Unit 1: Exploratory Data Analysis

(Ch 1.1, 1.3, 1.10-1.13, 2.4.3, 2.5)

Learning Objectives

At the end of this unit, students should be able to:

1. Define a population, sample, sample frame, variable of interest and identify these concepts in particular examples.
2. Describe what is meant by “statistical inference”.
3. Define, calculate, and interpret three measures of center.
4. Describe situations in which one measure might be better than another.
5. Define, calculate, and interpret three measures of variation.
6. Define, calculate, and interpret quantiles and percentiles.
7. Produce and interpret histograms; identify whether the distribution of a variable is unimodal or multimodal, based on a histogram.
8. Produce and interpret boxplots, and scatterplots.
9. Perform meaningful exploratory data analysis in R.

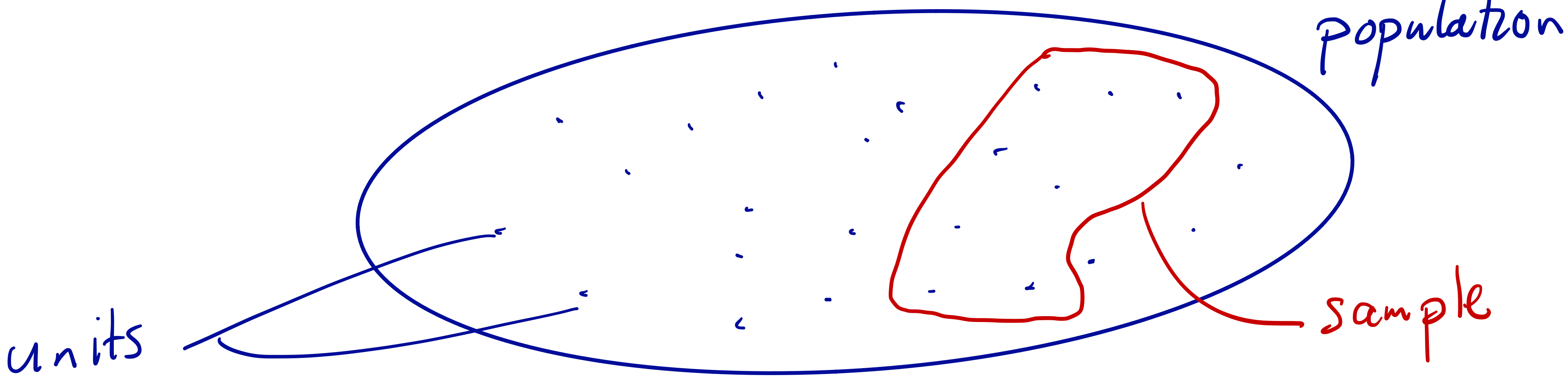
Populations and Samples

Statisticians hope to learn about some *characteristic/variable* in a *population*. But we often can't see the whole population; so, we investigate a *sample.*

Definition: A *population* is a collection of units (units can be people, widgets, servings of food, kittens, songs, Tweets, etc.)

Definition: A *sample* is a subset of the population.

Definition: A *characteristic/variable* of interest (VoI) is something to be measured for each unit.



Example: CU might want to study the average GPA of juniors who are engineering majors at CU. In this case, the Population is..? Reasonable Sample? VoI?

Populations and Samples: Examples

- Insurance company surveying damage in a particular town after hurricane…

Pop. : Addresses in this town

sample : Addresses actually observed (say, n of them)

VoI : Yes damage $\boxed{\text{or}}$ no damage

(sample frame??)

- Testing the strength of a picture hanger made by Kramerica Industries…

Populations and Samples

Statisticians learn about a characteristic in a population by studying a sample.

A major component of this course is to figure out how they make the jump from sample to population—Statistical Inference!

Exploratory Data Analysis (EDA)/Descriptive Statistics

Before we learn about inference, we're first going to learn how to explore data. This is helpful for summarizing, recognizing patterns, etc.

There are two main types of explorations: **numerical** and **graphical.**

Numerical Summaries: Sample Statistics

The calculation and interpretation of certain summarizing numbers can help us gain an understanding of the data.

statistic = value calc. from a sample

These sample numerical summaries are called *sample statistics*.

Sample Statistics: Measures of Centrality

Summarizing the “center” of the sample data is a popular and important characteristic of a set of numbers. The goal here is to capture something like the “typical” unit with respect to the VoI.

3 popular types of center:

- mean
- median
- mode

The Sample Mean

For a given set of numbers $\boxed{x_1,\ x_2,.\ .\ .,\ x_n,}$ the most familiar measure of the center is the *mean* (arithmetic average).

Sample mean x of observations x_1, x_2,. . ., x_n:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Advantages? Easy to calc.; uses all information (equally weighted)

Disadvantages? Not robust to outliers

The Sample Median

Median: Middle value when observations are ordered smallest to largest.

