

## 3: Summary Statistics

### Notation

Let's start by introducing some **notation**. Consider the following small data set:

21      42      5      11      30      50      28      27      24      52

- The symbol  $n$  represents the sample size ( $n = 10$ ).
- The **capital letter**  $X$  denotes the **variable**. (This is more compact than using descriptive labels such as AGE, for instance.)
- $x_i$  represents the  $i^{\text{th}}$  value of variable  $X$ . For the data that is above,  $x_1 = 21$ ,  $x_2 = 42$ , and so on.
- The symbol  $\Sigma$  (“sigma”) indicates that you should add all the values for the variable. For example,  $\Sigma x_i = x_1 + x_2 + \dots + x_{10} = 21 + 42 + 5 + 11 + 30 + 50 + 28 + 27 + 24 + 52 = 290$ .

### Measures of Central Location

#### Mean

When used without specification, “mean” refers to the **arithmetic average** of a data set. To calculate the mean, add all the values in the data set and divide by the number of observations.

We distinguish between the population mean and the sample mean. The **population mean**  $\mu$  (“mu”) is:

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$$

Using more compact notation:

$$\mu = \frac{1}{N} \sum x_i$$

where  $\Sigma x_i$  represents the sum of all values in the population and  $N$  represents the population size.

**Illustrative example.** The data set `populati.sav` is a complete listing of a population. It has 600 individuals. The sum of age values in this population is 17,703. Therefore,  $\mu = \frac{17.703}{600} = 29.5$  years.

The **sample mean** ( $\bar{x}$ ; pronounced “x bar”) is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

In more succinct notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

**Illustrative example.** The data set `sample.sav` on the prior page has  $\sum x_i = 290$  and  $n = 10$ .

Therefore  $\bar{x} = \frac{1}{10}(290) = 29$  years.

The formulas for population mean  $\mu$  and sample mean  $\bar{x}$  are [almost] identical. Whether you are calculating a population mean or sample mean is based on whether data represent the entire population or a sample. *In practice you should always assume you are working with a sample and not the entire population.*

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

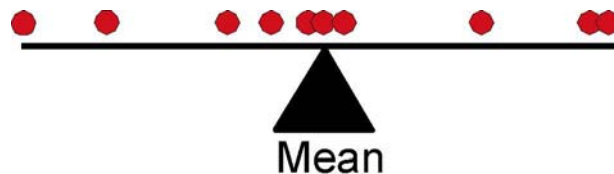


Fig:sumstat1.ai

The mean is like a seesaw. A small child can sit farther from the center of a seesaw in order to balance a large child sitting closer to the center. Similarly, a single value sitting far off the center of a distribution can have a profound effect on the mean.

The mean also tells you three things that you might want to know. It is a 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.

**Reporting the mean:** The mean should be rounded before it is reported. In general, report the mean (and other summary statistics) with one decimal more than the recording of the initial data. For example, if `AGE` is recorded in years, report the mean age to the nearest tenth of a year. Included units of measure with your summary statistics. The mean of our sample is 29.0 years.

## Median

The median is a different kind of average. It is the value that is greater than or equal to half of the values in the data set. It is the *middle point* of the data set.

The median is easy to determine if you first order the data from low to high. This is an **ordered array**. The distance from the lowest value in the ordered array to where a given point is situated in the array is called **depth**. The median has a depth of  $\frac{n+1}{2}$ .

- When  $n$  is even, this will fall between two values. Average these values to get the median.
- When  $n$  is odd, the depth  $\frac{n+1}{2}$  will fall right on a value. This will be your median.

**Illustrative example (populati.sav).** Since  $n = 10$  in the illustrative data, the median has a depth of  $(10 + 1) / 2 = 5.5$ . This places the median between the data points 27 and 28.

5	11	21	24	27	28	30	42	50	52
				↑					
				median					

When the depth of the median falls between two values, we average them to find the median. The median in this data set is  $(27 + 29) / 2 = 27.5$ .

**Illustrative example (tumorsize.sav):** MRI based estimates of the volume of 5 benign tumors (cubic centimeters) are:

4	7	8	11	12
		↑		
		median		

The median has a depth of  $(5 + 1) / 2 = 3$ . Therefore, the median is 8.

The median is *resistant* to outliers. For instance, if the biggest value in the above set had been 120 instead of 12, the median would still be 8.

