MATH 205: Statistical methods

Vu Dinh

Departments of Mathematical Sciences
University of Delaware

August 31st, 2022

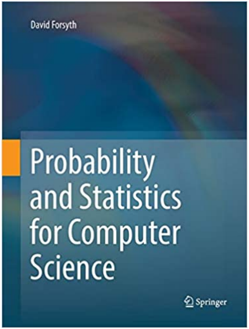
General information

- Lectures:

  MWF 3:35pm-4:25pm, Gore Hall Room 304

- Labs:
  - Section 050L: M 2:30pm - 3:20pm, Gore Hall Room 222
  - Section 051L: W 2:30pm - 3:20pm, Gore Hall Room 222
- Office hours
  - Tuesday 3:00pm - 4:30pm, Ewing Hall Room 312
  - Friday 1:30pm - 3:00pm, Ewing Hall Room 312
  - or by appointments

Textbooks



Lectures:
Probability and Statistics for Computer Science.
David Forsyth (2018)

Labs:
simpleR – Using R for Introductory Statistics.
John Verzani (2002)

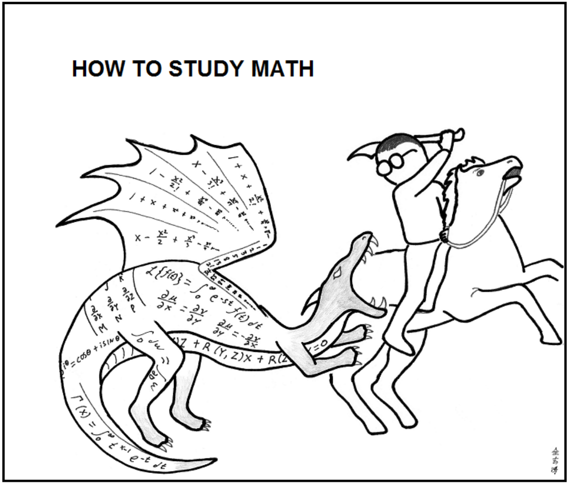
Other classroom settings

- The lectures will be recorded by UD Capture, accessible through Canvas.
  Note that there will be no camera in class, so work on the board wouldn't be seen in the records.

Evaluation

- Overall scores will be computed as follows:
  25% homework, 15% quizzes, 25% midterm, 35% final
- No letter grades will be given for homework, midterm, or final. Your letter grade for the course will be based on your overall score.
- The lowest homework scores and the lowest quiz score will be dropped.
- Letter grades you can achieve according to your overall score.
  - $\geq$ 90%: At least A
  - $\geq$ 75%: At least B
  - $\geq$ 60%: At least C
  - $\geq$ 50%: At least D

Homework



Don't just read it; fight it!

--- Paul R. Halmos

Homework

- There are 5 homework assignments throughout the semester
- Assignments will be posted on Monday (starting from the third week) and will be due on Friday of *the following week*, *at the beginning of* lecture.
- No late homework will be accepted.
- Your lowest homework scores will be dropped in the calculation of your overall homework grade.

Quizzes

- At the end of some chapter, there will be a short quiz during class.
- The quiz dates will be announced at least one class in advance.
- The lowest quiz score will be dropped.

Exams

- There will be an in-class midterm exam during the week of October 24-28. The exam consists of two parts: a written exam during the Oct 28 lecture, and the computational exam during the lab sessions of that week.
- Final exam (written) during the final week.

