

3: Summary Statistics

[Measures of Central Location](#) | [Five-Point Summaries](#) | [Measures of Spread](#)

Measures of Central Location

Notation

In the preceding chapter we used stem-and-leaf plots and frequency tables to learn about distributional shape, location. In this chapter we use numerical summaries to describe location and spread. (Shape is rarely described numerically.)

Illustrative data (sample.sav). Let us again consider 10 AGE values (years) selected at random from a population. Data are:

21 42 5 11 30 50 28 27 24 52

Let

n represent the **sample size** (e.g., $n = 10$)

X represent the **variable** (in this case, AGE), and

x_i represent the **value** of the i^{th} observation in the data set (e.g., $x_1 = 21$).

The symbol Σ (capital sigma) is the **summation sign**, indicating all values should be added. For the illustrative data set, $\Sigma x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} = 21 + 42 + 5 + 11 + 30 + 50 + 28 + 27 + 24 + 52 = 290$.

We introduce three different measures of central location: the mean, the median, and the mode.

Mean

The **mean** is the *arithmetic average* of the data set.

The **population mean** (μ ; pronounced “mu”) is:

$$m = \frac{\sum x_i}{N} = \frac{1}{N} \sum x_i \quad (3.1)$$

where Σx represents the sum of all values and N represents the population size.

Illustrative example (populati.sav). The data set `populati.sav` represents a small finite population of $N = 600$. The sum of all the age values in this population is 17,703. Therefore, $\mu = 17,703 / 600 = 29.505$.

When values for the entire population are unavailable, we work from the sample. The **sample mean** is denoted \bar{x} (“x bar”):

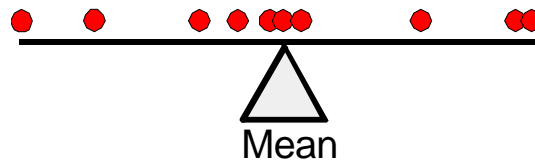
$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{n} \sum x_i \quad (3.2)$$

where $\sum x$ represents the sum of all values in the sample and n represents the sample size.

Illustrative example (sample.sav). For `sample.sav`, $\sum x_i = 290$ and $n = 10$. Therefore, $\bar{x} = 290 / 10 = 29.0$.

Notice that the operations specified in formula 3.1 and formula 3.2 are nearly identical: they both tell you to add all the values and divide by the number of observations. Thus, whether you are calculating a population mean or sample mean is based on whether data represent all possible values (the population) or a subset of all possible values (the sample).

Interpretation of the mean: A mean represents the gravitational center of a distribution. This is where the distribution would *balance*:



This is a reflection of three things that you might want to know. It is a valid reflection of:

1. An individual value drawn at random from the sample.
2. An individual value drawn at random from the population.
3. The population mean.

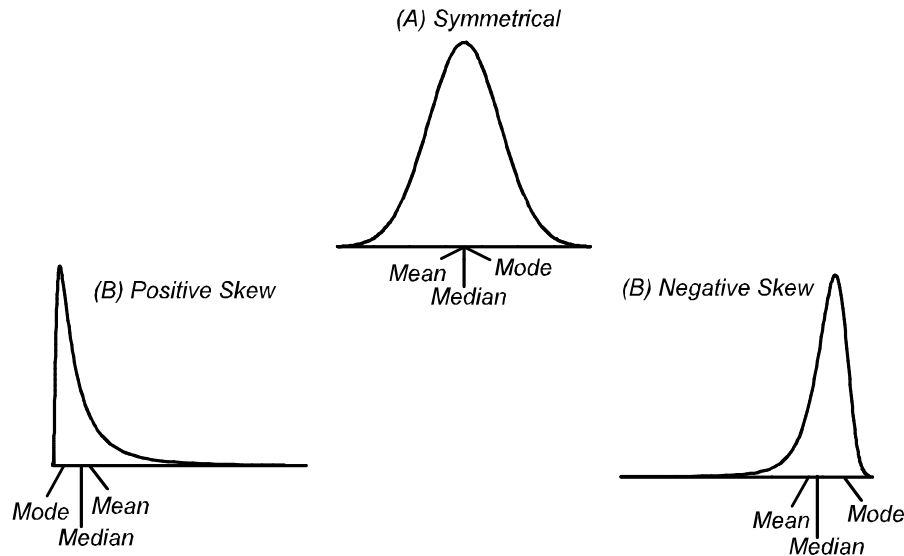
The mean tells you *nothing* about the spread or spread of the distribution.

Reporting statistical results:

- Statistical results should be rounded before reported. In general, the mean should be reported to one decimal place beyond the precision of the data. For example, if AGE is measured in years, the mean age should be reported to the nearest tenth of a year (e.g., 29.0 years).
- To derive one decimal place accuracy, calculations should carry at least three decimal places.
- It is important to indicate units when reporting statistics. The mean age is 29.0 *years*—not merely “29.0.”

