Chapter 1: Overview and Descriptive Statistics

1.1 Populations, Samples, and Processes

1.2 Pictorial and Tabular Methods in Descriptive Statistics

1.3 Measures of Location

1.4 Measures of Variability

1.1 Populations, Samples, and Processes

- Basic concepts

- Branches of Statistics

- Probability and Statistics

- Data collection

Basic Concepts

- **Unit**: A single entity whose characteristics are of interest.
- **Population**: The entire group of units under study.
  - e.g: The OU student body; Pounds of fish caught by all persons in a fishing derby.
  - **Census**: When desired information is available for all units in the population. Too costly or even infeasible when the population is large.
- **Sample**: A subset of the population selected in a prescribed manner.
  - e.g: 50 OU students selected randomly from the Department of Mathematics and Statistics; Pounds of fish caught by 25 selected persons in a fishing derby.
  - The purpose to collect a sample is to get useful information about the population.
  - The manner in which a sample is generated must be carefully designed.
  - **Sample size**: The number of units in the sample. The sample size determines the *level of randomness*.

- In a study, only certain characteristics of the objects in a population are of interest. A **variable** is any characteristic whose value may change from one unit to another in the population.
- A variable can be
  - **Categorical** (or qualitative, nonnumeric): When the characteristic under study concerns a qualitative trait that is only classified in categories and not numerically measured.
    - Nominal: gives names or labels to various categories and therefore has no order information. e.g., gender, colors, zip codes, social security numbers, etc.
    - Ordinal: keeps the properties of nominal plus the order information. e.g., evaluation {Poor, Fair, Good, Better, Best}.
  - **Quantitative** (or numeric): When the characteristic is measured on a numerical scale.
    - Discrete: a finite or countable number of possible values. e.g., counting numbers.
    - Continuous: infinite possible values from an interval of real number. e.g., physically measurable quantities of length, volume, time, mass, etc.

- A **Univariate** data set consists of observations on a single variable.
  - The types of transmission (automatic or manual) of 5 cars: A, M, M, A, A.
  - The numbers of years of higher education of 3 engineers: 5, 7, 4.
- A **Bivariate** data set consists of observations made on two variables.
  - For 10 families, $(x_1, y_1), \ldots, (x_{10}, y_{10})$, where

$$\begin{aligned} x_i &= \text{height of father in the } i\text{th family} \\ y_i &= \text{height of son in the } i\text{th family} \end{aligned}$$

- A **Multivariate** data set arises when observations are made on two or more variables.
  - Increasingly available and important in applications.

Information collected from 10 OU students

Obs	Sex	HrsSleep	Height	GPA
1	M	6.0	71	3.70
2	F	7.0	60	3.85
3	M	5.5	69	3.42
⋮				
10	F	6.5	58	3.57

- Unit: Any OU student.
- Population: The OU student body.
- Sample: 10 selected OU students.
- Sample size: 10.
- Variables: Sex, HrsSleep, Height, and GPA.
- Data type: multivariate.

Branches of Statistics

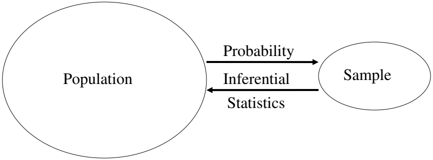
Descriptive statistics and **inferential statistics** are two branches of Statistics.

- **Descriptive statistics** summarize and describe important features of the data. Some methods in descriptive statistics are graphical in nature. Histograms, stem-and-leaf plots, boxplots, and scatterplots, etc. are primary graphs used in descriptive statistics. Other descriptive methods involve calculation of numerical summary measures such as means, standard deviations, and correlation coefficients, etc.
- Having obtained a sample information from a population, an investigator would frequently like to use sample information to draw some type of conclusion about the population. **A sample is a means to an end rather than an end in itself**. Techniques for generalizing conclusions from a sample to a population are gathered within **inferential statistics**.

Probability and Statistics

Although examples of statistical applications in daily life as well as scientific studies are drawn from widely different fields, a few common characteristics are readily discernible.

- First, in order to acquire new knowledge, relevant data must be collected.
- Second, some amount of variability in the data is unavoidable even though observations are made under the same or closely similar conditions.
- A third notable feature is that access to a complete set of data is either physically impossible or not feasible from a practical standpoint.



Getting conclusions about samples from their populations is called **probability analysis** and generalizing inferences from samples to populations is called **inferential statistics**. Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a sample from a given population. This is why we study probability before statistics.

