Measures of Central Tendency and Measures of Location

Measures of central tendency are numbers that tell us where the majority of values in the distribution are located. Also, we may consider these measures to be the center of the probability distribution from which the data were sampled.

The Arithmetic Mean

The arithmetic mean is the sum of the individual values in a data set divided by the number of values in the data set. We can compute a mean of both a finite population and a sample. For the mean of a finite population (denoted by the symbol μ), we sum the individual observations in the entire population and divide by the population size, N. When data are based on a sample, to calculate the sample mean (denoted by the symbol ($\bar{x}$) we sum the individual observations in the sample and divide by the number of elements in the sample, n. The sample mean is the sample analog to the mean of a finite population. Formulas for the population (4.1a) and sample means (4.1b) are shown below; also see Table 4.1.

TABLE 4.1. Calculation of Mean (Small Population, $N = 5$)

Index (i)	X
1	70
2	80
3	95
4	100
5	125
Σ	470

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{470}{5} = 94$$

Population mean (μ):

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (4.1a)$$

where X_i are the individual values from a finite population of size N.

Sample mean ($\bar{X}$):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (4.1b)$$

where Xi are the individual values of a sample of size n. The population mean (and also the population variance and standard deviation) is a parameter of a distribution. Means, variances, and standard deviations of finite populations are almost identical to their sample analogs.

Statisticians generally use the arithmetic mean as a measure of central tendency for numbers that are from a ratio scale (e.g., many biological values, height, blood sugar, cholesterol), from an interval scale (e.g., Fahrenheit temperature or personality measures such as depression), or from an ordinal scale (high, medium, low). The values may be either discrete or continuous; for example, ranking on an attitude scale (discrete values) or blood cholesterol measurements (continuous).

It is important to distinguish between a continuous scale such as blood cholesterol and cholesterol measurements. While the scale is continuous, the measurements we record are discrete values. For example, when we record a cholesterol measurement of 200, we have converted a continuous variable into a discrete measurement. The speed of an automobile is also a continuous variable. As soon as we state a specific speed, for example, 60 miles or 100 kilometers per hour, we have created a discrete measurement. This example becomes clearer if we have a speedometer that gives a digital readout such as 60 miles per hour.

The mean of grouped data:

For large data sets (e.g., more than about 20 observations when performing calculations by hand), summing the individual numbers may be impractical, so we use grouped data. When using a computer, the number of values is not an issue at all. The procedure for calculating a mean is somewhat more involved for grouped data than for ungrouped data. First, the data needs to be placed in a frequency table. We then apply Formula below, which specifies that the midpoint of each class interval (X) is multiplied by the frequency of observation in that class. The mean using grouped data is:

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

where Xi is the midpoint of the ith interval and fi is the frequency of observations in the ith interval. In order to perform the calculation specified by Formula, first we need to place the data from Table 4.2 in a frequency table, as shown in Table 4.3. For a review of how to construct such a table. From Table 4.3, we can see that $\sum fX = 9715$, $\sum f = n = 100$, and that the mean is estimated as 97.2 (rounding to the nearest tenth).

TABLE 4.2. Plasma Glucose Values (mg/dl) for a Sample of 100 Adults, Aged 20–74 Years

74	82	86	88	90	91	94	97	106	123
75	82	86	89	90	92	95	98	108	124
77	82	87	89	90	92	95	99	108	128
78	83	87	89	90	92	95	99	113	132
78	83	87	89	90	92	95	99	113	134
78	83	88	89	90	93	95	99	115	140
80	83	88	89	90	93	96	100	118	151
81	85	88	89	90	94	96	101	120	153
81	86	88	89	90	94	97	104	121	156
81	86	88	90	91	94	97	105	122	164

TABLE 4.3. Calculation of a Mean from a Frequency Table (Using Data from Table 4.2)

Class Interval	Midpoint (x)	f	fx
160–169	165.5	1	165.50
150–159	155.5	3	466.50
140–149	145.5	1	145.50
130–139	134.5	2	269.00
120–129	124.5	6	747.00
110–119	114.5	4	458.00
100–109	104.5	7	731.50
90–99	94.5	37	3496.50
80–89	84.5	33	2788.50
70–79	74.5	6	447.00
		100	9715.00

