MEDB 5501, Module11

2025-11-04

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
  - Examples:
    - Female=0, Male=1
    - Man=0, Woman=1
- Interpretation
  - Intercept is estimated average outcome for first category
  - Slope is the estimated average change
    - Second category average minus first category average

Speaker notes

Back in the previous modules, you saw a relationship between creating an indicator variable for a linear regression model and running a two-sample t-test. Just to review, when R sees a string that represents a categorical variable, it assigns the value of 0 to the level that appears first in alphabetical order and 1 to the level that appears last in alphabetical order. So a variable that has strings “male” and “female”, R will assign 0 to “female” and 1 to “male”. If the string were “man” and “woman”, R would assign 0 to “man” and 1 to “woman”.

The intercept represents the estimated average value of Y for the first or zero category. The slope represents the estimated average change in Y when you switch from the 0 category to the 1 category.

Example: Turtle experiment, 1

Sex	Fed	Fasted10	Fasted20
Male	42.8	42.4	38.9
Male	43.1	42.2	40.3
Male	40.4	40.8	37.5
Male	46.6	45.9	42.9
Female	42.2	42.4	39.7
Female	38.7	38.1	35.8
Female	35.3	34.3	32.3
Female	40.5	40.1	37.3

Speaker notes

Here is some data from a study of protein levels in turtles that are fed regularly and then subjected to short fasting periods. How would R assign an indicator variable for this data?

Turtle experiment, 2

Sex	Fed	Fasted10	Fasted20	Indicator
Male	42.8	42.4	38.9	1
Male	43.1	42.2	40.3	1
Male	40.4	40.8	37.5	1
Male	46.6	45.9	42.9	1
Female	42.2	42.4	39.7	0
Female	38.7	38.1	35.8	0
Female	35.3	34.3	32.3	0
Female	40.5	40.1	37.3	0

Speaker notes

R will assign 0 to the category level that appears first alphabetically, which is “Female”. It assigns 1 to the category level that appears second alphabetically, which is “Male”.

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
- Note: the first category in alphabetical order is zero for all indicator variables.

Speaker notes

If you have a categorical variable with more than 2 levels, you need more than 1 indicator. The number of indicators is always one less than the number of levels. So a category with 3 levels needs two indicators, a category with 6 levels needs 5 indicators.

How this is done in R differs from how it is done in SAS or SPSS. R looks at the alphabetical order. The first indicator is equal to 1 for the SECOND category level in alphabetical order. The next indicator is equal to 1 for the THIRD category level in alphabetical order. And so forth.

The way R defines things the first category level in alphabetical order is zero for each of the k-1 indicators.

Example with dietary cracker data

Cracker	Digested
control	1772.84
bran	1752.63
combo	2121.97
gum	2558.61
gum	2026.91
bran	2047.42
combo	2254.75
control	2353.21
combo	2153.36
gum	2331.19
bran	2547.77
...	

Speaker notes

This is a partial listing of a study of digested calorie counts for four differnt types of crackers: control, bran, combo, and gum. How would R assign indicator variables for this data?

With four levels, you need three indicator variables.

Dietary cracker data, first indicator variable

Cracker	Digested	i1
control	1772.84	0
bran	1752.63	0
combo	2121.97	1
gum	2558.61	0
gum	2026.91	0
bran	2047.42	0
combo	2254.75	1
control	2353.21	0
combo	2153.36	1
gum	2331.19	0
bran	2547.77	0

...

Speaker notes

If you put the category levels in alphabetical order (bran, combo, control, gum), then the second category level is “combo”. The first indicator variable is 1 for “combo” and 0 for the other levels.

Dietary cracker data, second indicator variable

Cracker	Digested	i1	i2
control	1772.84	0	1
bran	1752.63	0	0
combo	2121.97	1	0
gum	2558.61	0	0
gum	2026.91	0	0
bran	2047.42	0	0
combo	2254.75	1	0
control	2353.21	0	1
combo	2153.36	1	0
gum	2331.19	0	0
bran	2547.77	0	0
...			