Data Collection

The representativeness of sample to population is critical to statistical analysis. Random and unbiased ways of collecting data assure that the sample is a fair representative of the population. **Experimental and observational studies** are two most popular ways of data collection.

- In an **observational study**, the researcher records information concerning the subjects under study without any interference with the process that is generating the information. The researcher is a passive observer of the transpiring events.
- In an **experimental study**, the researcher actively manipulates certain variables associated with the study, called the **explanatory variables**, and then records their effects on the **response variables** associated with the experimental subjects.

Observational Study

Example (Example 1.1.1: Music and Good Grades)

- The Study
  - Comparing GPAs of music students and non-music students
- The Results
  - Music students: 3.59
  - Non-music students: 2.91

There are three basic types of observational studies: **sample survey, prospective study, and retrospective study**.

1. A sample survey is a study that provides information about a population at a particular point in time.
2. A prospective study is a study that observes a population in the present using a sample survey and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes.
3. A retrospective study is a study that observes a population in the present using a sample survey and also collects information about the subjects in the sample regarding the occurrence of specific outcomes that have already taken place.

Experimental Study

- There are two types of variables in an experimental study.
  - **Explanatory variables** (also called **independent variables** or **controlled variables** or **factors**) are selected by the researcher for comparison.
  - **Response variables** (also called **dependent variables**) are measurements or observations that are recorded but not controlled by the researcher.
- **Treatments** in an experimental study are the conditions constructed from combinations of factor levels.
- In experimental studies, there are many factors other than controlled factors potentially have influences on the response variables. Random assignment of subjects to treatments is critical.
- **Design of experiment** is also crucial in experimental studies such that the researcher follows a systematic plan established prior to running the experiment. The plan includes how all randomization is conducted, either the assignment of experimental units to treatments or the selection of units from the treatment populations.

A **designed experiment** is an investigation in which a specified framework is provided in order to observe, measure, and evaluate groups with respect to a designated response.

Example (Example 1.1.2: Commercially Raised Shrimp)

- A researcher is studying the conditions under which commercially raised shrimp reach maximum weight gain. Three water temperatures (25^o, 30^o, 35^o) and four water salinity levels (10%, 20%, 30%, 40%) were selected for the study. Shrimp were raised in containers with specified water temperatures and salinity levels. The weight gain of the shrimp in each container was recorded after 6-week study period.
- What are the factors and treatments?

1.2 Pictorial and Tabular Methods in Descriptive Statistics

- Subscripted variables
- Summation notation
- Display data from a single variable with tables and graphs
  - Frequency tables
  - Pie charts
  - Bar charts
  - Stem-and-leaf displays
  - Frequency tables
  - Histograms
  - Boxplots

Subscripted Variables

- Notation: In statistics, we almost always have multiple observations (values) of the same variable. To distinguish the different values of the variable, we add a subscript to the variable; the subscript is written as a small number to the lower right of the variable. The subscript has values of 1 to n, where n is the number values of the variable.
- Example: For the sample of pH measurements $\{6.3, 6.2, 5.9, 6.5\}$, let x be the variable storing pH, then the sample size $n = 4$, and

$$x_1 = 6.3, x_2 = 6.2, x_3 = 5.9, x_4 = 6.5.$$

- More generally, we can use a letter such as i, j, or k to denote the subscript, so (for instance), the "ith" value of the variable x would be denoted by x_i.

Summation Notation

- Very often it is necessary to denote (or express algebraically) the sum of some or all of the values of a subscripted variable. This is done with **summation notation**:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

- Example:

i	1	2	3	4
x_i	-2	4	6	5
y_i	3	7	-1	8

$$\sum_{i=1}^{4} x_i = ? \; \sum_{i=2}^{4} y_i = ? \; \sum_{i=1}^{4} x_i y_i = ? \; \sum_{i=1}^{4} x_i^2 = ? \; \left(\sum_{i=1}^{4} x_i\right)^2 = ?$$

Data from One Qualitative Variable

Example (Example 1.2.1: Job Danger)

The following table represents a summary of a study to determine which types of employment may be the most dangerous to their employees. The data were reported by National Safety Council in 1999. Approximately 3, 240, 000 workers suffered disabling injuries. Each of the 3, 240, 000 disabled workers was classified according to the industry group in which they were employed.

