Visualizing Data: Freq. Tables, Histograms

Engineering Statistics
Section 1.2

Josh Engwer

TTU

25 January 2016

Descriptive Statistics

Recall that Statistics consists of two broad branches:

- Descriptive Statistics
- Statistical Inference

The remainder of this chapter focuses squarely on Descriptive Statistics:

Definition

Descriptive Statistics is the organization, summary, visualization and presentation of data that conveys useful information about the data.

Descriptive Statistics involves:

- Data Visualization (this section)
- Numerical Summaries (Sections 1.3 & 1.4)

Data Visualizations

The following data visualizations will never be considered in this course:

- Stem-and-Leaf Displays
- Bar Charts
- Pie Charts
- Control Charts
- Scatter Plots
- Dot Plots
- Line Plots
- Contour Plots
- Radar Plots
- Violin Plots
- Heatmaps

So which data visualizations will be considered in this course??

- Frequency Tables (this section)
- Histograms (this section)
- Boxplots (Section 1.4)
- Frequency Polygons (Chapter 3)

Frequency Tables

Given a sample of eye colors:

H, Br, Br, Br, S, A, H, H, G, A, Bl, Bl, Br, Bl, A, Br, H, G, A, A, Br, Bl, G, Bl, Bl

Then the resulting frequency table is:

EYE COLOR	FREQUENCY	RELATIVE FREQUENCY
Amber (A)		
Blue (Bl)		
Brown (Br)		
Green (G)		
Hazel (H)		
Silver (S)		
TOTAL:		

Frequency Tables

Given a sample of eye colors:

H, Br, Br, Br, S, A, H, H, G, A, Bl, Bl, Br, Bl, A, Br, H, G, A, A, Br, Bl, G, Bl, Bl

Then the resulting frequency table is:

EYE COLOR	FREQUENCY	RELATIVE FREQUENCY
Amber (A)	5	
Blue (Bl)	6	
Brown (Br)	5	
Green (G)	3	
Hazel (H)	4	
Silver (S)	1	
TOTAL:	**24**	

The **frequency** entails from counting the # data points of a given category.

Then compute the total frequency.

Frequency Tables

Given a sample of eye colors:

H, Br, Br, Br, S, A, H, H, G, A, Bl, Bl, Br, Bl, A, Br, H, G, A, A, Br, Bl, G, Bl, Bl

Then the resulting frequency table is:

EYE COLOR	FREQUENCY	RELATIVE FREQUENCY
Amber (A)	5	$5/24 \approx 0.208$
Blue (Bl)	6	$6/24 = 0.250$
Brown (Br)	5	$5/24 \approx 0.208$
Green (G)	3	$3/24 = 0.125$
Hazel (H)	4	$4/24 \approx 0.167$
Silver (S)	1	$1/24 \approx 0.042$
TOTAL:	**24**	1.000

Each category's **relative frequency** is its frequency divided by the total freq.

Round decimals to three decimal places as necessary.

The total relative frequency should be very close to one (btw 0.998 & 1.002)

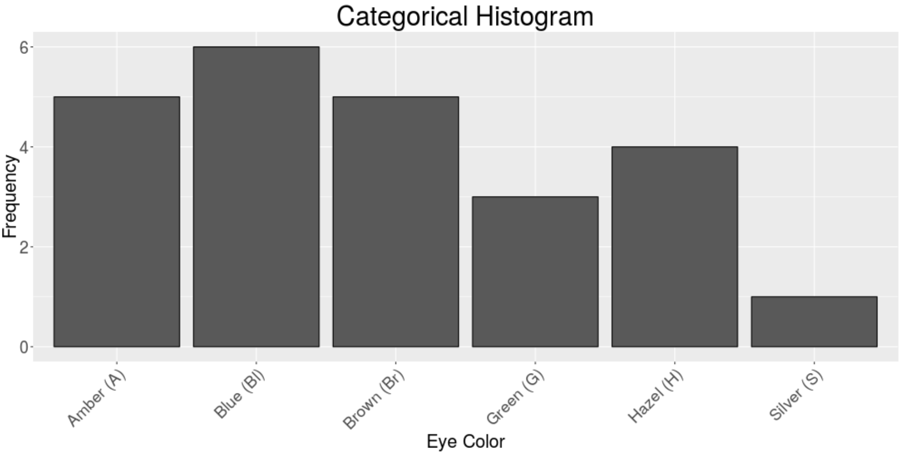
Frequency tables can also be made for numerical data (see the 1.2 Outline).

