Descriptive statistics are statistics calculated from the data and are used for describing, organizing, and summarizing data. The predictive and inferential part of statistics is based on inferential statistics, which we will consider in subsequent lectures. Obviously, we have to summarize and describe the data before we know enough about it to make inferences or predictions.

There are two basic types of descriptive statistics: measures of central tendency and measures of variability. Measures of central tendency attempt to represent a set of data or a frequency distribution by a single most representative or characteristic value that indicates the central value around which the other values tend to cluster. Measures of variability indicate the degree of closeness of the individual values to the measure of central tendency.

\[ STATISTICS \ MEASURING \ CENTRAL \ TENDENCY: \ MODE, \ MEDIAN, \ MEAN \]

Mode

The mode is the value that occurs most often in a set of data. The mode may not even exist, or there may be more than one mode; sometimes, either or both of these alternatives is true, depending on your point of view.

Example: The set of values 1, 1, 2, 2, 2, 3, 4, 5, clearly has a mode of 2, since 2 is the most frequently occurring value.

The mode is most useful for data that have been grouped into frequency distributions, since we can easily see the class with the largest frequency. We call this class the modal class and say that the mode is in this class.

Figures 2.1 and 2.2 from the previous lectures are reproduced below. For the quiz scores (Fig. 2.1), the modal class is 15 and the mode is 15. For the dental specialists (Fig. 2.2), the orthodontics class is the modal class.
The mode is of little value for very small data sets. However, with data sets containing large numbers of values, the mode may be a useful measure of central tendency.

The mode is applicable to frequency distributions of variables at all four levels of measurement.

Median

The median is another measure of central tendency that can be used to represent a set of data. The median of a set of values is the value that is both greater than or equal to half the values and less than or equal to half the values, i.e., it is the middle number in an ordered array.

Example: The set of values 1, 1, 3, 4, 5, 6, 7, 8, 10 has a median of 5, since 5 is the middle number; four of the values are less than 5 and four of the values are greater than 5.

Example: The set of values 1, 13, 17, 20, 35, 42, 59 has a median of 20, since 20 is the middle number.

Both of the above arrays had odd numbers of values. How do we determine the median for an array with an even number of values? Suppose we have the values 1, 2, 3, 5, 6, 7, 9, 10. By our definition, we see that the median could either be 5 or 6, since either of these satisfies the definition, but we want a unique number for the median. We prefer some number between 5 and 6, and the logical choice is the number that is exactly halfway between these two middle numbers. The rule is to select the number that is exactly between the two middle numbers and call it the median. Thus, the median for this set of values is 5.5.

Suppose we have the set of values 1, 2, 3, 6, 12, 17, 18, 20, 31, 40. The median falls between 12 and 17, so by our rule, the median is halfway between 12 and 17. Thus, 17 - 12 = 5, and half of 5 is 2.5, so we either add 2.5 to 12 or subtract 2.5 from 17 to get a median of 14.5.

The original quiz scores for Figure 2.1 are reproduced in Table 2.1. The median score is 15.5. Note that the median is not applicable to the data on dental specialists.

<table>
<thead>
<tr>
<th>Table 2.1. Quiz scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>9  15  16  17</td>
</tr>
<tr>
<td>11 15  16  18</td>
</tr>
<tr>
<td>13 15  16  18</td>
</tr>
<tr>
<td>13 15  17  19</td>
</tr>
<tr>
<td>14 15  17  20</td>
</tr>
</tbody>
</table>

In general, the median is used only for quantitative variables, i.e., with interval and ratio scales, since the median is a numerical value.

Sample vs. Population

A set of data is considered to be a sample of values from some larger population of possible values. The sample is the subset of values we actually observe. The population is the set of all possible values of the variable. There is a subtle distinction between everyday usage and the statistical
definitions of sample and population. Suppose we are measuring heights of sixth graders in Texas. In everyday usage, we would refer to a sample of sixth graders in Texas, i.e., we're talking about the sixth graders themselves. In statistics, although we frequently use the same everyday language, the term "sample" refers to the subset of heights actually observed and the term "population" refers to the set of all possible values of heights, not to the sixth graders themselves.