Speaker notes

The third category level in alphabetical order is “control”. The second indicator variable is assigned a value of 1 for “control” and 0 for the other category levels.

Dietary cracker data, third indicator variable

Cracker	Digested	i1	i2	i3
control	1772.84	0	1	0
bran	1752.63	0	0	0
combo	2121.97	1	0	0
gum	2558.61	0	0	1
gum	2026.91	0	0	1
bran	2047.42	0	0	0
combo	2254.75	1	0	0
control	2353.21	0	1	0
combo	2153.36	1	0	0
gum	2331.19	0	0	1
bran	2547.77	0	0	0

...

Speaker notes

The fourth category level in alphabetical order is “gum”. The third indicator is assigned a value of 1 for “gum” and 0 for the other categories.

Notice that the category level that appears first (“bran”) is left out. You could assign an indicator for it, but that would lead to a redundancy. Once you see that the cracker is not “combo”, “control”, or “gum”, then you know that it must be “bran”.

The category level that is always zero is called the reference category.

Interpretation with multiple indicator variables

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

Speaker notes

The interpretation of the regression estimates when you use multiple indicator variables is not too much different than your interpretation when you have a single indicator variable.

The intercept is the estimated average value of Y when all of the indicator variables are equal to zero. This is the estimated average value of Y for the reference level, which by default in R is the category level that appears first in alphabetical order.

The first slope represents an estimated change. It is the change when you move from the first category level to the second category level. This is essentially the difference between the second category level mean and the first category level mean.

The second slope also represents an estimated change, but for a different pair of categories. It is the difference between the third category level mean and the first category level mean.

This continues for any other indicator variables. Every comparison is a comparison to the reference level or the level that appears first in alphabetical order.

What if you want a different reference category?

Create your own indicator variables

```
dietary |>
  mutate(i1=as.numeric(Cracker=="bran")) |>
  mutate(i2=as.numeric(Cracker=="combo")) |>
  mutate(i3=as.numeric(Cracker=="gum")) -> dietary_1
```

or revise the order using the factor function.

```
new_order <- c("control", "bran", "combo", "gum")
dietary |>
  mutate(Cracker=factor(Cracker, levels=new_order)) -> dietary_2
lm()
```

Speaker notes

Now there is no special reason to make the first category level in alphabetical order the reference category. In the dietary cracker example, it may make more sense to compare every other cracker type to the “control” cracker type. This would require you to change the default in R.

There are two ways to do this. First, you can create your own indicator variables. The code above shows that the first indicator, i1, is 1 for “bran” and 0 for the other categories. The second indicator, i2, is 1 for “combo” and 0 for the other categories. The third indicator, i3, is 1 for “gum” and 0 for the other categories. This make the “control” level the only level which is 0 for every indicator. The results will be that the slope associated with i1, the first indicator, represents the estimated average difference between “bran” and “control”. The slope associate with i2 represents the estimated average difference between “combo” and “control”. The slope associated with i3 represents the estimated average difference between “gum” and “control”.

If you want R to create these indicator variables for you, list the category levels in a new order with the reference level being the first level in the list. Then use the factor function to assign this order, rather than an alphabetical order, to the category levels.

Example using fruitfly lifespans, 1

- Experiment with 125 cages
  - Does fruitfly mating affect average male lifespan?
  - Isolate a male fruitfly with
    - one or nine virgin females,
    - one or nine pregnant females, or
    - no females
  - Note: males will not mate with pregnant females

Speaker notes

Here is an experiment, apparently a real experiment, looking at average lifespan for male fruitflies. Fruitflies have fairly short lifespans, making it easy to run experiments like this.

The research question is whether mating affects the average lifetime of fruitflies.

