Descriptive Statistics

Petra Petrovics
DESCRIPTIVE STATISTICS

• **Definition:** Descriptive statistics is concerned only with collecting and describing data

• **Methods:**
  - statistical tables and graphs
  - descriptive measures

**Descriptive measure** – a single number that provides information about a set of data
Description of a Population

I. Central Tendency
   - mean
   - mode
   - median

II. Percentiles, Quartiles

III. Dispersion

IV. Shape
I.1. Means

- **Arithmetic mean** (average)
- **Geometric mean** – the ratio of any two consecutive numbers is constant
  - e.g. compound interest rate
- **Harmonic mean** – units of measurement differ between the numerator and denominator
  - e.g. miles per hour
- **Quadratic mean**
  - e.g. the form of standard deviation, mean of differences
Arithmetic Mean

- Typically referred to as mean.
- The most common measure of central tendency.
- It is the only common measure in which all the values play an equal role.
- Symbol: $\bar{x}$, called $X$-bar

Raw Data Expressions (simple formula):

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

Frequency Distribution Expressions (weighted formula):

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{\sum f_i}$$
Properties of Mean

- The sum of the differences from the mean is 0.
  \[ \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \]

- \( \sum_{i=1}^{n} (x_i - a)^2 \) is minimal, if \( a = \bar{x} \)

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( d_i = x_i - \bar{x} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-100</td>
</tr>
<tr>
<td>150</td>
<td>-50</td>
</tr>
<tr>
<td>210</td>
<td>+10</td>
</tr>
<tr>
<td>240</td>
<td>+40</td>
</tr>
<tr>
<td>300</td>
<td>+100</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>1000</td>
</tr>
<tr>
<td>( \bar{x} )</td>
<td>200</td>
</tr>
</tbody>
</table>
## Properties of Mean 2.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$x_i+50$</th>
<th>$x_i\cdot 1.1=y$</th>
<th>$Z=x+y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>150</td>
<td>110</td>
<td>210</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>165</td>
<td>315</td>
</tr>
<tr>
<td>210</td>
<td>260</td>
<td>231</td>
<td>441</td>
</tr>
<tr>
<td>240</td>
<td>290</td>
<td>264</td>
<td>504</td>
</tr>
<tr>
<td>300</td>
<td>350</td>
<td>330</td>
<td>630</td>
</tr>
<tr>
<td>$\Sigma$ 1000</td>
<td>1250</td>
<td>1100</td>
<td>2100</td>
</tr>
<tr>
<td>200</td>
<td>250</td>
<td>220</td>
<td>420</td>
</tr>
</tbody>
</table>

- If you add a constant ‘a’ to every $x_i$, the mean will be $a + \bar{x}$
- If you multiply every $x_i$ by a constant ‘b’, the mean will be $b\cdot\bar{x}$
- $x_1, x_2, ..., x_n \rightarrow \bar{x}$
- $y_1, y_2, ..., y_n \rightarrow \bar{y}$
- $x_1 + y_1; ...; x_n + y_n \rightarrow \bar{x} + \bar{y}$
Geometric Mean

The rate of change of a variable over time. The $n^{\text{th}}$ root of the product of $n$ values.

Raw Data Expressions:

$$\overline{X}_g = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

Frequency Distribution Expressions:

$$\overline{X}_g = \sqrt[n]{\prod_{i=1}^{n} X_i^{f_i}}$$
GDP in Hungary

<table>
<thead>
<tr>
<th>Period</th>
<th>Previous quarter = 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008. Q1</td>
<td>100.9</td>
</tr>
<tr>
<td>2008. Q2</td>
<td>99.8</td>
</tr>
<tr>
<td>2008. Q3</td>
<td>99.0</td>
</tr>
<tr>
<td>2008. Q4</td>
<td>98.1</td>
</tr>
</tbody>
</table>

Source: HCSO

Average growth rate:

$$x_g = \sqrt[4]{1.009 \cdot 0.998 \cdot 0.99 \cdot 0.981} = \sqrt[4]{0.978} = 0.994 = 99.4\%$$
Harmonic Mean

The harmonic mean of a set of $n$ numbers is found by adding up the reciprocals of the numbers, and then dividing $n$ by this sum.

