Analysis of fruitfly data

AUTHOR
Michael Dang

PUBLISHED
November 13, 2024

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

Load the tidyverse library

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

Comments on the code

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

```
vlist <- c(
  "id",
  "partners",
  "type",
  "longevity",
  "thorax",
  "sleep")
```

Comments on the code

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,…
```

Comments on the code

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

Notice that the two categorical variables, partners and type, are actually numbers rather than strings. To avoid having R treat these variables as if they were continuous, use the factor function in some of the code below.

Question 1

Create a subset of the fruitfly data by removing the age where type equals 9. Draw a clustered boxplot with sleep as the outcome and partners and type as the categorical predictors. Interpret this graph. Is there evidence of non-normality?

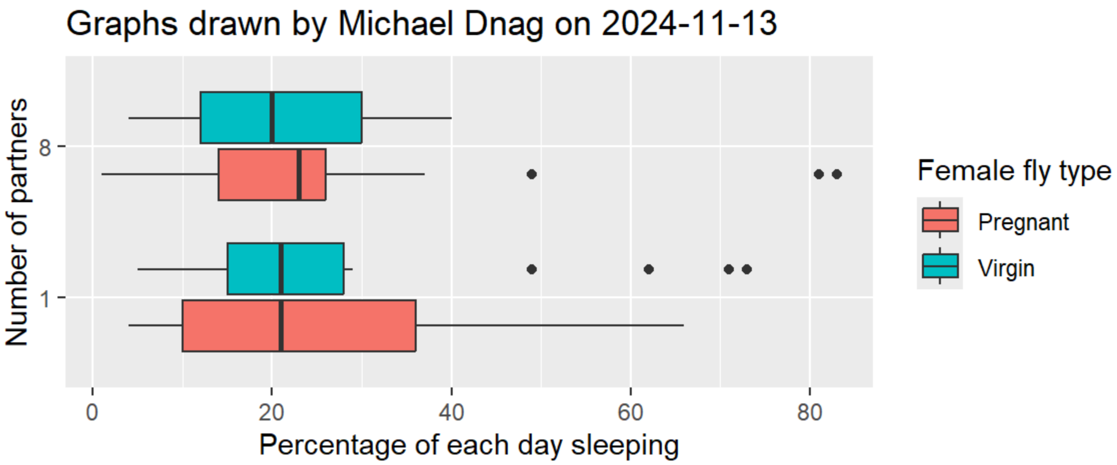
```
fly |>
  filter(type != 9) -> fly_subset
```

Comments on the code

If you exclude the pure control group (No females), you can analyze the two factors, partners and type individually. Partners has two category levels, 1 for when one female was included in the cage and 8 for when eight females were included in the cage. Type also has two category levels, 0 for pregnant female fly/flies and 1 for virgin fly/flies. A male fly will not mate with a pregnant females, so you can think of this as a second level of controls. The two factors are crossed, meaning that every possible combination of partners and type has outcomes measured.

Draw boxplot of longevity by partners and type

```
fly_subset |>
  ggplot(aes(factor(partners), sleep, fill=factor(type))) +
    geom_boxplot() +
    xlab("Number of partners") +
    ylab("Percentage of each day sleeping") +
    ggtitle("Graphs drawn by Michael Dnag on 2024-11-13") +
    labs(fill="Female fly type") +
    scale_fill_discrete(labels=c("Pregnant", "Virgin")) +
    coord_flip()
```



“Virgin” flies tend to sleep a higher median percentage of the day, with greater variability, while “Pregnant” flies show a narrower range and lower median. The data shows evidence of non-normality due to outliers and asymmetrical distributions in both groups.

Question 2

Calculate descriptive statistics for sleep (mean, standard deviation, and sample size) by the combination of the two categorical predictors, partners and type. Is there evidence of heterogeneity?

