Problem set 3

Siwei Dai

October 15, 2021

NOTE1: *Start with the file* `ps3_2021.Rmd` *(available from the github repository at https://github.com/ UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

NOTE2: *You will need to have a working LaTeX installation to compile your code.*

Question 1:

Consider the random process of flipping a fair coin three times.

(1a) Describe the sample space, Ω.

$$\Omega = \{(HHH), (HHT), (HTT), (HTH), (THH), (THT), (TTH), (TTT)\}$$

(1b) The random variable X that we're interested is the number of heads that we get from our random process. Write out the probability mass function for this random variable. *Hint: the coin is fair, so each of the events in the sample space above occurs with equal probability. Note how many heads we get in each event. Then look at the proportion of times we get no heads, one head, etc. These proportions are equal to the probability. List the number of heads under the x column. List the associated probabilities under the $P(X = x)$ column.*

x	$P(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8

(1c) Calculate the mean of this random variable. Please show your work.

$$E[X] = \sum_x x f(x)$$
$$= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8}$$
$$= \frac{3}{2}$$

(1c) Write out code to simulate this random process, where the output is a single realization of the random variable (i.e., a number that represents the number of heads in your coin flips).

```r
set.seed(60637)
# your code here
coin_flip <- sum(sample(
    x = c(0, 1),
    size = 3,
    replace = TRUE))
coin_flip
```

```
## [1] 2
```

(1d) Now run your random process so you sample from it 10,000 times [PLEASE DON'T OUTPUT ALL 10,000 OBSERVATIONS IN YOUR HOMEWORK, just save it to an R object]. What is the average number of heads across these 10k observations? This is the sample mean for a given sample.

```r
# your code here
omega <- c(0, 1, 2, 3)
probs <- c(1/8, 3/8, 3/8, 1/8)
coin_10k <- sample(
    x = omega,
    size = 10000,
    replace = TRUE,
    prob = probs)
coin_heads <- mean(coin_10k)
coin_heads
```

```
## [1] 1.506
```

(1e): Write your own function called `mymean()` to calculate the sample mean from a vector. Apply your function to your size 10k sample that you saved in the last problem. (Don't use `mean()` inside your function.)

```r
# your code here
mymean <- function(x) {
    sum(x) / length(x)
}
mymean(coin_10k)
```

```
## [1] 1.506
```

Question 2:

Using the same random process of flipping three fair coins, code the random variable Y as 1 if we get three heads, and 0 otherwise.

(2a) Write out the probability mass function for this random variable Y.

y	$P(X = x)$
0	7/8
1	1/8

(2a) Write out the joint probability mass function for the joint distribution of X and Y.

x	y	$P(X = x, Y = y)$
0	0	1/8
1	0	3/8
2	0	3/8
3	1	1/8

(2b) Write out the probability mass function for this random variable X *conditional* on Y.

x	y	$P(X = x \mid Y = y)$
0	0	1/7
1	0	3/7
2	0	3/7
3	1	1

Question 3:

(3a) Load the data set that you selected for your independent project. If your data set is not already in tibble format, transform it into a tibble. Print the data set so that we can see the top few observations and the column names and types.

```r
# your code here
datapath <- './ps_data/FullData.csv'
df <- as.tibble(read.csv(datapath))
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
```

```r
# The variables are in Chinese;
# essentially the variables included the name of the bureaucrat,
# when a particular experience started and ended,
# where he assumed the office and the level of the office,
# and other personal variables (alma mater, etc.)
head(df)
```

```
## # A tibble: 6 x 25
##    用户编码 姓名    经历序号 起始时间.YYYY.MM.~ 终止时间..YYYY.MM.~ 是否全国性组织
##       <int> <chr>      <int> <chr>             <chr>               <chr>
## 1        1 孙永春          1 1974/6/1          1976/12/1           否
## 2        1 孙永春          2 1976/12/1         1978/3/1            否
## 3        1 孙永春          3 1978/3/1          1978/10/1           否
## 4        1 孙永春          4 1978/10/1         1980/6/1            否
```

```
## 5        1 孙永春        5 1980/6/1        1984/8/1        否
## 6        1 孙永春        6 1984/8/1        1987/9/1        否
## # ... with 19 more variables: 是否中央派出 <chr>, 一级关键词编码 <int>,
## #   地方一级关键词 <chr>, 二级关键词编码 <int>, 地方二级关键词 <chr>,
## #   地区级别 <chr>, 录入信息备注 <lgl>, 地方三级关键词 <chr>, 基本大类别 <chr>,
## #   职务一级关键词编码 <int>, 职务一级关键词 <chr>, 录入信息备注.1 <lgl>,
## #   职务二级关键词 <chr>, 职务三级关键词 <chr>, 具体职务 <chr>, 级别 <chr>,
## #   标志位 <chr>, 学历 <chr>, 序号 <int>
```

