MEDB 5501, Module10

2024-10-22

Topics to be covered

- What you will learn
  - The analysis of variance model
  - The F-test
  - Assumptions
  - Confidence intervals
  - R code for analysis of variance
  - Sample size justification
  - R code for sample size justification

Review two-sample t-test

- $H_0: \ \mu_1 = \mu_2$
- $H_1: \ \mu_1 \neq \mu_2$
- $T = \frac{\bar{X}_1 - \bar{X}_2}{se}$
  - Accept H0 if T is close to zero.

Speaker notes

You saw how to compare two means last week. Here is the general framework.

What to do with three or more groups?

- $H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1: \ \mu_i \neq \mu_j$ for some i, j
- Accept H0 if the F ratio (defined below) is close to 1.

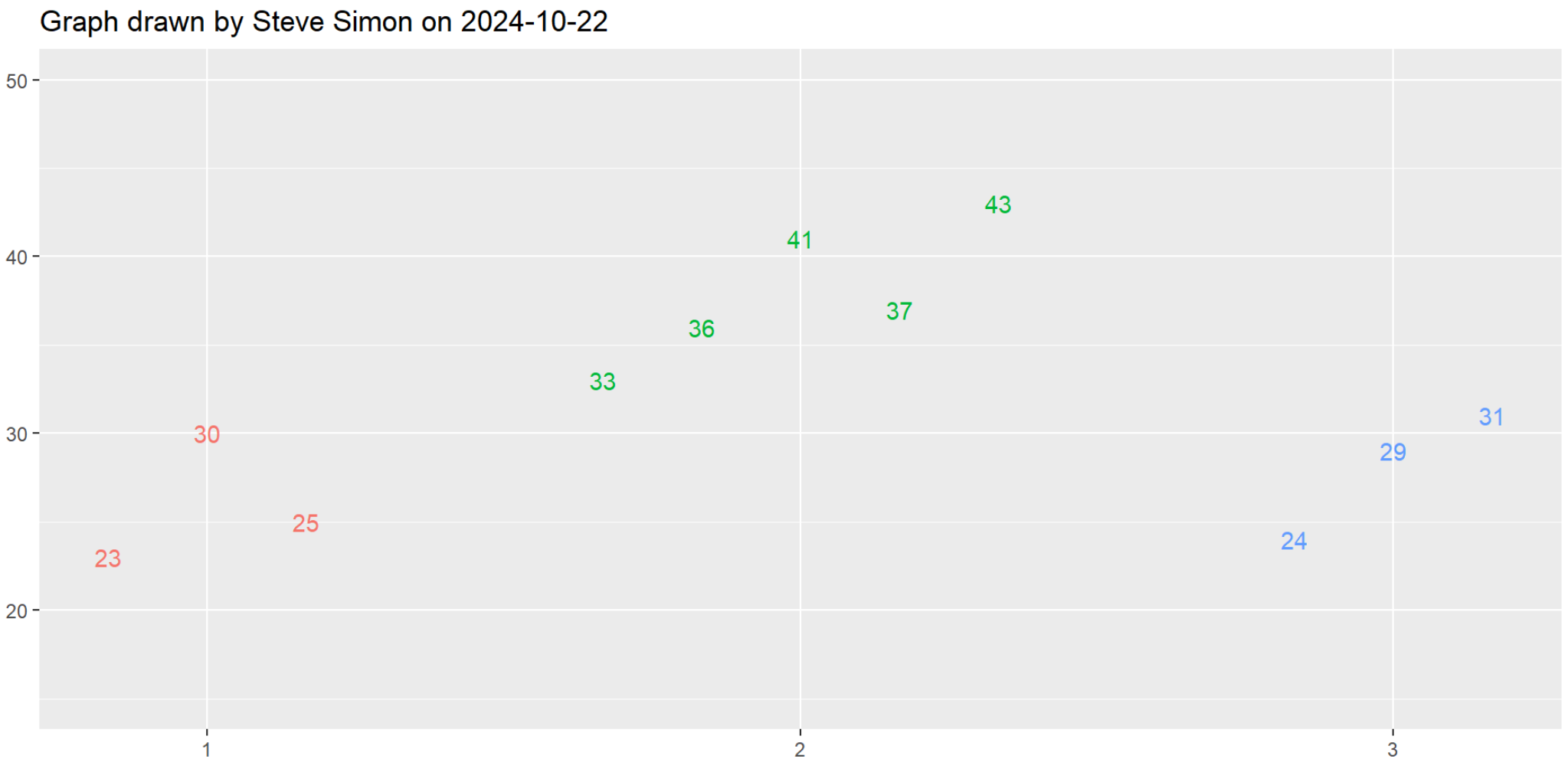
Artificial data

	g	y
1	1	23
2	1	30
3	1	25
4	2	33
5	2	36
6	2	41
7	2	37
8	2	43
9	3	24
10	3	29
11	3	31

Speaker notes

To motivate some of the calculations in Analysis of Variance, I created an artificial data set with numbers that are easy to work with.

