Incorporating R into STA107

An Introduction to Probability and Modelling Course

Asal Aslemand

Department of Mathematical and Computational Sciences
University of Toronto Mississauga

2025-02-19

About Me

Asal Aslemand
Statistics Enthusiast
UTM Alumnus

Artwork by Anna Ly

Motivation

- STA107 (UTM - MCS Department)
  - Course Description in Academic Calendar
    - Emphasizes the construction of discrete probability models for applications.
    - Expects the understanding of the concept of randomness and aspects of its mathematical representation.
  - Previous Course Offerings (Winter 2018 - Winter 2020) Included:
    - Introduction to Data
    - Normal Distribution
    - Foundation of Inference
- STA130 (Faculty of A & S - DoSS)
  - Course Description in Academic Calendar
    - Discusses the crucial role played by statistical reasoning in solving challenging problems.
    - Uses a combination of logical thinking, mathematics, computer simulation, and oral and written discussion and analysis.
  - Offered since Winter 2018

STA Curriculum Mapping

- Program Learning Outcomes
  - Foundation of Probability Theory
  - Statistical Inferential Reasoning (Theory & Implementation)
  - Multiple Perspectives
  - Data Collection
  - Data Visualization
  - Data Wrangling
  - Programming Skills
  - Statistical and Probability Modelling
  - Simulation
  - Data Analysis
  - Statistical Analysis
  - Applying Statistical and Probability Knowledge
  - Communication Skills (with Non-statisticians)
  - Communication Skills (with Statisticians)
  - Collaboration Skills
  - Critical Thinking
  - Ethical Practice
- Re-think STA107, our introductory course design/offering

Statistics and Data Science Education

- Foster active learning
  - Hands-on small-group activities (Kalain & Kasim, 2014)
  - Collaborative practice of statistics (Parke, 2008)
- Use technology to explore concepts
  - Computer simulations aid student understanding of concepts (Jamie, 2002)
  - Develop inferential reasoning by using visuals to show distribution of observed statistics under the null hypothesis (Chance et al., 2024)

Open Source Resources

- Speegle, D., & Clair, B. (2024). Probability, Statistics, and Data: A Fresh Approach Using R.
- Wagaman, A. S., & Dobrow, R. P. (2021). Probability with Applications and R. 2nd ed.
- Horgan, J. M. (2020). Probability with R.
- Irizarry, R. A. (2019). Introduction to Data Science. Data Wrangling and Visualization with R.
- Irizarry, R. A. (2019). Introduction to Data Science. Statistics and Prediction Algorithms Through Case Studies.
- Çetinkaya-Rundel, M., & Hardin. J. (2024). Introduction to Modern Statistics (2e).
- Timbers, T., Campbell, T., & Lee, M. (2024). Data Science. A First Introduction.

