Problem Set 7 Key

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(summarytools)
```

```
Attaching package: 'summarytools'

The following object is masked from 'package:tibble':

    view
```

```
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

```
library(lsr)
```

1 Problem 1

1. The data set "acupuncture.csv" includes data from the control group of a randomized controlled trial investigating the effect of acupuncture therapy on headache severity. Each participant was identified using an ID number in the id column. The headache severity was measured for each participant before receiving the treatment and at a 3-month follow-up. The column pk1 includes the baseline headache severity score of the participants while the column pk2 includes the headache severity score after 3 months.

```
acu <- read.csv("datasets/acupuncture.csv")
glimpse(acu)
```

```
Rows: 173
Columns: 5
$ X     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1~
$ id    <int> 112, 113, 114, 130, 131, 137, 138, 141, 149, 150, 161, 169, 184,~
$ group <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ pk1   <dbl> 9.25, 42.50, 24.25, 21.75, 14.50, 11.75, 56.50, 15.50, 49.25, 9.~
$ pk2   <dbl> 4.75, 34.50, 16.25, 2.00, 20.00, 11.25, 70.50, 5.75, 9.75, 8.75,~
```

   a. What test is the most appropriate to test whether there is a difference between the average headache severity scores of the participants before receiving the treatment and the 3-month follow-up? [1pt.]

Paired t-test because these scores were measured from the same participant.

   b. What are the statistical hypotheses for testing whether the average headache severity score in this cohort was lower at the 3-month follow-up compared to the baseline? [2pts.]

$$H_0 : \mu_d = 0; H_a : \mu_d > 0$$

where μ_d is defined as the average difference between baseline and 3-month follow-up.

> **!** Important
>
> If _d is the difference between the treatment between 3-month baseline and follow-up, then the hypothesis should be:

$$H_0 : \mu_d = 0; H_a : \mu_d < 0$$

c. What is the value of the test statistic and the corresponding number of degrees of freedom? [2pts.]

```
ttest <- t.test(acu$pk1,acu$pk2,paired=T,alternative="greater")
ttest
```

```
    Paired t-test

data:  acu$pk1 and acu$pk2
t = 6.9415, df = 172, p-value = 3.799e-11
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 4.916204      Inf
sample estimates:
mean difference
       6.453757
```

d. What is the p-value? [1pt.]

The p-value is $3.7991629 \times 10^{-11}$.

e. At a significance level of 0.001, do we have sufficient evidence that the average headache severity score in this cohort was lower at the 3-month follow-up compared to the baseline? [1pt.]

Yes, we have sufficient evidence that the average headache severity score in this cohort was lower at the 3-month follow-up compared to the baseline.

2. The file aaqol_subset.csv includes data from the 2015 Asian American Quality of Life survey implemented in Austin, TX. We are interested to test whether there is a difference in average age across the different Asian ethnicities represented in the study.

```
aaqol <- read.csv("datasets/aaqol_subset.csv")
glimpse(aaqol)
```

```
Rows: 1,000
Columns: 3
$ X         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
$ Age       <int> 21, 26, 45, 20, 21, 26, 66, 23, 26, 59, 28, 19, 61, 25, 69, ~
$ Ethnicity <chr> "Asian Indian", "Asian Indian", "Korean", "Vietnamese", "Chi~
```

a. What are the statistical hypotheses for testing whether there is a difference in average age across the different Asian ethnicities represented in the study? [2 pts.]

The null hypothesis is that the average age across the different Asian ethnicities are equal. The alternative hypothesis is that there is at least one Asian ethnic group with a different average age.

b. What is the value of the test statistic and the corresponding number of degrees of freedom? Hint: We might need more than one type of degrees of freedom. [3pts.]

```
mod_aov<- aov(Age~Ethnicity,data=aaqol)
summary(mod_aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Ethnicity     4   4361  1090.3   3.664 0.0057 **
Residuals   995 296109   297.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. What is the resulting p-value? [1 pt.]

The p-value is 0.0057.

d. Comment on the sufficiency of evidence based on the alternative hypothesis using a significance level of 0.05. Account for the context of the problem. [1 pt.]

At a significance level of 0.05, there is sufficient evidence that at least one Asian ethnic group has a different average age compared to the others.

e. Perform a post-hoc test involving pairwise comparisons between the average age of each Asian ethnic group. Apply the Tukey correction. Which pair/s of Asian ethnic groups yielded evidence of pairwise difference in average ages? Use a significance level of 0.05. [2pts.]

