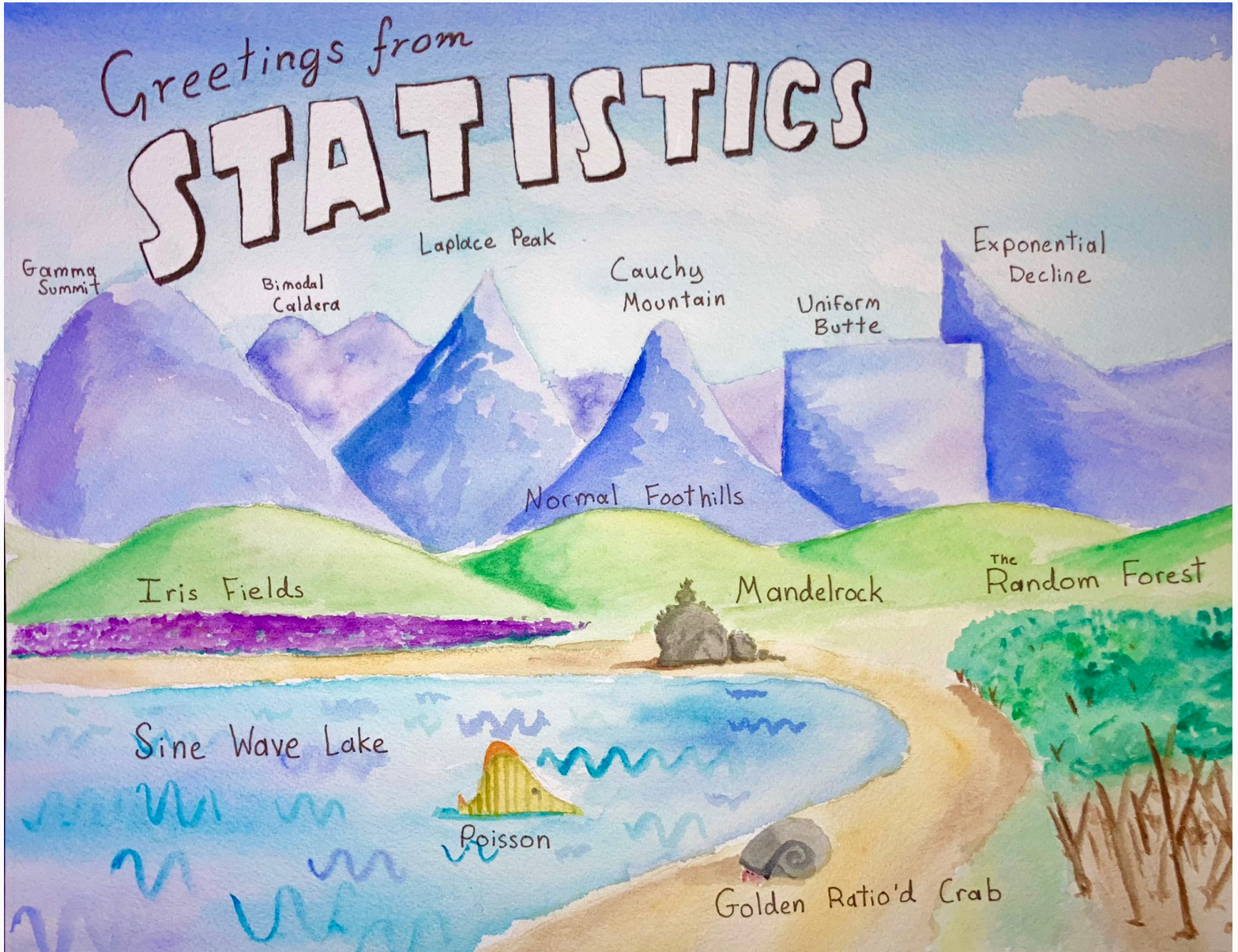
Introduction to DAV 5300

Computational Mathematics and Statistics

Jason Bryer, Ph.D.

August 27, 2024


Greetings from
STATISTICS
Gamma Summit
Bimodal Caldera
Laplace Peak
Cauchy Mountain
Uniform Butte
Exponential Decline
Normal Foothills
Iris Fields
Mandelrock
The Random Forest
Sine Wave Lake
Poisson
Golden Ratio'd Crab

@skyetetra

Agenda

- Introductions
- Syllabus
- Class meetups
- Course Schedule
- Assignments (how you will be graded)
- Software setup
- Brief introduction to R

While waiting, please complete this formative assessment:

A little about me...

- Earned my Ph.D. in Educational Pschology and Methodology from the University at Albany. Dissertation: A National Study Comparing Charter and Traditional Public Schools Using Propensity Score Analysis
- Assistant Professor at CUNY in Data Science and Information Systems
- Principal Investigator for a Department of Education Grant to develop and test the Diagnostic Assessment and Achievement of College Skills (www.DAACS.net)
- Authored over a dozen R packages including:
  - likert
  - sqlutils
  - timeline
- Specialize in propensity score methods. Three new methods/R packages developed include:
  - multilevelPSA
  - TriMatch
  - PSAboot

Also a Father...

Runner...

And photographer.

Syllabus

Syllabus and course materials are here:
https://fall2024.dav5300.net

We will use Canvas primary for submitting assignments only. Please submit PDFs.

PDFs are preferred for the homework as there is some LaTeX formatting in the R markdown files. The `tinytex` R package helps with install LaTeX, but you can also install LaTeX using MiKTeX (for Windows) and BasicTeX (for Mac).

Class Meetings

Class will meet every Tuesday.

In order to get the most out of this class attendance is required.

One Minute Papers - Complete the one minute paper after each Meetup (whether you watch live or watch the recordings). It should take approximately one to two minutes to complete.