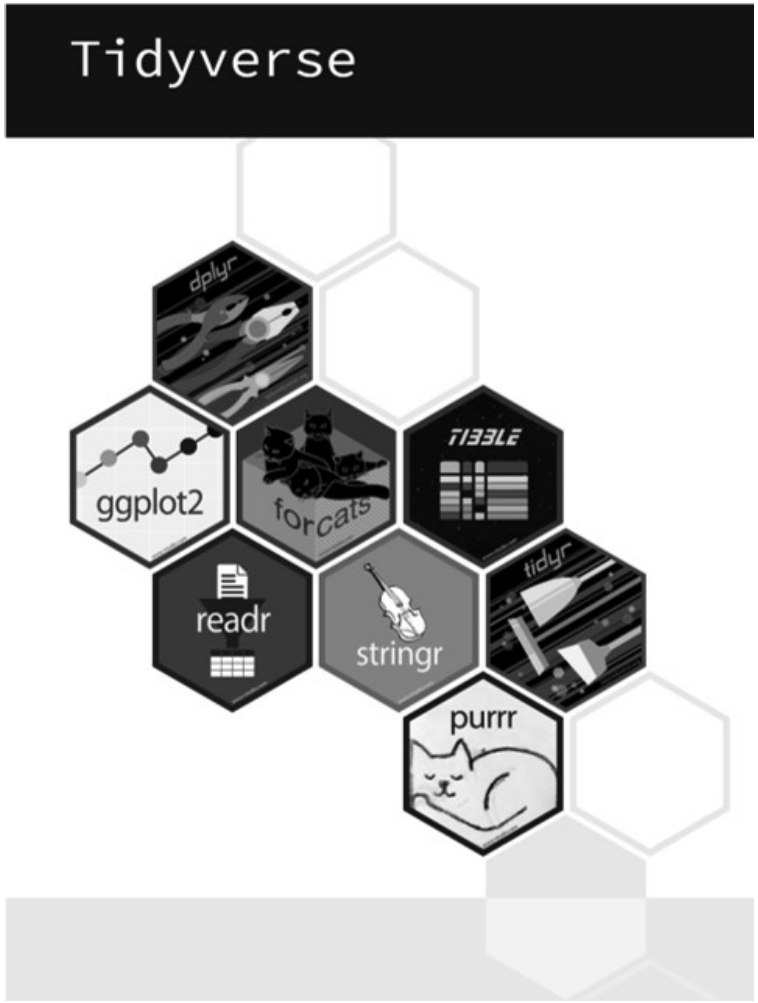
Winter 2025 STA107 Course

- Emphasizes a practical introduction to probability concepts and modelling.
- Uses statistical software R to illustrate simulating empirical probabilities for understanding randomness and variability.
- Emphasizes the role simulation plays in learning probability concepts and in practicing statistics.
- Emphasizes random variables and their probability distributions, beginning with the focus on discrete distributions, their relationships, and their connection to Normal distribution.
- Includes the concept of sampling distributions and how sample-based statistics like means, and proportions behave as random variables, and that with larger sample sizes, these distributions take on a shape resembling the Normal distribution.
- Covers practical tools for making statistical inference.

The statistics student tossed a coin throughout the Boston Marathon to study the behavior of a chance process in the long run.

R Statistical Computing

- Interact with R through RStudio via U of T JupyterHub
- Explore a variety of R packages, including the Tidyverse package.
- The tidyverse is a collection of R packages designed for data science.
- All packages share an underlying design philosophy, grammar, and data structures.



Course Assessments

Type	Description	Due Date	Weight
Homework Assignments	Probability Simulations using R	On-going	10%
Tutorial Participations	Tutorial Activities	On-going	10%
Midterm Test	Course Modules 1 to 6	March 1, 2025	25%
Infographic Activity	Infographic Quiz	March 23, 2025	8%
Infographic Design	Communicating Data	April 2, 2025	5%
Post-course Survey	Course Reflection	April 4, 2025	2%
Final Exam	Cumulative Final Exam	April Exam Period	40%
		Total	100%

Tutorial Structure

R Activity → Session Participation → Reflection

Preparing for tutorials by completing an R activity assignment.

Aligning with course content for relevance.

Collaborating with peers in small groups discussions.

Doing hands-on activities and completing worksheets.

Reflecting on understanding weekly course contents.

Completing a short anonymous graded survey in Quercus.

