# MATH 205: Statistical methods

Vu Dinh

Departments of Mathematical Sciences
University of Delaware

September 1st, 2021

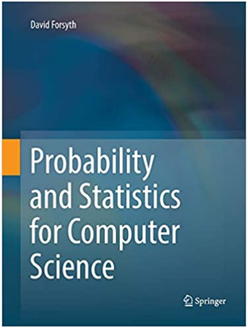
# General information

- Lectures:

  MW 3:35pm-4:50pm, Kirkbride 205

- Labs:
  - Section 050L: M 2:30pm - 3:20pm, Ewing 101
  - Section 051L: W 2:30pm - 3:20pm, Ewing 101
- Office hours
  - TTh 2:00pm - 3:30pm, Ewing 312
  - or by appointments

# Textbooks




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

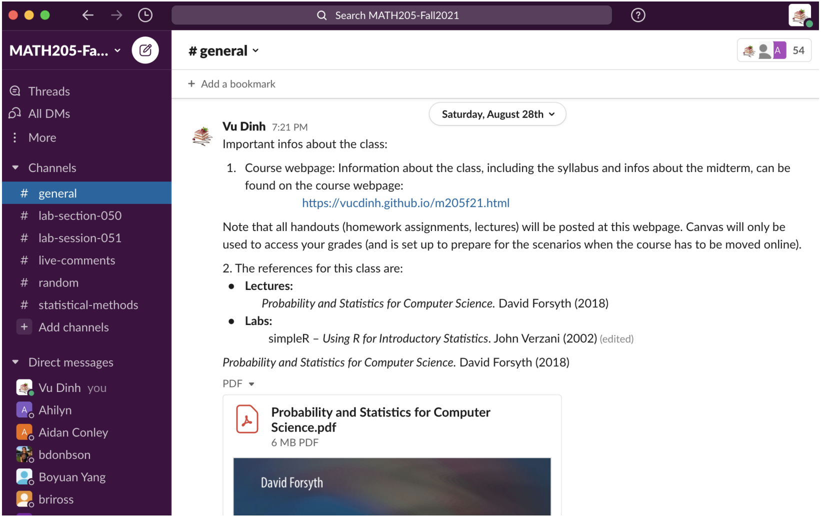
**Labs:**
*simpleR* – Using R for Introductory Statistics.
John Verzani (2002)

# The safety of our learning environment

We will adhere to the practice of wearing face masks and cleaning your seat and desk area at the beginning of class:

- Must wear a cloth mask that covers your nose and mouth
- Must not eat or drink in class
- Upon entering the classroom, wipe down your seat and desk area

# Communications



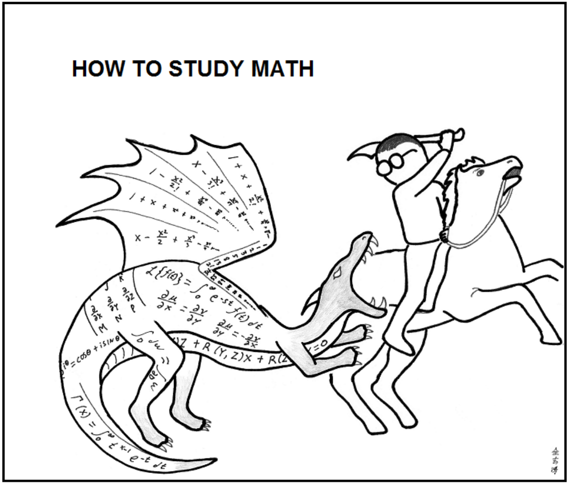
# Other classroom settings

- 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.
- The lab doesn't have UD Capture. I will provide a Zoom session for each lab.

# Evaluation

- Overall scores will be computed as follows:
  25% homework, 10% quizzes, 25% midterm, 40% 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 Wednesday 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 25-27. The exam consists of two parts: a written exam during the Oct 27 lecture, and the computational exam during the lab sessions of that week.
- Final exam (written) during the final week.

