MEDB 5501, Module11

2024-10-30

Topics to be covered

- What you will learn
  - Analysis of variance is linear regression
  - Log transformation
  - Kruskal-Wallis test
  - R code for analysis of variance
  - Your homework

Indicator variables, 1

- Two levels is less complex with three or more levels
  - R assigns 0 to first category (alphabetically)
  - R assigns 1 to second category
- Interpretation
  - Intercept is estimated average outcome for first category
  - Slope is the estimated average change
    - Second category average minus first category average

Indicator variables, 2

- With k levels, you need k-1 indicators
  - First indicator
    - R assigns 1 to second category (alphabetically)
    - R assigns 0 to all other categories
  - Second indicator
    - R assigns 1 to third category (alphabetically)
    - R assigns 0 to all other categories
  - And so on

Indicator variables, 3

- Interpretation
  - Intercept is estimated average outcome for first category
  - First slope is the estimated average change
    - Second category average minus first category average
  - Second slope is the estimated average change
    - Third category average minus first category average
  - And so on

Example using fruitfly lifespans, 1

- Experiment with 125 cages
  - Does fruitfly mating affect average male lifespan?
  - Isolate a male fruitfly with
    - virgin females,
    - pregnant females, or
    - no females
- Cages 1-25 have one male fruitfly, no female fruit flies
  - Set partners=0, type=9

Example using fruitfly lifespans, 2

- Cages 26-50 have one male fruitfly, one pregnant female
  - Set partners=1, type=0
  - Males will not mate with pregnant females
- Cages 51-75 have one male fruitfly, one virgin female
  - Set partners=1, type=1
- Cages 76-100 have one male fruitfly, eight pregnant females
  - Set partners=8, type=0
- Cages 101-125 have one male fruitfly, eight virgin females
  - Set partners=8, type=1

Listing of fruitfly.yaml, 1

```
---
data_dictionary: fruitfly.dat.txt

copyright: >
  This dataset is copyrighted by the authors of the Journal of Statistics
  Education article, but should be available for individual educational uses
  under the Fair Use provisions of copyright law.

description: >
  Does access to mating affect the lifespan of fruitflies? This data shows the
  longevity of male fruitflies in the presence or absence of female fruitflies
  to mate with. Male fruitflies were housed with 0, 1, or 8 females. In some
  groups, the females were pregnant and thus not available for mating. There
are
  two covariates, length of the thorax and percentage of time sleeping, that
  might also influence longevity.
```

Listing of fruitfly.yaml, 2

```
source: >
  Partridge, L., and Farquhar, M. (1981), "Sexual Activity and the Lifespan of
  Male Fruitflies," Nature, 294, 580-581.

  James A. Hanley & Stanley H. Shapiro. Sexual Activity and the Lifespan of
Male
  Fruitflies: A Dataset That Gets Attention. Journal of Statistics Education
  v.2, n.1 (1994)

additional_description:
  https://jse.amstat.org/datasets/fruitfly.txt

download_url:
  https://jse.amstat.org/datasets/fruitfly.dat.txt
```

Listing of fruitfly.yaml, 3

```
format:
  fixed width

varnames:
  not included

missing_value_code:
  not needed

size:
  rows: 125
  columns: 6
```

Listing of fruitfly.yaml, 4

```
vars:
  id:
    columns: 1-2

  partners:
    columns: 4
    label: Number of female partners
    values: 0, 1, or 8
```

