Measures of Variability
Measures of Dispersion (Variability)

- I previously introduced **central tendency** as a statistical measure that describes the center of the distribution of scores in a population or sample.

- However, distributions also vary in their **dispersion** or **variability**—the degree to which individual data points tend to vary from the central tendency.
Differences in Variability

- Left: $n=20$, $M=2.64$
- Right: $n=20$, $M=3.26$
Candidate Measures

- Range
- Interquartile range
- Average deviation
- Variance & standard deviation
Range

- The **range** is the difference between the highest and lowest scores in a population or sample
  - Example in common use: high and low daily temperatures

- The range is completely dependent on **outliers**, or extreme values
Range

N = 600
range = 3.72

N = 601
range = 8.28
Quantiles, Percentiles, and Percentile Ranks

• The relative location of individual scores within a distribution can be described by quantiles (or fractiles) and quantile ranks.

• **Quantiles** are measures that divide data into two or more equal parts, depending on their rank ordering
  – Percentiles, quartiles, and deciles are typical examples of quantiles

• The **percentile rank** for a particular $x$ value is the percentage of individuals with scores equal to or less than that $x$ value.

• When an $x$ value is described by its rank, it is called a **percentile**.
Interquartile Range & Other Range Statistics

• **Interquartile range (IQR)** is the range of the middle 50% of observations

• To compute upper ($Q_3$) and lower ($Q_1$) quartiles, first order the data by ascending values

• Method 1 (my preferred method):
  – Compute the lower and upper **quartile locations** or **ranks**
  – Lower quartile location ($25^{th}$ percentile) = $0.25 \times N$
  – Upper quartile location ($75^{th}$ percentile) = $0.75 \times N$
  – Round to **quartile location** to nearest whole number, the corresponding value is the **quartile**
Interquartile Range & Other Range Statistics

Computing interquartile range (continued)

- Method 2 (from the book):
  - Find the median and split the data into upper and lower halves
  - Find the median for the lower half: this is the lower quartile
  - Find the median for the upper half: this is the upper quartile

- Use either method to compute quartiles. For large data sets, they give similar answers

- Interquartile range (IQR) = Q₃ − Q₁
Interquartile Range

Example:

- $N=13$
- Lower quartile rank = $0.25 \times 13 = 3.25$
- Lower quartile ($Q_1$) = 2
- Upper quartile rank = $0.75 \times 13 = 9.75$
- Upper quartile ($Q_3$) = 4
- IQR = $Q_3 - Q_1 = 4 - 2 = 2$

<table>
<thead>
<tr>
<th>Scores</th>
<th>Sorted Scores</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>13</td>
</tr>
</tbody>
</table>
Interquartile Range & Other Range Statistics

• Although quartiles are often used, there’s nothing special about quartiles

• Any percentile can be used to compute range or to compute trimmed statistics

• A sample with a percentage of the extreme scores removed is called a trimmed sample

• Trimmed statistics are statistics (e.g., means or ranges) computed on trimmed samples

• All range measures suffer from the problem that they only consider extreme scores in a data set (or trimmed data set)
Average Deviation

- The **average deviation** of scores from the mean might seem like a good alternative measure that takes into account all the scores in a sample or population.

  \[
  \text{average deviation} = \frac{1}{N} \sum_{i} (x_i - M)
  \]

- However, there’s a problem. For example, compute the average deviation for this set of numbers

  \[
x = \{2, 1, 3, 4, 9, 5\}
  \]

  hint: \(M = 4.0\)
Variance & Standard Deviation

- One way to remedy this problem is to take the average of the squared deviations. The resulting value is called the variance.

  \[ SS = \sum_{i} (x_i - M)^2 \]

  population variance: \( \sigma^2 = \frac{SS}{N} \)

- To express the variability in terms of score units, we usually use the square root of this value, the standard deviation.

  population standard deviation: \( \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}} \)
Sample Variance & Standard Deviation

- When measuring variance and standard deviation for samples, we use a denominator of $n - 1$ instead of $N$.

  \[
  \text{sum of squared deviations: } SS = \sum_{i} (x_i - M)^2
  \]

  \[
  \text{sample variance: } s^2 = \frac{SS}{n - 1}
  \]

  \[
  \text{sample standard deviation: } s = \sqrt{s^2} = \sqrt{\frac{SS}{n - 1}}
  \]
Differences in Variability

$n=20$
$M=2.64$
$s=0.66$

$n=20$
$M=3.26$
$s=0.07$
Computational Formulae for Variance and SD

$$SS = \sum x^2 - \left( \frac{\sum x}{N} \right)^2$$

$$s^2 = \frac{SS}{n-1} \quad s = \sqrt{\frac{SS}{n-1}}$$
Measures of Variability

Computing Sample Variance and SD

| x  | 5 | 6 | 4 | 3 | 7 | 5 |

**Definitional Formula**

\[ SS = \sum (x - M)^2 \]

**Computational Formula**

\[ SS = \sum x^2 - \left( \frac{\sum x}{N} \right)^2 \]

\[ s^2 = \frac{SS}{N - 1} \]

\[ s = \sqrt{\frac{SS}{N - 1}} \]
The Mean & Variance as Estimators

- The **expected value** of a statistic is its long-range average over repeated sampling
  - Notated as $E()$. For example the expected value of $s$ is denoted $E(s)$

- In statistics, **bias** is a property of a statistic whose expected value does not equal the parameter it represents
  - $s$ is an **unbiased** estimator of $\sigma$ (i.e., $E(s) = \sigma$)
  - $M$ is an **unbiased** estimator of $\mu$ (i.e., $E(M) = \mu$)
A Note about N-1

Population  \( x = \{A,B,C\} \), with \( A=8 \), \( B=5 \), \( C=2 \)

We know that the actual variance \( \sigma^2 \) is 6.0, what happens if we sample randomly from the population and estimate \( \sigma^2 \) using the population variance formula?
A Note about N-1

- Because the sample variance $s^2$ is estimated using the sample mean (an estimate), it is computed with only $n-1$ degrees of freedom
  - given the sample mean and $n-1$ out of $n$ observations in a sample (e.g., $x_1, \ldots, x_{(n-1)}$), we can determine the value of the missing observation ($x_n$)

- As a result, the average squared deviation of a sample is a biased estimator of $\sigma^2$

$$E \left[ \frac{\sum (x - M)^2}{n} \right] = \frac{(n-1)\sigma^2}{n}$$
A Note about $N-1$

$$E \left[ \frac{\sum (x-M)^2}{n-1} \right] \frac{\sigma^2}{n-1} = E \left[ \frac{\sum (x-M)^2}{n-1} \right] = \sigma^2$$