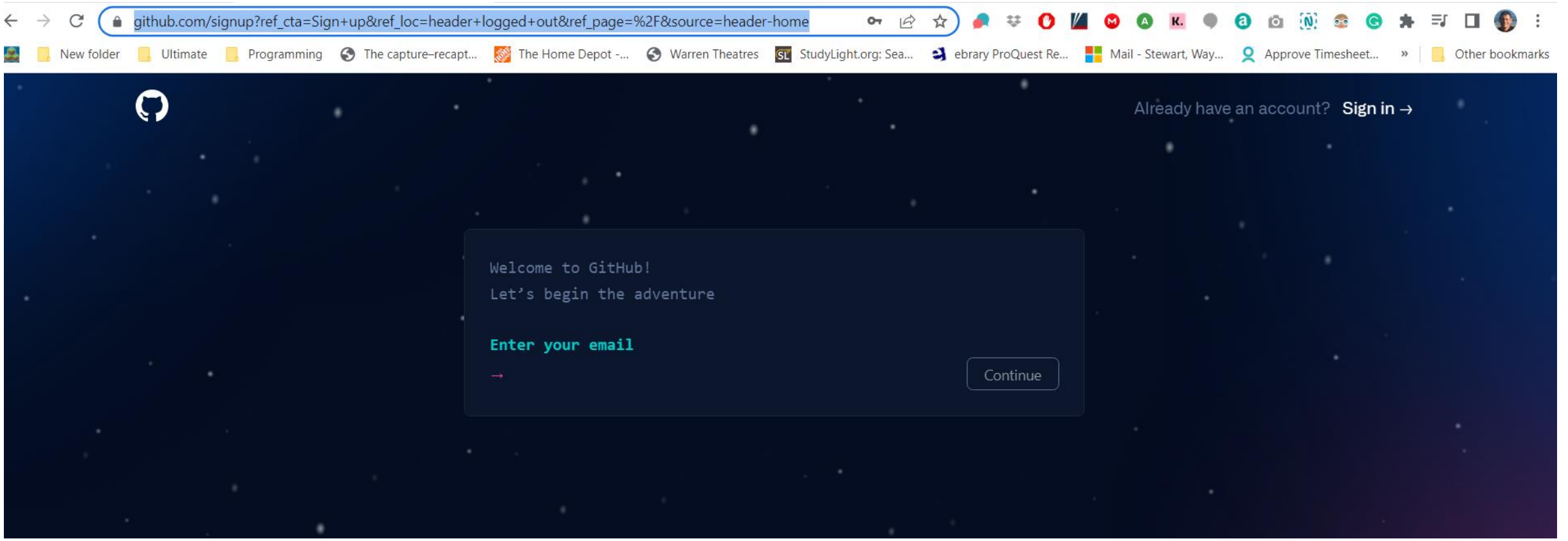
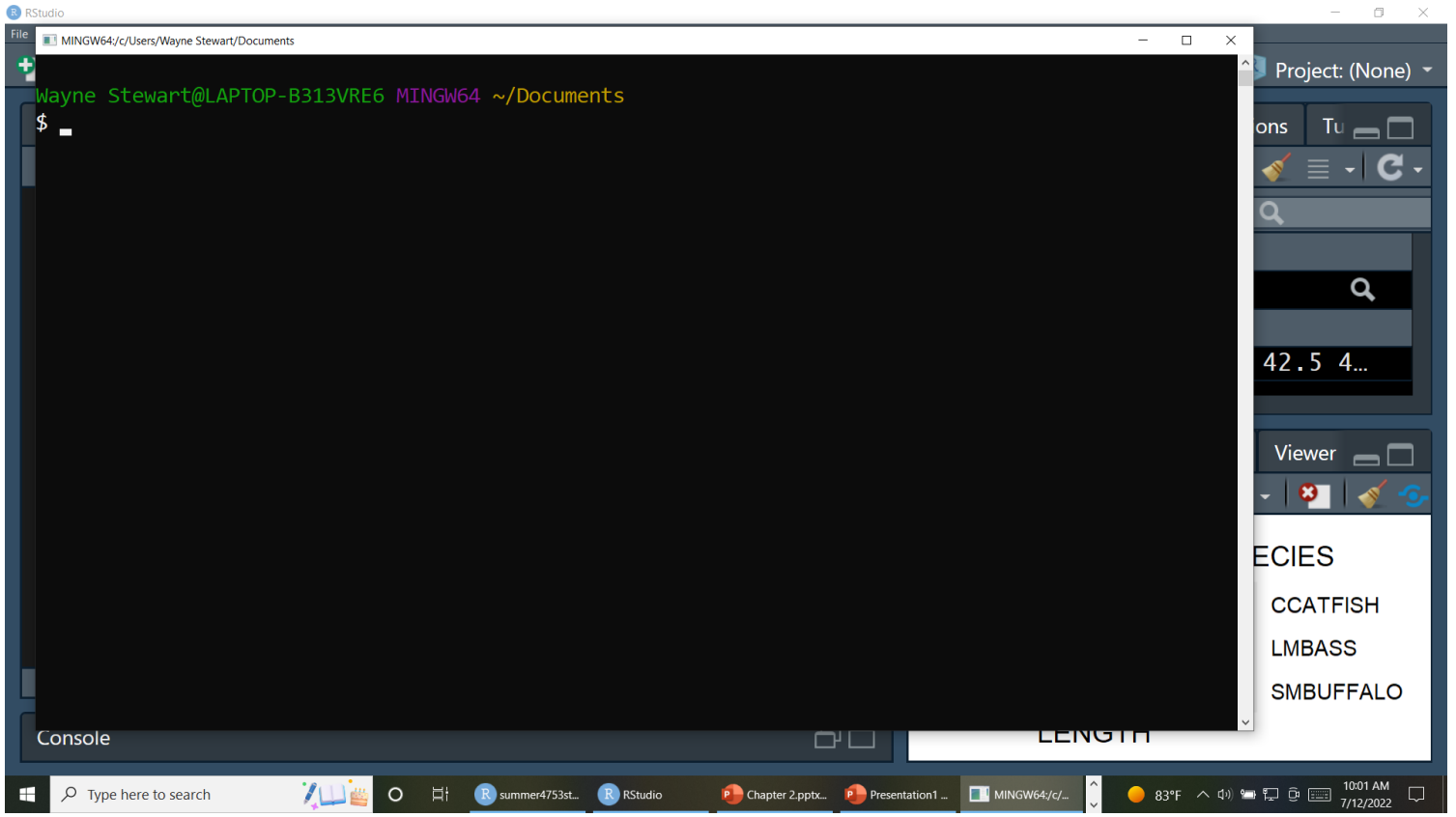
Chapter 2

Dr Wayne Stewart

Obtain account on GITHUB



Tools -> terminal



Configure GIT (Use terminal)

```
git config --global user.name 'Jane Doe'
git config --global user.email 'jane@example.com'
git config --global --list
```

Descriptive Statistics

OBJECTIVE

To present graphical and numerical methods for exploring, summarizing, and describing data

CONTENTS

2.1 Graphical and Numerical Methods for Describing Qualitative Data

2.2 Graphical Methods for Describing Quantitative Data

2.3 Numerical Methods for Describing Quantitative Data

2.4 Measures of Central Tendency

2.5 Measures of Variation

2.6 Measures of Relative Standing

2.7 Methods for Detecting Outliers

2.8 Distorting the Truth with Descriptive Statistics

STATISTICS IN ACTION
Characteristics of Contaminated Fish in the Tennessee River, Alabama

DDT data set

```
ddt <- read.csv(file.choose())
```

DDT

We now return to the U.S. Army Corps of Engineers study of fish contaminated from the toxic discharges of a chemical plant once located on the banks of the Tennessee River in Alabama. The study data are saved in the **DDT** file.

