COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 7: Histograms

Copyright © 2025 Barnard College

Sept 24, 2025

Reminders

- HW 1 is due Today
  - HW 2 due next week Wednesday
  - Remember to run the last cell

Barnard BC 1016 Submission Instructions

First, make sure you have run all cells up to this point. Next, run the last cell in this notebook (`grader.export(pd` `run_tests=True)`) to run a last set of tests.

Once you can view all the test results from running the `grader.export` cell, please export a PDF of your noteboo these instructions:

1. Go to File > Print
2. Select the option to export a PDF. In the preview, ensure that your entire notebook contents are visible including results.
3. Save the PDF and submit via Courseworks. Homeworks should be submitted to BC1016, and Lab assignments s submitted to BC1017

Note, please ignore the Data8 instructions about submitting a zip file - you should still run the grader.export a PDF instead as per the instructions above

Submission

Make sure you have run all cells in your notebook in order before running the cell below, so that all images/graphs ap output. The cell below will generate a zip file for you to submit. **Please save before exporting!**

```
[50]: # Save your notebook first, then run this cell to export your submission.
grader.export(pdf=False, run_tests=True)
```

```
Running your submission against local test cases...


Your submission received the following results when run against available test cases:

    q3_1_2 results: All test cases passed!

    q3_3_1 results: All test cases passed!

    q3_3_2 results: All test cases passed!

    q4_1_1 results: All test cases passed!

    q51 results: All test cases passed!

    q5_1_1 results: All test cases passed!
```

Your submission has been exported. Click here to download the zip file.

Type Markdown and LaTeX: α^2

Lecture Outline

- Histograms (continued)
- Census Demo
- Chart Selection Demo with Weather Data

Visualizing Categorical Data

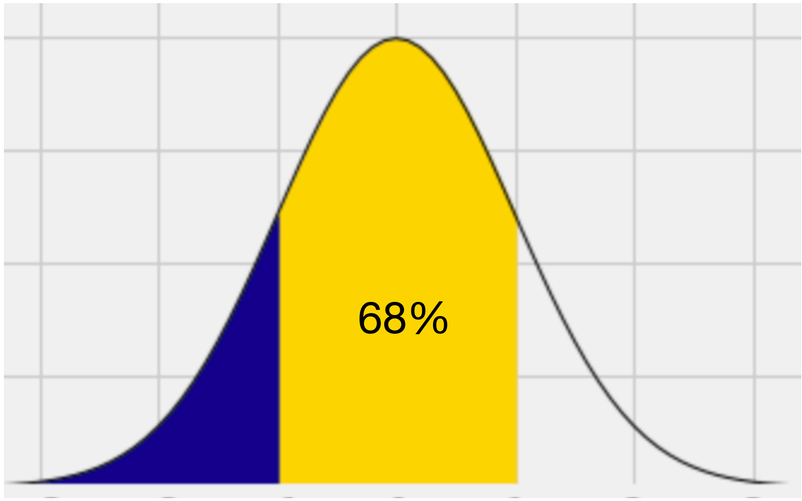
Area Principle for Histograms

Area Principle

Areas should be *proportional* to the values they represent

In a histogram, the area of each bar is the percent of individuals in the corresponding bin

(Later on in the course, we will approximate histograms with smooth curves)