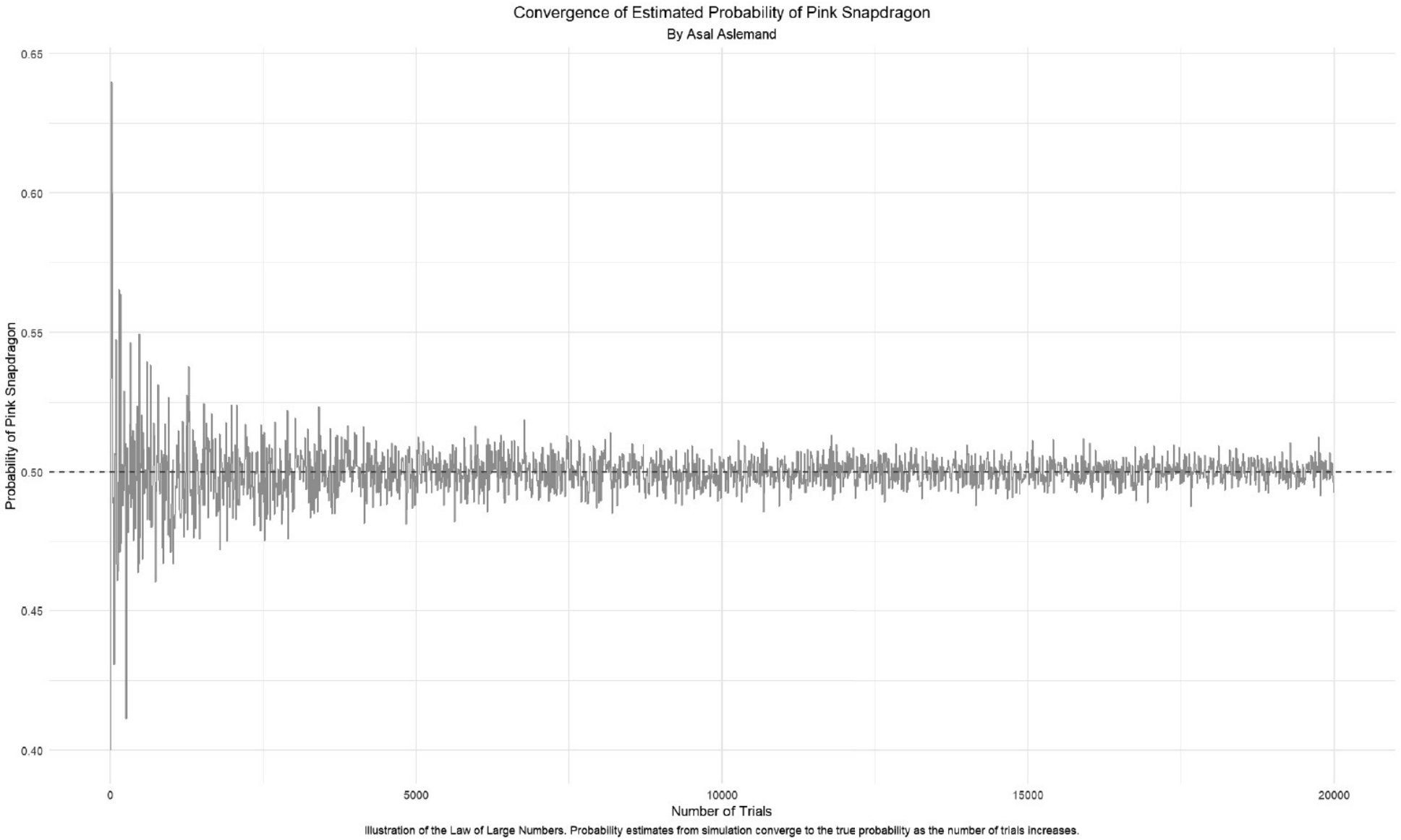
Snapdragon

In snapdragon (Antirrhinum majus), individual plants can be red flowered, pink flowered, or white flowered. According to a certain Mendelian genetic model, self pollination of pink-flowered plants should produce progeny that are red, pink, and white in the ratio 1:2:1. It is believed that the three colours occur with true probabilities 1/4, 1/2, and 1/4.

R Simulation

- Illustration of Empirical Probabilities:
  - Generate a large sample (n = 10000) from the probability distribution of coloured-flowered plants for snapdragons.
  - Estimate the probabilities for the three colours.
- Illustration of Law of Large Numbers:
  - Create a sequence of 5 to 20000 trial sizes.
  - For each trial size, generate a random sample from the probability distribution of coloured-flowered plants for snapdragons.
  - For each trial size, estimate the probability of pink coloured-flowered snapdragons.

Illustration of Law of Large Numbers



Students' Reflection

Attempts: 331 out of 337

I understand how to estimate probabilities by doing simulation in R.

Not at all	8 respondents	**2 %**
Very little	38 respondents	11 %
Somewhat well	102 respondents	30 %
Well	102 respondents	30 %
Very well	81 respondents	24 %
No Answer	6 respondents	2 %