Example using fruitfly lifespans, 2

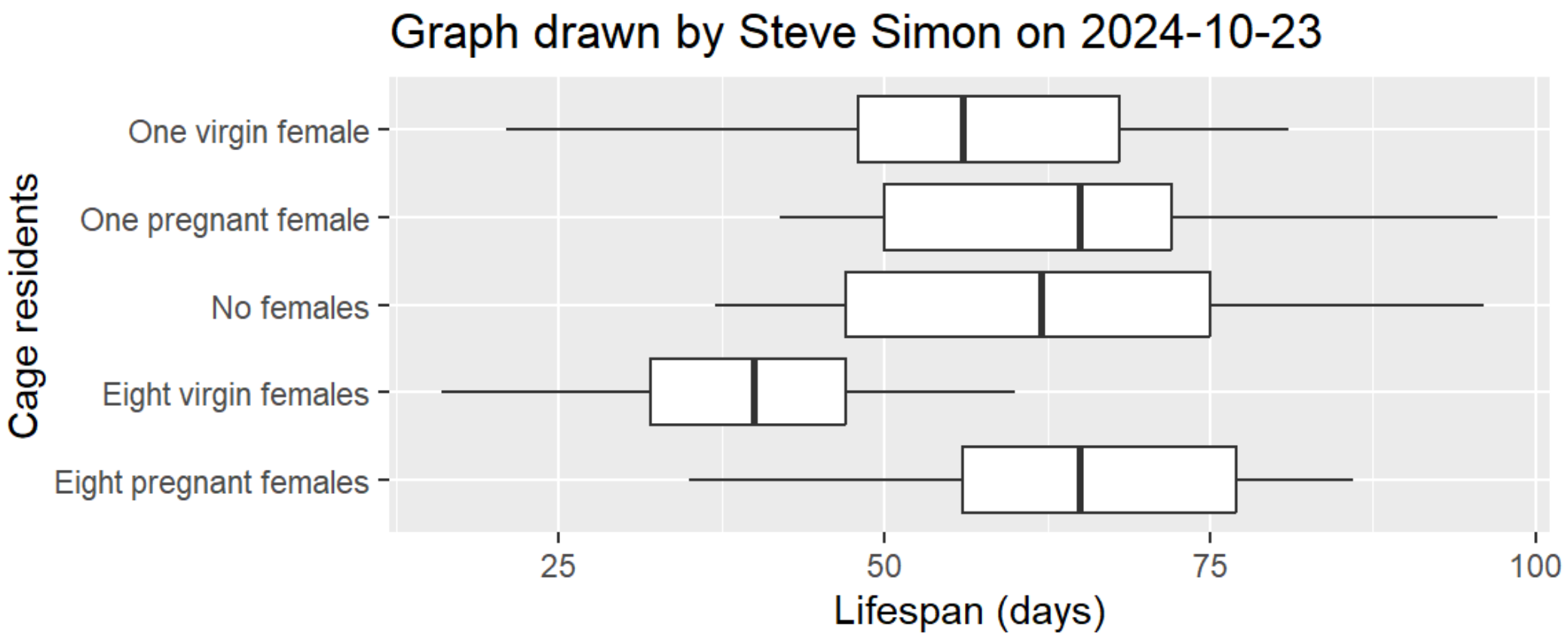
- Cages 1-25 have one male fruitfly, no female fruit flies
- Cages 26-50 have one male fruitfly, one pregnant female
- Cages 51-75 have one male fruitfly, one virgin female
- Cages 76-100 have one male fruitfly, eight pregnant females
- Cages 101-125 have one male fruitfly, eight virgin females

Speaker notes

There is an interesting series of experimental conditions. The male fruitfly in each of the first 25 cages is trapped alone and therefore not able to mate.

