Chapter 1

# Overview and Descriptive Statistics

---

1.1

## Populations, Samples, and Processes

---

### Populations and Samples

A *population* is a well-defined collection of objects.

When information is available for the entire population we have a *census*. A subset of the population is a *sample*.

---

### Data and Observations

*Univariate* data consists of observations on a single variable (*multivariate* – more than two variables).

---

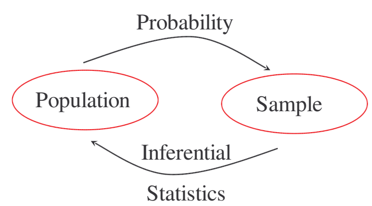
### Branches of Statistics

*Descriptive Statistics* – summary and description of collected data.

*Inferential Statistics* – generalizing from a sample to a population.

---

### Relationship Between Probability and Inferential Statistics




---

1.2

## Pictorial and Tabular Methods in Descriptive Statistics

---

### Stem-and- Leaf Displays

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List stem values in a vertical column.
3. Record the leaf for every observation.
4. Indicate the units for the stem and leaf on the disply.

---

### Stem-and-Leaf Example

Observed values:
9, 10, 15, 22, 9, 15, 16, 24,11

0 | 9 9
2 | 1 0 5 5 6
3 | 2 4

Stem: tens digit

Leaf: units digit

---

### Stem-and- Leaf Displays

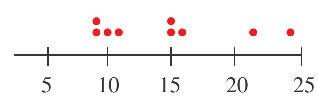
- Identify typical value
- Extent of spread about a value
- Presence of gaps
- Extent of symmetry
- Number and location of peaks
- Presence of outlying values

---

### Dotplots

Represent data with dots.

Observed values:
9, 10, 15, 22, 9, 15, 16, 24,11




---

### Types of Variables

A variable is *discrete* if its set of possible values constitute a finite set or an infinite sequence. A variable is *continuous* if its set of possible values consists of an entire interval on a number line.

---

### Histograms: Discrete Data

Determine the frequency and relative frequency for each value of $x$. Then mark possible $x$ values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency of that value.

---

Ex. Students from a small college were asked how many charge cards that they carry. $x$ is the variable representing the number of cards and the results are below.

| *x* | #people | *Rel. Freq* |
|---|---|---|
| 0 | 12 | 0.08 |
| 1 | 42 | 0.28 |
| 2 | 57 | 0.38 |
| 3 | 24 | 0.16 |
| 4 | 9 | 0.06 |
| 5 | 4 | 0.03 |
| 6 | 2 | 0.01 |

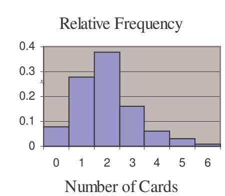
Frequency Distribution

---

### Histograms

Credit card results:

| *x* | *Rel. Freq.* |
|---|---|
| 0 | 0.08 |
| 1 | 0.28 |
| 2 | 0.38 |
| 3 | 0.16 |
| 4 | 0.06 |
| 5 | 0.03 |
| 6 | 0.01 |




---

### Histograms Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Then mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the relative frequency.

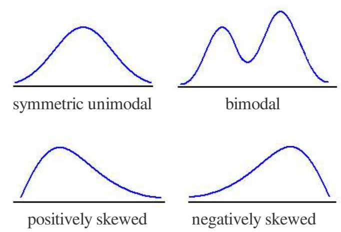
## Histograms (Continuous Data): Unequal Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using:

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting heights are called *densities* and the vertical scale is the *density scale*.

## Histogram Shapes




## 1.3 Measures of Location

## The Mean

The average (*mean*) of the $n$ numbers $x_1, x_2, \ldots, x_n$ is $\bar{x}$ where

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

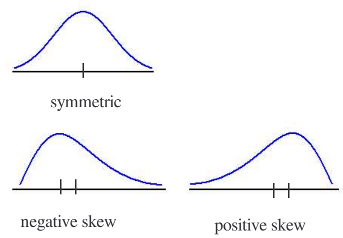
*Population mean*: $\mu$

## Median

The *sample median* is the middle value in a set of data that is arranged in ascending order. For an even number of data points the median is the average of the middle two.

*Population median*: ũ

## Three Different Shapes for a Population Distribution




## 1.4 Measures of Variability

## Sample Variance

*Variance* is a measure of the spread of the data.

The *sample variance* of the sample $x_1$, $x_2$, $\ldots x_n$ of $n$ values of $X$ is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

We refer to $s^2$ as being based on $n - 1$ degrees of freedom.

## Standard Deviation

*Standard deviation* is a measure of the spread of the data using the same units as the data.

The *sample standard deviation* is the square root of the sample variance:

$$s = \sqrt{s^2}$$

## Formula for $s^2$

An alternative expression for the numerator of $s^2$ is

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

## Properties of $s^2$

Let $x_1, x_2, \ldots, x_n$ be any sample and $c$ be any nonzero constant.

1. If $y_1 = x_1 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$
2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$,

where $s_x^2$ is the sample variance of the $x$'s and $s_y^2$ is the sample variance of the $y$'s.

## Upper and Lower Fourths

After the $n$ observations in a data set are ordered from smallest to largest, the *lower* (*upper*) *fourth* is the median of the smallest (largest) half of the data, where the median is included in both halves if $n$ is odd. A measure of the spread that is resistant to outliers is the *fourth spread (IQR)* $f_s$ = upper fourth – lower fourth.

## Outliers

Any observation farther than $1.5 f_s$ from the closest fourth is an *outlier*. An outlier is *extreme* if it is more than $3 f_s$ from the nearest fourth, and it is *mild* otherwise.

## Boxplots