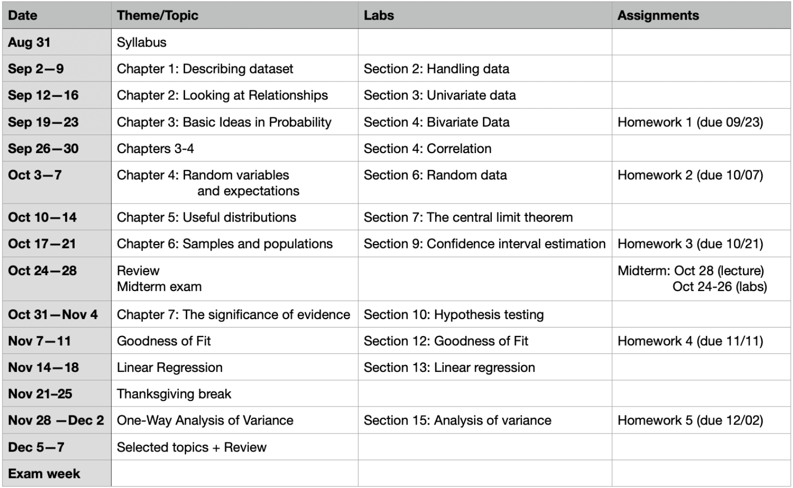
Data analysis

- Open-source statistical system R

  `http://cran.r-project.org/`

- You need to install R and RStudio on your personal laptops

Tentative schedule



Date	Theme/Topic	Labs	Assignments
Aug 31	Syllabus		
Sep 2—9	Chapter 1: Describing dataset	Section 2: Handling data	
Sep 12—16	Chapter 2: Looking at Relationships	Section 3: Univariate data	
Sep 19—23	Chapter 3: Basic Ideas in Probability	Section 4: Bivariate Data	Homework 1 (due 09/23)
Sep 26—30	Chapters 3-4	Section 4: Correlation	
Oct 3—7	Chapter 4: Random variables and expectations	Section 6: Random data	Homework 2 (due 10/07)
Oct 10—14	Chapter 5: Useful distributions	Section 7: The central limit theorem	
Oct 17—21	Chapter 6: Samples and populations	Section 9: Confidence interval estimation	Homework 3 (due 10/21)
Oct 24—28	Review Midterm exam		Midterm: Oct 28 (lecture) Oct 24-26 (labs)
Oct 31—Nov 4	Chapter 7: The significance of evidence	Section 10: Hypothesis testing	
Nov 7—11	Goodness of Fit	Section 12: Goodness of Fit	Homework 4 (due 11/11)
Nov 14—18	Linear Regression	Section 13: Linear regression	
Nov 21–25	Thanksgiving break		
Nov 28 —Dec 2	One-Way Analysis of Variance	Section 15: Analysis of variance	Homework 5 (due 12/02)
Dec 5—7	Selected topics + Review		
Exam week			

Chapter 1: Describing dataset

Categorical and continuous data

Statistics deal with the collection, organization, analysis, interpretation and presentation of data:

- Categorical:
  - data that records categories
  - each data item can take a (typically small) set of prescribed values
  - example: students' majors or programs
- Continuous:
  - can receive any value in a particular range
  - example: height or weight or body temperature

Dataset as d-tuples

- A d-tuple is an ordered list of d elements
- We think of a dataset as a collection of d-tuples
- Example:
  A dataset has entries for ID, Email, Name, Audit, Units, Program and Plan, Level, Grade, Weight for 55 students
  $\rightarrow d = 9$, $N = 55$.

Modern statistics is about high-dimensional data...

...but this class is about $d = 1$ and $d = 2$

- Chapter 1: Looking at 1D data
- Chapter 2: Looking at 2D data
- Confidence interval, hypothesis testing, goodness of fit: analyzing 1D data
- Linear regression: analyzing 2D data

Chapter 1: Describing univariate data

Summarizing univariate data:

- Mean
- Median
- Standard deviation
- Interquartile Range

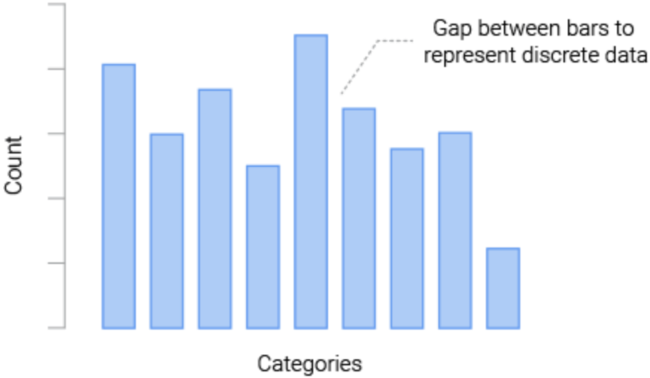
Visualizing univariate data:

