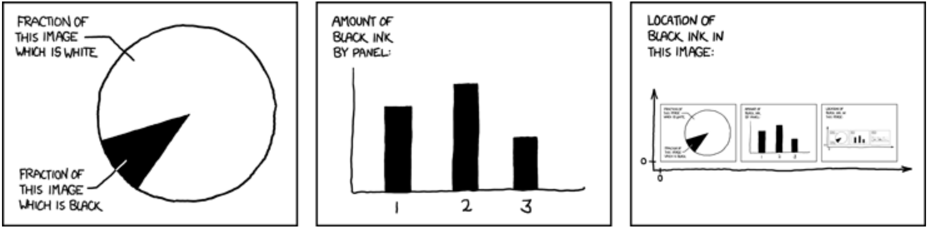
Ch. 1 – Descriptive Statistics and Overview



xkcd.com

Example 1 – Concrete Strength Data

The following measurements of flexural strength (in MPa) were taken from 27 specimens of high-performance concrete obtained by using superplasticizers and certain binders[1]:

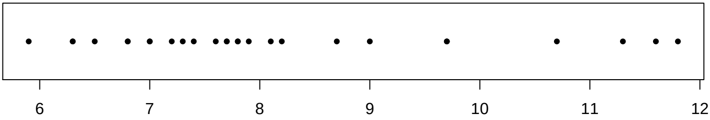
5.9, 6.3, 6.3, 6.5, 6.8, 6.8, 7.0, 7.0, 7.2, 7.3, 7.4, 7.6, 7.7, 7.7, 7.8, 7.8, 7.9, 8.1, 8.2, 8.7, 9.0, 9.7, 9.7, 10.7, 11.3, 11.6, 11.8

- How can we **visualize** this distribution?
- How can we quantify the **center** of the distribution?
- How can we quantify the **spread** of the distribution?

[1] From "Effects of Aggregates and Microfillers on the Flexural Properties of Concrete", **Magazine of Concrete Research**, 1997: 81–98

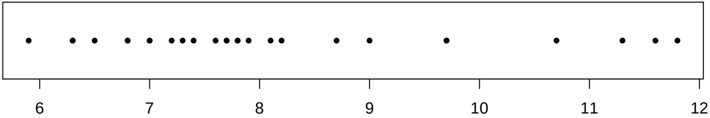
Dotplot

One simple method of visualization is a **dotplot** (or strip chart). The 27 observations are simply plotted on a line:

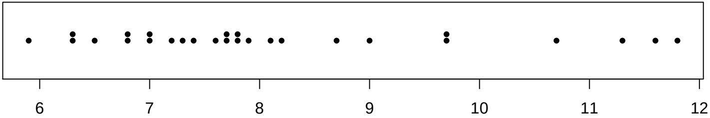


Dotplot

One simple method of visualization is a **dotplot** (or strip chart). The 27 observations are simply plotted on a line:



Identical values may be represented by stacking the dots:



Measures of center: Sample Mean and Median

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

Measures of center: Sample Mean and Median

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

2. The **sample median**: Sort the data from smallest to largest. If n is odd, then the sample median $\tilde{x}$ is the middle observation in the list; if n is even, then $\tilde{x}$ is the average of the two middle observations. Explicitly, if $x_1 \leq x_2 \leq \cdots \leq x_n$,

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

Measures of center: Sample Mean and Median

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

2. The **sample median**: Sort the data from smallest to largest. If n is odd, then the sample median $\tilde{x}$ is the middle observation in the list; if n is even, then $\tilde{x}$ is the average of the two middle observations. Explicitly, if $x_1 \leq x_2 \leq \cdots \leq x_n$,

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

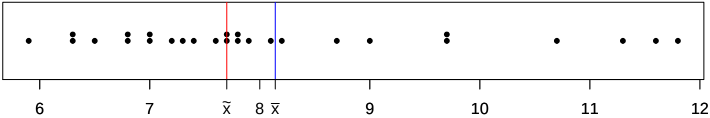
For example, given data 2, 3, 9, 11, 15, 17 we would have

$$\text{Sample mean: } \overline{x} = \frac{1}{6}(2 + 3 + 9 + 11 + 15 + 17) = 9.5$$

$$\text{Sample median: } \tilde{x} = \frac{1}{2}(9 + 11) = 10$$

Mean vs. Median

The sample mean and sample median for the strength data are $\overline{x} = 8.14$, $\tilde{x} = 7.7$.