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{9715.0}{100} = 97.15$$

Non Symmetric Distributions

Example 3

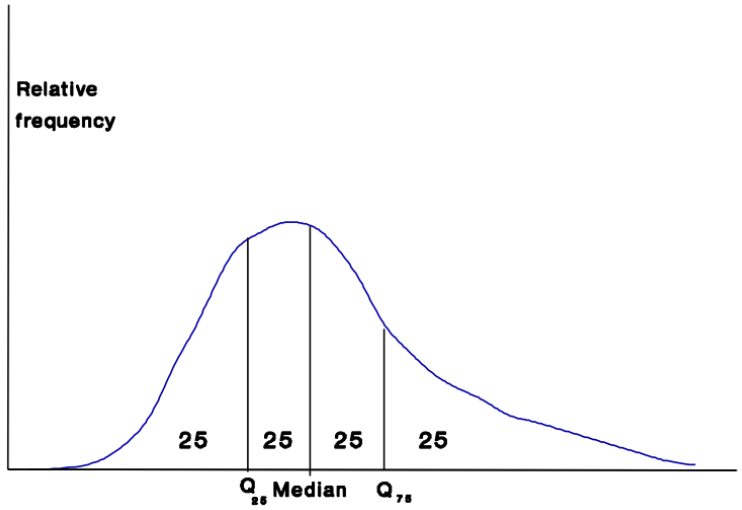
The number of days spent in hospital by 17 subjects after an operation, arranged in increasing size, were: 3 4 4 6 8 8 8 10 10 12 14 14 17 25 27 37 42 The distribution is not symmetric (asymmetric)

because low values are closer together and often repeated, compared with the string of high values. The mean is 14.6 days. This is not in the centre of the distribution.

The median and quartiles of a distribution

The median refers to the 50% point in a frequency distribution of a population. The median is an alternative measure of central value that works better for such a skewed distribution. It is the value which halves the distribution, with 50% of the observations below it and 50% above. The three values which divide the distribution into quarters are called the quartiles. The middle quartile is the median, and the distance between the lower quartile and the upper quartile, called the inter-quartile range, is used as a measure of spread.

For a distribution with a large number of observations the quartiles are most easily found from the cumulative relative frequency distribution (such as in the example above), by reading off the values that correspond to 25%, 50%, and 75%.



For a smaller number of observations the median can be found directly by arranging the observations in order from the lowest to the highest value and striking off values at both ends until only one or two remain. If one, this value is the median; if two the median is half way between them. The median is then used to divide the data into two halves and the medians of each of the halves found in the same way - these are the upper and lower quartiles. (If the median is the single central value, include it in each half).

For example 3, the median stay in hospital is 10 days. The 1st and 3rd quartiles are 8 and 17 days.

The median of grouped data

When data are grouped in a frequency table, the median is an estimate because we are unable to calculate it precisely. Thus, the following Formula is used to estimate the median from data in a frequency table 4.3:

median = lower limit of the interval + i(0.50n – cf)

where i = the width of the interval

n = sample size (or N = population size)

cf = the cumulative frequency below the interval that contains the median

TABLE 4.4. Calculation of a Mean from a Frequency Table (Using Data from Table 4.2)

Class Interval	*f*	*cf*
160–169	1	100
150–159	3	99
140–149	1	96
130–139	2	95
120–129	6	93
110–119	4	87
100–109	7	83
90–99	37	76
80–89	33	39
70–79	6	6

The sample median (an analog to the population median) is defined in the same way as a population median. For a sample, 50% of the observations fall below and 50% fall above the median. For a population, 50% of the probability distribution is above and 50% is below the median.

In Table 4.4, the lower end of the distribution begins with the class 70–79. The column "cf" refers to the cumulative frequency of cases at and below a particular interval. For example, the cf at interval 80–89 is 39. The cf is found by adding the numbers in columns f and cf diagonally; e.g., 6 + 33 = 39. First, we must find the interval in which the median is located. There are a total of 100 cases, so one-half of them (0.50n) equals 50. By inspecting the cumulative frequency column, we find the interval in which 50% of the cases (the 50th case) fall in or below: 90–99. The lower real limit of the interval is 89.5.