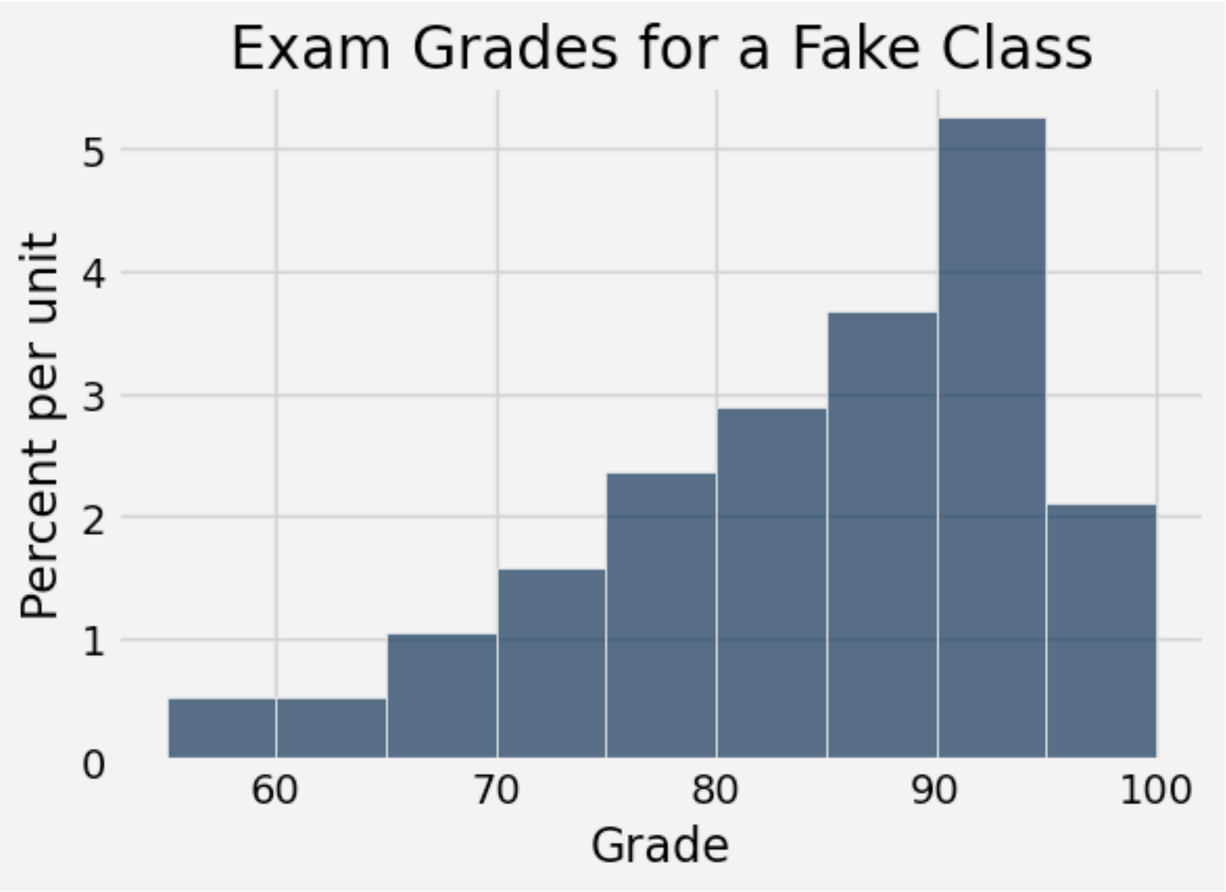


Histograms

The area of each bar is a percentage of the whole

The horizontal axis is a numerical distribution - the bins don’t need to be of equal size

The vertical axis is a rate (e.g., percent/year) - density

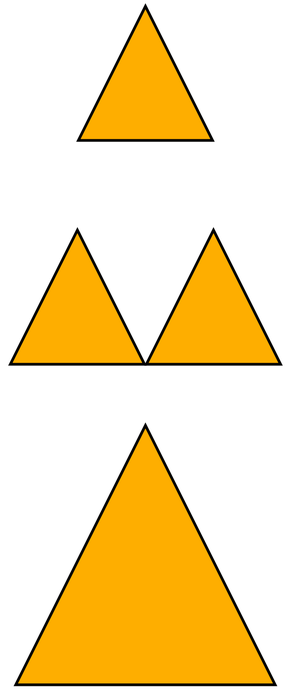


Bin size = 5

Area Principle

Areas should be proportional to the values they represent

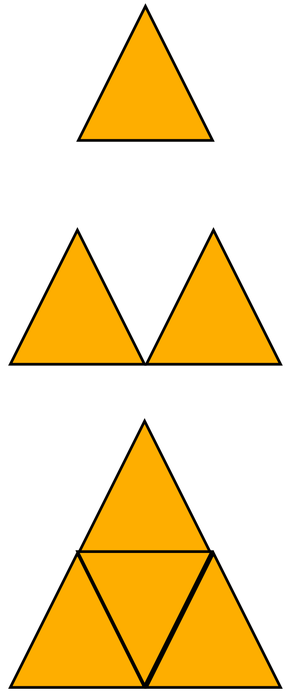
- For example
  - If you represent 20% of a population by:
  - Then 40% can be represented by:
  - But not by:



Area Principle

Areas should be proportional to the values they represent

- For example
  - If you represent 20% of a population by:
  - Then 40% can be represented by:
  - But not by:



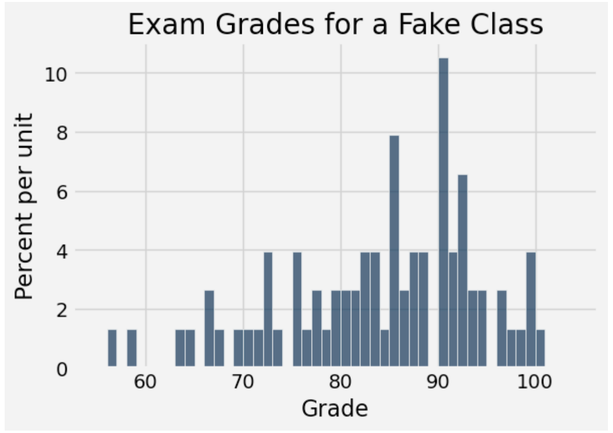
Area Principle

$$\text{area of bar} = \text{percent of entries in bin}$$

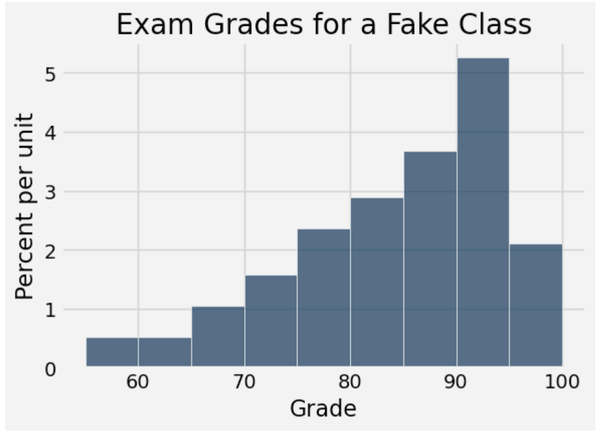
$$\text{area of bar} = (\text{height of bar}) \times (\text{width of bin})$$

$$\text{height of bar} = \frac{\text{area of bar}}{\text{width of bin}} = \frac{\text{percent of entries in bin}}{\text{width of bin}}$$

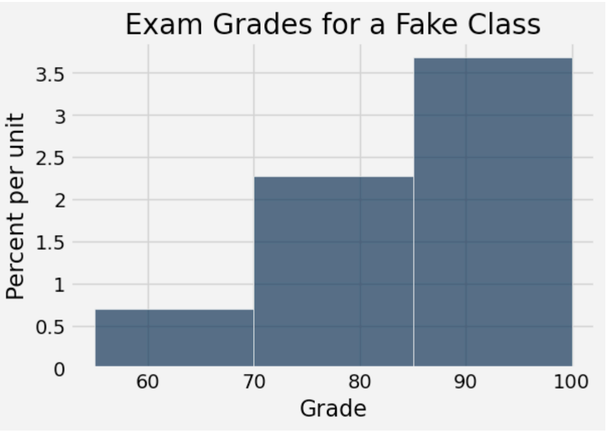
Bin size = 1



Bin size = 5



Bin size = 15



Area Principle

$$\text{area of bar} = \text{percent of entries in bin}$$

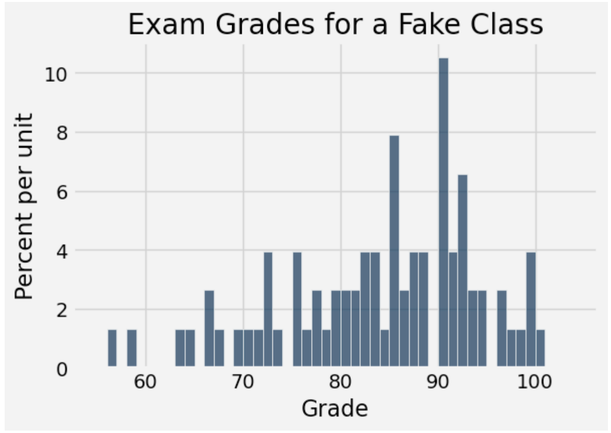
$$\text{area of bar} = (\text{height of bar}) \times (\text{width of bin})$$

$$\text{height of bar} = \frac{\text{area of bar}}{\text{width of bin}} = \frac{\text{percent of entries in bin}}{\text{width of bin}}$$

