

# Central Tendency, Variation, Normal Distribution

CCDATSCL | Data Science

## Descriptive Statistics

Descriptive statistics describe variables in a number of ways. It also allows us to quantify precisely these descriptions of the data.

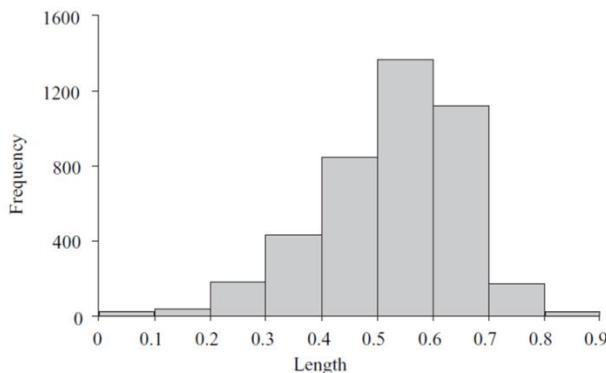


Figure 5.1. Histogram of variable Length

## Central Tendency

In addition to describing the shape of the distribution of a sample or population of measurements, we also describe the data set's central tendency.

A measure of central tendency represents the center or middle of the data. Sometimes we think of a measure of central tendency as a typical value.

1. Mean
2. Mode
3. Median

## Population Mean

One important measure of central tendency for a population of measurements is the population mean.

The population mean, which is denoted  $\mu$  and pronounced mew, is the average of the population measurements.

Class	Class Size	DS ClassSizes
Business Law	60	
Finance	41	
International Studies	15	
Management	30	
Marketing	34	

The mean  $\mu$  of this population of class sizes is

$$\mu = \frac{60 + 41 + 15 + 30 + 34}{5} = \frac{180}{5} = 36$$

## Population Parameters

In order to understand how to estimate a population mean, we must realize that the population mean is a population parameter.

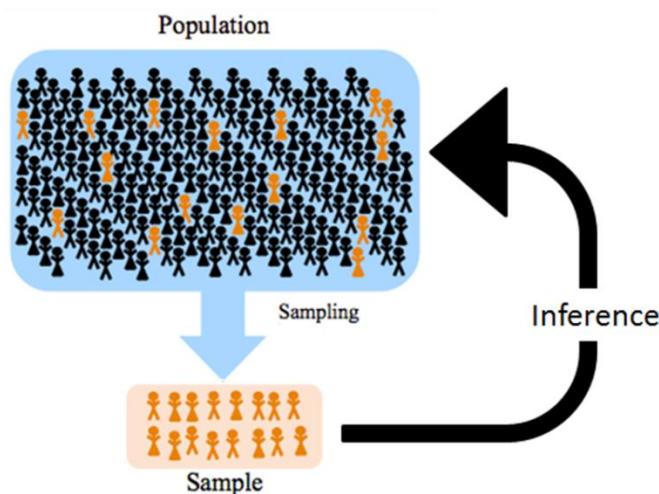
A population parameter is a number calculated using the population measurements that describes some aspect of the population. That is, a population parameter is a descriptive measure of the population.

## Estimating Population Parameters

The simplest way to estimate a population parameter is to make a point estimate, which is a one-number estimate of the value of the population parameter.

It should be an educated guess based on sample data.

One sensible way to find a point estimate of a population parameter is to use a sample statistic.



## Sample Statistic

A sample statistic is a number calculated using the sample measurements that describes some aspect of the sample. That is, a sample statistic is a descriptive measure of the sample.

## Sample Mean

The sample statistic that we use to estimate the population mean is the sample mean, which is denoted as  $\bar{x}$  (pronounced x bar) and is the average of the sample measurements.

The sample mean  $\bar{x}$  is defined to be

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

and is the point estimate of the population mean  $\mu$ .