- The mean may be strongly influenced by a few extreme observations, whereas the median is **robust** against such influence.
- If the largest measurement 11.8 were replaced 118, then the mean would increase to $\overline{x} = 12.07$, while the median would remain $\tilde{x} = 7.7$.

Sample Mean vs. Sample Median

If the sample median is robust but the sample mean isn't, why use the mean?

Sample Mean vs. Sample Median

If the sample median is robust but the sample mean isn't, why use the mean?

- In many cases, the sample mean is more **efficient** than the sample median: it achieves greater accuracy with a smaller sample size.

Sample Mean vs. Sample Median

If the sample median is robust but the sample mean isn't, why use the mean?

- In many cases, the sample mean is more **efficient** than the sample median: it achieves greater accuracy with a smaller sample size.
- To address the sample mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.

Sample Mean vs. Sample Median

If the sample median is robust but the sample mean isn't, why use the mean?

- In many cases, the sample mean is more **efficient** than the sample median: it achieves greater accuracy with a smaller sample size.
- To address the sample mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.
- To achieve a balanced tradeoff between efficiency and robustness, there are hybrid approaches such as a **trimmed mean**, e.g., discarding the top 10% and bottom 10% and then taking the sample mean of the remaining data.

Sample Mean vs. Sample Median

If the sample median is robust but the sample mean isn't, why use the mean?

- In many cases, the sample mean is more **efficient** than the sample median: it achieves greater accuracy with a smaller sample size.
- To address the sample mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.
- To achieve a balanced tradeoff between efficiency and robustness, there are hybrid approaches such as a **trimmed mean**, e.g., discarding the top 10% and bottom 10% and then taking the sample mean of the remaining data.
- We'll discuss these issues in Chapter 6 (Point Estimation).

Measure of Spread: Sample variance

In addition to describing the center of the data (using the mean or median), we also want to describe how "spread out" the data is. This can be measured using the **sample variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

Measure of Spread: Sample variance

In addition to describing the center of the data (using the mean or median), we also want to describe how "spread out" the data is. This can be measured using the **sample variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

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

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

Measure of Spread: Sample variance

In addition to describing the center of the data (using the mean or median), we also want to describe how "spread out" the data is. This can be measured using the **sample variance**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

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

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

Like the mean, the sample variance and sample standard deviation may be strongly influenced by extreme observations.

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

$$= \frac{(2-9.5)^2+(3-9.5)^2+(9-9.5)^2+(11-9.5)^2+(15-9.5)^2+(17-9.5)^2}{6-1}$$

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

$$= \frac{(2-9.5)^2+(3-9.5)^2+(9-9.5)^2+(11-9.5)^2+(15-9.5)^2+(17-9.5)^2}{6-1}$$

$$= 37.5$$

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} \\
&= \frac{(2-9.5)^2+(3-9.5)^2+(9-9.5)^2+(11-9.5)^2+(15-9.5)^2+(17-9.5)^2}{6-1} \\
&= 37.5
\end{aligned}
$$

Now we can calculate the sample standard deviation:

Calculating the Sample Variance and Standard Deviation

Given data 2, 3, 9, 11, 15, 17, we can calculate the sample variance (recall $\overline{x} = 9.5$):

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} \\
&= \frac{(2-9.5)^2+(3-9.5)^2+(9-9.5)^2+(11-9.5)^2+(15-9.5)^2+(17-9.5)^2}{6-1} \\
&= 37.5
\end{aligned}
$$

Now we can calculate the sample standard deviation:

$$s = \sqrt{37.5} \approx 6.12$$

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **first quartile**, or lower fourth.