Industry group	Number of Disabling Injuries (in 1, 000s)	Relative Frequency Percent of Total
Agriculture	130	4.01
Construction	470	14.51
Manufacturing	630	19.44
Transportation & Utilities	300	9.26
Trade	380	11.72
Services	750	23.15
Government	580	17.90

What are the right questions to ask?

1. How many and what percentage of individuals fall into each category?
2. Are individuals equally divided across categories, or do the percentages across categories follow some other interesting pattern?

- **Frequency (f)**: The number of population members in that class (counts).
- **Relative frequency (rf)**: The fraction of population members in that class (given as proportions or percentages, for example).

$$rf = f/n$$

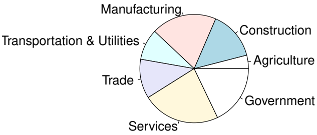
- A **frequency table** is a table whose first column displays each distinct outcome, second column displays that outcome's frequency, and third column displays the relative frequency.

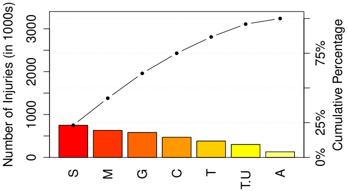
Graphical Displays of Qualitative Data

Graphs are the most important tool for examining data because they convey comparative information in ways that no table or description ever could. Trends, differences, and associations are effortlessly seen in the blink of an eye. The eye perceives immediately what the brain would take much longer to deduce from a table of numbers. Qualitative data are often presented graphically by a **pie chart** or a **bar chart**.

- A **pie chart** is a circle partitioned into segments such that the angle of each segment is proportional to the contribution of each category in the total value of a variable, where the total corresponds to 360°.
- A **bar chart** is a graph consisting of horizontal or vertical rectangles or bars, the height of each bar depicts the magnitude of the variable in each category.
- A **Pareto diagram**, named after Vilfredo Federico Damaso Pareto (Italian engineer, sociologist, economist, political scientist and philosopher), is a special bar chart where bars are arranged in a nonincreasing order with a line of cumulative percentages.

Job Danger





```
> pie(c(130,470,630,300,380,750,580), c("Agriculture","Construction","Manufacturing",
"Transportation & Utilities","Trade","Services","Government"),radius=1)
> x<-c(130,470,630,300,380,750,580)
> names(x)<-c("A","C","M","T.U","T","S","G")
> barplot(x)
> library(qcc)
> pareto.chart(x,ylab="Number of Injuries(in 1000s)",main="")
Pareto chart analysis for x
        Frequency   Cum.Freq.  Percentage Cum.Percent.
  S    750.000000  750.000000   23.148148    23.148148
  M    630.000000 1380.000000   19.444444    42.592593
  G    580.000000 1960.000000   17.901235    60.493827
  C    470.000000 2430.000000   14.506173    75.000000
  T    380.000000 2810.000000   11.728395    86.728395
  T.U  300.000000 3110.000000    9.259259    95.987654
  A    130.000000 3240.000000    4.012346   100.000000
```

Example (Example 1.2.2: Letter Grades)

The following data is the letter grades of 20 persons in a statistics discussion section:

C, D, F, A, A, B, B, A, A, C, C, C, D, F, B, D, B, C, B, C.

Try to construct a frequency table, a pie chart and a bar chart.

Data from One Quantitative Variable with Discrete Values

Example (Example 1.2.3: Daily Computer Stoppages)

The daily number of computer stoppages are observed over 30 days at a university computing center.

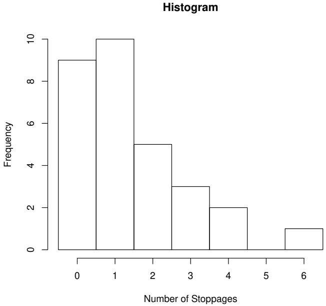
1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

What are the right questions to ask for quantitative data?

1. Are there individual values that provide interesting information because they are unique or stand out in some way?
2. What are the interesting summary measures that helps us understand the collection of individuals who were measured?

Discrete data could either be summarized by frequency tables or histograms.