## Example: sample mean

TABLE 3.1 A Sample of 50 Mileages 

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

$$\sum_{i=1}^{50} x_i = x_1 + x_2 + \dots + x_{50} = 30.8 + 31.7 + \dots + 31.4 = 1578$$

Therefore, the mean of the sample of 50 mileages is

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50} = \frac{1578}{50} = 31.56$$

This point estimate says we estimate that the **mean mileage** that would be obtained by all of the new midsize cars that will or could potentially be produced this year is 31.56 mpg. Unless we are extremely lucky, however, there will be sampling error.

That is, the point estimate 31.56 mpg, which is the average of the sample of fifty randomly selected mileages, will probably not exactly equal the population mean  $\mu$ , which is the average mileage that would be obtained by all cars.

## Sample Median

Intuitively, the median divides a population or sample into two roughly equal parts.

Consider a population or a sample of measurements, and arrange the measurements in increasing order. The median,  $M_m$ , is found as follows:

- 1 If the number of measurements is odd, the median is the middlemost measurement in the ordering.
- 2 If the number of measurements is even, the median is the average of the two middlemost measurements in the ordering.

## Example: sample median

For example, recall that Chris's five classes have sizes 60, 41, 15, 30, and 34. To find the median of the population of class sizes, we arrange the class sizes in increasing order as follows:

15    30    34    41    60

As another example, suppose that in the middle of the semester Chris decides to take an additional class—a sprint class in individual exercise. If the individual

exercise class has 30 students, then the sizes of Chris's six classes are (arranged in increasing order):

15    30    30    34    41    60

Because the number of classes is even, the median of the population of class sizes is the average of the two middlemost class sizes, which are circled. Therefore, the median is students  $(30 + 34)/2 = 32$ .

Note that, although two of Chris's classes have the same size, 30 students, each observation is listed separately (that is, 30 is listed twice) when we arrange the observations in increasing order.

## Sample Mode

A third measure of the central tendency of a population or sample is the mode.

The mode,  $M_o$ , of a population or sample of measurements is the measurement that occurs most frequently.

## Example: sample mode

For example, the mode of Chris's six class sizes is **30**. This is because more classes (two) have a size of 30 than any other size.

Sometimes the highest frequency occurs at more than one measurement. When this happens, two or more modes exist. When exactly two modes exist, we say the data are **bimodal** and may be reported as {first mode, second mode} or (first mode + second mode)/2.

When more than two modes exist, we say the data are **multimodal**.

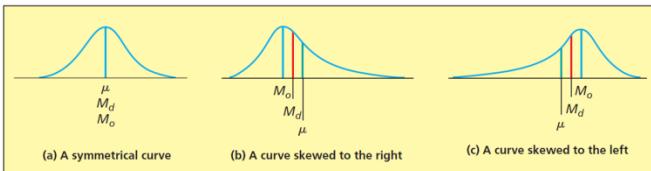
If data are presented in classes (such as in a frequency or percent histogram), the class having the highest frequency or percent is called the **modal class**.

## Comparing the mean, median, and mode

Often we construct a histogram for a sample to make inferences about the shape of the sampled population.

When we do this, it can be useful to **smooth out** the histogram and use the resulting relative frequency curve to describe the shape of the population. Relative frequency curves can have many shapes. Three common shapes are illustrated below:

FIGURE 3.3 Typical Relationships among the Mean  $\mu$ , the Median  $M_d$ , and the Mode  $M_o$

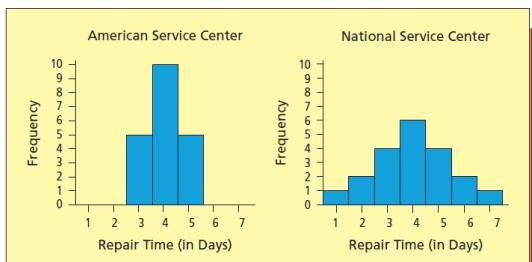


- Part (a)** - of this figure depicts a population described by a symmetrical relative frequency curve. For such a population, the mean ( $\mu$ ), median ( $M_d$ ), and mode ( $M_o$ ) are all equal. Note that in this case all three of these quantities are located under the highest point of the curve. It follows that when the frequency distribution of a sample of measurements is approximately **symmetrical**, then the sample mean, median, and mode will be nearly the same.
- Part (b)** - depicts a population that is skewed to the right. Here the population mean is larger than the population median, and the population median is larger than the population mode
- Part (c)** - depicts a population that is skewed to the left. Here the population mean is smaller than the population median, and the population median is smaller than the population mode.

