MEDB 5501, Module11

2025-11-04

Topics to be covered

- What you will learn
  - Analysis of variance is linear regression
  - R code for analysis of variance
  - Log transformation
  - R code for analysis of variance
  - 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 eight virgin females,
    - one or eight 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 or crowding 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.

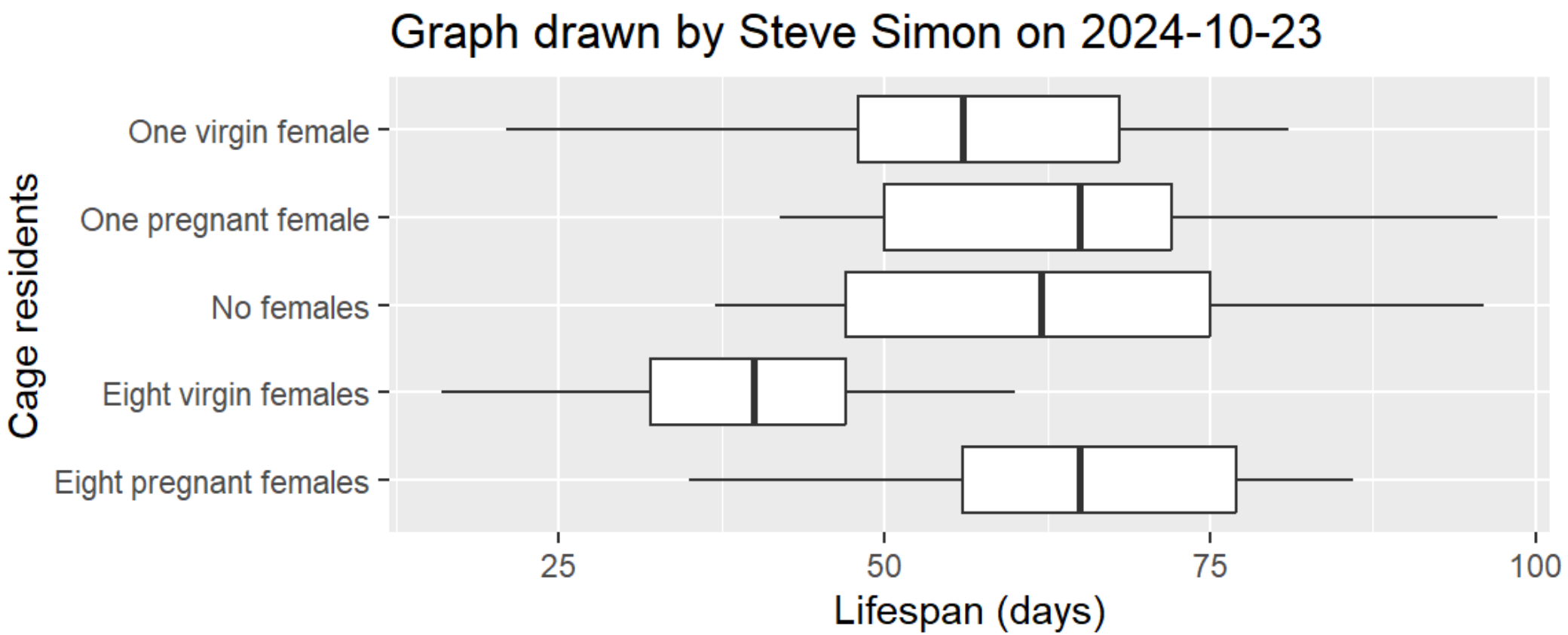
In the second set of cages, each male fruitfly shares a cage with a pregnant female fruitfly. Males will not mate with a pregnant female. So this serves as a second type of control group for the hypothesis that mating affects longevity. It also allows you test whether sharing a cage with another fly affects longevity.

In the third set of cages, the one male fruitfly shares the cage with a virgin female. By comparing this set to the first two sets, you can examine the hypothesis that mating decreases the average lifespan.

The fourth cage is where it gets really interesting. The one male fly shares a cage with EIGHT pregnant females. If there is an effect of crowding on longevity, surely you will see it here.