The key questions to be answered are: Where (i.e., what river or creek) are the different species most likely to be captured? What is the typical weight and length of the fish? What is the level of DDT contamination of the fish? Does the level of contamination vary by species? These questions can be partially answered by applying the descriptive methods of this chapter. Of course, the method used will depend on the type (quantitative or qualitative) of the variable analyzed.

To answer these questions we need to wrangle the data.
We will use subsetting rules and a package called “dplyr”

Plots

Qualitative Data

Definition 2.1

A **class** is one of the categories into which qualitative data can be classified.

Definition 2.2

The **category (or class) frequency** for a given category is the number of observations that fall in that category.

Definition 2.3

The **category (or class) relative frequency** for a given category is the proportion of the total number of observations n that fall in that category, i.e.,

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

FATAL

TABLE 2.1 Summary Frequency Table for Cause of Energy-Related Fatal Accidents

Category *(Cause)*	Frequency *(Number of Accidents)*	Relative Frequency *(Proportion)*
Coal mine collapse	9	.145
Dam failure	4	.065
Gas explosion	40	.645
Nuclear reactor	1	.016
Oil fire	6	.097
Other (e.g., Lightning, Power plant)	2	.032
Totals	62	1.000

Source: "Safety of nuclear power reactors." *World Nuclear Association*, May 2012.

Summary of Graphical Descriptive Methods for Qualitative Data

Bar Graph: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.

Pie Chart: The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the class relative frequency.

Pareto Diagram: A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged by height in descending order from left to right.

Plots

Quantitative data

EPAGAS

TABLE 2.2 EPA Mileage Ratings on 100 Cars

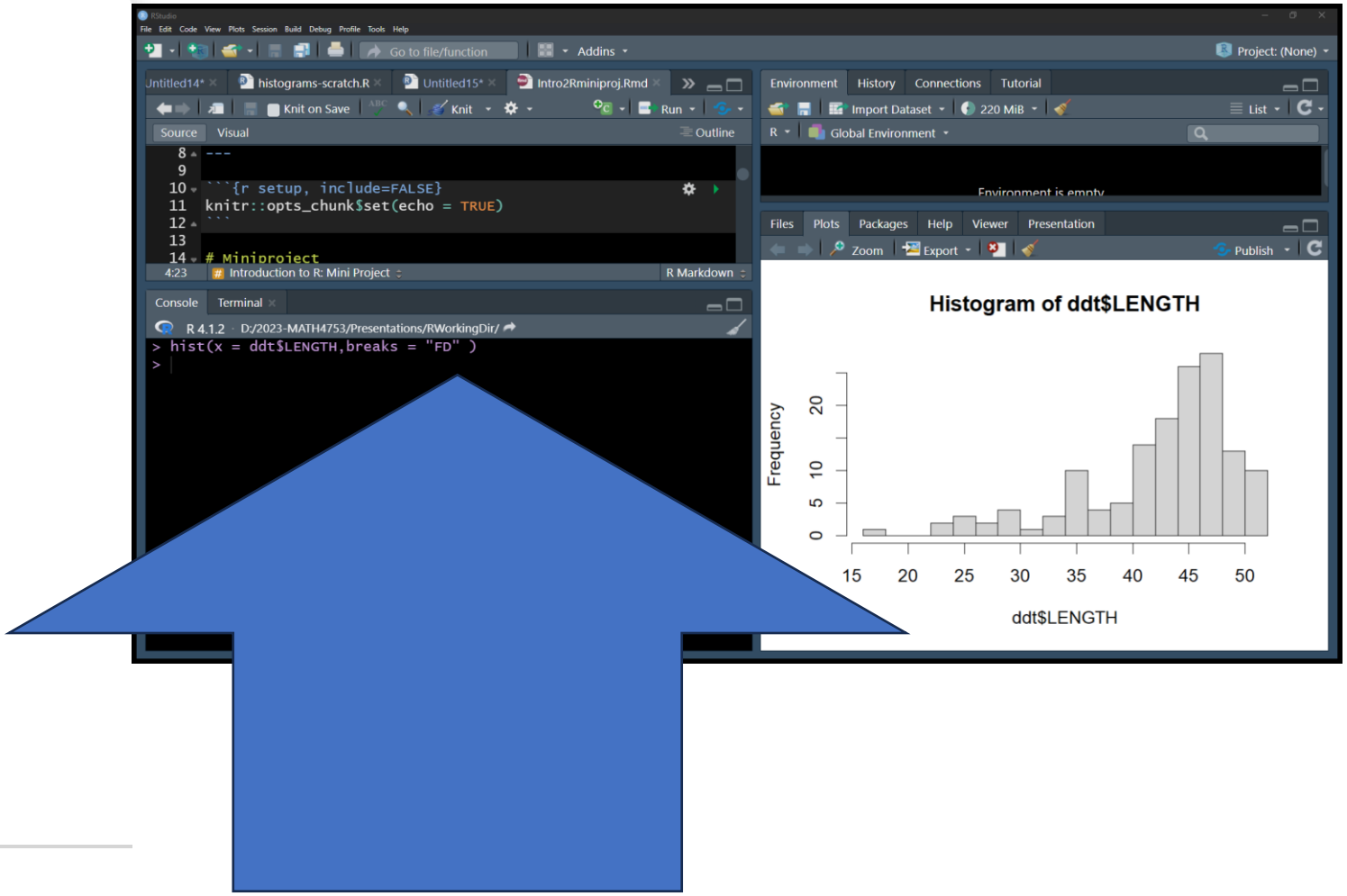
36.3	41.0	36.9	37.1	44.9	36.8	30.0	37.2	42.1	36.7
32.7	37.3	41.2	36.6	32.9	36.5	33.2	37.4	37.5	33.6
40.5	36.5	37.6	33.9	40.2	36.4	37.7	37.7	40.0	34.2
36.2	37.9	36.0	37.9	35.9	38.2	38.3	35.7	35.6	35.1
38.5	39.0	35.5	34.8	38.6	39.4	35.3	34.4	38.8	39.7
36.3	36.8	32.5	36.4	40.5	36.6	36.1	38.2	38.4	39.3
41.0	31.8	37.3	33.1	37.0	37.6	37.0	38.7	39.0	35.8
37.0	37.2	40.7	37.4	37.1	37.8	35.9	35.6	36.7	34.5
37.1	40.3	36.7	37.0	33.9	40.1	38.0	35.2	34.8	39.5
39.9	36.9	32.9	33.8	39.8	34.0	36.8	35.0	38.1	36.9

Histograms

Determining the Number of Classes in a Histogram

Number of Observations in Data Set	Number of Classes
Less than 25	5–6
25–50	7–10
More than 50	11–15

Histogram in R



The default for `breaks` is `"Sturges"`: see `nclass.Sturges`. Other names for which algorithms are supplied are `"Scott"` and `"FD"` / `"Freedman-Diaconis"` (with corresponding functions `nclass.scott` and `nclass.FD`). Case is ignored and partial matching is used. Alternatively, a function can be supplied which will compute the intended number of breaks or the actual breakpoints as a function of `x`.

Steps to Follow in Constructing a Histogram

Step 1 Calculate the range of the data:

$$\text{Range} = \text{Largest observation} - \text{Smallest observation}$$

Step 2 Divide the range into between 5 and 15 classes of equal width. The number of classes is arbitrary, but you will obtain a better graphical description if you use a small number of classes for a small amount of data and a larger number of classes for larger data sets (see the rule of thumb in the previous box). The lowest (or first) class boundary should be located below the smallest measurement, and the class width should be chosen so that no observation can fall on a class boundary.

Step 3 For each class, count the number of observations that fall in that class. This number is called the **class frequency**.

Step 4 Calculate each **class relative frequency:**

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of measurements}}$$

Step 5 The **histogram** is essentially a bar graph in which the categories are classes. In a **frequency histogram**, the heights of the bars are determined by the class frequency. Similarly, in a **relative frequency histogram**, the heights of the bars are determined by the class relative frequency.

Summary of Graphical Descriptive Methods for Quantitative Data

Dot Plot: The numerical value of each quantitative measurement in the data set is represented by a dot on a horizontal scale. When data values repeat, the dots are placed above one another vertically.