Schedule

Start	Topic
Tuesday, August 27, 2024	Intro to the Course
Tuesday, September 03, 2024	Intro to Data
Tuesday, September 10, 2024	Summarizing Data
Tuesday, September 17, 2024	Probability
Tuesday, September 24, 2024	Distributions
Tuesday, October 01, 2024	Foundation for Inference
Tuesday, October 08, 2024	Inference for Categorical Data
Tuesday, October 15, 2024	Inference for Numerical Data
Tuesday, October 22, 2024	Linear Regression
Tuesday, October 29, 2024	Maximum Likelihood Estimation
Tuesday, November 05, 2024	Multiple Regression
Tuesday, November 12, 2024	Conferences (online)
Tuesday, November 19, 2024	Predictive Modeling
Tuesday, November 26, 2024	NO MEETUP - Thanksgiving
Tuesday, December 03, 2024	Bayesian Analysis
Tuesday, December 10, 2024	Presentations
Tuesday, December 17, 2024	Final Exam

Assignments are due on Monday before the next class.

Textbooks

OpenIntro Statistics by David Diaz, Mine Çetinkaya-Rundel, and Christopher D Barr.

Learning Statistics with R by Danielle Navaro - We will only use the Bayesian chapter from this book.

Optional

R for Data Science by Hadley Wickham and Garrett Grolemund - Recommended reference for those new to R.

Assignments

Labs (30%) - Labs are designed to provide you an opportunity to apply statistical concepts using statistical software.

Textbook questions (15%) - The assigned questions from the textbook provide an opportunity to assess conceptional understandings.

Participation (10%) - You are expected to attend every class and to complete a one minute paper at the conclusion of class.

Data Project (25%) - In a group of 2 to 3 students will present the results of analysis using a data set of your choice. More details will be provided a few weeks into the class.

Final exam (20%) - A multiple choice exam will be given on the last day of class.

All assignments are due on Monday. Assignments submitted late will be penalized. Assignments will not be accepted more than one week after their due date.

Academic Integrity

With the exception of the data project, I expect you to complete all assignments (e.g. homework, labs) on your own. It is fine to ask questions of your peers and professor, but working together and/or sharing answers is not allowed.

Yeshiva's Policy

The submission by a student of any examination, course assignment, or degree requirement is assumed to guarantee that the thoughts and expressions therein not expressly credited to another are literally the student's own. Evidence to the contrary will result in appropriate penalties. For more information, visit https://www.yu.edu/academic-integrity.

Communication

- Email: jason.bryer@yu.edu.
- Canvas
- Office hours before and after class and by appointment.

Software Setup

Why R?

There are many languages data scientists use. R is specifically designed for statistics. We will leverage many R packages that are specifically designed to conduct, teach, and communicate statistical analysis.

To be a well rounded data scientists, I believe you need to have experience in both R and Python. For this course:

- Use R for the labs (they are designed to help you learn the core commands).
- You may use Python or R for the homework and data project.

Software

This is an applied statistics course so we will make extensive use of the R statistical programming language.

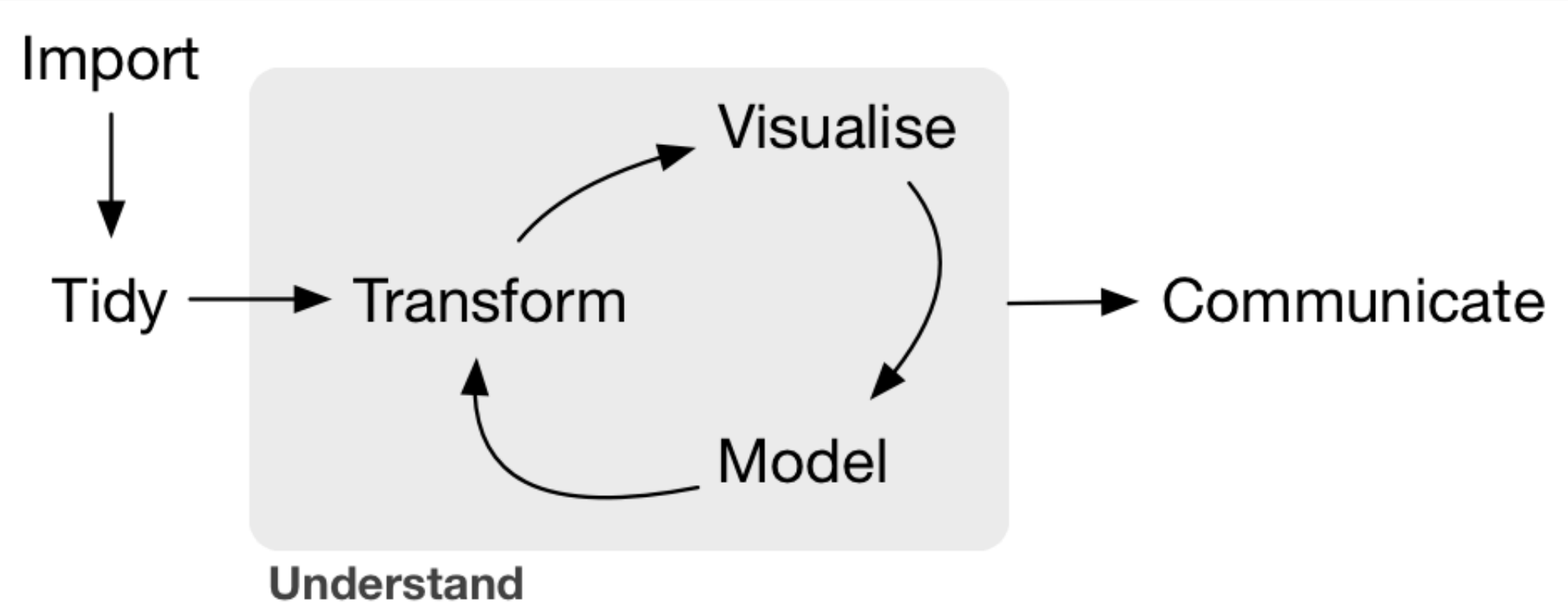
- Install R and RStudio on your own computer. I encourage everyone to do this at some point by the end of the semester.

