Chap 2 Notes: Summarizing Data by Numerical Measures: Center and Spread

J. Harner A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

Numerical Measures of Center of a Data Set

Definition: An average is a single value which represents all the data.

Types of averages:

1. Sample Mean
2. Sample Median
3. Sample Mode

Sample Mean

Definition: The sample (arithmetic) mean, denoted by $\overline{X}$, is obtained by adding all the observed values of a numeric variable and dividing by the sample size (n). It is the most commonly used measure of the center of a data set.

$$\overline{X} = \frac{\sum x}{n}$$

Example:

X values	14	23	8	19	41

$$\begin{aligned}\overline{X} &= \frac{\sum x}{n} \\ &= \frac{14 + 23 + 8 + 19 + 41}{5} \\ &= \frac{105}{5} \\ &= 21\end{aligned}$$

Population Mean

Definition: The population (arithmetic) mean, denoted by μ, is obtained by adding values measured or counted on each of the elements of the population and dividing by the number of elements in the population. Generally, μ cannot be computed since only the sample values are known.

$$\mu = \frac{\sum x}{N},$$

where N is the population size.

Note: The sample mean, $\overline{X}$, is an estimator of μ, the population mean.

Sample Median

Definition: The sample median, denoted by M (or $\tilde{X}$), is the middle value of the observed values of a numeric variable ranked from the lowest to the highest.

Note: If the sample contains n observations, the middle value, i.e., M, is the $\frac{1}{2}(n+1)^{\text{th}}$ ranked value.

Note: $\frac{1}{2}(n+1)^{\text{th}}$ is the position of the median in the ranked data.

Computing the Sample Median

The procedure (algorithm) for computing the sample median is:

1. Rank the data (from low to high values).
2. Determine the position of M, i.e., the $\frac{1}{2}(n+1)^{\text{th}}$ ranked value.
3. Locate the median M ($\tilde{X}$).

Sample Median Example I

Example:

X values	14	23	8	19	41
Ranked values	8	14	19	23	41

Position of M:

$$\begin{aligned} \frac{1}{2}(n+1) &= \frac{1}{2}(5+1) \\ &= 3 \end{aligned}$$

Locate M:

$$M = 19$$

Note: n is odd in this example.

Sample Median Example II

Example:

X values	43	26	37	19	52	80
Ranked values	19	26	37	43	52	80

Position of M:

$$\begin{aligned}\frac{1}{2}(n+1) &= \frac{1}{2}(6+1) \\ &= 3.5\end{aligned}$$

Locate M:

$$\begin{aligned}M &= \frac{37+43}{2} \\ &= \frac{80}{2} \\ &= 40\end{aligned}$$

Note: n is even in this example.

Sample Mode

Definition: The sample mode is the observed values which occurs with the greatest frequency.

Example:

X values	14	23	8	19	41

$$\text{Mode} = \text{None},$$

since each value occurs only once.

Do not write:

$$\text{Mode} = 0.$$

Sample Mode Example

X values	4	3	1	4	0	3	3
	1	4	0	1	1	2	2

Example:

Class	Freq
0	2
1	4
2	2
3	3
4	3

$$\text{Mode} = 1$$

Note: A set of data may have more than one mode.

Sample Bimodal Example

X values	5	8	5	6	2	7
	6	5	3	4	2	6

Example:

Class	Freq
2	2
3	1
4	1
5	3
6	3
7	1
8	1

$$\text{Mode} = 5$$

and

$$\text{Mode} = 6$$

Which Measure of Center Should be Used?

For the dataset:

X values	14	23	8	19	41

$$\begin{aligned} \overline{X} &= 21 \\ M &= 19 \\ \text{Mode} &= \text{None} \end{aligned}$$

Which “average” should we use?

Which “Average” Should We Use?

Which “average” should we use?

- Use $\overline{X}$ for (approximately) symmetric distributions;
- Use M ($\tilde{X}$) for “significantly” skewed distributions;
- Do not use the mode!

Note: Outliers affect the value of $\overline{X}$ much more than the value of M.

How to We Handle Outliers?

How to we handle outliers?