-0

Discrimination Index ⓘ

2% answered correctly

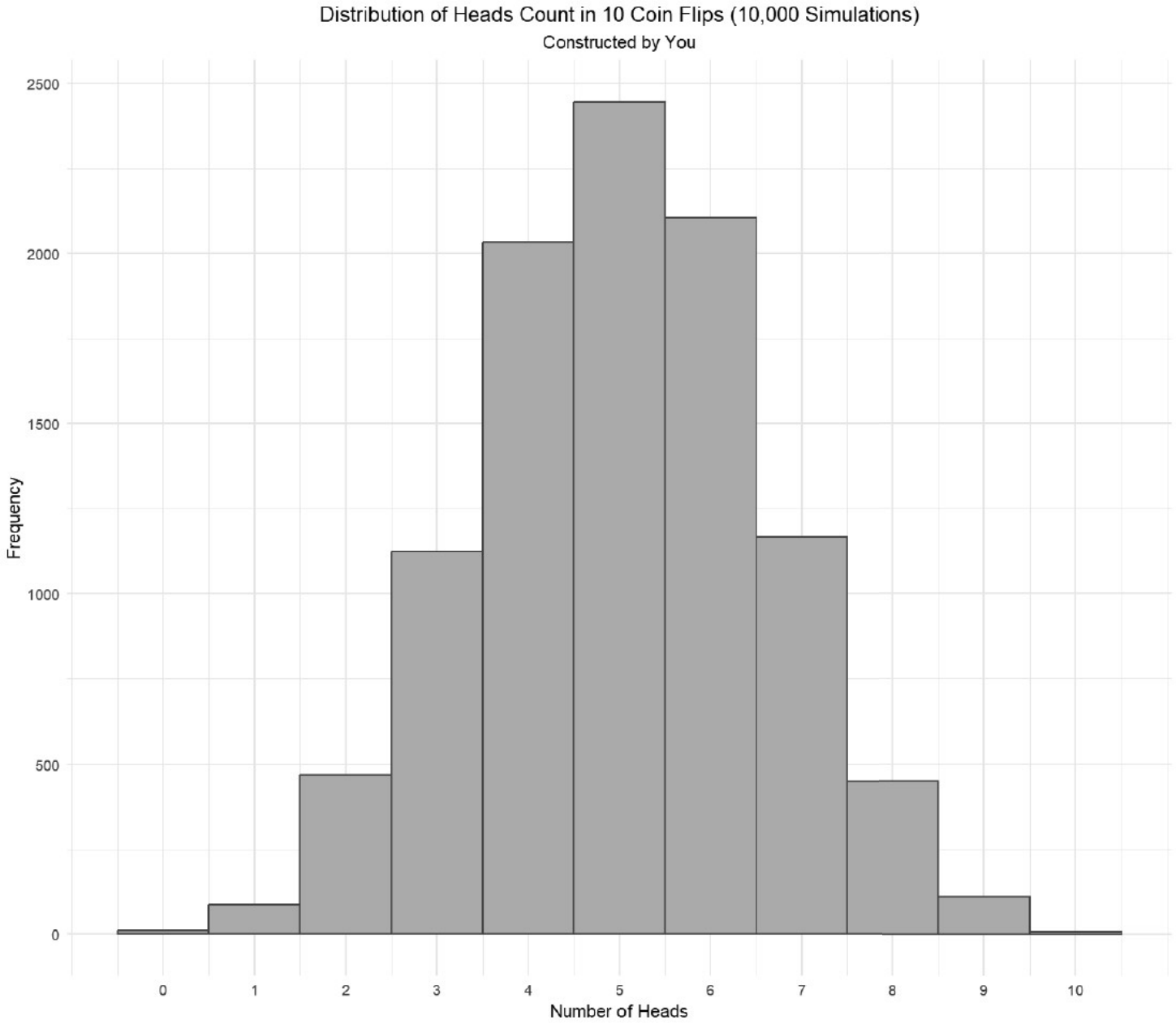
Coins and Cups: The Lady Tasting Tea

There is a famous story about a lady who claimed that tea with milk tastes different depending on whether the milk was added to the tea or the tea added to the milk. She preferred milk added first.
Let's suppose we decide to test the lady with ten cups of tea. We'll flip a coin to decide which way to prepare the cups. If we flip a head, we will pour the milk in first; if tails, we put the tea in first. Then we present the tea cups to the lady and have her state which ones she thinks were prepared each way.

R Simulation

- Let's figure out how hard it is to get 9 out of 10 correct just by guessing.
- Flip 10 coins. We'll call the heads correct guesses and the tails incorrect guesses. Then we'll flip 10 more coins, and 10 more, and .... That would get pretty tedious.
- Use R to perform 10,000 simulations of 10 coin flips. Counts the number of heads (correct guesses), ranging from 0 to 10.
- Display the probability distribution for the number of correct guesses.
- Estimate the probability that the lady could get 9 or 10 correct just by guessing. Interpret the result.
- Compare the estimated probability with the theoretical probability that the lady could get 9 or 10 correct just by guessing.

R Simulation



Students' Reflection

Attempts: 333 out of 334

I understand how to interpret counting methods by doing simulations in R.