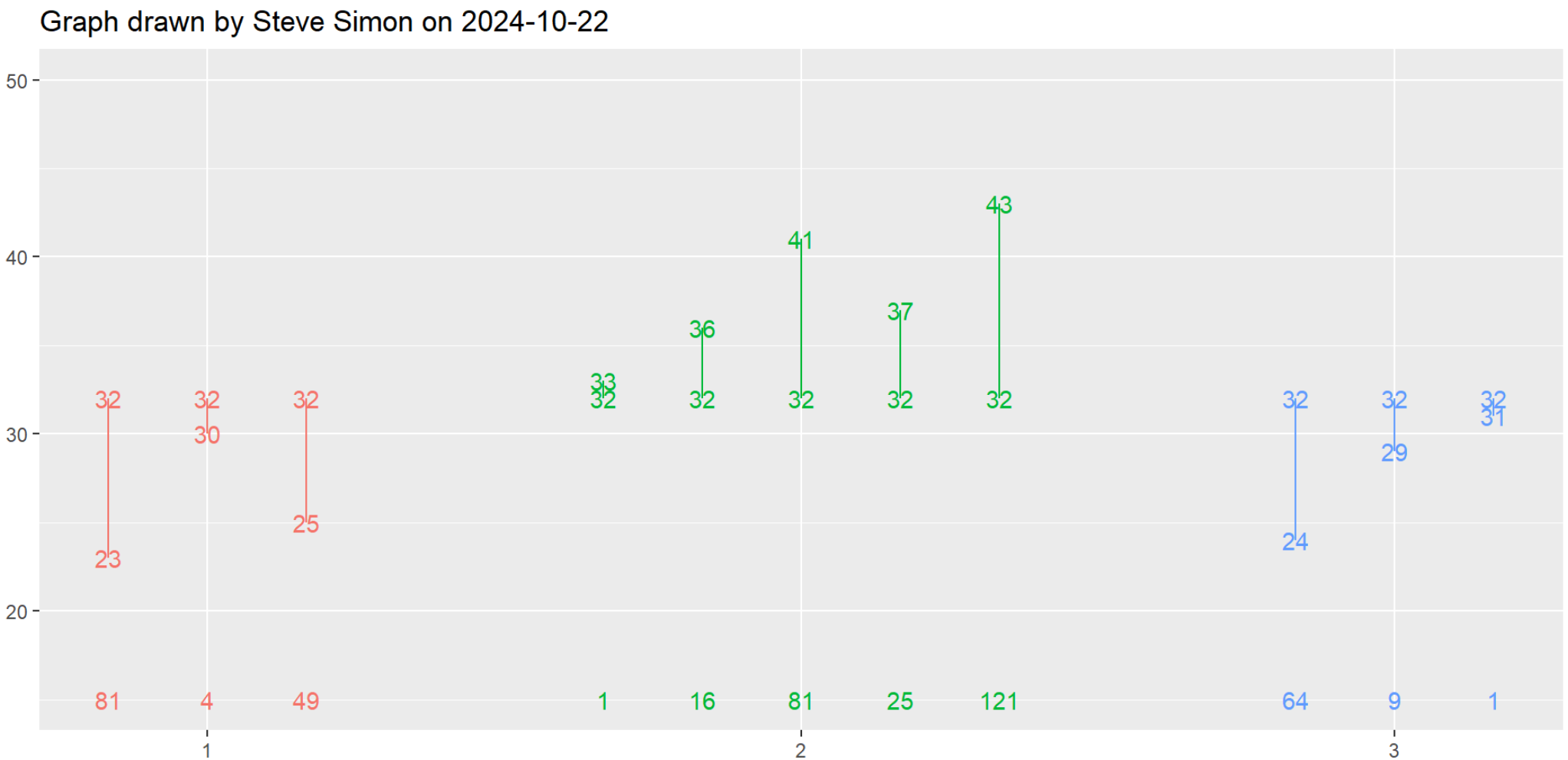
Scatterplot



Speaker notes

Here is a plot of the data.