Open source statistical system R

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

# Tentative schedule

| Date | Theme/Topic | Labs | Assignments |
|---|---|---|---|
| **Sep 1** | Syllabus | | |
| **Sep 8** | Chapter 1: Describing dataset | Section 2: Handling data | |
| **Sep 13 - 15** | Chapter 2: Looking at Relationships | Section 3: Univariate data | |
| **Sep 20-22** | Chapter 3: Basic Ideas in Probability | Section 4: Bivariate Data | Homework 1 (due 09/22) |
| **Sep 27-29** | Chapters 3-4 | Section 4: Correlation | |
| **Oct 4-6** | Chapter 4: Random variables and expectations | Section 6: Random data | Homework 2 (due 10/06) |
| **Oct 11-13** | Chapter 5: Useful distributions | Section 7: The central limit theorem | |
| **Oct 18-20** | Chapter 6: Samples and populations | Section 9: Confidence interval estimation | Homework 3 (due 10/20) |
| **Oct 25-27** | Review and midterm exam | | Midterm: Oct 27 (lecture), Oct 25-27 (labs) |
| **Nov 1-3** | Chapter 7: The significance of evidence | Section 10: Hypothesis testing | |
| **Nov 8-10** | Goodness of Fit | Section 12: Goodness of Fit | Homework 4 (due 11/10) |
| **Nov 15-17** | Linear Regression | Section 13: Linear regression | |
| **Nov 22-24** | Thanksgiving break | | |
| **Nov 29 - Dec 1** | One-Way Analysis of Variance | Section 15: Analysis of variance | Homework 5 (due 12/01) |
| **Dec 6-8** | 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

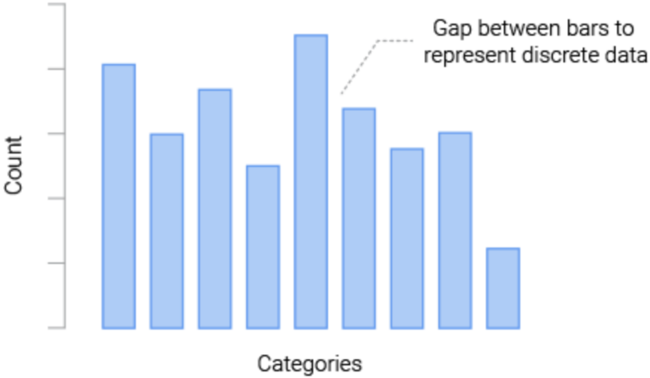
Visualizing univariate data:

- Bar chart
- Pie chart
- Histogram
- Box plot

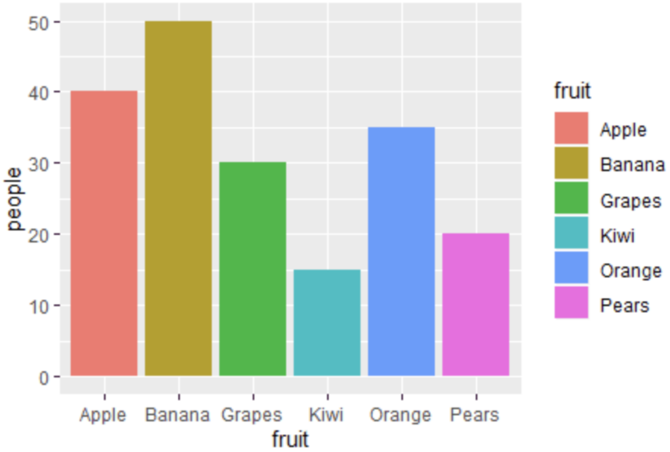
# Categorical data: bar charts

- 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
Count
Categories

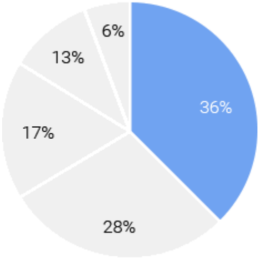
# Example: People's favorite fruit in a survey



# 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

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

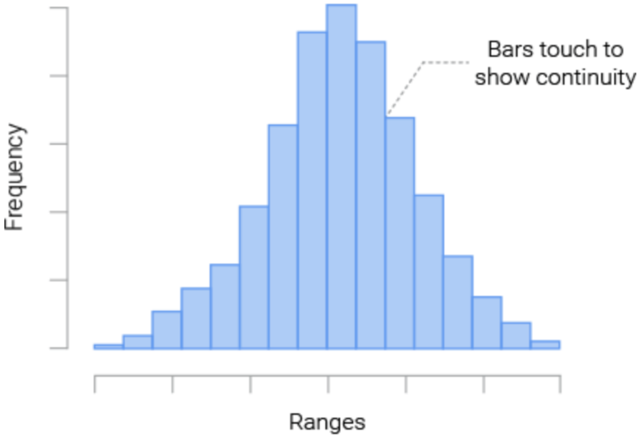



They are good if you are trying to compare parts of a single data series to the whole.

