Descriptive Statistics

Numerical Summary Measures

Describe Distribution with Numbers
• Measure of Center
• Measure of Variation
• Measure of Position

Measure of Central Tendency

Mean: the average value of the data.

If \( n \) observations are denoted by \( x_1, x_2, \ldots, x_n \), their (sample) mean is

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

Example: Birth weights (in lb) of 5 babies born from a group of women under certain diet.

7, 6, 8, 7, 7

Sol:
mean = \( \frac{7 + 6 + 8 + 7 + 7}{5} = \frac{35}{5} = 7 \)

[near the center of the data set]

Example: (number of hysterectomies performed by 15 male doctors)

27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

⇒ mean = 41.33

Weighted Mean

Example: (Grade point average)
A student received 3 A’s, 5 B’s, 2 C’s.

<table>
<thead>
<tr>
<th>Class (grade point, x)</th>
<th>Frequency (weight, w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Average grade point = \( \frac{3 \times 4 + 5 \times 3 + 2 \times 2}{3 + 5 + 2} = \frac{31}{10} = 3.1 \)

Weighted Mean

Weighted mean = \( \frac{w_1 x_1 + w_2 x_2 + \ldots + w_k x_k}{w_1 + w_2 + \ldots + w_k} = \frac{\sum w \cdot x}{\sum w} \)

Where \( w_1, w_2, \ldots \) are the weights and \( x_1, x_2, \ldots \) are the values (or class midpoint or class mark).
Descriptive Statistics

**Grouped Mean**

**Mean Estimation**

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency (w)</th>
<th>Class Mark (x)</th>
<th>w * x</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 - &lt; 110</td>
<td>1</td>
<td>100</td>
<td>1x100</td>
</tr>
<tr>
<td>110 - &lt; 130</td>
<td>2</td>
<td>120</td>
<td>2x120</td>
</tr>
<tr>
<td>130 - &lt; 150</td>
<td>3</td>
<td>140</td>
<td>3x140</td>
</tr>
<tr>
<td>150 - &lt; 170</td>
<td>1</td>
<td>160</td>
<td>1x160</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td></td>
<td>920</td>
</tr>
</tbody>
</table>

Estimated mean = \( \frac{920}{7} = 131.43 \)

**Median**: of a data set is

- the data value exactly in the middle of its ordered list if the number of pieces of data is odd,
- the mean of the two middle data values in its ordered list if the number of pieces of data is even.

[median is not influenced by outliers and is best for nonsymmetric distribution]

**Example**: (number of hysterectomies performed by 15 doctors)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

median = 34

**Example**: (Birth weights for 6 infants.)
5, 7, 6, 8, 5, 9

ordered list => 5, 5, 6, 7, 8, 9

median = \( (6+7)/2 = 6.5 \)

**Mode**: of a data set is the observation that occurs most frequently.

**Example 1**: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

Mode = 25

**Example 2**: (Blood type of 15 students)

Mode = A

A – 8
B – 3
O – 3
AB – 1
Descriptive Statistics

What is a Modal class?

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Freq</th>
<th>Cumulative R.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>90&lt; - 110</td>
<td>2</td>
<td>2/22 = .091</td>
<td>2/22</td>
</tr>
<tr>
<td>110&lt; - 130</td>
<td>4</td>
<td>4/22 = .182</td>
<td>6/22</td>
</tr>
<tr>
<td>130&lt; - 150</td>
<td>7</td>
<td>7/22 = .318</td>
<td>13/22</td>
</tr>
<tr>
<td>150&lt; - 170</td>
<td>3</td>
<td>3/22 = .136</td>
<td>16/22</td>
</tr>
<tr>
<td>170&lt; - 190</td>
<td>1</td>
<td>1/22 = .045</td>
<td>17/22</td>
</tr>
<tr>
<td>190&lt; - 210</td>
<td>0</td>
<td>0/22 = .000</td>
<td>17/22</td>
</tr>
<tr>
<td>210&lt; - 230</td>
<td>1</td>
<td>1/22 = .045</td>
<td>18/22</td>
</tr>
<tr>
<td>230&lt; - 250</td>
<td>0</td>
<td>0/22 = .000</td>
<td>18/22</td>
</tr>
<tr>
<td>250&lt; - 270</td>
<td>0</td>
<td>0/22 = .000</td>
<td>18/22</td>
</tr>
<tr>
<td>270&lt; - 290</td>
<td>1</td>
<td>1/22 = .045</td>
<td>19/22</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Mean ?

Median ?

Mode ?