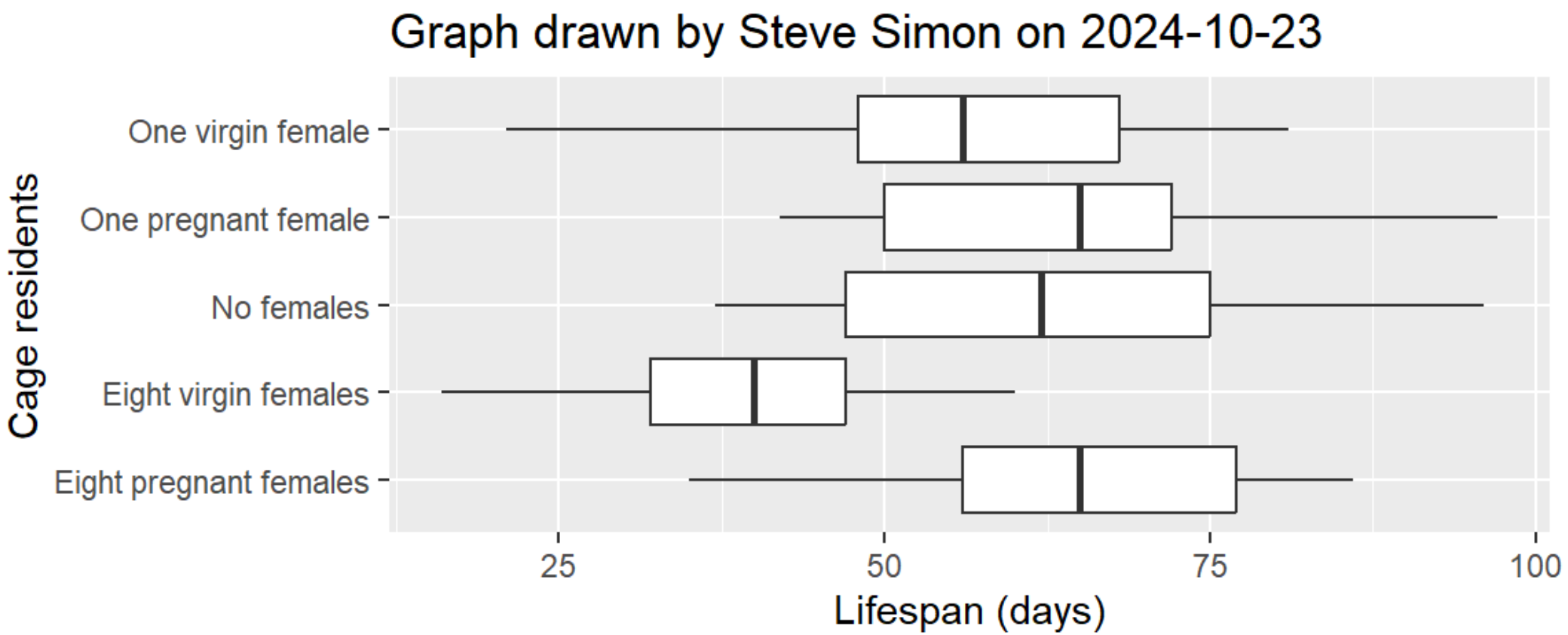
Listing of fruitfly.yaml, 5

```
type:
  columns: 6
  label: Type of female fruitfly
  values:
    0: newly pregnant female
    1: virgin female
    9: not applicable (when partners=0)

longevity:
  columns: 8-9
  label: Lifespan
  unit: days
  range: positive
```

Listing of fruitfly.yaml, 6

```
  thorax:
    columns: 11-14
    label: Length of thorax
    unit: mm
    range: positive

  sleep:
    columns: 16-17
    label: Percentage of each day sleeping
    range: 0 to 100
---
```

Fruitfly lifespan boxplots



Fruitfly lifespan group means

```
# A tibble: 5 × 4
  cage                  longevity_mn longevity_sd     n
  <chr>                        <dbl>        <dbl> <int>
1 Eight pregnant females        63.4         14.5    25
2 Eight virgin females          38.7         12.1    25
3 No females                    63.6         16.5    25
4 One pregnant female           64.8         15.7    25
5 One virgin female             56.8         14.9    25
```

Fruitfly lifespan analysis using aov

```r
m1 <- aov(longevity ~ cage, data=fly)
anova(m1)
```

```
Analysis of Variance Table

Response: longevity
           Df Sum Sq Mean Sq F value    Pr(>F)
cage        4  11939 2984.82  13.612 3.516e-09 ***
Residuals 120  26314  219.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fruitfly lifespan analysis using lm, 1

```r
m2 <- lm(longevity ~ cage, data=fly)
anova(m2)
```

```
Analysis of Variance Table

Response: longevity
           Df Sum Sq Mean Sq F value    Pr(>F)
cage        4  11939 2984.82  13.612 3.516e-09 ***
Residuals 120  26314  219.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fruitfly lifespan analysis using lm, 2

```
1 m2 <- lm(longevity ~ cage, data=fly)
2 tidy(m2)
```

```
# A tibble: 5 × 5
  term                      estimate std.error statistic  p.value
  <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                  63.4       2.96   21.4    9.12e-43
2 cageEight virgin females    -24.6       4.19   -5.88   3.73e- 8
3 cageNo females                0.200     4.19    0.0478 9.62e- 1
4 cageOne pregnant female       1.44      4.19    0.344  7.32e- 1
5 cageOne virgin female        -6.60      4.19   -1.58   1.18e- 1
```