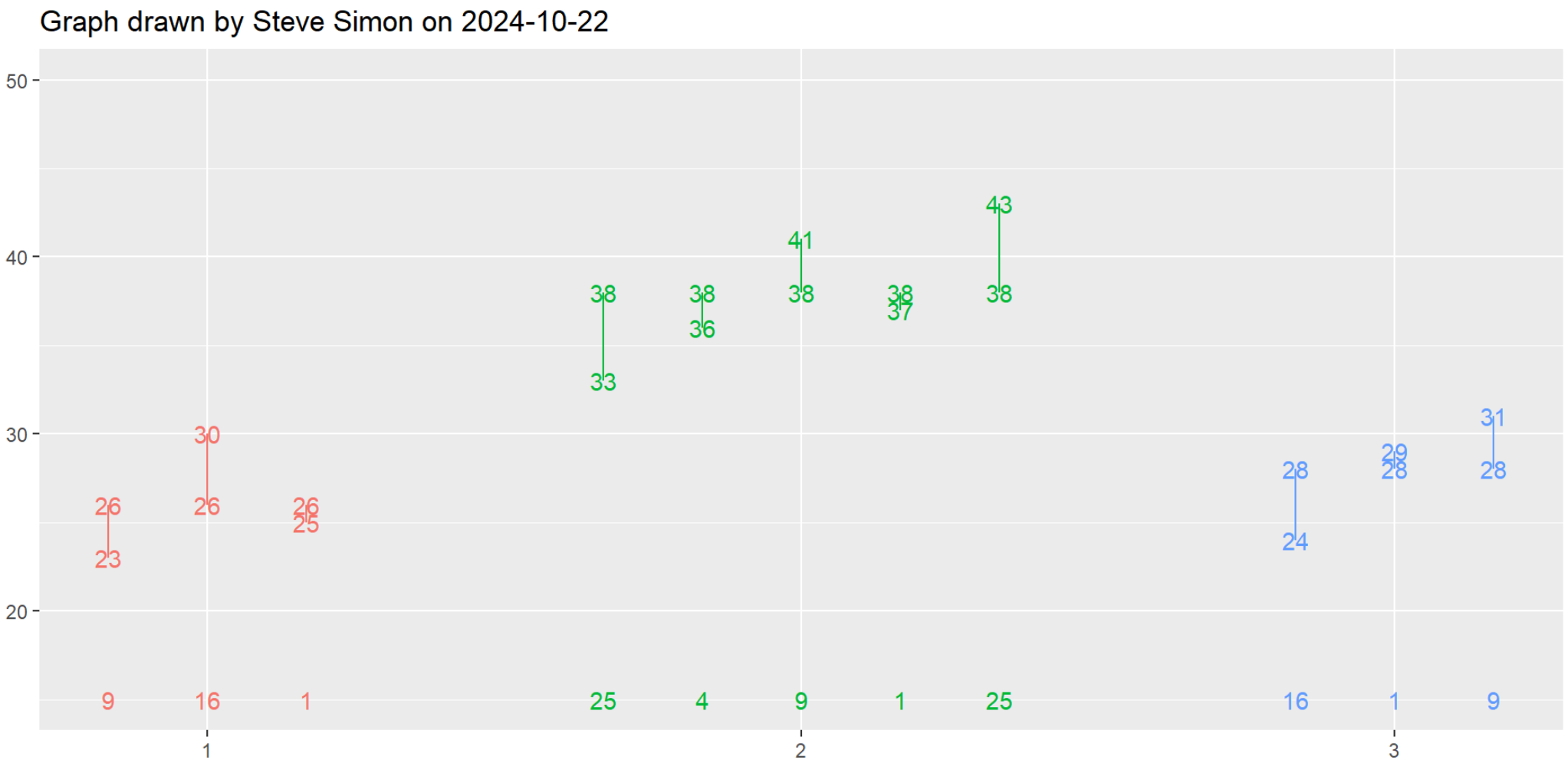
SS(Total) = 452


Graph drawn by Steve Simon on 2024-10-22
50
40
30
20
1
2
3
32
32
32
30
25
23
33
32
36
32
41
32
37
32
43
32
32
24
32
29
32
31
81
4
49
1
16
81
25
121
64
9
1

Speaker notes

Total sums of squares is the squared deviation between each individual value and the overall mean.