Not at all	4 respondents	**1 %**	
Very little	32 respondents	10 %	
Somewhat well	91 respondents	27 %	
Well	113 respondents	34 %	
Very well	93 respondents	28 %	
No Answer	1 respondent	0 %	

-0

Discrimination Index ?

1% answered correctly

Mites and Wilt Disease

Researchers were interested in if attack of a plant by one organism induced resistance to subsequent attack by a different organism. Individually potted cotton plants (Gossypium) were randomly allocated to two groups. Each plant in one group received an infestation of spider mites (Tetranychus); the other group were kept as controls. After 2 weeks the mites were removed, and all plants were inoculated with Verticillium, a fungus that causes wilt disease.

Mites and Wilt Disease

The accompanying table shows the numbers of plants that developed symptoms of wilt disease.

	Wilt Disease		
Mites	No	Yes	Total
No	4	17	21
Yes	15	11	26
Total	19	28	47

- Compare the proportion of plants in the study with mites to no mites that did not develop Wilt disease.
- Find the relative risk of not developing Wilt disease, comparing mites to no mites. Interpret this number in the context of this problem.
- Consider all possible values for the number of plants in the no-mites group that did not develop Wilt disease. Identify values that would be considered more unusual (rare or extreme) than the observed data. Provide two such extreme values.

