Chapter 4

Measures of Dispersion and Skewness

Outline of Chapter 4

Measures of Dispersion and Skewness

1. Measures of Dispersion
   - The Range
   - The Variance and Standard Deviation
   - The Coefficient of Variation
2. Measures of Skewness
   - Symmetry and Skewness
   - Common Measures of Skewness
3. Measures of Kurtosis
4. The Boxplot (Optional)

Reference: Chapter 8-9, 12 of Elementary Statistics by ACS

Part 1

Measures of Dispersion

Measure of Dispersion

A **measure of dispersion** is a descriptive summary measure that helps us characterize the data set in terms of how varied the observations are from each other.

- A **small value** indicates that the **observations are not too different from each other**; that is, there is a **concentration of observations about the center** of the distribution.
- A large value indicates that the **observations are very different from each other** or they are **widely spread out from the center**.
- The smallest possible value of a measure of dispersion should be 0. A zero measure should indicate the absence of variation.

Illustration

$$A = \{98,98,99,99,99,100,100,100,100,100,100,100,101,101,101,102,102\}$$
$$B = \{20,30,40,50,60,70,80,90,100,110,120,130,140,150,160,170,180\}$$

The means of both collections are equal to 100. Their medians are also equal to 100. (Verify)

- Which collection must have a higher measure of dispersion?
- For which collection is the mean a more reliable measure of central tendency? (**Reliable** in the sense that if we repeatedly select an observation at random from the collection, its value is usually not too different from the mean.)

The measure of dispersion serves as a measure of the reliability of the mean or median as measures of central tendency.

General Classifications

Measures of Dispersion can be generally classified into two:

1. **Measures of Absolute Dispersion**

A measure of absolute dispersion has the same unit as the observations. (Examples: range, interquartile range, standard deviation)

2. **Measures of Relative Dispersion**

A measure of relative dispersion has no unit and is therefore useful in comparing the variability of one distribution with another distribution. (Example: coefficient of variation)

Range

The **range** is the distance between the maximum value and the minimum value.

$$Range = Maximum - Minimum$$

Sometimes the range is presented by stating the smallest and the largest values.

Example:
Given the weights of 5 rabbits (in lbs): 8, 10, 12, 14, 15,find the range.

Solution:
The maximum is 15 pounds and the minimum is 8 pounds. Thus, the range of the weights of the rabbits is

$$\boldsymbol{Range} = maximum - minimum = 15 - 8 = 7 \text{ pounds}$$

We can also say that the weights of the rabbits range from 8 to 15 pounds.

Range: Characteristics

Strength:
It is a simple, easy-to-compute and easy-to-understand measure.

Weaknesses:
- It fails to communicate any information about the clustering or the lack of clustering of the values in the middle of the distribution since it uses only the extreme values (minimum and maximum).
- An outlier can greatly affect its value.
- It tends to be smaller for smaller collections than for larger collections.
- It cannot be approximated from frequency distributions with an open-ended class.
- It is not tractable mathematically.

Interquartile Range (IQR)

The **interquartile range (IQR)** is the difference between the third and first quartiles of the data set. That is,

$$IQR = Q_3 - Q_1$$

- The interquartile range reflects the **variability of the middle 50%** of the observations in the array.
- The IQR may be viewed as the **range for a trimmed data set** wherein the smallest 25% and the largest 25% of observations have been removed. This modified range addresses the weakness of the range's sensitivity towards outliers.
- A shortcoming of the IQR is that it could be 0 even if there is still some variation among the smallest 25% and largest 25% of all observations.

IQR: Example

A professor studying the variation in final exam scores (50 items) of Stat 101 students found that the first quartile score is 26 and the third quartile score is 38. Find the interquartile range.

Solution:

$$\boldsymbol{IQR} = Q_3 - Q_1 = 38 - 26 = 12$$

Note that if Q_1 and Q_3 are not given, you have to solve for them first.

Variance

The **population variance** is the mean of the squared deviations between each observed value and the mean.

Population Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

where Xi is the measure taken from the i^{th} element of the population/sample

μ is the population mean

$\bar{X}$ is the sample mean

N is the population size

n is the sample size

Variance

- The **population variance is a parameter** while the **sample variance is a statistic**.
- The squared difference of an observation from the mean gives us an **idea on how close this observation is to the mean**. A large squared difference indicates that the observation and the mean are far from each other while a small squared difference indicates that the observation and the mean are close to each other. A small variance indicates that the observations are highly concentrated about the mean so that it is appropriate to use the mean to represent all of the values in the collection.