Fruitfly lifespan analysis using lm, 3

```
# A tibble: 5 × 4
  cage                  longevity_mn reference_mn mean_difference
  <chr>                        <dbl>        <dbl>           <dbl>
1 Eight pregnant females        63.4         63.4            0
2 Eight virgin females          38.7         63.4          -24.6
3 No females                    63.6         63.4            0.200
4 One pregnant female           64.8         63.4            1.44
5 One virgin female             56.8         63.4           -6.6
```

Break #1

- What you have learned
  - Analysis of variance is linear regression
- What’s coming next
  - Log transformation

Analysis of variance model

- Sample 1: $Y_{11}, Y_{12}, \cdots, Y_{1n_1}$ are $N(\mu_1, \sigma)$
- Sample 2: $Y_{21}, Y_{22}, \cdots, Y_{2n_2}$ are $N(\mu_2, \sigma)$
- $\cdots$
- Sample k: $Y_{k1}, Y_{k2}, \cdots, Y_{kn_k}$ are $N(\mu_k, \sigma)$
- $H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1: \ \mu_i \neq \mu_j$ for some i, j

Violation of assumptions

- Non-normality
- Heterogeneity
- Lack of independence

When to consider a log transformation

- Only positive values
- Max/min > 3
- Skewed distribution
- Groups with larger means have more variation

Log transformation, 1

```
Analysis of Variance Table

Response: log_longevity
           Df  Sum Sq  Mean Sq F value    Pr(>F)
cage        4 0.97717 0.244293  15.846 1.935e-10 ***
Residuals 120 1.85004 0.015417
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

Although there are no problems with heterogeneity or non-normality with this particular dataset, here is an illustration of how to use a log transformation with analysis of variance.

Log transformation, 2

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

Fit: aov(formula = log_longevity ~ cage, data = log_fly)

$cage
                                                      diff         lwr
upr
One virgin female-Eight virgin females      0.1727239664  0.07545440
0.26999353
No females-Eight virgin females             0.2246277842  0.12735822
0.32189735
Eight pregnant females-Eight virgin females 0.2248176039  0.12754804
0.32208717
One pregnant female-Eight virgin females    0.2340081450  0.13673859
```

Speaker notes

Here are the results of the Tukey post hoc tests on the log scale. These intervals are difficult to interpret.

Log transformation, 3

	diff	lwr	upr
One virgin female-Eight virgin females	1.488415	1.1897464	1.862059
No females-Eight virgin females	1.677366	1.3407821	2.098444
Eight pregnant females-Eight virgin females	1.678099	1.3413683	2.099361
One pregnant female-Eight virgin females	1.717150	1.3725829	2.148215
No females-One virgin female	1.126948	0.9008122	1.409852
Eight pregnant females-One virgin female	1.127441	0.9012060	1.410468
One pregnant female-One virgin female	1.153677	0.9221777	1.443290
Eight pregnant females-No females	1.000437	0.7996874	1.251582
One pregnant female-No females	1.023718	0.8182967	1.280707
One pregnant female-Eight pregnant females	1.023271	0.8179391	1.280148

Speaker notes

Here are the results translated back to the original scale. The confidence intervals are intervals for the ratio of two geometric means.