Raw Data Expressions:

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

Frequency Distribution Expressions:

$$
\bar{X}_h = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{X_i}}, \text{ where } n = \sum_{i=1}^{k} f_i
$$
Relation between the Partitional Ratio and Dynamic Ratio

<table>
<thead>
<tr>
<th>Factories</th>
<th>Turnover (M Ft)</th>
<th>Partitional of turnover (%)</th>
<th>Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t₀</td>
<td>t₁</td>
<td>t₀ (%)</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>36</td>
<td>20</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>60</td>
<td>27</td>
</tr>
<tr>
<td>E</td>
<td>70</td>
<td>77</td>
<td>47</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>14.5</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>150</strong></td>
<td><strong>187.5</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
\[
\bar{V} = \frac{\sum A}{\sum B} = \frac{187.5}{150} = 1.25
\]

\[
\bar{V} = \frac{\sum B_i \cdot V_i}{\sum B_i} = \frac{30 \cdot 1.2 + 40 \cdot 1.5 + 70 \cdot 1.1 + 10 \cdot 1.45}{150} = 1.25
\]

\[
\bar{V} = \frac{\sum B_i \cdot V_i}{\sum B_i} = \frac{0.2 \cdot 1.2 + 0.27 \cdot 1.5 + 0.47 \cdot 1.1 + 0.06 \cdot 1.45}{1} = 1.25
\]

\[
\bar{V} = \frac{\sum A_i}{\sum \frac{A_i}{V_i}} = \frac{187.5}{\frac{36}{1.2} + \frac{60}{1.5} + \frac{77}{1.1} + \frac{14.5}{1.4}} = 1.25
\]
**Quadratic Mean**

\[ \bar{x}_q = \sqrt{\frac{n}{\sum_{i=1}^{n} x_i^2}} \]

\[ \bar{x}_q = \sqrt{\sum_{i=1}^{k} g_i \cdot x_i^2} \]

\[ \bar{x}_q = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot x_i^2}{\sum_{i=1}^{k} f_i}} \]
I.2. Median

- Statistic which has an equal number of variates above and below it \( \left( \frac{n+1}{2} \right) \)
- Raw Data Expressions: ranked value
- Independent from extreme values 🔷
- Just from data in order 🔸
- The „middle term”

I.3. Mode

- The value that occurs most frequently
- Typical value
I.3. Mode

- The value that occurs most frequently
- Typical value
<table>
<thead>
<tr>
<th>Measurement Scale</th>
<th>Best Measure of the ‘Middle’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal (Categorical)</td>
<td>Mode</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Median</td>
</tr>
<tr>
<td>Interval</td>
<td>Symmetrical data: Mean Skewed data: Median</td>
</tr>
<tr>
<td>Ratio</td>
<td>Symmetrical data: Mean Skewed data: Median</td>
</tr>
</tbody>
</table>
II. Percentiles and Quartiles

• The $P^{th}$ percentile of a group of members is that value below which lie $P\%$ ($P$ percent) of the numbers in the group.

• $Q_1$ (lower quartile): The first quartile is the 25th percentile. It is that point below which lie $\frac{1}{4}$ of the data.

• $Q_2$ (middle quartile): The median is the data below which lie half the data. It is the 50th percentile.

• $Q_3$ (upper quartile): The third quartile is the 75th percentile point. It is that below which lie 75 percent of the data.
III. Measures of Dispersion

1. Range
2. Interquartile Range
3. Population and Sample Standard Deviation
4. Population and Sample Variance
5. Coefficient of Variation
III.1. Range

- The **range** of a set of observations is the difference between the largest observation and the smallest observation.

\[ R = X_{\text{max}} - X_{\text{min}} \]

III.2. IQR

- Interquartile range: difference between the first and third quartiles.

\[ IQR = Q_3 - Q_1 \]
III.3. Standard Deviation

- The **standard deviation** is a measure of dispersion around the mean.
- A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.
- In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within 2 standard deviations.
Properties of Standard Deviation

• 0, if \( x = \text{constant} \)
• \( 0 \leq |\sigma| \leq \bar{x} \sqrt{N - 1} \)
• \( \sigma^2 = \bar{x}_q^2 - \bar{x}^2 \)
**Properties of Standard Deviation**

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$d_i = x_i - \bar{x}$</th>
<th>$d_i^2$</th>
<th>$y_i = x_i + 50$</th>
<th>$d_i = y_i - \bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-100</td>
<td>10000</td>
<td>150</td>
<td>-100</td>
</tr>
<tr>
<td>150</td>
<td>-50</td>
<td>2500</td>
<td>200</td>
<td>-50</td>
</tr>
<tr>
<td>210</td>
<td>+10</td>
<td>100</td>
<td>260</td>
<td>+10</td>
</tr>
<tr>
<td>240</td>
<td>+40</td>
<td>1600</td>
<td>290</td>
<td>+40</td>
</tr>
<tr>
<td>300</td>
<td>+100</td>
<td>10000</td>
<td>350</td>
<td>+100</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>1 000</td>
<td>0</td>
<td>$\Sigma$ 1 250</td>
<td>0</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>200</td>
<td>$\bar{y}$</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sigma^2 = 4 840$</td>
<td>$\sigma = 69.6$</td>
<td>$\sigma^2 = 4 840$</td>
<td>$\sigma = 69.6$</td>
</tr>
</tbody>
</table>