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **first quartile**, or lower fourth.
- The median of the larger half is called the **third quartile**, or upper fourth.

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **first quartile**, or lower fourth.
- The median of the larger half is called the **third quartile**, or upper fourth.
- Their difference is the **interquartile range (IQR)**, or fourth spread:

$$\text{IQR} = \text{third quartile} - \text{first quartile}$$

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **first quartile**, or lower fourth.
- The median of the larger half is called the **third quartile**, or upper fourth.
- Their difference is the **interquartile range (IQR)**, or fourth spread:

$$\text{IQR} = \text{third quartile} - \text{first quartile}$$

Example: Given data 2,3,5,6,8,9,12,13,15 the median is $\tilde{x} = 8$, the first quartile is 5, the third quartile is 12, and the interquartile range is 7.

Robust Measure of Spread: Interquartile Range

A robust measure of spread, called the **interquartile range (IQR)**, is calculated as follows:

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **first quartile**, or lower fourth.
- The median of the larger half is called the **third quartile**, or upper fourth.
- Their difference is the **interquartile range (IQR)**, or fourth spread:

$$\text{IQR} = \text{third quartile} - \text{first quartile}$$

Example: Given data 2,3,5,6,8,9,12,13,15 the median is $\tilde{x} = 8$, the first quartile is 5, the third quartile is 12, and the interquartile range is 7.

	Efficient	Robust
Measure of center	$\bar{x}$	$\tilde{x}$
Measure of spread	s	IQR

Five-number summary and Boxplot

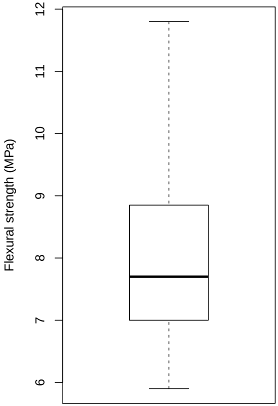
The **five-number summary** consists of

1. The minimum observation
2. The first quartile
3. The median
4. The third quartile
5. The maximum observation

For the concrete strength data:

$$5.90, 7.00, 7.70, 8.85, 11.80$$

This is shown graphically in a **boxplot**: The bottom and top of the box show the first and third quartiles; the horizontal line inside the box shows the median; the whiskers (dotted lines) extend to the minimum and maximum observations.



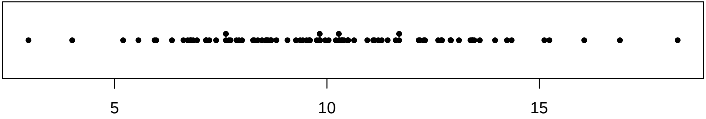
Example 2 – Power Usage

Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from a power company determined energy consumption (in BTU) for a sample of 90 homes over a fixed time interval:

2.97, 4.00, 5.20, 5.56, 5.94, 5.98, 6.35, 6.62, 6.72, 6.78, 6.80, 6.85, 6.94, 7.15, 7.16, 7.23, 7.39, 7.62, 7.62, 7.69, 7.73, 7.87, 7.93, 8.00, 8.26, 8.29, 8.37, 8.47, 8.56, 8.58, 8.61, 8.67, 8.69, 8.81, 9.07, 9.27, 9.37, 9.43, 9.52, 9.58, 9.60, 9.76, 9.82, 9.83, 9.83, 9.84, 9.96, 10.04, 10.21, 10.28, 10.28, 10.30, 10.35, 10.36, 10.40, 10.49, 10.50, 10.64, 10.95, 11.09, 11.12, 11.21, 11.29, 11.43, 11.62, 11.70, 11.70, 12.16, 12.19, 12.28, 12.31, 12.62, 12.69, 12.71, 12.91, 12.92, 13.11, 13.38, 13.42, 13.43, 13.47, 13.60, 13.96, 14.24, 14.35, 15.12, 15.24, 16.06, 16.90, 18.26

