Measures of Centrality and Dispersion

This module introduces key concepts of summarizing data: measures of centrality and dispersion.

Learning Objectives

After completion of this module, the student will be able to

- calculate central tendencies and measures of dispersion of data sets employing both formulas and spreadsheet functions (LO2a)
- interpret central tendencies and measures of dispersion in the context of a data set (LO2a)
- choose an appropriate graph to visualize the data (LO2a)
- use online resources to gather information about data (LO2b)

Knowledge and Skills

- sample mean, average, median
- range, percentile, quartile, interquartile range
- sample variance, sample standard deviation
- EXCEL functions for calculating descriptive statistics

Prerequisites

- real number line
- percentages
- graphing in Excel
Online Resources

- Location of data (quartiles, percentiles, interquartile range):
  - [http://cnx.org/content/m16314/latest/](http://cnx.org/content/m16314/latest/)
- Measures of the center of data (sample mean, median):
  - [http://cnx.org/content/m17102/latest/](http://cnx.org/content/m17102/latest/)
- Measures of dispersion (sample standard deviation)
  - [http://cnx.org/content/m17103/latest/](http://cnx.org/content/m17103/latest/)
- Summary of formulas: [http://cnx.org/content/m16310/latest/](http://cnx.org/content/m16310/latest/)

Data Sets

Below are seven distinct data sets from the Organization for Economic Co-operation and Development (OECD). Pick one data set that interests you and complete the following tasks that are common to all data sets and Task 2 that is listed below each data set.

Tasks

1. Give a short description of the OECD (history, members, and mission).
2. See below.
3. Find the following quantities
   a. Largest and smallest value
   b. Range (i.e., the difference between the largest and the smallest value)
   c. Average
   d. Median
   e. 25th percentile, 50th percentile, 75th percentile
   f. Middle 50%
   g. Interquartile range IQR (i.e., the range of the middle 50%): IQR=75th percentile – 25th percentile
   h. Sample standard deviation
   i. Sample variance
4. Visualize the data using an appropriate graph.
5. In one paragraph, describe one aspect of the data that you found surprising or particularly interesting.
Tab GDP: Gross Domestic Product  
http://lysander.sourceoecd.org/vl=890100/cl=22/nw=1/rpsv/health2007/1-3.htm

2. The data set lists the GDP per capita in PPP (purchasing power parities). Describe in your own words what this means.

Tab LEB: Life Expectancy at Birth  
http://puck.sourceoecd.org/vl=3331303/cl=20/nw=1/rpsv/health2007/2-1.htm

2. The data set lists the life expectancy at birth in years. Describe in your own words what this means.

Tab HEF: Health expenditures and financing  
http://puck.sourceoecd.org/vl=3061947/cl=19/nw=1/rpsv/health2007/5-1.htm

2. The data set lists the total per capita expenditures on health in PPP (purchasing power parities). Describe in your own words what this means.

Tab PPT: Practicing Physicians per 1000 Population  
http://puck.sourceoecd.org/vl=3140824/cl=13/nw=1/rpsv/health2007/4-2.htm

2. The data set lists the number of practicing physicians per 1000 population. Describe in your own words what this means.

Tab PNT: Practicing Nurses per 1000 Population  
http://puck.sourceoecd.org/vl=3140824/cl=13/nw=1/rpsv/health2007/4-2.htm

2. The data set lists the number of practicing nurses per 1000 population. Describe in your own words what this means.

Tab TAG: Trends in tertiary graduation and entry rates  
http://oberon.sourceoecd.org/vl=1423295/cl=19/nw=1/rpsv/factbook2009/09/01/02/index.htm
2. The data set lists the percentage of tertiary type A graduation rates. Describe in your own words what this means.

