

MEASURES OF CENTRAL TENDENCY

**DR. BORNWELL SIKATEYO
UNZA SCHOOL OF MEDICINE**

Learning objectives

At the end of this lecture, you should be able to:

- Appreciate the use and calculation of means, medians and modes which are measures of central tendency

Measures of Central Tendency

- A *measure of central tendency* is a *univariate* statistic that indicates, in one manner or another,
 - the *average* or *typical* observed value of a variable in a data set, or
 - put otherwise, the *center* of the frequency distribution of the data.

How to Calculate Averages

- We consider three measures of central tendency that are appropriate for three different levels of measurement (nominal, ordinal, and interval).
- These are the mean, median and mode.
- The measures of central tendency, can be calculated from three different “starting points”:
 - a *list of observed values* (i.e., raw data, e.g., one column in a data spread sheet);
 - a *frequency table* (or bar graph);
 - a *histogram* and.

The Mode

- The *mode* (or *modal value*) of a variable in a data set is the value of the variable that is observed *most frequently* in that data (or, given a continuous frequency curve, is at the point of *greatest density*).
 - *Note*: the mode is the *value* that is observed most frequently, not the *frequency* itself.
- The mode is defined for *every* type of variable [i.e., *nominal*, *ordinal*, *interval*, or *ratio*].
 - However, the mode is used as a measure of central tendency primarily for nominal variables only.

How To Calculate the Mode

Cases = Students (n = 13)

Variable = # Problem Sets Turned In

Raw Data (Observed Values)

4 5 5 3 4 5 5 3 1 2 3 5 5

Given a list of observed values (raw data):

Construct a frequency table (see next slide).

How To Calculate the Mode (cont.)

Frequency Distribution

	<u>Values</u>	<u>Abs. Frq.</u>	<u>Rel. Frq.</u>	<u>Cum. Frq</u>	
	0	0	0%	0%	100%
	1	1	8%	8%	100%
	2	1	8%	16%	92%
	3	3	23%	39%	84%
	4	2	15%	54%	61%
(Value with the greatest [absolute or relative] frequency	Mode 5	6	46%	100%	46%
	Total	13	100%		

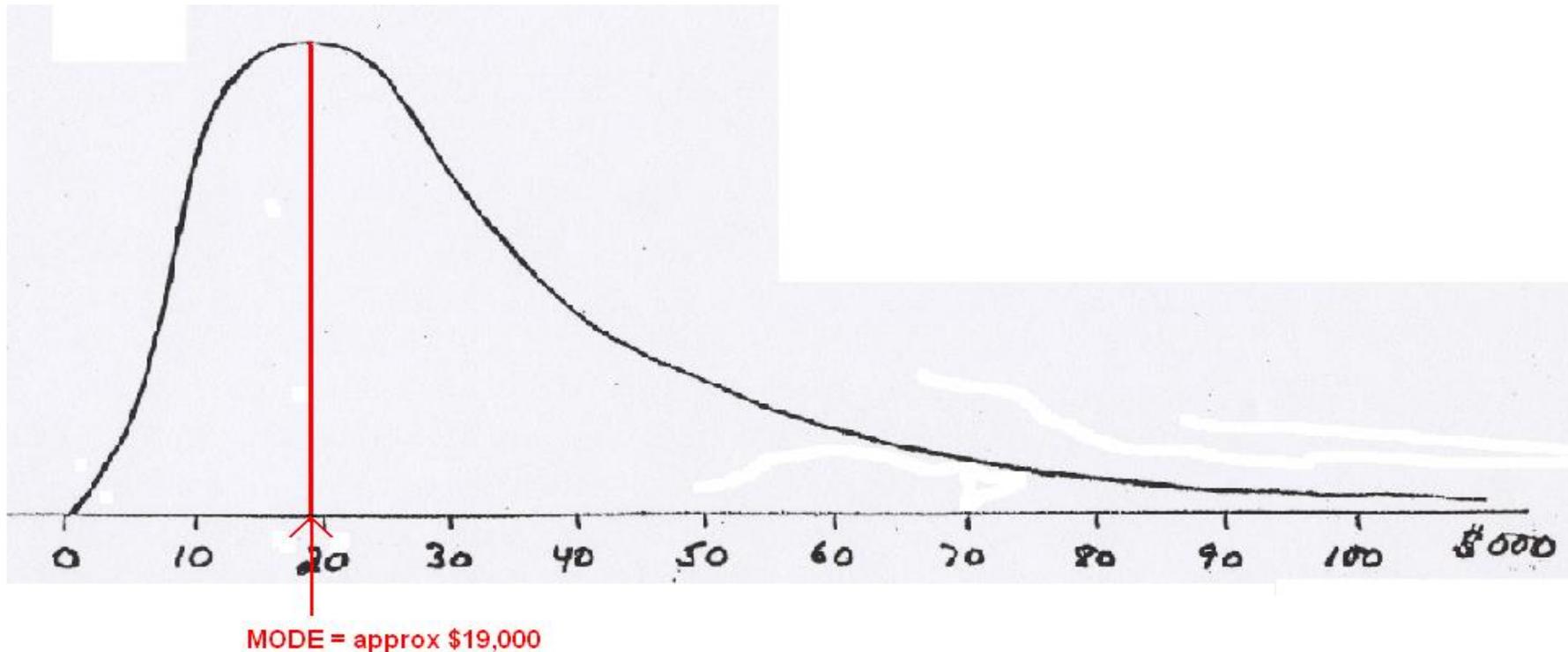
- Given a frequency table or bar graph (or having constructed one):
 - observe which value in the table (or graph) has the *greatest* (absolute or relative) *frequency*;
 - the most frequent *value* (not the *frequency* itself) is the mode.