Value	Frequency	Relative Frequency
0	9	0.30
1	10	0.33
2	5	0.17
3	3	0.10
4	2	0.07
6	1	0.03
Total	30	1



Constructing a Histogram for discrete data

1. Determine the frequency and relative frequency of each x value.
2. Mark possible x values on a horizontal scale.
3. Above each value, draw a rectangle:
    - leave no space between rectangles;
    - the rectangles should have equal width;
    - the y-axis can be frequency or relative frequency.

```
### R code for Example 1.2.3
X<-c(1,3,1,1,0,1,0,1,1,0,2,2,0,0,0,1,2,1,2,0,0,1,6,4,3,3,1,2,4,0)
table(X)
 0  1  2  3  4  6
 9 10  5  3  2  1
hist(X,br=seq(-0.5,6.5,1),xlab="Number of Stoppages",main="Histogram",right=F)
```

Data from One Quantitative Variable with Continuous Values

Example (Example 1.2.4: Coyote Lengths)

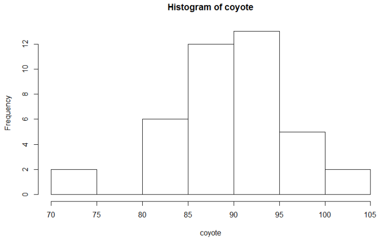
Below is a data set of 40 female coyote lengths:

93.0	97.0	92.0	101.6	93.0	84.5	102.5	97.8	91.0	98.0
93.5	91.7	90.2	91.5	80.0	86.4	91.4	83.5	88.0	71.0
81.3	88.5	86.5	90.0	84.0	89.5	84.0	85.0	87.0	88.0
86.5	96.0	87.0	93.5	93.5	90.0	85.0	97.0	86.0	73.7

Constructing a Histogram for continuous data

1. Determine the frequency and relative frequency for each equal-sized class.
2. Mark the class boundaries on a horizontal measurement axis.
3. Above each class interval, draw a rectangle whose height is the corresponding frequency or relative frequency.

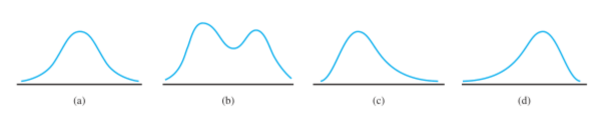
Class Intervals	Frequency	Relative Frequency
[70, 75)	2	0.05
[75, 80)	0	0
[80, 85)	6	0.15
[85, 90)	12	0.30
[90, 95)	13	0.325
[95, 100)	5	0.125
[100, 105)	2	0.05
Total	40	1



```
### R code for Example 1.2.4
coyote = c(93.0,97.0,92.0,101.6,93.0,84.5,102.5,97.8,91.0,98.0,93.5,91.7,90.2,91.5,80.0,86.4,
91.4,83.5,88.0,71.0,81.3,88.5,86.5,90.0,84.0,89.5,84.0,85.0,87.0,88.0,86.5,96.0,
87.0,93.5,93.5,90.0,85.0,97.0,86.0,73.7)
# Freq Histogram
h = hist(coyote,right=FALSE)
# Relative Freq Histogram
h$counts = h$counts/sum(h$counts)
plot(h,ylab="Relative Frequency")
```

Shapes of Histograms

- A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks. A histogram with more than two peaks is said to be **multimodal**.
- A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail. A unimodal histogram is **negatively skewed** if the left or lower tail is stretched out compared with the right or upper tail.



Stem-and-Leaf Displays

A **stem-and-leaf display**, or **stemplot** is a simple device to obtain an informative visual representation of the data set.

How to Construct a Stemplot

1. Select one or more leading digits for the stem values. The trailing digits become the leaves. e.g.,

$$47 \rightarrow \underbrace{4}_{\text{stem}}\,\underbrace{7}_{\text{leaf}}$$

$$265 \rightarrow \underbrace{26}_{\text{stem}}\,\underbrace{5}_{\text{leaf}}$$

2. List possible stem values in a vertical column.
3. Record the leaf for each observation beside the corresponding stem value, and arrange leaves in increasing order.
4. Indicate the units for stems and leaves someplace in the display.

Example (Example 1.2.5: Binge Drinking)

The article "Health and Behavioral Consequences of Binge Drinking in College" (J. of the Amer. Med. Assoc., 1994: 1672-1677) reported on a comprehensive study of heavy drinking across the United States. A binge episode was defined as five or more drinks in a row for male and four or more for females. The data are percentages of undergraduate students who are binge drinkers.