- Eliminate an outlier if it is a “mistaken” data value;
- Do not eliminate an outlier if it an actual data value.

We can expect a few outliers in any reasonably sized set of data.

Numerical Summaries

To summarize or represent data numerically, we need:

- a measure of center (average);
- a measure of dispersion (spread or variation);
- a measure of skewness.

Variation is important and will be covered in this subsection.

Measuring Spread

The spread can reasonably be based on how the X values vary about $\overline{X}$. But how?

Since $\sum(X - \overline{X}) = 0$, as can be shown algebraically, the sum of the deviations about the sample mean is not a measure of spread.

Two choices seem plausible:

- $\sum(X - \overline{X})^2$, the sum of squared deviations;
- $\sum |X - \overline{X}|$, the sum of absolute deviations.

The first choice, i.e., the sum of squared deviations, is widely used for data which is approximately symmetrically distributed..

The Sample Variance

The sample variance is the average of the sum of squared deviations. However, instead of dividing by n, which would give a true mean of the squared deviations, $n - 1$ is used for statistical reasons, which will be explained later.

$$
\begin{aligned}
s^2 &= \frac{\sum (X - \overline{X})^2}{n-1} \\
&= \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}.
\end{aligned}
$$

Note: $\sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$, as can be shown algebraically, and the right-hand side is easier to compute on a calculator.

The Sample Standard Deviation

The sample standard is the square root of the variance. That is:

$$\begin{aligned} s &= \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}} \\ &= \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}. \end{aligned}$$

Standard Deviation Example

We want to compute the sample standard deviation for the dataset:

X values 14 23 8 19 41

$$\overline{X} = \frac{\sum X}{n} = \frac{14 + 23 + 8 + 19 + 41}{5} = \frac{105}{5} = 21$$

$$\begin{aligned}\sum(X - \overline{X})^2 &= (14 - 21)^2 + (23 - 21)^2 + \cdots + (41 - 21)^2 \\ &= 49 + 4 + \cdots + 400 \\ &= 626\end{aligned}$$

Thus:

$$\begin{aligned} s &= \sqrt{\frac{626}{5 - 1}} \\ &= \sqrt{156.5} \\ &= 12.5 \end{aligned}$$

Interpretation of s

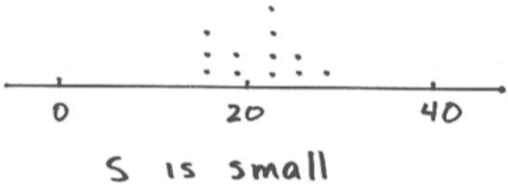
Values of s indicate the spread of the data:

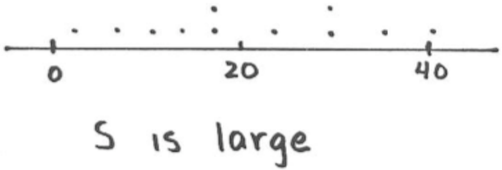
1. $s \approx 0 \Longrightarrow$ little or no variation in the data, i.e., nearly all values are the same;
2. s "small" $\Longrightarrow$ the data values are not widely dispersed;
3. s "large" $\Longrightarrow$ the data values are widely dispersed.

Notes:

1. The unit of measurement associated with s is the same as the unit of the variable.
2. Round off s to one more decimal place than the data.

Plots of the Spread





The Population Standard Deviation

The population variance and standard deviation are given by:

$$\begin{aligned}\sigma^2 &= \frac{\sum(X-\mu)^2}{N} \\ \sigma &= \sqrt{\frac{\sum(X-\mu)^2}{N}}\end{aligned}$$

Since generally we do not know the value of μ, we cannot compute σ^2 or σ.
Thus we use:

1. s^2, the sample variance, to estimate σ^2, the population variance;
2. s, the sample standard deviation, to estimate σ, the population standard deviation.

Linear Transformations

Suppose we need to change the units of measurement, e.g., from lbs to kg or from F to C.

How would such a change affect the mean, median, and standard deviation?