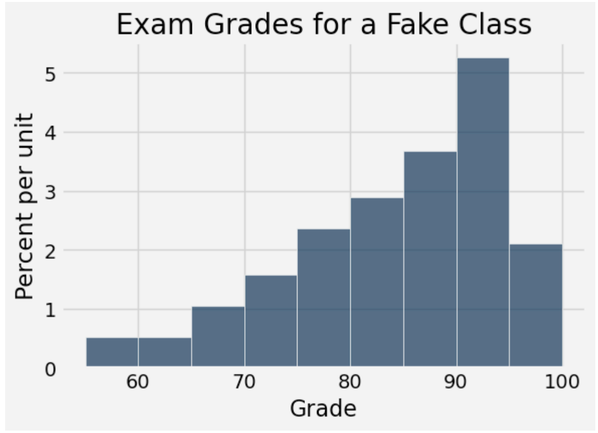
when bin size is 1:

$$\text{height} = \frac{\text{area of bar}}{\text{width of bin}} = \frac{\text{area of bar}}{1} = \frac{\%\text{ of entries in bin}}{1}$$

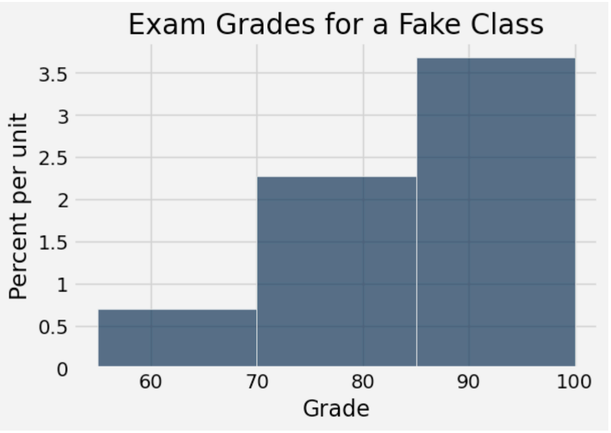
Bin size = 1



Bin size = 5



Bin size = 15



Area Principle

$$\text{area of bar} = \text{percent of entries in bin}$$

$$\text{area of bar} = (\text{height of bar}) \times (\text{width of bin})$$

$$\text{height of bar} = \frac{\text{area of bar}}{\text{width of bin}} = \frac{\text{percent of entries in bin}}{\text{width of bin}}$$

When bin size is 1:

$$\text{height} = \frac{\text{area of bar}}{\text{width of bin}} = \frac{\text{area of bar}}{1} = \frac{\%\text{ of entries in bin}}{1}$$

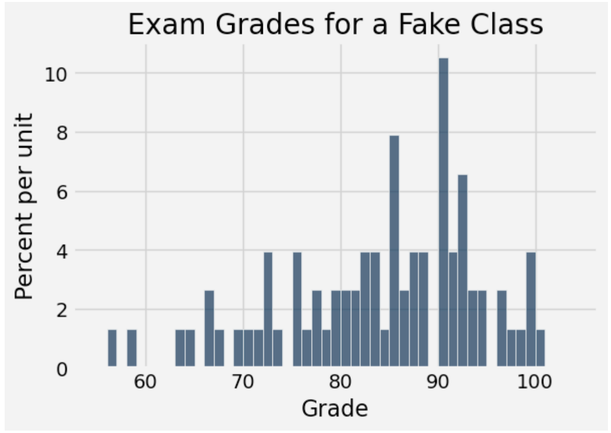
When bin size is 5:

$$\text{height} = \frac{\%\text{ entries in bin}}{5}$$

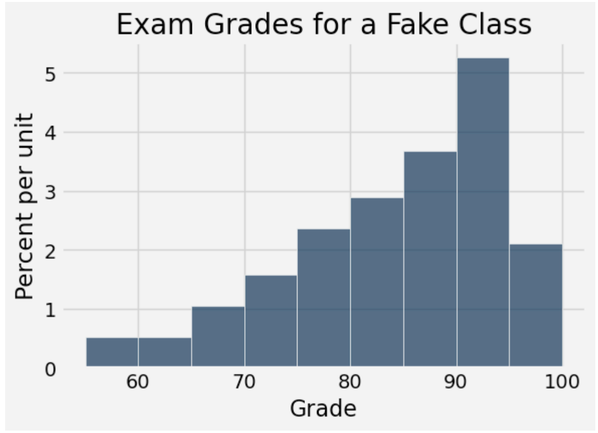
When bin size is 15:

$$\text{height} = \frac{\%\text{ entries in bin}}{15}$$

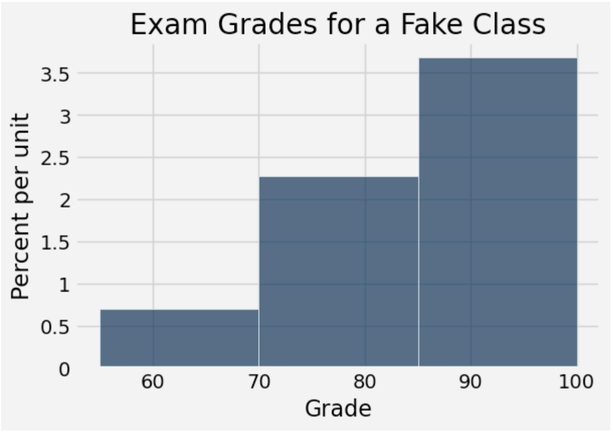
Bin size = 1



Bin size = 5



Bin size = 15

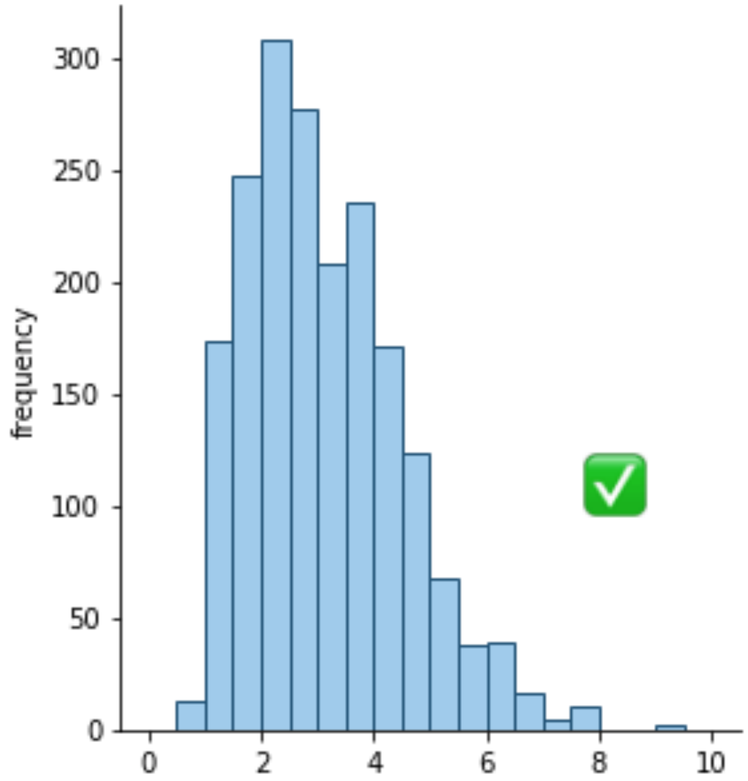


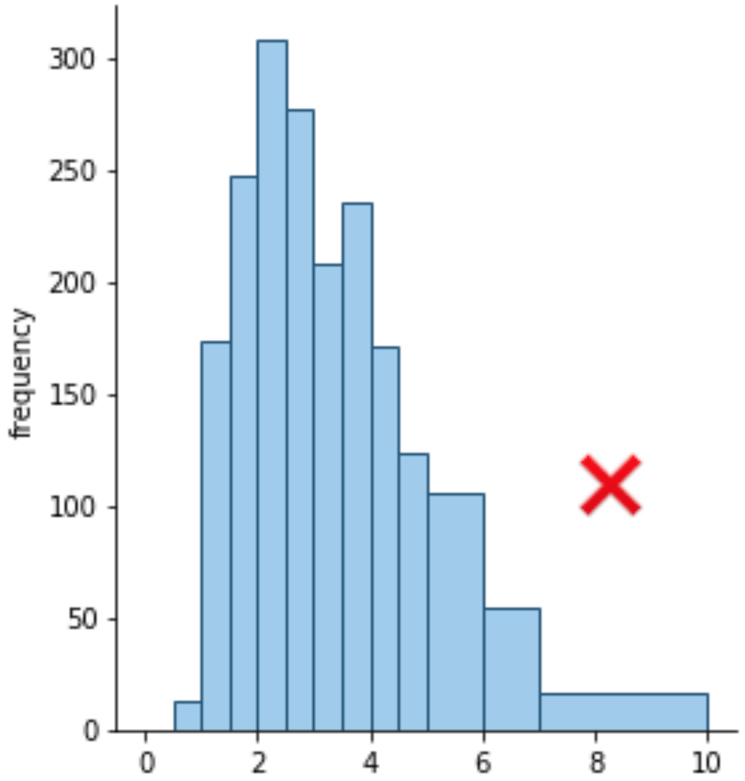
Unequal Bin Sizes