47	34	21	38	27	36	35	42	30	33
43	51	33	18	31	33	35	45	43	55

Stemplot: $1|8 = 18$

$$\begin{array}{r|l} 1 & \\ 2 & \\ 3 & \\ 4 & \\ 5 & \end{array} \quad \Rightarrow \quad \begin{array}{r|l} 1 & 8 \\ 2 & 1\ 7 \\ 3 & 4\ 8\ 6\ 5\ 0\ 3\ 3\ 1\ 3\ 5 \\ 4 & 7\ 2\ 3\ 5\ 3 \\ 5 & 1\ 5 \end{array} \quad \Rightarrow \quad \begin{array}{r|l} 1 & 8 \\ 2 & 1\ 7 \\ 3 & 0\ 1\ 3\ 3\ 3\ 4\ 5\ 5\ 6\ 8 \\ 4 & 2\ 3\ 3\ 5\ 7 \\ 5 & 1\ 5 \end{array}$$

```
### R code for Example 1.2.5
binge<-c(47,34,21,38,27,36,35,42,30,33,43,51,33,18,31,33,35,45,43,55)

stem(binge)

  The decimal point is 1 digit(s) to the right of the |

  1 | 8
  2 | 17
  3 | 0133345568
  4 | 23357
  5 | 15
```

A stem-and-leaf display conveys information about the following aspects of the data.

1. Identification of a typical or representative value
2. Extent of spread about the typical value
3. Presence of any gaps in the data
4. Extent of symmetry in the distribution of values
5. Number and location of peaks
6. Presence of any *outliers* - values far from the rest of the data

1.3 Measures of Location: Statistical Ideas*

- Consider a list of n measurements $x_1, x_2, \cdots, x_n$. We are looking for a single summary τ such that τ (Greek letter, read "tau") is close or representative to x_i's.

- Measure the total discrepancy between τ and x_i's, $d = \sum_{i=1}^{n} d(x_i, \tau)$:
  - $d(x_i, \tau) = 1$ if $\tau \neq x_i$, and 0 otherwise and $d = \sum_{i=1}^{n} d(x_i, \tau)$.
  - $d(x_i, \tau) = |x_i - \tau|$ and $d = \sum_{i=1}^{n} |x_i - \tau|$.
  - $d(x_i, \tau) = (x_i - \tau)^2$ and $d = \sum_{i=1}^{n} (x_i - \tau)^2$.

- Minimize the total discrepancy: The **mode** minimizes the number of times of nonzero deviations from the summary; the **median** minimizes the sum of the absolute deviations from the summary; and the **mean** minimizes the sum of the squared deviations from the summary.

White, John Myles (2013, Modes, Medians and Means: A Unifying Perspective, at `http://www.johnmyleswhite.com/notebook/2013/03/22/modes-medians-and-means-an-unifying-perspective`

Schwertman, N.C., Gilks, A.J., and Cameron, J.A. (1990, A simple noncalculus proof that the median minimizes the sum of the absolute deviations, *The American Statistician* 44(1), 38-39).

Measures of Location: Quantitative Data

Mean

- The *sample mean* of a set of n measurements $x_1, x_2, \cdots, x_n$ is the sum of these measurements divided by n.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- The *population mean* is the average of all values in the population, usually denoted by μ.

Median

- The *(sample) median*, denoted by $\tilde{x}$, of a set of n measurements x_1, x_2, $\cdots$, x_n is the midpoint, the number such that half the observations are smaller and the other half are larger.

How to find the median

1. Sort the observations in increasing order:

$$x_{(1)}, x_{(2)}, \cdots , x_{(n)}.$$

2. When n is odd,

$$\tilde{x} = \text{median}(x_1, x_2, \cdots , x_n) = x_{(\frac{n+1}{2})}.$$

3. when n is even,

$$\tilde{x} = \text{median}(x_1, x_2, \cdots , x_n) = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

- The *population median* is a middle value in the population, a number such that half the population are smaller and the other half are greater, usually denoted by $\tilde{\mu}$.

Example (Example 1.3.1)

A data set has $n = 12$ observations.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
2.3	8.8	3.9	4.1	6.4	5.9	4.2	2.9	1.3	5.1	1.7	3.5

Example (Example 1.3.1)

A data set has $n = 12$ observations.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
2.3	8.8	3.9	4.1	6.4	5.9	4.2	2.9	1.3	5.1	1.7	3.5

Sort in increasing order:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
1.3	1.7	2.3	2.9	3.5	3.9	4.1	4.2	5.1	5.9	6.4	8.8

Example (Example 1.3.1)

A data set has $n = 12$ observations.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
2.3	8.8	3.9	4.1	6.4	5.9	4.2	2.9	1.3	5.1	1.7	3.5

