VISUALIZING DATA: FREQ. TABLES & HISTOGRAMS [DEVORE 1.2]

- **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 **frequency** entails from counting the # data points of that category.

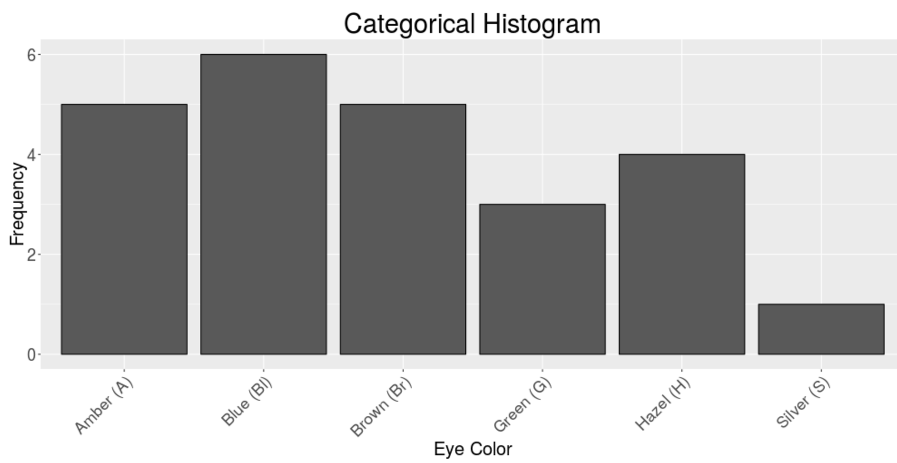
Compute the **total frequency**: $5 + 6 + 5 + 3 + 4 + 1 = \mathbf{24}$

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

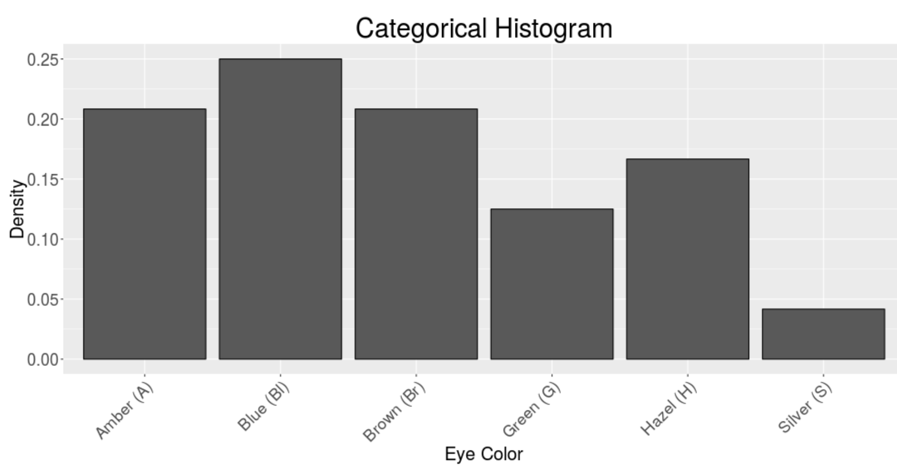
The **total relative frequency** should be very close to one (i.e. between 0.998 & 1.002)

Frequency tables can also be made for numerical data. (see EX 1.2.2 & EX 1.2.3 in this outline)

- **HISTOGRAMS FOR CATEGORICAL DATA:** From the above sample the resulting histogram is:



or using **density** (which is the same as **relative frequency** for categorical data) on the vertical axis



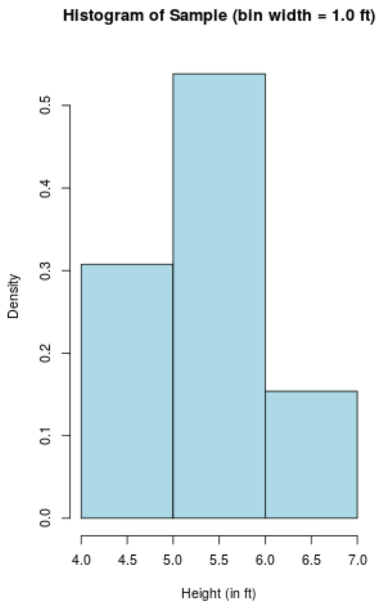
Finally, the vertical axis could be **percent**. (i.e. multiply the relative freq. or density by 100%)