How To Calculate the Mode (cont.)

- Notice that the modal number of problem sets turned in is 5, even though most students turned in fewer than 5,
- So if we recoded the variable to create just two dichotomous categories:
 - (a) turned in all 5
 - (b) did not turn in all 5
- the latter (i.e. did not turn in all 5) category becomes the *modal category*

How To Calculate the Mode (continued)

- *Given a continuous frequency curve:*
 - the mode is the *value* of the variable under the *highest point* of the frequency curve (the point with the greatest density of observed values).



Advantages of the mode

- It is easy to understand and calculate.
- It is not affected by extreme large or small values.
- It can be useful for qualitative data.
- Can be located graphically.

Disadvantages of the mode

- Its computation is not based on all values as is the case for the mean.
- It will not be well defined if the data consists of small number of values (it is possible that there can be more than one modal value).
- It is not capable of further mathematical treatment.
- Sometimes the data may not have a mode at all.

The Median

- The *median* (or *median value*) of a variable in a data set is:
 - the value in the *middle* of the observations, in the sense that no more than *half* of the cases have *lower* values and no more than half of the cases have *higher* values or,
 - more generally, such that no more than half of the cases have values that lie on either side of the median value.

The Median

- Given a quite precisely measured continuous variable and a very large number of cases, we can in practice say that:
 - half the cases have lower values and half have higher values (e.g., LEVEL OF INCOME).
 - Equivalently, the median value is the value of the case at the 50th percentile of the distribution.
- The median is defined if and only if the variable is *at least ordinal* in nature [i.e., *ordinal, interval, or ratio*].
- We can therefore rank all (non-missing) observations in terms of lower to higher values.

The Median (continued)

- *Given a list of observed values (raw data):*
 1. rank order the cases in terms of their observed values (e.g., from lowest to highest);
 2. identify the value of the case right at the *middle of this rank-ordered list, and*
 3. the value of this case is the median value; *or*
 4. construct a frequency table and find where the *cumulative* frequency crosses the 50% mark.

The Median (continued)

Raw Data

4 5 5 3 4 5 5 3 1 2 3 5 5

Rank Ordered Observed Values

1 2 3 3 3 4 4 5 5 5 5 5 5

The Median (continued)

Rank Ordered Observed Values

1 2 3 3 3 4 4 5 5 5 5 5 5

(observed value in the middle of the ranking)

Five cases (less than half) have lower values and six cases (less than half) have higher values

Frequency Distribution

Values	Abs. Frq.	Rel. Frq.	Cum. Frq	
0	0	0%	0%	100%
1	1	8%	8%	100%
2	1	8%	16%	92%
3	3	23%	39%	84%
4	2	15%	54%	61%
5	6	46%	100%	46%
Total	13	100%		

(Value at which the cumulative frequency crosses the 50% threshold)

MEDIAN →

The Median (continued)

- If the number of cases is even, there is no observed value at the exact middle of the list.
- Look at the pair of observations closest to the middle of the list.
- If they have the same value, that value is the median.
- If they have different values, every value in the *interval* bounded by these two values meets the definition of a median; but conventionally the median in this event is defined as the *midpoint* of the interval.

The Median (continued)

Original Raw Data

4 5 5 3 4 5 5 3 1 2 3 5 5

Delete one observed 5 (n = 12)

1 2 3 3 3 4 4 5 5 5 5 5 ~~5~~

(Middle Pair of
Observed Values)

Median = 4

Add one observed 5 (n = 14)

1 2 3 3 3 4 4 5 5 5 5 5 5 5

(Middle Pair of
Observed Values)

Strictly: Median = interval from 4 to 5

But conventionally: median = midpoint of interval, i.e., 4.5

But [important] median \neq midpoint of entire range of
observations (i.e., 3 in this data)

The Mean

- The *mean* (or *mean value*) of a variable in a set of data is the *result of adding up all the observed values of the variable and dividing by the number of cases* (i.e., the “average” as the term is most commonly used).
- It is the most used measure of location.