You will also need to have LaTeX installed as well in order to create PDFs. The `tinytex` R package helps with this process:

```
install.packages('tinytex')
tinytex::install_tinytex()
```

Introduction to R

Workflow



Source: Wickham & Grolemund, 2017

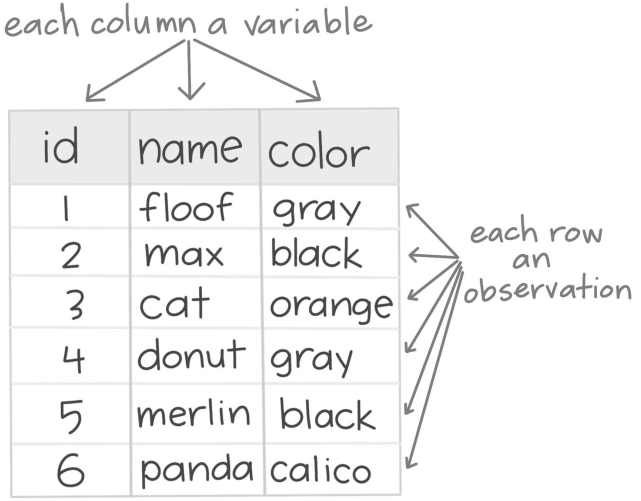
Tidy Data

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



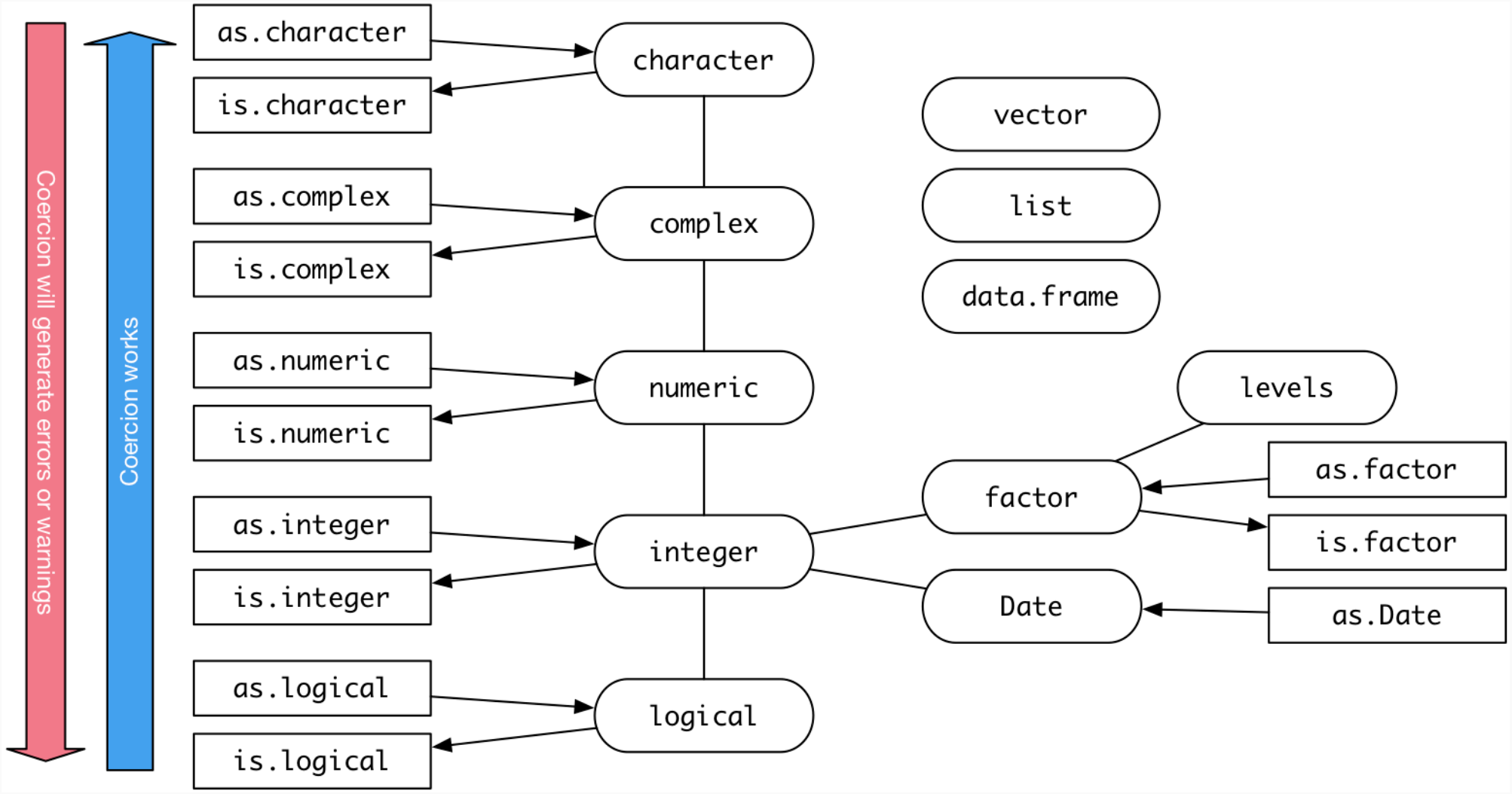
Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

See Wickham (2014) Tidy data.

Types of Data

- Numerical (quantitative)
  - Continuous
  - Discrete
- Categorical (qualitative)
  - Regular categorical
  - Ordinal

Data Types in R



Data Types / Descriptives / Visualizations

Data Type	Descriptive Stats	Visualization
Continuous	mean, median, mode, standard deviation, IQR	histogram, density, box plot
Discrete	contingency table, proportional table, median	bar plot
Categorical	contingency table, proportional table	bar plot
Ordinal	contingency table, proportional table, median	bar plot
Two quantitative	correlation	scatter plot
Two qualitative	contingency table, chi-squared	mosaic plot, bar plot
Quantitative & Qualitative	grouped summaries, ANOVA, t-test	box plot