The first cage includes eight virgin flies with the one male fly. Lots of crowding and lots of opportunities for mating. It will be interesting to tabulate the results from this cage and compare it to the previous cage with eight pregnant females which has lots of crowding and no opportunities for mating. it will also be interesting to compare it to groups 2 and 3. The presence of only female fly means a lot less crowding. Finally, there is an interesting comparison to the first group, the group with no female flies and therefore no crowding.

Fruitfly lifespan boxplots



Speaker notes

Here are the boxplots of the five groups. Notice that R alphabetizes these groups. You can change this using the factor function as noted above.

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
```

Speaker notes

The means show the same pattern. The lifespans tend to be much lower in the cages with eight virgin females. The second lowest mean is with the cages with one virgin female. This tends to support the belief that it is mating rather than crowding that affects longevity.

The formal statistical models will be covered in the next video.

Break #1

- What you have learned
  - Analysis of variance is linear regression
- What’s coming next
  - R code for analysis of variance

R code for analysis of variance

Refer to the program simon-5501-11-fruitfly.qmd.

Break #2

- What you have learned
  - R code for analysis of variance
- 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

Speaker notes

Let’s revisit the discussion of assumptions. There are k samples and you are testing the hypothesis that the population means associated with these samples are all equal versus the alternative that at least two of the means differ from each other.

Violation of assumptions

- Non-normality
- Heterogeneity
- Lack of independence

Speaker notes

The possible violations of the assumptions in an analysis of variance model are non-normality, heterogeneity, and lack of independence.

In some settings, a log transformation can remedy the first two of these violations.

When to consider a log transformation

- Only positive values
- Max/min > 3
- Right (positive) skewed distribution
- Groups with larger means have more variation

Speaker notes

Some practical guidance helps you decide whether you should consider a log transformation.