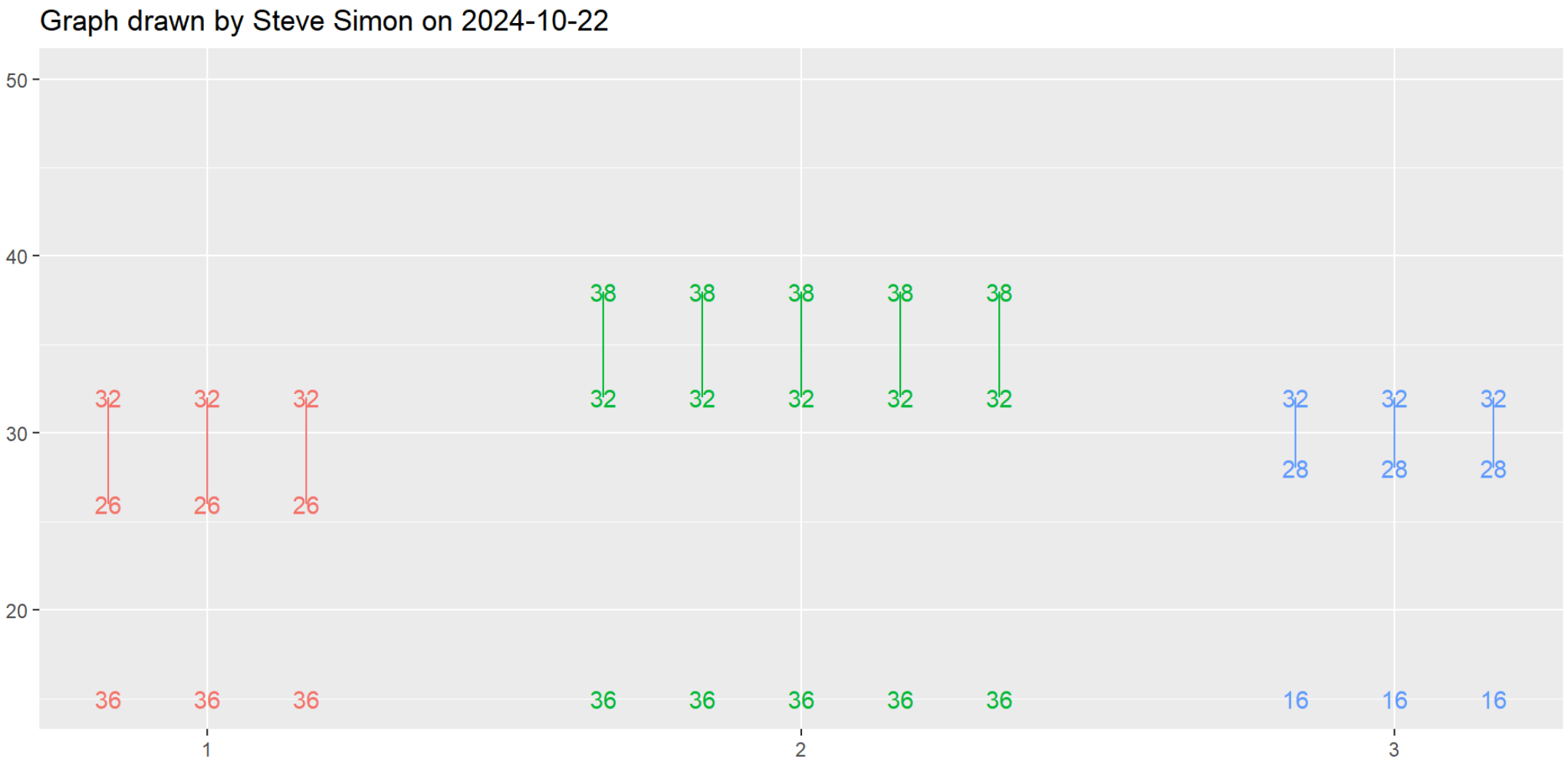
SS(Within) = 116



Speaker notes

Within sums of squares is the squared deviation between each individual value and the group means.

SS(Between) = 336



Speaker notes

Between SS is the squared deviation between the group means and the overall mean.

Degrees of freedom

- For SS(Total), df = 10
- For SS(Within), df = 8
- For SS(Between), df = 2
- In general,
  - N = number of observations total
  - k = number of groups
    - Total df = N-1
    - Within df = N-k
    - Between df = k-1

Speaker notes

The concept of degrees of freedom is tricky. It is the number of “independent” observations, or the number of observations minus the number of estimated parameters.

For Total SS, you have 11 observations, but one estimated parameter, the overall mean of 32. The degrees of freedom is 11-1 = 10.

For Within SS, you also have 11 observations, but there are 3 estimated parameters, the three group means. The degrees of freedom is 11-3 = 8

For Between SS, you only have three observations, the three group means. There is one estimated parameter, the overall mean. The degrees of freedom is 3-1 = 2.