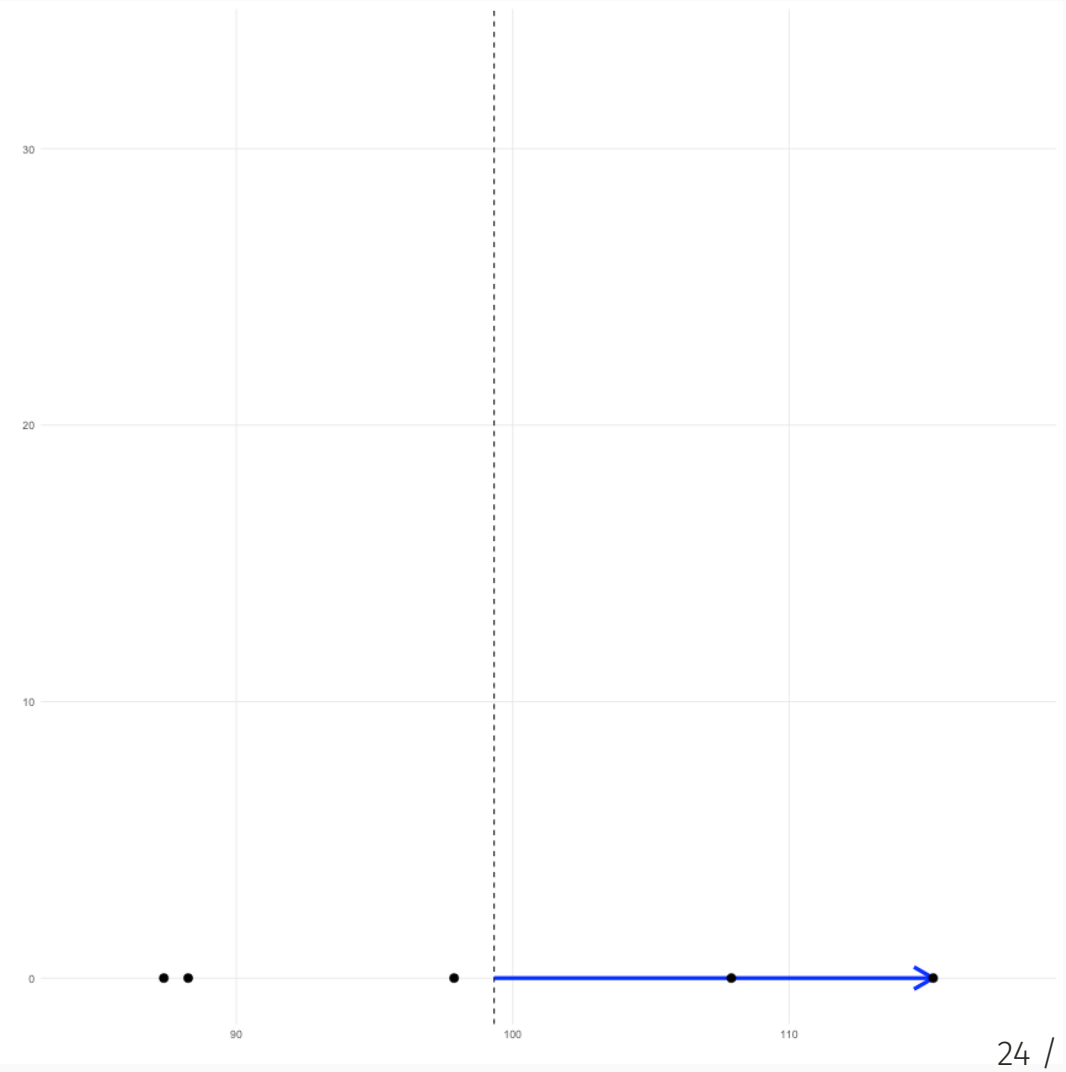
Variance

Population Variance:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

Consider a dataset with five values (black points in the figure). For the largest value, the deviance is represented by the blue line ($x_i - \bar{x}$).

See also:
https://shiny.rit.albany.edu/stat/visualizess/
https://github.com/jbryer/VisualStats/