Comparison of the Mean, Median, and Mode

The mean, median, and mode are equivalent when the distribution is unimodal and symmetrical. However, with asymmetry, the median is approximately one-third the distance between the mean and mode:



The mean, median, and mode offer different advantages and disadvantages. The mean offers the advantages of familiarity and efficiency. It also has advantages when making inferences about a population mean (covered in Chapter 5.) However, the mean is *markedly* influenced by asymmetry and outliers. Under such circumstances, the median is a more “stable” quantification of the distribution’s center. An often cited example of this advantage come when considering the salary of employees, where the salary of highly paid executives skews the average income toward a misleadingly high value. Another example is the average price of homes (in which case high priced homes skew the data in a positive direction). In such circumstances, the median is less likely to be misinterpreted, and is therefore the preferred measure of central location.

A procedure used to diagnose asymmetry of a distribution is to compare its mean and median. When the mean is greater than the median, the distribution has a positive skew. When the mean is about equal to the median, the distribution is symmetrical. When the mean is less than the median, the distribution has a negative skew:

mean > median ↔ positive skew
mean ≅ median ↔ symmetry
mean < median ↔ negative skew

5-Point Summaries

A **quantile** is any of several ways of dividing the total number of observations into equally sized groups (i.e., each group having the same number of observations). For example, *percentiles* divide a data set into 100 equally sized groups, *quintiles* divide a data set into 5 equally sized groups, and *quartiles* divide a data set into 4 equally sized groups.

A good picture of a distribution can be achieved by dividing it into four equally sized groups. Each dividing point for groups thus divided is called a **quartile**. The first quartile marks the bottom quarter of the data, the second quartile marks the middle of the data, and the third quartile marks the top quarter of the data.

With small data sets, the exact location of quartiles must be interpolated. There are several methods of interpolation. Tukey's (1977, p. 33) method recipe for interpolating quartiles—which he calls **hinges***—is:

- (A) Put the data in rank order
- (B) Divide the data into two groups by finding its median
- (C) Find the median of the low group. This is the first quartile (Q1)
- (D) Find the median of the high group. This is the third quartile (Q3)

For `sample.sav`:

```
5      11      21      24      27 | 28      30      42      50      52
                        median
```

The low group consisting of {5, 11, 21, 24, 27} has a middle value of 21. This is the **first quartile (Q1)**, also called the **lower hinge**. The high group consisting of {28, 30, 42, 50, 52} and has a middle value of 42. This is the **third quartile (Q3)** or **upper hinge**.

The **five-point summary** of a distribution is its

- Q0 = Minimum
- Q1 = First quartile
- Q2 = Median
- Q3 = Third quartile
- Q4 = Maximum

The five-point summary for `sample.sav` is 5, 21, 27.5, 42, 52.

* The term hinge implies this is where the data “folds” upon itself.

Illustrative example #2: Consider this second illustrative data example:

1.47 2.06 2.36 3.43 3.74 3.78 3.94

Here, $n = 7$. The median has a depth of $(7+1)/2 = 4$. This value is 3.43. When the median is actually in the data set (whenever n is odd), it is included in both the low group and high groups when splitting the data.

In this case, the *low group* is:

1.47 2.06 2.36 3.43

The “folding point” (hinge) of this low group is $Q1$: in this instance, average of 2.06 and 2.36, or 2.21.

The *high group* is

3.43 3.74 3.78 3.94

The hinge of this high group is $Q3$: in this instance the average of 3.74 and 3.78, or 3.76.

Thus, the five-point summary is: 1.47, 2.21, 3.43, 3.76, 3.94.

Box-and-Whiskers Plot

Box-and-whiskers plots display five-point summaries and “outside values” in graphical form. A procedure for constructing a boxplot is:

- (A) Determine the 5-point summary for the data.
- (B) Next to an axis, draw a box extending from $Q1$ to $Q3$.
- (C) Inside the box, draw a line that locates the median.
- (D) Calculate the interquartile range as follows:

$$IQR = Q3 - Q1 \quad (3.4)$$

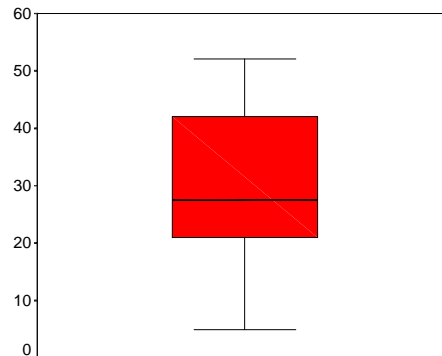
- (E) Calculate fences 1.5 hinge-spreads below and above the hinges:

$$\begin{aligned} \text{Fence}_{\text{lower}} &= Q1 - (1.5)(IQR) \\ \text{Fence}_{\text{upper}} &= Q3 + (1.5)(IQR) \end{aligned} \quad (3.5)$$

