Chapter 2 – Describing Distributions with Numbers

Example: Phyllis had 6 assignment grades in her Stat2331 class:
86 88 92 44 89 90.

Idea: We want a few numbers that can summarize a list of observations.

Measures of Center

I. Mean: Sum of all the observations divided by the number of observations in the list.

II. Median: The middle number when the observations are put in order. To calculate median:
   a) sort observations from smallest to largest
   b) if n is odd (n = number of observations)
      median = middle value of the sorted list
      \[ \frac{n + 1}{2} \text{ th observation up from the bottom of the list} \]
   c) if n is even median = mean of middle two observations

Example: Phyllis grades (cont’d)

Her mean grade is:

Her median grade is:

Issue: Which one to use?

Q: Does the mean, 81.5, give a good idea of her “typical” grade?

What about the median, 88.5?
Suppose we exclude the outlier 44. We have

<table>
<thead>
<tr>
<th></th>
<th>$M$</th>
<th>$\bar{x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>With the outlier 44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without the outlier 44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

→ The mean is pulled by extreme observations or outliers. So it is not a resistant measure of center.

→ The median is not pulled by the outliers. So it is a resistant measure of center.

**Q:** In figuring Phyllis’ grade, her instructor needs to know total HW points she scored but lost this info. Which would be more useful to him

a) her mean score  

b) her median score

→ When the distribution is skewed:

For “typical” value, use  
But if interested in total, use

**Issue:** How do they relate to each other on various distributions?

Symmetric:

Right-skewed:

Left-skewed:
The mean is pulled by the extremes in the tail, but median is not.

**Example.** In a city, incomes have mean $45,000 and median $31,000. Which of these is a possible histogram of incomes?

**Issue:** How are the values spread around the mean?

**Example.** At two diners, 3 monthly water bills (in $) are:

*Granny's:* 101, 100, 99

*Pop's:* 90, 110, 100

How do the bills of the two diners differ?
Measures of Spread

I. **Standard deviation** ($s$): measures the “average” distance between each observation and the mean of the data. To calculate $s$:

(a) **Calculate variance** ($s^2$):
- Compute squared distance between each observation and the mean of the data.
- Find an “adjusted” average of these distances.

(b) Take square root of the variance to get $s$.

**Ex:** Water bills of diners (cont’d).

**Granny’s**

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$x_i - \bar{x}$</th>
<th>$(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variance ($s^2$) = 

SD ($s$) =

Similarly, for Pop’s we get $s^2 = \_\_\_\_\_\_$ and $s = \_\_\_\_$.
Behavior & Properties of Standard Deviation

- $s$ measures the spread about the mean and should be used only when the mean is chosen to measure the center, i.e.,

- $s$ is in the same units as the original observations, e.g.,

- $s \geq 0$. The more the data are spread out from the mean, the smaller / higher the $s$.

- If all numbers in a list are same, $s = ____$. 

- How would outliers affect the $s$?

- $s$ is / is not a resistant measure. Should not be used to describe a

**Ex:** Phyllis HW grades contd.: 86 88 92 44 89 90

<table>
<thead>
<tr>
<th></th>
<th>$\bar{x}$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>With the outlier 44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without the outlier 44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ex:** Choose 4 numbers from the list 0, 1, ..., 10, repeats allowed, such that
(a) they have the smallest possible SD.

(b) they have the largest possible SD.

**II. Five-Number Summary:** Uses five numbers to divide the whole distribution in four equal parts. One-quarter of all the observations are covered between two consecutive numbers.
The five numbers are min, $Q_1$, $Q_2$, $Q_3$, & max. $Q_1$, $Q_2$, & $Q_3$ are the 1st, 2nd & 3rd quartiles of a distribution.

$Q_1$ is the number at or below which 25% of the observations fall.

$Q_2$ is the number at or below which 50% of the observations fall. So $Q_2 =$

$Q_3$ is the number at or below which 75% of the observations fall.

To calculate the quartiles:

1. **Sort the data** in increasing order and locate the median $M$ of this list. This is $Q_2$.

2. Find the median of the ordered observations located to the left of the overall median $M$. This is $Q_1$.

3. Find the median of the ordered observations located to the right of the overall median $M$. This is $Q_3$.

**Example. Home Run Data**

<table>
<thead>
<tr>
<th>Babe Ruth</th>
<th>54 22 41 35 41 34 60 46 46 54 47 25 49 59 46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger Maris</td>
<td>26 13 33 61 23 8 28 14 39 16</td>
</tr>
</tbody>
</table>

The five-number summaries are:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maris</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We can graphically display the five-number summary by a **boxplot**.

A boxplot consists of:

- A central box that spans the distance between $Q_1$ and $Q_3$.
- Horizontal lines marking min., median and max.
- Vertical lines extending from the bottom and top of the box to the min. and max. observations, respectively.

Boxplots of Home run data:

**Notes:**

1. Boxplots are best used for side-by-side comparison of more than one distribution.
2. Make sure that you include a numerical scale while drawing boxplot.
3. For skewed distributions, use five-number summary and boxplots.
4. Boxplots give an indication of the symmetry or skewness of a distribution. In a symmetric distribution, the first and third quartiles are equally distant from the median.
Example:

Consider the box plot at the left, which summarizes the systolic blood pressures of 39 adult males.

Q: Estimate the 5-number summary from the boxplot.

\begin{align*}
\text{Min} &= \\
\text{Q}_1 &= \\
\text{Q}_2 &= \\
\text{Q}_3 &= \\
\text{Max} &=
\end{align*}

Five number summary and boxplot give us an idea about the whole distribution. To get an idea about the spread only, we can use \textbf{Inter-Quartile Range (IQR)}:

\begin{align*}
\text{Ex: Home Run data} \\
\text{IQR for Ruth} &= \\
\text{IQR for Maris} &=
\end{align*}
Outline of Chapter 1 and 2

Graphing Distributions

1) Categorical Data
   a) 
   b) 

2) Quantitative Data
   a) 
   b) 
   c) 
   A distribution is described by its
   1) 
   2) 
   3) 
   4) 

Numerical Summaries

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>5-Num/IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resistant Measure?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use for Symmetric Dist.?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use for Skewed Dist.?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>