Histograms for Categorical Data

Given a sample of eye colors:

Br, G, Br, Br, S, A, H, H, G, A, Bl, Bl, Br, Bl, A, Br, H, G, A, A, Br, Bl, G, Bl, Bl

Then the resulting histogram is:

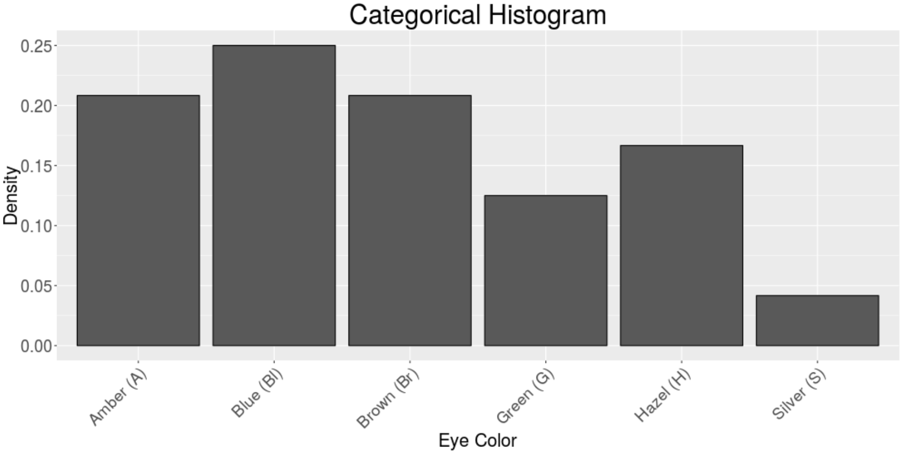


Histograms for Categorical Data

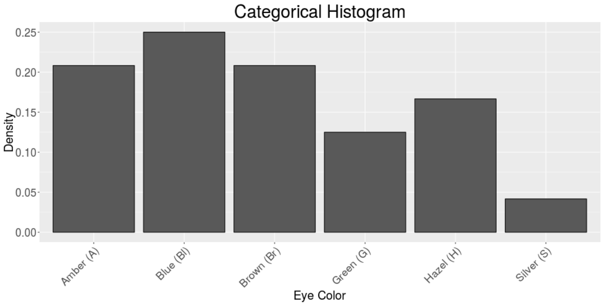
Given a sample of eye colors:

Br, G, Br, Br, S, A, H, H, G, A, Bl, Bl, Br, Bl, A, Br, H, G, A, A, Br, Bl, G, Bl, Bl

Then the resulting histogram is:



Histograms for Categorical Data



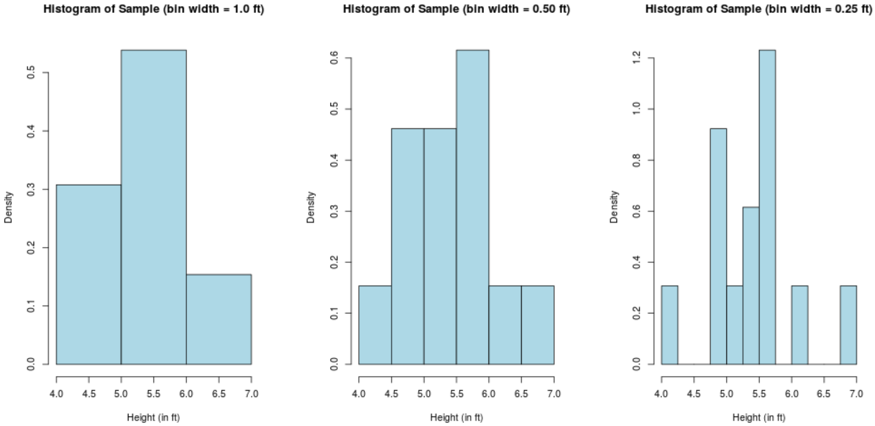
The vertical axis of a histogram is always one of the following:

- Frequency (count of each category/bin)
- Relative Frequency $=$ (Frequency)$/$(Total Frequency)
- Percent (%) $=$ Relative Frequency $\times\ 100\%$
- Density
  - Categorical Data: Density $=$ Relative Frequency
  - Numerical Data: Density $=$ (Relative Frequency)$/$(Bin Width)

Histograms for Discrete Data (Equal Bin Widths)