How to calculate the Mean

$$\text{mean of } X = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}$$

Or mean = $\frac{\text{sum of value}}{\text{\# of observations}}$

The Mean of 20, 30, 40 is therefore:

$$(20+30+40)/3 = 30.$$

Mean or Median

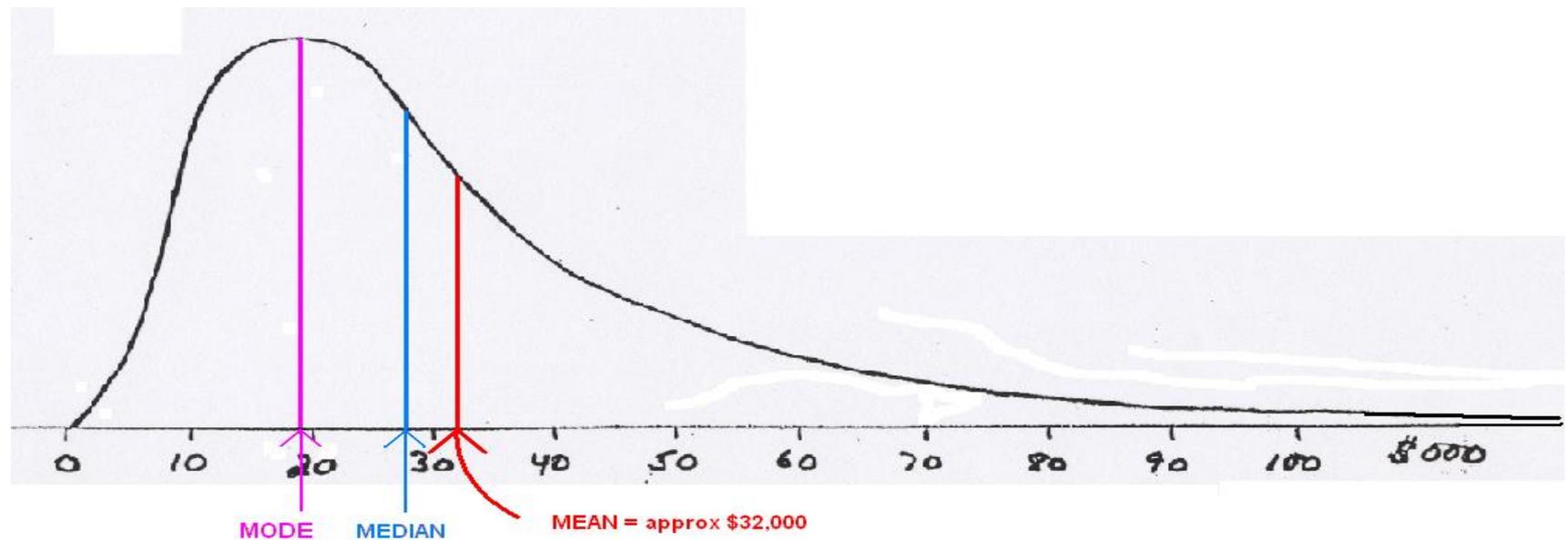
- **Strength of median:** The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for **highly skewed distributions**.
- **Weakness of median:** It only relies on the central values and ignores all the other data.
 - It is also less amenable than the mean to statistical tests.

Mean or Median

- e.g. family income.
- **mean** of 20, 30, 40, and 990 is $(20+30+40+990)/4 = 270$.
- The **median** of these four observations is $(30+40)/2 = 35$.
- Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Median vs. Mean Values (continued)

- If the distribution of the data is *skewed*, the mean is pulled (relative to the median) in the direction of the long thin tail.
- For example, income is distributed in a highly skewed fashion, with a long thin tail in the direction of higher income. Thus mean income is typically considerably higher than median income.



Choosing a measure of central tendency

- Depends on the nature of the distribution.
- For continuous variables in a unimodal and symmetric distribution the mean, median and mode are identical.
- With a skewed distribution the median may be more useful
- For statistical analyses the mean is the preferred measure.

Summary

In this session, we have looked at:

- Calculation and interpretation of measures of central tendency
 - Mean
 - Median
 - Mode
- The appropriate application of measures of Central Tendency

Summary

Quick definitions

- **Mode**
 - the most frequently occurring score
- **Median**
 - the mid-point of a set of ordered scores
- **Mean**
 - the result of dividing the arithmetic sum of scores by the number of scores

Acknowledgement

- Some of the notes adapted from Mr. Y Banda, UNZA School of Nursing Sciences.

Conclusion

Thank you for your attention