## Mode

The mode is the most frequently occurring value in a data set.

AGE	Freq	%	
3	2	0.3%	
4	9	1.4%	
5	28	4.3%	
6	37	5.7%	
7	54	8.3%	
8	85	13.0%	
<b>9</b>	<b>94</b>	<b>14.4%</b>	← Mode (most frequent value)
10	81	12.4%	
11	90	13.8%	
12	57	8.7%	
13	43	6.6%	
14	25	3.8%	
15	19	2.9%	
16	13	2.0%	
17	8	1.2%	
18	6	0.9%	
19	3	0.5%	
Total	654	100.0%	

The mode in this dataset is 9 since it appears more frequently than any other value.

Note: The mode is unreliable in all but large data sets.

## 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:

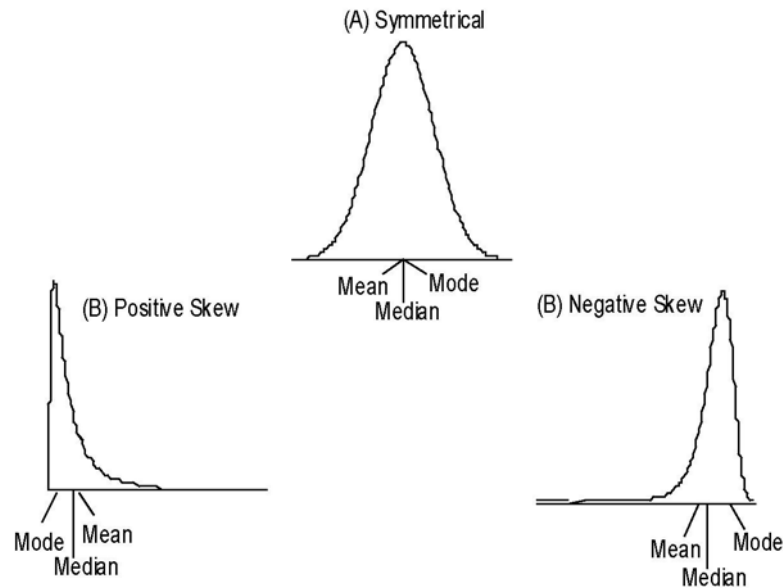


Fig:sumstats2.ai

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. On the downside, the mean is markedly influenced by extreme skewness and outliers. Under circumstances of extreme skewness, the median is a more “stable.” 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.

You can judge the asymmetry of a distribution by comparing 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  $\cong$  median ↔ symmetry  
mean < median ↔ negative skew

In summary, the mean is preferred when data are symmetrical and do not have outliers. In other instances, the median is often preferred measure of central location.

# Measures of Spread

## **Range**

One simple way to measure spread is to provide the smallest (**minimum**) and largest (**maximum**) values in the data set. The difference of these values is the sample **range**.

$$\text{range} = \text{maximum} - \text{minimum}$$

The range of the sample in the stemplot below is  $52 - 5 = 47$ .

```
| 0 | 5
| 1 | 1
| 2 | 1478
| 3 | 0
| 4 | 2
| 5 | 02
×10
```

The sample range is a crude and biased measure of spread. If you use the range, you should supplement it with at least one other measure of spread.

## **Standard Deviation and Variance**

The **standard deviation** is the most common measure of spread. A **deviation** is the difference between a value and the mean:  $x_i - \bar{x}$ . We take each deviation in a data set and square it:  $(x_i - \bar{x})^2$ . This makes the negative or positive sign associated with the deviation unimportant. We then sum the squared deviations to derive a statistic known as the **sum of squares (SS)**:

$$SS = \sum (x_i - \bar{x})^2$$

The **population variance** is the average sum of squares:  $\sigma^2 = \frac{SS}{N}$

The **sample variance** is  $s^2 = \frac{SS}{n-1}$

The sample variance divides the sum of squares by  $n - 1$  instead of  $n$ . When  $n$  is large,  $n - 1 \approx n$ , so the numerical results from the two formulas will be similar. However, when  $n$  is small, the sample variance formula will give a bigger result than the population variance formula. This is necessary to derive an unbiased estimate of the population variance. The number  $n-1$  is called the **degree of freedom** of the variance. We lose one degree of freedom when using  $\bar{x}$  to estimate  $\mu$ .