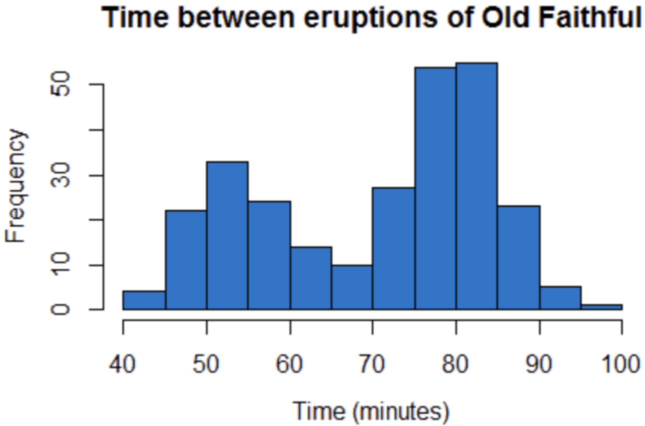
# Continuous data: histograms

- 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)

# Summarizing univariate data

- Mean
- Median
- Standard deviation
- Variance
- Interquartile Range

# Mean

**Definition 1.1 (Mean)** Assume we have a dataset $\{x\}$ of $N$ data items, $x_1, \ldots, x_N$. Their mean is

$$\text{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

# Properties of the Mean

**Useful Facts 1.1 (Properties of the Mean)**

- Scaling data scales the mean: or

$$\text{mean}(\{kx_i\}) = k\text{mean}(\{x_i\}).$$

- Translating data translates the mean: or

$$\text{mean}(\{x_i + c\}) = \text{mean}(\{x_i\}) + c.$$

- The sum of signed differences from the mean is zero: or,

$$\sum_{i=1}^{N}(x_i - \text{mean}(\{x_i\})) = 0.$$

# Median

**Definition 1.4 (Median)** The median of a set of data points is obtained by sorting the data points, and finding the point halfway along the list. If the list is of even length, it's usual to average the two numbers on either side of the middle. We write

$$\text{median}(\{x\})$$

for the operator that returns the median.

Median is not affected by outliers

The risk of developing iron deficiency is especially high during pregnancy. The problem with detecting such deficiency is that some methods for determining iron status can be affected by the state of pregnancy itself. Consider the following data on transferrin receptor concentration for a sample of women with laboratory evidence of overt iron-deficiency anemia ("Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy," *Amer. J. Clin. Nutrit.*, 1991: 1077–1081):

$$x_1 = 15.2 \quad x_2 = 9.3 \quad x_3 = 7.6 \quad x_4 = 11.9 \quad x_5 = 10.4 \quad x_6 = 9.7$$
$$x_7 = 20.4 \quad x_8 = 9.4 \quad x_9 = 11.5 \quad x_{10} = 16.2 \quad x_{11} = 9.4 \quad x_{12} = 8.3$$

The list of ordered values is

$$7.6 \quad 8.3 \quad 9.3 \quad 9.4 \quad 9.4 \quad 9.7 \quad 10.4 \quad 11.5 \quad 11.9 \quad 15.2 \quad 16.2 \quad 20.4$$

Since $n = 12$ is even, we average the $n/2 =$ sixth- and seventh-ordered values:

$$\text{sample median} = \frac{9.7 + 10.4}{2} = 10.05$$

# Measures of variability: deviation from the mean

**Definition 1.2 (Standard Deviation)** Assume we have a dataset $\{x\}$ of $N$ data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is:

$$\begin{aligned}\text{std}(\{x_i\}) &= \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \text{mean}(\{x\}))^2} \\ &= \sqrt{\text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})}.\end{aligned}$$

# Properties of the standard deviation

**Useful Facts 1.2 (Properties of Standard Deviation)**

- Translating data does not change the standard deviation, i.e. $\mathsf{std}(\{x_i + c\}) = \mathsf{std}(\{x_i\})$.
- Scaling data scales the standard deviation, i.e. $\mathsf{std}(\{kx_i\}) = k\mathsf{std}(\{x_i\})$.
- For any dataset, there can be only a few items that are many standard deviations away from the mean. For $N$ data items, $x_i$, whose standard deviation is $\sigma$, there are at most $\frac{1}{k^2}$ data points lying $k$ or more standard deviations away from the mean.
- For any dataset, there must be at least one data item that is at least one standard deviation away from the mean, that is, $(\mathsf{std}(\{x\}))^2 \leq \max_i (x_i - \mathsf{mean}(\{x\}))^2$.

The standard deviation is often referred to as a scale parameter; it tells you how broadly the data spreads about the mean.

**Definition 1.3 (Variance)** Assume we have a dataset $\{x\}$ of $N$ data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\mathsf{var}(\{x\}) = \frac{1}{N}\left(\sum_{i=1}^{i=N}(x_i - \mathsf{mean}(\{x\}))^2\right)$$

$$= \mathsf{mean}\left(\{(x_i - \mathsf{mean}(\{x\}))^2\}\right).$$