Stem-and-Leaf Display: The numerical value of the quantitative variable is partitioned into a "stem" and a "leaf." The possible stems are listed in order in a column. The leaf for each quantitative measurement in the data set is placed in the corresponding stem row. Leaves for observations with the same stem value are listed in increasing order horizontally.

Histogram: The possible numerical values of the quantitative variable are partitioned into class intervals, where each interval has the same width. These intervals form the scale of the horizontal axis. The frequency or relative frequency of observations in each class interval is determined. A vertical bar is placed over each class interval with height equal to either the class frequency or class relative frequency.

Numerical methods for describing

Quantitative data

Definition 2.4

A **statistic** is a numerical descriptive measure computed from sample data.

Definition 2.5

A **parameter** is a numerical descriptive measure of a population.

Definition 2.6

The **arithmetic mean** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the average of the measurements:

$$\frac{\sum_{i=1}^{n} y_i}{n}$$

Typically, the symbol $\bar{y}$ is used to represent the **sample mean** (i.e., the mean of a sample of n measurements), whereas the Greek letter μ represents the **population mean**.

Central tendency

Definition 2.7

The **median** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the middle number when the measurements are arranged in ascending (or descending) order, i.e., the value of y located so that half the area under the relative frequency histogram lies to its left and half the area lies to its right. We will use the symbol m to represent the *sample median* and the symbol τ to represent the *population median*.

Calculating the Median of Small Sample Data Sets

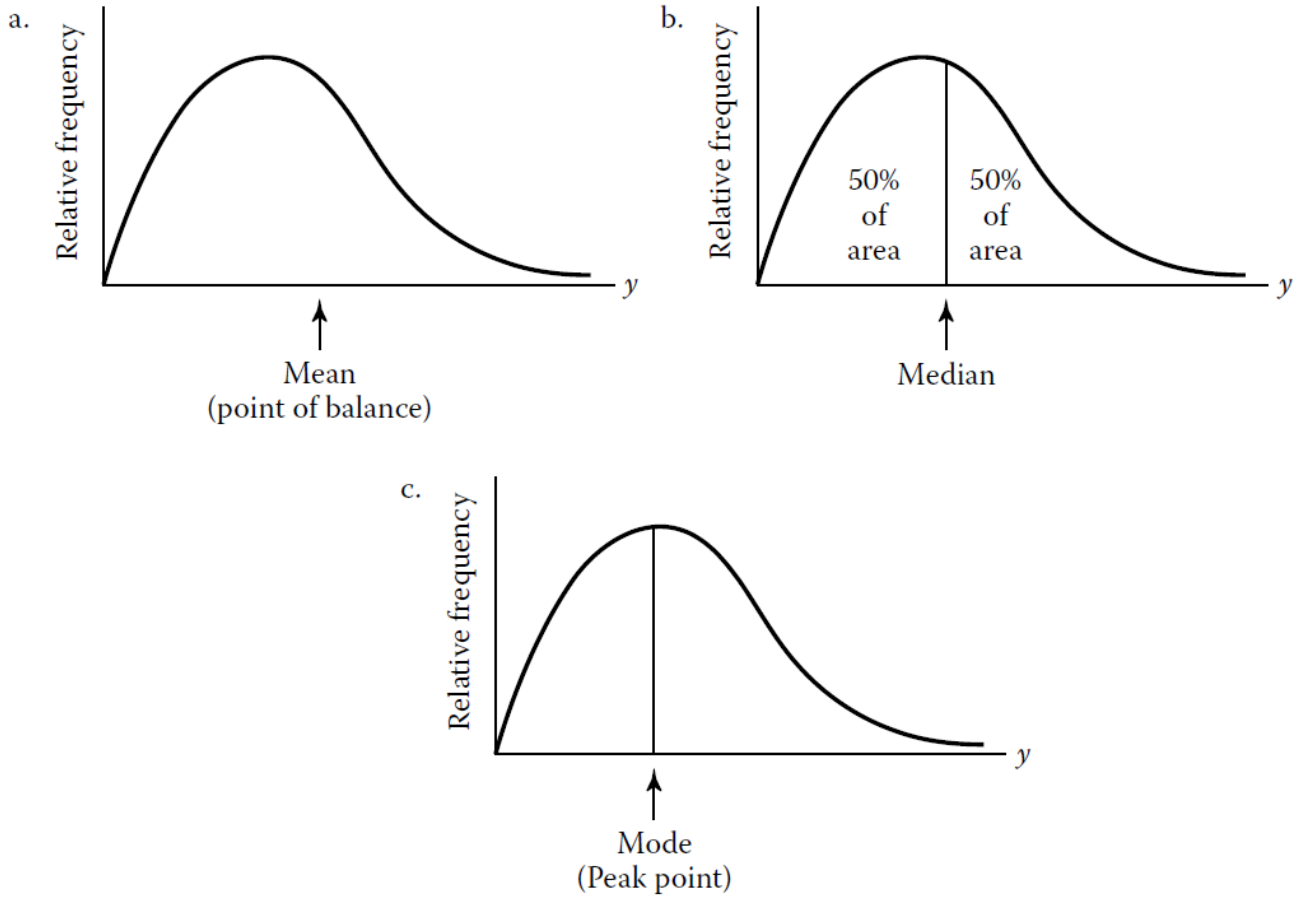
Let $y_{(i)}$ denote the ith value of y when the sample of n measurements is arranged in ascending order. Then the sample median is calculated as follows:

$$m = \begin{cases} y_{[(n+1)/2]} & \text{if } n \text{ is odd} \\ \dfrac{y_{(n/2)} + y_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Definition 2.8

The **mode** of a set of n measurements, $y_1, y_2, \ldots, y_n$, is the value of y that occurs with the greatest frequency.

FIGURE 2.9
Interpretations of the mean, median, and mode for a relative frequency distribution



Example 2.3

Comparing the Mean, Median, and Mode — Earthquake Aftershocks

EARTHQUAKE

Problem: Seismologists use the "aftershock" to describe the smaller earthquakes that follow a main earthquake. Following the Northridge earthquake in 1994, the Los Angeles area experienced 2,929 aftershocks in a three-week period. The magnitudes (measured on the Richter scale) of these aftershocks as well as their inter-arrival times (in minutes) were recorded by the U.S. Geological Survey. (The data are saved in the **EARTHQUAKE** file.) Find and interpret the mean, median, and mode for both of these variables. Which measure of central tendency is better for describing the magnitude distribution? The distribution of inter-arrival times?

Measures of variation

Definition 2.9

The **range** is equal to the difference between the largest and the smallest measurements in a data set:

$$\text{Range} = \text{Largest measurement} - \text{Smallest measurement}$$

Definition 2.10

The **variance** of a **sample** of n measurements, $y_1, y_2, \ldots, y_n$, is defined to be

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}$$

The **population variance** is defined to be

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n}$$

for a finite population with n measurements.

Definition 2.11

The **standard deviation** of a **sample** of n measurements is equal to the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The **population standard deviation** is $\sigma = \sqrt{\sigma^2}$.

Example 2.4

Computing measures of variation

Find the range, variance and standard deviation for the $n = 5$ sample observations: 1, 3, 2, 2, 4.