Variance

- The **sample variance is NOT the mean of the squared deviations of the observations from the mean**. The denominator of the sample variance is not n (the size of the sample); rather, it is (n-1). The reason for using (n-1) as the divisor is that in Inferential Statistics, the corresponding statistic with n as the divisor tends to underestimate the population variance. Using the divisor (n-1) is used to make up for this tendency to underestimate.
- The **unit of the variance is the square of the units of measures in the data set**. Thus, strictly speaking, the variance is not a measure of absolute dispersion. Often, it is desirable to return to the original units of measure and so **it is the standard deviation that is presented**.

Standard Deviation

The **standard deviation** is the positive square root of the variance.

- The **population standard deviation** for a finite population with N elements, denoted by the Greek letter σ (lower case sigma) is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

- The **sample standard deviation** for a sample with n elements, denoted by the letter s is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Standard Deviation: Example

Given the IQ of 7 students in the sample, compute for the sample standard deviation.

Let Xi = weight of i^{th} student in the sample, n = no. of students = 7.

Xi	Xi-107	$(Xi-107)^2$
100	-7	49
99	-8	64
110	3	9
105	-2	4
112	5	25
107	0	0
116	9	81
$\sum(X_i - 107)^2$		232

$$(\bar{X} = 107)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - 107)^2}{7-1}} = \sqrt{\frac{232}{6}} = 6.2$$

Computational Formula of the Variance

If the mean is a rounded figure then the propagation of rounding errors is very fast when we use the definitional formula to compute the variance. We can avoid this by using the following computational formula for the variance:

$$\sigma^2 = \frac{N \cdot \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2}{N^2}$$

$$s^2 = \frac{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Computational Formula: Example

Using the same data set on the IQ of 7 students in the sample, we compute the sample standard deviation using the computational formula.

X_i	X_i^2
100	10000
99	9801
110	12100
105	11025
112	12544
107	11449
116	13456
$\sum_{i=1}^{7} X_i = 749$	$\sum_{i=1}^{7} X_i^2 = 80375$

$$s = \sqrt{\frac{7\sum_{i=1}^{7} X_i^2 - \left(\sum_{i=1}^{7} X_i\right)^2}{7(7-1)}}$$

$$= \sqrt{\frac{7(80375) - (749)^2}{42}} = 6.2$$

Mathematical Properties of the SD

Property 1

If each observation of a set of data is transformed by the **addition** (or subtraction) of a constant c to each observation, the standard deviation of the new set of data is the same as the standard deviation of the original set of data.

$$s_{transformed} = s_{original}$$

Property 2

If each observation of a set of data is transformed by the **multiplication** (or division) of a constant c to each observation, the standard deviation of the new set of data is equal to the standard deviation of the original set of data multiplied (or divided by) $|c|$.

$$s_{transformed} = |c| s_{original}$$

Mathematical Properties of the SD

Example

The weights (in milligrams) of ants in a sample are as follows:
Sample data = {3, 8, 35, 1, 50}
Its mean is 19.4 mg. and its standard deviation is 21.892 mg.

If each measurement is converted to grams (1000 mg=1 g), what will be the new mean? the standard deviation?

If each ant gained 2 mg. in weight, what will be the new mean? the standard deviation?

Characteristics of the SD

- It **uses every observation** in its computation.
- It may be **distorted by outliers**. This is because squaring large deviations from the mean will give more weight to these outliers.
- It is amenable to algebraic treatment.
- It is **always nonnegative**. A value of 0 implies the absence of variation.
- Level of measurement must at least be **interval** for the standard deviation to be interpretable.

Variation of 2 or more Distributions

- Consider the following sample of weights of ants in milligrams $\{3, 8, 35, 1, 50\}$. Its standard deviation is 21.892 mg.
- This time consider this sample of weights of elephants in grams $\{6000000, 5999999, 5999998, 6000001, 6000002\}$. Its standard deviation is 1.581 grams.

- Can we use the standard deviations of the two collections to compare the variation of the observations of these two collections?
- We cannot use measures of absolute dispersion to compare the variation of the observations of two or more collections when (i) the units are different or (ii) the means are very different from each other.

Measure of Relative Dispersion

- **Measures of relative dispersion** are measures of dispersion that have no unit of measurement and are used to compare the scatter of one distribution with the scatter of another distribution.