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

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

The one-step formula for the standard deviation is

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

**Illustrative example.** Recall the 10 ages in SAMPLE.SAV:

21      42      5      11      30      50      28      27      24      52

The sample mean  $\bar{x}$  is 29.0. Keep things tidy when calculating the standard deviation:

Values	Deviations	Squared Deviation
21	21 - 29 = -8	-8 <sup>2</sup> = 64
42	42 - 29 = 13	13 <sup>2</sup> = 169
5	5 - 29 = -24	-24 <sup>2</sup> = 576
11	11 - 29 = -18	-18 <sup>2</sup> = 324
30	30 - 29 = 1	1 <sup>2</sup> = 1
50	50 - 29 = 21	21 <sup>2</sup> = 441
28	28 - 29 = -1	-1 <sup>2</sup> = 1
27	27 - 29 = -2	-2 <sup>2</sup> = 4
24	24 - 29 = -5	-5 <sup>2</sup> = 25
52	52 - 29 = 23	23 <sup>2</sup> = 529
Sum of values = 290	Sum of Deviations = 0	Sum of Squares = 2134

The sample variance is  $s^2 = \frac{SS}{n-1} = \frac{2134}{10-1} = 237.1111$  years<sup>2</sup>. The standard deviation

$s = \sqrt{s^2} = \sqrt{237.1111} = 15.398$  years. Round this to 15.4 years.

**Notes:**

1. Always report the standard deviation when you report the mean.
2. When all values in the data set are the same, there is no spread and standard deviation  $s = 0$ . In all other instances, the standard deviation is a positive value. When  $s$  is large, there is a lot of spread in the data.
3. Standard deviation  $s$  has the same units as the data. This is not true of variance  $s^2$ , which has “units squared.”
4. The standard deviation is sensitive to outlier and skews.
5. When the data are Normal (“bell-shaped”) 68% of values lie within 1 standard deviation of the mean, 95% of values lie within 2 standard deviations of the mean, and 99.7% of values lie within 3 standard deviations of the mean.
6. For distributions that are *not* Normal, **Chebychev’s rule** says at least  $\frac{3}{4}$  of the values lie within 2 standard deviations of the mean.

## The Interquartile Range

The **first quartile (Q1)** is one-quarter up the ordered array. Q1 is also called the 25<sup>th</sup> percentile because it is larger than 25% of the observations. The **third quartile (Q3)** is three-quarters up the array. Q3 is also called the 75<sup>th</sup> percentile, because it is larger than 75% of the observations.

When the data set is large ( $n \geq 100$ ), it is easy to find the quartiles. With small data sets, the exact location of quartiles must be interpolated. We use **Tukey's hinges** to interpolate quartiles. To determine hinges:

- (A) Put the data in rank order and locate the median of the data set.
- (B) Divide the data set into two groups: a low group and a high group. When  $n$  is odd, the median should be placed in both groups.
- (C) Find the middle value (median) of the low group. This is Q1.
- (D) Find the middle value (median) of the high group. This is Q3.

Once the quartiles have been determined, calculate the **inter-quartile range (IQR)**.

$$\text{IQR} = Q3 - Q1$$

A good summary of locations in the distribution is provided by the points that divide the data it into four equally-sized groups. This is the **5-point summary** is made of:

- Q0  $\equiv$  Quartile 0 (the *minimum*)
- Q1  $\equiv$  Quartile 1 (bigger than 25% of the data points)
- Q2  $\equiv$  Quartile 2 (the median)
- Q3  $\equiv$  Quartile 3 (bigger than 75% of the data points)
- Q4  $\equiv$  Quartile 4 (the *maximum*)

**Illustrative example #1.** The ordered array for data in this sample is:

5      11      21      24      27    |    28      30      42      50  
52

The low group of {5, 11, 21, 24, 27} has a middle value (Q1) of 21. The high group of {28, 30, 42, 50, 52} has a middle value (Q3) of 42. Thus, the five-point summary for these data is 5, 21, 27.5, 42, 52. The IQR  $Q3 - Q1 = 42 - 21 = 21$ .

