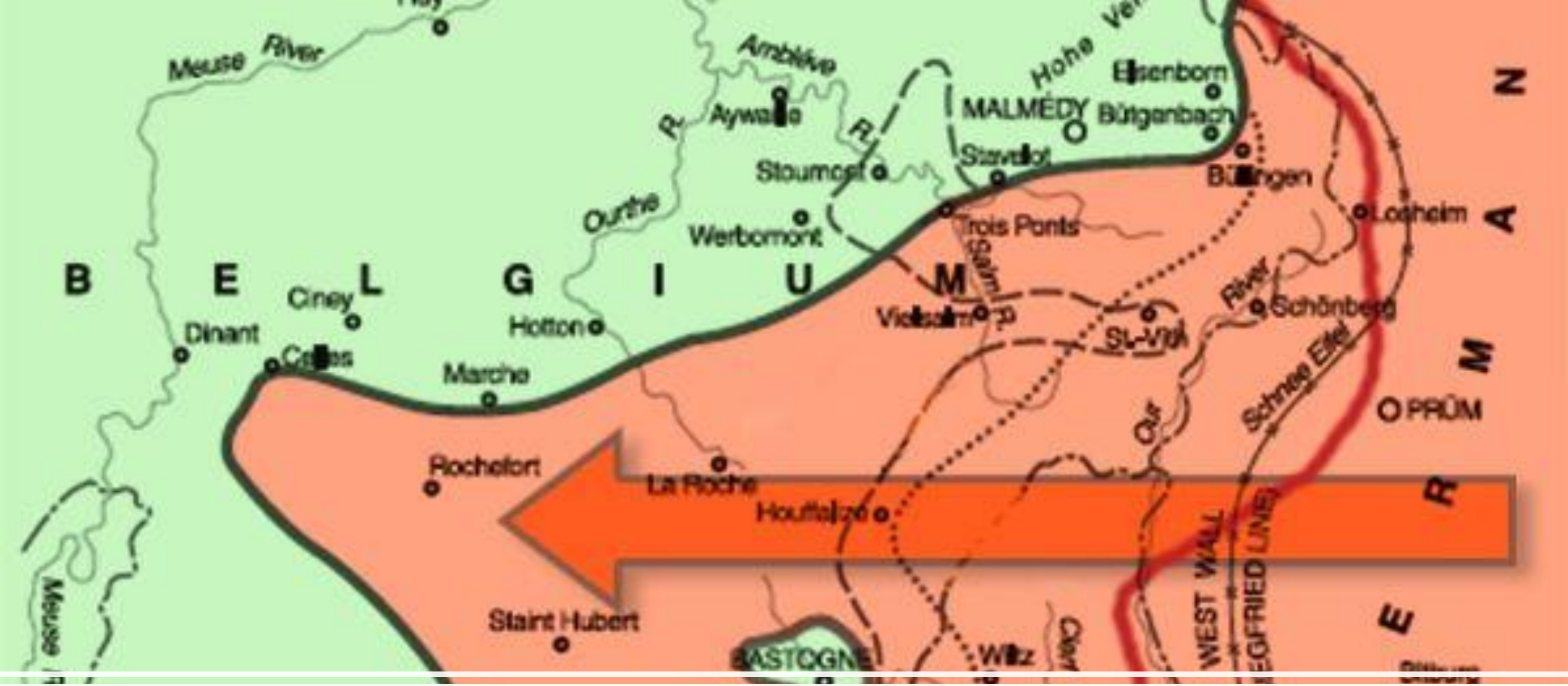
MATH 4753

Rrrrrrr Shivers me Timber
Its Rrrrr time for some
Ere Rrrrrrr
rmarkdown



The Battle of the Bulge

Tiger 2

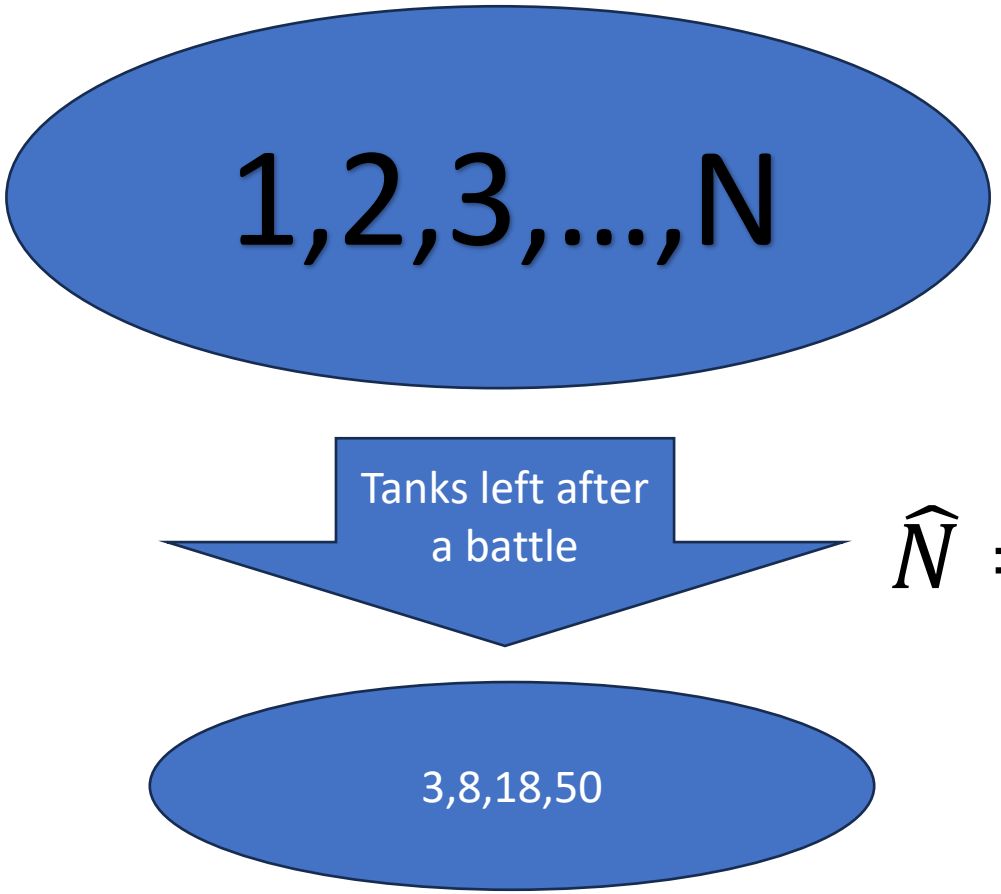
Where were the tanks made?

The Tiger I and Tiger II tanks were primarily manufactured at the Henschel factory in Kassel, Germany.

Serial numbers

Classical

Bayesian



$$\hat{N} = m + \frac{m}{k} - 1$$

$$\hat{N} = \frac{(m-1)(k-1)}{k-2}$$

m = max(sample), k = sample size

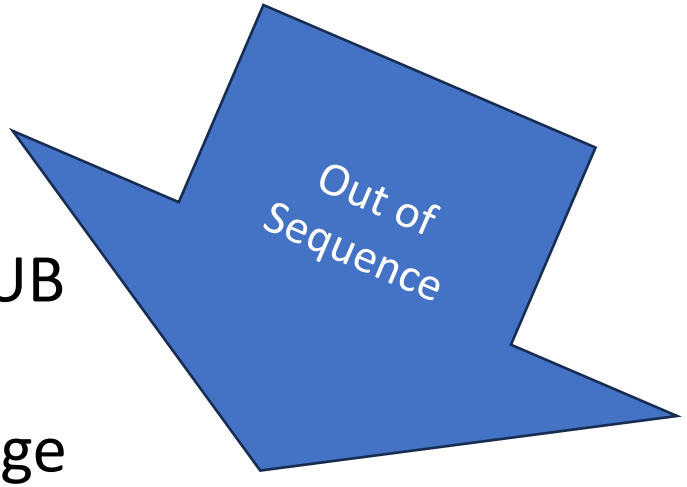
How did the estimation go?

Estimates for some specific months are given as:[7]

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

Today (Possibly)

1. Introduce the course
2. Canvas
3. Syllabus
4. Expectations
5. GIT and GITHUB
6. Clone Intro2R
7. Make R package
8. Lab 1 (Lab order 1,2,3,4,5,6,7,)
9. Chapters 1 and 2 (possibly)