The **coefficient of variation (CV)** is a measure of relative dispersion and is defined as:

Population CV	Sample CV
$\frac{\sigma}{\mu} \times 100\%$	$\frac{s}{\bar{X}} \times 100\%$

Note: The coefficient of variation describes the standard deviation as a percentage of the mean. (Example: CV=10% indicates that the standard deviation is 10% of the mean). Consequently, the CV is not interpretable when the mean is negative and is undefined when mean is 0

Example

Suppose we want to buy a stock and we can select from one out of the two. The prices of stock 1 and stock 2 per share are 2100 PhP and 650 PhP respectively. Let us say that for the past months, we compiled data on a sample of prices of stock 1 and stock 2 at the close of trading and we have the following statistics:

	Stock 1	Stock 2
Mean	2095	665
Standard Deviation	450	80

Solution:

We compute for the coefficient of variation to know which stock has more variable price.

$$CV_{stock\ 1} = \frac{450}{2095} \times 100\% = 21.5\%, \qquad CV_{stock\ 2} = \frac{80}{665} \times 100\% = 12.0\%$$

The z-score

The **z-score** or **standard score** helps determine the relative position of an observed value in the collection where the observed value is below or above the mean and it also measures how far the observed value is from the mean in terms of the size of the standard deviation.

Population z-score	Sample z-score
$\frac{X - \mu}{\sigma}$	$\frac{X - \bar{X}}{s}$

Note: The z-score is NOT a measure of dispersion!

The z-score

We can use the standard score two compare two or more observed values from different data sets.

- Different with respect to the mean and/or standard deviation; different units

We can also use the standard score in identifying possible outliers in our dataset.

- Rule of Thumb: If $|z| \geq 3$, then it is a possible outlier

Remarks

- A **positive z-score** measures the number of standard deviations an observation is above the mean, and a **negative z-score** measures the number of standard deviations an observation is below the mean. A z-score of 0 means that the observation is equal to the mean.

- The z-score **has no unit** which makes it possible to compare the z-scores computed using different collections

Example

The mean grade in Statistics 101 is 70% and the standard deviation is 10%, whereas in Math 17, the mean grade is 80% and the standard deviation is 20%. Mark got a grade of 75% in Stat 101 and a grade of 90% in Math 17. In which subject did Mark perform better if we consider the grades of the other students in the two subjects?

Example

Solution:

$$z_{Stat\ 101} = \frac{75 - 70}{10} = 0.5, \qquad z_{Math\ 17} = \frac{90 - 80}{20} = 0.5$$

If we consider the grades of the other students in the two subjects, Mark's score in Stat 101 is just as good as his score in Math 17. Based on the z-scores, Mark's scores in both subjects are 0.5 standard deviations above their respective mean scores.

Exercise:
Answer problems found in Exercise 8.1, 8.2 and 8.3. (Not graded)