Solution

The range is simply the difference between the largest (4) and smallest (1) measurement, i.e.,

$$\text{Range} = 4 - 1 = 3$$

To obtain the variance and standard deviation we must first calculate $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} y_i^2$:

$$\sum_{i=1}^{n} y_i = 1 + 3 + 2 + 2 + 4 = 12 \qquad \sum_{i=1}^{n} y_i^2 = (1)^2 + (3)^2 + (2)^2 + (2)^2 + (4)^2 = 34$$

Then the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{34 - \dfrac{(12)^2}{5}}{4} = 1.3$$

and the sample standard deviation is

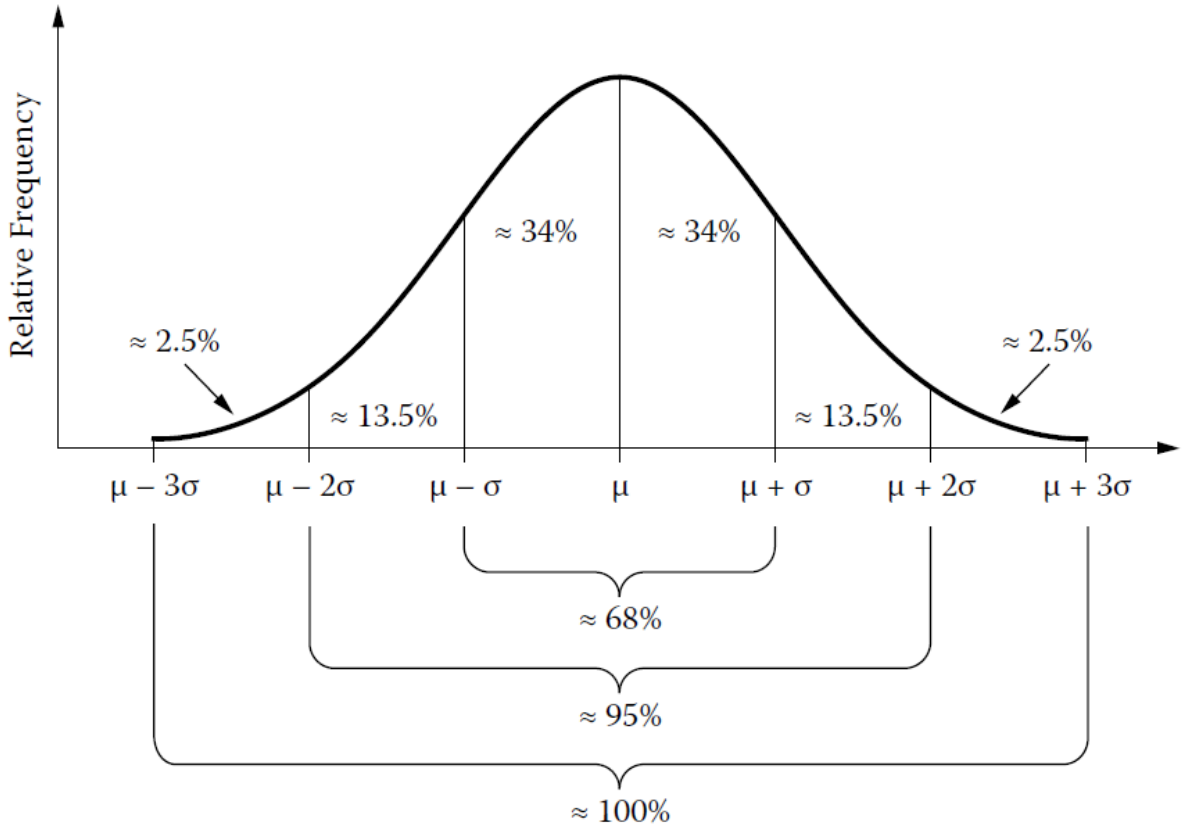
$$s = \sqrt{s^2} = \sqrt{1.3} = 1.1402$$

The empirical rule

The Empirical Rule

If a data set has an approximately mound-shaped, symmetric distribution, then the following rules of thumb may be used to describe the data set (see Figure 2.12a):

1. Approximately 68% of the measurements will lie within 1 standard deviation of their mean (i.e., within the interval $\bar{y} \pm s$ for samples and $\mu \pm \sigma$ for populations).
2. Approximately 95% of the measurements will lie within 2 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 2s$ for samples and $\mu \pm 2\sigma$ for populations).
3. Almost all the measurements will lie within 3 standard deviations of their mean (i.e., within the interval $\bar{y} \pm 3s$ for samples and $\mu \pm 3\sigma$ for populations).


Relative Frequency
≈ 2.5%
≈ 13.5%
≈ 34%
≈ 34%
≈ 13.5%
≈ 2.5%
μ – 3σ
μ – 2σ
μ – σ
μ
μ + σ
μ + 2σ
μ + 3σ
≈ 68%
≈ 95%
≈ 100%

Chebyshev's Rule

Chebyshev's Rule

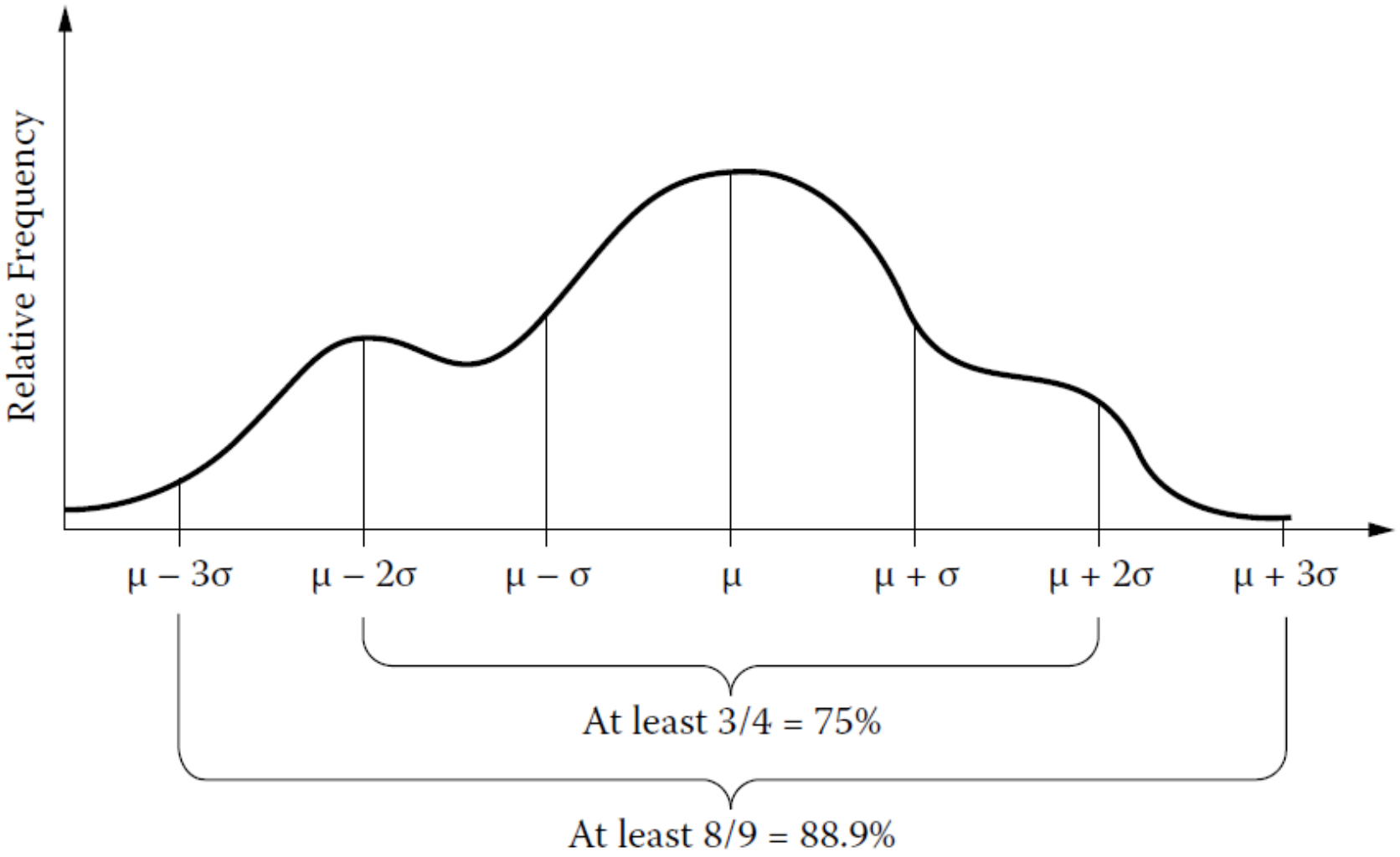
Chebyshev's Rule applies to any data set, regardless of the shape of the frequency distribution of the data (see Figure 2.12b).

a. It is possible that very few of the measurements will fall within 1 standard deviation of the mean, i.e., within the interval $(\bar{y} \pm s)$ for samples and $(\mu \pm \sigma)$ for populations.

b. At least $\frac{3}{4}$ of the measurements will fall within 2 standard deviations of the mean, i.e., within the interval $(\bar{y} \pm 2s)$ for samples and $(\mu \pm 2\sigma)$ for populations.

c. At least $\frac{8}{9}$ of the measurements will fall within 3 standard deviations of the mean, i.e., within the interval $(\bar{y} \pm 3s)$ for samples and $(\mu \pm 3\sigma)$ for populations.

d. Generally, for any number k greater than 1, at least $(1 - 1/k^2)$ of the measurements will fall within k standard deviations of the mean, i.e., within the interval $(\bar{y} \pm ks)$ for samples and $(\mu \pm k\sigma)$ for populations.