**Illustrative example #2.** Consider this new ordered array of  $n = 7$ .

1.47      2.06      2.36      3.43      3.74      3.78      3.94

The median is 3.43. The low group consists of {1.47, 2.06, 2.36, 3.43}. Notice that the median appears in both the low group and the high group Q1 is the average of the adjacent middle values of 2.06 and 2.36, which is  $(2.06 + 2.36) \div 2 = 2.21$ . The high group is {3.43, 3.74, 3.78, 3.94}. Q3 is the average of the adjacent middle values of 3.74 and 3.78, which is  $(3.74 + 3.78) \div 2 = 3.76$ . The five point summary is 1.47, 2.21, 3.43, 3.76, 3.95. The IQR  $= Q3 - Q1 = 3.76 - 2.21 = 1.55$

## **Boxplots**

### **Box-and-Whiskers Plot**

Box-and-whiskers plots display five-point summaries and potential outliers in graphical form. To construct a boxplot

1. Determine the **5-point summary** for the data.
2. Draw on graph paper a box extending from Q1 to Q3.
3. Inside the box, draw a line that locates the **median**.
4. Calculate the **interquartile range** ( $IQR = Q3 - Q1$ )
5. Calculate **fences** 1.5 hinge-spreads below and above the hinges:
  - a. The lower fence  $Fence_{Lower} = Q1 - (1.5)(IQR)$
  - b. The upper fence  $Fence_{Upper} = Q3 + (1.5)(IQR)$
  - c. Do *not* plot these fences.
6. Any value above the upper fence is an **upper outside value**. Any values below the lower fence is a **lower outside value**. Plot these values separate points on the graph.
7. 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**. Draw 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).** Recall the ordered array for AGE in sample.sav

5      11      21      24      27      28      30      42      50      52

1. The 5-point summary is 5, 21, 27.5, 42, 52 (see prior analysis)
2. The box extends from 21 to 42.
3. A line in the box locates the median at 27.5.
4. The  $IQR = 42 - 21 = 21$ .
5. Fences
  - a.  $F_L = 21 - (1.5)(21) = -10.5$
  - b.  $F_U = 42 + (1.5)(21) = 73.5$ .
6. 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.
7. Since there are no outside values, the upper inside value is the maximum (52) and the lower inside is the minimum (5). Whiskers are drawn from the hinges to the inside values.

The boxplot is:

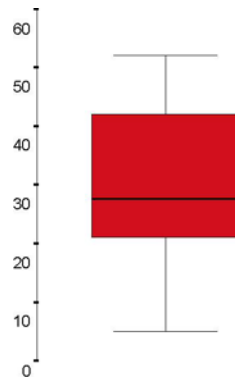


Fig: boxplot1.ai

**Interpretation of boxplots:** Boxplots show less detail than stemplots, but still provide insight into the central location, spread, and shape of a distribution. When you look at a boxplot, consider the following elements:

- **Central location:** The line in the box locates the median. In addition, the box locates the middle 50 percent of the data.
- **Spread:** The length of the box is called the **hinge-spread**. This corresponds to the IQR is a good quantifier of the distribution's spread. In addition, The whiskers from tip-to-tip (the "whisker-spread") quantifies the distribution's spread. The maximum and minimum are visible as well.
- **Shape:** Shape is difficult to judge except when the sample is large, in which case symmetry or lack of symmetry will be visible.

**Illustrative Example #2:** Let us look at a new data set with values:

3      21      22      24      25      26      28      29      31      51

1. The five-point summary is: 3, 22, 25.5, 29, 51
2. The box extends from 22 to 29
3. The median is marked at 25.5
4. The IQR (“hinge-spread”) =  $29 - 22 = 7$
5.  $Fence_{Upper} = 29 + (1.5)(7) = 39.5$ .  $Fence_{Lower} = 22 - (1.5)(7) = 11.5$ .
6. There is one value outside the upper fence (51). There is one value outside of the lower fence (value of 3). These points are plotted separately.
7. 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:

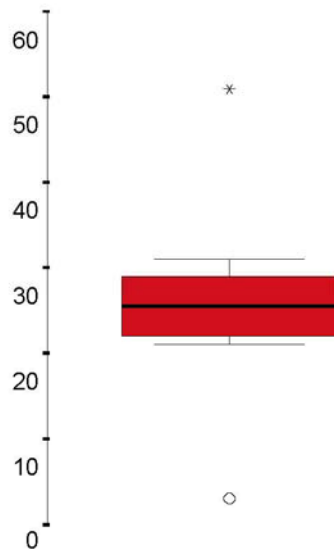


Fig: boxplot2.ai