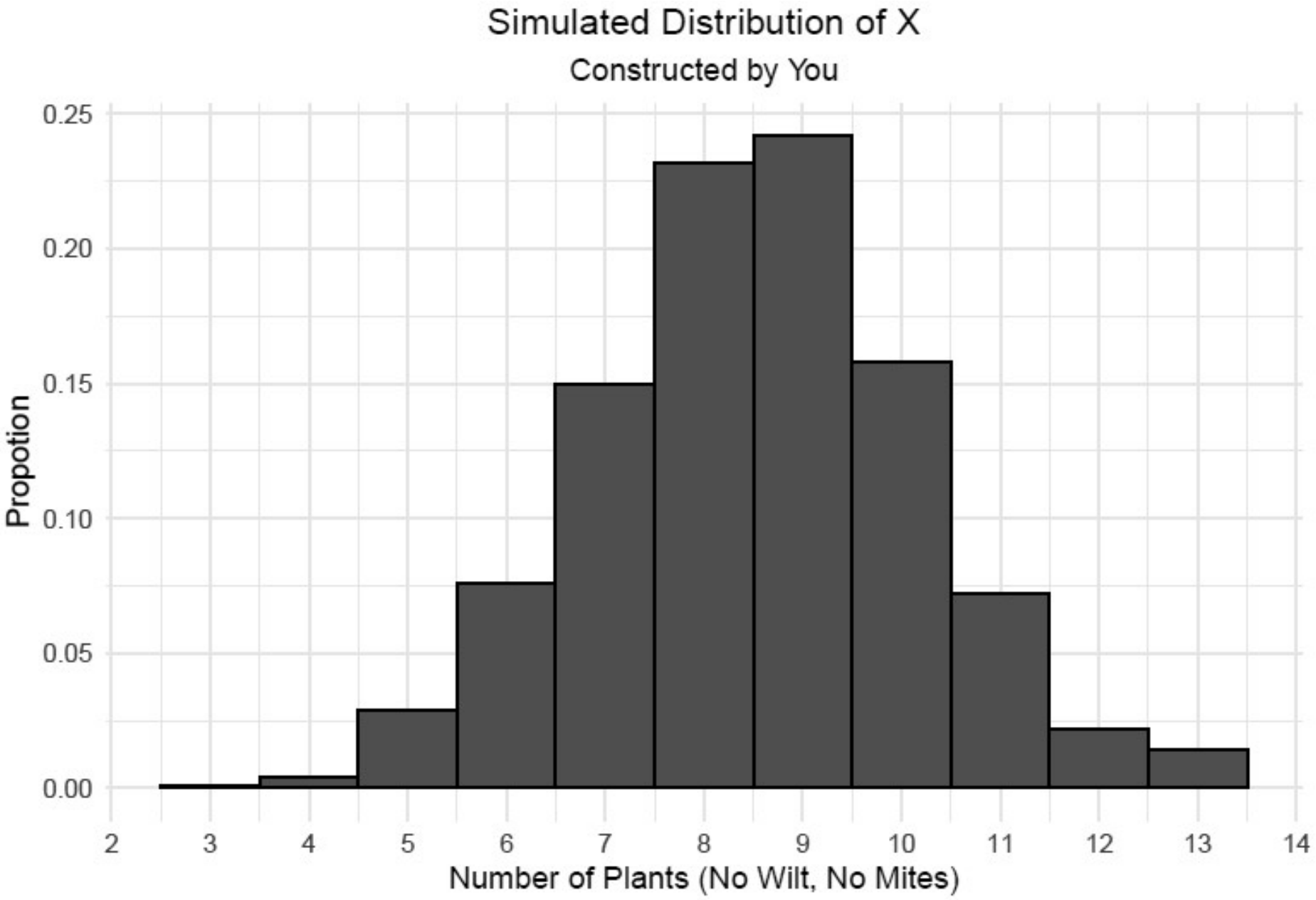
Physical Simulation

- Select 47 cards from your deck:
  26 red card (mites) and
  21 black (no mites)
- Shuffle the cards well.
- Deal out 19 cards.
  These represent the number plants without Wilt disease.
- Count the number of black cards among those 19.
  These represent the number of plants with no mites that did not develop Wilt disease.
- Repeat steps 2-4 four more times (five total).

R Simulation

- Use R to perform 1000 simulations (shuffle the mites distribution).
- In each simulation (each shuffle), count the number of plants in the no mites group that did not develop Wilt disease.
- Estimate the proportion of trials that an extreme event occurs.
- Displays the results using a frequency table, proportion table, and a histogram.
- Use R Markdown to compile results into a PDF document.

R Simulation



Students' Reflection

Attempts: 315 out of 319

-0

Discrimination Index

I can interpret the results from R simulations when estimating the probability of extreme events.

Not at all	7 respondents	**2 %**
Very little	20 respondents	6 %
Somewhat well	91 respondents	29 %
Well	113 respondents	35 %
Very well	84 respondents	26 %
No Answer	4 respondents	1 %

2% answered correctly

Law of Anomalous Numbers

In 1881, Simon Newcomb, a Canadian American astronomer and applied mathematician, noticed that fellow scientists using the logarithm tables were looking up numbers starting with 1 more often than numbers starting with 2, numbers with first digit 2 more often than 3, and so on. Newcomb proposed a law that the probability of a single number N being the first digit of a number was equal to $\log(N+1) - \log(N)$.

In 1938 Frank Benford, a physicist, examined 20 diverse data sets such as surface area of rivers, American League baseball statistics, atomic weights of elements, numbers appearing in Reader's Digest articles, and the street addresses, with total of 20,229 observations. He found that the number one appeared as a first digit 0.306 of the time in all cases, which is about equal to the logarithm of 2.

Probability Distribution of Leading Digits

Let X be the number denoting the leading digits 1, 2, . . . , 9. Benford's observations indicated that a likely distribution of first digits would look like the one shown in the table below:

Leading Digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Physical Simulation