## Variation

In addition to estimating a population's central tendency, it is important to estimate the variation of the population's individual values.

FIGURE 3.13 Repair Times for Personal Computers at Two Service Centers



## Case study: variation of repair times

Each portrays the distribution of 20 repair times (in days) for personal computers at a major service center. Because the mean (and median and mode) of each distribution equals four days, the measures of central tendency do not indicate any difference between the American and National Service Centers.

However, the repair times for the American Service Center are clustered quite closely together, whereas the repair times for the National Service Center are spread farther apart (the repair time might be as little as one day, but could also be as long as seven days).

Therefore, we need measures of variation to express how the two distributions differ.

## Range

One way to measure the variation of a set of measurements is to calculate the range.

Consider a population or a sample of measurements. The **range** of the measurements is the largest measurement minus the smallest measurement.

In Figure 3.13, the smallest and largest repair times for the American Service Center are three days and five days; therefore, the range is  $5 - 3 = 2$  days.

On the other hand, the range for the National Service Center is  $7 - 1 = 6$  days. The National Service Center's larger range indicates that this service center's repair times exhibit more variation.

In general, the range is not the best measure of a data set's variation.

In general, to fully describe a population's variation, it is useful to estimate intervals that contain different percentages (for example, 70 percent, 95 percent, or almost 100 percent) of the individual population values.

## Population Variance and STDev

### The Population Variance and Standard Deviation

The population variance  $\sigma^2$  (pronounced *sigma squared*) is the average of the squared deviations of the individual population measurements from the population mean  $\mu$ .

The population standard deviation  $\sigma$  (pronounced *sigma*) is the positive square root of the population variance.

The more spread out the population measurements, the larger is the population variance, and the larger is the population standard deviation. And vice versa.

When a population is too large to measure all the population units, we estimate the population variance and the population standard deviation by the sample variance and the sample standard deviation.

## Case study: Chris' class

For example, consider again the population of Chris's class sizes this semester. These class sizes are 60, 41, 15, 30, and 34. To calculate the variance and standard deviation of these class sizes, we first calculate the population mean to be

$$\mu = \frac{60 + 41 + 15 + 30 + 34}{5} = \frac{180}{5} = 36$$

Next, we calculate the deviations of the individual population measurements from the population mean  $\mu = 36$  as follows:

$$(60 - 36) = 24 \quad (41 - 36) = 5 \quad (15 - 36) = -21 \quad (30 - 36) = -6 \quad (34 - 36) = -2$$

Then we compute the sum of the squares of these deviations:

$$(24)^2 + (5)^2 + (-21)^2 + (-6)^2 + (-2)^2 = 576 + 25 + 441 + 36 + 4 = 1082$$

Finally, we calculate the population variance  $\sigma^2$ , the average of the squared deviations, by dividing the sum of the squared deviations, 1,082, by the number of squared deviations, 5. That is,  $\sigma^2$  equals  $1,082/5 = 216.4$ . Furthermore, this implies that the population standard deviation  $\sigma$  (the positive square root of  $\sigma^2$ ) is  $\sqrt{216.4} = 14.71$ .

Class	Class Size	DS ClassSizes
Business Law	60	
Finance	41	
International Studies	15	
Management	30	
Marketing	34	

## Variance vs Standard Deviation

- Standard deviation looks at how spread out a group of numbers is from the mean, by looking at the square root of the variance.
- The variance measures the average degree to which each point differs from the mean—the average of all data points.
- Since, standard deviation is the square root of variance, **the value is in the same units as the mean**.
- A normal distribution with mean = 10 and sd = 12 is exactly the same as a normal distribution with mean = 10 and variance = 144.

