STAT 344 Group Project (Part I)

Lucas Qin, Frederick Wang, Isabel Wilson

2023-10-17

Reading in Data

```
all_data_files = list.files("data")
all_df = read.csv(paste0("data/", all_data_files[1]))

for (i in 2 : length(all_data_files)) {
  all_df = rbind(all_df, read.csv(paste0("data/", all_data_files[i])))
}

filtered_df = all_df %>% filter(Section == "OVERALL") %>%
  mutate(CourseNum = ifelse(
    is.na(Detail), Course, paste0(Course, Detail)
  )) %>%
  select(-Campus, -Year, -Session, -Section, -Professor, -Course, -Detail)

filtered_df %>% slice(1:10)
```

```
##    Subject                                                 Title Enrolled
## 1     AANB                             Topics in Animal Welfare       11
## 2     ACAM                       Asian Canadians in Popular Culture       50
## 3     ACAM                               Dis/Orienting Asian Canada       48
## 4     ACAM                          Selected Topics in ACAM Studies       53
## 5     ACAM                          Selected Topics in ACAM Studies       12
## 6     ACAM                     Asian Canadian Community-Based Media       18
## 7     ACAM Directed Studies in Asian Canadian and Asian Migration        9
## 8     ADHE                                          Teaching Adults      164
## 9     ADHE                          Institutions of Adult Education      142
## 10    ADHE       Developing Short Courses, Workshops and Seminars      166
##          Avg   Std.dev High Low X.50 X50.54 X55.59 X60.63 X64.67 X68.71 X72.75
## 1  92.72727  2.611165   95  90    0      0      0      0      0      0      0
## 2  77.96000 10.604119   93  46    1      2      1      1      2      2      5
## 3  77.60417 17.594414   91   5    3      1      0      0      1      1      3
## 4  80.24528 13.027937   92   8    1      1      1      0      0      4      5
## 5  80.91667  3.553701   86  75    0      0      0      0      0      0      2
## 6  84.94444  3.621378   90  77    0      0      0      0      0      0      0
## 7  94.33333  2.958040   98  90    0      0      0      0      0      0      0
## 8  81.92073 12.560521   97   0    2      2      3      0      4      9     10
## 9  84.06338 11.298161   98   0    2      0      0      7      0      2      9
## 10 83.60241 12.079474   98   0    2      0      0      1      6      5      8
```

```
##    X76.79 X80.84 X85.89 X90.100 CourseNum
## 1       0      0      0      11       550
## 2       3     17     12       4       250
## 3      10      8     16       5       300
## 4       2     14     18       7      320B
## 5       3      5      2       0      320D
## 6       1      6      9       2       350
## 7       0      0      0       9      447C
## 8      12     40     47      35       327
## 9       4     33     35      50       328
## 10     18     34     38      54       329
```

```
N = nrow(filtered_df)
N
```

```
## [1] 4263
```

Since we don't have access to the population standard deviation to guess the standard error of the sample we will be taking, we decided to use our personal transcript standard deviations to estimate the sample standard error. We found our transcript standard deviations to be around 7 percent, so we decided to guess the sample standard error to be 7 as well.

```
margin_of_error = 1 # desired width is 2%
sample_stdev_guess = 7 # intuitive guess since we don't have previous studies
n = (1/(1/(qnorm(0.975)^2*sample_stdev_guess^2) + 1/N)) %>% ceiling()
n
```

```
## [1] 181
```

```
set.seed(1)
srs = sample(filtered_df$Avg, n)
srs
```

```
##   [1] 74.56250 71.29213 91.22581 82.38462 75.32812 88.62500 90.33333 77.70563
##   [9] 93.63636 74.67647 78.22605 75.90698 82.68966 83.72727 83.57576 83.21176
##  [17] 75.77686 76.36039 86.11111 91.98361 62.62740 86.45455 76.51220 69.29897
##  [25] 84.25000 83.87619 81.00000 80.87143 93.33333 90.90000 75.72131 90.20833
##  [33] 78.67857 72.79200 83.06897 81.03704 75.95000 78.44578 66.02703 80.45238
##  [41] 81.35294 81.75000 85.73620 73.94118 87.06780 85.92683 86.07143 80.51515
##  [49] 91.38462 95.75000 79.85714 75.74747 84.56250 75.88889 91.58333 86.65000
##  [57] 76.47059 92.33333 83.11111 90.69565 71.75000 84.98291 79.82143 87.87500
##  [65] 79.30000 82.53191 82.73256 87.57143 89.92683 80.47619 95.19048 75.70968
##  [73] 88.13228 92.00000 74.70992 71.26263 82.66667 92.50000 77.78571 76.13793
##  [81] 87.58333 79.35435 86.69444 74.53982 93.10000 72.71951 93.66667 90.72727
##  [89] 80.63158 90.87500 69.23427 76.96667 69.82143 87.30000 68.88806 90.55556
##  [97] 80.44928 76.21818 72.16667 85.70588 85.88889 83.40441 93.02128 87.37500
## [105] 88.07317 85.54545 79.70513 81.17290 78.73077 80.45000 71.55556 79.50000
## [113] 77.26667 72.25989 84.60000 60.23077 91.27273 83.31250 87.26316 84.89062
## [121] 83.45238 88.54545 95.61538 88.64474 80.14545 81.80952 73.81690 78.90625
## [129] 74.73585 73.81767 67.89908 92.25000 81.10526 85.45161 88.30000 78.05882
## [137] 74.00000 75.92000 95.50000 94.14286 81.72308 73.29412 85.64286 84.80645
## [145] 88.00000 84.33333 88.55556 80.25000 88.65714 72.52941 69.92453 84.71429
```

```
## [153] 87.23077 79.77778 84.55556 89.83333 88.20000 78.57143 86.46667 72.50000
## [161] 82.13043 86.65217 76.41917 78.73077 78.76471 78.38095 82.10169 89.16667
## [169] 71.21429 75.18919 88.64706 72.67188 80.32967 76.21164 84.00000 77.33333
## [177] 90.36735 89.94737 91.66667 84.67188 74.38596
```

```
sample_mean =  mean(srs); sample_mean
```

```
## [1] 81.99828
```

```
sample_se = sqrt((1-n/N)*sd(srs)^2/n); sample_se
```

```
## [1] 0.5199058
```

```
ci_lb = sample_mean - qnorm(0.975)*sample_se
ci_ub = sample_mean + qnorm(0.975)*sample_se

conf_int = c(ci_lb, ci_ub)
c("Confidence Interval for Average UBC Grades Across All Classes in 2021",
  "Winter Using Simple Random Sample",
  conf_int)
```

```
## [1] "Confidence Interval for Average UBC Grades Across All Classes in 2021"
## [2] "Winter Using Simple Random Sample"
## [3] "80.9792785841943"
## [4] "83.0172719455626"
```