Skewed to the Right

Measure of Dispersion (variability)

Range = largest data value – smallest data value

Sample from group I (diet program I):
7, 6, 8, 7, 7
=> mean = (7 + 6 + 8 + 7 + 7) / 5 = 35/5 = 7

Sample from group II (diet program II):
3, 4, 8, 9, 11
=> mean = (3 + 4 + 8 + 9 + 11) / 5 = 35/5 = 7

Does the mother’s diet program affect the birth weights of babies?

Is there any difference between the two samples?

range of sample I = 8 - 6 = 2
range of sample II = 11 - 3 = 8

Variance and Standard Deviation

Measure the spread of the data around the center of the data.

Example: Birth weights (in lb) of 5 babies born from a group of women under diet program II.
3, 4, 8, 9, 11 => mean = \( \bar{x} = 7 \)

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Deviation from mean ( x_i - \bar{x} )</th>
<th>Squared Dev. ( (x_i - \bar{x})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3 – 7 = – 4</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>4 – 7 = – 3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8 – 7 = 1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>9 – 7 = 2</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>11 – 7 = 4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>46</td>
</tr>
</tbody>
</table>

Sample Variance = 46/4 = 11.5 lb,
Sample Standard Deviation = \( \sqrt{46/4} = 3.39 \) lb.

Descriptive Stat - 3
Descriptive Statistics

If \( n \) observations are denoted by \( x_1, x_2, ..., x_n \), their variance and standard deviation are

Sample Variance: \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \)  
(unbiased estimator for variance of an infinite population.)

Sample Standard Deviation: \( s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \)

Sample Mean: \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \)

A Short Cut formula:

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

Data, \( x_i \)  
<table>
<thead>
<tr>
<th>( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>35</td>
</tr>
</tbody>
</table>

What is the standard deviation of the weights of babies from the sample of mothers who received diet program I?

Data: 7, 6, 8, 7, 7

\( s^2 = \frac{(0+1+1+0+0)}{(5-1)} = \frac{1}{4} \)

\( s = \sqrt{\frac{1}{4}} = 0.71 \)

Does the mother’s diet program affect the birth weights of babies?

Diet I: mean = 7, \( s = 0.71 \)
Diet II: mean = 7, \( s = 3.39 \)

Population Parameters

If \( N \) observations are denoted by \( x_1, x_2, ..., x_n \) are all the observation in a finite population, their mean, \( \mu \), variance \( \sigma^2 \), and standard deviation, \( \sigma \), are

Population Mean: \( \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \)

Population Variance: \( \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \)

Population Standard Deviation: \( \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \)

About \( s \) (sample standard deviation):

- \( s \) measures the spread around the mean.
- the larger \( s \) is, the more spread out the data are.
- if \( s = 0 \), then all the observations must be equal.
- \( s \) is strongly influenced by outliers.

The Use of Mean and Standard Deviation:

- Describe distribution
- Understand the center and the spread of the distribution
Descriptive Statistics

### Profit Margin (1972-1981)

<table>
<thead>
<tr>
<th>Company</th>
<th>$\bar{X}$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Water Works</td>
<td>7.61</td>
<td>.68</td>
</tr>
<tr>
<td>Brown &amp; Sharpe</td>
<td>7.62</td>
<td>7.39</td>
</tr>
<tr>
<td>Campbell Soup</td>
<td>13.65</td>
<td>1.05</td>
</tr>
<tr>
<td>McDonald’s</td>
<td>20.04</td>
<td>1.02</td>
</tr>
<tr>
<td>Pam American</td>
<td>-9.8</td>
<td>14.18</td>
</tr>
</tbody>
</table>

### Birth Weight Data

<table>
<thead>
<tr>
<th>Unit: oz</th>
<th>Mean, $\bar{X}$</th>
<th>Standard Deviation, $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother Drank Alcohol</td>
<td>102.2</td>
<td>25</td>
</tr>
<tr>
<td>Mother Did not Drink Alcohol</td>
<td>118.4</td>
<td>20</td>
</tr>
</tbody>
</table>

### Properties of a symmetric and bell-shaped (Normal) distribution:

- the distribution is symmetric about its mean ($\mu$).
- 68% of the area is between $\mu - \sigma$ and $\mu + \sigma$.
- 95% of the area is between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- 99.7% of the area is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

### Chebychev’s inequality

There is at least $1 - (1/k^2)$ of the data in a data set lie within $k$ standard deviation of their mean.
**Descriptive Statistics**

**Example:** Heart rates for asthmatic patients in a state of respiratory arrest has a mean of 140 beats per minute and a standard deviation of 35.5 beats per minute. What percentage of the population of this type of patients have heart rates lie between two standard deviations of the mean in a state of respiratory arrest?

It will be at 75%, because \( k = 2 \), and \( 1 - \left(\frac{1}{2^2}\right) = \frac{3}{4} = 75\% \).