Relative Frequency
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ
At least 3/4 = 75%
At least 8/9 = 88.9%

Example 2.5

Applying Rules for Describing the Distribution of Iron Ore Contents

IRONORE

Refer to Example 2.2 (p. 34) and the data on percent iron content of iron-ore specimens. Use a rule of thumb to describe the distribution of iron content measurements. In particular, estimate the number of the 390 iron-ore specimens that have iron content measurements that fall within 2 standard deviations of the mean.

TABLE 2.4 Applying Rules of Thumb to the 390 Iron Content Measurements

k	$\bar{y} \pm ks$	Expected Proportion Using Empirical Rule	Expected Proportion Using Chebyshev's Rule	Actual Proportion
1	(65.05, 66.43)	$\approx$ .68	at least 0	.744
2	(64.36, 67.12)	$\approx$ .95	at least .75	.947
3	(63.67, 67.81)	$\approx$ 1.00	at least .889	.980

Measures of relative standing

Definition 2.12

The **100*p*th percentile** of a data set is a value of y located so that $100p\%$ of the area under the relative frequency distribution for the data lies to the left of the $100p$th percentile and $100(1-p)\%$ of the area lies to its right. (*Note:* $0 \leq p \leq 1$.)

Definition 2.13

The **lower quartile**, Q_L, for a data set is the 25th percentile.

Definition 2.14

The **midquartile** (or median), m, for a data set is the 50th percentile.

Definition 2.15

The **upper quartile**, Q_U, for a data set is the 75th percentile.

Example 2.6

Finding Quartiles—Ingot Freckling

Freckles are defects that sometimes form during the solidification of alloy ingots. A freckle index has been developed to measure the level of freckling on the ingot. A team of engineers conducted several experiments to measure the freckle index of a certain type of superalloy (*Journal of Metallurgy*, Sept. 2004). The data for $n = 18$ alloy tests is shown in Table 2.5. Create a stem-and-leaf display for the data and use it to find the lower quartile for the 18 freckle indexes.

FRECKLE

TABLE 2.5 Freckle Indexes for 18 Superalloys

30.1	22.0	14.6	16.4	12.0	2.4	22.2	10.0	15.1
12.6	6.8	4.1	2.5	1.4	33.4	16.8	8.1	3.2

Source: Yang, W. H., et al., "A freckle criterion for the solidification of superalloys with a tilted solidification front," *Journal of Metallurgy*, Vol. 56, No. 9, Sept. 2004 (Table IV).

Z-score

Definition 2.16

The ***z*-score** for a value y of a data set is the distance that y lies above or below the mean, measured in units of the standard deviation:

$$\textit{Sample } z\textit{-score}: \quad z = \frac{y - \bar{y}}{s}$$

$$\textit{Population } z\textit{-score}: \quad z = \frac{y - \mu}{\sigma}$$

Can this method of detection be improved?

Leave one out statistics

$$Z_{-i} = \frac{x_i - \bar{x}_{-i}}{s_{-i}}$$

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

histograms-scratch.R | z-boxplots.R | ch2.R | wrangling_data.R | ch3.R

Source on Save | Run | Source

```
81
82  zloo <- function(x){ # x vector
83    n <- length(x)
84
85    X <- matrix(data = rep(x,n ),
86                nrow = n,
87                ncol = n,
88                byrow = FALSE)
89
90    diag(X) <- NA # Place NA's on the diagonal
91
92    muloo <- colMeans(x = X,
93                      na.rm = TRUE)
94
95    sdloo <- apply(X = X,
96                   MARGIN = 2,
97                   sd,
98                   na.rm  = TRUE) # put options for the FUN in last place
99
100   zloo <- (x-muloo)/sdloo
101
102   z <- scale(x = x,
103              center = TRUE,
104              scale = TRUE)[,1]
105
106   zz <- matrix(c(x,z,zloo, muloo,sdloo), nrow = n, ncol = 5, byrow = FALSE)
107
108   colnames(zz) <- c("x","z","zloo","muloo","sdloo")
109
110   outloo <- zz[abs(zloo) > 3,]
111   outz <- zz[abs(z) > 3,]
112
113   newcol <- ifelse(test = abs(zloo) > 3,
114                    yes = "Red",
115                    no = ifelse(test = abs(zloo) >=2 & abs(zloo) <= 3,
116                                yes =  "Pink",
117                                no = "Green"))
118
119   pairs(x = zz, pch = 21, bg = newcol, main = "Red outliers, loo")
120
121   list(zz = zz, outloo = outloo, outz = outz )
122 }
123
124 zloo(x = epg$MPG)
```

125:1 (Top Level) R Script

Console

Environment | History | Connections | Tutorial

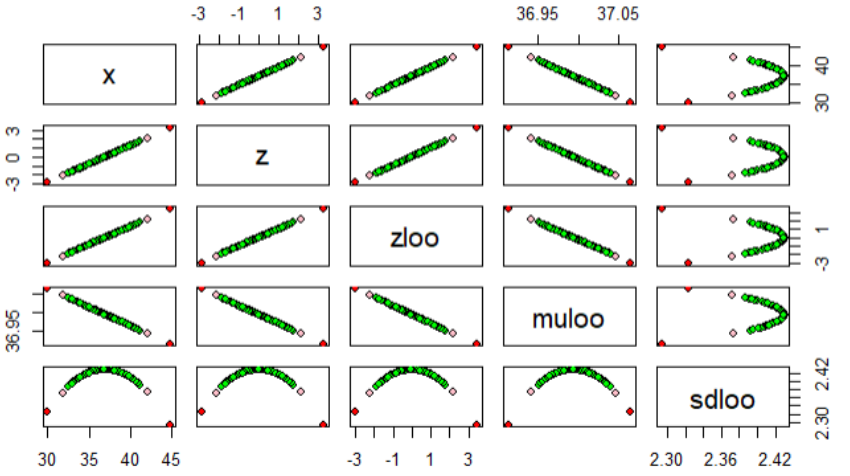
Import Dataset | 300 MiB | List

R | Global Environment

D	num [1:10] 0.89 0.83 1.31 1.48 0.97 0.46 1.23…
Dc	Ord.factor w/ 3 levels "(0,0.527]"<"(0.527,1.…
dird	"D:/MATH4773-5773/DATA/Excel/"
files	chr [1:371] "ACCIDENTS.XLS" "ACIDRAIN.XLS" "A…
freq	'table' int [1:6(1d)] 9 4 40 1 6 2
inc	0.526666666666667
l	int [1:371] 13 12 13 12 13 14 9 10 11 12 ...
newnames	chr [1:371] "ACCIDENTS" "ACIDRAIN" "AFLATOXIN…
prop	'table' num [1:6(1d)] 0.1452 0.0645 0.6452 0.…
tab	'table' int [1:3(1d)] 1 5 4
x	num [1:100] 36.3 41 36.9 37.1 44.9 36.8 30 37…
z	num [1:100] -0.287 1.6568 -0.0389 0.0438 3.26…
Functions	
myconvert	function (xl)
zloo	function (x)