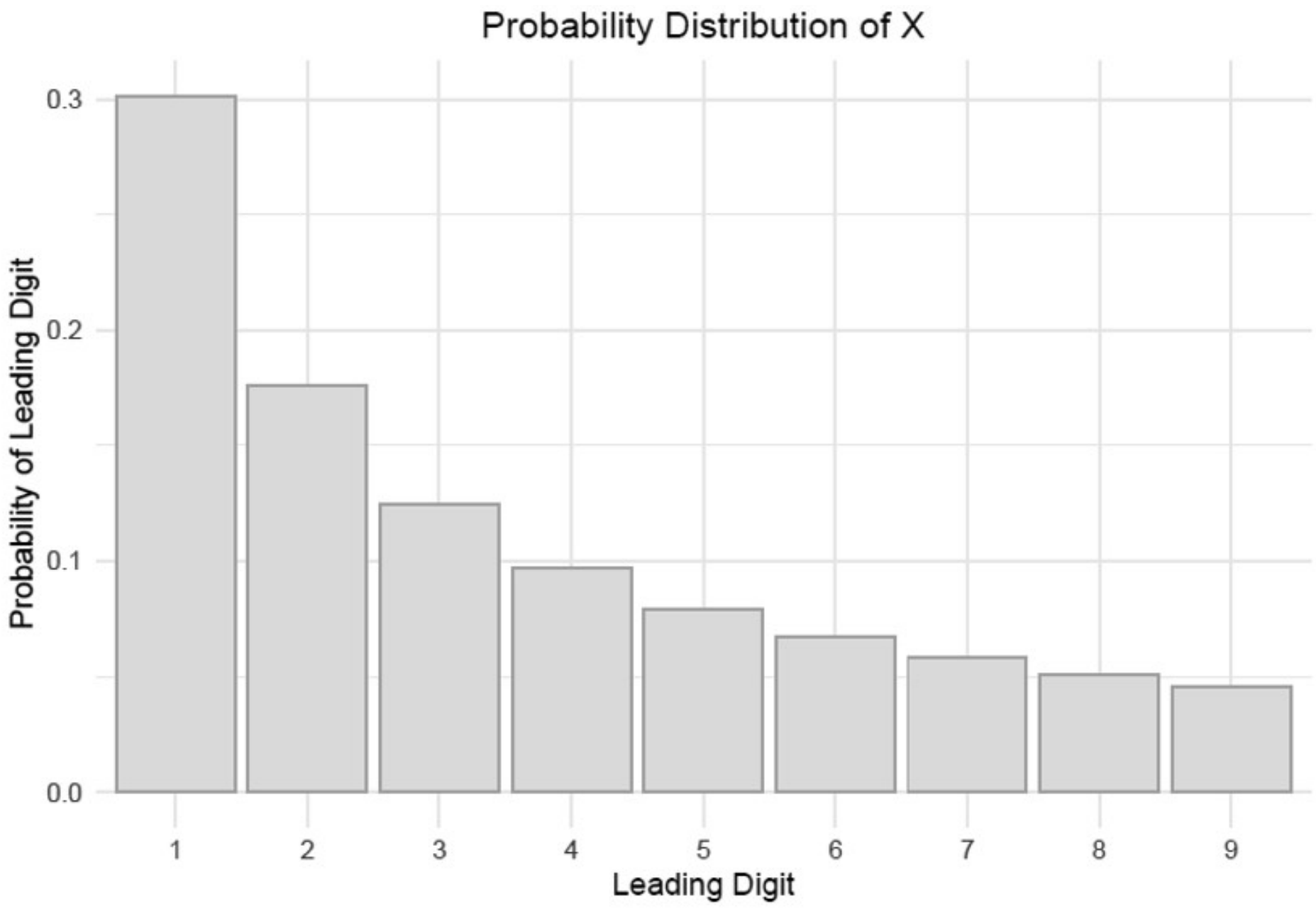
Watch Dr. Hanna Fry's video. Make notes of her live experiment.