1. Adding/Subtracting the same constant from each data value shifts the center (mean or median) by the same amount. The spread (standard deviation) is unchanged.
2. Multiplying each data value by a positive constant c also multiplies the mean, median, and standard deviation by the same factor c.

LifeStats: See Example 2.12

Measuring Skewness

Skewness is a measure of how much a distribution deviates from symmetry.

One possible measure of skewness is:

$$\text{Skewness} = \frac{3(\overline{X} - \tilde{X})}{s}$$

1. Skewness $\approx 0 \Longrightarrow$ the distribution is approximately symmetric;
2. Skewness $> 1 \Longrightarrow$ the distribution is markedly positively skewed;
3. Skewness $< 1 \Longrightarrow$ the distribution is markedly negatively skewed.

Skewness Example

For the data:

X values	14	23	8	19	41

We found $\overline{X} = 21$, $\tilde{X} = 19$, and $s = 12.5$.

Thus,

$$\begin{aligned} \text{Skewness} &= \frac{3(\overline{X} - \tilde{X})}{s} \\ &= \frac{3(21 - 19)}{12.5} \\ &= \frac{6}{12.5} \\ &= 0.48 \end{aligned}$$

The distribution of data is slightly positively skewed.

Normal Approximation

Quartiles

Quartiles are numbers which partition the data into 4 subgroups of (approximately) equal size.

- Q_1 is called the lower (or first) quartile. $\approx$ 25% of the data values are $\leq Q_1$.
- Q_2 is called the second quartile (or $\tilde{X}$). $\approx$ 50% of the data values are $\leq Q_2$.
- Q_3 is called the upper (or third) quartile. $\approx$ 75% of the data values are $\leq Q_3$.
- Q_4 is called the fourth quartile. $\approx$ 100% of the data values are $\leq Q_4$.

The Sample Interquartile Range (IQR) is defined by:

$$\text{IQR} = Q_3 - Q_1$$

Quartile Example

Ranked X values ($n = 12$):

$$4 \quad 5 \quad 8 \;|\; 11 \quad 16 \quad 23 \;|\; 24 \quad 29 \quad 31 \;|\; 38 \quad 41 \quad 44$$

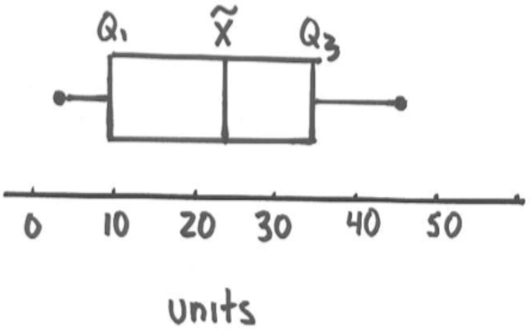
- $Q_2 = 23.5$ is the median of the entire data set.
- $Q_1 = 9.5$ is the median of the data left of Q_2.
- $Q_3 = 34.5$ is the median of the data right of Q_2.

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 34.5 - 9.5 \\ &= 25 \end{aligned}$$

Boxplots

The five-number summary statistics (the minimum value, $Q_1, \tilde{X}, Q_3$, and the maximum value) can be used to draw a boxplot.

For the previous (quartile) example:



Boxplots can be used to graphically represent the distribution of the data (LifeStats: Example 2.14).

Outlier Boxplots

An outlier boxplot is used to identify (potential) outliers.

Although the box is the same as the box in the ordinary boxplot, the "tails" are plotted differently.

To construct an outlier boxplot:

1. Draw a box using Q_1 and Q_3 as the sides;
2. Compute the IQR, i.e., $d = Q_3 - Q_1$;
3. Draw vertical lines at $1.5 \times d$ above Q_3 and $1.5 \times d$ below Q_1;
4. Draw vertical lines at $3 \times d$ above Q_3 and $3 \times d$ below Q_1.

Outlier Boxplot Example

For the previous (quartile) example:

$$\begin{aligned} d &= Q_3 - Q1 \\ &= 34.5 - 9.5 \\ &= 25 \\ 1.5 \times d &= 37.5 \\ 3 \times d &= 75 \end{aligned}$$