In the second set of cages, each male fruitfly shares a cage with a pregnant female fruitfly

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
new_order <- c(
  "No females",
  "One pregnant female",
  "One virgin female",
  "Eight pregnant females",
  "Eight virgin females")

fly |>
  mutate(cage=factor(cage, levels=new_order)) -> fly_1

m1 <- aov(longevity ~ cage, data=fly_1)
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_1)
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 tidy(m2)
```

```
# A tibble: 5 × 5
  term                      estimate std.error statistic  p.value
  <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                  63.6       2.96   21.5    6.76e-43
2 cageOne pregnant female       1.24      4.19    0.296  7.68e- 1
3 cageOne virgin female        -6.80      4.19   -1.62   1.07e- 1
4 cageEight pregnant females   -0.200     4.19   -0.0478 9.62e- 1
5 cageEight virgin females    -24.8       4.19   -5.93   2.98e- 8
```

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    0.2348081459  0.13753858 0.33207771
No females-One virgin female                0.0519038178 -0.04536575 0.14917338
Eight pregnant females-One virgin female    0.0520936375 -0.04517593 0.14936320
```

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.

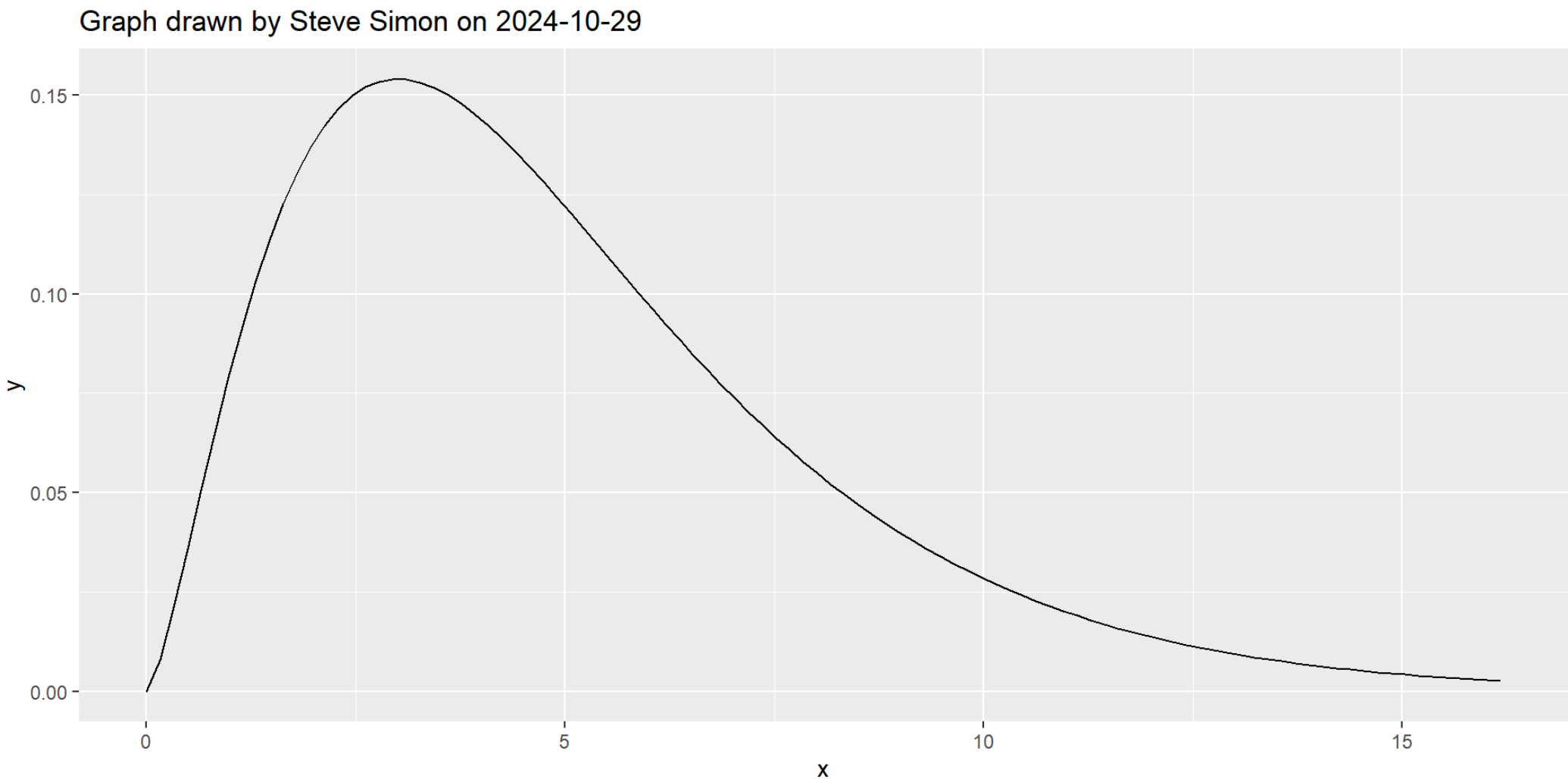
Break #2

- 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

```
1 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

```
## Datasets

-   fly: original data from fruitfly.txt
-   fly_1: re-order the categories in the variable cage

## Models

-   m1: analysis of variance, longevity ~ cage
-   m2: linear regression, same variables
-   m3: analysis of variance, make no females the reference category
```

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

````
## Load the tidyverse library

```{r}

#| label: setup
#| message: false
#| warning: false

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

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

````
## List the variable names

```{r}
#| label: variable-list

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

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

```
#### Comments on the code

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, 6

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

```{r}
#| label: read

fly <- read_table(
  "../data/fruitfly.txt",
  col_types="nnnnnn",
  col_names=vlist)
glimpse(fly)
```
````

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

```
#### Comments on the code

The read_table function is part of the readr library, which is included when you specify the tidyverse library. You can use this function when you have one or more blanks or tabls between each data value.
```

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

````
## Create cage groups

```{r}
#| label: cage

fly |>
  mutate(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")) -> fly_1