Here is a point that requires discussion. Previously, we stated that the measurements from a continuous scale represent discrete values. The numbers placed in the frequency table were continuous numbers rounded off to the nearest unit. The real limits of the class interval are halfway between adjacent intervals. As a result, the real limits of a class interval, e.g., 90–99, are 89.5 to 99.5. The width of the interval (i) is (99.5 – 89.5), or 10. Thus, placing these values in Formula yields:

median = 89.5 + 10[(0.50)(100) – 39] = 97.47

Another way of calculating median of grouped data is:

$$\text{Median} = l + \left[\frac{\frac{n}{2} - c}{f}\right] \text{X } h$$

Where,

l = lower limit of median class

n = total number of observations

c = cumulative frequency of the preceding class

f = frequency of median class

h = class size (upper limit - lower limit)

The Mode

The mode refers to the class (or midpoint of the class) that contains the highest frequency of cases. In Table 4.4, the modal class is 90–99. When a distribution is portrayed graphically, the mode is the peak in the graph. Many distributions are multimodal, referring to the fact that they may have two or more peaks. Such multimodal distributions are of interest to epidemiologists because they may indicate different causal mechanisms for biological phenomena, for example, bimodal distributions in the age of onset of diseases such as tuberculosis, Hodgkins disease, and meningococcal disease. Figure 4.1 illustrates unimodal and bimodal distributions.

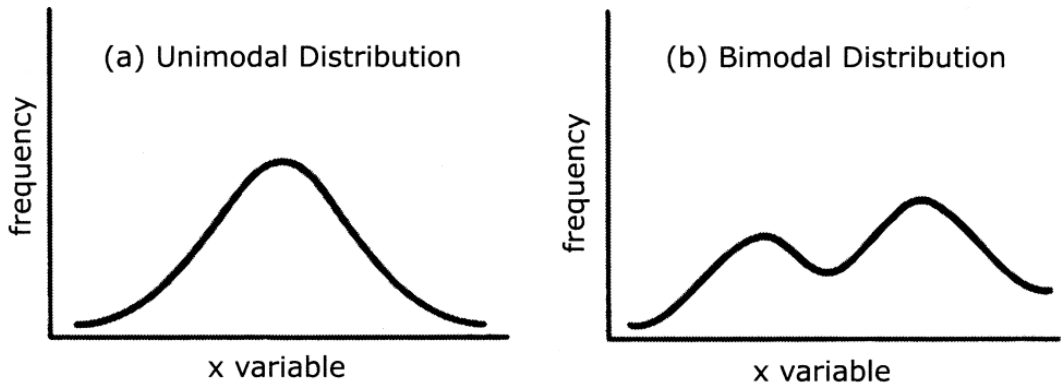


Figure 4.1. Unimodal and bimodal distribution curves. (Source: Authors.)

Geometric Mean

The geometric mean is a statistical measure that summarizes the central tendency of a **multiplicative** data set. Unlike the arithmetic mean (average), which simply sums the values and divides by the number of values, the geometric mean calculates the **nth root of the product** of all the values.

Applications of Geometric Mean:

- **Growth rates:** Analyzing the average growth rate of an investment or population over a period.
- **Index numbers:** Constructing stock market indices or other economic indicators that reflect proportional changes.
- **Biology:** Studying the average change in population size or cell growth over time.

Calculating the Geometric Mean:

There are two main ways to calculate the geometric mean:

1. **Formula:**

$$GM = \sqrt[n]{X_1 X_2 X_3 \cdots X_n}$$

- **n:** represents the number of data points.
- **x_1 to xn:** represent the individual data values.

2. Logarithmic method:

1. Take the **logarithm** (base doesn't matter but consistency is key) of each data value.
2. Calculate the **average** of the logarithms.
3. Take the **antilogarithm** of the average.

Both methods will produce the same result.