Mean

In statistics, we distinguish between a sample mean and the population mean because the sample mean is calculated from some subset of values, while the population mean is calculated from the entire population of values.

We calculate the mean of a sample by the following formula, which also defines the sample mean:

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

In this formula, $\overline{X}$ denotes the sample mean, $n$ is the number of observations or values in the sample (the sample size), the Greek letter $\sum$ (sigma) is shorthand for "add up all the values from $i = 1$ to $i = n$, where $i$ is an index that counts the values, and the $X_i$ are the individual data values. This is called summation notation.

Thus, without the use of the sigma or summation notation, we could write

$$\overline{X} = \frac{1}{n} (X_1 + X_2 + X_3 + \cdots + X_n)$$

where the three dots between the third and last terms indicate that the middle terms have been omitted to save writing, because it is obvious from the rule or expression what these terms should be. Thus, the use of the three dots save writing, but the use of summation notation saves even more writing. Both are simply methods of abbreviation.

Example: Suppose our sample of data consists of the four values 3, 6, 5, 6. The sample mean is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{4} \sum_{i=1}^{4} X_i = \frac{1}{4} (3 + 6 + 5 + 6) = \frac{20}{4} = 5.0$$

Example: The mean of the 20 rat weights in Table 2.5 from the second lecture is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{20} (214 + 221 + \cdots + 283) = 257.75$$
The mean is properly used with quantitative data only, i.e., interval and ratio variables. It is often used with ordinal variables such as ranks, even though it shouldn’t be.

The population mean is calculated exactly the same way as the sample mean, but we use two different symbols to make it clear that we are referring to the population and not a sample. We use the Greek letter \( \mu \) for the population mean (instead of \( \bar{X} \)) and upper case \( N \) for the population size (instead of \( n \), the sample size.) The population mean is defined by the formula

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} X_i
\]

Sample mean for grouped data

Suppose we did not have the original rat weight data of Table 2.5 (from Lecture 2) but we did have the histogram of Figure 2.3 (reproduced below). If we want to know the mean of the data, we can estimate it fairly closely by calculating a mean from the grouped data. We calculate the mean from the grouped data by using the class mark as the representative value for each class. Remember that we have made approximations by lumping the values within each class and considering them as equivalent to each other, so if we now ask what is the most representative value in each class, the midpoint or class mark is the most sensible choice. Since we don’t have all 20 original data values, we multiply each class mark by the number of values in that class and add all these products, then divide by 20 as before. Thus, we have 1 weight of 215 g in the third class, 2 weights of 225 g in the second class, 1 weight of 235 g in the third class, 2 weights of 245 g in the fourth class, etc. The total is

\[
1 \times 215 + 2 \times 225 + 1 \times 235 + 2 \times 245 + 2 \times 255 + 5 \times 265 + 5 \times 275 + 2 \times 285 = 5170
\]

Then we divide by 20 to get the mean, \( \overline{X} = \frac{5170}{20} = 258.5 \).

If we now let \( X_i \) be the class mark for class \( i \), and \( n_i \) be the frequency of the \( i \)th class, we may define the mean for grouped data by

\[
\overline{X} = \frac{1}{n} \sum_{i=1}^{k} X_i n_i
\]

where the total sample size is \( n = n_1 + n_2 + \cdots + n_k \), and \( k \) is the number of classes.
Suppose we have only the histogram of Figure 2.4 from Lecture 2. As a second example of the mean of grouped data, we have

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{k} X_i n_i = \frac{1}{20} (3 \times 220 + 3 \times 240 + 7 \times 260 + 7 \times 280)
\]

\[= 258.00\]

Again this answer is very close to the mean of 257.75 for the original (ungrouped) data.