```
````

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

```
#### Comments on the code

The case_when function is part of the dplyr library, which is included when you
specify the tidyverse library. This function evaluates a series of tests (listed on
the left side of the tilde) and assigns the particular value to the first test that
evaluates to TRUE. If no test evaluates to TRUE, the case_when function assigns a
missing value.
```

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

````
## Calculate descriptive statistics

```{r}
#| label: longevity-means

fly_1 |>
  group_by(cage) |>
  summarize(
    longevity_mn=mean(longevity),
    longevity_sd=sd(longevity),
    n=n())
```
````

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

```
#### Interpretation of the output

The mean lifespan is 39 days for the "Each virgin females" category, which is much lower than the other means. The means lifespan is 57 days for the "One virgin female" category and this is slightly lower than the remained three categories. The standard deviations are reasonably small and more or less consistent across all groups.
```

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

````
## Draw boxplot

```{r}
#| label: longevity-boxplot
#| fig.width: 6
#| fig.height: 2.5

fly_1 |>
  ggplot() +
  aes(longevity, cage) +
  geom_boxplot() +
  labs(
    caption="Steve Simon, 2024-10-23, CC0",
    x="Lifespan (days)",
````
````

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

```
#### Interpretation of the output

The boxplot shows a roughly normal distribution with no outliers.

## One factor analysis of variance for longevity

```{r}
#| label: longevity-one-factor-anova

m1 <- aov(longevity ~ cage, data=fly_1)
tidy(m1)
```
```

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

```
#### Interpretation of the output

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.

## Linear model for longevity, 1

```{r}
#| label: longevity-lm-1

m2 <- lm(longevity ~ cage, data=fly_1)
anova(m2)
```
```

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

````
#### Interpretation of the output

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

## Linear model for longevity, 2

```{r}
#| label: longevity-lm-2

tidy(m2)
```
````

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

```
#### Comments on the code

The test statistics and p-values listed here do not make any adjustments for multiple comparisons. Some researchers think this is fine. Others prefer that you use the Tukey post hoc comparisons, which insures that the overall Type I error rate is less than 0.05.
```

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

```
#### Interpretation of the output

The linear model creates indicator variable for four out of the five category
levels. The reference category is "Eight pregnant females" which is the first
category level in alphabetical order.