First, do you have only positive values? The log transformation does not work with zeros or negative values.

Is there a good spread in the data? The log transformation squeezes the large values and stretches the small values, but only if there is a lot of difference between the large and small values.

Look at the largest and smallest values in the dataset across all the groups. If they differ by a factor of three or more, then a log transformation might help.

Is the data in each group skewed to the right or skewed positive? Does it tend to produce extreme values, but only on the high end? Then the stretching and squeezing will bring those high end outliers closer to the rest of the data.

Finally, do the groups with larger means also have more variation? If so, the log transformation will squeeze the groups with the larger standard deviations more, which might tend to equalize the variations.

To state this negatively, skip the log transformation if

- there are zeros and/or negative values,
- if the ratio of the largest to smallest value is less than 3,
- if the distribution in each group is symmetric or skewed left, or
- if the groups with the largest means do not show more variation.

How logarithms convert multiplication and division

- $log(a \times b) = log(a) + log(b)$
  - $log(a/b) = log(a) - log(b)$

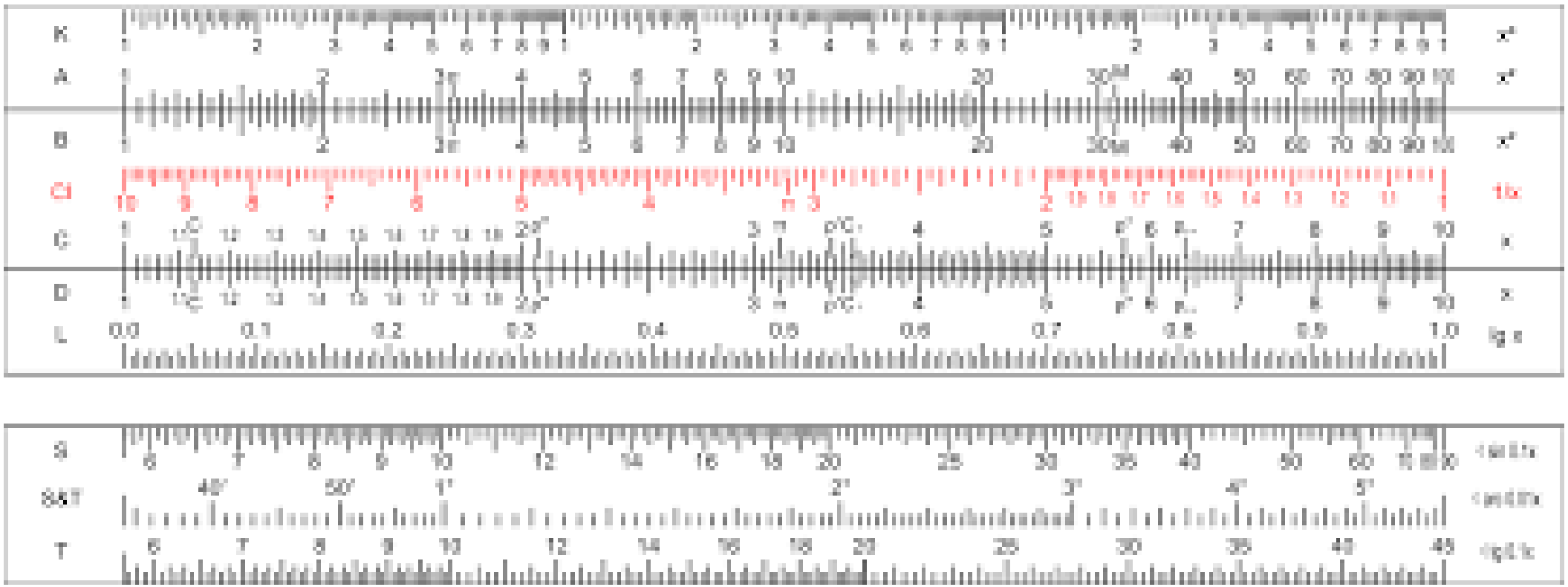
Speaker notes

An interesting property of logarithms is how they handle multiplication and division.

The log of a times b equals the log of a plus the log of b. In other words, the log of a product is equal to the sum of the logs. The log converts multiplication to addition.

Similarly, the log of a divided by b equals the log of a minus the log of b. The log converts division to subtraction.

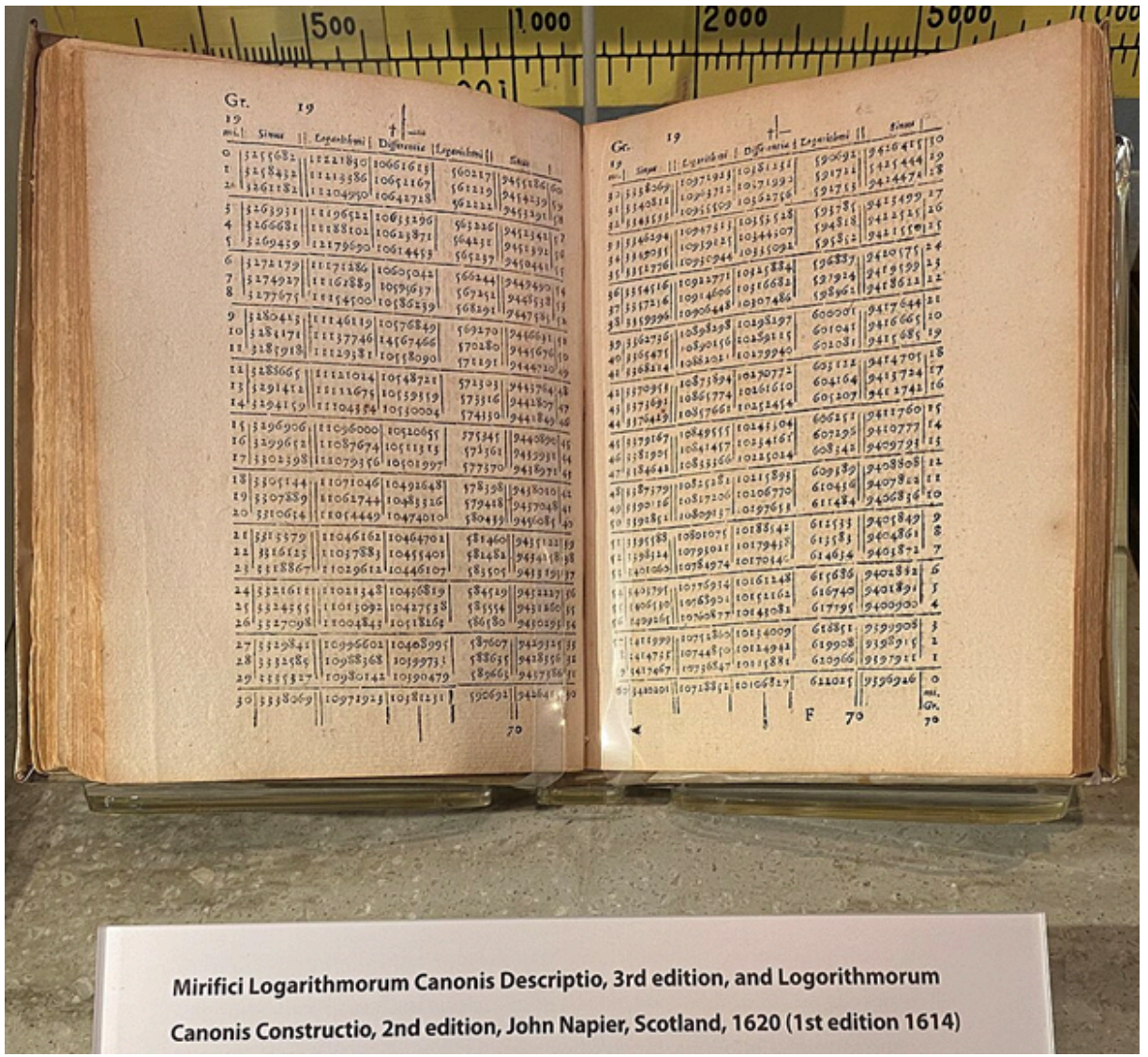
The slide rule uses this property



Speaker notes

A slide rule used to be commonly used for mathematical calculations before we had pocket calculators. It used logarithmic spacing to allow easy computation of multiplications and divisions.

The book of logarithms



Mirifici Logarithmorum Canonis Descriptio, 3rd edition, and Logorithmorum Canonis Constructio, 2nd edition, John Napier, Scotland, 1620 (1st edition 1614)

Speaker notes

