Measures of Central Tendency

describe the center of the distribution

Measures of Variability

describe the spread of the data around the center
Parameters -

descriptive measures for a population

Statistics -

descriptive measures for a sample
Measures of Central Tendency

- Mode
- Median
- Mean
Mode
- the measurement that occurs most often (with the highest frequency)

Example

Data set

<table>
<thead>
<tr>
<th>Data value</th>
<th>How many observations with this value in the data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
</tr>
</tbody>
</table>
Useful for discrete variables.

For continuous variables we can define the MODAL INTERVAL to be the class interval with the highest frequency.

The mode is taken as the midpoint of the modal interval, it is an approximation of the mode of the actual sample measurement.
Mode Characteristics

♦ There can be more than one mode for a data set
♦ Not influence by extreme measurements
♦ Modes of subset cannot be combined to determine the mode of the complete data set
♦ Can be used in qualitative and quantitative data
♦ In group data the mode can change depending on the categories (classes) used
Median

-the middle value (after ordering the data from smallest to largest)

If the sample has an **odd** number of observations the median is the middle value

**Example**

Data set -original

| 4.4 | 4.9 | 4.2 | 4.4 | 4.8 | 4.9 | 4.8 | 4.5 | 4.3 | 4.8 | 4.7 | 4.4 | 4.2 |

Data set -ordered

| 4.2 | 4.2 | 4.3 | 4.4 | 4.4 | 4.4 | 4.5 | 4.7 | 4.8 | 4.8 | 4.8 | 4.9 | 4.9 |

The median is **4.5**
If the sample has an even number of observations the median is the average of the middle two observations when the measurements are arranged from lowest to highest.

**Example**

Data set - original

| 95 | 86 | 78 | 90 | 62 | 73 | 89 | 92 | 84 | 76 |

Data set - ordered

| 62 | 73 | 76 | 78 | 84 | 86 | 89 | 90 | 92 | 95 |

The median is \( \frac{84 + 86}{2} = 85 \)
Median Characteristics

♦ There is only one median for a data set

♦ Not influenced by extreme measurements

♦ Medians of subsets cannot be combined to determine the median of the complete data set

♦ Can be used only for quantitative data
Mean

- Most commonly used measure of central tendency
- Refers to as “average value” or “arithmetic mean”
- Sum of the measurements divided by the total number of measurements.
Population mean and sample mean are denoted with different symbols.

Population mean

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

Sample mean

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]
Example

15 observations - \( n = 15 \)

<table>
<thead>
<tr>
<th>( y_i )</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>55.20</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>18.06</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>28.16</td>
</tr>
<tr>
<td>( y_4 )</td>
<td>44.14</td>
</tr>
<tr>
<td>( y_5 )</td>
<td>61.61</td>
</tr>
<tr>
<td>( y_6 )</td>
<td>4.88</td>
</tr>
<tr>
<td>( y_7 )</td>
<td>180.29</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( y_{14} )</td>
<td>9.98</td>
</tr>
<tr>
<td>( y_{15} )</td>
<td>82.73</td>
</tr>
</tbody>
</table>

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \\
= \frac{1}{15} (55.20 + 18.06 + \cdots + 82.73) = \frac{2,483.56}{15} = 165.57
\]

Generally the mean should be carried NO MORE than one decimal place beyond the data set.
For a random sample selected from a population the sample mean is an **estimate** of a population mean.

Mean is influenced by extreme values (outliers).

**Trimmed mean** - drop the lowest and the highest observations and compute the mean.

5% trimmed mean
10% trimmed mean

**Example** - 10% trimmed mean

\[
\bar{y} = \frac{2,483.56 - 4.88 - 807.8}{13} = 128.53
\]

Median is a 50% trimmed mean
Mean Characteristics

♦ There is only one mean for a data set

♦ It is influence by extreme measurements

♦ Means of subset can be combined to determine the mean of the complete data set

♦ Can be used only for quantitative data
Measures of variability (or dispersion)

Why measures of central tendency alone are not enough to describe the data?
Range

It is the simplest measure of data variation

Range = Largest observation minus Smallest observation

♦ Only two measurements regardless of the number of observations
♦ Reveals nothing with respect of the way in which the bulk of the observations are dispersed within the interval.
♦ Not stable
♦ Very sensitive to a single unusual value

Example

Range = $807.8 - 4.88 = $802.92$
Percentile

The $p^{th}$ percentile of a set of $n$ measurements arranged in order of magnitude is the value that has at most $p\%$ of the measurements below it and at most $(100 - p)\%$ above it.

Used commonly to report test results.

How to calculate percentile -

$$p\% = \frac{100(j - 0.5)}{n}$$
Lower and Upper Quartiles

Lower quartile = 25\textsuperscript{th} percentile

Middle quartile = 50\textsuperscript{th} percentile or median

Upper quartile = 75\textsuperscript{th} percentile

**Interquartile range**
The difference in between the upper and the lower quartiles

\[ \text{IQR} = 75\textsuperscript{th} \text{ percentile} - 25\textsuperscript{th} \text{ percentile} \]

useful for comparing variabilities of two or more data sets, outlier resistant
Deviation from the center of the distribution or mean

\[ y - \bar{y} \]

**Example** Ott page 87 Fig. 3.23
Variance

Population variance - mean of squared deviations from the mean data value

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

\( N \) - population size

\( \sigma^2 \) - population variance
Sample variance - the variance of a set of \( n \) measurements \( y_1, y_2, y_3, \ldots, y_n \) with mean \( \bar{y} \) and the sum of squared deviations divided by \( n-1 \)

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

\( n \) - sample size

\( S^2 \) - sample variance

This is an unbiased estimator of the population variance
The formula can look like this

\[ s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{df} \]

where \( df \) is degrees of freedom.

**Degrees of freedom** - the number of statistically independent pieces of information in the sum of squares.

Or we can say that degrees of freedom is equal to the number of observations used to estimate the sum of squares minus the number of parameters that need to be estimated prior to calculating the sum of squares.

Variance is measured in square of the units of the original data.
Standard deviation

Positive square root of the variance

Population standard deviation $\sigma$

Sample standard deviation $s$

Basically provides same information as the variance, but in the same units than the original data. Variance is needed to compute the standard deviation

For symmetrical ("mound-shaped") histogram:

about 68% of all observations are within one standard deviation distance from the mean
about 95% of all observations are within two standard deviation distance from the mean

Standard deviation can be approximated as range/4
Coefficient of variation

measures the variability in the values of the data relative to the variability of the mean

\[ CV\% = \frac{s}{\bar{y}} \times 100\% \]

CV has no units, but is expressed as a percentage.

Relative measure of variation that is independent of the unit of measure.

Example

Soil organic matter content ranges from 4.5% to 7.5%, mean is 6%, std deviation is 0.7
Soil hydraulic conductivity ranges from 0.001 in/hr to 1 in/hr, mean is 0.01 in/hr, std deviation is 0.012

\[ CV_{organic\ matter} = \frac{0.7}{6} \times 100\% = 12\% \]

\[ CV_{hydraulic\ cond.} = \frac{0.012}{0.01} \times 100\% = 120\% \]
Descriptive statistics in SAS

PROC MEANS

PROC UNIVARIATE

Example - soybean oil and protein contents measured in seed samples of the three soybean varieties