(3b) What do you think is the appropriate unit of observation in your data? Is your data set already formatted so that each row describes a unique unit of observation? If not, what does each row describe?

I want to investigate why certain bureaucrats, after being removed from office or being demoted, can be re-promoted while others didn't. An appropriate unit of observation should be the bureaucrat who got demoted, with variables denoting when he was demoted, if after demotion he has been promoted again, and personal variables (connections with upper-level elites, alma mater, etc.)

Currently, each row denotes a unique experience within a political elite's career, i.e. what one's job was at a given year, along with the personal variables. After adding new variables (if he has been demoted, if after demotion he has been promoted again) by mutating, and filtering the observations to leave the demoted bureaucrats within the dataset, the dataset would be fine.

[**Extra credit:** if your data set will need to be reshaped using `pivot_longer()` or `pivot_wider()`, try reshaping it now. If it doesn't need to be reshaped, you can try reshaping it anyhow. Give your reshaped columns informative names. Explain what the unit of observation is in your reshaped data set.]

```r
# The following reshaping is to integrate each observation from one career experience
# to an observation on one bureaucrat, with extra variables denoting the position he had
# in each career phase
df_tidy <- df %>%
    # translate the variables into English
    # exper_num refers to the career stage
    rename('name' = 姓名,
           'position_ori' = 级别,
           'exper_num' = 经历序号) %>%
    select(name, position_ori, exper_num) %>%
    # recode the position from strings into numeric; larger the number, higher the position
    mutate('position_numeric' = case_when(position_ori == '无级别' ~ 0,
                                          position_ori == '小于副处' ~ 1,
                                          position_ori == '副处' ~ 2,
                                          position_ori == '正处' ~ 3,
                                          position_ori == '副厅' ~ 4,
                                          position_ori == '正厅' ~ 5,
                                          position_ori == '副部' ~ 6,
                                          position_ori == '正部' ~ 7,
                                          position_ori == '副国' ~ 8,
                                          position_ori == '正国' ~ 9)) %>%
    select(name, position_numeric, exper_num) %>%
    pivot_wider(
        names_from = exper_num,
        names_prefix = 'Experience No.',
        values_from = position_numeric)
# Each observation now describes the levels of every office a bureaucrat assumed in his career.
# The columns denote the nth position the bureaucrat had in his career.
head(df_tidy)
```

```
## # A tibble: 6 x 38
##   name   `Experience No.1` `Experience No.2` `Experience No.3` `Experience No.4`
##   <chr>             <dbl>             <dbl>             <dbl>             <dbl>
## 1 孙永春               1                 1                 1                 1
## 2 雷建国               0                 0                 0                 1
## 3 黄胜                 0                 0                 1                 1
## 4 吴翠云               1                 1                 1                 1
## 5 陈先运               0                 0                 0                 0
## 6 郭兆信               0                 0                 1                 1
## # ... with 33 more variables: Experience No.5 <dbl>, Experience No.6 <dbl>,
## #   Experience No.7 <dbl>, Experience No.8 <dbl>, Experience No.9 <dbl>,
## #   Experience No.10 <dbl>, Experience No.11 <dbl>, Experience No.12 <dbl>,
## #   Experience No.13 <dbl>, Experience No.14 <dbl>, Experience No.15 <dbl>,
## #   Experience No.16 <dbl>, Experience No.17 <dbl>, Experience No.18 <dbl>,
## #   Experience No.19 <dbl>, Experience No.20 <dbl>, Experience No.21 <dbl>,
## #   Experience No.22 <dbl>, Experience No.23 <dbl>, Experience No.24 <dbl>, ...
```