**MEASURES OF VARIABILITY: RANGE, VARIANCE, STANDARD DEVIATION, COEFFICIENT OF VARIATION**

**Range**

We define the range of the data as the difference between the highest and the lowest values.

Range = highest value - lowest value

The highest data value in Figure 2.1 is 20, while the lowest is 9. Thus, the range of these quiz scores is

\[\text{range} = \text{highest} - \text{lowest} = 20 - 9 = 11.\]

Note that in ordinary conversation, many people would say that the values range from 9 to 20; this statement does not specify the range in the statistical sense. The range is properly specified by the definition above.

One limitation of the range is that most of the data might be clustered fairly close together, but one or two extreme values would result in a large range, suggesting more variability in the data than is warranted. Further, the range uses only two of the observed values. We would be better off using all the data in our sample to calculate a measure of variability.

**Variance**

If the mean is our preferred measure of central tendency, it makes sense to define the variability in terms of the differences or deviations of each of the values from the mean, i.e., the variability is some measure of the distances between each of the data values and the mean. As a first step, we can define all these distances, deviations, or differences as \(d_i = X_i - \bar{X}\) where \(d_i\) denotes the difference between the value \(X_i\) and the mean, \(\bar{X}\). Suppose we consider the set of values in the \(X_i\) column of Table 3.1 as our data.
Table 3.1. Deviations from the mean as a measure of variability.

<table>
<thead>
<tr>
<th>i</th>
<th>X_i</th>
<th>d_i = X_i - \bar{X}</th>
<th>d_i^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>+2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>+3</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>+1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \sum X_i = 49 \quad \sum d_i = 0 \quad \sum d_i^2 = 28 \]

\[ \bar{X} = 7 \]

We first calculate the mean of these data values, \( \bar{X} = 7.0 \). Then we calculate the differences, \( d_i \) in the third column of the table. The squared differences are shown in the fourth column of Table 3.1 and their sum is 28. Thus, if we take the average of this sum of squared deviations, we might have a reasonably good measure of variability. Instead of dividing the total sum of squares by \( n \), the number of values, we divide by \( n - 1 \). In general, for any sample size \( n \), the degrees of freedom is \( n - 1 \), or

\[ d.f. = n - 1 = \text{degrees of freedom} \]

In summary, the sample variance, \( S^2 \), is defined and calculated by either of the formulas below:

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} d_i^2 \]

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

The variance of the data in Table 3.1 is

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

\[ = \frac{1}{7-1} \left[ (6-7)^2 + (4-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (10-7)^2 + (8-7)^2 \right] \]

\[ = \frac{1 + 9 + 4 + 0 + 4 + 9 + 1}{6} \]

\[ = 4.667 \]
Shortcut or machine formula for the sample variance

With large data sets, it is faster to calculate the variance by

\[ s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right] \]

This formula looks more complicated than the ones above, but is mathematically identical to them. The term

\[ \sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \]

means square every data value and add up these squared values. The term

\[ (\sum_{i=1}^{n} x_i)^2 = (x_1 + x_2 + \cdots + x_n)^2 \]

means add up all the data values, then square this total. Using the data from Table 3.1 with this formula,

\[ s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right] \]
\[ = \frac{1}{6} \left[ 6^2 + 4^2 + \cdots + 8^2 - (6 + 4 + \cdots + 8)^2 \right] \]
\[ = \frac{1}{6} \left[ 371 - (49)^2 \right] \]
\[ = 4.667 \]

Population variance

The population variance is defined as

\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 \]

i.e., we square the deviation of every value from the population mean, \( \mu \), then divide by the population size, \( N \).
Standard deviation

The standard deviation is the square root of the variance.

\[ S = \sqrt{S^2} = \text{sample standard deviation} \]

\[ \sigma = \sqrt{\sigma^2} = \text{population standard deviation} \]

In the example above, \( S = \sqrt{S^2} = \sqrt{4.667} = 2.16 \).