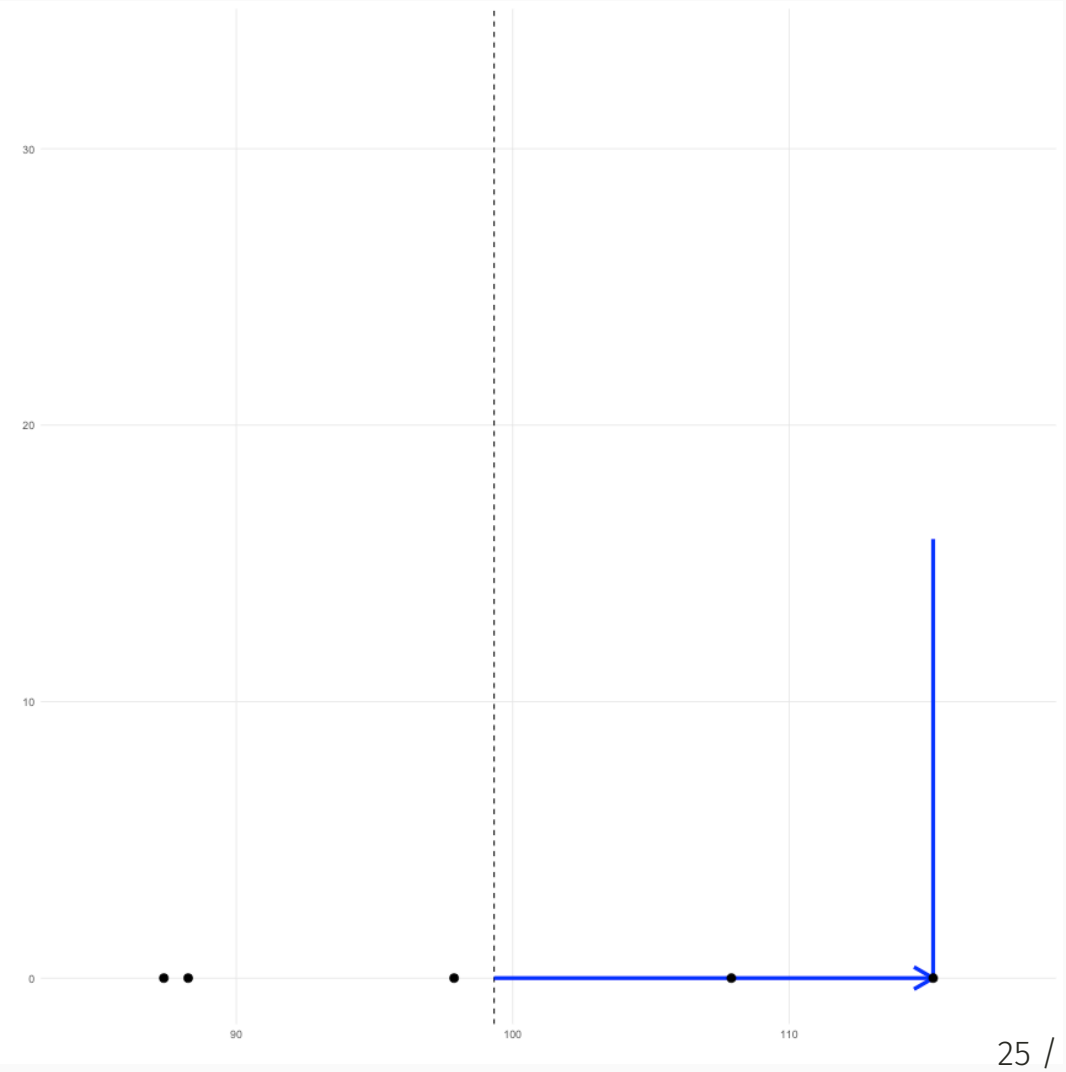


Variance (cont.)

Population Variance:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

In the numerator, we square each of these deviances. We can conceptualize this as a square. Here, we add the deviance in the *y* direction.

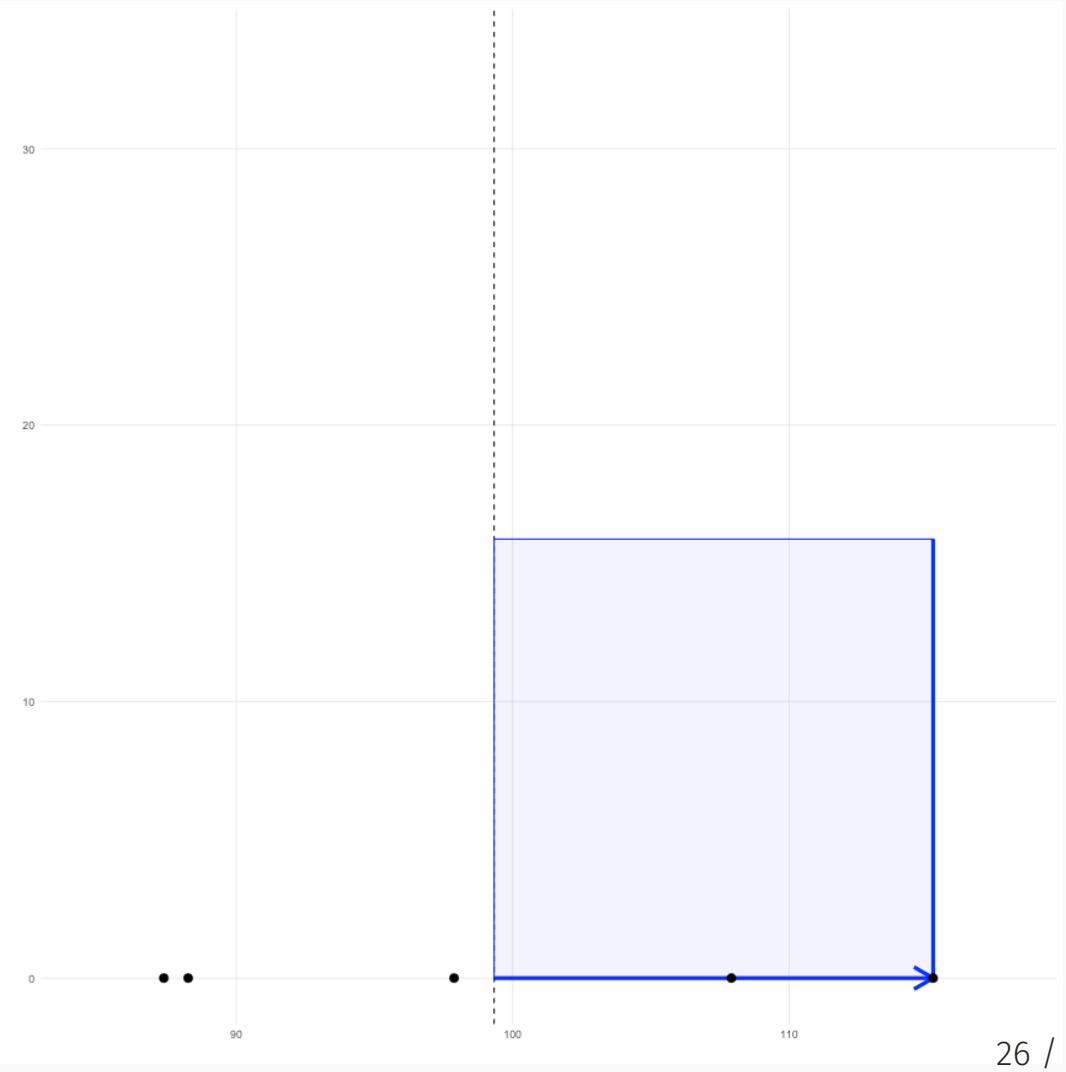


Variance (cont.)

Population Variance:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

We end up with a square.