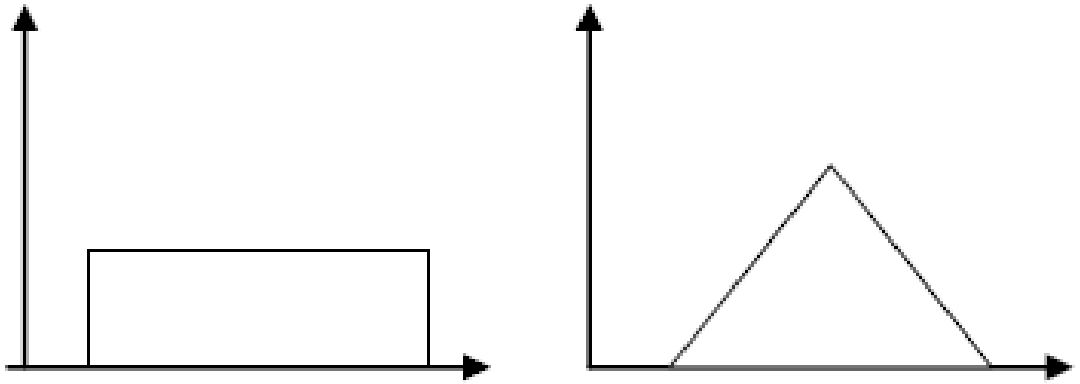
Part 2

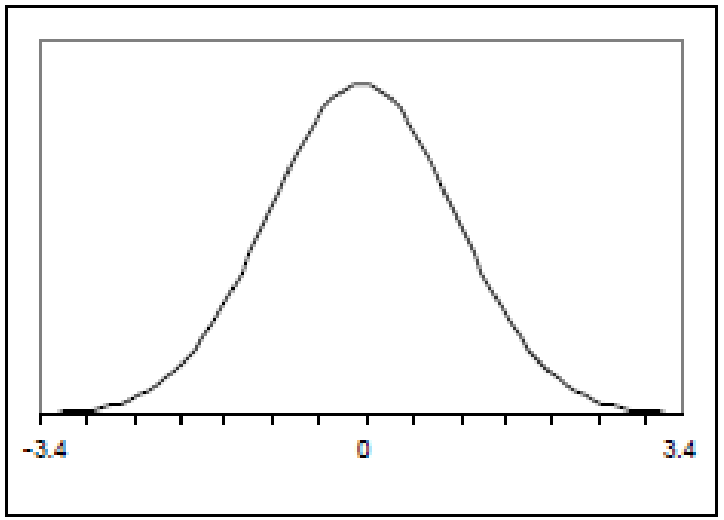
Measures of Skewness

Symmetric and Skewed Distribution

If it is possible to divide the histogram at the center into two identical halves, wherein each half is a mirror image of the other, then it is called a **symmetric distribution**. Otherwise, it is called a **skewed distribution**.

Examples:

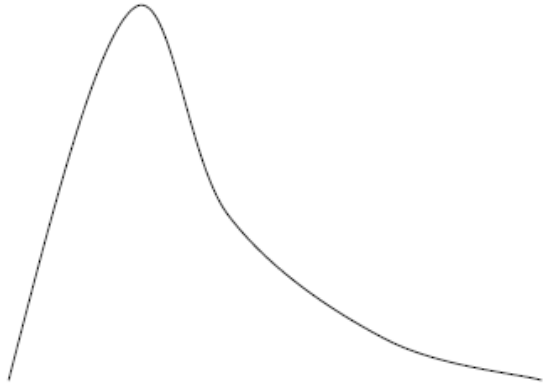




Skewness

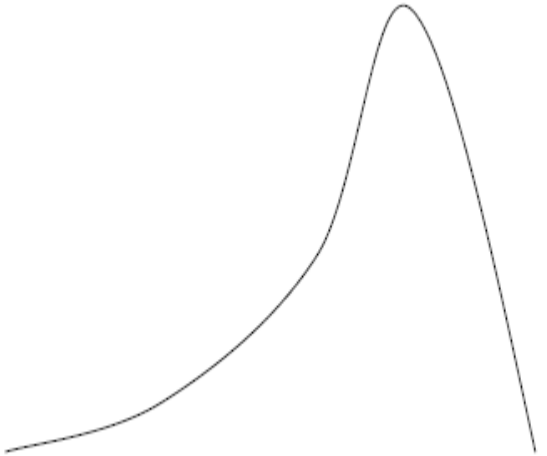
1. Positively Skewed or Skewed to the Right

If the concentration of the values is at the left-end of the distribution and the upper tail of the distribution stretches out more than the lower tail, then the distribution is said to be positively skewed or skewed to the right.