Going back even further in history, mathematical calculations often relied on a book of logarithms. Mathematical calculations involving multiplication and/or division were time consuming to be done by hand. Not terribly so, but the calculations needed, for example, to make astronomical predictions relied on so many of these calculations.

It was faster to convert the two numbers into logarithms, add or subtract, and then convert back using antilogarithms.

What a back transformation does

- $antilog(a) + antilog(b) = antilog(a \times b)$
- $antilog(a) - antilog(b) = antilog(a/b)$
  - For log base ten, antilog(x) means 10^x
  - For log base two, antilog(x) means 2^x
- A difference of means on the log scale becomes a ratio of means after back transformation.

Speaker notes

You reverse the property of log transformation with back transformation. A sum of two antilogs is the antilog of the product. A difference of two antilogs is the antilog of the ratio.

Recall that antilog, the reversal of the log transformation means raising 10 to the power if you used base 10 logarithms. If you used base 2 logarithms it means raising 2 to the power.

From a statistical perspective what this means is that back transformation converts a difference of means into a ratio of means.

Interpretation of the back transformed confidence intervals

- A confidence interval is a range of plausible values.
  - A back transformed confidence interval is a range of plausible ratios.
- Does the interval contain the ratio of 1?
  - The two groups are statistically the same.
- Does the interval contain only values larger than 1?
  - The first group is statistically larger
- Does the interval contain only values smaller than 1?
  - The first group is statistically smaller

Speaker notes

Ask yourself if the interval contains the value of 1. A ratio of 1 implies equality. So if 1 is a plausible value, then it is plausible to behave as if the two population means are equal.

If the confidence interval only includes values larger than 1, then it is plausible to behave as if the population mean of the first group is larger than the population mean of the second group.

If the confidence interval only includes values smaller than 1, then it is plausible to behave as if the population mean of the first group is smaller than the population mean of the second group.

Log transformation, 1

