Problem Set 10 Key

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

```
-- 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    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- 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
```

1 Problem 1

The dataset penguins preloaded in R includes data on the size and sex of adult penguins in the Palmer Archipelago. Suppose we are interested in testing for a correlation between body mass and flipper length.

```
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
$ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
$ bill_len    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, ~
$ bill_dep    <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, ~
```

```
$ flipper_len <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180,~
$ body_mass   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, ~
$ sex         <fct> male, female, female, NA, female, male, female, male, NA, ~
$ year        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

a. Estimate the Spearman correlation coefficient for the sample. [2 pts.]

```
ctest <- cor.test(penguins$body_mass,penguins$flipper_len,method = "spearman")
```

```
Warning in cor.test.default(penguins$body_mass, penguins$flipper_len, method =
"spearman"): Cannot compute exact p-value with ties
```

```
ctest
```

```
    Spearman's rank correlation rho

data:  penguins$body_mass and penguins$flipper_len
S = 1066875, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8399741
```

The Spearman correlation coefficient is 0.8399741.

b. Perform a hypothesis test that tests whether there is a correlation between body mass and flipper length based on the Spearman correlation coefficient. What is the p-value? [2pts.]

The p-value is <2.2e-16.

c. Is there evidence of a correlation between body mass and flipper length? Use a significance level of 0.01. [1 pt.]

We reject the null hypothesis. There is sufficient evidence of a correlation between body mass and flipper length.

2 Problem 2

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.

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

```
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. Use a nonparametric test 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? Is the alternative hypothesis one-sided or two-sided? [3pt.]

The alternative hypothesis is two-sided.

H_0 : The median of the paired difference of headache scores is zero./ The distribution of the paired difference of headache scores is centered at zero.

H_a : The median of the paired difference of headache scores is not zero./ The distribution of the paired difference of headache scores is notcentered at zero.

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

```
wtest <- wilcox.test(x=acupuncture$pk1,y=acupuncture$pk2,paired=T)
wtest
```

```
    Wilcoxon signed rank test with continuity correction

data:  acupuncture$pk1 and acupuncture$pk2
V = 12240, p-value = 4.799e-14
alternative hypothesis: true location shift is not equal to 0
```

The p-value is $4.7992427 \times 10^{-14}$.

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

We reject the null hypothesis at 0.001 level of significance. We have sufficient evidence that the average headache seveiry score differs at the 3-month follow-up compared to the baseline.

3 Problem 3

The data set in hflashes.csv contains data from a 14-year cohort study by Freeman et al (2001) that investigated the occurrence of hot flashes in 375 participants.

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

```
Rows: 375
Columns: 14
$ pt       <int> 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 23, 24~
$ ageg     <int> 2, 3, 1, 1, 2, 3, 2, 2, 2, 3, 2, 1, 1, 1, 1, 3, 2, 1, 1, 1, 2~
$ aagrp    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0~
$ edu      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1~
$ d1       <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0~
$ f1a      <int> 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1~
$ pcs12    <dbl> 56.80537, 59.18338, 57.73952, 55.83575, 55.89324, NA, 55.5009~
$ hotflash <int> 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0~
$ bmi30    <int> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ estra    <dbl> 106.710, 31.250, 13.410, 10.640, 24.060, 37.305, 26.320, 24.1~
$ fsh      <dbl> 3.005, 11.195, 14.545, 5.530, 9.780, 10.290, 7.960, 4.775, 7.~
$ lh       <dbl> 2.980, 5.760, 5.595, 2.260, 2.600, 3.395, 3.570, 2.095, 3.660~
$ testo    <dbl> 7.680, 11.930, 24.375, 8.280, 4.050, 8.275, 15.995, 12.340, 1~
$ dheas    <dbl> 61.225, 104.920, 117.450, 36.850, 11.165, 100.360, 76.780, 83~
```

a. Use a non-parametric test to determine if there is a difference in the distribution of baseline estradiol measurements (Variable estra) between current (Variable f1a=1) and non-current smokers (Variable f1a=0). [3 pts.]

We use the independent 2-sample Wilcoxon test, also known as the Mann-Whitney U test. The hypotheses are:

H_0 : The distribution/median of baseline estradiol measurements for current and non-current smokers are the same. H_a : The distribution/median of baseline estradiol measurements for current and non-current smokers are not the same.

```
wtest <- wilcox.test(estra~f1a,data=hflash)
wtest
```

```
    Wilcoxon rank sum test with continuity correction

data:  estra by f1a
W = 15390, p-value = 0.3415
alternative hypothesis: true location shift is not equal to 0
```

b. What is the resulting p-value? [1 pt.] The p-value is 0.341461.

c. Comment on the strength/sufficiency of evidence against the null hypothesis. [1pt]

We fail to reject the null hypothesis. There is insufficient evidence that the distribution/median of baseline estradiol measurements for current and non-current smokers are not the same.