Basic Data Analysis

Definition: in statistics, the term **data** refers to a set containing values of a numerical or qualitative variable.

Examples of data sets:

(a) the high temperatures for each day in New York City during the year 2015,

(b) the fasting blood glucose level of a group of 30 individuals on particular day,

(c) the running times of a computer program upon each execution of that program,

(d) the final GPAs of NYCCT graduates in Spring 2016,

(e) the times that it takes an airplane to travel a particular route in the last 30 days along with the average head wind speed each day,

(f) the list of political parties of the presidential candidate each voter selected in the 2012 election,

Exercise: For each example above, describe the type of data, for example, is it quantitative or qualitative? What are the possible (realistic) values for the data? What are appropriate units, if any?

Note: repetitions in a data set should not be ignored, as each element of a data set is implicitly associated with a particular observation. For example, if a high temperature of 80 degrees is observed on a day in May and on a day in June, the value 80 should be counted for both days (i.e., it should appear twice in the data set). This does not contradict the example from the probability review: $\{1, 2, 3, 4, 5, 6\} = \{1, 1, 2, 2, 2, 3, 4, 5, 6, 5, 6\}$ (why not?).

Measures of Central Tendency, Dispersion, and Skewness

In this section, let the general data set be: $\{x_1, x_2, x_3, \ldots, x_n\}$, in which $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

Definitions: (Measures of Central Tendency)

(i) The **mean** or **average** of a data set is denoted $\bar{x}$ and is given by the formula

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

(ii) The **median** of a data set is the number with the property that when the data set is ordered, half of the values lie below this number (and half of the values lie above it). If the data set has an even number of entries, it is customary to take the median as the average of the two entries in the middle of the ordered data set.

(iii) The **mode** of a data set is the data entry that occurs with the greatest frequency. A data set may have no mode, one mode, or more than one mode. If there are two modes, the data is said to be *bimodal*.

Exercises:

1. The following data represent the round-trip flight prices for a particular route as of 8:00 am on each day of a particular week: $\{288, 497, 497, 327, 332, 588, 291\}$. Answer the following:
   (a) Find the mean of this data set.
   (b) Find the median of this data set.
   (c) Find the mode of this data set.
   (d) Compare the mean, median, and mode. Which is the smallest? Which is the largest? Which do you believe best represents this data set, and why?
2. The following data represent the age of students in a particular section of a freshman calculus course: $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$. Answer the following:
   (a) Find the mean of this data set.
   (b) Find the median of this data set.
   (c) Find the mode of this data set.
   (d) Compare the mean, median, and mode. Which is the smallest? Which is the largest? Which do you believe best represents this data set, and why?

Definitions: (Measures of Dispersion)

(i) The **range** of a data set is

$$\text{Range} \;=\; \text{Maximum data entry} \;-\; \text{Minimum data entry}$$

(ii) The **variance** of a data set is the average of the sum of the squared deviation of each entry from the mean:

$$\text{Var} \;= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(iii) The **standard deviation** of a data set is the square root of the variance. That is,

$$\text{Standard Deviation} \;= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

1. The following data represent the round-trip flight prices for a particular route as of 8:00 am on each day of a particular week: $\{288, 497, 497, 327, 332, 588, 291\}$. Answer the following:

(a) Find the range of this data set.

(b) Find the variance of this data set.

(c) Find the standard deviation of this data set.

(d) What are the units of the range, variance, and standard deviation of this data set?

2. The following data represent the age of students in a particular section of a freshman calculus course: $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$. Answer the following:

(a) Find the range of this data set.

(b) Find the variance of this data set.

(c) Find the standard deviation of this data set.

(d) What are the units of the range, variance, and standard deviation of this data set?

Definitions:

(i) An **outlier** of a data set is a value that is far removed from most other values in the data set. There is no standard formula or definition for determining which value(s) is an outlier. An example of a formula will be given later.

(ii) A data set is said to be **skewed to the right** or has a **positive skew** if it contains several entries that are considerably larger than most of the other entries in the data set.