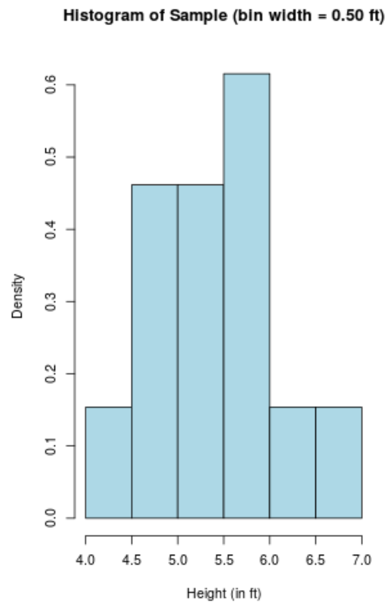
VISUALIZING DATA: MORE HISTOGRAMS [DEVORE 1.2]

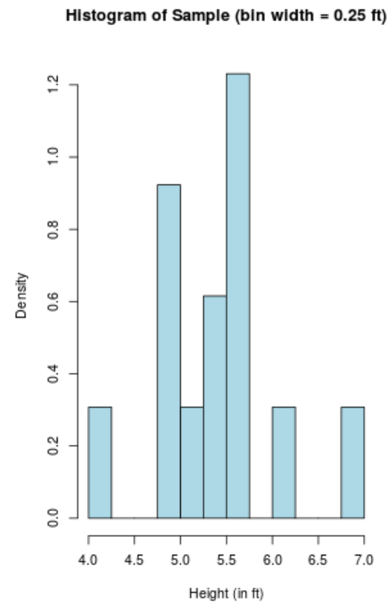
- **HISTOGRAMS FOR DISCRETE NUMERICAL DATA (EQUAL BIN WIDTHS):**

  Given a sample of heights (in ft): 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:







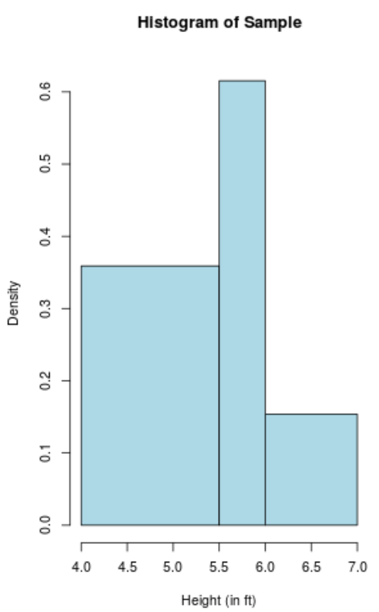
  Pick a bin width that avoids gaps (right figure) and "overlumping" (left figure).

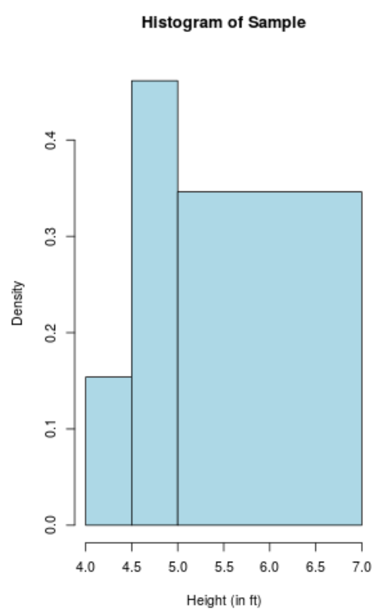
  For discrete numerical data, bin widths will be chosen a priori & the vertical axis is **always density** $= \dfrac{\text{relative frequency}}{\text{bin width}}$

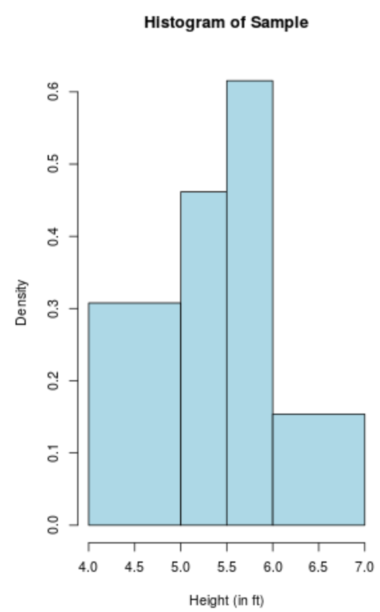
- **HISTOGRAMS FOR DISCRETE NUMERICAL DATA (UNEQUAL BIN WIDTHS):**

  Given a sample of heights (in ft): 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 discrete numerical data, bin widths will be chosen a priori & the vertical axis is **always density** $= \dfrac{\text{relative frequency}}{\text{bin width}}$

VISUALIZING DATA: MODALITY & SKEWNESS [DEVORE 1.2]