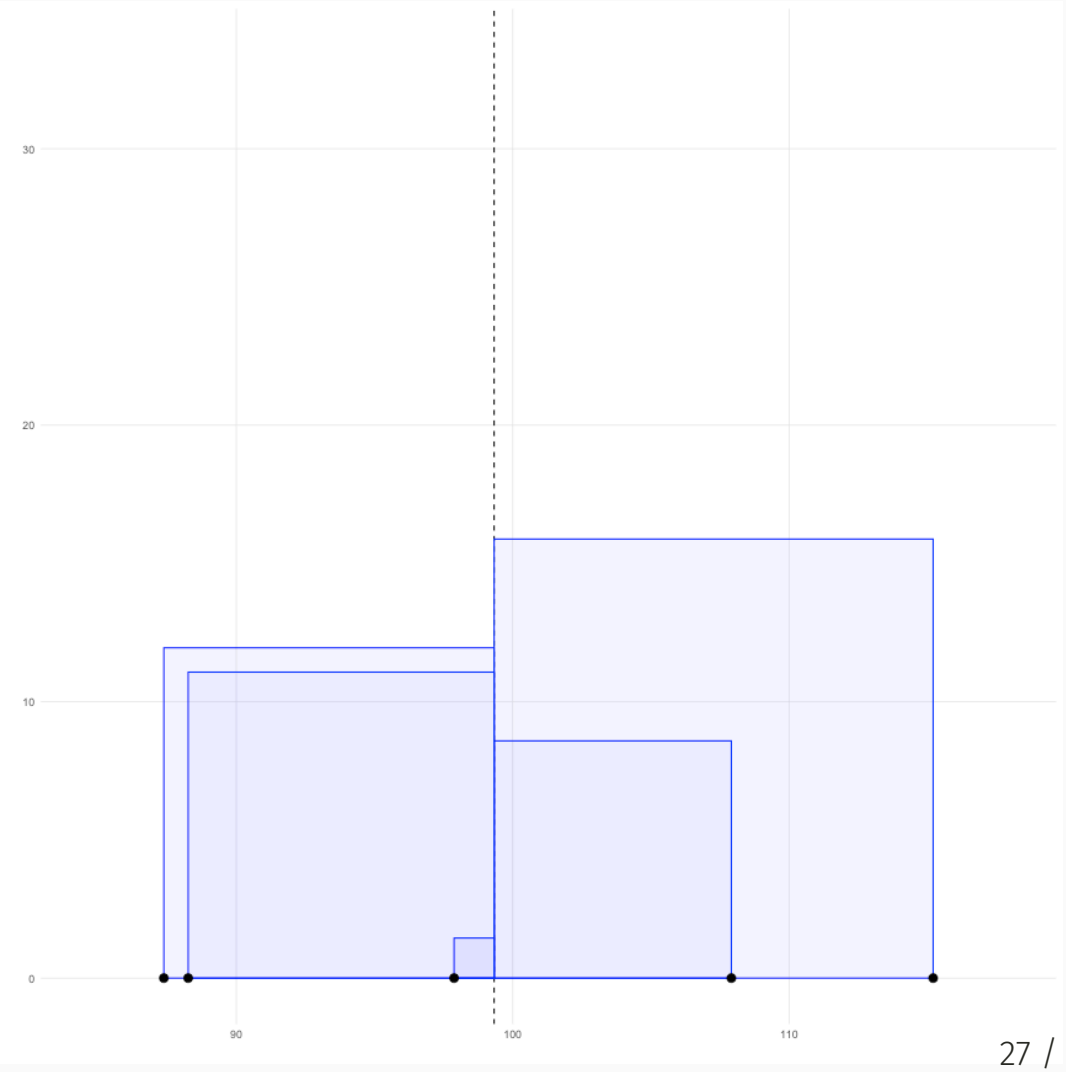


Variance (cont.)

Population Variance:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

We can plot the squared deviance for all the data points. That is, each component in the numerator is the area of each of these squares.

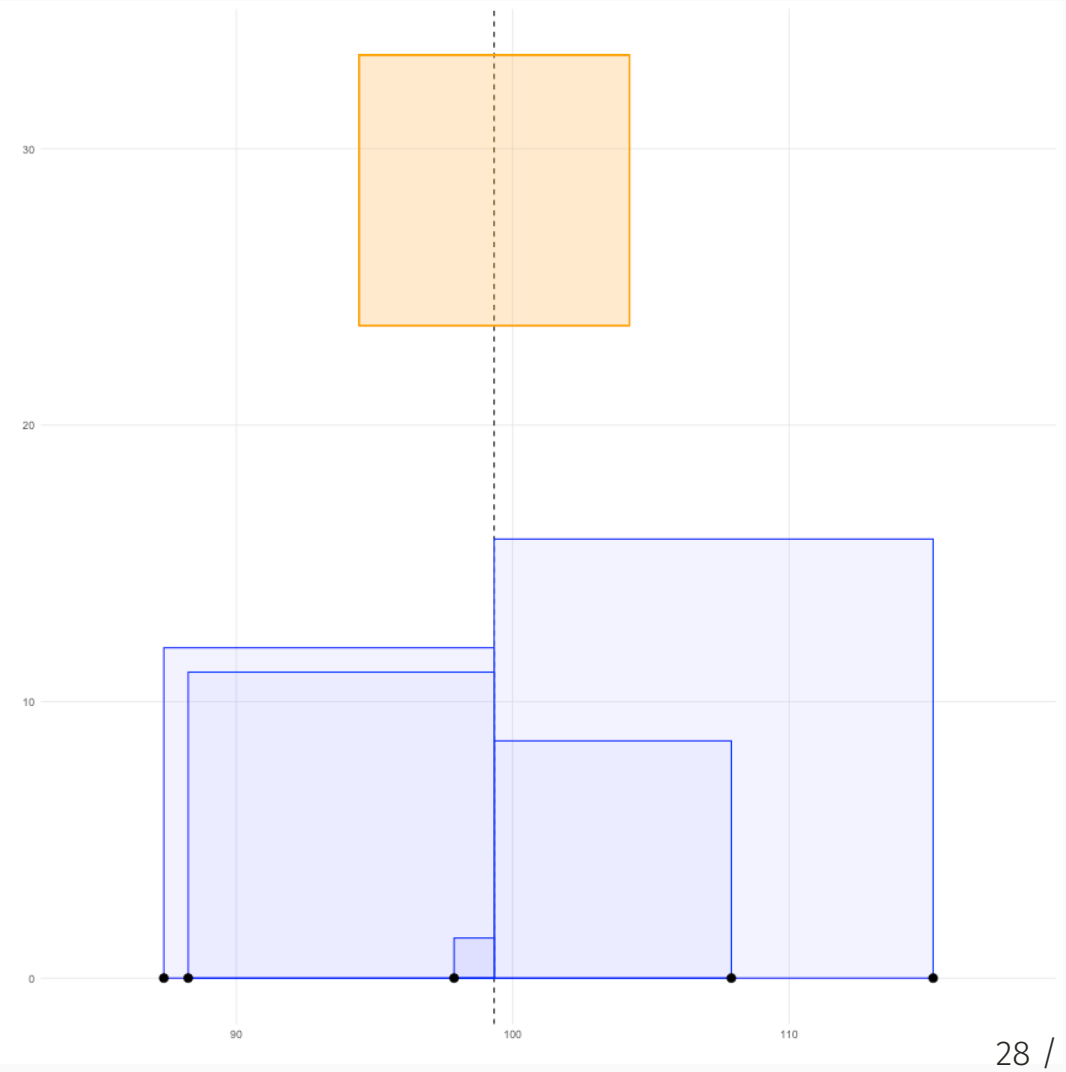


Variance (cont.)