Given a sample: 4.9, 4.9, 5.0, 5.7, 6.2, 5.3, 5.2, 5.5, 5.6, 5.7, 5.7, 4.1, 6.8

Here are three histograms using equal bin widths:



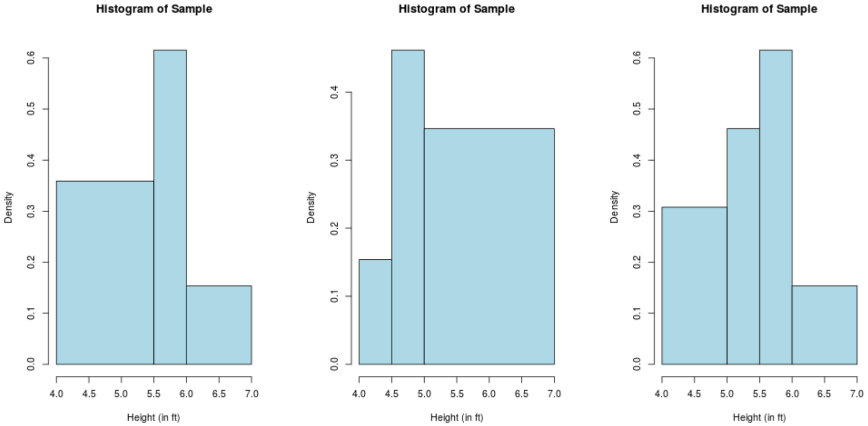
Pick a bin width that avoids gaps (right figure) and ”overlumping” (left figure).

For this course, bin widths will be chosen a priori.

Histograms for Discrete Data (Unequal Bin Widths)

Given a sample: 4.9, 4.9, 5.0, 5.7, 6.2, 5.3, 5.2, 5.5, 5.6, 5.7, 5.7, 4.1, 6.8

Here are three histograms using unequal bin widths:

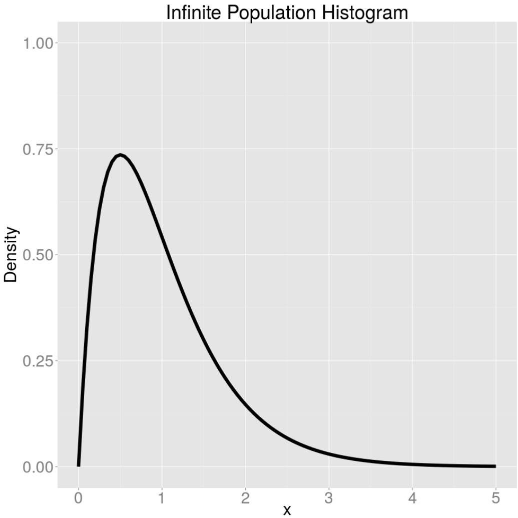


Unequal bin widths are useful when there are some isolated data points.
For this course, bin widths will be chosen a priori.

Histograms for Continuous Numerical Data

Continuous numerical data usually refer to infinite populations.

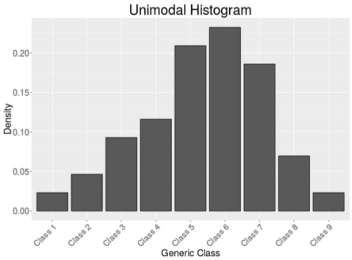
A histogram for an infinite population looks like a smooth curve:



Modality of Discrete & Categorical Data

Definition

A dataset is **unimodal** if its histogram has exactly one peak.
A dataset is **bimodal** if its histogram has exactly two peaks.
A dataset is **multimodal** if its histogram has many peaks.

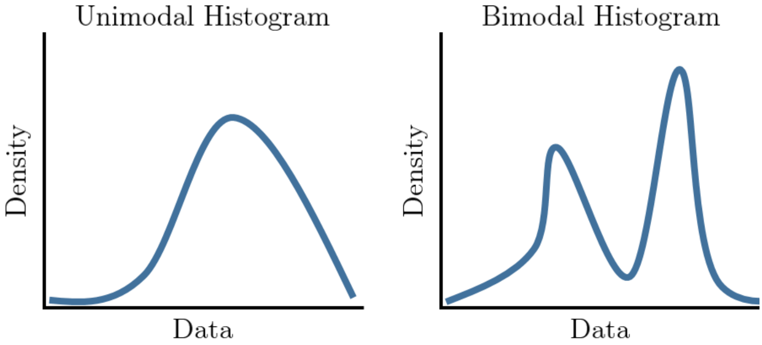


See page 22 of the textbook for an example of **multimodal** discrete data.

