Dehshini Ali

Problem set 4

i.

```
>library(tidyverse)
>balt621 <- read_csv("balt621.csv")
>balt621 %>%
group_by(season) %>%
summarize(n_pm10 = sum(!is.na(pm10)), mean_pm10 = mean(pm10, na.rm=TRUE),
n_mortality = sum(!is.na(death)), mean_mortality = mean(death))
```

Table showing each season and the number of days of PM10 observations, number of mortalities, mean number of mortalities and mean PM10 values.

season	n_pm10	mean_pm10	n_mortality	mean_mortality
Autumn	559	-2.97	1748	19.2
Spring	563	0.740	1729	18.2
Summer	571	4.58	1748	17.7
Winter	551	-2.93	1715	20.8

ii.

```
pm10winter <- filter(balt621, season=="Winter")
quintiles = quantile(pm10winter$pm10, c(0,.2,.4,.6,.8,1), na.rm=TRUE)
pm10winter$pm_group = cut(pm10winter$pm10, breaks=quintiles, labels=1:5)
table(pm10winter$pm_group)
```

PM10 strata quintiles and number of days

For winter

PM10 strata	1	2	3	4	5
Number of days	110	110	110	110	110

For autumn

PM10 strata	1	2	3	4	5
Number of days	111	112	111	112	112

For summer

PM10 strata	1	2	3	4	5
Number of days	114	114	114	114	114

For spring

PM10 strata	1	2	3	4	5
Number of days	112	112	113	112	113

iii.

```
pm10winter %>% filter(pm_group==1) %>%
summarize(mean=mean(death), sd=sd(death), n=n())

pm10winter %>% filter(pm_group==5) %>%
summarize(mean=mean(death), sd=sd(death), n=n())
```

For winter

lowest quintile

mean	sd	n
22.23636	5.303499	110

Highest quintile

mean	sd	n
21.35455	5.35456	110

For summer

Lowest quintile

mean	sd	n
17.63158	4.154671	114

Highest quintile

mean	sd	n
19.92982	5.678276	114

For spring

Lowest quintile

mean	sd	n
18.55357	4.191023	112

Highest quintile

mean	sd	n
18.99115	5.008021	113

For autumn

Lowest quintile

mean	sd	n
19.81982	5.377053	111

Highest quintile

mean	sd	n
20.52679	5.42759	112

iv.

```
pm10winter.15 = pm10winter %>% filter(pm_group==1 | pm_group==5)
t.test(death ~ pm_group, data=pm10winter.15, var.equal=FALSE)
```

For winter

```
Welch Two Sample t-test
data:  death by pm_group
t = 1.2272, df = 217.98, p-value = 0.2211
alternative hypothesis: true difference in means between group 1 and group 5 is not equal to 0
95 percent confidence interval:
 -0.5344249  2.2980613
sample estimates:
mean in group 1 mean in group 5
       22.23636        21.35455
```

Null hypothesis: true difference in means between group 1 (lowest PM10 pollution days) and group 5 (highest PM10 pollution days) is equal to zero.

Fail to reject null hypothesis because P value is greater than set alpha level of 0.05.

```
pm10spring.15 = pm10spring %>% filter(pm_group==1 | pm_group==5)
t.test(death ~ pm_group, data=pm10spring.15, var.equal=FALSE)
```

For spring

```
Welch Two Sample t-test
data:  death by pm_group
t = -0.71099, df = 216.92, p-value = 0.4779
alternative hypothesis: true difference in means between group 1 and group 5 is not equal to 0
95 percent confidence interval:
 -1.6506036  0.7754455
sample estimates:
mean in group 1 mean in group 5
       18.55357        18.99115
```

Null hypothesis: true difference in means between group 1 (lowest PM10 pollution days) and group 5 (highest PM10 pollution days) is equal to zero.