Sort in increasing order:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
1.3	1.7	2.3	2.9	3.5	3.9	4.1	4.2	5.1	5.9	6.4	8.8

Since the number of observations is even,

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{3.9 + 4.1}{2} = 4.$$

Other Measures of Location: Quantitative Data

- **Mode**: The mode is the value which appears most often. It is not necessarily unique. And there could be no mode, one mode, or more than one mode.
- **Quartiles**: Roughly speaking, quartiles divide the data into four equal parts, with observations above the third quartile (Q_3, also called upper fourth) constituting the upper quarter of the data set, the second quartile (Q_2) being identical to the median, and the first quartile (Q_1, also called lower fourth) of the data separating the lower quarter from the upper three-quarters.
- **Percentiles**: The $100p$th percentile is a value such that after the data are ordered from the smallest to the largest, at least $100p\%$ of the observations are at or below this value and at least $100(1-p)\%$ are at or above the value. Roughly, if np is not integer, the $100p$th percentile is $x_{(\lceil np \rceil)}$; if np is an integer the $100p$th percentile is $[x_{(np)} + x_{(np+1)}]/2$.

Quartiles

- Order the data from the smallest to the largest and separate the smaller half from the larger half. If n is odd, exclude the median $\tilde{x}$ in both halves.
- The *lower fourth* Q_1 is the median of the smaller half and the *upper fourth* Q_3 is the median of the larger half. The *fourth spread* f_s is defined as

$$f_s = \text{upper fourth} - \text{lower fourth}.$$

That is, $f_s = Q_3 - Q_1 = IQR$ (Interquartile Range).

Example (Example 1.3.2)

Data set

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
2	4	3	4	6	5	4	-6	5

Sort in increasing order:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
-6	2	3	4	4	4	5	5	6

$Q_1 = \text{median}\{-6, 2, 3, 4\} = 2.5$, $Q_3 = \text{median}\{4, 5, 5, 6\} = 5$, $f_s = Q_3 - Q_1 = 2.5$.

- **Five-number Summary**: minimum, lower fourth Q_1, median $\tilde{x}$, upper fourth Q_3, maximum.
- **Outliers**: Any observation farther than $1.5f_s$ from the closest fourth: $> Q_3 + 1.5f_s$ or $< Q_1 - 1.5f_s$.
    - *Extreme*: more than $3f_s$ from the nearest fourth.
    - *Mild*: otherwise.
- **Trimmed mean**: A $100\alpha\%$ trimmed mean (denoted by $\bar{x}_{\text{tr}(100\alpha)}$) is computed by eliminating the smallest $100\alpha\%$ and the largest $100\alpha\%$ of the data and then averaging what is left over. The trimmed mean is a compromise between $\bar{x}$ and $\tilde{x}$. If $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ are the order statistics, the α-trimmed mean is defined as

$$\bar{x}_{\text{tr}(100\alpha)} = \frac{1}{n - 2\lfloor n\alpha \rfloor} \sum_{k=\lfloor n\alpha \rfloor + 1}^{n-\lfloor n\alpha \rfloor} x_{(k)},$$

where $\lfloor n\alpha \rfloor$ is the largest integer less than or equal to $n\alpha$.

A trimmed mean with a moderate trimming percentage (someplace between 5% and 25%) will yield a measure of center that is neither as sensitive to outliers as is the mean nor as insensitive as the median.

Example (Example 1.3.3: Copper Content)

The production of Bidri is a traditional craft of India. Bidri wares (bowls, vessels, and so on) are cast from an alloy containing primarily zinc along with some copper. Consider the following 26 observations on copper content (%) for a sample of Bidri artifacts in London's Victoria and Albert Museum (Craddock, P. T., 2005, Enigmas of bidri, *Surface Engineering* 21(5-6), 333-339): 2.0, 2.4, 2.5, 2.6, 2.6, 2.7, 2.7, 2.8, 3.0, 3.1, 3.2, 3.3, 3.3, 3.4, 3.4, 3.6, 3.6, 3.6, 3.6, 3.7, 4.4, 4.6, 4.7, 4.8, 5.3, 10.1.

```
### R code for Example 1.3.3
> copper<-c(2.0,2.4,2.5,2.6,2.6,2.7,2.7,2.8,3.0,3.1,3.2,3.3,
3.3,3.4,3.4,3.6,3.6,3.6,3.6,3.7,4.4,4.6,4.7,4.8,5.3,10.1)
> mean(copper)
[1] 3.653846
> median(copper)
[1] 3.35
> fivenum(copper)  # Note: R uses a different definition for Q1 and Q3
[1]  2.00  2.70  3.35  3.70 10.10
> mean(copper,trim=1/13)
[1] 3.418182
```

Mean v.s. Median

- The more symmetric the distribution is, the more they are close together.
- In a skewed distribution, mean is pulled toward the longer tail.
- Median is more resistant. Outliers and skewed distributions can alter the mean a lot.