Chapter 1

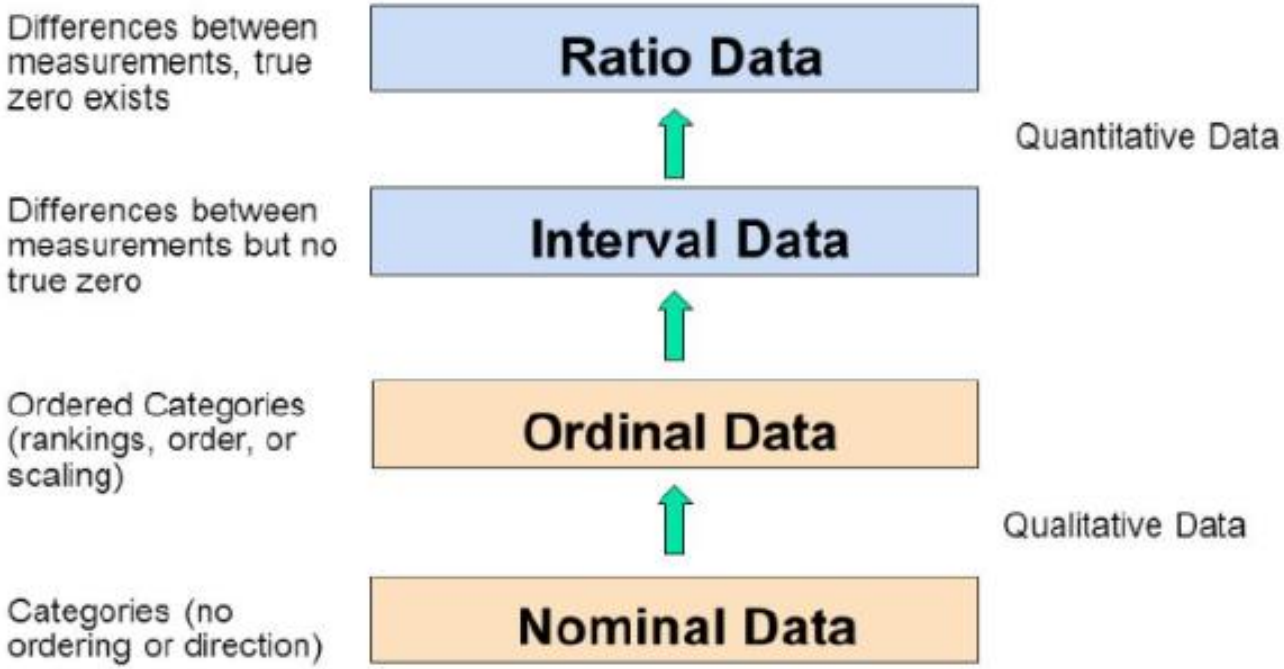
Dr. Wayne Stewart

STATISTICS

for Engineering and the Sciences

SIXTH EDITION

Make sure you do the quiz



Chapter 1

Introduction
“The role of statistics in data analysis”

OBJECTIVE

To identify the role of statistics in the analysis of data from engineering and the sciences

CONTENTS

Definition 1.1

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data.

You will need to be “competent” in R

- This means for this course that you will need to understand
  - Vectors
  - Matrices
  - Data frames
  - Subsetting
  - Assignment
  - Creating Functions
  - Plotting
    - Base R
    - Ggplot

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

WLS.R | lab12025.Rmd | LS12025.Rmd | brms-three-means.R | Untitled7* | exam-analysis.R | Jags_Project_MLR.R | L1Rhelp.R | LiveExample.R

Knit on Save | Knit | Run

Source | Visual | Outline

````
13 ```
14
15 # Introduction
16
17 Lab 1 getting up to speed.
18
19 ```{r}
20 ddt <- read.csv("DDT.csv")
21 ylm <- lm(LENGTH ~ WEIGHT, ddt)
22 r <- residuals(ylm)
23 RSS <- sum(r^2)
24 ```
25
26
27 $\LaTeX$
28
29 $$RSS=\sum_{i=1}^n r_i^2 = `r RSS`$$
30
31
32 ## Tasks
````

Using an R object in Latex

$$RSS = \sum_{i=1}^{n} r_i^2 = `r RSS`$$

29:37 | # Introduction | R Markdown

Dr. Wayne Stewart

2025-01-15

Introduction

Lab 1 getting up to speed.

```
ddt <- read.csv("DDT.csv")
ylm <- lm(LENGTH ~ WEIGHT, ddt)
r <- residuals(ylm)
RSS <- sum(r^2)
```

$\LaTeX$

$$RSS = \sum_{i=1}^{n} r_i^2 = 3870.625477$$

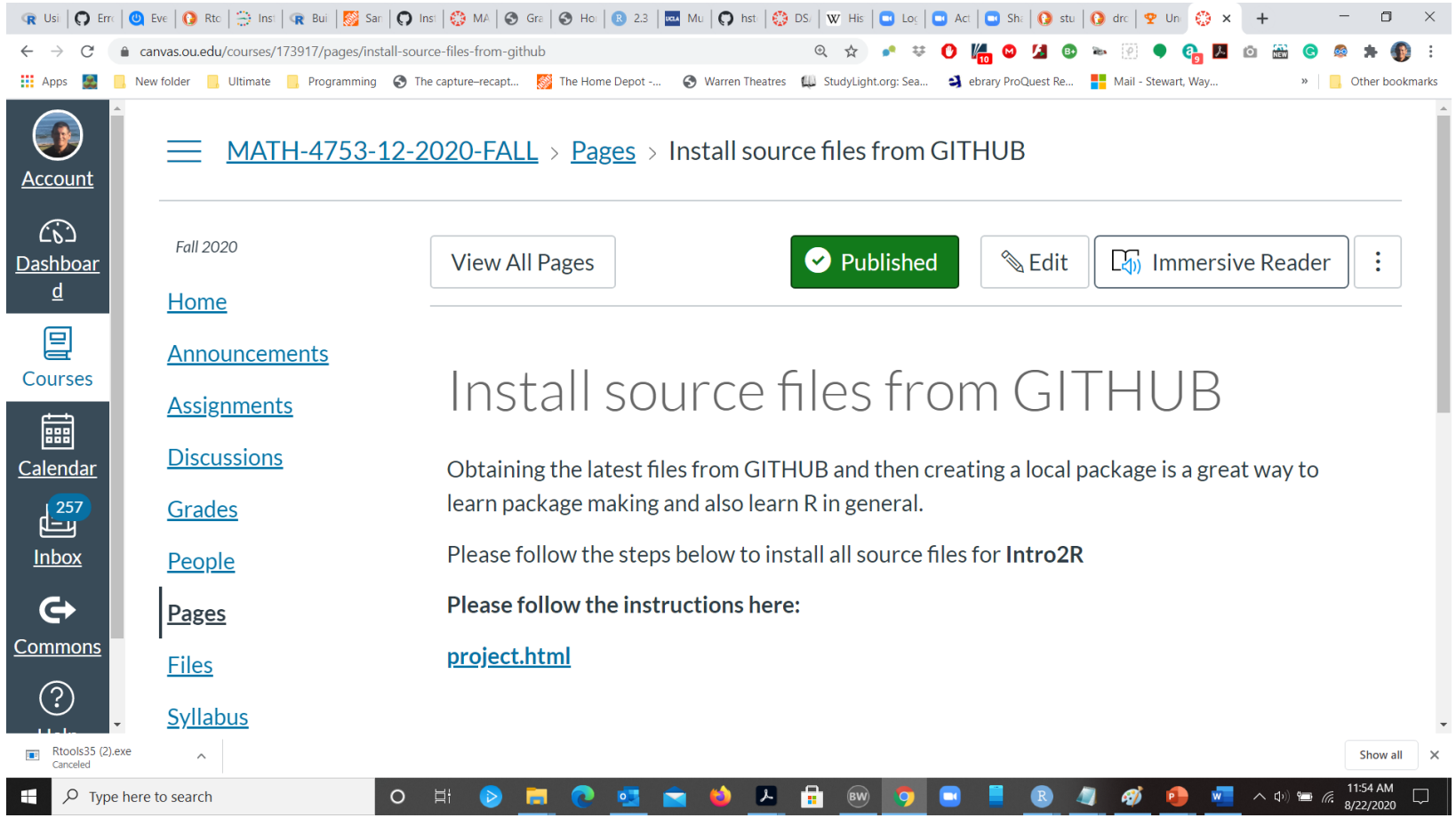
Tasks

Task 1

```
getwd()
```

```
## [1] "D:/2023-MATH4753/Laboratories/Lab1Rmd"
```

Make sure you clone package “Intro2R”



Install my package

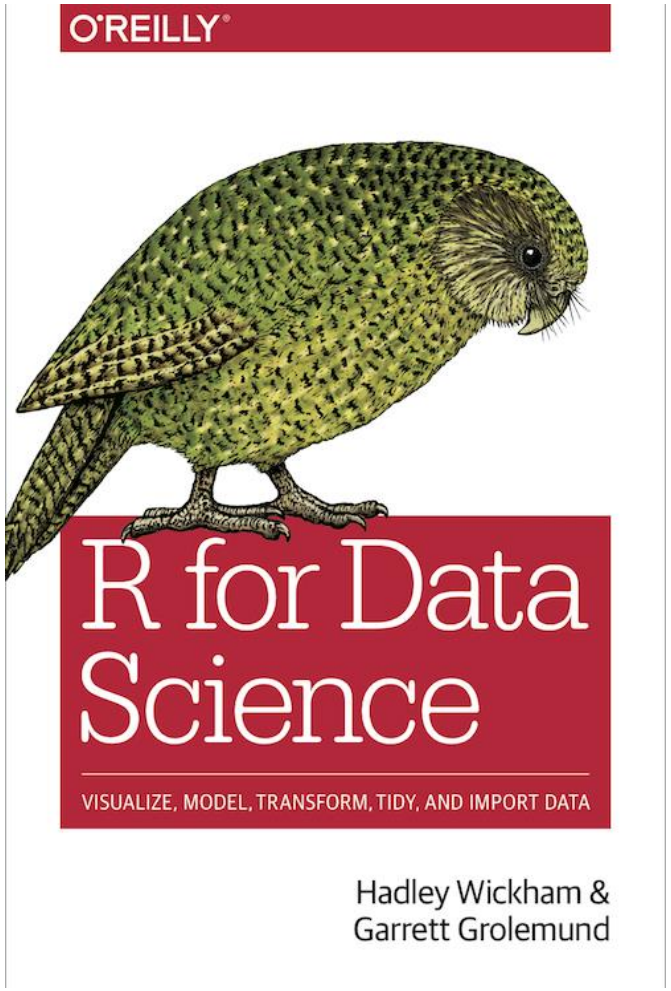
```
devtools::install_github("https://github.com/MATHSTATSOU/Intro2R.git", build_vignettes = TRUE,force = TRUE)
```

