#10/11: Histograms and Bar Graphs

September 20th, 2019

Karle Flanagan and Wade Fagen-Ulmschneider

Towards Data Visualization

As we've seen previously, data frames and tables are a powerful way of organizing and displaying summary data. However, large tables of numbers can be difficult to interpret, no matter how organized they are. Sometimes it is much easier to interpret graphs than numbers.

Before we start looking at different ways to visualize data, let's talk about the two main types of data that we will analyze in this class:

TYPES OF VARIABLES

At the beginning of any statistical analysis, it is important for you to identify the nature of the data that are available. One key distinction is the difference between *quantitative* and *qualitative* data.

_____ data are naturally measured numerically. Many of the variables that data scientists study are *quantitative* or *numerical*. Their values are numbers on which you can perform arithmetic. Quantitative data can be further divided into two types:

_____ Variables:	**_____ Variables:**

Qualitative (categorical) data are not numerically measured. Instead, the data fall into descriptive categories.

List a few qualitative variables:

Examples: ***What type of variables are the following?***

a) Occupation:

b) The number of deaths in US prisons last year:

c) City You Were Born In:

d) Height:

e) Zip Code:

f) Number of Illini wins in football against Purdue:

*__Quantitative__ data can be displayed using a *histogram* because the data range can be divided into numeric intervals; areas make sense.

*__Qualitative__ data must be displayed using more generic *bar graphs*, because the data cannot be divided into numeric intervals, and areas have no real meaning.

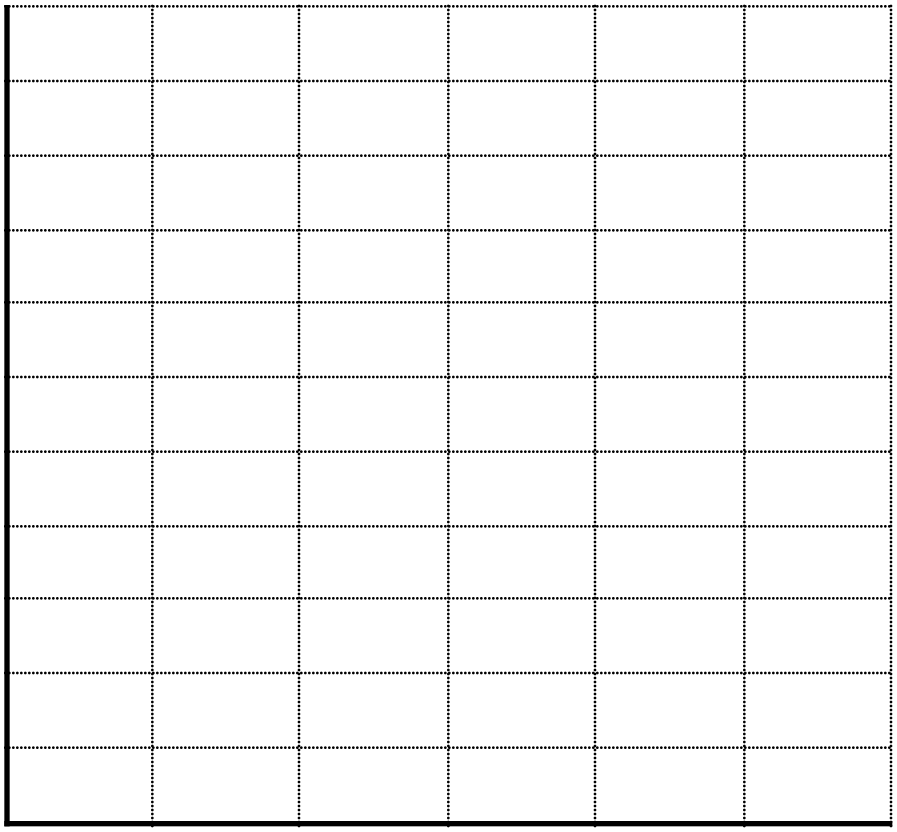
Bar Graphs vs. Histograms

A _____ is a familiar way of visualizing qualitative (categorical) variables/distributions. It displays a bar for each category. The bars are equally spaced and equally wide. You can create bar graphs with horizontal bars or vertical bars.

Bars use height to depict frequencies or counts. The widths of the blocks in bar graphs don't mean anything. They're all the same regardless of the size of the category.

Example 1: Draw a bar graph using the following data from the GPA dataset. First, find the number of sections in each category, then label your x and y-axis, and lastly, draw the bar graph.

Category:	**STAT**	**CS**	**IS**	**MATH**	**CHEM**	**PHYS**
# of Sections:						