Example (Example 1.3.4*)

Suppose you are in a class of 40 students. You get a score of 79 in the midterm exam. You also know others' scores. The distribution is that 1 outstanding student gets 91, 1 poor guy gets only 10, and all of the rest 37 get 80.

- Mean:

$$\bar{x} = \frac{79 + 91 + 10 + 80 \times 37}{40} = 78.5.$$

- Median: $\tilde{x} = 80$.
- Trimmed mean:

$$\begin{aligned} \bar{x}_{\mathrm{tr}(2.5)} &= \frac{80 \times 37 + 79}{38} = 79.97. \\ \bar{x}_{\mathrm{tr}(5)} &= 80. \end{aligned}$$

Measures of Location: Qualitative Data

When the data is qualitative, *sample proportion* is often used.

- Let x be the number of observations in the sample falling in the category of interest. Then, $\hat{p} = x/n$ is the sample proportion (or relative frequency) in this category.
- The counterpart is p (population proportion).

Flip #	1	2	3	4	5	6	7	8	9	10
Head or not	1	0	1	0	0	1	0	0	0	1

The sample proportion of head is

$$\hat{p} = \frac{4}{10}.$$

1.4 Measures of Variability

Only a measure of location is far from enough.

Example (Example 1.4.1)

A large bakery regularly orders cartons of Maine blueberries from two suppliers. The target weight of the cartons is supposed to be 22 ounces. Samples of cartons from two suppliers were weighted. Here are the data:

Supplier I: 17, 22, 22, 22, 22, 22, 27

Supplier II: 17, 19, 20, 22, 22, 27, 27

Which supplier is more satisfactory?

- **Range**: the maximum value minus the minimum.
- The **(sample) variance** of n observations $x_1, x_2, ..., x_n$ is

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

  - A shortcut formula to compute s^2 is

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

  - The counterpart is σ^2, the *population variance*.
- The **(sample) standard deviation** is

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.}$$

The *population standard deviation* is σ.

- To compute the sample variance and standard deviation of blueberry weights from Supplier II:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
17	$17 - 22 = -5$	$(-5)^2 = 25$
19	$19 - 22 = -3$	$(-3)^2 = 9$
20	$20 - 22 = -2$	$(-2)^2 = 4$
22	$22 - 22 = 0$	$0^2 = 0$
22	$22 - 22 = 0$	$0^2 = 0$
27	$27 - 22 = 5$	$5^2 = 25$
27	$27 - 22 = 5$	$5^2 = 25$
$\bar{x} = 22$	sum $= 0$	sum $= 88$

- $s^2 = 88/(7-1) = 14.67, s = \sqrt{14.67} = 3.83.$
- Using the shortcut method:

$$\sum x_i^2 = 3476$$

$$s^2 = \frac{1}{7-1}\left(3476 - (7)(22^2)\right) = 14.67.$$

Example (Example 1.4.2: Variation of Copper Content, Example 1.3.3 cont'd)

```
> copper<-c(2.0,2.4,2.5,2.6,2.6,2.7,2.7,2.8,3.0,3.1,3.2,3.3,
  3.3,3.4,3.4,3.6,3.6,3.6,3.6,3.7,4.4,4.6,4.7,4.8,5.3,10.1)
#Range
> diff(range(copper))
[1] 8.1
#Variance
> var(copper)
[1] 2.394585
#Standard deviation
> sd(copper)
[1] 1.547445
```

Properties of the mean (median) and standard deviation

- $n - 1$ is called the degrees of freedom.
- s measures variability about the mean and should be used only when the mean is chosen as the measure of center.
- s is always zero or greater than zero. $s = 0$ only when there is no variability. This happens only when all observations have the same value. Otherwise, $s > 0$.
- As the observations become more variable about their mean, s gets larger.
- s has the same units of measurement as the original observations. For example, if you measure weight in kilograms, both the mean $\bar{x}$ and the standard deviation s are also in kilograms. This is one reason to prefer s to the variance s^2, which would be in squared kilograms.
- Like the mean $\bar{x}$, s is not resistant. A few outliers can make s very large.