The science of statistics is commonly applied to two types of problems:

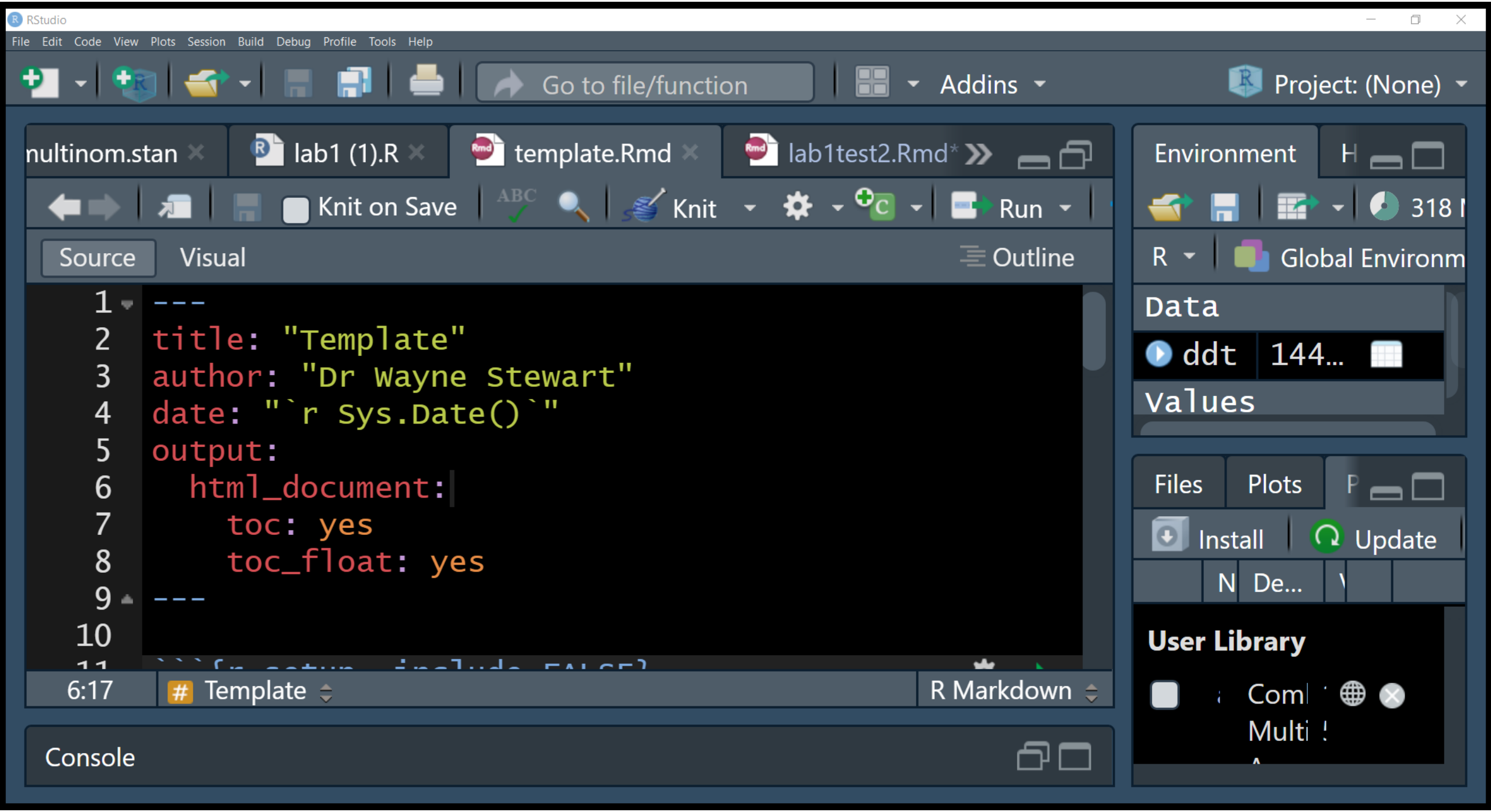
1. Summarizing, describing, and exploring data
2. Using sample data to infer the nature of the data set from which the sample was selected

Plenty of resources – mostly labs!!

Free books



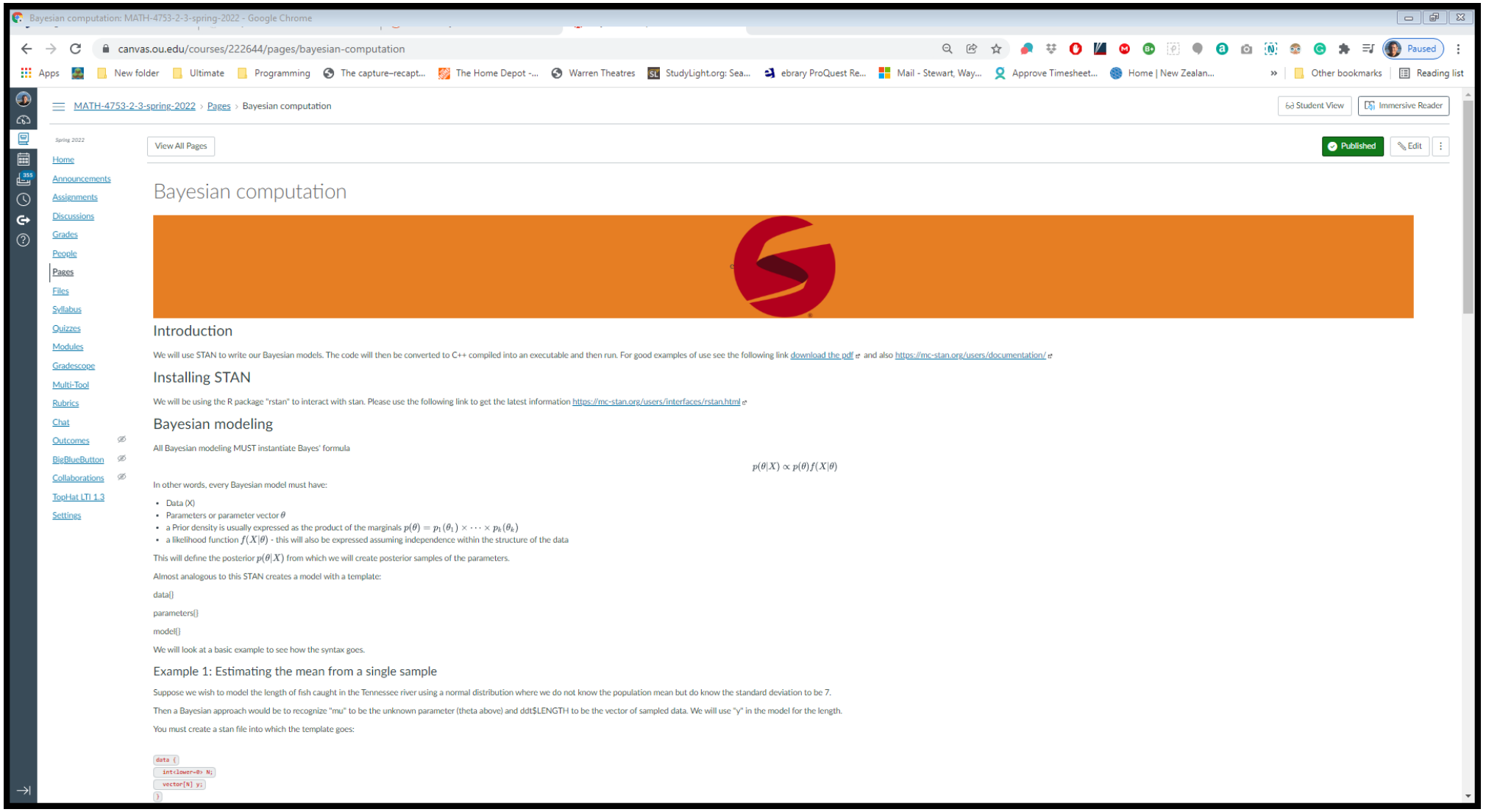
https://bookdown.org/


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
nultinom.stan
lab1 (1).R
template.Rmd
lab1test2.Rmd*
Knit on Save
Knit
Run
Source
Visual
Outline
1 ---
2 title: "Template"
3 author: "Dr Wayne Stewart"
4 date: "`r Sys.Date()`"
5 output:
6 html_document:
7 toc: yes
8 toc_float: yes
9 ---
10
6:17
Template
R Markdown
Console
Environment
318
R
Global Environm
Data
ddt
144...
Values
Files
Plots
Install
Update
N
De...
User Library

New 2024 and continues this semester. Bayesian STATS using STAN

https://canvas.ou.edu/courses/222644/pages/help

You will learn the basic Bayesian Algorithm



Definition 1.2

The branch of statistics devoted to the organization, summarization, and description of data sets is called **descriptive statistics.**

Definition 1.3

The branch of statistics concerned with using sample data to make an inference about a large set of data is called **inferential statistics**.

Definition 1.4

A statistical **population** is a data set (usually large, sometimes conceptual) that is our target of interest.

Definition 1.5

A **sample** is a subset of data selected from the target population.

Definition 1.6

The object (e.g., person, thing, transaction, specimen, or event) upon which measurements are collected is called the **experimental unit**. (*Note:* A population consists of data collected on many experimental units.)

Definition 1.7

A **variable** is a characteristic or property of an individual experimental unit.