- (F) Any values above the upper fence is an **upper outside value**. Any values below the lower fence is a **lower outside value**. Outside values are plotted as separate points on the graph.
- (G) The largest value still inside the upper fence is called the **upper inside value**. The smallest value still inside the lower fence is the **lower inside value**. Drawn whiskers from the upper extent of the box (upper hinge) to the upper inside value. maximum, and from the lower extent of the box (“bottom hinge”) to the minimum.

Illustrative example (sample.sav).

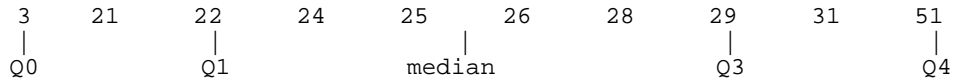
- (A) The 5-point summary for `sample.sav` is 5, 21, 27.5, 42, 52.
- (B) The box extends from 21 to 42.
- (C) A line in the box locates the median at 27.5.
- (D) The $IQR = 42 - 21 = 21$.
- (E) $Fence_{Upper} = 42 + (1.5)(21) = 73.5$. $Fence_{Lower} = 21 - (1.5)(21) = -10.5$.
- (F) No value is more than 73.5, so there are no upper outside values. No value is less than 10.5, so there are no lower outside values.
- (G) Since there are no upper outside values, the upper inside value is the maximum. Since there are no lower outside values, the lower inside is the minimum.
- (H) Whiskers are drawn from the upper hinge to the upper inside value and from the lower hinge to the lower inside value.



Interpretation:

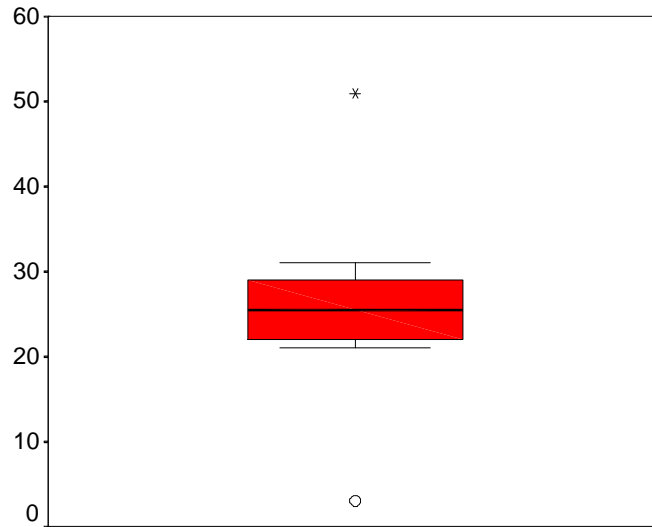
- The **box** locates the middle 50 percent of the data.
- The **median** in the box locates the distribution's center.
- The length of the box (**hinge-spread**) quantifies the distribution's spread.
- The whiskers from tip-to-tip (**whisker-spread**) quantifies distribution's spread.
- When the sample is moderate to large, the distribution's shape can be judged as to **symmetry** in terms of location of the box to the whiskers and length of the whiskers.
- Presence of **outside values** suggests the distribution might have long tails.

Boxplot Example 2: Let us look at a new data set with values:



- (A) The five-point summary is: 3, 22, 25.5, 29, 51.
- (B) The box extends from 22 to 29.
- (C) The median is marked at 25.5.
- (D) The IQR (“hinge-spread”) = $29 - 22 = 7$.
- (E) $Fence_{Upper} = 29 + (1.5)(7) = 39.5$. $Fence_{Lower} = 22 - (1.5)(7) = 11.5$.
- (F) There is one value outside the upper fence (51). There is one value outside of the lower fence (3). These points are plotted separately.
- (G) The highest value still inside the upper fences is 31. The upper whisker is drawn from the upper hinge (29) to the upper inside value (31). The lowest value still inside the lower fence is 21, thus demarcating the lower whisker. The lower whisker extends from the lower hinge (22) to this lower-inside-value (21).

The boxplot looks like this:



Measures of Spread

Variance

Both the variance and standard deviation are based on **deviations** of data points around the distribution's mean. Let d_i represent the deviation of data point i :

$$d_i = x_i - \bar{x} \quad (3.6)$$

Illustrative example (verysmall.sav). Consider the very small data set {1, 3}. This data set has $n = 2$ and $\bar{x} = 2$. Thus, $d_1 = (1 - 2) = -1$ and $d_2 = (3 - 2) = +1$.