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Source

Console | Terminal | Background Jobs

R 4.1.2 · D:/2023-MATH4753/Presentations/Chapter 2/

```
 [68,] 38.7  0.705571804  0.710897931 36.97677 2.424022
 [69,] 39.0  0.829646564  0.836726928 36.97374 2.421653
 [70,] 35.8 -0.493817546 -0.496898532 37.00606 2.427177
 [71,] 37.0  0.002481495  0.002493869 36.99394 2.430202
 [72,] 37.2  0.085198002  0.085626016 36.99192 2.430112
 [73,] 40.7  1.532736873  1.559179903 36.95657 2.400900
 [74,] 37.4  0.167914509  0.168776100 36.98990 2.429852
 [75,] 37.1  0.043839749  0.044058789 36.99293 2.430178
 [76,] 37.8  0.333347523  0.335199847 36.98586 2.428824
 [77,] 35.9 -0.452459293 -0.455191141 37.00505 2.427663
 [78,] 35.6 -0.576534053 -0.580393975 37.00808 2.426078
 [79,] 36.7 -0.121593265 -0.122208813 36.99697 2.430019
 [80,] 34.5 -1.031474841 -1.042291027 37.01919 2.416976
 [81,] 37.1  0.043839749  0.044058789 36.99293 2.430178
 [82,] 40.3  1.367303859  1.387418038 36.96061 2.406913
 [83,] 36.7 -0.121593265 -0.122208813 36.99697 2.430019
 [84,] 37.0  0.002481495  0.002493869 36.99394 2.430202
 [85,] 33.9 -1.279624361 -1.296884319 37.02525 2.409816
 [86,] 40.1  1.284587352  1.302000220 36.96263 2.409657
 [87,] 38.0  0.416064030  0.418508507 36.98384 2.428055
 [88,] 35.2 -0.741967067 -0.747769959 37.01212 2.423367
 [89,] 34.8 -0.907400081 -0.915779681 37.01616 2.419972
 [90,] 39.5  1.036437831  1.047361500 36.96869 2.416848
 [91,] 39.9  1.201870845  1.216864522 36.96465 2.412227
 [92,] 36.9 -0.038876758 -0.039070920 36.99495 2.430183
 [93,] 32.9 -1.693206896 -1.727097810 37.03535 2.394395
 [94,] 33.8 -1.320982615 -1.339548134 37.02626 2.408471
 [95,] 39.8  1.160512592  1.174396342 36.96566 2.413447
 [96,] 34.0 -1.238266108 -1.254290733 37.02424 2.411118
 [97,] 36.8 -0.080235012 -0.080637755 36.99596 2.430122
 [98,] 35.0 -0.824683574 -0.831686502 37.01414 2.421756
 [99,] 38.1  0.457422283  0.460194721 36.98283 2.427607
[100,] 36.9 -0.038876758 -0.039070920 36.99495 2.430183

$outloo
      x        z      zloo    muloo    sdloo
[1,] 44.9  3.269784  3.481459 36.91414 2.293825
[2,] 30.0 -2.892596 -3.039662 37.06465 2.324155

$outz
      x        z      zloo     muloo     sdloo
44.900000  3.269784  3.481459 36.914141  2.293825

>
```

Environment | History | Connections | Tutorial

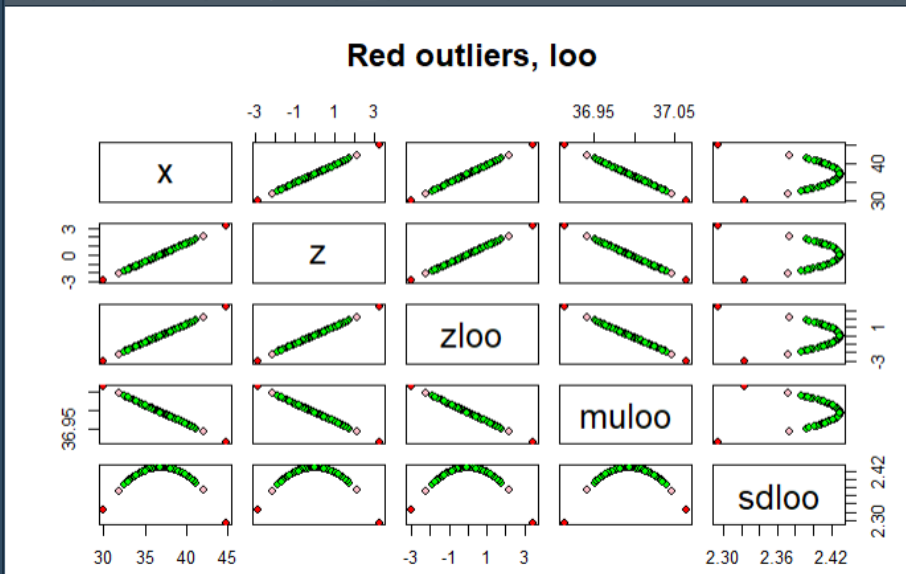
Import Dataset | 301 MiB | List

R | Global Environment

D	num [1:10] 0.89 0.83 1.31 1.48 0.97 0.46 1.23…
Dc	Ord.factor w/ 3 levels "(0,0.527]"<"(0.527,1.…
dird	"D:/MATH4773-5773/DATA/Excel/"
files	chr [1:371] "ACCIDENTS.XLS" "ACIDRAIN.XLS" "A…
freq	'table' int [1:6(1d)] 9 4 40 1 6 2
inc	0.526666666666667
l	int [1:371] 13 12 13 12 13 14 9 10 11 12 ...
newnames	chr [1:371] "ACCIDENTS" "ACIDRAIN" "AFLATOXIN…
prop	'table' num [1:6(1d)] 0.1452 0.0645 0.6452 0.…
tab	'table' int [1:3(1d)] 1 5 4
x	num [1:100] 36.3 41 36.9 37.1 44.9 36.8 30 37…
z	num [1:100] -0.287 1.6568 -0.0389 0.0438 3.26…
Functions	
myconvert	function (xl)
zloo	function (x)