Definition 1.8

A **measure of reliability** is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

Four Elements of Descriptive Statistical Problems

1. The population or sample of interest
2. One or more variables (characteristics of the population or sample units) that are to be investigated
3. Tables, graphs, or numerical summary tools
4. Identification of patterns in the data

Five Elements of Inferential Statistical Problems

1. The population of interest
2. One or more variables (characteristics of the experimental units) that are to be investigated
3. The sample of experimental units
4. The inference about the population based on information contained in the sample
5. A measure of reliability for the inference

Types of Data

Definition 1.9

Quantitative data are those that are recorded on a naturally occurring numerical scale, i.e., they represent the quantity or amount of something.

Definition 1.10

Qualitative data are those that cannot be measured on a natural numerical scale, i.e., they can only be classified into categories.

*A finer breakdown of data types into nominal, ordinal, interval, and ratio data is possible. **Nominal** data are qualitative data with categories that cannot be meaningfully ordered. **Ordinal** data are also qualitative data, but a distinct ranking of the groups from high to low exists. **Interval** and **ratio** data are two different types of quantitative data. For most statistical applications (and all the methods presented in this introductory text), it is sufficient to classify data as either quantitative or qualitative.

Learn more about the difference between nominal, ordinal, interval and ratio data with this video by NurseKillam

OK to compute....	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratios, coefficient of variation	No	No	No	Yes

Does measurement scale matter for data analysis?

Knowing the measurement scale for your variables can help prevent mistakes like taking the average of a group of zip (postal) codes, or taking the ratio of two pH values. Beyond that, knowing the measurement scale for your variables doesn't really help you plan your analyses or interpret the results.

Note that sometimes, the measurement scale for a variable is not clear cut. What kind of variable is color? In a psychological study of perception, different colors would be regarded as nominal. In a physics study, color is quantified by wavelength, so color would be considered a ratio variable. What about counts?

Sampling

Once you decide on the type of data—quantitative or qualitative—appropriate for the problem at hand, you'll need to collect the data. Generally, you can obtain the data in three different ways:

1. Data from a *published source*
2. Data from a *designed experiment*
3. Data from an *observational study* (e.g., a *survey*)

Definition 1.11

A **designed experiment** is a data-collection method where the researcher exerts full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned the *treatment* and an untreated (or, *control*) group.

Definition 1.12

An **observational study** is a data-collection method where the experimental units sampled are observed in their natural setting. No attempt is made to control the characteristics of the experimental units sampled. (Examples include *opinion polls* and *surveys*.)

Definition 1.11

A **designed experiment** is a data-collection method where the researcher exerts full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned the *treatment* and an untreated (or, *control*) group.

Definition 1.12

An **observational study** is a data-collection method where the experimental units sampled are observed in their natural setting. No attempt is made to control the characteristics of the experimental units sampled. (Examples include *opinion polls* and *surveys*.)

Definition 1.13

A **representative sample** exhibits characteristics typical of those possessed by the population of interest.

Definition 1.14

A **simple random sample** of n experimental units is a sample selected from the population in such a way that every different sample of size n has an equal chance of selection.

```
> choose(100,5) [1] 75287520
```

Example 1.3

Obtaining a Simple Random Sample for Strength Testing

Suppose you want to randomly sample 5 glass-fiber strips from a lot of 100 strips for strength testing. (Note: In Chapter 3 we demonstrate that there are 75,287,520 possible samples that could be selected.) Use a random number generator to select a simple random sample of 5 glass-fiber strips.

```
> sample(1:100, 5, replace = FALSE)
[1] 43 75 58 55 28
```

Simple Random Sampling

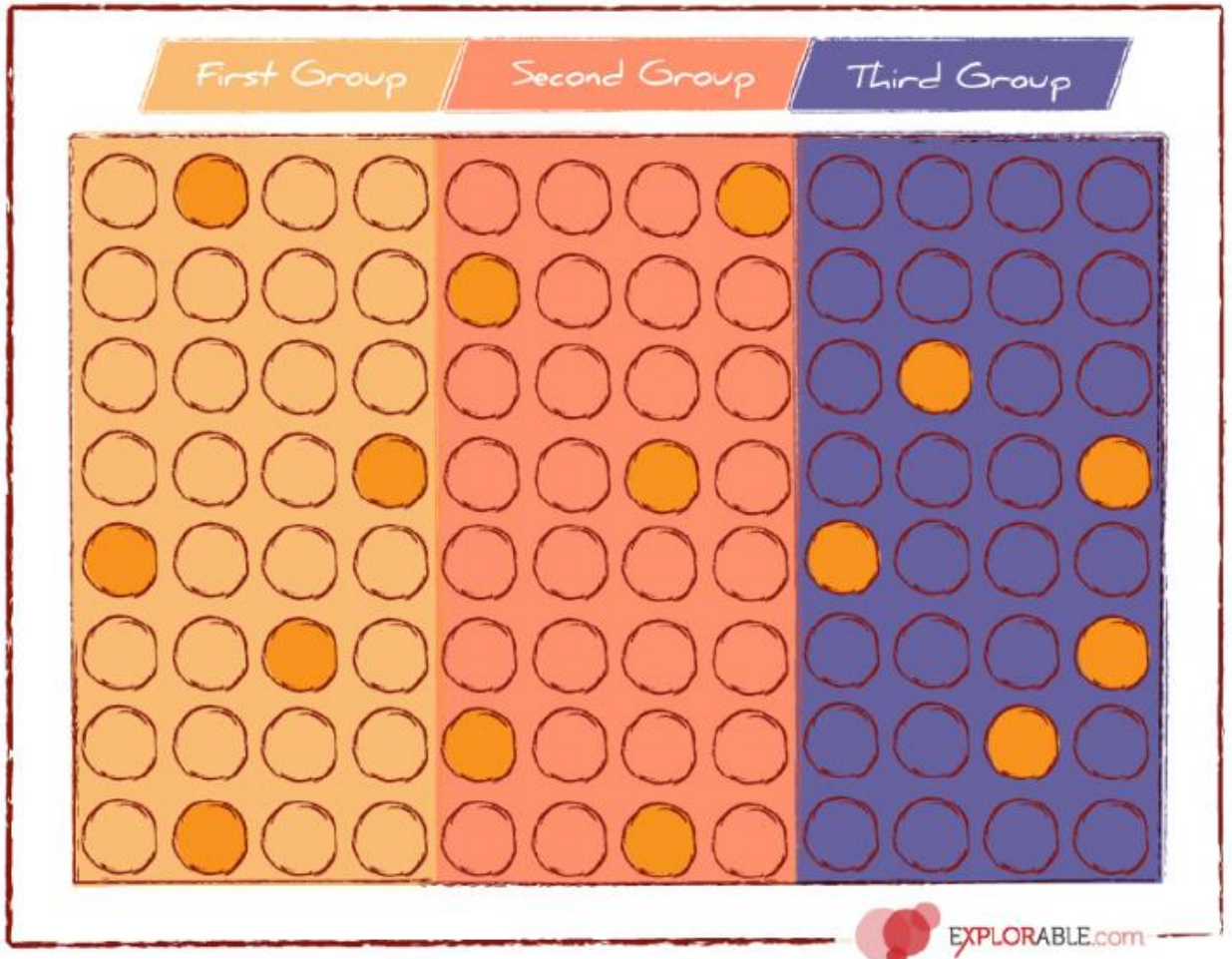
Random Sampling Designs

In addition to simple random samples, there are more complex random sampling designs that can be employed. These include (but are not limited to) **stratified random sampling, cluster sampling**, and **systematic sampling**.