The sum of statistical deviations will always equal zero (since they balance around the center of a distribution). Therefore, we will base our measure of variability on the square of the deviations. This makes the signs of the deviations unimportant. (The square of a negative number is a positive). Then sum of the squared deviation is known as a **sum of squares (SS)**:

$$SS = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.7)$$

The sum of squares for the very small data set {1, 3} is: $SS = -1^2 + +1^2 = 1 + 1 = 2$.

The **population variance** (σ^2 ; sigma squared) is the mean sum of squares:

$$s^2 = \frac{SS}{N} \quad (3.8)$$

Assuming `verysmall.sav` represents a full population, $\sigma^2 = 2/2 = 1$.

The **sample variance** (s^2) is:

$$s^2 = \frac{SS}{(n - 1)} \quad (3.9)$$

Illustrative example (sample.sav). For the illustrative data set `sample.sav`, $n = 10$, $\bar{x} = 29.0$, and $SS = (21-29)^2 + (42-29)^2 + (5-29)^2 + (11-29)^2 + (30-29)^2 + (50-29)^2 + (28-29)^2 + (27-29)^2 + (24-29)^2 + (52-29)^2 = 2134$. Thus, $s^2 = \frac{2134}{(10-1)} = 237.1111$.

Interpretation: Because variance carries units *squared* (e.g., years²), it is rarely interpreted directly. Instead, we take square root of the variance, which is called the standard deviation.

Standard Deviation

The **population standard deviation** (σ) is the square root of the population variance:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{N}} \quad (3.10)$$

For `verysmall.sav` ($N = 2$, $\sigma^2 = 1$), the standard deviation $\sigma = \sqrt{1} = 1$.

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

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} \quad (3.11)$$

For `sample.sav` ($n = 10$, $s^2 = 237.111$), $s = \sqrt{(237.1111 \text{ years}^2)} = 15.4$ years.

Interpretation: By rooting the variance, units of “years squared” revert to “years.” This helps in interpreting the standard deviation as a measure of spread. Still, interpreting a standard deviation is not as easy as interpreting a mean.

One thing to keep in mind is that distributions with big standard deviations have more variability than distributions with small standard deviations. For example, if the standard deviation of the age in one population is 15 years and the standard deviation in a different population is 2 years, the first population has much more age variability than the second population.

But how do we interpret a single standard deviation? One way to interpret a single standard deviation is to indicate the percent of data that falls within a specified number of standard deviations of the mean. We have two rules for applying this approach.

The first rule for interpreting standard deviations applies to **normal** distributions.[†] When this is the case:

- 68% of values lie within 1 standard deviation of the mean. These boundaries are $\mu \pm \sigma$.
- 95% of values lie within 2 standard deviations of the mean. These boundaries are $\mu \pm 2\sigma$.
- Nearly all values lie within 3 standard deviations of the mean. These boundaries are $\mu \pm 3\sigma$.

For example, if ages is normally distributed with a mean (μ) of 30 and standard deviation (σ) of 10, then 68% of the population will be in the age range 30 ± 10 (20 to 40), 95% will be in the age range 30 ± 20 (10 to 50), and nearly all will be in the age range 30 ± 30 (0 to 60).

For distributions that are *not* normal, **Chebyshev's rule** applies, which states

[†] The normal distribution is introduced in the next chapter. For now let us note that a normal distribution is symmetrical and bell-shaped.

- *At least 75%* of the values lie within 2 standard deviations from the mean
- *At least seven-eighths* lie within 3 standard deviations from the mean

For a population with a mean age of 30 years and standard deviation of 10 years, for instance, we know with that *at least 75%* of the values lie in the range 30 ± 20 (10 to 50) and at least seven-eighths lie in the range 30 ± 30 (0 to 60) .

Interquartile Range (IQR)

The interquartile range (IQR) (introduced earlier) is also an excellent measure of spread. Recall that the IQR is the difference in quartiles:

$$IQR = Q3 - Q1 \quad (3.12)$$

For the illustrative data, $Q1 = 21$ and $Q3 = 42$. Thus, $IQR = 42 - 21 = 21$.

Interpretation: The interquartile range is relatively unaffected by outliers and is relatively “robust.”[‡] Groups with large *IQRs* have greater variability than groups with smaller *IQRs*. A good way to compare group *IQRs* is with side-by-side boxplot!

SPSS: Statistics in this chapter are computed with Analyze > Descriptive Statistics > Explore and selecting the variable you want to explore. SPSS specifics and peculiarities are documented in lab.

[‡] “Robustness” implies resistant to the influence of outliers and distributional asymmetry.