In general, if you let N represent the total number of observations across all groups and let k represent the number of groups, then the degrees of freedom are N-1, N-k, and k-1.

R calculations of sums of squares

```r
m2 <- aov(y ~ factor(g), data=aa)
tidy(m2) |>
  select(term, df, sumsq, meansq)
```

```
# A tibble: 2 × 4
  term          df sumsq meansq
  <chr>      <dbl> <dbl>  <dbl>
1 factor(g)      2   336  168
2 Residuals      8   116   14.5
```

Speaker notes

Here is the analysis of variance table as computed by R. I have left out the p-value and F-ratio, which you will hear about in the next video.

Break #1

- What you have learned
  - The analysis of variance model
- What’s coming next
  - The F-test

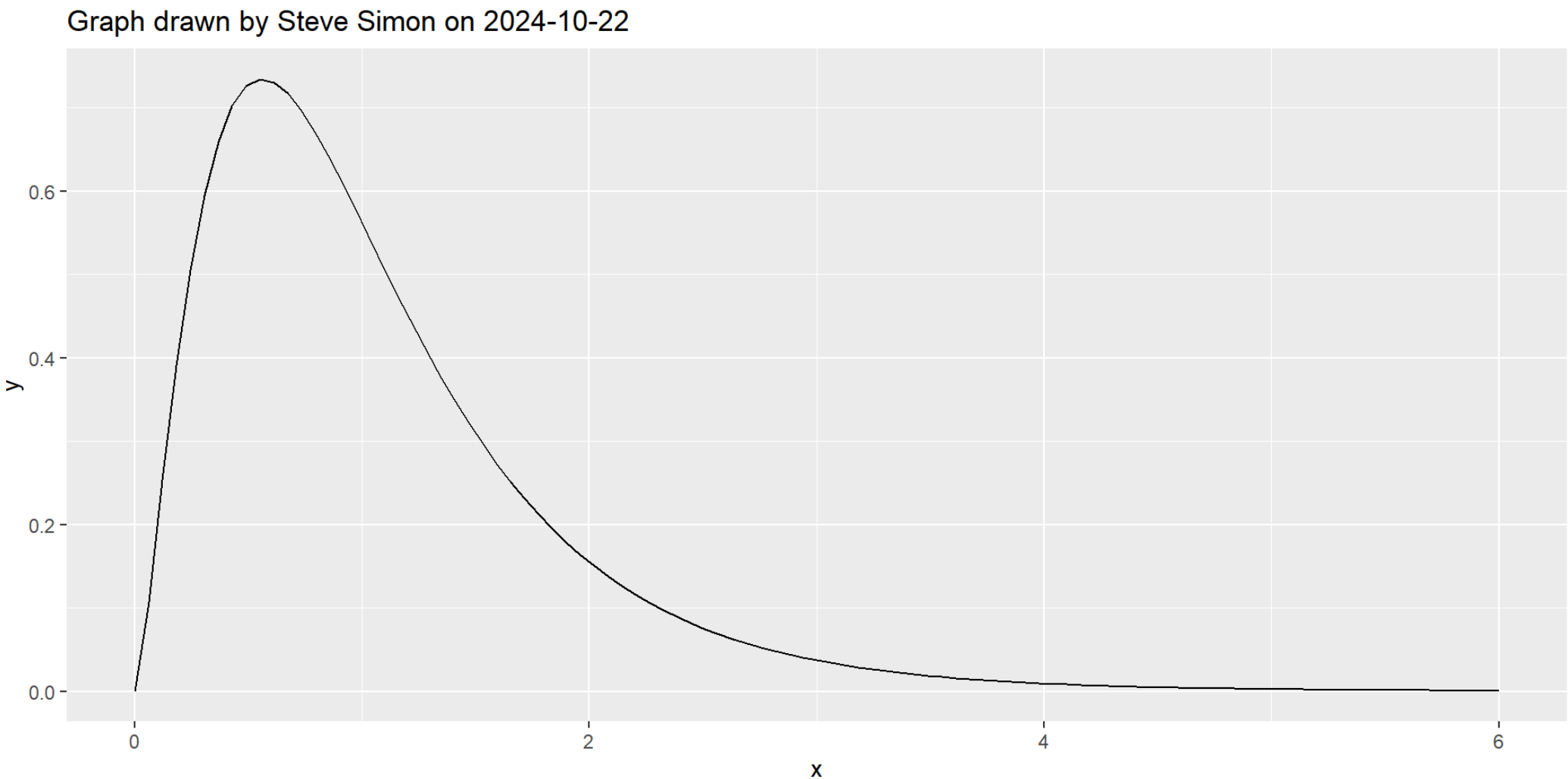
The F-distribution

- Ratio of two measures of variation
- Two measures are comparable if F is close to 1
- F distribution is skewed and can never be negative
- Two measures of degrees of freedom
  - df_1 = degrees of freedom for numerator
  - df_2 = degrees of freedom for denominator

Speaker notes

The F distribution appears many times in Statistics when you are comparing two different measures of variation.