Choose every kth value

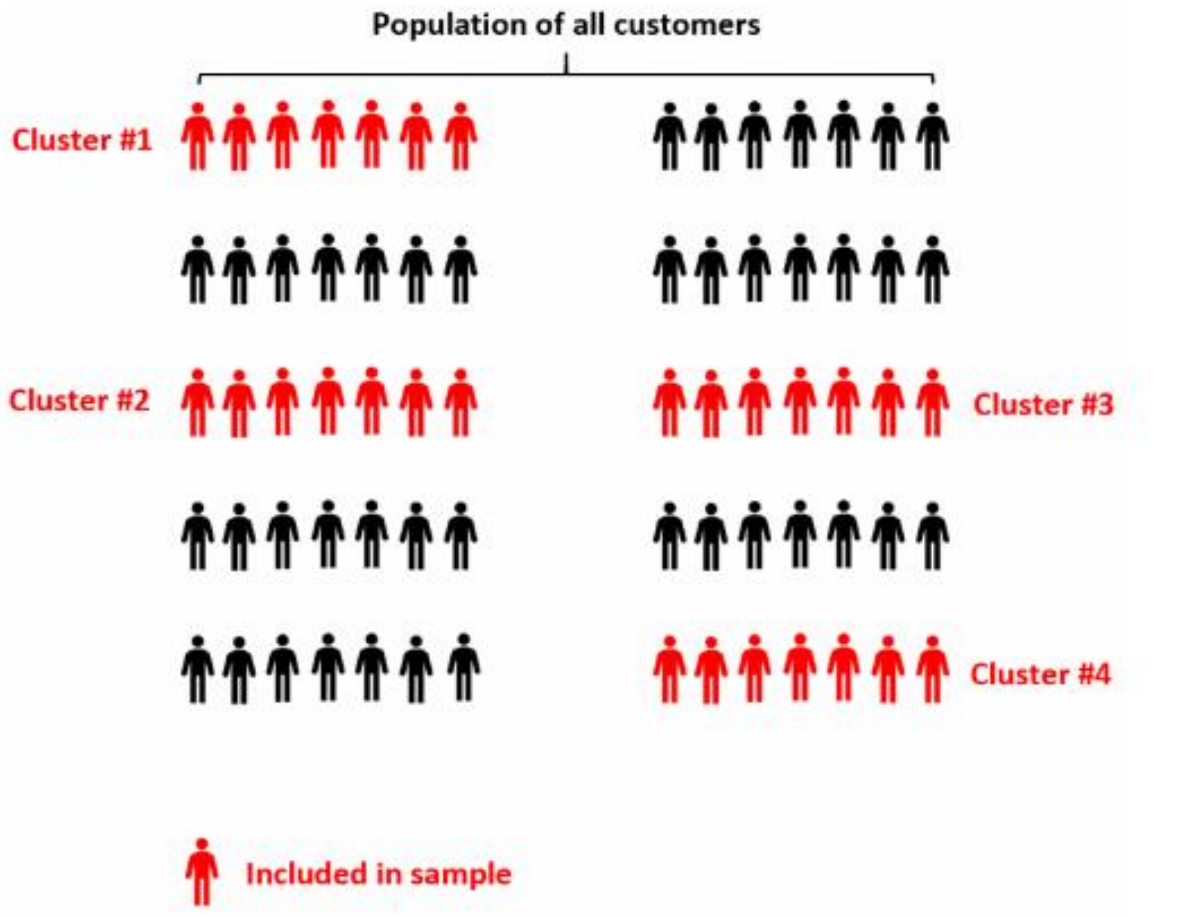
Stratified Sampling

Stratified sampling is a probability sampling technique wherein the researcher divides the entire population into different subgroups or strata, then randomly selects the final subjects proportionally from the different strata.

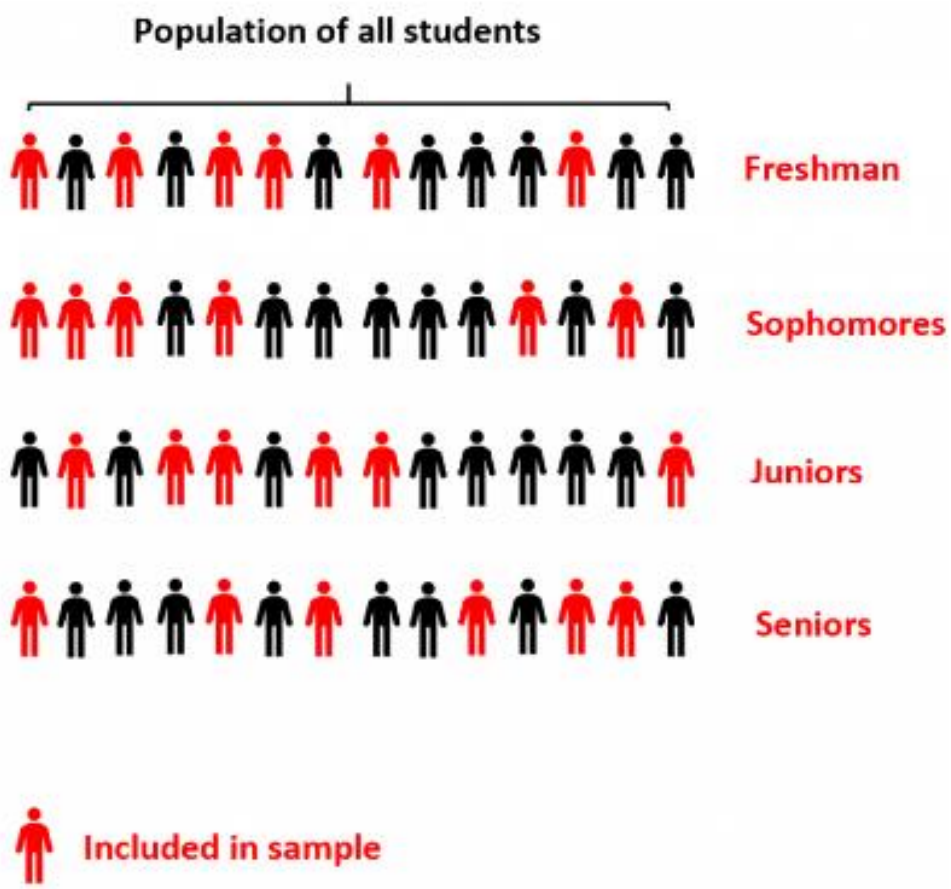


Cluster sampling

https://www.statology.org/cluster-sampling-vs-stratified-sampling/



Stratified sampling



https://www.statology.org/cluster-sampling-vs-stratified-sampling/

Required R skills

Data wrangling

Plotting

Example of sampling: Swiss municipalities

The Swiss municipalities population

Description

This population provides information about the Swiss municipalities in 2003.

Usage

```
data(swissmunicipalities)
```

Format

A data frame with 2896 observations on the following 22 variables:

CT

Swiss canton.

REG

Swiss region.

COM

municipality number.

Nom

municipality name.

Cluster sampling

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

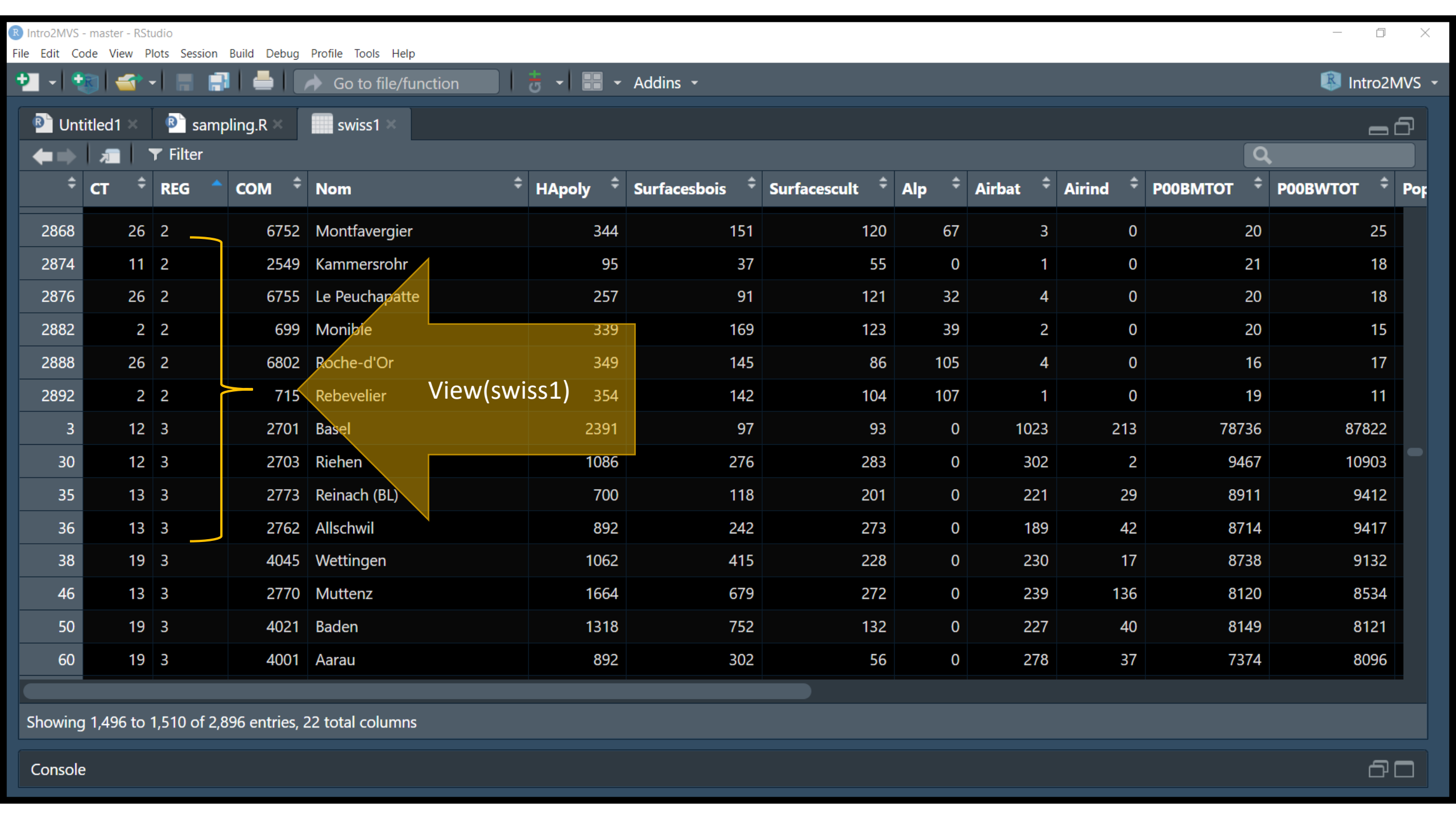
Go to file/function | Addins | Project: (None)

.-Questions-answers-2022.Rmd | vaers_analysis.R | suitability.Rmd | Untitled12* | glm-table.R | ch11-lda.R | sampling.R* | swiss1