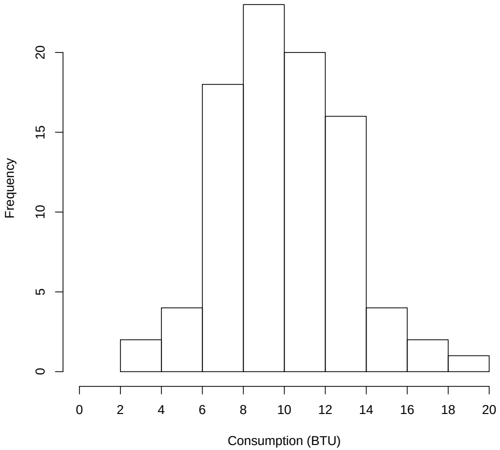
Example 2 – Power Usage

The dotplot is less effective here: many dots are packed close together, making it difficult to see where the dots are most concentrated:



Histogram

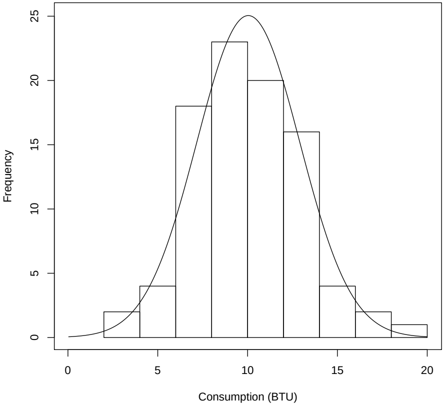
A **histogram** is constructed by grouping the data into bins of equal width and counting the number of observations within each bin:



Here the bins are the intervals (0,2], (2,4], (4,6], (6,8], . . . , (16,18], (18,20].

Normal Distribution

This histogram appears to be a fairly close fit to the symmetric, bell-shaped density curve of a **normal distribution**:



Normal distributions are very common in statistics and will be discussed in Chapter 4.

Stem-and-leaf diagram

3	0
4	0
5	269
6	03678889
7	2224667799
8	03345666778
9	134456688888
10	0023333445569
11	11234677
12	223367799
13	144456
14	023
15	12
16	19
17	
18	3

A **stem-and-leaf diagram** is constructed by grouping the data according to their first one or two digits (the **stem**), shown on the left part of the diagram, while the remaining digit (the **leaf**) is displayed on the right.

Here the first few measurements are 3.0, 4.0, 5.2, 5.6, 5.9, 6.0, 6.3, etc.

A stem-and-leaf diagram is similar to a "sideways" histogram. It is visually less appealing but provides more precise information about the data.

Population vs. Sample

- A **population** is a group that we want to study.

Population vs. Sample

- A **population** is a group that we want to study.
- A **sample** is a subset of a population that is observed and measured.

Population vs. Sample

- A **population** is a group that we want to study.
- A **sample** is a subset of a population that is observed and measured.
- For example, Gallup regularly polls American adults about their opinions on various topics. The relevant *population* consists of the group of all American adults, consisting of approximately 240 million people. For any given poll, however, the *sample* typically consists of at most a few thousand people who are actually contacted by Gallup and polled.

Population vs. Sample

- A **population** is a group that we want to study.
- A **sample** is a subset of a population that is observed and measured.
- For example, Gallup regularly polls American adults about their opinions on various topics. The relevant *population* consists of the group of all American adults, consisting of approximately 240 million people. For any given poll, however, the *sample* typically consists of at most a few thousand people who are actually contacted by Gallup and polled.
- In the concrete-strength example, only 27 specimens of concrete were tested. The 27 specimens form the *sample* which the researchers directly measure. The *population* consists of all such concrete specimens have been or could be formed by the same process and under the same conditions.

Estimation: Parameters vs. Statistics

- A **parameter** is a quantity which describes a population.

Estimation: Parameters vs. Statistics

- A **parameter** is a quantity which describes a population.
- A **statistic** is a quantity computed from sample data, used to estimate a population parameter.

