ECON 21020 PSet 3

Phalgun Garimella

Question 6a

```
caschool = read_excel("caschool.xlsx")
```

There are 420 observations in the data set.

Question 6b

```
income = caschool$avginc * 1000
```

i.

The variable `income` measures average district income in dollars (as opposed to thousands of dollars like in `avginc`).

ii.

```
mean(caschool$avginc)
```

```
## [1] 15.32
```

```
sd(caschool$avginc)
```

```
## [1] 7.226
```

The first number displayed is the mean of `avginc`.

The second number displayed is the standard deviation of `avginc`.

Both values are in thousands of dollars.

iii.

```
mean(income)
```

```
## [1] 15317
```

```
sd(income)
```

```
## [1] 7226
```

The first number displayed is the mean of `income`.

The second number displayed is the standard deviation of `income`.

Both values are in dollars.

Given our result in the previous part, the mean and standard deviation for `income` are what we expect because `income` is simply a scaled version of `avginc`. When data is scaled by a constant (in this case it was by a factor of 1000), the mean and standard deviation of that data are also scaled by the same constant. Hence, our results make sense.

Question 6c

i.

```
mean(caschool$math_scr)
```

```
## [1] 653.3
```

The number displayed above is the mean math score across all districts.

ii.

```
math_max_20 = c()
counter1 = 0
for (i in 1:420) {
  if (caschool$str[i] <= 20) {
    math_max_20 = c(math_max_20, caschool$math_scr[i])
    counter1 = counter1 + 1
  }
```

```
}
fraction1 = counter1 / 420
mean1 = mean(math_max_20)
print(fraction1)
```

```
## [1] 0.5786
```

```
print(mean1)
```

```
## [1] 655.7
```

The first number displayed is the fraction of districts that have an average class size of 20 or fewer students.

The second number displayed is the mean math score in districts with average class size of 20 or fewer students.

iii.

```
math_min_20 = c()
counter2 = 0
for (i in 1:420) {
  if (caschool$str[i] > 20) {
    math_min_20 = c(math_min_20, caschool$math_scr[i])
    counter2 = counter2 + 1
  }
}
fraction2 = counter2 /420
mean2 = mean(math_min_20)
print(fraction2)
```

```
## [1] 0.4214
```

```
print(mean2)
```

```
## [1] 650.1
```

The first number displayed is the fraction of districts that have an average class size of more than 20 students.

The second number displayed is the mean math score in districts with average class size of more than 20 students.

iv.

We can derive our answer in (i) using our answers in (ii) and (iii). In particular, we can take a weighted average of the mean math score in districts for both groups of average class size using the two fractions calculated earlier.

```
print(fraction1 * mean1 + fraction2 * mean2)
```

```
## [1] 653.3
```

The result of our calculation is as desired: the mean math score across all districts.

v.

Let us denote the population mean math score in districts with average class size of 20 or fewer students as $\mu_{str \leq 20}$.

Let us denote the population mean math score in districts with average class size of more than 20 students as $\mu_{str > 20}$.

Our hypotheses, using the population level conditional expectations described above, are as follows:

$H_0 : \mu_{str \leq 20} = \mu_{str > 20}$

$H_1 : \mu_{str \leq 20} \neq \mu_{str > 20}$

Since we are faced with a difference of means problem, we can conduct a two-sided, two-sample t-test. We can calculate our test statistic using the formula $T = |\frac{\bar{Y}_{str>20} - \bar{Y}_{str \leq 20}}{\sqrt{\frac{s^2_{str>20}}{N_{str>20}} + \frac{s^2_{str \leq 20}}{N_{str \leq 20}}}}|$, where:

$\bar{Y}_{str>20}$ is the sample mean math score in districts with average class size of more than 20 students.

$\bar{Y}_{str \leq 20}$ is the sample mean math score in districts with average class size of 20 or fewer students.

$s^2_{str>20}$ is the sample variance of the math score in districts with average class size of more than 20 students.

$s^2_{str \leq 20}$ is the sample variance of the math score in districts with average class size of 20 or fewer students.

$N_{str>20}$ is the sample size of the districts with average class size of more than 20 students.

$N_{str \leq 20}$ is the sample size of the districts with average class size of 20 or fewer students.

```
t_stat = abs((mean2 - mean1) /
               sqrt(var(math_min_20) / counter2 + var(math_max_20) / counter1))
print(t_stat)
```

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

Given the value of our test statistic displayed above, we can now find our p-value using $df = 420 - 2 = 418$ and make a conclusion at the 10% level.

```
2 * pt(q = t_stat, df = 418, lower.tail = FALSE)
```

```
## [1] 0.001922
```

Since the p-value displayed above is less than 0.1, we reject the null hypothesis at the 10% level in favor of the alternate hypothesis that the population mean math score in districts with average class size of 20 or fewer students is different from the population mean math score in districts with average class size of more than 20 students.

vi.

```
cov(caschool$avginc, caschool$math_scr)
```

```
## [1] 94.78
```

```
cov(income, caschool$math_scr)
```

```
## [1] 94779
```

The first number displayed is the covariance between `avginc` and `math_scr` with units thousands of dollars · points.

The second number displayed is the covariance between `income` and `math_scr` with units dollars · points.

The two covariances are different due to properties of covariance. If we consider $Cov[aX, Y]$ where a is some constant, we observe $Cov[aX, Y] = \mathbb{E}[aXY] - \mathbb{E}[aX]\mathbb{E}[Y] = a\mathbb{E}[XY] - a\mathbb{E}[X]\mathbb{E}[Y] = a(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) = aCov[X, Y]$. Hence, since we scale `avginc` by a constant to get `income`, we can see that the covariance between `income` and `math_scr` is scaled by the same constant compared to the covariance between `avginc` and `math_scr`.

vii.

```
cor(caschool$avginc, caschool$math_scr)
```

```
## [1] 0.6994
```

```r
cor(income, caschool$math_scr)
```

```
## [1] 0.6994
```

The first number displayed is the correlation between `avginc` and `math_scr`.

The second number displayed is the correlation between `income` and `math_scr`.

Since these two values are correlations, they are unit-less.

The two correlations are the same due to properties of correlation – since we normalize each covariance by the standard deviation of both metrics, the scaling of `income` compared to `avginc` is irrelevant. If we consider $Corr[aX, Y]$ where a is some constant, we observe $Corr[aX, Y] = \frac{Cov[aX,Y]}{\sigma_{aX}\sigma_Y} = \frac{aCov[X,Y]}{a\sigma_X\sigma_Y} = \frac{Cov[X,Y]}{\sigma_X\sigma_Y} = Corr[X, Y]$. Hence, since we scale `avginc` by a constant to get `income`, we can see that the correlation in either case remains the same.