The Sample Median

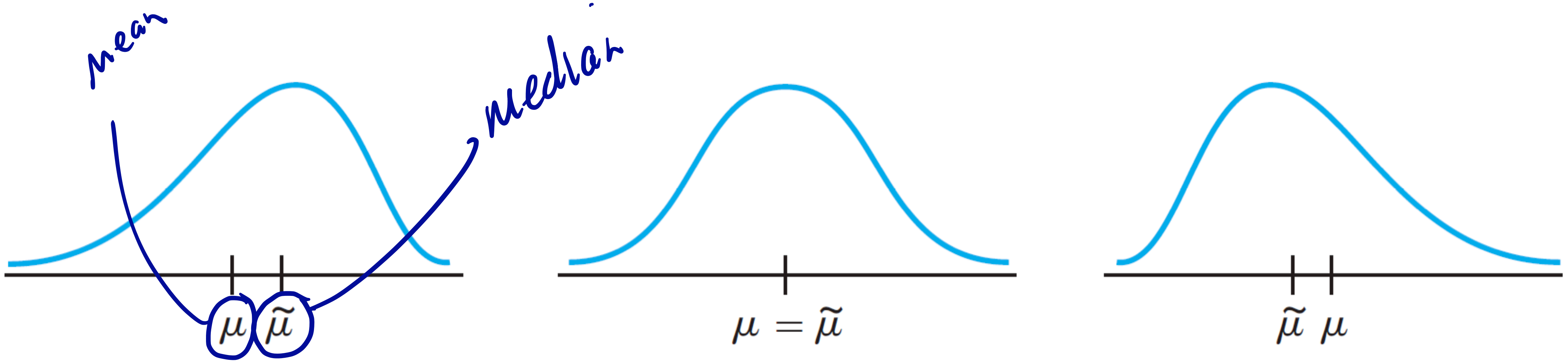
Median: Middle value when observations are ordered smallest to largest.

To calculate: Order the *n* observations smallest to largest (repeated values included and find the middle one.

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ ordered value}\textit{s} \end{cases}$$

The Mean vs. the Median

The population mean and median will not generally be identical.



(a) Negative skew (b) Symmetric (c) Positive skew

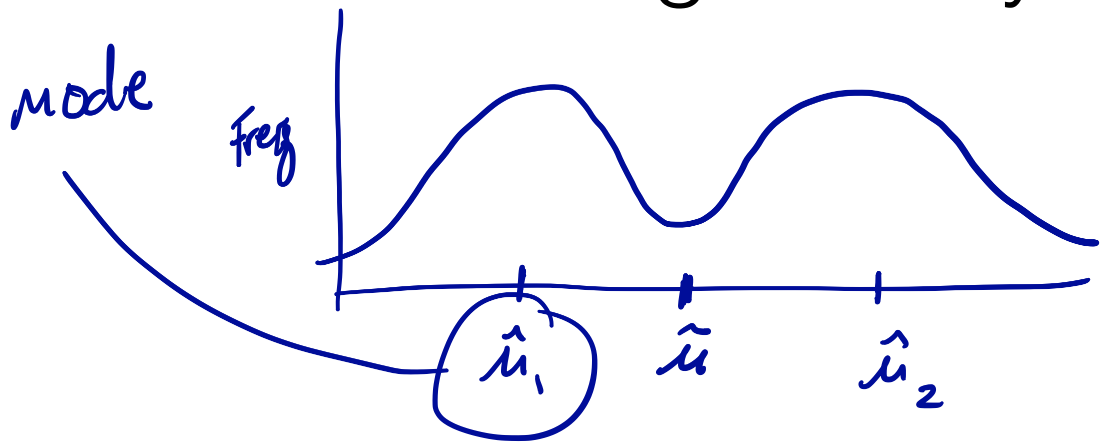
Three different shapes for a population distribution

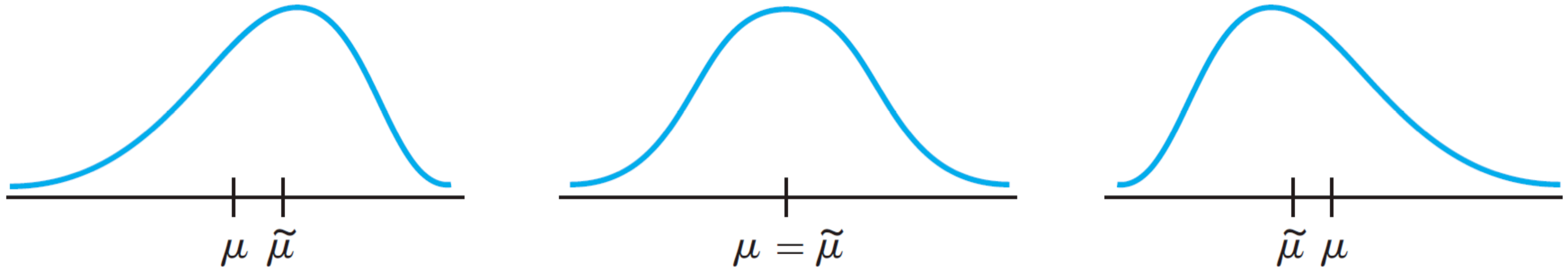
NZNO

Thurs: 9-11am, $\boxed{\text{2-4pm}}$

The Mean vs. the Median

The population mean and median will not generally be identical.