The intercept, 63 days, is the estimated average longevity in days for the "Eight
pregnant females" category.

The first slope is -25. The estimated average longevity decreases by 25 days when you
switch from the "Eight pregnant females" category to the "Eight virgin females"
category.

The second slope is 0.2. The estimated average longevity increases by 0.2 days when
```

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

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

```{r}
#| label: re-order-1
new_order <- c(
  "No females",
  "One pregnant female",
  "Eight pregnant females",
  "One virgin female",
  "Eight virgin females")

fly_1 |>
  mutate(cage=factor(cage, levels=new_order)) -> fly_2
```
```

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

```
#### Comments on the code

The factor function converts a variable a new variable with number codes and labels.
The order of labels is designated by the levels argument.

In this example, the code creates a reference category of "no females" instead of the
default reference category based on alphabetical order.

## Re-order cage groups, 2

```{r}
#| label: re-order-2

m3 <- lm(longevity ~ cage, data=fly_2)
```
```

Listing of simon-5501-11-fruitfly.qmd, 20

```
#### Interpretation of the output

The estimated average longevity is 64 days for the cage with no females. The cages
with one and eight pregnant females change only slightly: one day or less on average.
The cage with one virgin female has 6.8 days less in longevity on average and the
cage with eight virgin females has 25 days in average longevity.

Do not interpret the p-values because they are not adjusted for multiple comparisons.

## Tukey post-hoc test

```{r}
m4 <- aov(longevity ~ cage, data=fly_2)
TukeyHSD(m4)
```
```

Listing of simon-5501-11-fruitfly.qmd, 21

```
#### Interpretation of the output

There are ten possible comparisons among the five groups. Six of the confidence
intervals contain zero. This indicates that there is no statistically significant
difference between

-   one pregnant female and no females
-   eight pregnant females and no females
-   eight pregnant females and one pregnant female
-   one virgin female and one pregnant female
-   one virgin female and one pregnant female
-   one virgin female and eight pregnant females

There is a statistically significant decrease in average longevity for
```

Listing of simon-5501-11-fruitfly.qmd, 22

````
## Log transformation, 1

```{r}
#| label: log-longevity-anova-1

fly_2 |>
  mutate(log_longevity=log10(longevity)) -> fly_3

m3 <- aov(log_longevity ~ cage, data=fly_3)
tidy(m3)
```
````

Listing of simon-5501-11-fruitfly.qmd, 23

```
#### Interpretation of the output

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, 24

````
## Log transformation, 2

```{r}
#| label: log-longevity-anova-2

m3 |>
  TukeyHSD() |>
  tidy() -> m4

m4 |>
  select(contrast, estimate, conf.low, conf.high) |>
  mutate(estimate=10^estimate) |>
  mutate(conf.low=10^conf.low) |>
  mutate(conf.hight=10^conf.high)
````
````

Listing of simon-5501-11-fruitfly.qmd, 25

```
#### Comments on the code

It is better to transform these results back to the original scale of measurement
before intepreting the confidence intervals.

#### Interpretation of the output

When you back-transform after a log transformation, the confidence intervals are
confidence intervals for a ratio of means. If the confidence interval includes the
value of 1, there is no statistically significant difference between the two groups.

In this example, there are four statistically significant differences, all involving
the cage with eight virgin females. The fruitflies with eight virgin females have a
longevity that 40% shorter on average frutiflies with no females. The relevant
```

Listing of simon-5501-11-fruitfly.qmd, 26

````
## Kruskal-Wallis test

```{r}
#| label: kw

kruskal.test(longevity ~ cage, data=fly_2)
```


#### Interpretation of the output

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, 27

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

```{r}
#| label: save

save(
  fly,
  fly_3,
  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