- If you add a constant ‘a’ to every $x_i$, the standard deviation will be the same.
## Properties of Standard Deviation

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$d_i = x_i - \bar{x}$</th>
<th>$d_i^2$</th>
<th>$y_i = x_i \cdot 1.1$</th>
<th>$d_i = y_i - \bar{y}$</th>
<th>$d_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-100</td>
<td>10 000</td>
<td>110</td>
<td>-110</td>
<td>12 100</td>
</tr>
<tr>
<td>150</td>
<td>-50</td>
<td>2 500</td>
<td>165</td>
<td>-55</td>
<td>3 025</td>
</tr>
<tr>
<td>210</td>
<td>+10</td>
<td>100</td>
<td>231</td>
<td>+11</td>
<td>121</td>
</tr>
<tr>
<td>240</td>
<td>+40</td>
<td>1 600</td>
<td>264</td>
<td>+44</td>
<td>1 936</td>
</tr>
<tr>
<td>300</td>
<td>+100</td>
<td>10 000</td>
<td>330</td>
<td>+110</td>
<td>12 100</td>
</tr>
<tr>
<td>$\Sigma$ 1000</td>
<td>0</td>
<td>24 200</td>
<td>$\Sigma$ 1 100</td>
<td>29 282</td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$ = 200</td>
<td></td>
<td></td>
<td>$\bar{y}$ = 220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$ = 4 840</td>
<td></td>
<td></td>
<td></td>
<td>$\sigma^2$ = 5 856.4</td>
<td></td>
</tr>
<tr>
<td>$\sigma$ = 69.6</td>
<td></td>
<td></td>
<td></td>
<td>$\sigma$ = 76.52</td>
<td></td>
</tr>
</tbody>
</table>

- If you multiply every $x_i$ by a constant ‘$b$’, the standard deviation will be $b \times \sigma$
III.4. Variance

- **Variance** of a set of observations: the average squared deviation of the data points from their mean.

- Population variance:
  \[ \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{\sum_{i=1}^{n} f_i} \]

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

III.5. Coefficient of Variation

- The measure of dispersion around the mean in %.

  \[ V = \frac{\sigma}{\bar{X}} \quad V = \frac{S}{\bar{X}} \]
IV. Measures of Shape

- **Skewness** is a measure of the degree of asymmetry of a frequency distribution.
- **Kurtosis** is a measure of the flatness (versus peakedness) of a frequency distribution.
IV.1. Kurtosis

The measure of the extent to which observations cluster around the central point.

Positive – cluster more and have longer tails

Negative – cluster less and have shorter tails

For a normal distribution, the value of the kurtosis statistic is zero.
IV.2. Skewness

\[ A = \frac{\bar{X} - Mo}{\sigma} \]

\[ F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} \]

Skewed to the left (long right tail)

- \( Mo < Me < \bar{X} \)
- \( A > 0 \)

Symmetry

- \( Me = Mo = \bar{X} \)

Skewed to the right

- \( \bar{X} < Me < Mo \)
- \( A < 0 \)
The box plot is a set of five summary measures of the distributions of the data:
- the median of the data
- the lower quartile
- the upper quartile
- the smallest observation
- the largest observation
+ asymmetry
Box&Whiskers

- Smallest observation within 1.5(IQR) of lower hinge
- Lower quartile (hinge)
- IQR
- Median
- Upper quartile (hinge)
- Largest observation within 1.5(IQR) of upper hinge

Source: Aczel [1996]
Elements of Box Plot

- **Outlier**: Data point not below inner fence.
- **Smallest data point not below inner fence**: Smallest data point within the inner fence.
- **Half the data are within the box**: Half the data are within the interquartile range (IQR) from the lower quartile $Q_L$ to the upper quartile $Q_U$.
- **Largest data point not exceeding inner fence**: Largest data point within the inner fence.
- **Suspected outlier**: Data point outside the inner fence.

**Outer fence**:
- $Q_L - 3(IQR)$
- $Q_U + 3(IQR)$

**Inner fence**:
- $Q_L - 1.5(IQR)$
- $Q_U + 1.5(IQR)$

**Median**:
- $Q_L$ (lower quartile)
- $Q_U$ (upper quartile)

**IQR (Interquartile Range)**:
- $Q_U - Q_L$

Source: Aczel [1996]
Data sets A and B seem to be similar; sets C and D are not similar.

Source: Aczel [1996]
Box Plot

The highest salary

The least standard deviation

Employment Category

Clerical  Custodial  Manager

Q₃
Me
Q₁
Thanks for your attention!