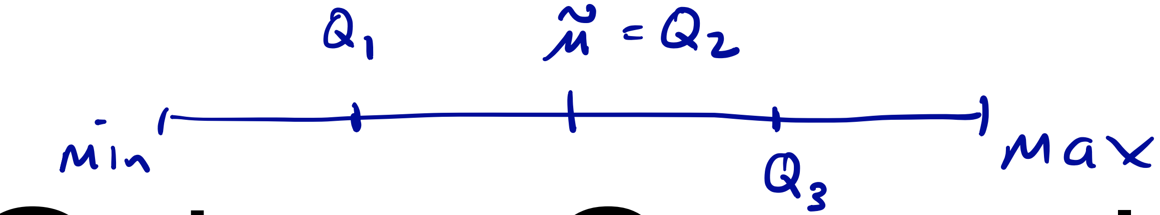
(a) Negative skew

(b) Symmetric

(c) Positive skew

Three different shapes for a population distribution

Which population characteristic is most important?



Other Sample Measures

Mode: most frequently occurring value.

Quartiles: divide the data set into **four** equal parts (how is this calculated?)

Quantile

Percentiles: A data set can be even more finely divided. What does "percentile" mean?

Example calculations of the median and quartiles:

Data: 34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.

$\tilde{x} = 26$

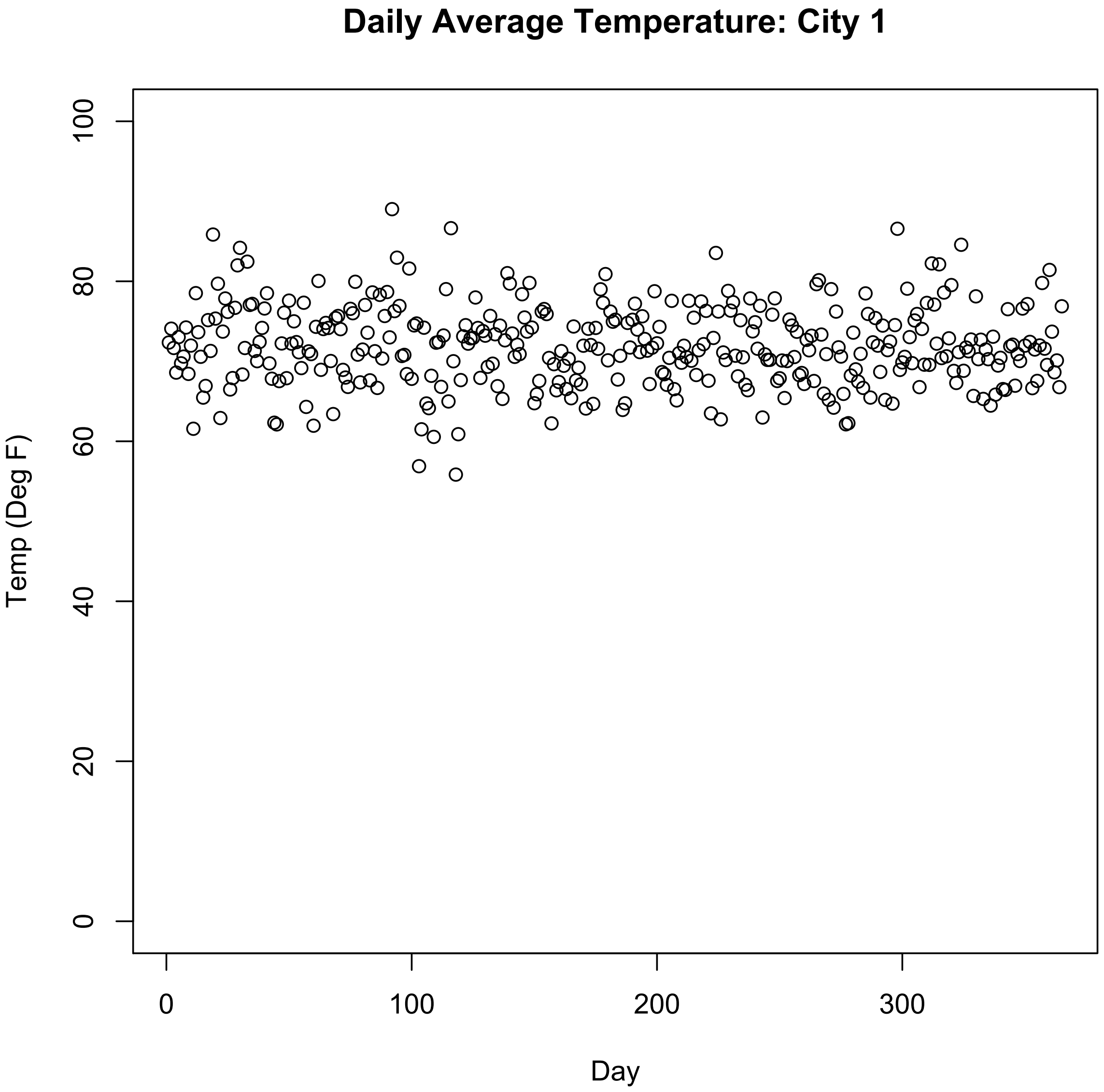
Variability

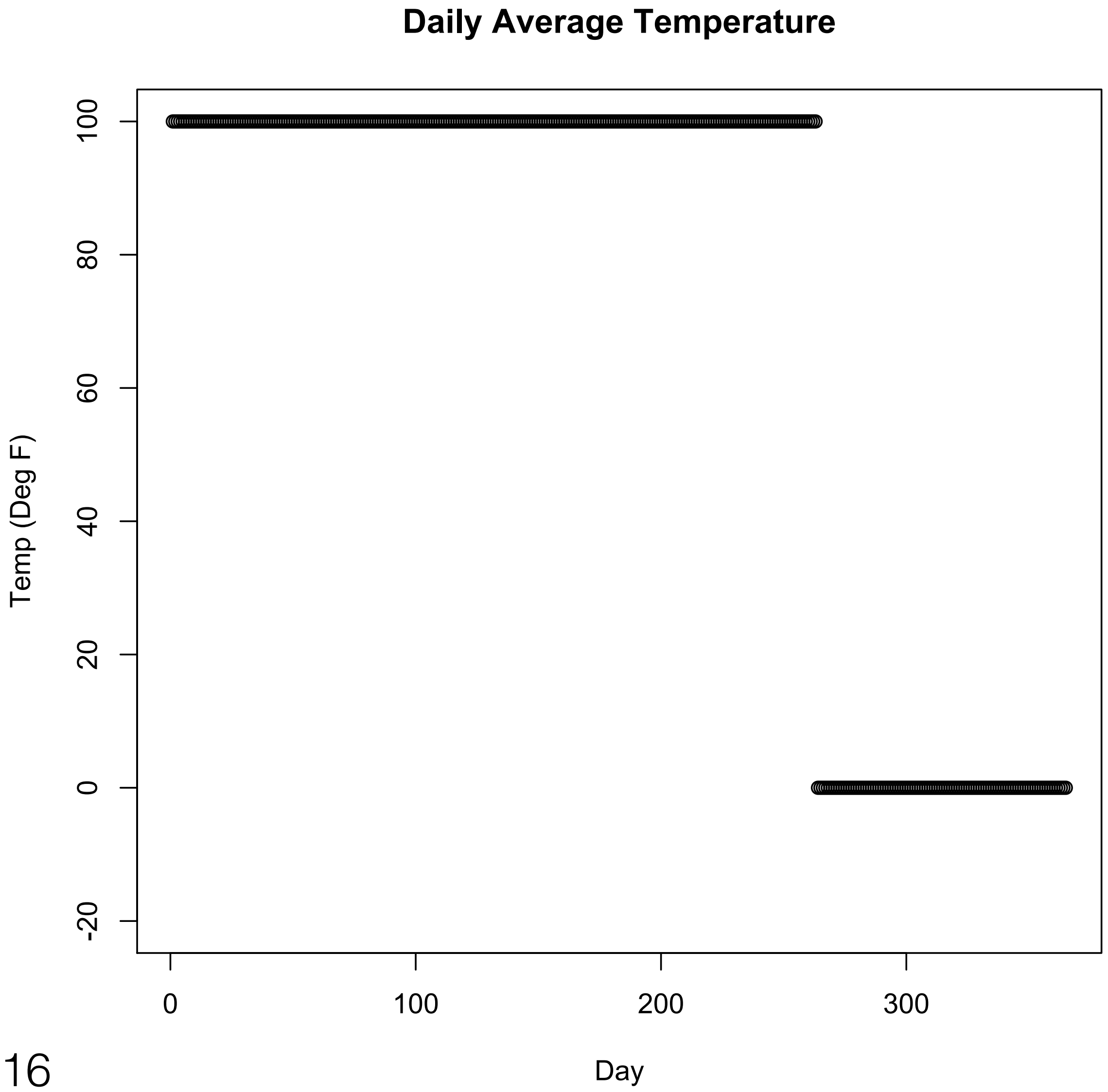
So far, we’ve learned techniques for visualizing our data and measures of center. What about the *spread* of the data?

Example: A tail of two cities.