Tab EOH: Expenditure on health as a percentage of GDP (2006 or latest available year)
http://titania.sourceoecd.org/vl=1675245/cl=14/nw=1/rpsv/factbook2009/10/02/01/index.htm

2. The data set lists the expenditure on health as a percentage of GDP (public and private). Describe in your own words what this means.
Introduction to Descriptive Statistics in Excel

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Practice Data Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ID</td>
<td>Data</td>
<td>Largest value</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>Smallest value</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>6</td>
<td>Range</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>7</td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>Median</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>4</td>
<td>25th Percentile</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>7</td>
<td>50th Percentile</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>4</td>
<td>75th Percentile</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>3</td>
<td>Middle 50%</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>7</td>
<td>Interquartile range</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>5</td>
<td>Sample standard deviation</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>4</td>
<td>Sample variance</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>18</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>19</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>20</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>21</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

You can find the data set in the table above under the Practice tab in the MeasureCentralityDispersion.xlsx worksheet. We will use this data set to illustrate how to calculate in Excel the measures of centrality and dispersion that are introduced in this worksheet.

1. **Largest Value**: the largest value is the maximum value of all the data points

   enter in Cell E3: `=MAX(E4:E24)`

2. **Smallest Value**: the smallest value is the minimum value of all the data points

   enter in Cell E4: `=MIN(E4:E24)`
3. **Range**: the range is the difference between the maximum and the minimum value

   enter in Cell E5 =E3-E4

4. **Average**: the average of \( n \) numbers \( x_1, x_2, \ldots, x_n \) is calculated as \( \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \)

   enter in Cell E6 =AVERAGE(E4:E24)

5. **Median**: the median of \( n \) numbers \( x_1, x_2, \ldots, x_n \) is the middle value of the ordered numbers (from smallest to largest) when \( n \) is odd and is the average of the two middle values when \( n \) is even. In Excel a single function calculates the median of a set of data

   enter in Cell E7 =MEDIAN(E4:E24)

6. **Percentiles**: a percentile is a number below which a certain percent of observations fall. For instance, the 25th percentile of a data set is the value of the observation below which 25% of the values fall. In Excel, the function calculating percentiles is given as “=PERCENTILE(array,k)” where \( \text{array} \) is the range of cells of the data and \( k \) is the percentile (e.g., 0.25 for the 25th percentile).

   enter in Cell E8 =PERCENTILE(E4:E24,0.25)
   enter in Cell E9 =PERCENTILE(E4:E24,0.5)
   enter in Cell E10 =PERCENTILE(E4:E24,0.75)

7. **Middle 50%**: The middle 50% are the numbers between the 25th and the 75th percentile, that is, the numbers between the values in Cell E8 and Cell E10.

   enter in Cell E11 “value in Cell E8” …”value in Cell E10”

8. **Interquartile Range**: The interquartile range is the difference between the 75th and 25th percentile.

   enter in Cell 12 =E10-E8

9. **Sample Standard Deviation**: the sample standard deviation of \( n \) numbers \( x_1, x_2, \ldots, x_n \) is calculated as

\[
s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}
\]

   enter in Cell E13 =STDEV(E4:E24)

10. **Sample Variance**: the sample variance of \( n \) numbers \( x_1, x_2, \ldots, x_n \) is calculated as
\[ s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \]

It is the square of the sample standard deviation. There are two ways to calculate this in Excel.

enter in Cell E14 \[=\text{VAR(E4:E24)}\]

or if you already calculated the sample standard deviation (in Cell E13),

enter in Cell E14 \[=\text{E13}^2\]

Solution:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest value</td>
<td>9</td>
</tr>
<tr>
<td>Smallest value</td>
<td>2</td>
</tr>
<tr>
<td>Range</td>
<td>7</td>
</tr>
<tr>
<td>Average</td>
<td>5.190476</td>
</tr>
<tr>
<td>Median</td>
<td>5</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>4</td>
</tr>
<tr>
<td>50th Percentile</td>
<td>5</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>7</td>
</tr>
<tr>
<td>Middle 50%</td>
<td>4...7</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>3</td>
</tr>
<tr>
<td>Sample standard deviation</td>
<td>1.965173</td>
</tr>
<tr>
<td>Sample variance</td>
<td>3.861905</td>
</tr>
</tbody>
</table>