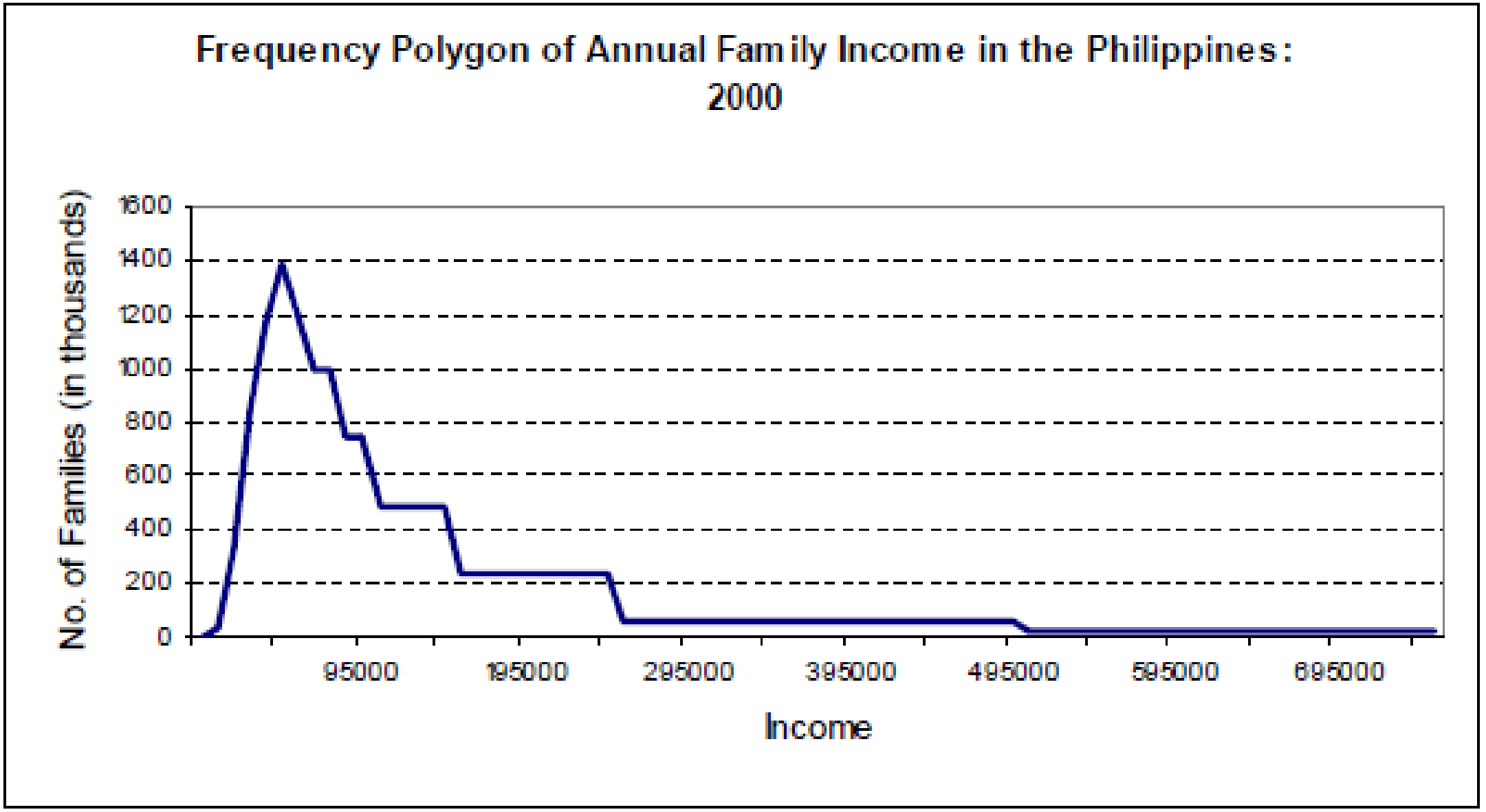


2. Negatively Skewed or Skewed to the Left

If the concentration of the values is at the right-end of the distribution and the lower tail of the distribution stretches out more than the upper tail then the distribution is said to be negatively skewed or skewed to the left.



Example


Frequency Polygon of Annual Family Income in the Philippines: 2000
No. of Families (in thousands)
1600
1400
1200
1000
800
600
400
200
0
95000
195000
295000
395000
495000
595000
695000
Income

Example

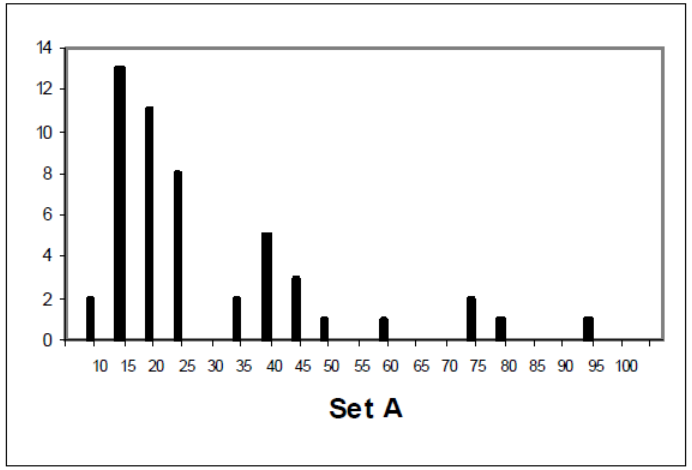
Below are two different sets of test scores. Set A will remind you of the results of a very difficult Physics exam that only a few brilliant students can answer while the rest of the class is clueless on what to answer. On the other hand, Set B will remind you of the results of a relatively easy exam with a few poor-performing students.

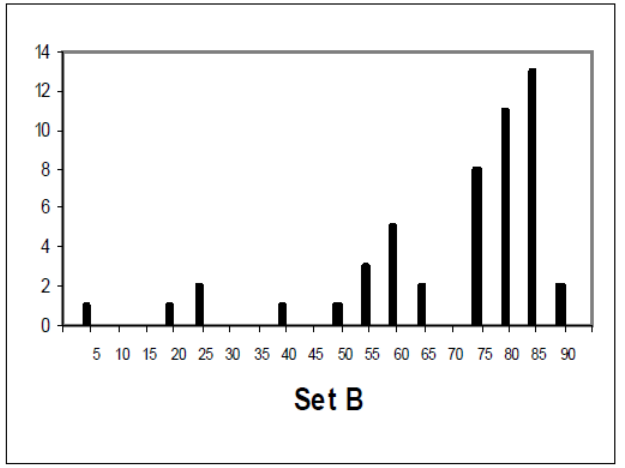
Set A

10 10 15 15 15 15 15 15 15 15 15 15 15 15 15 20 20 20 20 20 20 20 20 20 20
20 25 25 25 25 25 25 25 25 35 35 40 40 40 40 40 45 45 45 50 60 75 75 80 95