Population Variance:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

The variance is therefore the average of the area of all these squares, here represented by the orange square.



Population versus Sample Variance

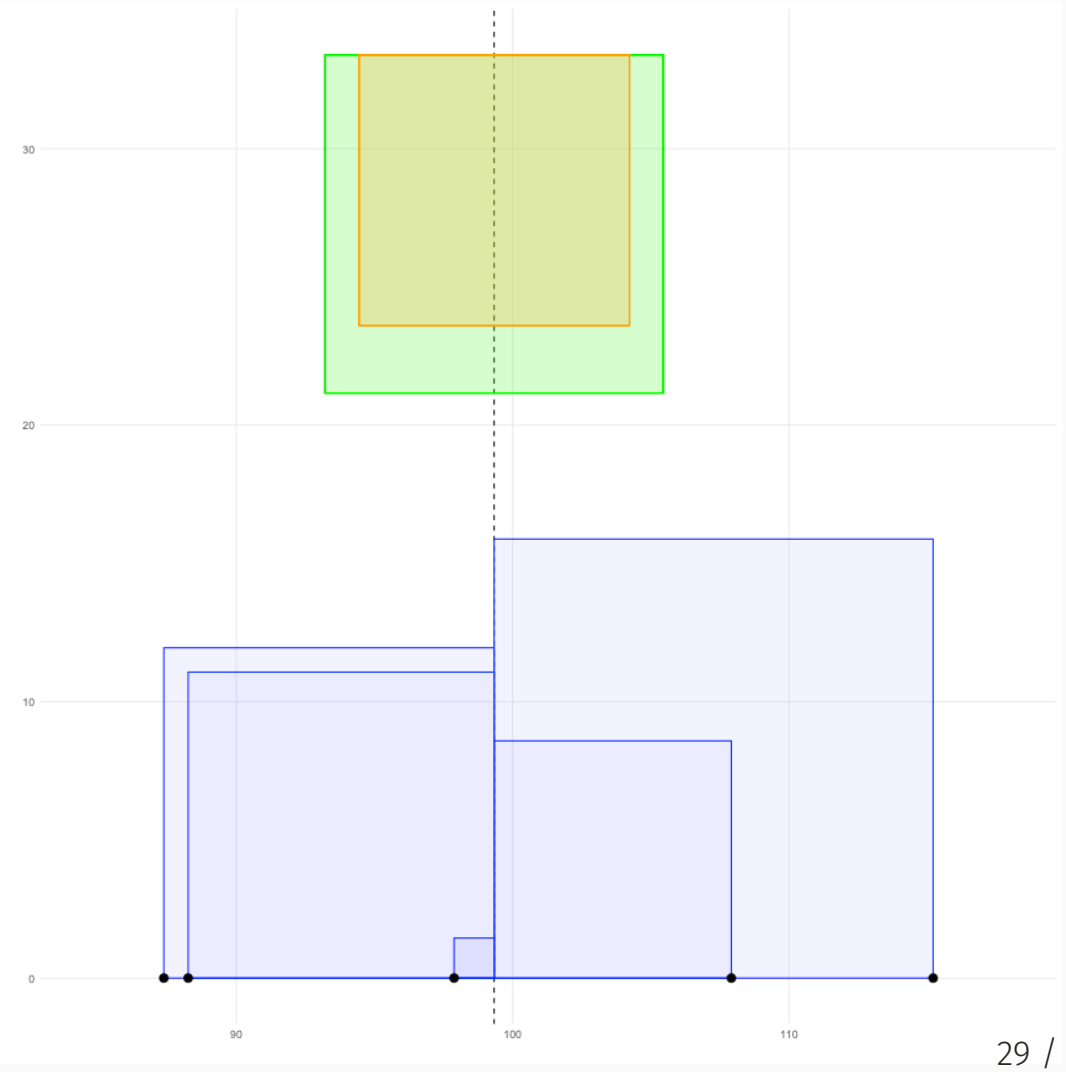
Typically we want the sample variance. The difference is we divide by $n-1$ to calculate the sample variance. This results in a slightly larger area (variance) then if we divide by n.

Population Variance (yellow):

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

Sample Variance (green):

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$



Robust Statistics

Consider the following data randomly selected from the normal distribution:

```
set.seed(41)
x <- rnorm(30, mean = 100, sd = 15)
mean(x); sd(x)
```

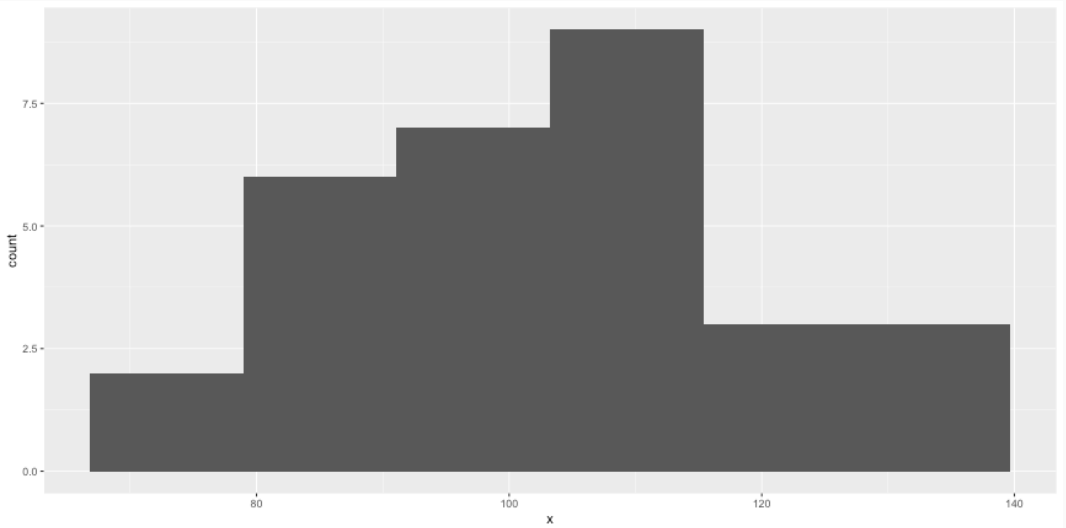
```
## [1] 103.1934
```

```
## [1] 16.8945
```