(iii) A data set is said to be **skewed to the left** or has a **negative skew** if it contains several entries that are considerably less than most of the other entries in the data set.

The definitions above are not precise. In general, skewness measures asymmetry in a data set, and there exist various formulas to calculate skewness.

Exercises:

1. The following data represent the round-trip flight prices for a particular route as of 8:00 am on each day of a particular week: $\{288, 497, 497, 327, 332, 588, 291\}$. Answer the following:

(a) Find potential outliers of this data set.

(b) Do you think this data is skewed to the right, skewed to the left, or not skewed?

2. The following data represent the age of students in a particular section of a freshman calculus course: $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$. Answer the following:

(a) Find potential outliers of this data set.

(b) Do you think this data is skewed to the right, skewed to the left, or not skewed?

Graphical Representation of Data

In this section, let the general data set again be: $\{x_1, x_2, x_3, \ldots, x_n\}$, in which $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

Definitions: (Measures of Position)

(i) The **quartiles** of data set are the (three) values that partition the ordered data set into *four* equally sized sets.

(ii) The **first quartile** that can be denoted as Q_1 is the quartile such that approximately 1/4 of the data values are less than Q_1 (and thus, approximately 3/4 of the values are greater than Q_1).

(iii) The **second quartile** that can be denoted as Q_2 is the quartile such that approximately 1/2 of the data values are less than Q_2 (and thus, approximately 1/2 of the values are greater than Q_2). What is another name for Q_2?

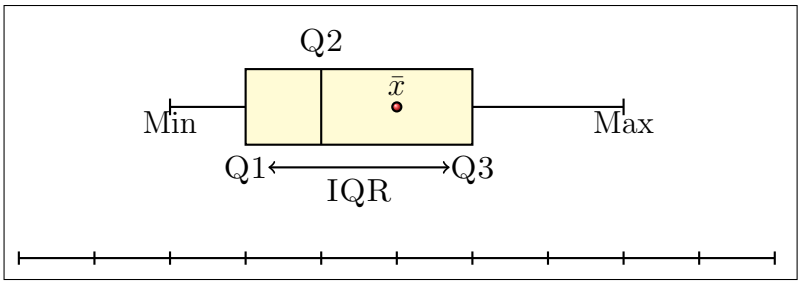
(iv) The **third quartile** that can be denoted as Q_3 is the quartile such that approximately 3/4 of the data values are less than Q_3 (and thus, approximately 1/4 of the values are greater than Q_3).

(v) The **interquartile range** (IQR) is a measure of dispersion given by

$$\text{IQR} = Q_3 - Q_1$$

Note that this is the range of the middle half of the values of a data set.

Graphing: Box-and-Whisker Plot
A box-and-whisker plot generally displays the minimum data entry, Q_1, Q_2, Q_3, and the maximum data entry of a data set. To construct a box-and-whisker plot, draw a number line representing the data entries and the box-and-whisker diagram above the number line as follows:



An *outlier* in the data, if any, is usually not included in the whisker part. If there is an outlier, it is represented by a dot (outside of the whisker). A common formula for determining outliers is: x_i is an outlier if it is less than $Q_1 - 1.5 \times \text{IQR}$ or if it is greater than $Q_3 + 1.5 \times \text{IQR}$

Exercises:

1. The following data represent the age of students in a particular section of a freshman calculus course: $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$. Answer the following:
   (a) Find the quartiles, Q_1, Q_2, Q_3 of this data.
   (b) Find the interquartile range of this data set.
   (c) Determine if this data has any outliers.
   (d) Graphically represent this data in a box and whisker plot.
2. The house prices for houses on the market in a given town are (in units of $10,000$) $\{19, 31, 11, 9, 25, 26, 21, 28, 22, 14, 8, 17, 35, 16, 10, 15, 19, 25, 45, 7\}$. Answer the following:
   (a) Find the mean, median, and mode of this data set. Which do you think best represents the data and why?
   (b) Find the range of the data set.
   (c) Find the variance of the data set.
   (d) Find the standard deviation of the data set.
   (e) Find the quartiles, Q_1, Q_2, Q_3 of this data.
   (f) Find the interquartile range of this data set.
   (g) Determine if this data has any outliers.
   (h) Graphically represent this data in a box and whisker plot.

