Analysis of fruitfly data

This program reads data on fruit fly longevity. Find more information in the data dictionary.

This code was written by Steve Simon and Leroy Wheeler on 2024-10-29 and is placed in the public domain.

Load the tidyverse library

```
library(broom)
library(tidyverse)
```

For most of your programs, you should load the tidyverse library. The broom library converts your output to a nicely arranged dataframe. The messages and warnings are suppressed.

List the variable names

```
fn <- "https://jse.amstat.org/datasets/fruitfly.dat.txt"
vlist <- c(
  "id",
  "partners",
  "type",
  "longevity",
  "thorax",
  "sleep")
```

When a dataset does not have variables on the first line, you need to specify them in the code.

Read the data and view a brief summary

```
fly <- read_fwf(
  "../data/fruitfly.txt",
  col_types="nnnnnn",
```

```
  fwf_widths(
    widths=c(2, 2, 2, 3, 5, 3),
    col_names=vlist))
glimpse(fly)
```

```
Rows: 125
Columns: 6
$ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
$ partners  <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, …
$ type      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ longevity <dbl> 35, 37, 49, 46, 63, 39, 46, 56, 63, 65, 56, 65, 70, 63, 65, …
$ thorax    <dbl> 0.64, 0.68, 0.68, 0.72, 0.72, 0.76, 0.76, 0.76, 0.76, 0.76, …
$ sleep     <dbl> 22, 9, 49, 1, 23, 83, 23, 15, 9, 81, 12, 15, 37, 24, 26, 17,…
```

The fruitfly dataset has a fixed width format (fwf). You need to specify the columns that each variable uses.

Create cage groups

```
fly$cage <-
  case_when(
    fly$partners==0 & fly$type==9 ~ "No females",
    fly$partners==1 & fly$type==0 ~ "One pregnant female",
    fly$partners==1 & fly$type==1 ~ "One virgin female",
    fly$partners==8 & fly$type==0 ~ "Eight pregnant females",
    fly$partners==8 & fly$type==1 ~ "Eight virgin females")
```

The five categories represent different combinations of partners and type.

Question 1: Review the fruitfly analysis discussed in this module. There is a second variable, sleep, that might be influenced by the presence or absence of virgin or pregnant females. Compute descriptive statistics for sleep levels in each of the five groups. Interpret these statistics

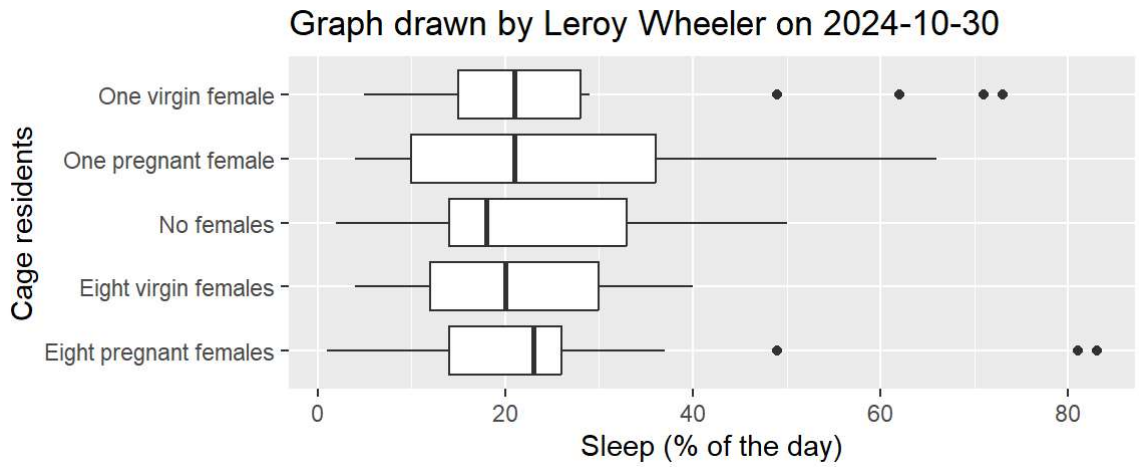
```
fly |>
  group_by(cage) |>
  summarize(
    sleep_mn=mean(sleep),
    sleep_sd=sd(sleep),
    n=n())
```

```
# A tibble: 5 × 4
  cage                   sleep_mn sleep_sd     n
  <chr>                     <dbl>    <dbl> <int>
1 Eight pregnant females     25.2     19.8    25
2 Eight virgin females       20.8     10.7    25
3 No females                 21.6     12.5    25
4 One pregnant female        24.1     16.7    25
5 One virgin female          25.8     18.4    25
```

The mean sleep time is similar among all groups but the cage containing eight virgin females might be a little low compared to the others. The standard deviations are consistent across all groups and off by no more than a factor of 2.