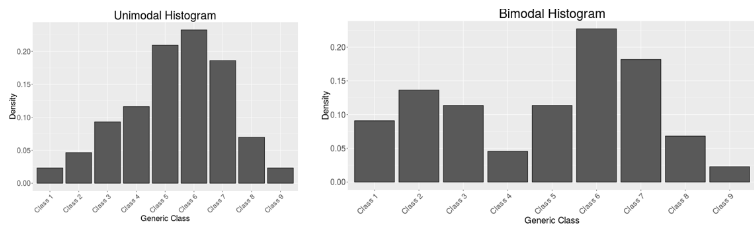
- **MODALITY OF DATA (DEFINITION):**

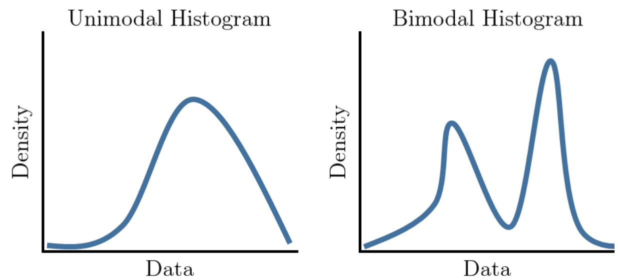
  A dataset/sample/population is **unimodal** if its histogram has exactly one peak.

  A dataset/sample/population is **bimodal** if its histogram has exactly two peaks.

  A dataset/sample/population is **multimodal** if its histogram has many peaks.

- **MODALITY OF DATA (EXAMPLES):** (see pgs 22-23 of textbook for examples of multimodal data)





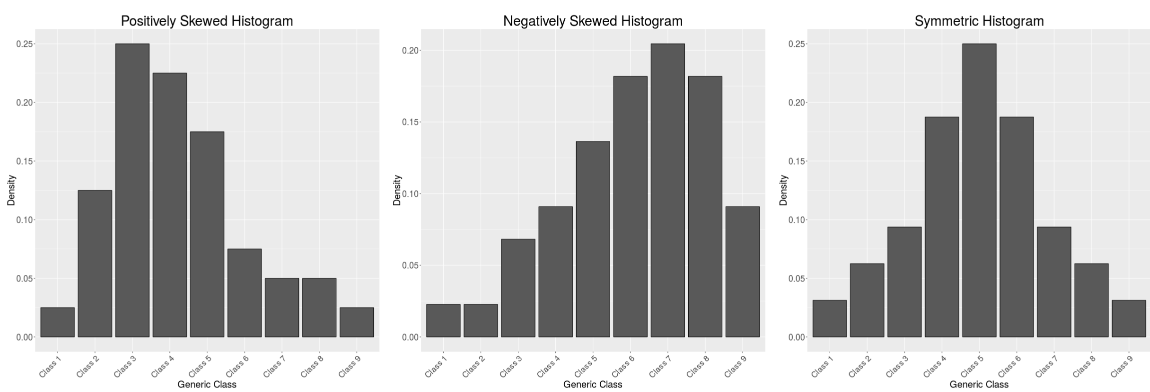
- **SKEWNESS OF DATA (DEFINITION):**

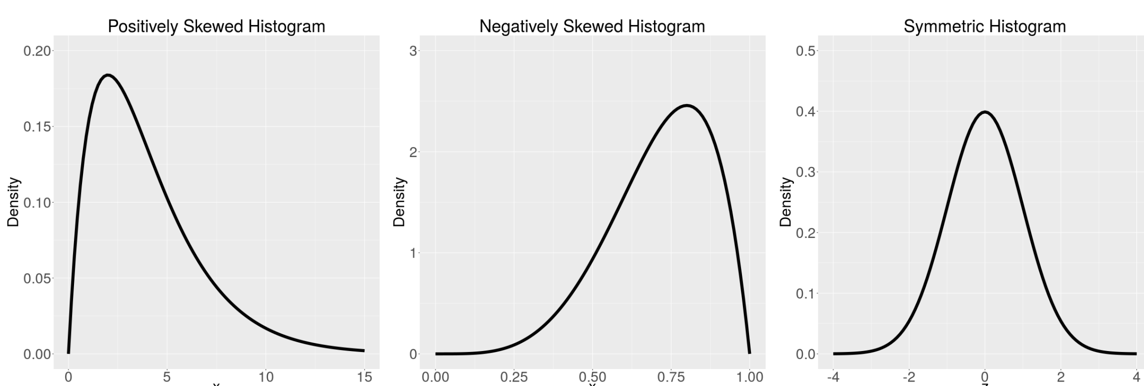
  A dataset/sample/population is **positively skewed** if its histogram has a long upper tail.

  A dataset/sample/population is **negatively skewed** if its histogram has a long lower tail.