Filter

	CT	REG	COM	Nom	HApoly	Surfacesbois	Surfacescult	Alp	Airbat	Airind	P00BMTOT	P00BWTOT	Pop020
1	1	4	261	Zurich	8781	2326	967	0	2884	260	175836	187437	5732
2	25	1	6621	Geneve	1593	67	31	0	773	60	83449	94515	3242
3	12	3	2701	Basel	2391	97	93	0	1023	213	78736	87822	2816
4	2	2	351	Bern	5162	1726	1041	0	1070	212	59727	68907	1939
5	22	1	5586	Lausanne	4136	1635	714	0	856	64	58621	66293	2429
6	1	4	230	Winterthur	6787	2807	1827	0	972	238	43788	46695	1894
7	17	5	3203	St.Gallen	3940	1139	1222	0	812	134	34987	37639	1433
8	3	6	1061	Luzern	1581	408	183	0	524	27	27324	32172	953
9	2	2	371	Biel (BE)	2123	976	196	18	463	108	23447	25208	912
10	2	2	942	Thun	2158	425	694	0	523	137	19071	21306	812
11	2	2	355	Koniz	5099	1567	2621	0	515	53	17952	19830	786
12	24	2	6421	La Chaux-de-Fonds	5566	1572	2221	867	467	56	17535	19481	831
13	10	2	2196	Fribourg	930	164	138	0	301	40	16980	18567	686
14	14	5	2939	Schaffhausen	3102	1467	643	0	467	115	16026	17602	695
15	18	5	3901	Chur	2809	1456	477	49	354	82	15683	17306	688
16	24	2	6458	Neuchatel	1810	998	197	7	323	26	15542	17372	635

Showing 1 to 16 of 2,896 entries, 22 total columns



	CT	REG	COM	Nom	HApoly	Surfacesbois	Surfacescult	Alp	Airbat	Airind	P00BMTOT	P00BWTOT	Pop
2868	26	2	6752	Montfavergier	344	151	120	67	3	0	20	25	
2874	11	2	2549	Kammersrohr	95	37	55	0	1	0	21	18	
2876	26	2	6755	Le Peuchapatte	257	91	121	32	4	0	20	18	
2882	2	2	699	Monible	339	169	123	39	2	0	20	15	
2888	26	2	6802	Roche-d'Or	349	145	86	105	4	0	16	17	
2892	2	2	715	Rebevelier	354	142	104	107	1	0	19	11	
3	12	3	2701	Basel	2391	97	93	0	1023	213	78736	87822	
30	12	3	2703	Riehen	1086	276	283	0	302	2	9467	10903	
35	13	3	2773	Reinach (BL)	700	118	201	0	221	29	8911	9412	
36	13	3	2762	Allschwil	892	242	273	0	189	42	8714	9417	
38	19	3	4045	Wettingen	1062	415	228	0	230	17	8738	9132	
46	13	3	2770	Muttenz	1664	679	272	0	239	136	8120	8534	
50	19	3	4021	Baden	1318	752	132	0	227	40	8149	8121	
60	19	3	4001	Aarau	892	302	56	0	278	37	7374	8096	

Intro2MVS - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Intro2MVS

Source

Console | Terminal | Background Jobs

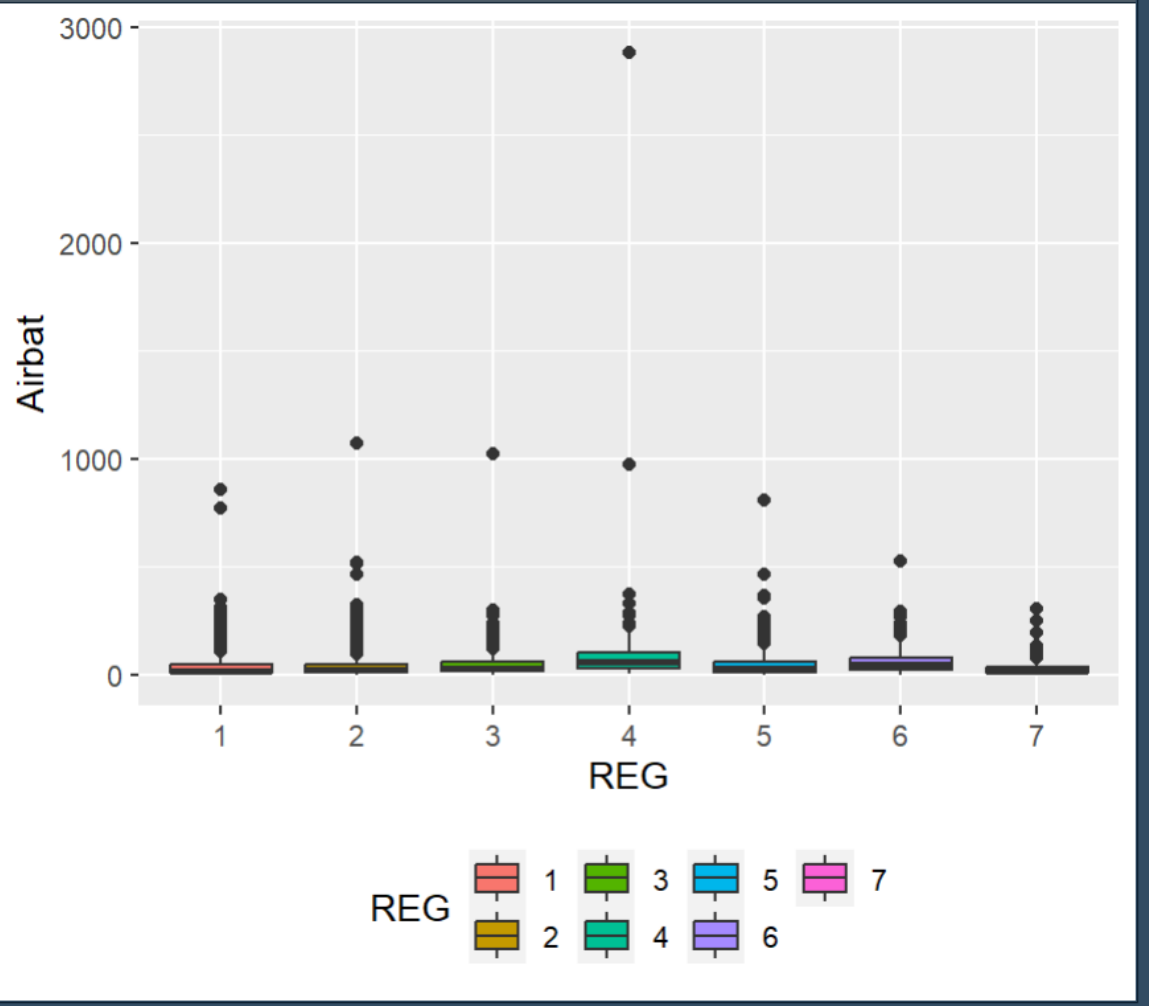
R 4.1.2 · D:/RPACKAGES/Intro2MVS/

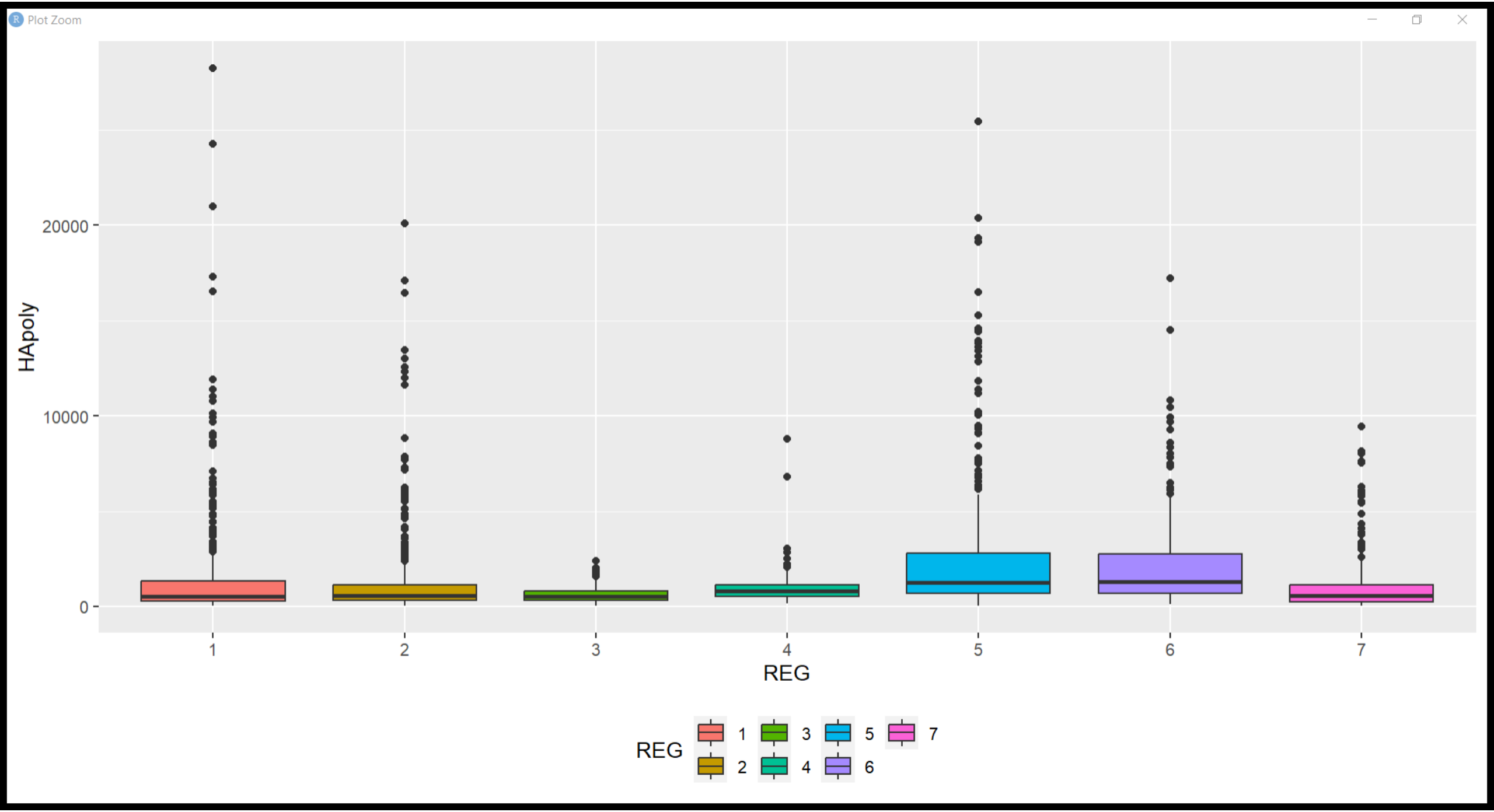
```
> # Cluster sampling
> library(Intro2R)
> library(dplyr)
> library(ggplot2)
> library(sampling)
>
> data("swissmunicipalities") # brng in the data from the packag
e
> swiss1 <- swissmunicipalities %>% mutate(REG = factor(REG)) #
remake REG as a factor
> cl <- sampling::cluster(data = swiss1, # invoke the cluster fu
nction from the package sampling (::)
+                 clustername = "REG",
+                 size = 3,
+                 method = "srswor",
+                 description = TRUE)
Number of selected clusters: 3
Number of units in the population and number of selected units:
2896 1381
>
> g <- ggplot(swiss1 ) # use ggplot layers to construct plot
> g <- g + geom_boxplot(aes(x = REG, y = Airbat, fill = REG)) #
add boxplot layer
> g <- g + theme( legend.position = "bottom") # add a theme modi
fication
> g # print the plot object
```