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

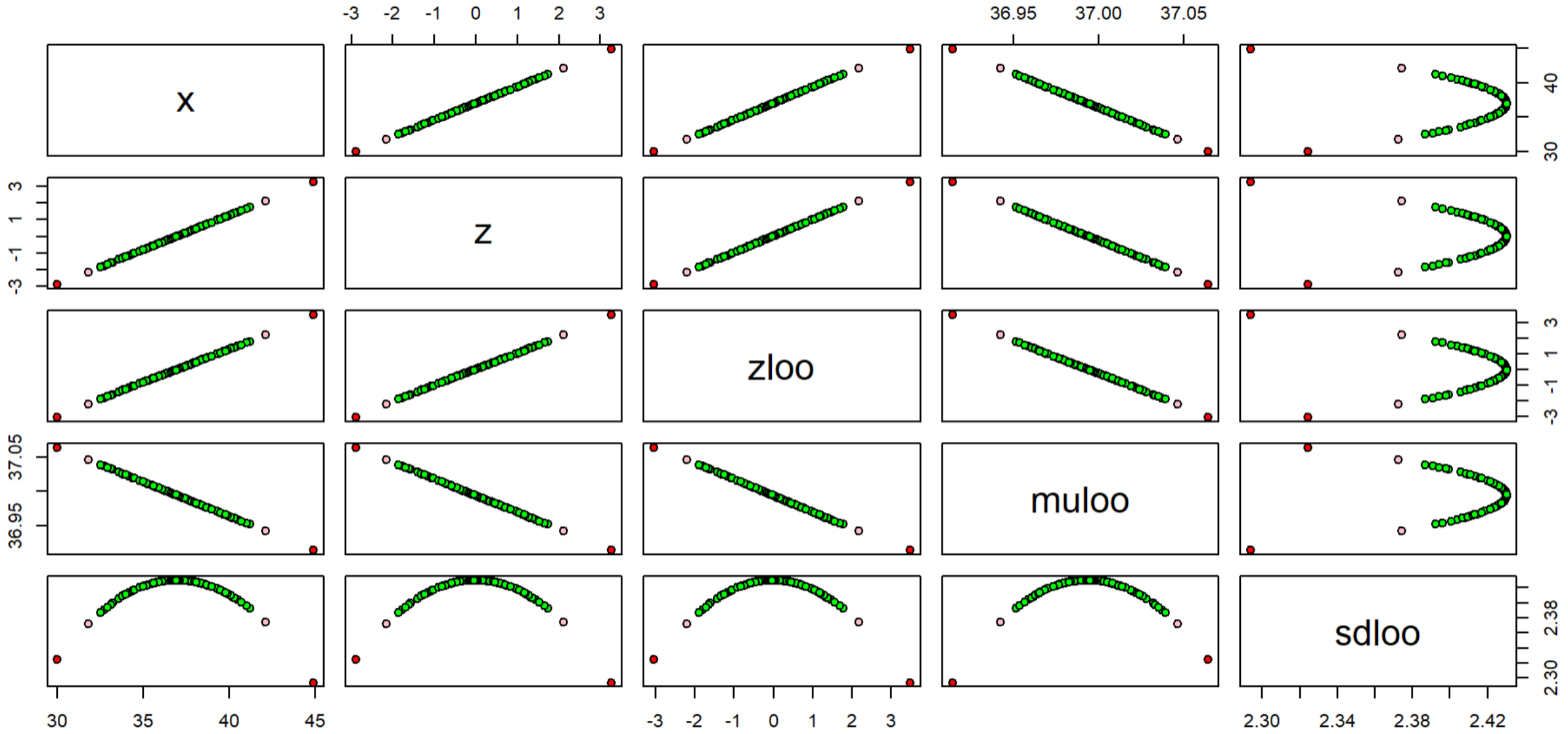
Source

Console | Terminal | Background Jobs

R 4.1.2 · D:/2023-MATH4753/Presentations/Chapter 2/

```
> zstuff$outloo
        x         z      zloo    muloo    sdloo
[1,] 44.9  3.269784  3.481459 36.91414 2.293825
[2,] 30.0 -2.892596 -3.039662 37.06465 2.324155
> zstuff$outz
        x         z      zloo     muloo     sdloo
44.900000  3.269784  3.481459 36.914141  2.293825
>
```

Red outliers, loo



Outliers and their detection

Definition 2.17

An observation y that is unusually large or small relative to the other values in a data set is called an **outlier**. Outliers typically are attributable to one of the following causes:

1. The measurement is observed, recorded, or entered into the computer incorrectly.
2. The measurement comes from a different population.
3. The measurement is correct, but represents a rare (chance) event.

Example 2.8

Deleting Outliers—
Energy-Related Fatalities

FATAL

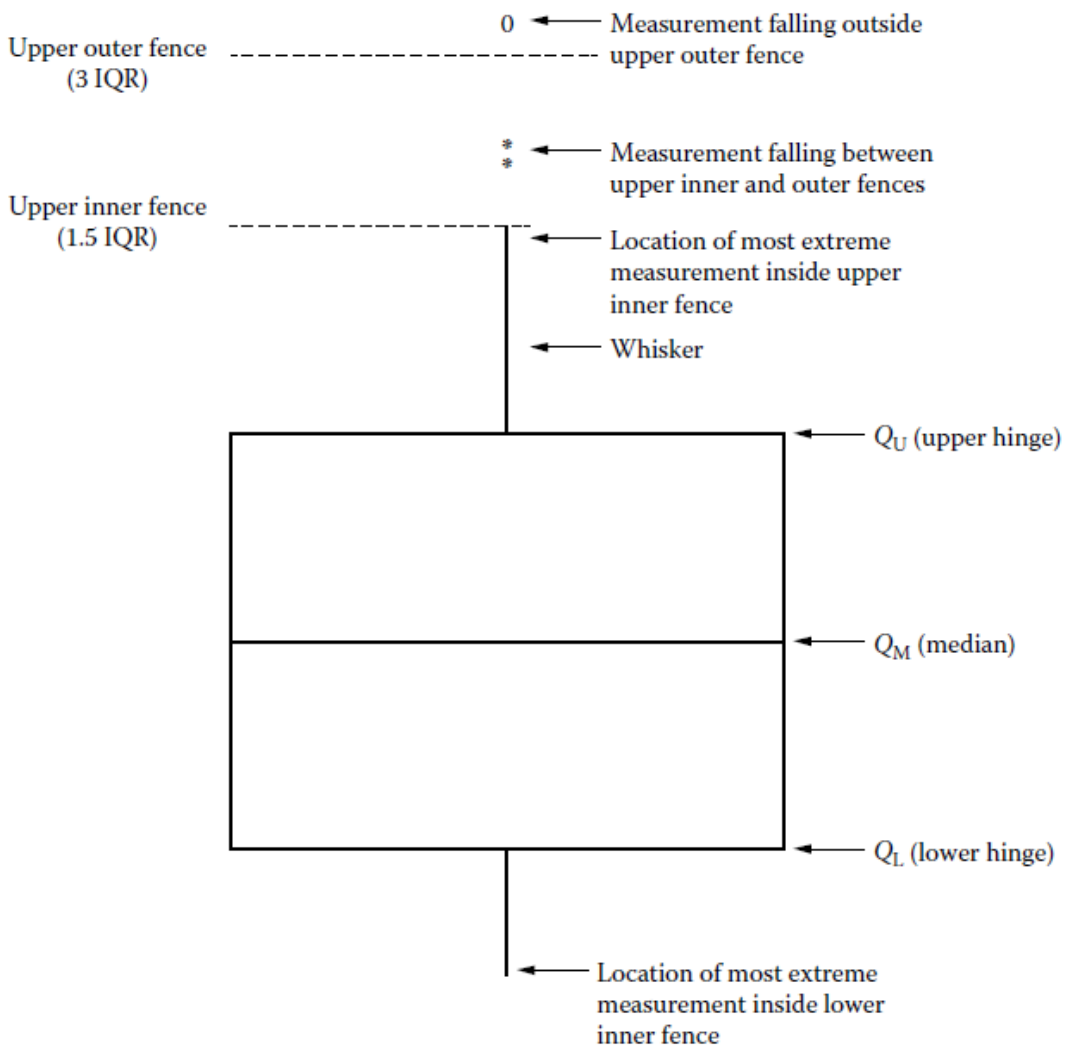
Refer to the sample data on 62 energy-related accidents worldwide since 1979 that resulted in multiple fatalities. (The data are saved in the FATAL file.) In addition to the cause of the fatal energy-related accident, the data set also contains information on the number of fatalities for each accident. The first observation in the data set is a dam failure accident that occurred in India in 1979, killing 2500 people. Is this observation an outlier?

Using boxplots for outlier detection

Definition 2.18

The **interquartile range**, IQR, is the distance between the upper and lower quartiles:

$$\text{IQR} = Q_{\text{U}} - Q_{\text{L}}$$


0
Measurement falling outside upper outer fence
Upper outer fence (3 IQR)
*
*
Measurement falling between upper inner and outer fences
Upper inner fence (1.5 IQR)
Location of most extreme measurement inside upper inner fence
Whisker
Q_U (upper hinge)
Q_M (median)
Q_L (lower hinge)
Location of most extreme measurement inside lower inner fence

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

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

Files | Plots | Packages | Help | Viewer | Presentation

R: Box Plots | Find in Topic

If multiple groups are supplied either as multiple arguments or via a formula, parallel boxplots will be plotted, in the order of the arguments or the order of the levels of the factor (see `factor`).

Missing values are ignored when forming boxplots.

Value

List with the following components:

`stats` a matrix, each column contains the extreme of the lower whisker, the lower hinge, the median, the upper hinge and the extreme of the upper whisker for one group/plot. If all the inputs have the same class attribute, so will this component.