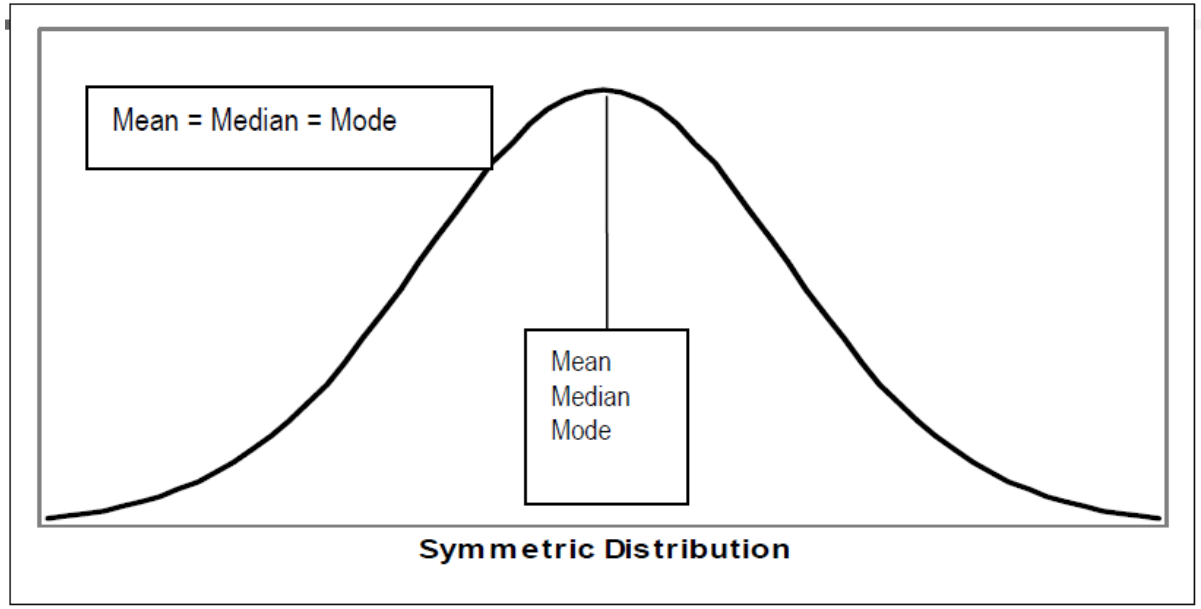
Set B

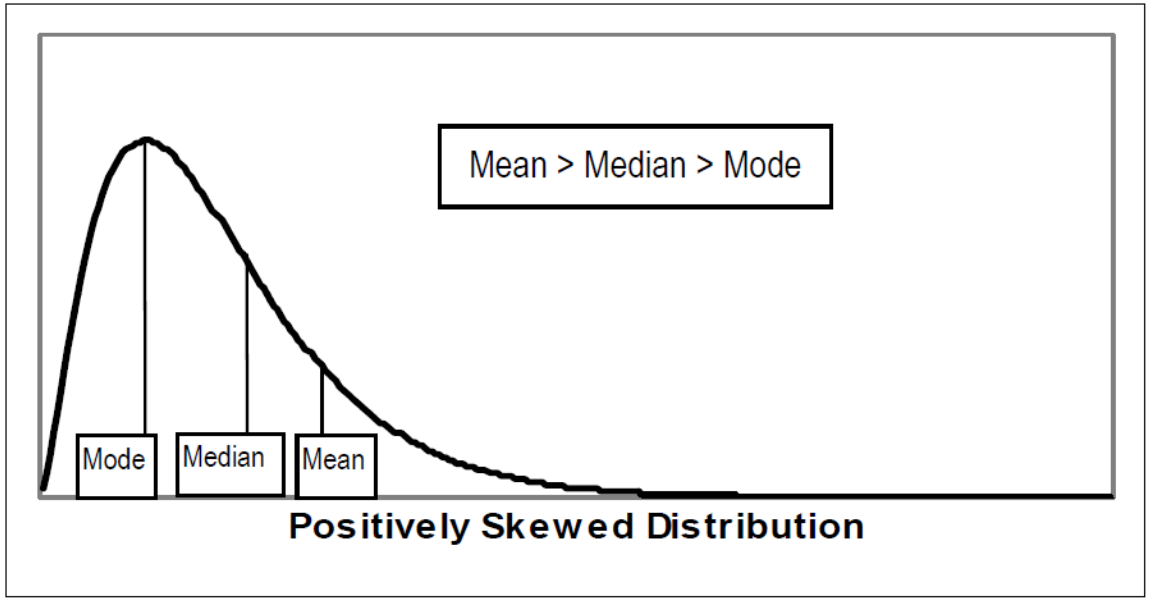
5 20 25 25 40 50 55 55 55 60 60 60 60 60 65 65 75 75 75 75 75 75 75 75 80
80 80 80 80 80 80 80 80 80 85 85 85 85 85 85 85 85 85 85 85 85 90 90