Modality of Continuous Data

Definition

A dataset is **unimodal** if its histogram has exactly one peak.
A dataset is **bimodal** if its histogram has exactly two peaks.
A dataset is **multimodal** if its histogram has many peaks.

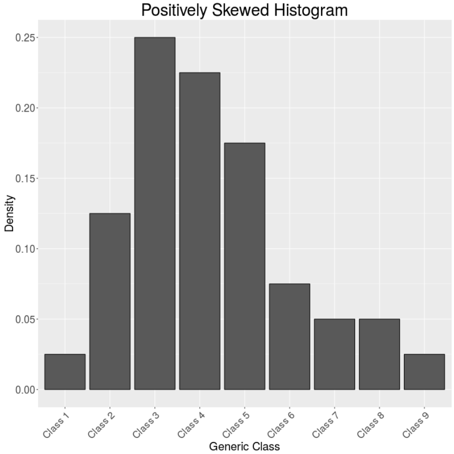


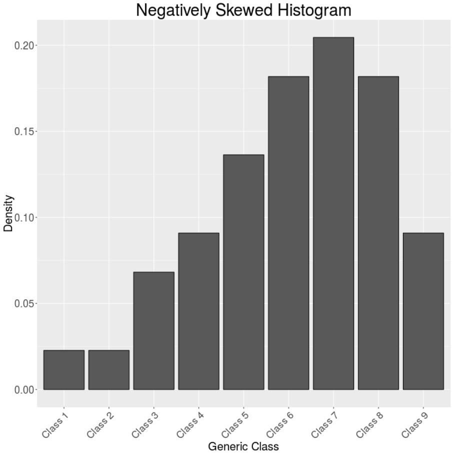
See page 23 of the textbook for an example of **multimodal** continuous data.

Skewness of Discrete & Categorical Data

Definition

A dataset is **positively skewed** if its histogram has a long upper tail.
A dataset is **negatively skewed** if its histogram has a long lower tail.

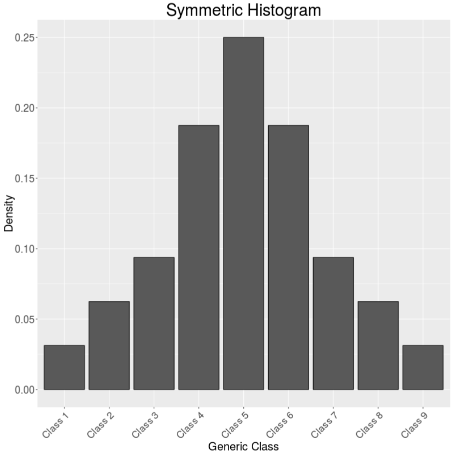




Skewness of Discrete & Categorical Data

Definition

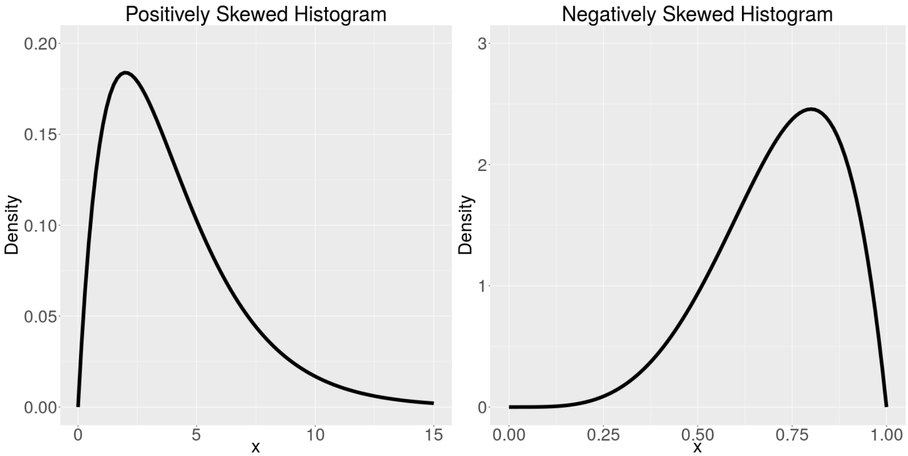
A dataset is **symmetric** if its histogram's left half and right half are mirror images of each other.