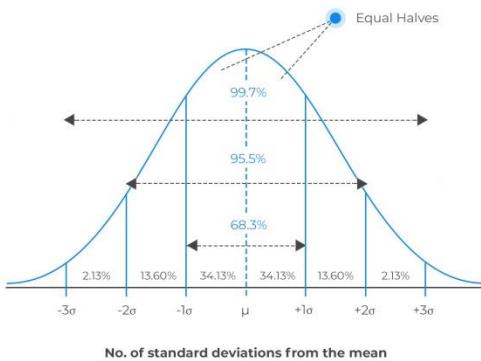
## Notations

	Sample Size	Mean	Standard Deviation	Variance
Population	N	$\mu$	$\sigma$	$\sigma^2$
Sample	n	$\bar{x}$	s	$s^2$

## The Normal Distribution

A normal distribution has a **bell-shaped density curve** described by its mean and standard deviation . The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation.

## Shape of the normal distribution



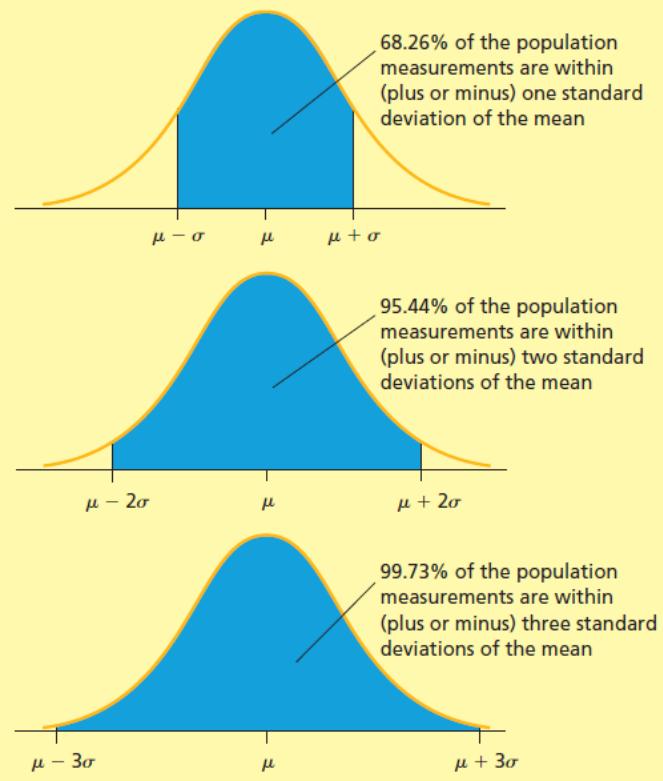
## Empirical Rule

One type of relative frequency curve describing a population is the normal curve.

The normal curve is a symmetrical, bell-shaped curve and is illustrated on the right.

If a population is described by a normal curve, we say that the population is normally distributed.

### (a) The Empirical Rule



### The Empirical Rule for a Normally Distributed Population

- If a population has mean  $\mu$  and standard deviation  $\sigma$  and is described by a normal curve, then, as illustrated in Figure 3.14(a),
- 1 68.26 percent of the population measurements are within (plus or minus) one standard deviation of the mean and thus lie in the interval  $[\mu - \sigma, \mu + \sigma] = [\mu \pm \sigma]$
  - 2 95.44 percent of the population measurements are within (plus or minus) two standard deviations of the mean and thus lie in the interval  $[\mu - 2\sigma, \mu + 2\sigma] = [\mu \pm 2\sigma]$
  - 3 99.73 percent of the population measurements are within (plus or minus) three standard deviations of the mean and thus lie in the interval  $[\mu - 3\sigma, \mu + 3\sigma] = [\mu \pm 3\sigma]$

**FIGURE 6.4** How the Mean  $\mu$  and Standard Deviation  $\sigma$  Affect the Position and Shape of a Normal Probability Curve