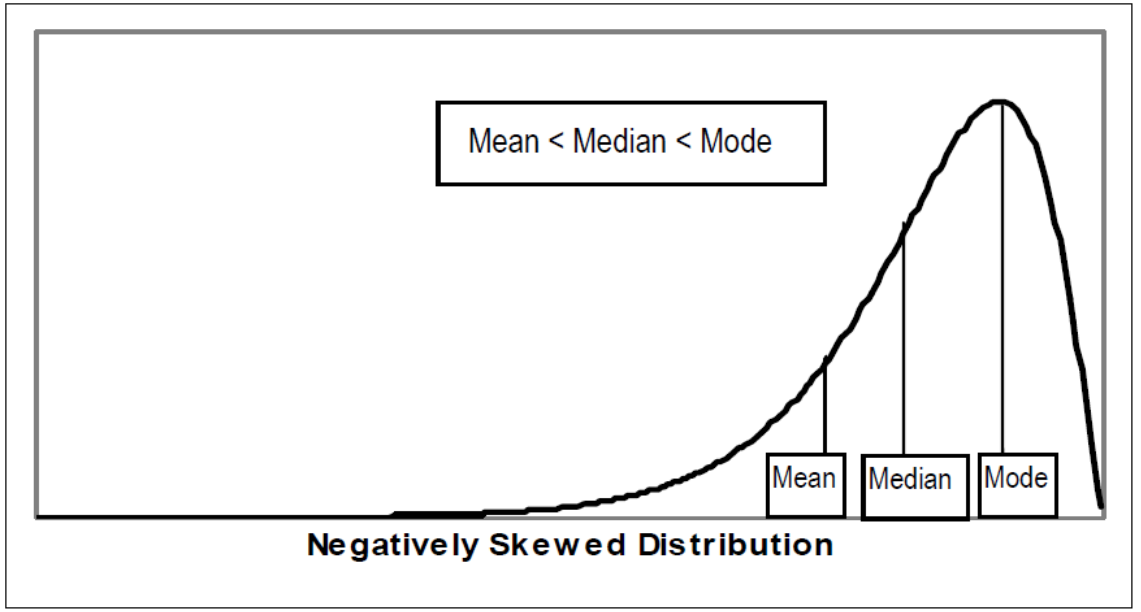




Relationship of Mean, Md and Mode







Importance

- Skewness sometimes **presents a problem in the analysis** of data because it **can adversely affect the behavior of certain summary measures**.
- For this reason, certain procedures in statistics depend on **symmetry assumptions**.
- It would be inappropriate to use these procedures in the presence of severe skewness.
- Sometimes we need to perform special preliminary adjustments, such as **transformations**, before analyzing skewed data.
- Other times, we need to look for procedures that are not affected by skewness.
- **What is important is that, at the onset, we are already able to detect skewness in order to prevent contamination of subsequent analysis**. Or else, we will only end up with spurious conclusions.

Measure of Skewness

A **measure of skewness** is a single value that indicates the degree and
direction of asymmetry.

Interpretation:

Direction of Skewness

Sk = 0: symmetric
Sk > 0: positively skewed
Sk < 0: negatively skewed

Degree of Skewness

The farther |Sk| is from 0, the more skewed the distribution.

Coefficient of Skewness

- Pearson's first coefficient of skewness for a sample

$$Sk_1 = \frac{\bar{X} - Mode}{s}$$

- Pearson's second coefficient of skewness for a sample

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

- Coefficient of skewness based on the third moment

$$Sk_3 = \frac{\sum_{i=1}^{N}(X_i - \mu)^3 / N}{\sigma^3}$$

$$Sk_3 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3 / n}{\left(s\sqrt{\frac{n-1}{n}}\right)^3}, \qquad Sk_3^* = \frac{\sqrt{n(n-1)}}{n-2} Sk_3$$