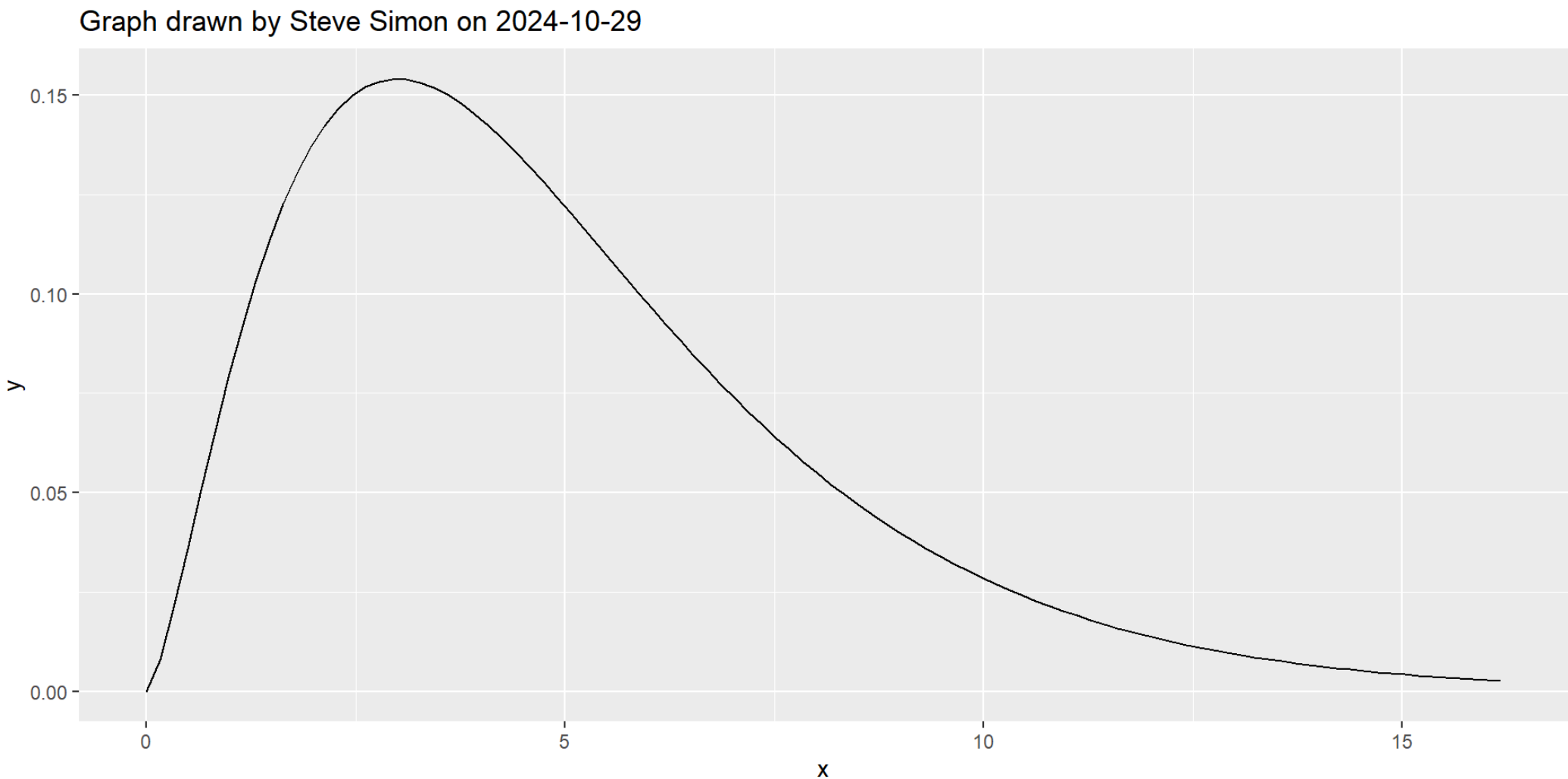
Break #2

- What you have learned
  - Log transformation
- What’s coming next
  - Kruskal-Wallis test

The Chi-squared distribution

- Often denoted as χ^2_{df}
- Has a single parameter, degrees of freedom
  - Never negative
  - Skewed right
  - Mean equals degrees of freedom
- Calculations in R
  - pchisq(x, df) = $P[\chi^2_{df} < x]$
  - qchisq(p, df) = p^{th} quantile, $\chi^2_{df,p}$

The Chi-squared distribution



Speaker notes

This is a graph of the Chi-squared distribution with 5 degrees of freedom.

Computing the Kruskal-Wallis test

- Rank the observations, $R(X_{ij})$
  - 1 for smallest, 2 for second smallest, etc.
  - Compute the average rank in each group, $\bar{R}_i$
  - Compute the overall rank, $\bar{R}$
  - T = $(N-1)\frac{\Sigma n_i(\bar{R}_i - \bar{R})^2}{\Sigma\Sigma(R(X_{ij}) - \bar{R})^2}$

Speaker notes

The Kruskal-Wallis test is similar to the Mann-Whitney-Wilcoxon test. You rank the data from low to high, calculate an average rank in each group. Look at how much deviation the group rank averages are from the overall rank averages.

Decision rule for Kruskal-Wallis test, 1

- Accept H_0 if T <- $\chi^2_{df,1-\alpha}$
  - df = k-1
- Accept H_0 if p-value > α
  - p-value = $P[\chi^2_{df} > T]$

Decision rule for Kruskal-Wallis test, 1

- Null hypothesis is difficult to define
  - Does not involve population means
  - Some claim involvement of population medians
  - Stochastic dominance
    - $P[X_{aj} > X_{bj}] > 0.5$ for some a and b.