Variability

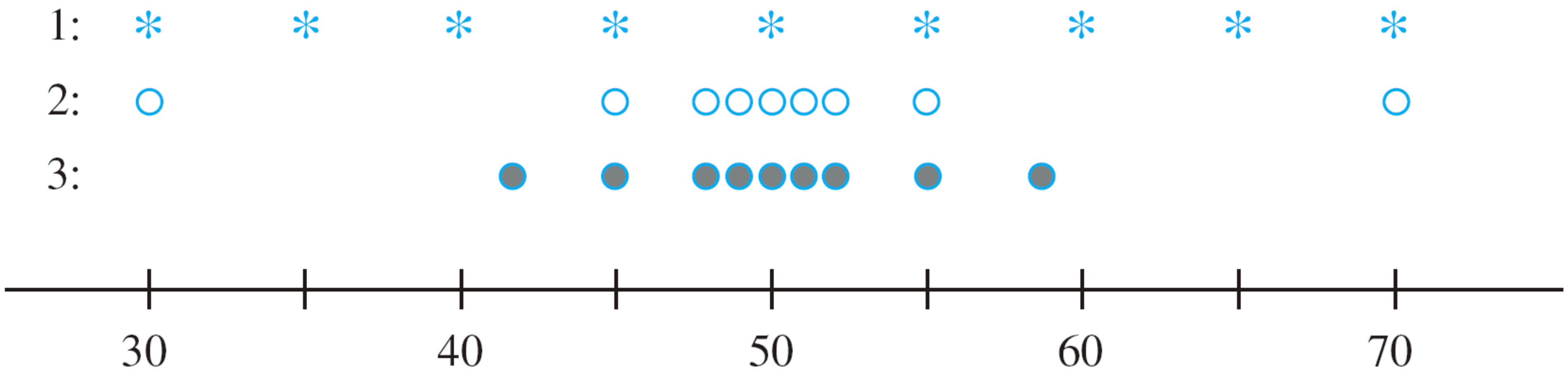
So far, we've learned techniques for visualizing our data and measures of center. What about the *spread* of the data?





Variability

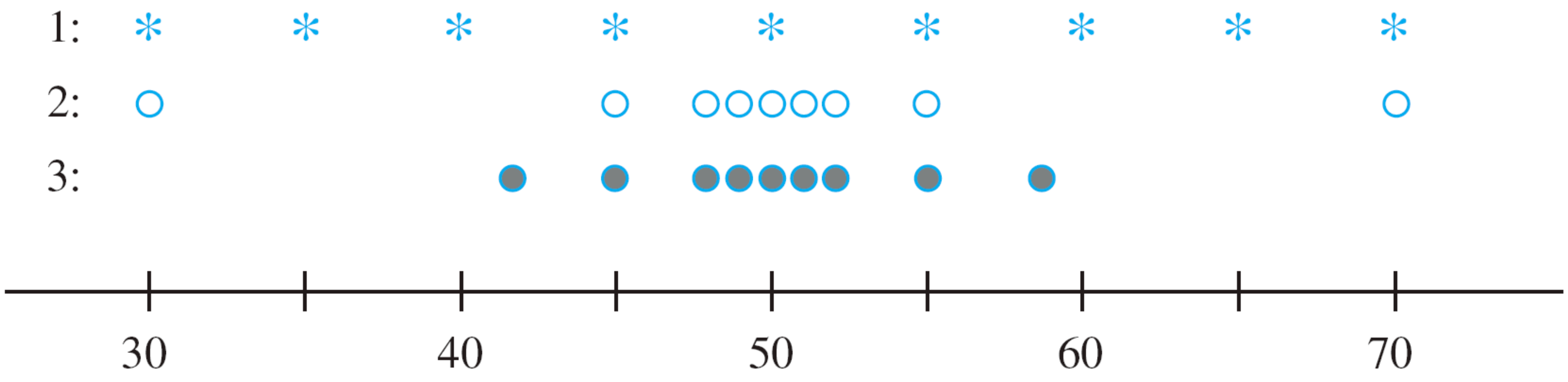
Simplest measure of variability: The range.



Samples with identical measures of center but different amounts of variability

Variability

Simplest measure of variability: The range. = Max-min



Samples with identical measures of center but different amounts of variability

What are disadvantages of the range?

$x_1, \dots, x_n$

Variability

Can we combine the deviations into a single quantity by finding the average deviation?

$x_i - \bar{x}$, $i = 1, \dots, n$

A more robust measure of variation takes into account deviations from the mean:

$$\left.\begin{array}{l} X_1, \ldots, X_n \longrightarrow S_X^2 \\ Y_1, \ldots, Y_n \longrightarrow S_Y^2 \end{array}\right\}$$

Variability

The sample variance, denoted by $\underline{\ s^2\ }$, is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (\text{units}^2)$$

The sample standard deviation, denoted by s, is the (positive) square root of the variance:

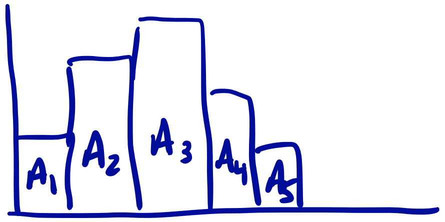
$$s = \sqrt{s^2} \qquad (\text{units})$$

Note that $\underline{\ s^2\ }$ and $\underline{\ s\ }$ are both nonnegative. The unit for $\underline{\ s\ }$ is the same as the unit for each of the $\underline{\ x_i\ }$.