Graph of the F distribution



Speaker notes

Here is a picture of the F distribution.

The F-test

- $H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1: \ \mu_i \neq \mu_j$ for some i, j
  - Compute F = MS(Between) / MS(Within)
  - Accept H0 if the F < $F(1-\alpha, k-1, N-k)$
- Only reject H_0 for large values

Speaker notes

You only reject the null hypothesis for large values of F. If the variation between groups is about equal to the variation within groups or if it is much smaller than the variation within groups, you should accept the null hypothesis.

The p-value for the F test

- p-value = P[F(k-1, N-k) > F]
- Accept H_0 if p-value > α.

Speaker notes

R will compute a p-value for you and you compare the p-value to alpha. With a large p-value (greater than alpha), you should accept the null hypothesis. With a small p-value (less than or equal to alpha), you should reject the null hypothesis.

R code for the F test

```
# A tibble: 2 × 6
  term          df sumsq meansq statistic  p.value
  <chr>      <dbl> <dbl>  <dbl>     <dbl>    <dbl>
1 factor(g)      2   336  168        11.6  0.00434
2 Residuals      8   116   14.5      NA   NA
```

Because the F-ratio is large and the p-value is less than 0.05, you should reject H_0 and conclude that there is a difference among the means.

R-squared

- $R^2 = \frac{SS(Between)}{SS(Total)}$
  - Proportion of variation explained by groups

```
glance(m1)$r.squared
```

```
[1] 0.7433628
```

Approximately 74% of the variation in measurements can be accounted for by the grouping.

Break #2

- What you have learned
  - The F-test
- What’s coming next
  - Assumptions

Important assumptions

- Same as independent-samples t-test
  - Normality
  - Equal variances
  - Independence
- Note: unequal sample sizes is not a violation of assumptions
  - But does lead to some tedious complications

How to check for non-normality

- Boxplots
  - Look for evidence of skewness or outliers

How to check for heterogeneity

- Descriptive statistics
  - Look for one standard deviation much larger than another
- Boxplots
  - Look for one box that is much wider than another

How to check for independence

- Qualitative assessment of how data was collected

Optional: analysis of residuals, 1

- Predicted value, $\hat{Y}_{ij}$
  - $\hat{Y}_{ij} = \bar{Y}_i$
- Residual, e_{ij}
  - e_{ij}= Observed - Predicted
  - $e_{ij} = Y_{ij} - \hat{Y}_{ij}$
  - $e_{ij} = Y_{ij} - \bar{Y}_i$

Optional: analysis of residuals, 2

- Normal probability plot (QQ plot) of residuals
- Histogram of residuals
- Plot predicted values versus residuals.

Break #3

- What you have learned
  - Assumptions
- What’s coming next
  - Confidence intervals

Review multiple comparisons issue

- Type I error: rejecting the null hypothesis when the null hypothesis is true.
  - Multiple simultaneous hypotheses increase the Type I error rate.
- E_1 = Type I error for Hypothesis 1
- E_2 = Type I error for Hypothesis 2
  - $P[E_1 \cup E_2] = P[E_1] + P[E_2] - P[E_1 \cap E_2]$
  - $P[E_1 \cup E_2] \leq P[E_1] + P[E_2]$
  - $P[E_1 \cup E_2] \leq 2\alpha$

Bonferroni adjustment

- For m hypotheses
  - $P[E_1 \cup \ldots \cup E_m] \leq m\alpha$
- Test each hypothesis at α/m
  - Preserves overall Type I error rate
- Example, 3 simultaneous hypotheses
  - Reject H0 if p-value < 0.05/3 or 0.0167

Tukey post hoc tests

- If you reject H0, which values are unequal
  - With k groups, there are k(k-1)/2 comparisons
- Studentized range (Tukey test)
  - Requires equal sample sizes per group
  - Uses harmonic mean approximation for slightly unequal sample sizes.
    - Do not use harmonic means if seriously different sample sizes.
- TukeyHSD