Application of Kruskal-Wallis test to fruitfly longevity

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

```
    Kruskal-Wallis rank sum test

data:  longevity by cage
Kruskal-Wallis chi-squared = 37.961, df = 4, p-value = 1.142e-07
```

Speaker notes

There a five groups, so four degrees of freedom. The test statistic is much larger than the degrees of freedom and the p-value is small. Reject the null hypothesis and conclude that there are differences between at least two of the five groups.

Break #3

- What you have learned
  - Kruskal-Wallis test
- What’s coming next
  - R code for analysis of variance

Listing of simon-5501-11-fruitfly.qmd, 1

```
---
title: "Analysis of fruitfly data"
format:
  html:
    embed-resources: true
---

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

[dd]: https://github.com/pmean/data/blob/master/files/fruitfly.yaml

This code was written by Steve Simon on 2024-10-23 and is placed in the public
domain.
```

Listing of simon-5501-11-fruitfly.qmd, 2

````
## Load the tidyverse library

```{r setup}
#| message: false
#| warning: false
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.
````

Listing of simon-5501-11-fruitfly.qmd, 3

````
## List the variable names

```{r variable-list}
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.
````

Listing of simon-5501-11-fruitfly.qmd, 4

````
## Read the data and view a brief summary

```{r read}
fly <- read_fwf(
  "../data/fruitfly.txt",
  col_types="nnnnnn",
  fwf_widths(
    widths=c(2, 2, 2, 3, 5, 3),
    col_names=vlist))
glimpse(fly)
```

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

Listing of simon-5501-11-fruitfly.qmd, 5

````
## Create cage groups

```{r cage}
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.
````

Listing of simon-5501-11-fruitfly.qmd, 6

````
## Calculate descriptive statistics

```{r longevity-means}
fly |>
  group_by(cage) |>
  summarize(
    longevity_mn=mean(longevity),
    longevity_sd=sd(longevity),
    n=n())
```

The mean lifespan is much lower for the eight virgin females group. The
standard
deviations are reasonably small and more or less consistent across all groups.
````

Listing of simon-5501-11-fruitfly.qmd, 7

````
## Draw boxplot

```{r longevity-boxplot}
#| fig.width: 6
#| fig.height: 2.5
fly |>
  ggplot(aes(cage, longevity)) +
    geom_boxplot() +
    ggtitle("Graph drawn by Steve Simon on 2024-10-23") +
    xlab("Cage residents") +
    ylab("Lifespan (days)") +
    coord_flip()
```

The boxplot shows a roughly normal distribution with no outliers.
````

Listing of simon-5501-11-fruitfly.qmd, 8

````
## One factor analysis of variance for longevity

```{r longevity-one-factor-anova}
m1 <- aov(longevity ~ cage, data=fly)
tidy(m1)
```

The F-ratio is large and the p-value is small. Conclude that there is a
difference among some or all of the population mean lifespans.
````

Listing of simon-5501-11-fruitfly.qmd, 9

```
## Linear model for longevity, 1

```{r longevity-lm-1}
m2 <- lm(longevity ~ cage, data=fly)
anova(m2)
```

You can use linear regression to reach the same conclusion. The sums of squares,
degrees of freedom, F-ratio, and p-value all match.
```

Listing of simon-5501-11-fruitfly.qmd, 10

````
## Linear model for longevity, 2

```{r longevity-lm-2}
tidy(m2)
```

The linear model creates indicator variable for four out of the five category levels. The regression slopes represent the estimated average difference in means between each category and the reference category.
````

Listing of simon-5501-11-fruitfly.qmd, 11

````
## Linear model for longevity, 3

```{r longevity-lm-3}
fly |>
  group_by(cage) |>
  summarize(longevity_mn=mean(longevity)) |>
  mutate(reference_mn=first(longevity_mn)) |>
  mutate(mean_difference=longevity_mn-reference_mn)
```

This table shows how to interpret the intercept and the four slope terms. The
intercept is the estimated average lifespan for the first level of the
categorical variable. The slope terms are the estimated average changes from
the
other levels of the categorical variable and the first level.
````

Listing of simon-5501-11-fruitfly.qmd, 12

````
## Re-order cage groups, 1

```{r re-order-1}
fly$cage_1 <-
  factor(
    fly$cage,
    levels = c(
      "No females",
      "One pregnant female",
      "Eight pregnant females",
      "One virgin female",
      "Eight virgin females"))
```

Compare every category to the No females category (the control condition).
````