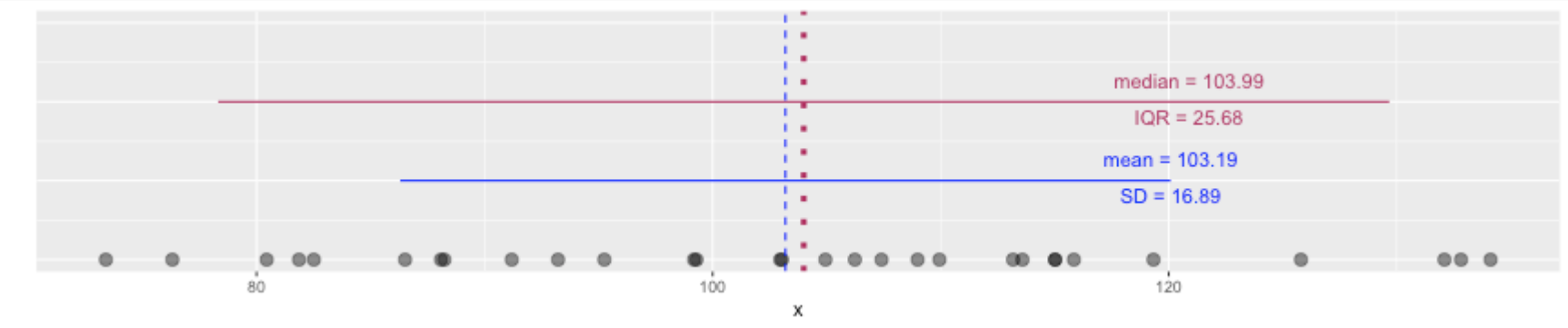
```
median(x); IQR(x)
```

```
## [1] 103.9947
```

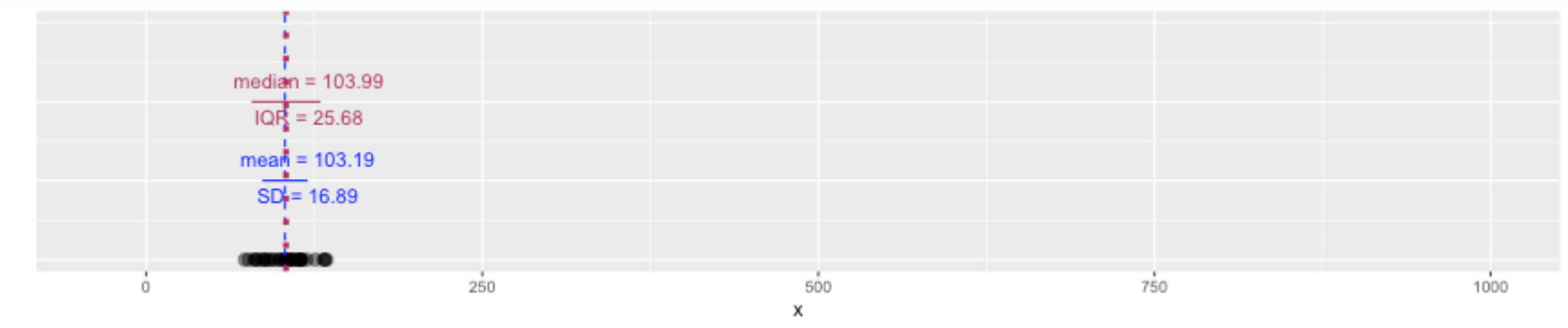
```
## [1] 25.68004
```



Robust Statistics

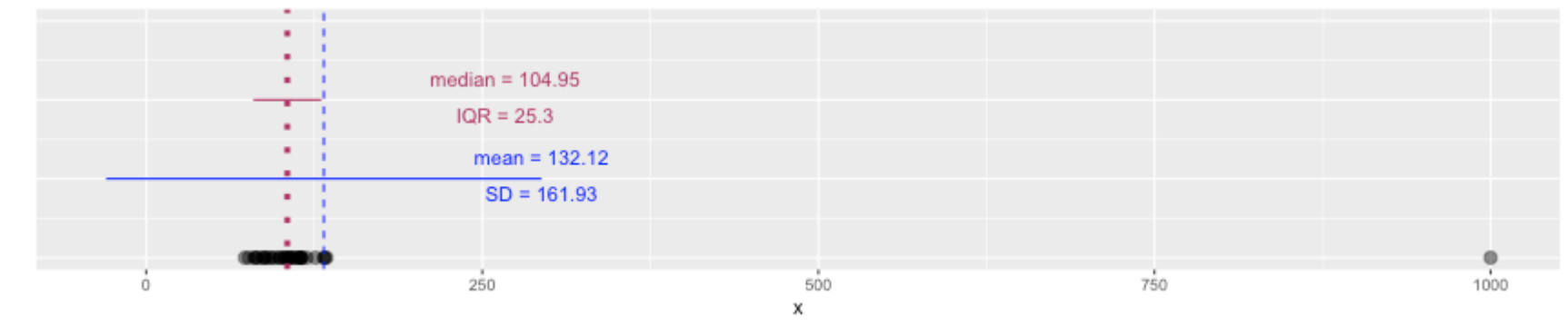


Robust Statistics



Let's add an extreme value:

```
x <- c(x, 1000)
```



Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Good luck with the semester!

jason.bryer@yu.edu
@jbryer
@jbryer@vis.social
github.com/jbryer/DAV5300-2024-Spring