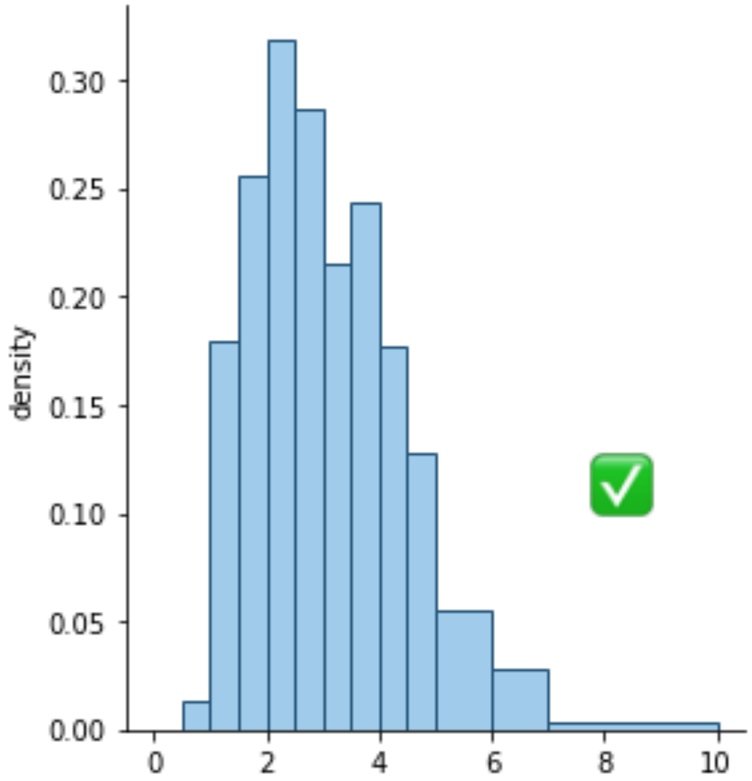
Bin sizes don’t need to be equal

- unequal bin size is often used for better representing tails

For unequal bin sizes - vertical axis now represents ***density*** rather than frequency