Skewness of Continuous Data

Definition

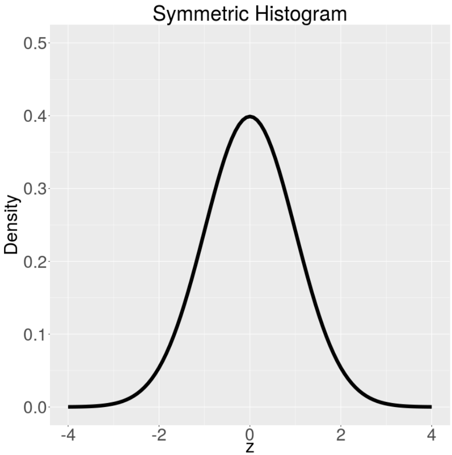
A dataset is **positively skewed** if its histogram has a long upper tail.
A dataset is **negatively skewed** if its histogram has a long lower tail.



Skewness of Continuous Data

Definition

A dataset is **symmetric** if its histogram's left half and right half are mirror images of each other.

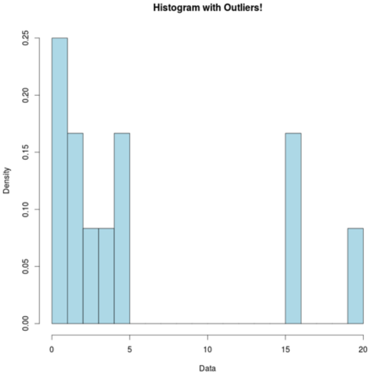


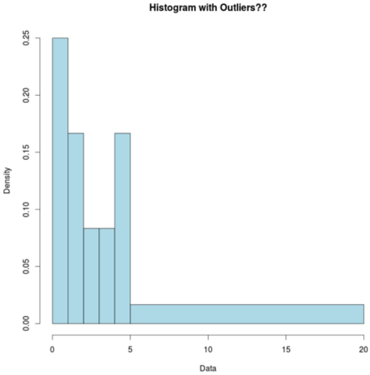
Outlier(s) in Discrete & Categorical Data

Definition

A data point in a dataset is an **outlier** if it is ”far away” from ”most” of the data.

Consider the dataset: 1,5,2,2,1,4,1,3,20,5,16,16





The left histogram (with equal bin widths) suggest that 16 & 20 are outliers.
But identifying outliers is unclear with the right histogram (unequal bin widths).

Outlier(s) in Discrete & Categorical Data

Definition

A data point in a dataset is an **outlier** if it is ”far away” from ”most” of the data.

- Outliers are essentially extreme values of a dataset or sample.
- Outliers often occur due to catastrophic measurement errors:
  - Instrumentation terribly mis-calibrated
  - Instrumentation malfunctions during measurement
  - Person deliberately lying in a survey
  - Person deliberately exagerating measurements or counts
- However, not all outliers are due to errors:
  - House prices
  - Exam scores
- Histograms are not always effective in revealing outliers.
- Better visual and numerical methods for identifying outliers in Section 1.4
- Outliers are never really considered for continuous data.

Textbook Logistics for Section 1.2

- Difference(s) in Terminology:

TEXTBOOK TERMINOLOGY	SLIDES/OUTLINE TERMINOLOGY
Frequency Distribution	Frequency Table
Rectangle Height	Bin Height
Class(es)	Bin(s)
Class Interval(s)	Bin(s)
Class Width	Bin Width

- Difference(s) in Notation:

CONCEPT	TEXTBOOK NOTATION	SLIDES/OUTLINE NOTATION
Bin Widths	$3.0- < 3.5$	$[3.0, 3.5)$

Textbook Logistics for Section 1.2

- Ignore ”Stem-and-Leaf Displays” section (pg 13-15)
  - Stem-and-leaf displays were popular prior to the 1980’s.
  - Stem-and-leaf displays were useful when computers were text-only.
  - Stem-and-leaf displays will never be used in this course.
- Ignore ”Dotplots” section (pg 15-16)
  - Dotplots are effectively histograms but with stacked dots instead of bars.
  - Unfortunately, there’s no freedom in choosing appropriate bin widths.
  - Dotplots are only useful for small samples.
  - Dotplots will never be used in this course.
- Ignore ”Multivariate Data” section (pg 24)
  - Multivariate Data can be hard or impossible to visualize
  - Multivariate data is never considered in this course.

Fin

Fin.