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

- **SKEWNESS OF DATA (EXAMPLES):**





VISUALIZING DATA: OUTLIERS [DEVORE 1.2]

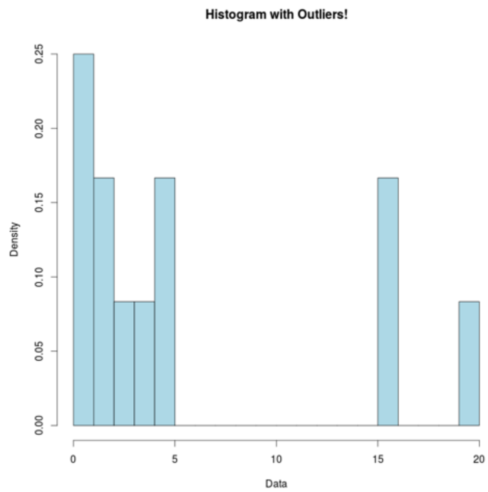
- **OUTLIER(S) IN DISCRETE NUMERICAL DATA (DEFINITION):**

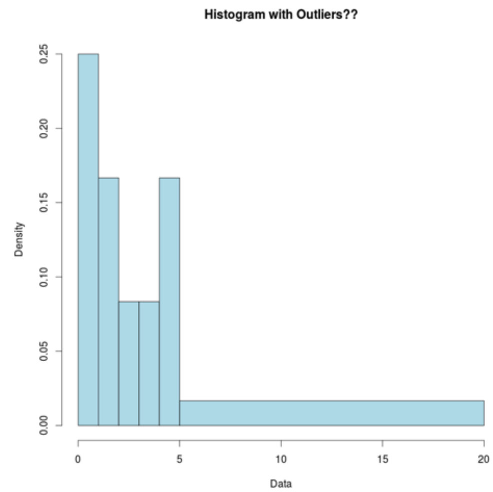
A data point in a dataset is an **outlier** if it is ”far away” from ”most” of the data.

- **OUTLIER(S) IN DISCRETE NUMERICAL DATA (EXAMPLE):** Consider the dataset:

1, 5, 2, 2, 1, 4, 1, 3, 20, 5, 16, 16

Then here are two histograms for the data:





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).

- **OUTLIERS (REMARKS):**
  - 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 rarely considered for categorical data.
  - Outliers are never considered for continuous data.

$\boxed{\textbf{EX 1.2.1:}}$ Given the following sample of college majors in a Linear Algebra course:

HIST, IE, CHE, ME, CS, IE, CHE, ME, MATH, MATH, IE, ME, ECO, ME, CHE, CS, MATH, ME, MATH, CS

a) Construct the frequency table for the data.

b) Construct a histogram for the data with **density** as the vertical axis.

c) What **proportion** of the students are MATH majors?

d) What **percent** of the students are Chemical Engineering (CHE) majors?

e) Which major has the most students?

EX 1.2.2: Given the following dataset of completed bowling games: 2, 7, 3, 1, 2, 3, 6, 0, 12, 4, 3, 2, 3, 3, 4, 5

Here, the characteristic observed for each bowling game is the **number of strikes**.

Moreover, suppose the **bin widths** (**class widths**) are **all equal** where each bin width equals **one**.

This choice of bin widths results in the following **bins** (**classes**):

0-<1, 1-<2, 2-<3, 3-<4, 4-<5, 5-<6, 6-<7, 7-<8, 8-<9, 9-<10, 10-<11, 11-<12, 12-<13

— OR IF YOU PREFER INTERVAL NOTATION —

$$[0, 1),\ [1, 2),\ [2, 3),\ [3, 4),\ [4, 5),\ [5, 6),\ [6, 7),\ [7, 8),\ [8, 9),\ [9, 10),\ [10, 11),\ [11, 12),\ [12, 13)$$

a) Construct the frequency table for the data.

b) Construct a histogram for the data with **density** as the vertical axis.

c) Describe the modality & skewness of the data.

d) Looking at the histogram, does there appear to be any outliers? If so, identify them.

EX 1.2.3: Given the following dataset of completed bowling games: 2, 7, 3, 1, 2, 3, 6, 0, 12, 4, 3, 2, 3, 3, 4, 5

Here, the characteristic observed for each bowling game is the **number of strikes**.

Moreover, suppose the **bin widths** (**class widths**) are **unequal** in way which leads to the following **bins** (**classes**):

0-<1, 1-<3, 3-<6, 6-<13

— OR IF YOU PREFER INTERVAL NOTATION —

$$[0, 1),\ [1, 3),\ [3, 6),\ [6, 13)$$

a) Construct the frequency table for the data.

b) Construct a histogram for the data with **density** as the vertical axis.

c) What proportion of the bowling games have at most five strikes?

d) What percent of the bowling games have at least six strikes?