Fail to reject null hypothesis because P value is greater than set alpha level of 0.05.

```
pm10summer.15 = pm10summer %>% filter(pm_group==1 | pm_group==5)
t.test(death ~ pm_group, data=pm10summer.15, var.equal=FALSE)
```

For summer

```
Welch Two Sample t-test
data:  death by pm_group
t = -3.4876, df = 207.04, p-value = 0.0005951
alternative hypothesis: true difference in means between group 1 and group 5 is not equal to 0
95 percent confidence interval:
 -3.5974049 -0.9990863
sample estimates:
mean in group 1 mean in group 5
      17.63158       19.92982
```

Null hypothesis: true difference in means between group 1 (lowest PM10 pollution days) and group 5 (highest PM10 pollution days) is equal to zero.

reject null hypothesis because P value is less than set alpha level of 0.05.

```
pm10autumn.15 = pm10autumn %>% filter(pm_group==1 | pm_group==5)
t.test(death ~ pm_group, data=pm10autumn.15, var.equal=FALSE)
```

For autumn

```
Welch Two Sample t-test
data:  death by pm_group
t = -0.9771, df = 221, p-value = 0.3296
alternative hypothesis: true difference in means between group 1 and group 5 is not equal to 0
95 percent confidence interval:
 -2.1328719  0.7189401
sample estimates:
mean in group 1 mean in group 5
      19.81982       20.52679
```

Null hypothesis: true difference in means between group 1 (lowest PM10 pollution days) and group 5 (highest PM10 pollution days) is equal to zero.

Fail to reject null hypothesis because P value is greater than set alpha level of 0.05.

v. 95% C.I by hand

$(x_1\text{-}x_5)$ □ $t_{a/2,\text{ df}}(s.e_{x1\text{-}x5})$

We are comparing 2 means, we don’t know the population variances. Ideally, we do the variance test (F statistic) to determine if the population variances are equal or not and so if we should pool the variances or not.
“Given the large sample sizes, there is no need to pool the variances”

$$s.e_{x1\text{-}x5} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

a = 0.05
df = modified df.
t approx equal to z when sample sizes are large.

For winter

(22.24-21.355) +/- 1.98($\sqrt{\frac{5.3^2}{110} + \frac{5.35^2}{110}}$)

0.885 +/- 1.96(0.718)
0.885 +/- 1.407
(-0.522) , 2.292

for summer

(17.63-19.93) +/- 1.98($\sqrt{\frac{4.15^2}{114} + \frac{5.68^2}{114}}$)
-2.3 +/- 1.96(0.659)
-2.3 +/- 1.292
-3.592 , (-1.008)

for spring

(18.55-18.99) +/- 1.98($\sqrt{\frac{4.19^2}{112} + \frac{5.01^2}{113}}$)
-0.44 +/- 1.96(0.616)
-1.647 , 0.767

for autumn

(19.82-20.53) +/- 1.98($\sqrt{\frac{5.34^2}{111} + \frac{5.43^2}{112}}$)

-0.71 +/- 1.96(0.721)
-2.123 , 0.703

vi.

Mean of mortalities (confidence interval [lower limit, upper limit])

	Summer	Winter	Spring	Autumn
Lowest quintile, **mean (interval)**	17.63158 (16.86, 18.41)	22.23636 (21.25, 23.22)	18.55357 (17.77, 19.33)	19.81982 (18.82, 20.82)
Highest quintile, **mean (interval)**	19.92982 (18.88, 20.98)	21.35455 (20.35, 22.35)	18.99115 (18.06, 19.92)	20.52679 (19.53, 21.53)
Difference, **mean (interval)**	-2.3 (-3.60, -1.0)	0.89 (-0.53, 2.30)	-0.44 (-1.65, 0.78)	-0.71 (-2.13, 0.72)

vii.

When controlling for seasonality, mortality is not significantly affected by pollution except in summer where high pollution days have a higher mean mortality compared to lower pollution days (mean difference= -2.3, confidence interval = -3.6, -1.0).

As the seasons get hotter, high pollution days have higher mean mortalities as compared to lower pollution days.
Mean differences: winter 0.89 (-0.53, 2.30) > spring -0.44 (-1.65, 0.78) > autumn -0.71 (-2.13, 0.72) > summer -2.3 (-3.6, -1.0)
This may signify an effect modification.