Example of Tukey post hoc test with artificial data

```
# A tibble: 3 × 2
      g y_mean
  <dbl>  <dbl>
1     1     26
2     2     38
3     3     28

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ factor(g), data = aa)

$`factor(g)`
    diff        lwr      upr     p adj
2-1   12   4.053770 19.94623 0.0064090
3-1    2  -6.884155 10.88416 0.8011989
3-2  -10 -17.946230 -2.05377 0.0171787
```

Interpretation, 1

- Group 2 (mean=38) is larger than Group 1 (mean=26)
- Group 3 (mean=28) and Group 1 (mean=26) are not different
- Group 3 (mean=28) is smaller than Group 2 (mean=38)

Interpretation, 2

- Group 2 (mean=38) is larger than Group 1 (mean=26)
- Group 2 (mean=38) is larger than Group 3 (mean=28)
- Group 1 (mean=28) and Group 3 (mean=26) are not different

Alternatives to Tukey post hoc tests, 1

- Bonferroni adjustment
  - Works for unequal sample sizes per group
  - Works for unequal variances
- Dunnett's test
  - Treatment versus multiple controls
- Scheffe's test
  - Works for complex comparison
    - Example $\mu_1 \ vs. \ \frac{\mu_2+\mu_3+\mu_4}{3}$

Controversies over Tukey/Bonferroni adjustment

- Increases Type II errors
- Impractical for large values of m

Break #4

- What you have learned
  - Confidence intervals
- What’s coming next
  - R code for analysis of variance

wolf-river-pollution data dictionary, 1

```
data_dictionary: wolf-river-pollution

 format:
  txt: tab-delimited

 varnames: first row of data

 missing_value_code: not needed

 description: |
  Ten water samples were taken at three different depths in Wolf River. Two
pollutants, Aldrin and HCB, were measured in each sample.
```

wolf-river-pollution data dictionary, 2

```
additional_description: https://gksmyth.github.io/ozdasl/general/wolfrive.html

 download_url: https://gksmyth.github.io/ozdasl/general/wolfrive.txt

 source: |
  OzDASL (Australian Data and Story Library), a repository for various data
sets useful for teaching.

 copyright: |
  Unknown. You should be able to use this data for individual educational
purposes under the Fair Use guidelines of U.S. copyright law.

 size:
  rows: 30
  columns: 3
```

wolf-river-pollution data dictionary, 3

```
Depth:
  label: Location of water sample
  values:
  - Surface
  - Middepth
  - Bottom
  scale: ordinal
```

wolf-river-pollution data dictionary, 4

```
Aldrin:
  label: Concentration of Aldrain
  units: Not specified
  scale: ratio

 HCB:
  label: Concentration of HCB
  units: Not specified
  scale: ratio
```

simon-5501-10-river.qmd, 1

```
---
title: "Analysis of wolf river pollution"
format:
  html:
    embed-resources: true
---

This program reads data on the relationship sampling depth and two pollutant
concentrations. Find more information in the [data dictionary][dd].

[dd]: https://github.com/pmean/datasets/blob/master/wolf-river-pollution.yaml

This program was written by Steve Simon on 2024-10-20 and is placed in the
public domain.
```

simon-5501-10-river.qmd, 2

````
## Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages
and warnings are suppressed.

```{r setup}
#| message: false
#| warning: false
library(broom)
library(tidyverse)
```
````

simon-5501-10-river.qmd, 3

````
## Read the data

```{r read-1}
river <- read_tsv(
  file="../data/wolf-river-pollution.txt",
  col_names=TRUE,
  col_types="cnn")
names(river) <- tolower(names(river))
glimpse(river)
```
````

simon-5501-10-river.qmd, 4

````
## Draw boxplots

```{r box-1}
#| fig.width: 6
#| fig.height: 2.5
river |>
  ggplot(aes(depth, aldrin)) +
    geom_boxplot() +
    xlab("River depth") +
    ylab("Aldrin concentration") +
    ggtitle("Graph drawn by Steve Simon on 2024-10-20") +
    coord_flip()
```

The deeper you sample, the higher the concentration of Aldrin. The variation
````

simon-5501-10-river.qmd, 5

````
## Descriptive statistics

```{r descriptives-1}
river |>
  group_by(depth) |>
  summarize(
    aldrin_mn=mean(aldrin),
    aldrin_sd=sd(aldrin),
    n=n())
```

The bottom samples have the highest average concentration and the highest
amount of variability.
````

simon-5501-10-river.qmd, 6

```
## Analysis of variance table

```{r aov-1}
m1 <- aov(aldrin ~ depth, data=river)
tidy(m1)
```
```

The F-ratio is large and the p-value is less than alpha. You should reject the null hypothesis and conclude that at least two means differ from one another.