The variance is calculated by squaring the differences \( X_i - \bar{X} \). Since the differences \( X_i - \bar{X} \) will have the same units as both \( X_i \) and \( \bar{X} \) (e.g., if \( X_i = 5 \text{ lb} \) and \( \bar{X} = 2 \text{ lb} \), then \( X_i - \bar{X} = 3 \text{ lb} \)), when we square \( X_i - \bar{X} \), we also square the units. Thus, if \( X_i - \bar{X} = 3 \text{ lb} \), then \( (X_i - \bar{X})^2 = 9 \text{ lb}^2 \). It is rather difficult to conceive of our variability in terms of "pounds squared." Thus, the standard deviation is used to give us a measure of variability in terms of our original units. If the variance is 100 \( \text{lb}^2 \), the standard deviation is 10 \text{ lb}.

Coefficient of Variation

The coefficient of variation is a measure of relative variability, with the ratio of the sample standard deviation to the sample mean expressed as a percent. The coefficient of variation is defined as

\[ C.V. = \frac{S}{\bar{X}} \times 100\% \]

The coefficient of variation, \( C.V. \), is useful in making comparisons of very different sets of data. It literally gets around the old adage that "you can't compare apples and oranges." You can with the \( C.V. \), since dividing the standard deviation by the mean cancels out the units of measurement and results in a dimensionless number.

For example, suppose we have calculated the means and standard deviations of heart rate and body temperature of a sample of patients. Which of these physiological variables is inherently more variable, heart rate or body temperature? Suppose we have calculated a mean temperature \( \bar{X} = 98.4^\circ \text{F} \) with a standard deviation \( S = 0.5^\circ \text{F} \), and a mean heart rate \( \bar{Y} = 78 \text{ beats/min} \) with \( S = 9 \text{ beats/min} \). The \( C.V. \) for temperature is

\[
C.V. \text{-temp.} = \frac{S}{\bar{X}} \times 100\% = \frac{0.5^\circ \text{F}}{98.4^\circ \text{F}} \times 100\% = 0.5\%
\]

while the \( C.V. \) for heart rate is

\[
C.V. \text{-H.R.} = \frac{S}{\bar{Y}} \times 100\% = \frac{9 \text{ beats/min}}{78 \text{ beats/min}} \times 100\% = 11.5\%
\]

Thus, our data indicate that heart rate, with the larger \( C.V. \), is more variable than body temperature.
Summary

Descriptive statistics are statistics calculated from the data, used for describing, organizing, and summarizing the data.

A sample is a subset of values actually observed, taken from some larger population of all possible values of the variable being studied.

Measures of central tendency indicate the most characteristic or typical value in the data and include the mode, median, and mean.

Mode: the most frequently occurring value; applicable to frequency distributions on all four scales of measurement.

Median: the value that is both greater than and less than half the data values; applicable to quantitative variables (interval, ratio scales.)

Mean:

The sample mean is

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

The population mean is

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

The (sample) mean for grouped data is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} X_i n_i$$

The mean is properly applicable only to quantitative variables (interval, ratio scales) although it often is used with ordinal variables such as ranks.

Measures of variability indicate the degree of closeness of the individual values to the measure of central tendency and include the range, variance, standard deviation, and coefficient of variation.

Range = highest value - lowest value
The range is applicable only to quantitative variables (interval, ratio scales).

Variance

The sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$
Where \( n-1 \) = degrees of freedom

The sample variance is calculated by

\[
S^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{n}{n} (\sum_{i=1}^{n} x_i)^2 \right]
\]

The sample standard deviation is \( S = \sqrt{S^2} \)

The population variance is

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2
\]

The population standard deviation is \( \sigma = \sqrt{\sigma^2} \)

The variance and standard deviation are properly used with quantitative variables (interval, ratio scales).

**Coefficient of variation (C.V.)**

\[
C.V. = \frac{S}{\bar{X}} \times 100\%
\]

The C.V. is applicable to quantitative variables (interval, ratio scales).