```
fly_subset |>
  group_by(type, partners) |>
  summarize(
    sleep_mn=mean(sleep),
    sleep_sd=sd(sleep),
    n=n()) -> fly_means
fly_means
```

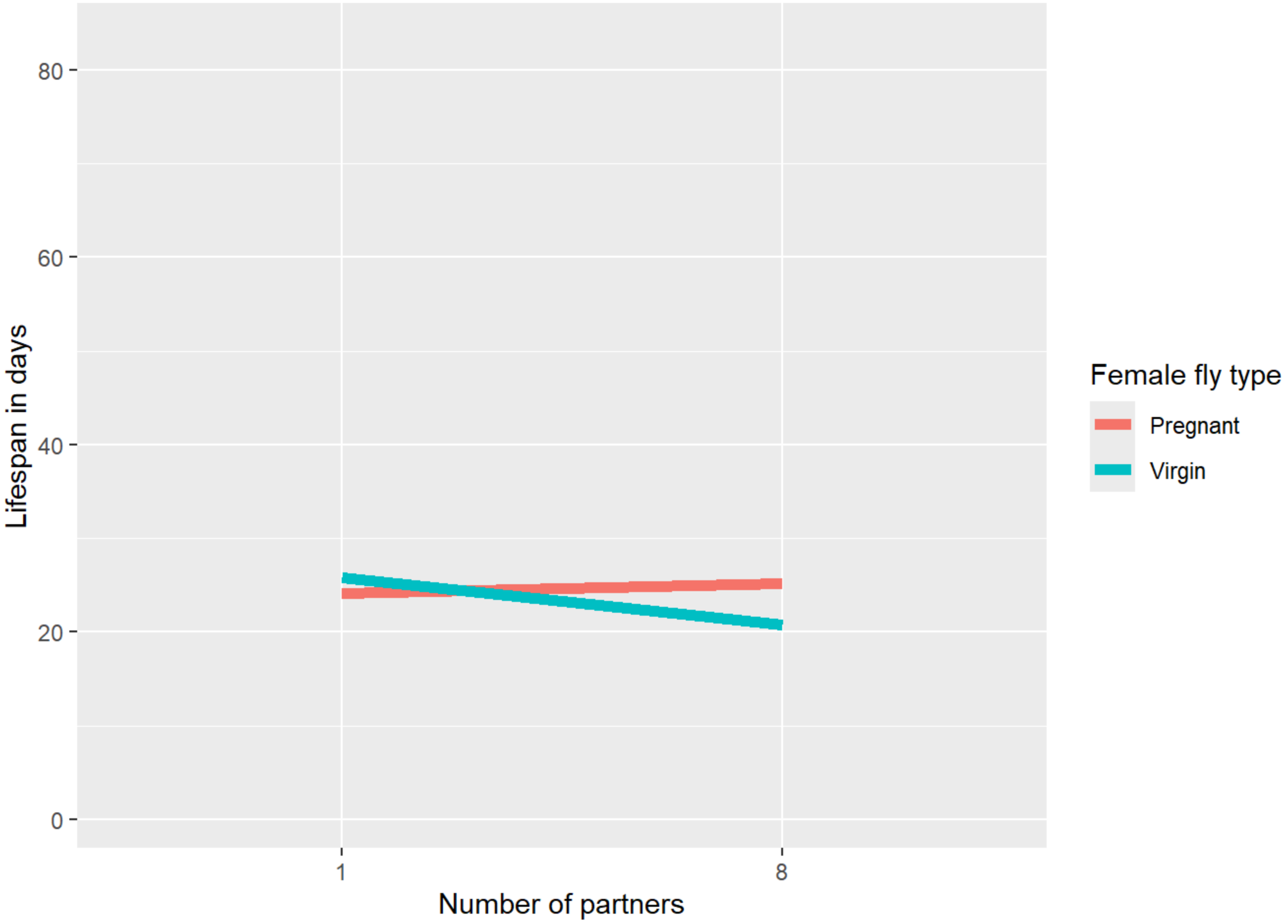
```
# A tibble: 4 × 5
# Groups:   type [2]
   type partners sleep_mn sleep_sd     n
  <dbl>    <dbl>    <dbl>    <dbl> <int>
1     0        1     24.1     16.7    25
2     0        8     25.2     19.8    25
3     1        1     25.8     18.4    25
4     1        8     20.8     10.7    25
```

Yes, there is evidence of heterogeneity in this data. The standard deviations for sleep percentage vary notably between groups, with values ranging from 10.74 to 19.83.

Question 3

Draw a line graph for the mean sleep levels compared by type and partners. Is there evidence of an interaction?

```
fly_means |>
  ggplot(aes(
    factor(partners),
    sleep_mn,
    group=factor(type),
    color=factor(type))) +
    geom_line(linewidth=2) +
    expand_limits(y=range(fly_subset$sleep)) +
  xlab("Number of partners") +
  ylab("Lifespan in days") +
  labs(color="Female fly type") +
  scale_color_discrete(labels=c("Pregnant", "Virgin"))
```



There is little evidence of a strong interaction between the number of partners and female fly type on lifespan. Both “Pregnant” and “Virgin” flies exhibit nearly parallel lines with a slight decrease in lifespan as the number of partners increases.

Question 4

Analyze the sleep variable using a two factor analysis of variance with an interaction. Present and interpret the analysis of variance table.

```
m1 <- aov(sleep ~ factor(partners)*factor(type), data=fly_subset)
anova(m1)
```

```
Analysis of Variance Table

Response: sleep
                              Df  Sum Sq Mean Sq F value Pr(>F)
factor(partners)               1    96.0   96.04  0.3408 0.5607
factor(type)                   1    46.2   46.24  0.1641 0.6863
factor(partners):factor(type)  1   231.0  231.04  0.8198 0.3675
Residuals                     96 27054.3  281.82
```

There is no evidence for a statistically significant interaction between the number of female partners and the type of partners (pregnant or virgin) nor their interaction significantly affects the amount of sleep. This imply no substantial impact of these factors on sleep variability.

Question 5

What factors might make you consider using a log transformation for the sleep variable? Do not run such an analysis but tell us whether you think the data would warrant such a transformation?

Ans:

- The factor that make me consider using a log transformation for sleep variable is the data distribution is skewed or if there are large differences in variability (heteroscedasticity) between groups. In the provided data, there are instances of both high and low sleep values with varying standard deviations across groups.
- I think the data should consider a log transformation, because the data is left-skewed and has extreme outliers based on the plot on Question 1.