Estimation: Parameters vs. Statistics

- A **parameter** is a quantity which describes a population.
- A **statistic** is a quantity computed from sample data, used to estimate a population parameter.
- Gallup ran a poll in December 2014, asking 805 American adults, "Do you approve or disapprove of the way Congress is handling its job?" 16% said "Yes". This 16% is a *statistic*, called the **sample proportion** $\hat{p}$. The percentage of all American adults who would answer "Yes" is a parameter, called the **population proportion** p. The parameter p is not known but is probably within a few percentage points of 16%.

Estimation: Parameters vs. Statistics

- A **parameter** is a quantity which describes a population.
- A **statistic** is a quantity computed from sample data, used to estimate a population parameter.
- Gallup ran a poll in December 2014, asking 805 American adults, "Do you approve or disapprove of the way Congress is handling its job?" 16% said "Yes". This 16% is a *statistic*, called the **sample proportion** $\hat{p}$. The percentage of all American adults who would answer "Yes" is a parameter, called the **population proportion** p. The parameter p is not known but is probably within a few percentage points of 16%.
- In the concrete-strength example, the 27 specimens had an average strength of 8.14 MPa. This is a *statistic*, the **sample mean** $\overline{x}$. The mean strength of all potential specimens is a *parameter*, the **population mean** μ. Likewise, the sample standard deviation s is a statistic, estimating the **population standard deviation** σ.

Statistical Inference

Statistical inference is the method of drawing *probabilistic* conclusions about a population based on sample data.

Statistical Inference

Statistical inference is the method of drawing *probabilistic* conclusions about a population based on sample data.

- A **confidence interval** is a range of values containing a population parameter, with a specific degree of confidence. For example, in the case of the Gallup poll, we can say with 95% confidence that between 12% and 20% of American adults approved of the way Congress was handling its job. We will discuss confidence intervals in Ch. 7 and Ch. 9.

Statistical Inference

Statistical inference is the method of drawing *probabilistic* conclusions about a population based on sample data.

- A **confidence interval** is a range of values containing a population parameter, with a specific degree of confidence. For example, in the case of the Gallup poll, we can say with 95% confidence that between 12% and 20% of American adults approved of the way Congress was handling its job. We will discuss confidence intervals in Ch. 7 and Ch. 9.
- A **hypothesis test** is a way to test whether a theory is consistent with the data or not. For example, in the Gallup poll, one might hypothesize that a different proportion of Republicans would answer "Yes" compared to Democrats. However, the data gives no significant support for this ($P = 0.53$). We will discuss hypothesis tests in Ch. 8.

Statistical Inference

Statistical inference is the method of drawing *probabilistic* conclusions about a population based on sample data.

- A **confidence interval** is a range of values containing a population parameter, with a specific degree of confidence. For example, in the case of the Gallup poll, we can say with 95% confidence that between 12% and 20% of American adults approved of the way Congress was handling its job. We will discuss confidence intervals in Ch. 7 and Ch. 9.
- A **hypothesis test** is a way to test whether a theory is consistent with the data or not. For example, in the Gallup poll, one might hypothesize that a different proportion of Republicans would answer "Yes" compared to Democrats. However, the data gives no significant support for this ($P = 0.53$). We will discuss hypothesis tests in Ch. 8.
- We cannot be certain that the conclusions of statistical inference are correct, but we can quantify the degree of uncertainty using *probability* (Chs. 2 through 5).

Key concepts

- Ways to display the distribution of a numerical variable:
  - dotplot, boxplot, histogram, stem-and-leaf diagram
- Measures of center: sample mean, sample median
- Measures of spread: sample variance, sample standard deviation, interquartile range
- Statistical inference: using *sample* data to draw probabilistic conclusions about a broader *population*.
  - Sample *statistics* are used to estimate population *parameters*.

Parameter	Statistic
population mean μ	sample mean $\overline{x}$
population standard deviation σ	sample standard deviation s
population proportion p	sample proportion $\hat{p}$