Analysis of fruitfly data

AUTHOR
Michael Dang

PUBLISHED
November 6, 2024

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

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, 8, …
$ type      <dbl> 0, 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())
```

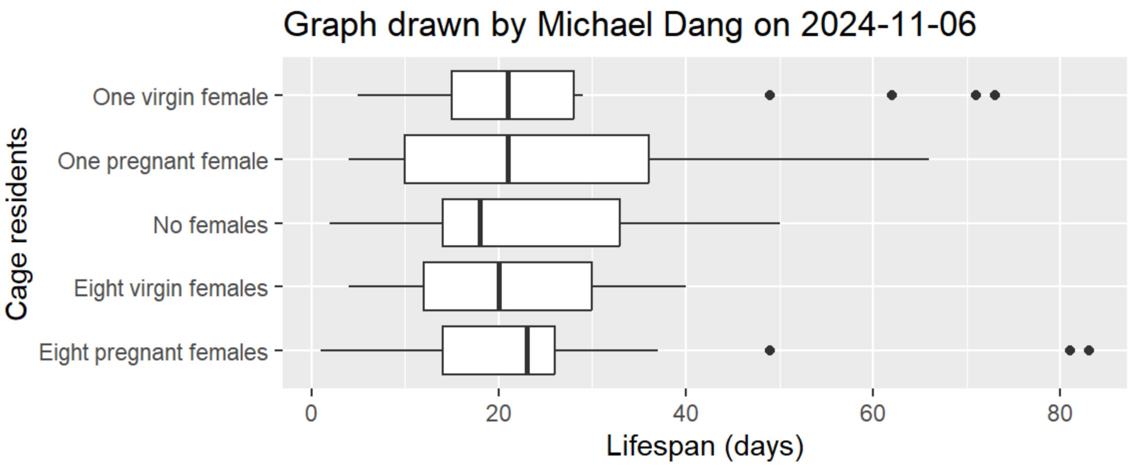
```
# A tibble: 5 × 4
  cage                   sleep_mn sleep_sd     n
  <chr>                     <dbl>    <dbl> <int>
1 Eight pregnant females     25.2     19.8    25
2 Eight virgin females       20.8     10.7    25
3 No females                 21.6     12.5    25
4 One pregnant female        24.1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

The mean sleep duration is much lower for the group with eight virgin females. Standard deviations vary but are generally consistent, indicating similar levels of variability across all groups.

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 Michael Dang on 2024-11-06") +
    xlab("Cage residents") +
    ylab("Lifespan (days)") +
    coord_flip()
```



The boxplot shows a left-skewed with many outliers

Question 4

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.

Answer:

- Based on the box plot, seem like the assumption of normality is violate because “One virgin female” and “Eight pregnant females” have outliers and high degree of variability.
- Also the standard deviation across groups are not consistent, hence homogenity may get violated.
- Assume sample were collected iid hence independence is met.

Therefore, assumption of ANOVA are not met due to lack of normality and homogenity.

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.

```
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 small and the p-value is large. Conclude that there is not statistical difference among some or all of the population mean lifespans.

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.

```
t1 <- TukeyHSD(m1, order = TRUE)
t1
```

```
  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
```

Since all the interval contains 0, we conclude that there are no significant differences in sleep levels between any of the cage groups.

Question 6

Conduct a Kruskal-Wallis test. Interpret your results.

```
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 Kruskal-Wallis chi-squared: 0.35, is a very low test statistics. Also the p-value is 0.98 indicates that there is no statistically significant difference in sleep levels among the different cage group.