Question 2: Draw a boxplot for sleep levels in each group. Interpret the boxplots.

```
fly |>
  ggplot(aes(cage, sleep)) +
    geom_boxplot() +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-30") +
    xlab("Cage residents") +
    ylab("Sleep (% of the day)") +
    coord_flip()
```



The boxplot also shows distribution of sleep time of the males in the different cages. The distributions vary of sleep time varies quite a bit in the different groups and there are some cages with significant outliers.

Question 3: Based on the previous two questions, do you believe that the assumptions of analysis of variance are met. Proceed with all of the remaining questions regardless of your conclusion here.

The assumptions of ANOVA require normal distribution, equal variance and independence of sample data. From the box plots we see some deviance from perfect normality but we see no more than a 2-fold difference in standard deviations between any two groups. Outliers to the right suggest that a log transformation before conducting ANOVA might be helpful. We will also assume independence between the samples.

Question 4: Conduct a single factor analysis of variance, using sleep as the dependent variable and cage as the categorical predictor variable. Print an analysis of variance table. Interpret the F-ratio and the p-value.

```r
m1 <- aov(sleep ~ cage, data=fly)
tidy(m1)
```

```
# A tibble: 2 × 6
  term         df  sumsq meansq statistic p.value
  <chr>     <dbl>  <dbl>  <dbl>     <dbl>   <dbl>
1 cage          4   487.   122.     0.474   0.755
2 Residuals   120 30778.   256.    NA      NA
```

The F-ratio is less than one and the p-value is greater than our alpha of 0.05. We conclude that there is no difference among the population mean sleep time.

Question 5: Calculate and interpret confidence intervals using the Tukey post hoc comparisons. Which intervals include 0 and which do not. Provide a general conclusion about which groups, if any, differ from one another.

```r
t3 <- TukeyHSD(m1, ordered=TRUE)
t3
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = sleep ~ cage, data = fly)

$cage
                                           diff        lwr      upr     p adj
No females-Eight virgin females            0.80 -11.746125 13.34613 0.9997793
One pregnant female-Eight virgin females   3.32  -9.226125 15.86613 0.9484003
Eight pregnant females-Eight virgin females 4.40  -8.146125 16.94613 0.8675467
One virgin female-Eight virgin females     5.00  -7.546125 17.54613 0.8042420
One pregnant female-No females             2.52 -10.026125 15.06613 0.9809592
Eight pregnant females-No females          3.60  -8.946125 16.14613 0.9316881
```

```
One virgin female-No females                4.20  -8.346125 16.74613 0.8858467
Eight pregnant females-One pregnant female  1.08 -11.466125 13.62613 0.9992758
One virgin female-One pregnant female       1.68 -10.866125 14.22613 0.9959201
One virgin female-Eight pregnant females    0.60 -11.946125 13.14613 0.9999297
```

All 95% confidence intervals have a lower value that is negative and an upper value which is positive which means that a zero value in the difference between each group is a distinct possibility. Additionally all the p values are large. With this information and from the previous ANOVA test, we conclude that all groups have similar sleep times.

Question 6: Conduct a Kruskal-Wallis test. Interpret your results.

```r
kruskal.test(sleep ~ cage, data=fly)
```

```
    Kruskal-Wallis rank sum test

data:  sleep by cage
Kruskal-Wallis chi-squared = 0.34861, df = 4, p-value = 0.9865
```

The chi-squared value is not greater than the df = 4 value and the p value is greater than our alpha =0.05 so we conclude there is no difference in sleep time using the Kruskal-Wallis rank sum test.