Measures of Variability (Dispersion)

- Summarized a distribution's shape and central tendency
- Now focus on summarizing a distribution's dispersion or variability:

Range

- the range is the difference between the highest and the lowest values
  - range = MAX(data) - MIN(data)
- The range is sensitive to extreme values
Quartiles

- Any distribution can be divided into quartiles.
- Quartiles divide sorted data into four subsets.
  - Each subset contains (1/4) or 25% of the data.
- The range between the first quartile and the third quartile (holding 50% of the data) is the interquartile range (IQR).

Finding Quartiles

- First, rank your data from smallest to largest value.
- Second, find the median location \((N+1)/2\).
- Find the quartile location: \(Q = (\text{median location}) + 1\) / 2.
- 1st quartile is the quartile location from the lowest score, the third quartile is the quartile location from the highest score.
- Interquartile Range (IQR) is the difference between 1st & 3rd quartiles.

Interquartile Range

- Interquartile Range (IQR) is
  - The difference between 3rd & 1st quartile.
  - Equals range of the middle 50% of the distribution.
  - Equals the range of 25% trimmed sample.
- Stable against outliers.
  - But doesn’t describe the spread of the full data set.
  - Two distributions could have same IQR but very different ranges.
Boxplots [aka box-and-whisker plots]

1. Find the median (located at \( \frac{N+1}{2} \)) → draw a horizontal line
2. Find Q1 and Q3 ("hinges") → draw lines at each
3. Draw a line from the top and bottom of the box to the farthest point that is no more than 1.5x the IQR from the box = whiskers
4. Draw asterisks to represent points that lie outside whiskers, i.e. outliers [none in this data set]

1st Quartile (25%) = 107
2nd Quartile (50%) = 112
3rd Quartile (75%) = 115
max whisker length = 1.5 \times 8 = 12

Langlois & ;, 1990

Composite of 4 faces
Composite of 32 faces

Showing “spread” in box plots

Data from Table 5.1

Showing Skew in Box plots

Positive Skew
Negative Skew

Data Set 1
Data Set 2

Score

Data Set 1
Data Set 2

Score
A Note About Outliers

- Outliers are extreme values in data
- Sometimes they represent real data
- Sometimes they are experimental errors
  - e.g., the subject fell asleep while during experimental task
  - e.g., the result was recorded incorrectly by the experimenter
  - e.g., the experimental equipment malfunctioned
- Outliers based on experimental errors should be removed from data
- Outliers representing real data should be kept and considered in analysis
- Always be wary of studies that throw out lots of data!

Variability around the mean

- How can we measure variation around the mean?
- Common sense: average deviation from mean: \[ \sum (Y_i - \bar{Y}) \]
  - Surprisingly, this idea doesn’t work:
    - the sum of deviations from mean is always ZERO
    - positive and negative deviations always cancel each other
  - So the sum of deviations is the same regardless of variability!
- One solution, get rid of the sign of the deviation
  - use absolute values [Mean Absolute Deviation]
  - squared deviations [Variance]

Variance

- Calculate the sum of squared deviations
- Divide by (n-1), not
- Result is the sample variance:
  \[ S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \]
  - n is the number of observations in the sample
  - (n-1) is the degrees of freedom
- Why (n-1) and not n?
  - using n-1 makes \( s^2 \) an unbiased estimator of population variance, \( \sigma^2 \)

Standard Deviation

- One problem with variance is that the units aren’t the same as for our \( Y_i \) values (e.g. if \( Y_i \) is in cm, \( s^2 \) in cm$^2$)
  - Also, the value for the variance is much higher than any of the deviations of the individual data points
- We can instead calculate the standard deviation (S.D.) of the sample, which is:
  \[ S = \sqrt{S^2} \]
Sample Statistics vs. Population Parameters

• The mean of our sample, \( \bar{Y} \), is an estimate of the mean of our population of interest, \( \mu \).

• The standard deviation of our sample, \( s \), is an estimate of the standard deviation of our population \( \sigma \).

• If a statistic is a good estimate of a parameter, we say it is unbiased.

---

Statistics vs. Parameters

<table>
<thead>
<tr>
<th></th>
<th>Central Tendency</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Statistic</strong></td>
<td>( \bar{Y} = \frac{\sum Y}{n} )</td>
<td>( s = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n-1}} )</td>
</tr>
<tr>
<td><strong>Population Parameter</strong></td>
<td>( \mu = \frac{\sum Y}{N} )</td>
<td>( \sigma = \sqrt{\frac{\sum (Y - \mu)^2}{n}} )</td>
</tr>
</tbody>
</table>

Note: the denominator for \( s \) contains \( n-1 \) because we have estimated \( \bar{Y} \) to calculate it. The denominator for \( \sigma \) contains \( n \) because we directly calculated \( \mu \).

---

Part 2 - Variability (summary)

• Common measures of variability:
  - range
  - interquartile range (IQR)
  - variance & standard deviation

• Quartiles
• Box plots
• Sample Statistics vs. Population Parameters