- The Empirical Rule (68-95-99.7): when the distribution appears symmetric and bell-shaped

  1. Approximately 68% of the data lie with in $\bar{x} \pm s$.
  2. Approximately 95% of the data lie with in $\bar{x} \pm 2s$.
  3. Approximately 99.7% of the data lie with in $\bar{x} \pm 3s$.

- *Chebyshev's Theorem: any distribution

  - At least $(1 - \frac{1}{k^2})$ of the data lie within $\bar{x} \pm ks$, where $k > 1$.

Example (Example 1.4.3*: Who Are the Speediest Drivers?)

A survey in Penn State University of 87 male students results the following data: 55 60 80 80 80 80 85 85 85 85 90 90 90 90 90 92 94 95 95 95 95 95 95 100 100 100 100 100 100 100 100 100 101 102 105 105 105 105 105 105 105 105 109 110 110 110 110 110 110 110 110 110 110 110 112 115 115 115 115 115 115 120 120 120 120 120 120 120 120 120 120 124 125 125 125 125 125 125 130 130 135 140 140 140 140 145 150

```
speed<-c(55,60,80,80,80,80,85,85,85,85,90,90,90,90,90,92,94,95,95,95,95,95,95,100,100,100,100,100,100,100,
100,100,101,102,105,105,105,105,105,105,105,105,109,110,110,110,110,110,110,110,110,110,110,110,112,115,115,
115,115,115,115,120,120,120,120,120,120,120,120,120,120,124,125,125,125,125,125,125,130,130,135,140,140,140,
140,145,150)
> mean(speed);sd(speed)
[1] 107.6897
[1] 17.68149
#Emprical/Chebyshev
> sum(speed>=mean(speed)-sd(speed)&speed<=mean(speed)+sd(speed))/87
[1] 0.7241379
> sum(speed>=mean(speed)-2*sd(speed)&speed<=mean(speed)+2*sd(speed))/87
[1] 0.954023
> sum(speed>=mean(speed)-3*sd(speed)&speed<=mean(speed)+3*sd(speed))/87
[1] 1
```

- If a fixed number c is added to all measurements in a data set, then the mean (median) of the new measurements is c plus the original mean (resp. c plus the original median)
- If all measurements in a data set are multiplied by a fixed number d, then the mean (median) of the new measurements is d times the original mean (resp. d times the original median)
- If a fixed number c is added to all measurements in a data set, then the sample variance and standard deviation of the new measurements are the same as that of the original ones.
- If all measurements in a data set are multiplied by a fixed number d, then the variance of the new measurements is equal to d^2 times the original one; the standard deviation of the new measurements is equal to $|d|$ times the original one.

Example (Example 1.4.4)

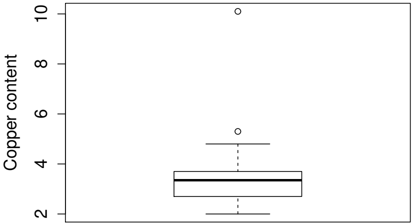
- A local Community College charges tuition:

  \$73 per credit plus a flat **fee of \$35** per semester.
  Ex: 12 credits would be billed $\$73(12) + \$35 = \$911$

- From a sample of students with $\bar{x} = 16.65$ credits and $s = 2.96$ credits. What would be the mean and standard deviation of the students' tuitions?
- Mean: $73(16.65) + 35 = 1250.45$.
- Standard deviation: $73(2.96) = 216.08$.

Boxplot: A Graph Display of Five Number Summary

Boxplot or box-and-whisker plot (invented by Mary Eleanor Spear, 1952, Charting Statistics, p. 166 and popularized by John Wilder Tukey, 1977, Exploratory data analysis) is a simple yet powerful tool to display a single batch of data, to study symmetry, "longtailedness," and distributional assumptions; to compare parallel batches of data; and to supplement more complex displays with univariate information.



f_s or IQR (interquartile range) is the difference between upper fourth (Q_3) and lower fourth (Q_1). Any observational farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth and it is **mild** otherwise.

```
### R code
copper<-c(2.0,2.4,2.5,2.6,2.6,2.7,2.7,2.8,3.0,3.1,3.2,3.3,
  3.3,3.4,3.4,3.6,3.6,3.6,3.6,3.7,4.4,4.6,4.7,4.8,5.3,10.1)
boxplot(copper)
```