Example: Calculation of the SD.

Data (units in dollars): 2,4,3,5,6,4. $= X$

$$s_X^2 = 2 \Rightarrow s_X = \sqrt{2} \approx 1.41$$



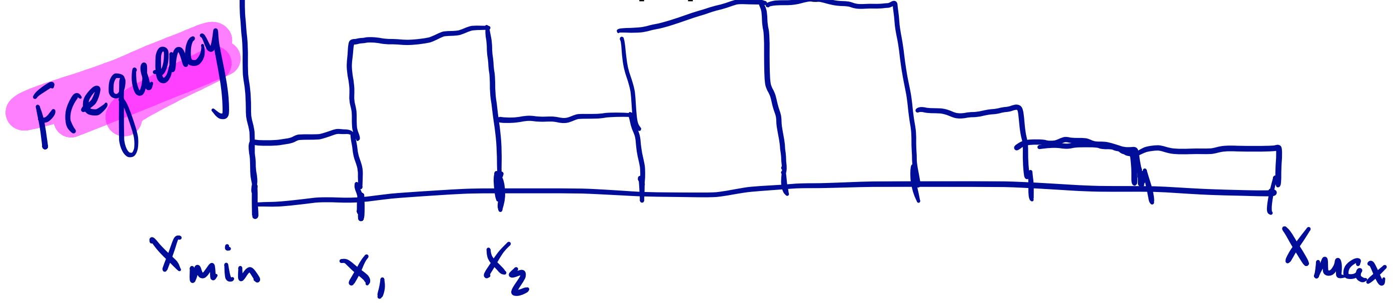
$\sum_{i=1}^{5} A_i = 1$

Graphics: Histograms

A **histogram** is a graphical representation of the distribution of numerical data.

Construct a histogram:

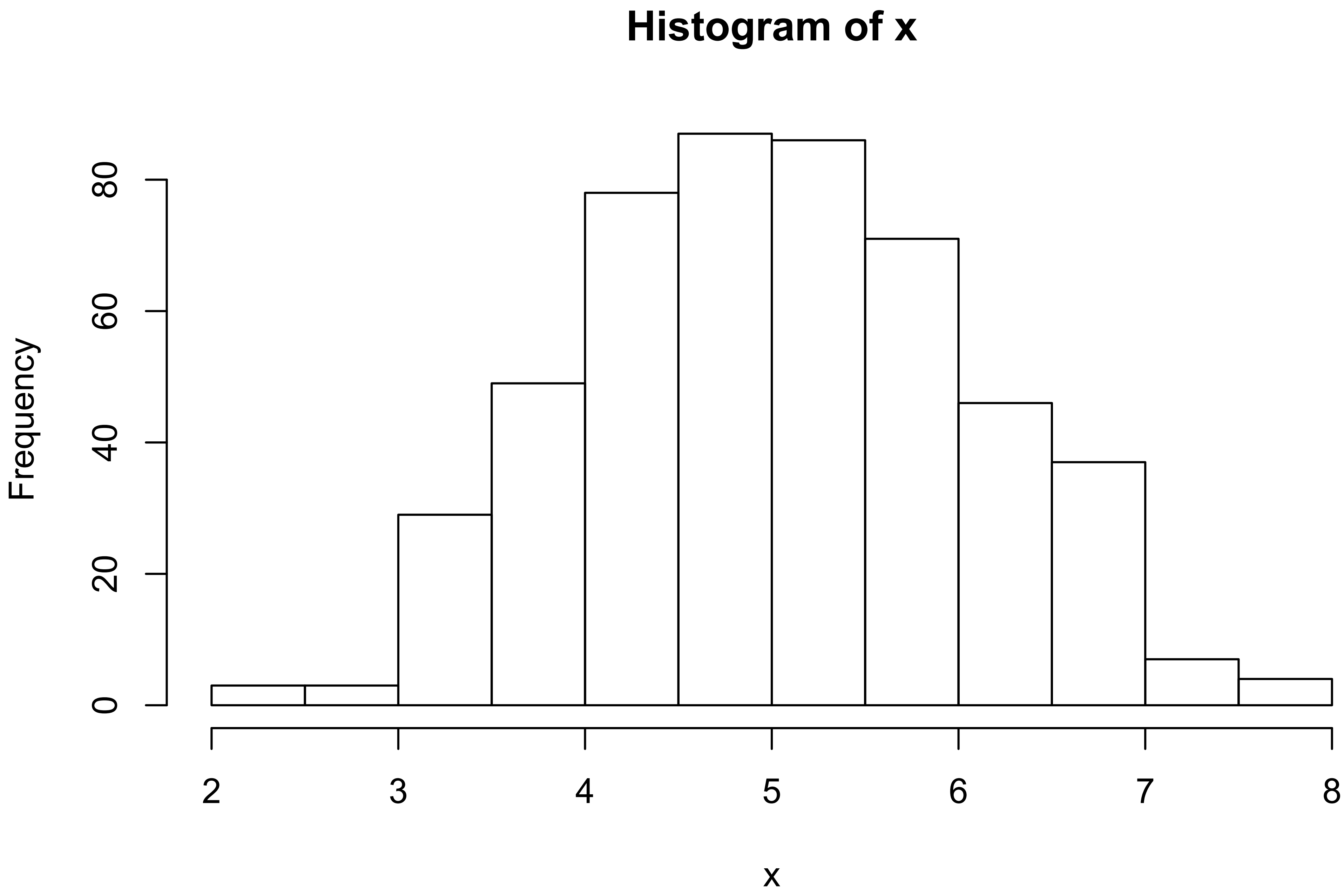
“Bin” the measured values of the VoI. (The bins are usually consecutive, non-overlapping, and are usually equal size.)