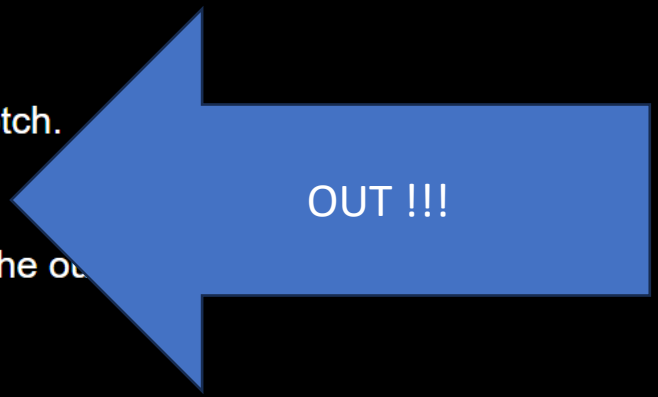
`n` a vector with the number of (non-`NA`) observations in each group.

`conf` a matrix where each column contains the lower and upper extremes of the notch.

`out` the values of any data points which lie beyond the extremes of the whiskers.

`group` a vector of the same length as `out` whose elements indicate to which group the ou

`names` a vector of names for the groups.



References

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

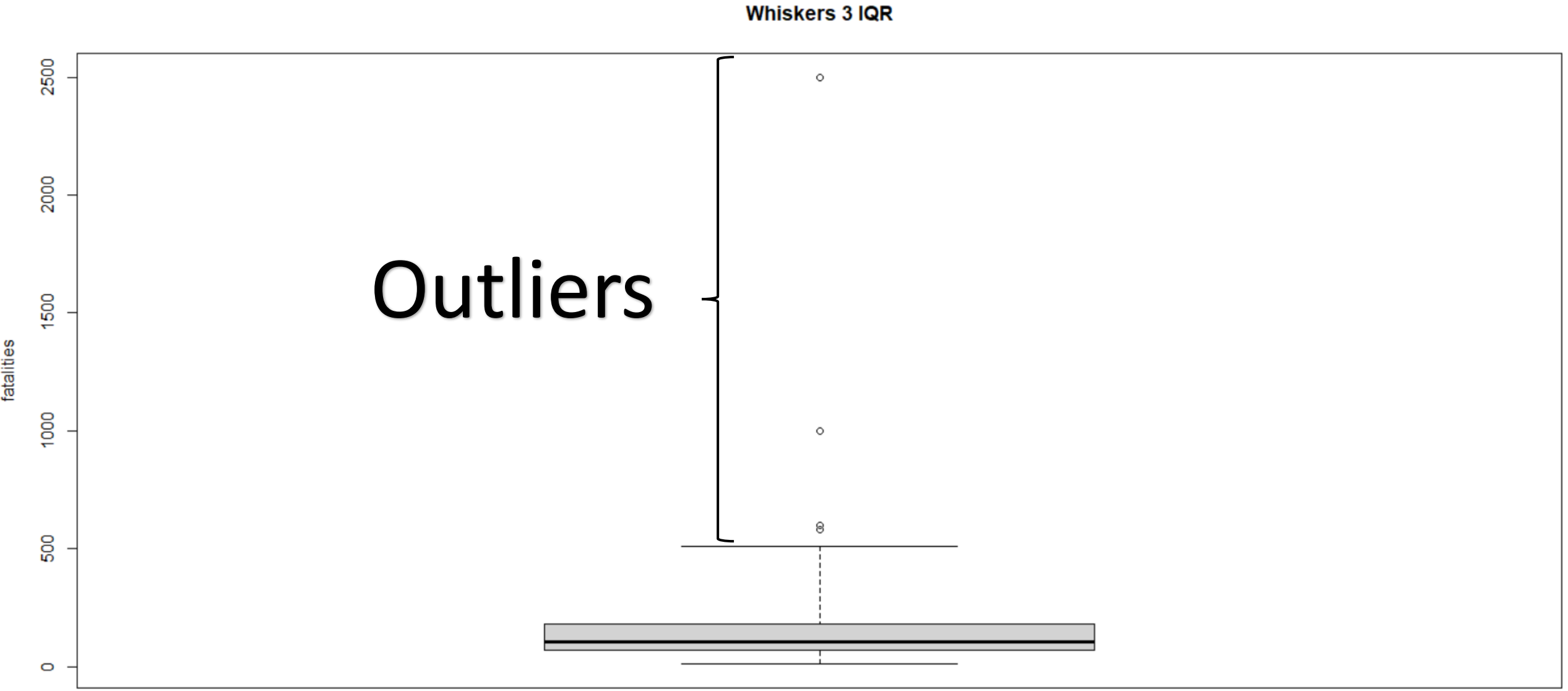
Go to file/function | Addins | Project: (None)

Source

Console | Terminal | Render | Background Jobs

R 4.1.2 · D:/2023-MATH4753/Presentations/RWorkingDir/

```
> #fatal
> graphics.off()
> windows();b3 <- boxplot(fatal, range = 3, main = "Whiskers 3 IQR", ylab = "fatalities")
> b3$out # outliers
[1] 2500 1000  600  580
> windows();b15 <- boxplot(fatal, range = 1.5, main = "Whiskers 1.5 IQR", ylab = "fatalities")
> b15$out
[1] 2500 1000  508  498  600  580  500  500
> setdiff(b15$out,b3$out) # possible outliers
[1] 508 498 500
>
```


Whiskers 3 IQR
Outliers
fatalities
0
500
1000
1500
2000
2500

Quick Review

Key Terms

Arithmetic mean 39
Bar graph 23
Box plots 69
Category frequency 23
Category relative frequency 23
Chebyshev's Rule 47
Class 22
Class interval 31
Dot plot 33
Empirical Rule 47
Hinges 56
Histogram 29
Inner fences 56
Interquartile range (IQR) 56
Lower quartile 52
Mean 68
Measures of central tendency 39
Measures of relative standing 39
Measures of variation 39
Median 68
Midquartile 52
Mode 68
Mound-shaped distribution 41
100*p*th percentile 52
Outer fences 57
Outlier 45
Parameter 39
Pareto diagram 24
Percentile 52
Pie chart 23
Population mean 39
Population standard deviation 46
Population variance 46
Range 46
Sample mean 39
Skewness 41
Standard deviation 46
Statistic 39
Stem-and-leaf display 29
Upper quartile 52
Variance 46
Whiskers 56
z-score 52

Key Formulas

$\frac{\text{Category frequency}}{n}$	Category relative frequency 23
$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	Sample mean 39
$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$	Sample variance 46
$s = \sqrt{s^2}$	Sample standard deviation 46
$z = \frac{y - \bar{y}}{s}$	Sample z-score 52
$z = \frac{y - \mu}{\sigma}$	Population z-score 52
$\text{IQR} = Q_U - Q_L$	Interquartile range 56
$Q_L - 1.5(\text{IQR})$	Lower inner fence 56
$Q_U + 1.5(\text{IQR})$	Upper inner fence 56
$Q_L - 3(\text{IQR})$	Lower outer fence 57
$Q_U + 3(\text{IQR})$	Upper outer fence 57

Chapter Summary Notes

- Graphical methods for qualitative data: **pie chart, bar graph**, and **Pareto diagram**
- Graphical methods for quantitative data: **dot plot, stem-and-leaf display**, and **histogram**
- Numerical measures of central tendency: **mean, median**, and **mode**

- Numerical measures of variation: **range, variance**, and **standard deviation**
- Sample numerical descriptive measures are called **statistics**.
- Population numerical descriptive measures are called **parameters**.
- Rules for determining the percentage of measurements in the interval (mean) $\pm$ 2 (std. dev.): **Chebyshev's Rule** (at least 75%) and **Empirical Rule** (approximately 95%)
- Measures of relative standing: **percentile score** and z**-score**
- Methods for detecting outliers: **box plots** and z**-scores**