Environment | History | Connections | Build | Git | Tutorial

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export




Plot Zoom
HApoly
20000
10000
0
1
2
3
4
5
6
7
REG
REG
1
2
3
4
5
6
7

Intro2MVS - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Intro2MVS

Source

Console | Terminal | Background Jobs

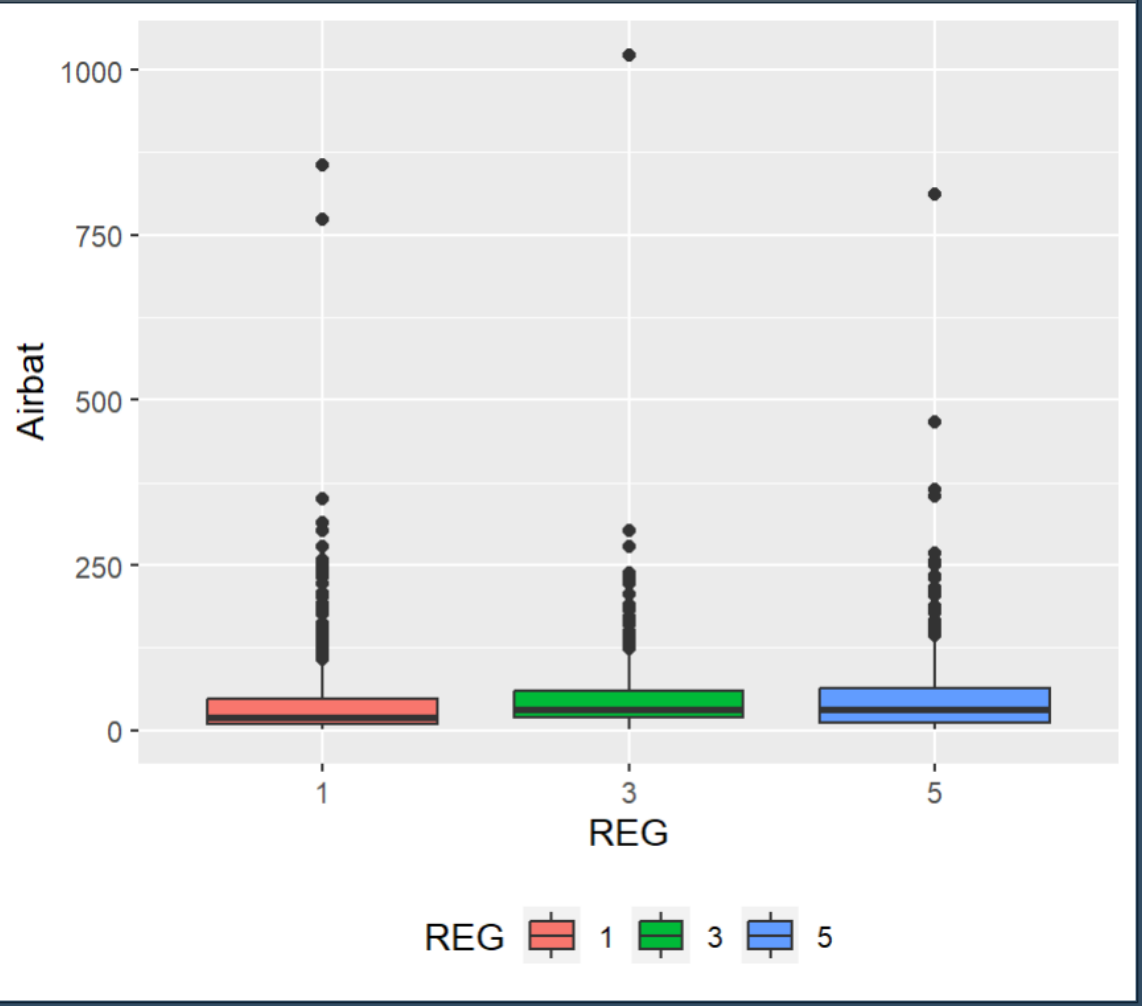
R 4.1.2 · D:/RPACKAGES/Intro2MVS/

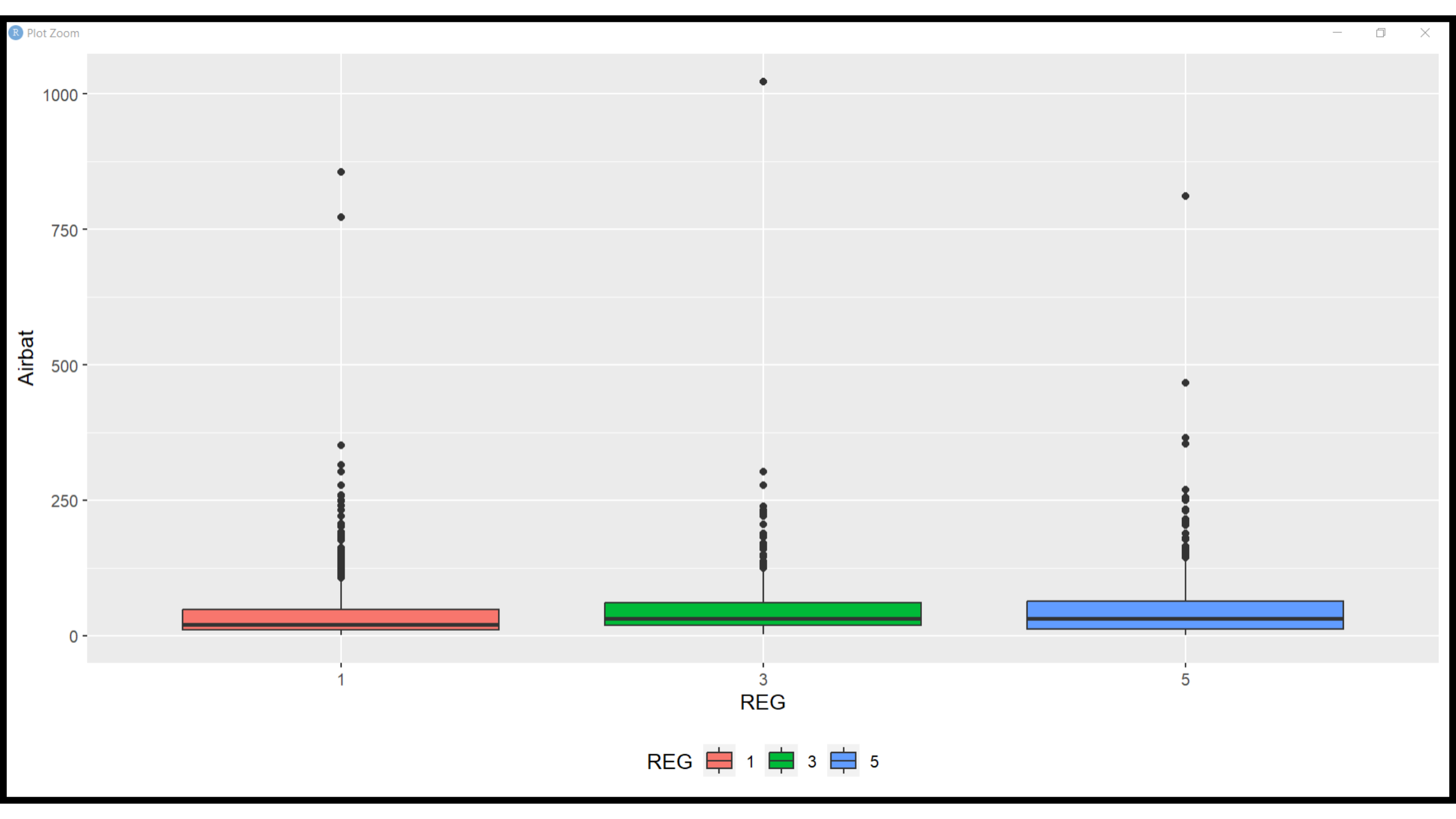
```
> 3/7 # proportion of categories selected
[1] 0.4285714
> data <- getdata(swiss1, cl)
> data2 <- within(data, REG <- factor(REG)) # a different way to r
eform REG as a factor
> g <- ggplot(data2, aes(x = REG, y = Airbat, fill = REG))
> g <- g + geom_boxplot()
> g <- g + theme( legend.position = "bottom" )
> g
>
```

