.5303315  0.4991572      3198
```

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(blackhisp)" = mean(blackhisp, na.rm=T),
            "sd(blackhisp)" = sd(blackhisp),
            "n(blackhisp)" = n())
```

```
Source: local data frame [2 x 4]
```

```
  any_ins mean(blackhisp) sd(blackhisp) n(blackhisp)
1       0       0.1716667            NA         2689
2       1       0.1465824            NA         3198
```

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(educper)" = mean(educper, na.rm=T),
            "sd(educper)" = sd(educper),
            "n(educper)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(educper) sd(educper) n(educper)
1       0      12.10483          NA       2689
2       1      11.93619          NA       3198
```

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(ghindx)" = mean(ghindx, na.rm=T),
            "sd(ghindx)" = sd(ghindx),
            "n(ghindx)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(ghindx) sd(ghindx) n(ghindx)
1       0     70.95892         NA      2689
2       1     69.93396         NA      3198
```

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(cholest)" = mean(cholest, na.rm=T),
            "sd(cholest)" = sd(cholest),
            "n(cholest)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(cholest) sd(cholest) n(cholest)
1       0      207.0904          NA       2689
2       1      204.1117          NA       3198
```

2. Two variables in question 1 can be considered as baseline outcome variables. For these two variables, test the null hypothesis that there is no mean difference ($H_0 : \mu = 0$). (10 points) Hint: you would want to use the following t-statistic for testing a difference in means:

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0 - \mu}{\hat{SE}(\bar{Y}^1 - \bar{Y}^0)}.$$

```
with(rand, t.test(ghindx ~ any_ins, var.equal = TRUE))
```

```
    Two Sample t-test
```

```
data:  ghindx by any_ins
t = 2.2492, df = 4357, p-value = 0.02455
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1315492 1.9183730
sample estimates:
mean in group 0 mean in group 1
       70.95892        69.93396
```

```
with(rand, t.test(cholest ~ any_ins, var.equal = TRUE))
```

```
    Two Sample t-test

data:  cholest by any_ins
t = 1.771, df = 2854, p-value = 0.07666
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3191487  6.2765898
sample estimates:
mean in group 0 mean in group 1
       207.0904        204.1117
```

- **Based on the tests, one can reject the null hypothesis that difference in population expectation of `ghindx` is zero ($t>2$). This is evidence showing that pre-treatment outcome is not balance, perhaps because the experiment is not well conducted. Another outcome variable, `cholest`, however, shows balance results ($t<2$). We might need to assess more pre-treatment outcome or characteristic variables to ensure that the experiment had created a treatment group and control group that are comparable.**

3. For the two post-treatment variables, repeat the exercises you just did (compute group mean, standard deviation, and run a t-test). Note that the results would be a bit different from Table 1.4 in the textbook. (10 points)

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(ghindxx)" = mean(ghindxx, na.rm=T),
            "sd(ghindxx)" = sd(ghindxx),
            "n(ghindxx)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(ghindxx) sd(ghindxx) n(ghindxx)
1       0      70.11108          NA       2689
2       1      68.10746          NA       3198
```

```
rand %>%
  group_by(any_ins) %>%
  summarise("mean(cholestx)" = mean(cholestx, na.rm=T),
            "sd(cholestx)" = sd(cholestx),
            "n(cholestx)" = n())
```

```
Source: local data frame [2 x 4]

  any_ins mean(cholestx) sd(cholestx) n(cholestx)
1       0       200.8954           NA        2689
2       1       201.8630           NA        3198
```

```
with(rand, t.test(ghindxx ~ any_ins, var.equal = TRUE))
```

```
    Two Sample t-test

data:  ghindxx by any_ins
t = 4.7109, df = 5081, p-value = 2.532e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.169813 2.837423
sample estimates:
mean in group 0 mean in group 1
       70.11108        68.10746
```

```
with(rand, t.test(cholestx ~ any_ins, var.equal = TRUE))
```

```
    Two Sample t-test

data:  cholestx by any_ins
t = -0.6963, df = 4511, p-value = 0.4863
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.691713  1.756589
sample estimates:
mean in group 0 mean in group 1
       200.8954        201.8630
```

- **Based on the test, we find statistically significant difference in post-treatment health index ($t = 4.7109 > 2$), but statistically insignificant difference in post-treatment cholestral level (-2$< t =$ -0.69633 < 2)**

4. Interpret the results you obtain (the mean comparison results for characteristic variables, pre-treatment outcome variables, and post-treatment outcome variables). (10 points)

- **Based on the results we have obtained, there's evidence suggesting that the experiment is not well conducted, although we need to perform more tests to see if that's the case. The experiment results are also not that consistent. In particular, it seems on average people's genral health decreased after receiving the treatment, but cholestral level show no difference after the treatment. Perhaps examing more outcome variables will give a more complete picture.**

Appendix: Stata and R Hints

Computing Group Means

- *In Stata, you can use the* `tabstat` *command. For example,*

```
use rand_initial.dta
tabstat female, by(any_ins) stat(mean sd n)
```

- *In R, you can use the* `group_by()` *function and* `summarise()` *function in the* `dplyr` *package. For example,*

```r
library(rio)
rand_initial = import("rand_initial.dta")

library(dplyr)
rand_initial %>%
  group_by(any_ins) %>%
  summarise(
    meanFemale = mean(female),
    sdFemale = sd(female),
    count = n()
  )
```

t-test

- *In Stata, the command for running a test is* `ttest`*. For example,*

```
ttest cholest, by(any_ins)
```

- *In R, the corresponding function is* `t.test()`*. For example,*

```r
t.test(cholest ~ any_ins, var.equal = TRUE)
```