```
mod_lm <- lm(Age~Ethnicity,data=aaqol)
em_mod_lm <- emmeans(mod_lm,~Ethnicity)
em_mod_lm
```

```
 Ethnicity    emmean   SE  df lower.CL upper.CL
 Asian Indian   39.6 1.14 995     37.3     41.8
 Chinese        43.6 1.05 995     41.5     45.6
 Filipino       41.0 1.68 995     37.7     44.4
 Korean         45.4 1.23 995     43.0     47.8
```

```
 Vietnamese     43.6 1.24 995     41.2     46.1

Confidence level used: 0.95
```

```
contrast(em_mod_lm,method="pairwise",adjust="tukey")
```

```
 contrast                  estimate   SE  df t.ratio p.value
 Asian Indian - Chinese     -4.0217 1.55 995  -2.602  0.0706
 Asian Indian - Filipino    -1.4737 2.03 995  -0.725  0.9507
 Asian Indian - Korean      -5.8251 1.67 995  -3.483  0.0047
 Asian Indian - Vietnamese  -4.0620 1.68 995  -2.419  0.1111
 Chinese - Filipino          2.5480 1.98 995   1.286  0.7002
 Chinese - Korean           -1.8034 1.61 995  -1.119  0.7965
 Chinese - Vietnamese       -0.0403 1.62 995  -0.025  1.0000
 Filipino - Korean          -4.3514 2.08 995  -2.089  0.2254
 Filipino - Vietnamese      -2.5883 2.09 995  -1.240  0.7282
 Korean - Vietnamese         1.7631 1.74 995   1.013  0.8494

P value adjustment: tukey method for comparing a family of 5 estimates
```

At a significance level of 0.05, Asian Indian (39.6) and Korean (45.4) participants yielded evidence of a pairwise difference in average ages, with an estimated difference of -5.82, test statistic of 3.483, and p-value of 0.0047. Other comparisons did not yield any evidence of a difference.

3. The data set socmed.csv includes the PHQ-9 scores of freshman and senior students. The students were also asked about their social media use (socmed, levels = low, high) and gaming habits (levels =yes,no) on their electronic devices.

```
socmed <- read.csv("datasets/socmed.csv")
glimpse(socmed)
```

```
Rows: 32
Columns: 5
$ X      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
$ year   <chr> "F", "F", "F", "F", "S", "S", "S", "S", "F", "F", "F", "F", "S"~
$ socmed <chr> "Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low", "High",~
$ gaming <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", ~
$ y      <int> 4, 5, 3, 4, 5, 3, 5, 5, 9, 9, 5, 8, 13, 10, 9, 9, 2, 4, 3, 3, 4~
```

a. Test whether there is an interaction between social media use and gaming time on their electronic devices. Use a significance level of 0.05. [2pts.]

```r
mod_aov <- aov(y~socmed + gaming + socmed:gaming,data=socmed)
summary(mod_aov)
```

```
              Df Sum Sq Mean Sq F value   Pr(>F)
socmed         1  94.53   94.53  43.841 3.50e-07 ***
gaming         1  57.78   57.78  26.797 1.71e-05 ***
socmed:gaming  1  13.78   13.78   6.391   0.0174 *
Residuals     28  60.38    2.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is sufficient evidence of an interaction effect between social media use and gaming habits at the 0.05 significance level. The test statistic was F(1,28)=6.391, and a p-value of 0.0174.

> **Note**
>
> Optional: We can also report the partial eta-squared as an effect size.
>
> ```r
> etaSquared(mod_aov)
> ```
>
> ```
>                    eta.sq eta.sq.part
> socmed         0.41741410   0.6102481
> gaming         0.25514006   0.4890241
> socmed:gaming 0.06085277   0.1858407
> ```
>
> The partial eta-square for the interaction effect is 0.186.

b. Provide the estimate of the group means for each social media*gaming level combination. [2pts.]

```r
mod_lm <- lm(y~socmed + gaming + socmed:gaming,data=socmed)
em_mod_lm <- emmeans(mod_lm,~socmed|gaming)
em_mod_lm
```

```
gaming = No:
 socmed emmean    SE df lower.CL upper.CL
 High     5.00 0.519 28     3.94     6.06
 Low      2.88 0.519 28     1.81     3.94

gaming = Yes:
 socmed emmean    SE df lower.CL upper.CL
```

```
 High      9.00 0.519 28     7.94    10.06
 Low       4.25 0.519 28     3.19     5.31

Confidence level used: 0.95
```

```
plot(em_mod_lm)
```