```
# A tibble: 10 × 7
   term  contrast                null.value estimate conf.low conf.high adj.p.value
   <chr> <chr>                        <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
 1 cage  One pregnant female…             0  1.02e-2  -0.0871    0.107      9.98e-1
 2 cage  Eight pregnant fema…             0  1.90e-4  -0.0971    0.0975     1.00e+0
 3 cage  One virgin female-N…             0 -5.19e-2  -0.149     0.0454     5.79e-1
 4 cage  Eight virgin female…             0 -2.25e-1  -0.322    -0.127      3.18e-8
 5 cage  Eight pregnant fema…             0 -9.99e-3  -0.107     0.0873     9.99e-1
 6 cage  One virgin female-O…             0 -6.21e-2  -0.159     0.0352     3.97e-1
 7 cage  Eight virgin female…             0 -2.35e-1  -0.332    -0.138      7.64e-9
 8 cage  One virgin female-E…             0 -5.21e-2  -0.149     0.0452     5.75e-1
 9 cage  Eight virgin female…             0 -2.25e-1  -0.322    -0.128      3.10e-8
10 cage  Eight virgin female…             0 -1.73e-1  -0.270    -0.0755     2.74e-5
```

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.

Here are the results of the Tukey post hoc tests on the log scale. You can examine whether these intervals include or exclude the value of zero, but further interpretation, using log-days as the unit of measurement, is tricky.

Log transformation, 3