Environment | History | Connections | Build | Git | Tutorial

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export




Plot Zoom
1000
750
500
250
0
Airbat
1
3
5
REG
REG
1
3
5

Intro2MVS - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Intro2MVS

Untitled1 | sampling.R* | swiss1

Source on Save | Run | Source

```
swiss1 <- swissmunicipalities %>% mutate(REG = factor(REG)) # remake REG as a factor
cl <- sampling::cluster(data = swiss1, # invoke the cluster function from the package sampling (::)
                        clustername = "REG",
                        size = 3,
                        method = "srswor",
                        description = TRUE)

g <- ggplot(swiss1 ) # use ggplot layers to construct plot
g <- g + geom_boxplot(aes(x = REG, y = Airbat, fill = REG)) # add boxplot layer
g <- g + theme( legend.position = "bottom") # add a theme modification
g # print the plot object
3/7 # proportion of categories selected
data <- getdata(swiss1, cl)
data2 <- within(data, REG <- factor(REG)) # a different way to reform REG as a factor
g <- ggplot(data2, aes(x = REG, y = Airbat, fill = REG))
g <- g + geom_boxplot()
g <- g + theme( legend.position = "bottom" )
g
```

33:1 (Top Level) R Script

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console | Terminal | Background Jobs

R 4.1.2 · D:/2023-MATH4753/Presentations/Chapter 1/

```
> library(sampling)
> data("swissmunicipalities")
> table(swissmunicipalities$REG)

  1   2   3   4   5   6   7
589 913 321 171 471 186 245
> cl <- sampling::cluster(swissmunicipalities,clustername = "REG", size = 3, method = "srswor", description = TRUE)
Number of selected clusters: 3
Number of units in the population and number of selected units: 2896 978
> head(cl)
  REG ID_unit      Prob
1   3      63 0.4285714
2   3     668 0.4285714
3   3    1507 0.4285714
4   3    1499 0.4285714
5   3    1155 0.4285714
6   3     423 0.4285714
>
```

Stratified sampling

Intro2MVS - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Intro2MVS

Source

Console | Terminal | Background Jobs

R 4.1.2 · D:/RPACKAGES/Intro2MVS/

```
> str <- sampler::ssamp(df = swiss1, n = 20, strata = REG)
> str %>% group_by(REG) %>% summarize(n())
# A tibble: 7 x 2
  REG   `n()`
  <fct> <int>
1 1         4
2 2         6
3 3         2
4 4         1
5 5         3
6 6         1
7 7         2
> swiss1 %>% group_by(REG) %>% summarise(nu = n())
# A tibble: 7 x 2
  REG      nu
  <fct> <int>
1 1       589
2 2       913
3 3       321
4 4       171
5 5       471
6 6       186
7 7       245
>
```

BIAS

Definition 1.15

Selection bias results when a subset of experimental units in the population have little or no chance of being selected for the sample.

Definition 1.16

Nonresponse bias is a type of selection bias that results when data on all experimental units in a sample are not obtained.

Definition 1.17

Intentionally selecting a biased sample in order to produce misleading statistics is considered **unethical statistical practice**.

If the household isn't reached in the morning, try again in the **afternoon** on the **same day**

If the household isn't reached in the afternoon, try again in the **evening** on the **same day**

Morning
7:00am-11:00am

Afternoon
11:00am-3:00pm

Evening
3:00pm-7:00pm

If the household isn't reached in the evening, try again in the **morning** on the **next day**

Try calling back the household in the next time slot for up to **7 attempts**

Read the examples

EXAMPLE 1.5

Method of Data Collection—Study of a Reinforced Concrete Building

As part of a cooperative research agreement between the United States and Japan, a full-scale reinforced concrete building was designed and tested under simulated earthquake conditions in Japan. For one part of the study (published in the *Journal of Structural Engineering*), several U.S. design engineers located on the west coast were asked to evaluate the new design. Of the 48 engineers surveyed, 75% believed the shear wall of the structure to be too lightly reinforced.

a. Identify the data-collection method.
b. Identify the target population.
c. Are the sample data representative of the population?

Definition 1.17

Statistical thinking involves applying rational thought and the science of statistics to critically assess data and inferences.

MATH 4753 + 4773

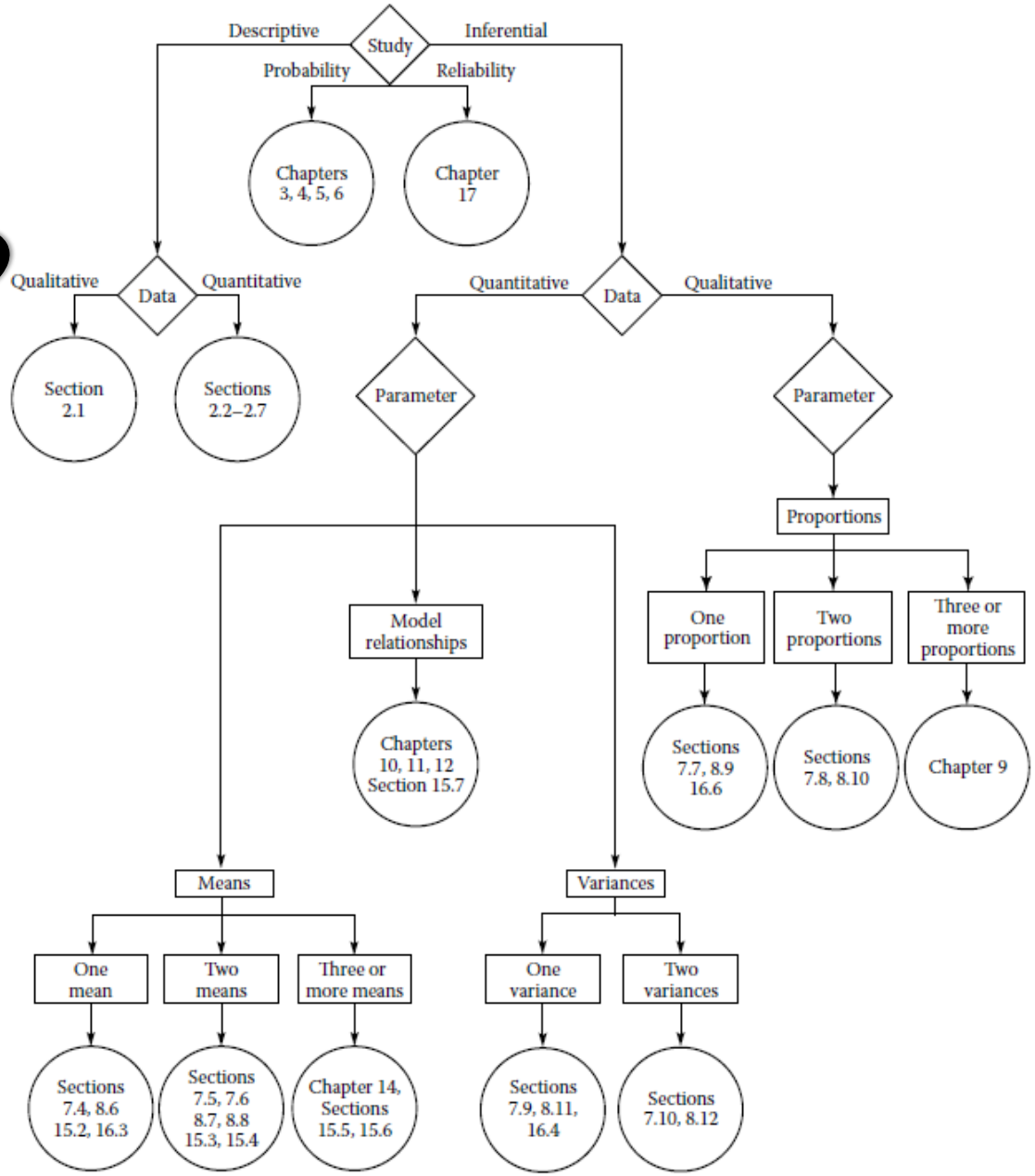


FIGURE 1.3
Flowchart of statistical methods described in the text

Quick Review

Key Terms

Data 2
Descriptive statistics 3
Designed experiment 8
Experimental unit 4
Inference 3
Inferential statistics 3
Measure of reliability 5
Measurement error 4
Nonresponse bias 13
Observational study 9
Population 3
Qualitative data 6
Quantitative data 6
Random number generator 9
Reliability 4
Representative sample 9
Sample 3
Selection bias 13
Simple random sample 9
Statistical thinking 16
Statistics 2
Survey 8
Variable 4

Chapter Summary Notes

- Two types of statistical applications: ***descriptive*** and ***inferential***
- Fundamental elements of statistics: ***population, experimental units, variable, sample, inference, measure of reliability***
- ***Descriptive statistics*** involves summarizing and describing data sets.
- ***Inferential statistics*** involves using a sample to make inferences about a population.
- Two types of data: ***quantitative*** and ***qualitative***
- Three data collection methods: **published source, designed experiment, observational study**.
- Types of random sampling: **simple random sample, stratified random sampling, cluster sampling,** and **systematic sampling**.