ECON 21020 Tabord-Meehan Pset 3 Question 6

```
options(digits=4,scipen=100)
```

Section (a)

```
data <- read_xlsx("caschool.xlsx")
data
```

```
## # A tibble: 420 x 18
##    `Observation Nu~ dist_cod county district gr_span enrl_tot teachers calw_pct
##               <dbl>    <dbl> <chr>  <chr>    <chr>      <dbl>    <dbl>    <dbl>
##  1                1    75119 Alame~ Sunol G~ KK-08        195    10.9     0.510
##  2                2    61499 Butte  Manzani~ KK-08        240    11.1    15.4
##  3                3    61549 Butte  Thermal~ KK-08       1550    82.9    55.0
##  4                4    61457 Butte  Golden ~ KK-08        243    14      36.5
##  5                5    61523 Butte  Palermo~ KK-08       1335    71.5    33.1
##  6                6    62042 Fresno Burrel ~ KK-08        137     6.40   12.3
##  7                7    68536 San J~ Holt Un~ KK-08        195    10      12.9
##  8                8    63834 Kern   Vinelan~ KK-08        888    42.5    18.8
##  9                9    62331 Fresno Orange ~ KK-08        379    19      32.2
## 10               10    67306 Sacra~ Del Pas~ KK-06       2247   108      79.0
## # ... with 410 more rows, and 10 more variables: meal_pct <dbl>,
## #   computer <dbl>, testscr <dbl>, comp_stu <dbl>, expn_stu <dbl>, str <dbl>,
## #   avginc <dbl>, el_pct <dbl>, read_scr <dbl>, math_scr <dbl>
```

Answer: We have 420 observations

Section (b)

```
data <- data %>% mutate(income = avginc * 1000)
data
```

```
## # A tibble: 420 x 19
##    `Observation Nu~ dist_cod county district gr_span enrl_tot teachers calw_pct
##               <dbl>    <dbl> <chr>  <chr>    <chr>      <dbl>    <dbl>    <dbl>
##  1                1    75119 Alame~ Sunol G~ KK-08        195    10.9     0.510
##  2                2    61499 Butte  Manzani~ KK-08        240    11.1    15.4
##  3                3    61549 Butte  Thermal~ KK-08       1550    82.9    55.0
##  4                4    61457 Butte  Golden ~ KK-08        243    14      36.5
##  5                5    61523 Butte  Palermo~ KK-08       1335    71.5    33.1
##  6                6    62042 Fresno Burrel ~ KK-08        137     6.40   12.3
##  7                7    68536 San J~ Holt Un~ KK-08        195    10      12.9
```

```
##  8                8    63834 Kern   Vinelan~ KK-08        888    42.5    18.8
##  9                9    62331 Fresno Orange ~ KK-08        379    19      32.2
## 10               10    67306 Sacra~ Del Pas~ KK-06       2247   108      79.0
## # ... with 410 more rows, and 11 more variables: meal_pct <dbl>,
## #   computer <dbl>, testscr <dbl>, comp_stu <dbl>, expn_stu <dbl>, str <dbl>,
## #   avginc <dbl>, el_pct <dbl>, read_scr <dbl>, math_scr <dbl>, income <dbl>
```

Part (i): The variable `income` measures average district income, denominated in dollars.

```
avginc_mean = mean(data$avginc)
avginc_sd = sd(data$avginc)
```

Part (ii): The mean of `avginc` is 15.3166 and the standard deviation of `avginc` is 7.2259.

```
inc_mean = mean(data$income)
inc_sd = sd(data$income)
```

Part (iii): The mean of `income` is 15316.5881 and the standard deviation of `income` is 7225.8898.

The mean and standard deviations of `income` are 1000 times the mean and standard deviation of `avginc`, which is what I would expect.

Section (c)

```
mean_math = mean(data$math_scr)
```

Part (i): The mean math score is 653.3426.

```
#From Stack Overflow: Learned that you can find proportions by taking the mean
#of a vector of boolean (true/false) values. Almost like an indicator var.
#https://stackoverflow.com/questions/68485739/calculating-proportion-of-values-using-condition-and-grou

data_new <- data %>%
  mutate(is_large = ifelse(data$str > 20, 1, 0)) %>%
  group_by(is_large) %>%
  summarize(math = mean(math_scr), varmath = var(math_scr), n = n()) %>%
  mutate(frac = n/sum(n))

data_new
```

Part (ii):

```
## # A tibble: 2 x 5
##   is_large  math varmath     n  frac
## *    <dbl> <dbl>   <dbl> <int> <dbl>
## 1        0  656.    374.   243 0.579
## 2        1  650.    304.   177 0.421
```

243/420 schools have class sizes of 20 students or less, and the mean math score among these schools is 655.7.

Part (iii): Per the summary table above, 177/420 schools have class sizes of more than 20 students, and the mean math score among these schools is 650.1.

Part (iv): In math: The overall mean we recovered in part 1 should be equal to a weighed sum of the group means recovered in part 2 and part 3 where the weights are the fraction of the total observations that fall into each group.

Part (v):

$H_0 : E[Math|is\_large = 0] = E[Math|is\_large = 1] \Rightarrow E[Math|is\_large = 0] - E[Math|is\_large = 1] = 0 H_a : E[Math|is\_$

To simplify notation: Let LG be a variable describing the math scores of districts with large class sizes, and SM be a variable describing the math scores of districts with small classes. Then we can conduct a two-sample test.

The test statistic T_n is as follows:

$$T_n = \left| \frac{\bar{SM} - \bar{LG} - 0}{\sqrt{\frac{\hat{\sigma_{LG}}}{n_{LG}} + \frac{\hat{\sigma_{SM}}}{n_{SM}}}} \right|$$

In code:

```
data_new <- data_new %>% arrange(is_large) #ensure that small is before large
```

```
mean_sm = data_new$math[[1]] #get relevant col of first row
mean_lg = data_new$math[[2]] #get relevant col of second row

var_sm = data_new$varmath[[1]]
var_lg = data_new$varmath[[2]]

count_sm = data_new$n[[1]]
count_lg = data_new$n[[2]]

diff = mean_sm - mean_lg
se = sqrt(var_sm/count_sm + var_lg/count_lg)

T_n = abs(diff/se)
T_n
```

```
## [1] 3.122
```

Compare to the critical value at $c_{1-0.1/2} = c_{0.95}$

```
crit_val <- qnorm(0.95)
```

Because `T_n` (3.1218) is greater than the `critical value` (1.6449), we will reject the null hypothesis at the 10% significance level.

```
cov_avg = cov(data$avginc,data$math_scr)
cov_inc = cov(data$income,data$math_scr)
```

Part (vi): The covariance with `avginc` is 94.7795 but the covariance with `income` 94779.4973. They are not the same, because the coviariance is not unitless. It is sensitive to changes in units (i.e. to scalar multiplication across all observations)

```
corr_avg = cor(data$avginc,data$math_scr)
corr_inc = cor(data$income,data$math_scr)
```

Part (vi) The correlation with `avginc` is 0.6994 and the correlation with `income` is 0.6994. They are the same because the correlation coefficient normalizes by the variances so as to be insensitive to changes in numeraires or other units.