```
# A tibble: 10 × 4
   contrast                                  estimate conf.low conf.high
   <chr>                                        <dbl>    <dbl>     <dbl>
 1 One pregnant female-No females               1.02     0.818     1.28
 2 Eight pregnant females-No females            1.00     0.800     1.25
 3 One virgin female-No females                 0.887    0.709     1.11
 4 Eight virgin females-No females              0.596    0.477     0.746
 5 Eight pregnant females-One pregnant female   0.977    0.781     1.22
 6 One virgin female-One pregnant female        0.867    0.693     1.08
 7 Eight virgin females-One pregnant female     0.582    0.466     0.729
 8 One virgin female-Eight pregnant females     0.887    0.709     1.11
 9 Eight virgin females-Eight pregnant females  0.596    0.476     0.746
10 Eight virgin females-One virgin female       0.672    0.537     0.841
```

Speaker notes

Here are the results translated back to the original scale. The confidence intervals are intervals for the ratio. Notice that six of the ten intervals include the value of 1.

Let’s interpret the first and last intervals.

The ratio of average longevity comparing one pregnant female to no females is 1.02, or 2% larger. The confidence interval ranges from 0.82 or 18% smaller to 1.28 or 28% larger. There is no statistically significant change in longevity between fruitflies caged alone and fruitflies caged with one pregnant female.

The ratio of average longevity comparing eight virgin females to one virgin female is 0.67 or 33% smaller. The confidence interval ranges from 0.54 or 46% smaller to 0.84 or 16% smaller. The average longevity of fruitflies with 8 virgin females is significantly shorter than the average longevity of fruitflies with only 1 virgin female.

Break #3

- What you have learned
  - Log transformation
- What’s coming next
  - R code for analysis of variance

R code for log transformed analysis of variance

Refer to the program simon-5501-11-fruitfly.qmd.

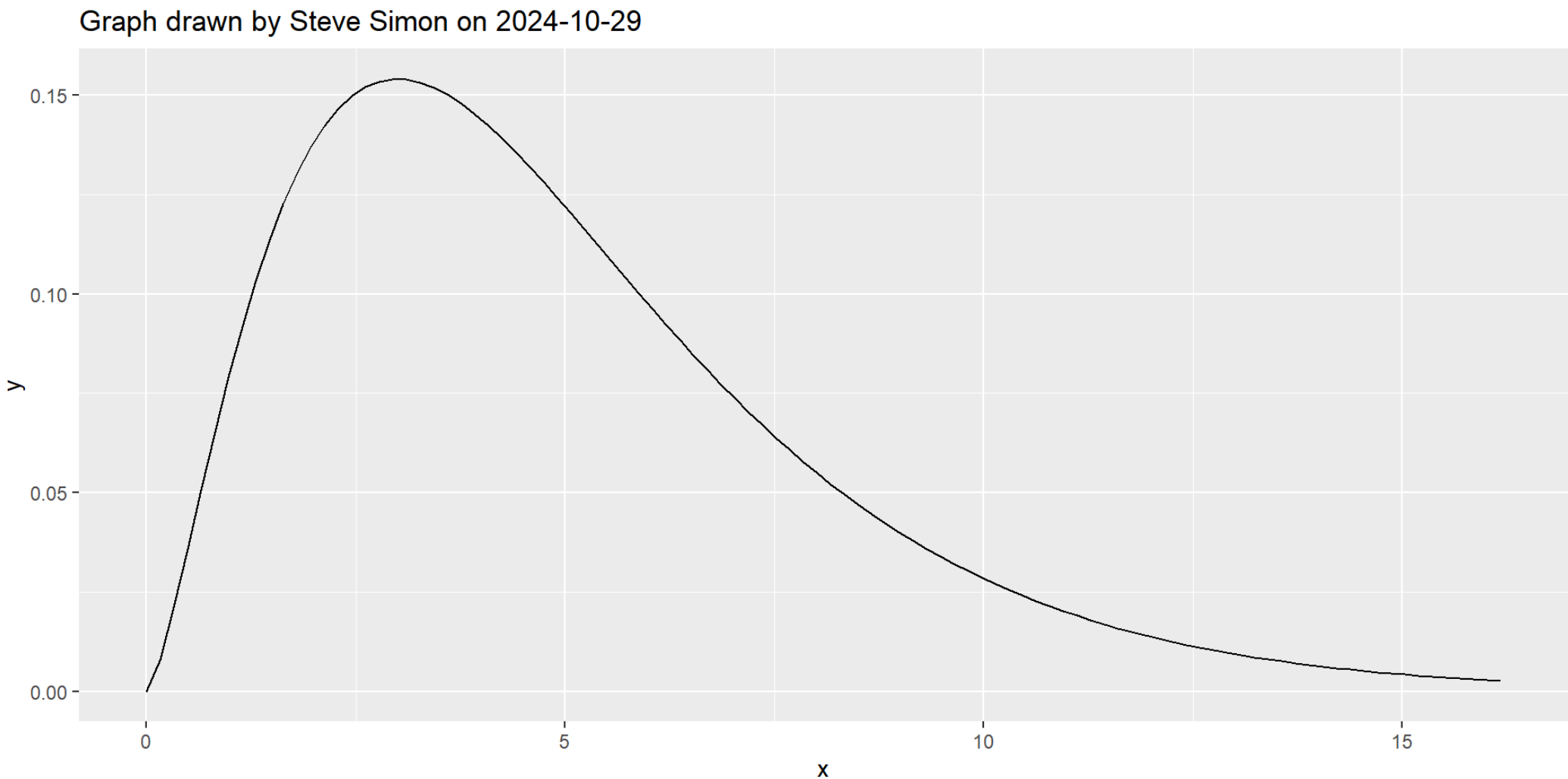
Break #4

- What you have learned
  - R code for analysis of variance
- 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.

Speaker notes

There is a technical issue with the Kruskal-Wallis test and nonparametric tests in general. Strictly speaking, they are not really tests of populations means. Some people describe the hypothesis in terms of population medians. Others talk about stochastic dominance. This is a distinction that some people prefer to sidestep, but others can get fussy about it.

If you are working with someone who is a stickler for details, use the concept of stochastic dominance. If they are not sticklers, just talk vaguely about one group possibly being “larger” in some sense than another group.

Application of Kruskal-Wallis test to fruitfly longevity

```r
kruskal.test(longevity ~ cage, data=fly_1)
```

```
    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 #5

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

R code for the Kruskal-Wallis test

Refer to the program simon-5501-11-fruitfly.qmd.

Break #6

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

Your homework

Refer to the file simon-5501-11-directions on my github site.

Summary

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