Frequency histogram: count how many values fall into each bin/ interval and draw accordingly.

Density histogram: count how many values fall into each bin, and adjust the height such that the sum of the area of each bin equals 1.

Graphics: Histograms



Histograms: Example

Charity is a big business in the United States. The Web site charitynavigator.com gives information on roughly 5500 charitable organizations.

Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities.

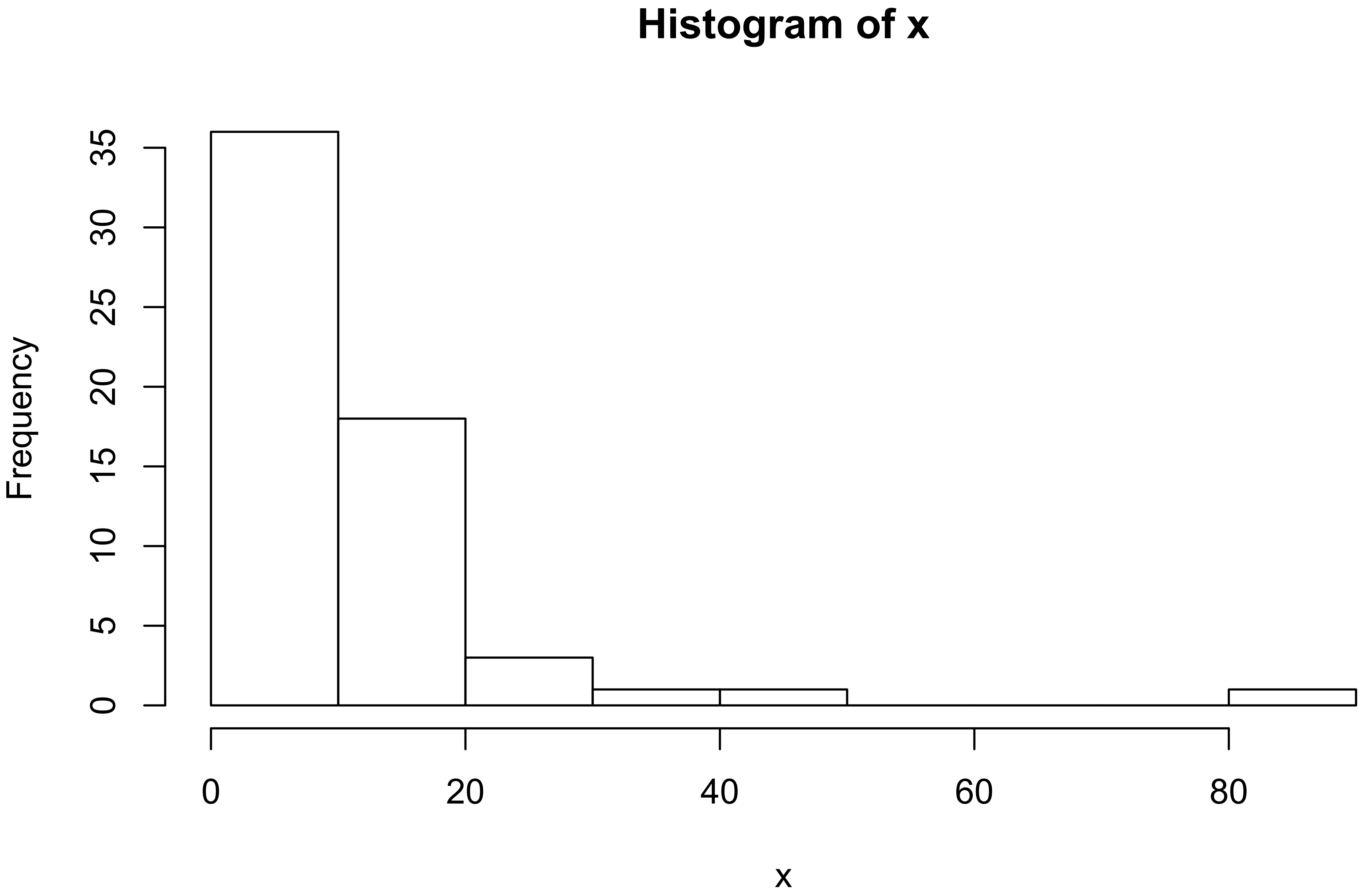
Histograms: Example

Here are the data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Histograms: Example

hist(x)