**Example:** Heart rates example: mean=144, s.d.=35.5

\[ k = 2 \]
\[ 75\% = 1 - \left(\frac{1}{2^2}\right) \]

\[ 140 - 2 \times 35.5 = 69 \]
\[ 140 + 2 \times 35.5 = 211 \]

What about within three standard deviations?

Heart rates example: mean=144, s.d.=35.5

\[ k = 3 \]
\[ 89\% = 1 - \left(\frac{1}{3^2}\right) \]

\[ 144 - 3 \times 35.5 = 33.5 \]
\[ 144 + 3 \times 35.5 = 246.5 \]

**Coefficient of Variation (C.V.):** is the standard deviation expressed as a percentage of the mean. It is a unit-free measure of dispersion. It provides a measurement for comparing relative variability of data sets from different scales.

\[ \text{C.V.} = \frac{s}{\bar{x}} \cdot 100\% \]

**Example:** As part of the Berkeley Guidance Study. The heights (in cm) and weights (in kg) of 13 girls were measured at age two. The average height was 86.6 cm with a s.d.= 2.9 and average weight is 12.6 kg with a s.d.= 1.4.

C.V. (height) = \( \frac{2.9}{86.6} \times 100\% = 3.3\% \)

C.V. (weight) = \( \frac{1.4}{12.6} \times 100\% = 11.1\% \)

More variability is weight than in height

**Measure of Position**

Standard Score, Percentile, Quartile
If $x$ is an observation from a distribution that has mean $\mu$, and standard deviation $\sigma$, the standardized value of $x$ is,

$$z = \frac{x - \mu}{\sigma}$$

"$\mu + 3\sigma$" has a z-score 3, since it is 3 s.d. from mean.

If a distribution has a mean 10 and a s.d. 2, the value 7 has a z-score $-1.5$.

$$z = \frac{7 - 10}{2} = -1.5.$$
Descriptive Statistics

Example: [even number of data] \( (n = 22) \)

\[ \begin{align*}
6, 60, 61, 63, 64, 64, 65, 65, 65, 66, 67, 69, 71, 71, 71, 72, 72, 72, 72, 73, 74, 75
\end{align*} \]

\[ Q_1 = 64 \quad \text{Median} = 68 \quad Q_3 = 72 \]

Measure of spread:

**Interquartile range** \( (IQR) = Q_3 - Q_1 \)

\[ IQR = 72 - 64 = 8 \]

---

The five-number summary

- **Minimum value**
- **\( Q_1 \)**
- **Median**
- **\( Q_3 \)**
- **Maximum value**

---

Example: (data sheet without outlier “6”)

\[ \begin{align*}
60, 61, 63, 64, 64, 65, 65, 65, 66, 67, 69, 71, 71, 71, 72, 72, 72, 72, 73, 74, 75
\end{align*} \]

\[ \text{Min} = 60, \quad Q_1 = 64.5, \quad \text{Median} = 69, \quad Q_3 = 72, \quad \text{Max} = 75. \]

---

With 6 in the data:

\[ \begin{align*}
6, 60, 61, 63, 64, 64, 65, 65, 65, 66, 67, 69, 71, 71, 71, 72, 72, 72, 72, 73, 74, 75
\end{align*} \]

\[ Q_1 = 64 \quad \text{Median} = 68 \quad Q_3 = 72 \]

\[ IQR = 72 - 64 = 8 \]

---

**Inner and outer fences for outliers**

- The inner fences are located at a distance of 1.5 \( IQR \) below \( Q_1 \)
  
  **lower** inner fence = \( Q_1 - 1.5 \times IQR \)
  
  and at a distance of 1.5 \( IQR \) above \( Q_3 \)
  
  **upper** inner fence = \( Q_3 + 1.5 \times IQR \).

- The outer fences are located at a distance of 3 \( IQR \) below \( Q_1 \)
  
  **lower** outer fence = \( Q_1 - 3 \times IQR \)
  
  and at a distance of 3 \( IQR \) above \( Q_3 \)
  
  **upper** outer fence = \( Q_3 + 3 \times IQR \).
Descriptive Statistics

Mild and Extreme outliers
- Data values falling between the inner and outer fences are considered **mild outliers**.
- Data values falling outside the outer fences are considered **extreme outliers**.

When outliers exist, the whisker extended to the smallest and largest data values **within the inner fence**.

Boxplot
- Find the five-number-summary and make a boxplot for the following data:
  13 72 78 40 50 56 50 52 57
  69 130 142 51 52

Remarks:
- If the distribution of the data is symmetric, then the **mean** and **median** will be about the same.
- The five-number summary is best for nonsymmetric data.
- The **median** and **quartiles** are not influenced by outliers.
- The **mean** and **standard deviation** are most appropriate to use only if the data are symmetric because both of these measures are easily influenced by outliers.