Graphing: A Stem-and-Leaf Plot
A stem-and-leaf plot generally displays all data entries, separating each number into a *stem* and a *leaf*. The stems are written in increasing order in the leftmost column, followed by a vertical bar, and leaves are listed in increasing order to the right of each stem, one leaf per each data entry. For example, if most data entries are between 0 and 100, the leaves may be the units digit and the stems the hundreds and tens digits (e.g. in 35, the stem could be 3 and the leaf could be 5). Thus, a stem-and-leaf plot contains all original data entries.

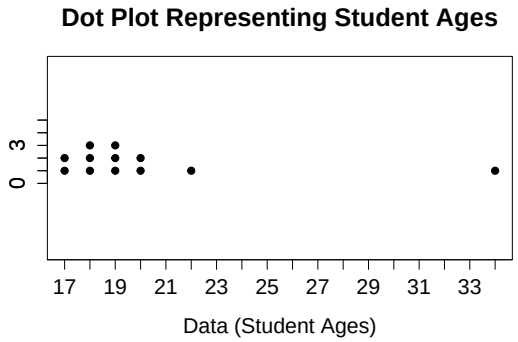
Example: Represent $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$ in a stem-and-leaf plot.

1	7	7	8	8	8	9	9	9
2	0	0	2					
3	4							

Key: $1|7 = 17$

Graphing: A Dot Plot
A dot plot displays all data entries as a dot for each entry above the appropriate value in a number line.

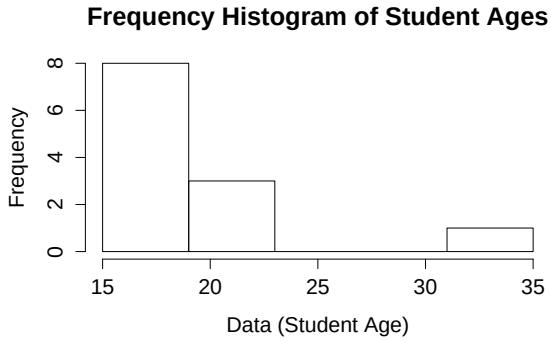
Example: Represent $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$ in a dot plot.



Graphing: A Frequency Histogram

A frequency histogram is a graphical representation of data in which data entries are placed into *classes* (intervals) represented typically on the horizontal axis, and a vertical bar is drawn for each class with height equal to the *frequency* (number) of entries in that class. The term bar graph can also refer to a frequency histogram.

Example: Represent $\{19, 18, 18, 17, 19, 20, 22, 19, 18, 17, 34, 20\}$ in a frequency histogram.



Exercises:

1. The house prices for houses on the market in a given town are (in units of $10,000$) $\{19, 31, 11, 9, 25, 26, 21, 28, 22, 14, 8, 17, 35, 16, 10, 15, 19, 25, 45, 7\}$. Answer the following:
   (a) Represent this data in a stem-and-leaf plot.
   (b) Represent this data in a dot plot.
   (c) Represent this data in a frequency histogram.
   (d) From which of the plots above can the mean be determined?
   (e) From which of the plots above can the median be determined?
   (f) From which of the plots above can the mode be determined?
   (g) From which of the plots above can the range be determined?