Listing of simon-5501-11-fruitfly.qmd, 13

```
## Re-order cage groups, 2

```{r re-order-2}
m3 <- lm(longevity ~ cage_1, data=fly)
tidy(m3)
```

There is one group, eight virgin females, that is statistically significantly different from the no female control group.
```

Listing of simon-5501-11-fruitfly.qmd, 14

````
## Log transformation, 1

```{r log-longevity-anova-1}
fly |>
  mutate(log_longevity=log10(longevity)) -> log_fly
m3 <- aov(log_longevity ~ cage, data=log_fly)
tidy(m3)
```

Although there are no problems with heterogeneity or non-normality, here is an illustration of how to use a log transformation with analysis of variance.
````

Listing of simon-5501-11-fruitfly.qmd, 15

```
## Log transformation, 2

```{r log-longevity-anova-2}
t3 <- TukeyHSD(m3, ordered=TRUE)
t3
```
```

Listing of simon-5501-11-fruitfly.qmd, 16

````
## Log transformation, 3

```{r log-longevity-anova-3}
t3$cage |>
  data.frame() |>
  select(diff, lwr, upr) |>
  mutate(
    diff=10^diff,
    lwr=10^lwr,
    upr=10^upr)
```
````

Listing of simon-5501-11-fruitfly.qmd, 17

````
## Log transformation, 4

```{r log-longevity-anova-4}
t3$cage |>
  data.frame() |>
  select(diff, lwr, upr) |>
  mutate(
    diff=10^(-diff),
    upr=10^(-lwr),
    lwr=10^(-upr))
```
````

Listing of simon-5501-11-fruitfly.qmd, 18

```
## Kruskal-Wallis test

```{r kw}
kruskal.test(longevity ~ cage, data=fly)
```

It may not be needed with this particular dataset, but this is an illustration of how to use the Kruskal-Wallis test.
```

Listing of simon-5501-11-fruitfly.qmd, 19

````
## Save important files for later use

```{r save}
save(
  fly,
  log_fly,
  file="../data/fruitfly.RData")
```
````

Break #4

- What you have learned
  - R code for analysis of variance
- What’s coming next
  - Your homework

Listing of simon-5501-11-directions.md, 1

```
---
title: "Directions for 5501-11 programming assignment"
---

This programming assignment was written by Steve Simon on 2024-10-08 and is
placed in the public domain.
```

Listing of simon-5501-11-directions.md, 2

```
## Program

-   Download the [program][tem]
    -   Store it in your src folder
-   Modify the file name
    -   Use your last name instead of "simon"
-   Modify the documentation header
    -   Add your name to the author field
    -   Optional: change the copyright statement

[tem]: https://github.com/pmean/classes/blob/master/biostats-1/11/src/simon-5501-11-fruitfly.qmd
```

Listing of simon-5501-11-directions.md, 3

```
## Data

-   Download the [data][dat] file
    -   Store it in your data folder
-   Refer to the [data dictionary][dic], if needed.

[dat]: https://github.com/pmean/data/blob/main/files/fruitfly.txt
[dic]: https://github.com/pmean/data/blob/main/files/fruitfly.yaml
```

Listing of simon-5501-11-directions.md, 4

```
## 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

## Question 2

Draw a boxplot for sleep levels in each group. Interpret the boxplots.
```

Listing of simon-5501-11-directions.md, 5

```
## 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.

## 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.
```

Listing of simon-5501-11-directions.md, 6

```
## 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.

## Question 6

Conduct a Kruskal-Wallis test. Interpret your results.
```

Listing of simon-5501-11-directions.md, 7

```
## Your submission

-   Save the output in html format
    -   Make sure that you include your name on all graphs
    -   Write interpretations that match your analysis, not the original
analysis
-   Convert the html file to pdf format.
-   Make sure that the pdf file includes
    -   Your last name
    -   The number of this course
    -   The number of this module
-   Upload the file
```

Listing of simon-5501-11-directions.md, 8

```
## If it doesn't work

Please review the [suggestions if you encounter an error page][sim3].

[sim3]: https://github.com/pmean/classes/blob/master/general/suggestions-if-you-encounter-an-error.md
```

Summary

- What you have learned
  - Analysis of variance is linear regression
  - Log transformation
  - Kruskal-Wallis test
  - R code for analysis of variance
  - Your homework