Can Maths Predict the Future, Dr. Hanna Fry at Ada Lovelace Day 2014

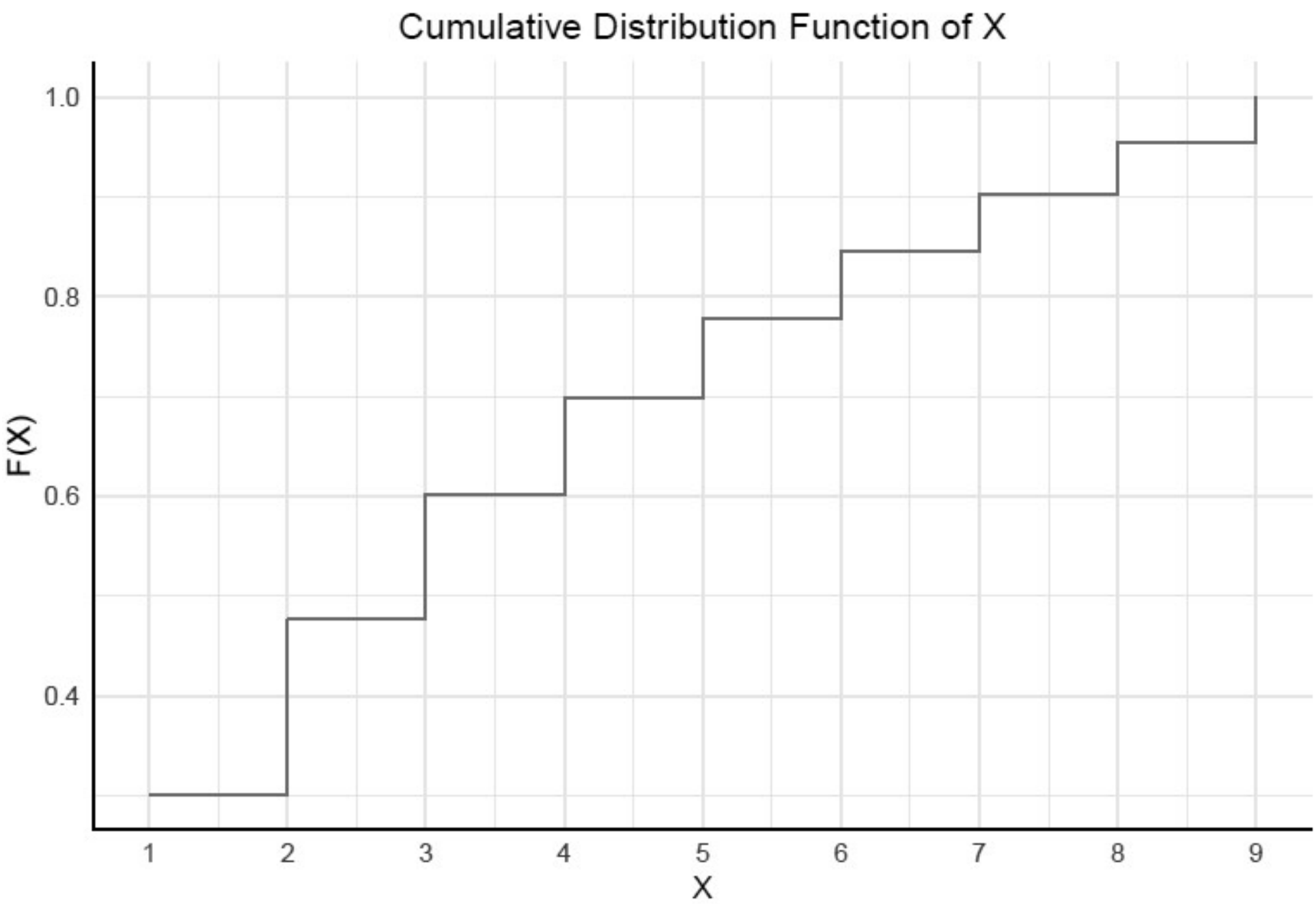
R Simulation

- Generate 20229 single digit random numbers between 1 to 9.
- Estimate probabilities for X.
- Displays the results using a frequency table, a proportion table, a histogram plot of estimated probabilities of X, and a plot of estimated cumulative distribution function of X.
- Use R Markdown to compile results into a PDF document.

R Simulation



R Simulation



Students' Reflection

Attempts: 340 out of 342

-0

Discrimination Index ⓘ

I understand how to use R to generate (simulate) large random samples from the probability distribution function of a random variable X and analyze how close the estimated probabilities get to the true probability values.

Not at all	10 respondents	**3 %**	
Very little	27 respondents	8 %	
Somewhat well	103 respondents	30 %	
Well	110 respondents	32 %	
Very well	90 respondents	26 %	
No Answer	2 respondents	1 %	

3% answered correctly

Teaching Assistants' Reflections

- Students are more engaged, especially with tutorials closely tied to assignments.
- Tutorials helps students see relevance, share ideas, and collaborate.
- Hands-on activities, like the card simulation, makes learning more interactive.
- Adding R provides another way to explain concepts engagingly.
- Having access to R allows for interactive demonstrations.
- R assignments let students learn beyond lectures.
- Providing R codes with video demos helps students learn while being assessed.
- Graded worksheets encourage effort in a low-stress, collaborative setting.

References (Literature Review)

Chance, B., McGaughey, K., Chung, S., Goodman, A., Roy, S., & Tintle, N. (2024). *Simulation-Based Inference: Random Sampling vs. Random Assignment? What Instructors Should Know.* Journal of Statistics and Data Science Education, 1–10.

Jamie, D. M. (2002). *Using Computer Simulation Methods to Teach Statistics: A Review of the Literature.* Journal of Statistics Education, 10(1).

Kalaian, S. A., & Kasim, R. M. (2014). *A Meta-Analytic Review of Studies of the Effectiveness of Small-Group Learning Methods on Statistics Achievement.* Journal of Statistics Education, 22(1).

Parke, C. S. (2008). *Reasoning and Communicating in the Language of Statistics.* Journal of Statistics Education, 16(1).

Acknowledgement