- Bar chart
- Pie chart
- Histogram
- Box plot

Categorical data: bar charts

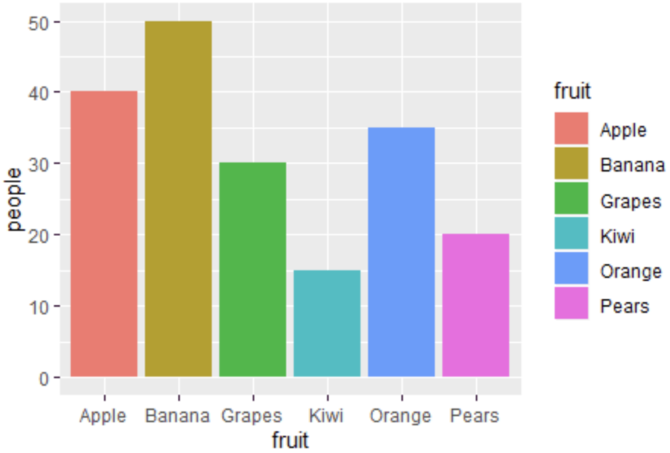
- A bar chart is a set of bars, one per category
- the height of each bar is proportional to the number of items in that category
- the height could be given by the frequency, or the proportion

Bar charts



Gap between bars to represent discrete data

Example: People's favorite fruit in a survey

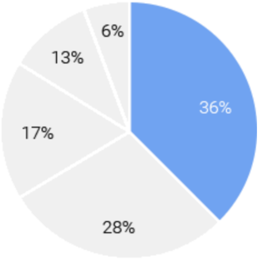


Categorical data: pie charts

- each slice of the pie corresponds to one category
- the area of the slice is proportional to the number of items in that category

Bar charts

A Pie Chart is a special chart that shows relative sizes of data using **pie slices**.

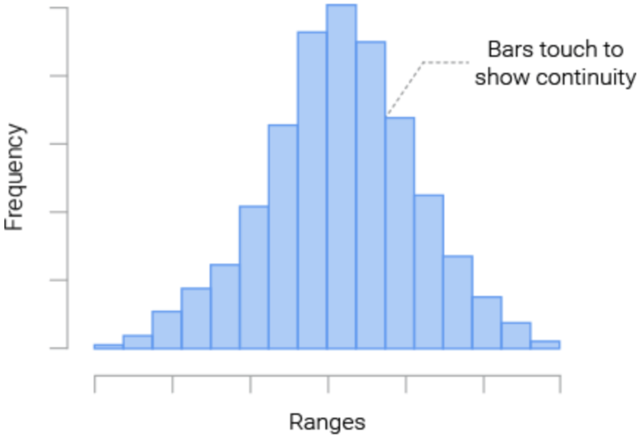


They are good if you are trying to compare parts of a single data series to the whole.

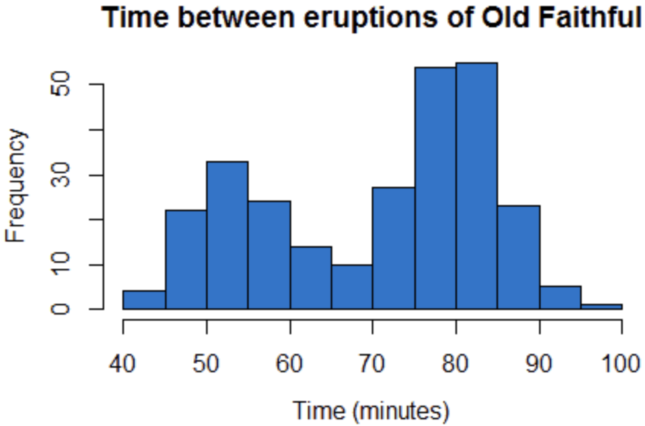
Continuous data: histograms

- a simple generalization of a bar chart
- We divide the range of the data into intervals, which do not need to be equal in length
- We then build a set of boxes, one per interval. Each box sits on its interval on the horizontal axis.
- The area of the box is proportional to the number of elements in the box.

Histograms



Histograms


Time between eruptions of Old Faithful
Frequency
0
10
30
50
40
50
60
70
80
90
100
Time (minutes)