Graphics: Histograms

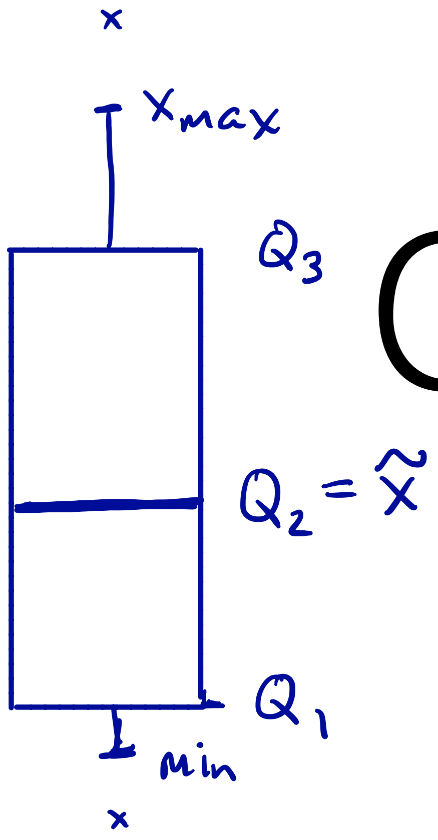
Histograms come in a variety of shapes.

Unimodal histogram: single peak

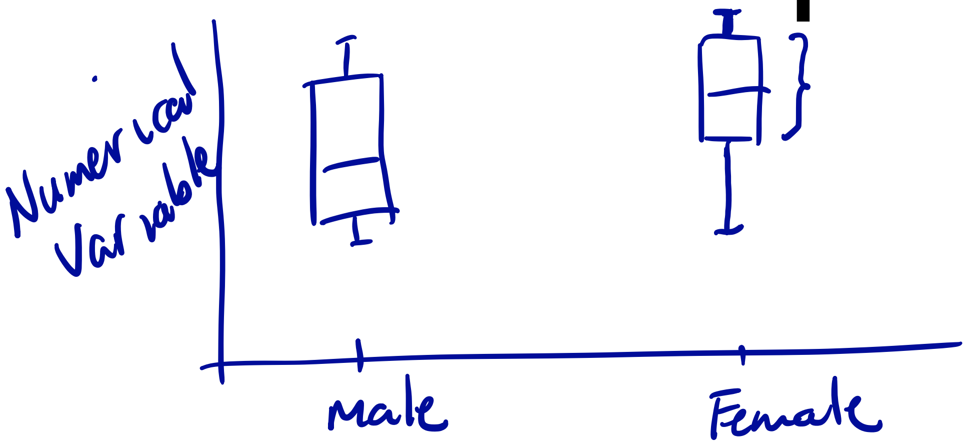
Bimodal histogram: two different peaks. Can occur when the data set consists of observations on two quite different kinds of individuals or objects.

Multimodal histogram: many different peaks

Other types: Symmetric histograms, Positively skewed histograms, Negatively skewed histograms



Graphics: Boxplots



A **boxplot** is a convenient way of graphically depicting groups of numerical data through the five number summary: minimum, first quartile, median, third quartile, and maximum.

Example: Drawing a boxplot by hand.

Classwork

Let $\bar{x} = \frac{1}{n}\sum x_i$

Answer the following questions for a sample data set with *n* values. What happens to the mean when:

$$\frac{1}{n}\sum(x_i - 3) = \frac{1}{n}\sum x_i - \frac{1}{n}\cdot n\cdot 3 = \bar{x} - 3$$

1. 3 is subtracted from every value in the data set?

2. Every value in the data set is multiplied by 3? $\frac{1}{n}\sum 3x_i = 3\cdot\frac{1}{n}\sum x_i = 3\bar{x}$

3. Every value in the data set is divided by 3? $\frac{1}{n}\sum \frac{x_i}{3} = \frac{1}{3}\bar{x}$

4. 3 is subtracted from the minimum value and 3 is added to the maximum value in the data set? $\frac{1}{n}\left[x_{(1)} - 3 + \sum_{i=2}^{n-1} x_{(i)} + x_{(n)} + 3\right] = \frac{1}{n}\sum x_{(i)} = \bar{x}$

Answer the questions above for the standard deviation.