simon-5501-10-river.qmd, 7

```
## Pairwise tests

```{r pairwise-1}
TukeyHSD(m1)
```

There is a statistically significant difference in average concentration of Aldrin between the surface measurements and the bottom measurements. The confidence interval, however, is very wide, indicating a large amount of sampling error.

Although the average middepth measurements are larger than the surface measurements and smaller than the bottom measurements, the difference of about 0.8 to 1.0 units is not statistically significant.
```

simon-5501-10-river.qmd, 8

Sample size calculation scenario

You want to replicate this study at a different site and want a lot more precision and power. If the amount of sampling error (mean squared within) is similar, and the populations means are very close (4.8 for the surface, 5.0 for the middepth, and 5.2 for the bottom), what sample size would you need to achieve 90% power with an alpha level of 0.05?

simon-5501-10-river.qmd, 9

````
## Sample size calculation, R code

```{r sample-size}
v <- var(c(4.8, 5.0, 5.2))
power.anova.test(
    groups=3,
    n=NULL,
    between.var=v,
    within.var=1.39,
    sig.level=0.05,
    power=0.90)
```
````

simon-5501-10-river.qmd, 10

```
## Sample size calculation, interpretation

A sample size of 221 measurements per depth level would provide 90% power for
detecting a difference between means of 4.8, 5.0, and 5.2 in aldrin
concentration. This assumes that the variation within groups is similar to the
previous study (1.39) and an alpha level of 0.05.
```

simon-5501-10-river.qmd, 11

```
## Recalculate sample size, new scenario

There is no way in heaven or earth that you can afford to make 221
measurements at each depth. So give up the idea that you can detect changes in
means across such a narrow range. Suppose that you want to be able to detect
differences among means that are 4.5, 5.0, and 5.5. If you could live with
that and if everything remains the same, what sample size would you need?
```

simon-5501-10-river.qmd, 12

````
## Recalculate sample size, R code

```{r recalculate}
v <- var(c(4.5, 5.0, 5.5))
power.anova.test(
    groups=3,
    n=NULL,
    between.var=v,
    within.var=1.39,
    sig.level=0.05,
    power=0.90)
```
````

simon-5501-10-river.qmd, 13

```
## Recalculate sample size, interpretation.

Much better! A sample size of 37 measurements per depth level would provide
90% power for detecting a difference between means of 4.5, 5.0, and 5.5 in
aldrin concentration. This assumes that the variation within groups is similar
to the previous study (1.39) and an alpha level of 0.05.
```

Break #5

- What you have learned
  - R code for analysis of variance
- What’s coming next
  - Sample size justification

Using the power.anova.test function to estimate sample size

What if n is outside your budget?

- Increase between.var
- Increase Type I error rate (sig.level)
- Increase Type II error rate (decrease power)
- Decrease number of groups

Break #6

- What you have learned
  - Sample size justification
- What’s coming next
  - R code for sample size justification

Break #7

- What you have learned
  - R code for sample size justification
- What’s coming next
  - Your homework

simon-5501-10-directions.md, 1

```
---
title: "Directions for 5501-10 programming assignment"
---

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

simon-5501-10-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/general/simon-5501-08-
sway.md
```

simon-5501-10-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/wolf-river-pollution.txt
[dic]: https://github.com/pmean/data/blob/main/files/wolf-river-pollution.yaml
```

simon-5501-10-directions.md, 4

Question 1

Compare the average hcb concentrations between the surface, middepth and bottom sampling locations using analysis of variance. Be sure to include appropriate descriptive statistics and graphs. Comment on the assumptions needed for this test, but do not conduct any alternative analyses. If there is a statistically significant difference among the three means, use the Tukey post-hoc comparison to identify where the differences lie.

simon-5501-10-directions.md, 5

```
## Question 2

You want to run a sample size calculation for a replication of this experiment
using hcb as the outcome measure. Assume that the sample means for hcb are
similar at surface and middepth, but higher at the bottom (4.8 for the
surface, 4.8 for middepth, and 5.2 for the bottom). What sample size would you
need to achieve 90% power at an alpha level of 0.05.
```

simon-5501-10-directions.md, 6

```
## Your submission

-   Save the output in html format
-   Convert it 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
```

simon-5501-10-directions.md, 7

```
## 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
  - The analysis of variance model
  - The F-test
  - Assumptions
  - Confidence intervals
  - R code for analysis of variance
  - Sample size justification
  - R code for sample size justification
  - Your homework