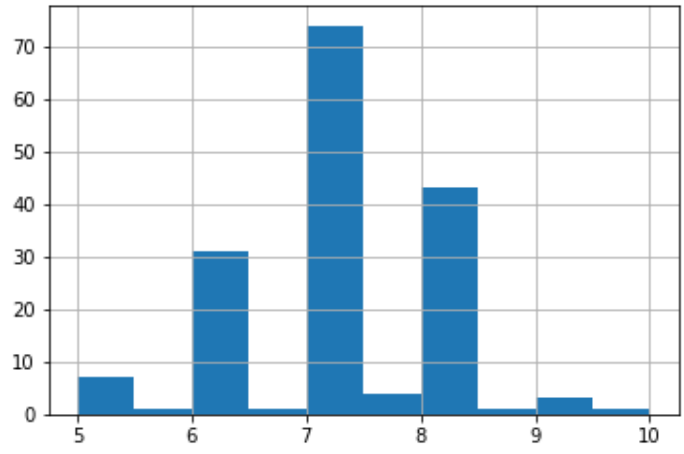
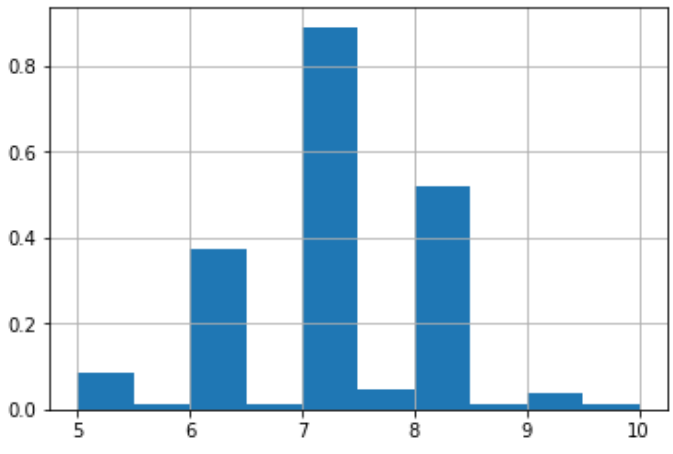
A _____ is a common way of visualizing quantitative variables/distributions. Unlike bar graphs, the x-axis of a histogram is always drawn to scale. Histograms can help us visualize the shape of a distribution.
There are two types of histograms that are commonly used:

1) _____ **Histogram:** Frequency histograms show how often different values in a dataset occur. The width of each block indicates the size of the interval and the height of each bar indicates the number of observations in each interval (or bin). Frequency histograms are very *simple* to interpret and read.

2) _____ **Histogram:** Density histograms use areas to depict percentages. The width of each block indicates the size of the interval and the area of each block indicates what percentage of the data belongs to that category. The height indicates density or how crowded the block is. Density histograms are used a lot in *statistics*.

IMPORTANT NOTE: In a density histogram, the total area of all the blocks is 1 or 100%.

Example 2: Here are histograms representing the hours of sleep from the "Hello" dataset:

Frequency Histogram	**Density Histogram**
	
`df["Hours of Sleep"].hist()`	`df["Hours of Sleep"].hist(density=True)`

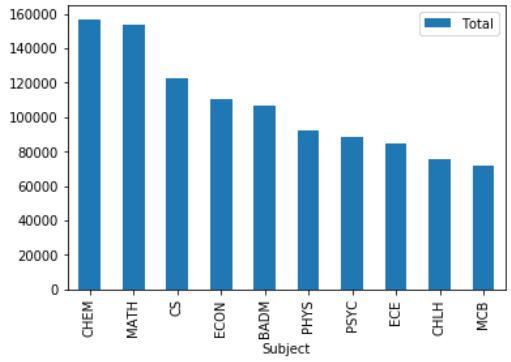
a) Is the number of students who got 5-6 hours of sleep greater than, less than, or equal to the number who got 8-10 hours of sleep?

b) How many students get 8 or more hours of sleep?

c) Suppose we all get 1 extra hour of sleep after exam week is over. How would the histogram change?

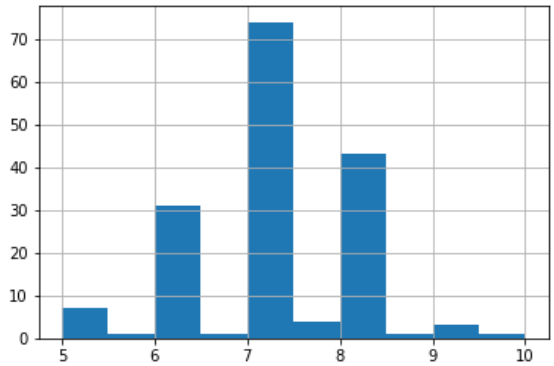
Bar Graphs and Histograms in Python!

Creating a Bar Graph

Python:	**General Syntax:**
	`df.plot.bar()`
	
	Example Shown: *(using GPA Dataset)*
	`g = df.groupby("Subject")`
	`df_subject = g.agg("sum")`
	`df_subject["Total"] = df_subject["A+"] +`
	`df_subject["A"] + ...`
	
	`df_subject = df_subject[ ["Total"] ]`
	`df_subject = df_subject.nlargest(10, "Total")`
	`df_subject.plot.bar()`
		

Description:	Plots the numeric values in of a DataFrame as a Bar Chart.	

Creating a Histogram

Python:	**General Syntax:**
	`df.hist()`
	
	Options:
	`bins = ##` `(# of bins)`
	`density = True` `(for %s)`
	
	Example Shown: *(using Hello dataset)*
	`df["Hours of Sleep"].hist()`
		

Description:	Plots the numeric values in of a DataFrame as a histogram.	

Data Science Discovery – Things To Be Doing:

1. **New EC Survey:** "Perception of Probability Words", +1 if completed by Sunday at 11:59pm
2. **Homework #3** is due Sunday, Sept. 22; Homework #4 out this weekend
3. **lab_gpa** due Monday (Sept. 23) by 11:59pm
4. **Midterm Exam 1:** *Monday, Sept. 30 - Wednesday, Oct. 2*
    a. Sign up for your slot to take the exam on https://cbtf.engr.illinois.edu/sched/
    b. More details next week, including practice exams!
5. **Open Office Hours:** Every M/W/R/F from 4:00pm - 6:30pm in 23 Illini Hall
6. **Extra Credit +1 Notebook** every lecture: out after lecture, due 11:30am before next lecture