Example:

Calculate the geometric mean of the following data set: {2, 4, 8, 16}

Which Measure Should You Use?

Each of the measures of central tendency has strengths and weaknesses. The mode is difficult to use when a distribution has more than one mode, especially when these modes have the same frequencies. In addition, the mode is influenced by the choice of the number and size of intervals used to make a frequency distribution.

The median is useful in describing a distribution that has extreme values at either end; common examples occur in distributions of income and selling prices of houses. Because a few extreme values at the upper end will inflate the mean, the median will give a better picture of central tendency.

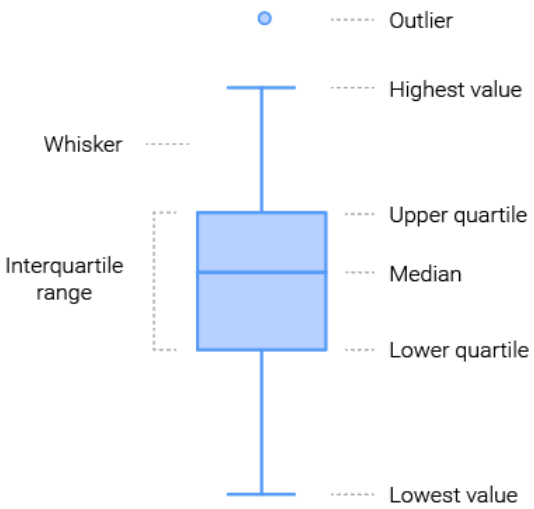
Finally, the mean often is more useful for statistical inference than either the mode or the median. For example, we will see that the mean is useful in calculating an important measure of variability: variance. The mean is also the value that minimizes the sum of squared deviations (mean squared error) between the mean and the values in the data set, a point that will be discussed in later chapters and that is exceedingly valuable for statistical inference.

Box and Whisker Plots?

Box and whisker plots, also known as boxplots, are graphical representations of data distribution. They provide a concise way to visualize the spread, center, and skewness of a dataset, making them valuable tools for exploratory data analysis and data comparison.

Components of a Box and Whisker Plot:

- **Box:** Represents the middle 50% of the data, also known as the interquartile range (IQR).
    - **Median:** The line dividing the box in half, representing the middle value when the data is ordered from least to greatest.
    - **Upper Quartile (Q3):** The upper boundary of the box, marking the value above which 25% of the data falls.
    - **Lower Quartile (Q1):** The lower boundary of the box, marking the value below which 25% of the data falls.
- **Whiskers:** Extend from the box and depict the range of data beyond the quartiles.
    - Ideally, they extend to the **minimum** and **maximum** values within 1.5 times the IQR.
    - **Outliers:** Values beyond 1.5 times the IQR are considered outliers and are typically plotted individually.



Creating a Box and Whisker Plot:

1. **Order your data:** Arrange your data points from least to greatest.
2. **Calculate the quartiles:**
    - **Q1:** Find the median of the lower half of the data.
    - **Q3:** Find the median of the upper half of the data.
3. **Calculate the interquartile range (IQR):** IQR = Q3 - Q1.
4. **Identify the median:** The median is the middle value in your ordered data set.
5. **Determine the whisker boundaries:**
    - Upper whisker boundary: Q3 + (1.5 X IQR)
    - Lower whisker boundary: Q1 – (1.5 X IQR)
6. **Plot the box:**
    - Draw a rectangle with the bottom at Q1, the top at Q3, and a line in the middle at the median.
7. **Plot the whiskers:**
    - Draw lines extending from the top and bottom of the box to the whisker boundaries.
8. **Plot outliers:**
    - If any values fall outside the whisker boundaries, plot them as individual points beyond the whiskers.

Interpreting a Box and Whisker Plot:

- **Center:** The median indicates the central tendency of the data.
- **Spread:** The IQR represents the middle 50% of the data, highlighting the spread of the data.
- **Skewness:** If the box is not symmetrical, the data is skewed. A skewed distribution has a longer whisker on one side than the other.
- **Outliers:** Identify any data points that fall outside the whisker boundaries, which may require further investigation.