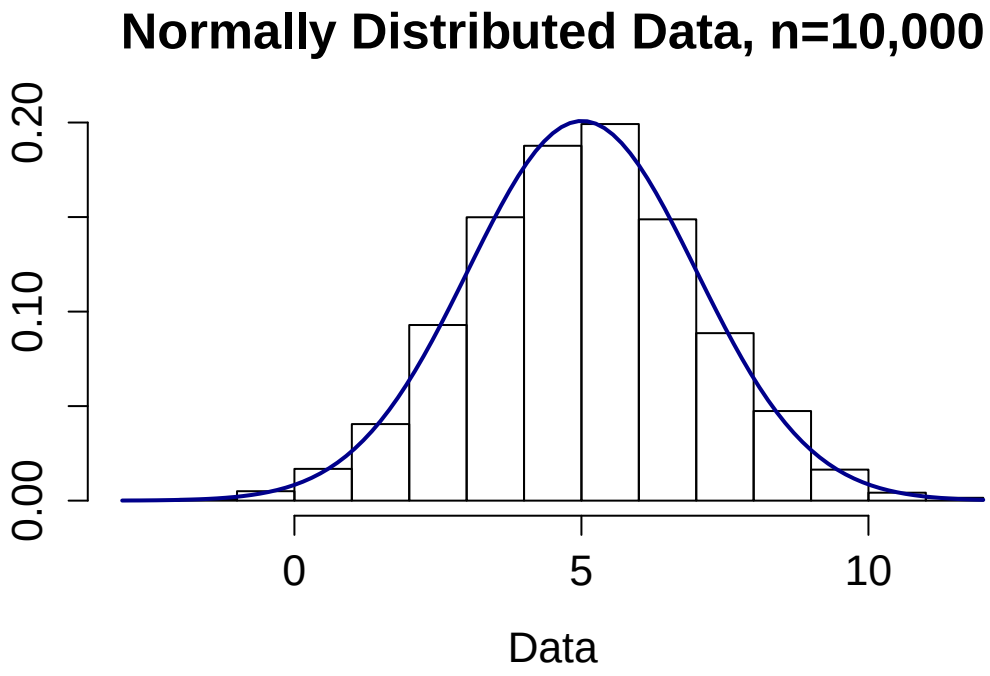
Normally Distributed Data

Definition: a data set is said to be approximately "normally distributed" if the tops of the bars of the histogram with the vertical axis entries scaled so that the total area of the bars is 1 approximately follows the bell-shaped curve given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note that this definition is imprecise. Here μ represents the mean of the data and σ the standard deviation of the data.

Example:
Data represented by the following histogram with appropriately scaled vertical axis is approximately normally distributed. The bell-shaped normal curve is drawn for reference.



Properties of Normally Distributed Data:

(i) The mean and median are (approximately) equal.

(ii) The data set is (approximately) symmetric.

(iii) About 68% of the data lies in the interval $(\mu - \sigma, \mu + \sigma)$ (within 1 standard deviation of the mean).

(iv) About 95% of the data lies in the interval $(\mu-2\sigma, \mu+2\sigma)$ (within 2 standard deviations of the mean).

(v) About 99.7% of the data lies in the interval $(\mu-3\sigma, \mu+3\sigma)$ (within 3 standard deviations of the mean).

Definition: The **z-score** is a number that represents how many standard deviations a data entry is away from the mean. A z-score for a data entry x is computed using the formula

$$z = \frac{x - \mu}{\sigma}$$

A positive z-score indicates that x is greater than the mean, and a negative z-score indicates that x is less than the mean.

Exercises: Suppose that x is data entry from a normally distributed data set with $\mu = 10$ and $\sigma = 3$.

1. Find the z-score of x if $x = 15$.
2. Find the z-score of x if $x = 16$.
3. Find the z-score of x if $x = 10$.
4. Find the z-score of x if $x = 4$.
5. Find the z-score of x if $x = 1$.
6. Find x with a z-score of 1.5.
7. Find x with a z-score of -1.5.
8. Find x with a z-score of 1.
9. Approximately what percentage of data entries are greater than $x = 13$?
10. Approximately what percentage of data entries are less than $x = 4$?
11. Approximately what percentage of data entries are between $x = 7$ and $x = 13$?
12. Approximately what percentage of data entries are less than $x = 5$ (a calculator is needed)?
13. Approximately what percentage of data entries are greater than $x = 8$ (a calculator is needed)?
14. Approximately what percentage of data entries are between $x = 5$ and $x = 8$ (a calculator is needed)?

Exercises: Which of the following histograms could represent normally distributed data? Right skewed data? Left skewed data? If possible and applicable, approximate the mean, median, and standard deviation of the data in each histogram.