https://www.atlassian.com/data/charts/histogram-complete-guide

Calculating Heights

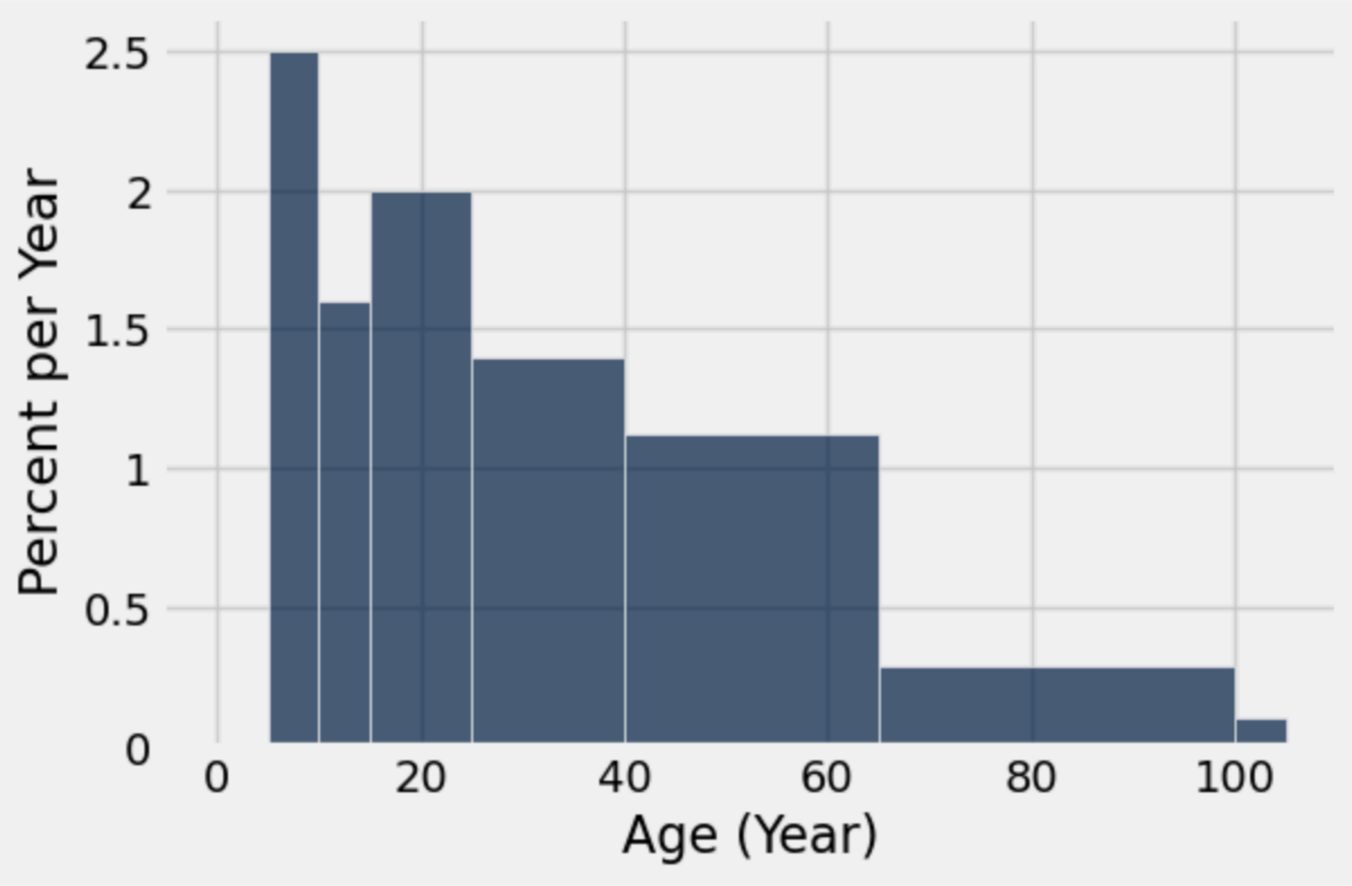
The [40, 65) bin contains 59/200 items

- The bin is 29.5% (59/200) of the whole

- The bin width is 65-40 = 25 years

- Height = $\frac{\text{29.5 percent}}{\text{25 years}}$

$$= 1.18\% \text{ per year}$$



Area Notebook Demo

Bar Chart vs Histogram

Bar Chart

- Distribution of categorical variable
- Length of bars is proportional to the frequency / percent of individuals

Histogram

- Distribution of numerical variable
- Horizontal axis is numerical, bins can be unequal
- Area of bars is proportional of percent of individuals, height measures density

Charts Summary

Type	Syntax	Description
Line graph	`.plot(x_axis, y_axis)`	Sequential data
Scatter Plot	`.scatter(x_axis, y_axis)`	Relation between two numerical values
Bar Chart	`.barh(column_label)`	Distribution of one categorical variable (already grouped)
Histogram	`.hist(column_label, unit, bins)`	Distribution of one numerical variable

Chart Selection Exercise

We have NYC weather data from 2019 as shown below (from Kaggle)

Which type of chart (line, scatter, bar, histogram) would best help you answer to each question?

- Do days with hotter highs also tend to have hotter lows?
- How do the number of rainy days compare with the number of snowy days?
- What percent of days have a high of at least 75 degrees?

date	tmax	tmin	tavg	condition
1/1/19	60	40	50	rainy
2/1/19	41	35	38	
3/1/19	45	39	42	
4/1/19	47	37	42	
5/1/19	47	42	44.5	rainy
6/1/19	49	32	40.5	
7/1/19	35	26	30.5	
8/1/19	47	35	41	rainy
9/1/19	46	35	40.5	rainy
10/1/19	35	30	32.5	

Census Demo

Next Class

- Functions, Groups, Pivots, and Joins