Remarks

- **Pearson's first coefficient of skewness** is a function of the **mode**. This becomes a problem if the mode does not exist or the collection is too small so that the mode is not a stable measure of central tendency.
- **Pearson's second coefficient of skewness** was based on Karl Pearson's empirical derivation on the distance of the **median** and the **mean** as compared to the distance of the mode and the mean.
- **Sk_3 is sensitive to outliers**
- Coefficient of Skewness based on the Quartiles

$$Sk_4 = \frac{(Q_3 - Md) - (Md - Q_1)}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1}$$

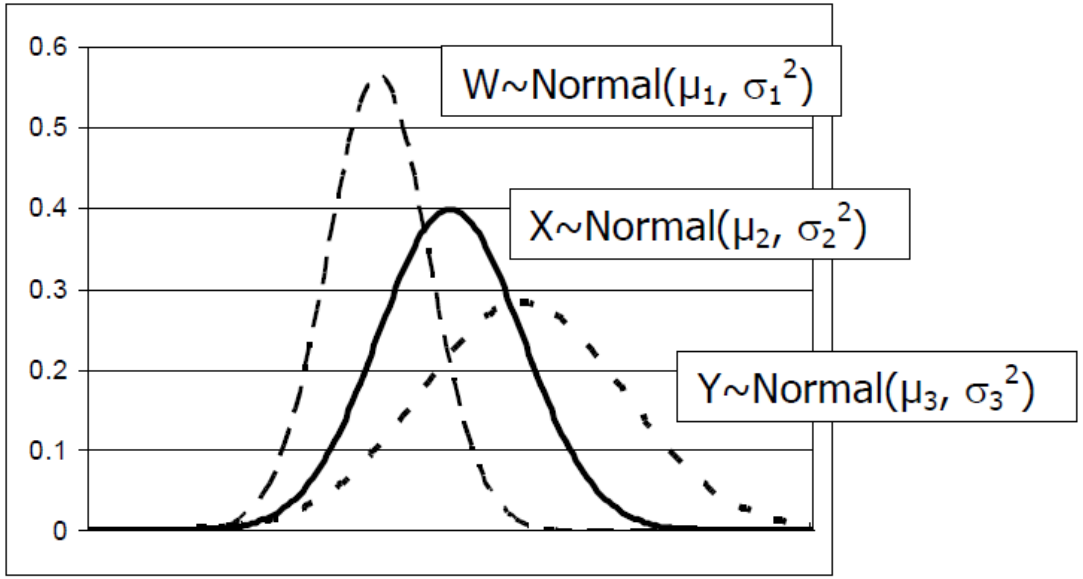
  - Unlike the coefficient of skewness based on the third moment, this measure is not sensitive to the presence of possible influential outlying values. Its value does not disproportionately inflate with the presence of a single unusually large or unusually small value.

Part 3

Measures of Kurtosis

Normal Distribution

- The normal distribution is one of the most important distributions in Statistics. It is a bell-shaped curve that is symmetric about its mean, μ. Its tails approach the x-axis on both sides but will never touch them. The area below any normal curve is equal to 1.
- The location of the center of the normal curve and its spread is determined by the mean and variance of the normal distribution.



Graph of 3 normally distributed random variables where $\mu_1 < \mu_2 < \mu_3$ while $\sigma_1^2 < \sigma_2^2 < \sigma_3^2$

Types of Kurtosis

Karl Pearson introduced the following terms to classify a unimodal distribution according to the shape of its hump as compared to a normal distribution with the same variance:

1. Mesokurtic

- hump is the same as the normal curve
- It is neither too flat nor too peaked

2. Leptokurtic

- curve is more peaked about the mean and the hump is narrower than the normal curve
- prefix “lepto" came from the Greek word leptos meaning small or thin.

3. Platykurtic

- curve is less peaked about the mean and the hump is flatter than the normal curve
- prefix “platy" came from the Greek word platus meaning wide or flat.

Remarks

1. **Leptokurtic** curve has **thicker** tails than normal. **Platykurtic** curve has **thinner** tails than normal.

2. In a **leptokurtic** curve, the sharper peak implies a **higher concentration of values around the mode** compared to a normal distribution of the same variance. Thus, in order to achieve equal variability, the leptokurtic curve must have thicker tails, or **more observations on the tails**, to compensate for the sharper peak. We can then say that the leptokurtic distribution's variance is attributed to a few observations that highly deviate from the mode.

3. In a **platykurtic** curve, the flatter peak implies **lower concentration of values around the mode** compared to a normal distribution of the same variance. Thus, in order to achieve equal variability, the platykurtic curve must have thinner tails. The platykurtic distribution's variance is attributed to many observations that moderately deviate from the mode.

Importance

- It can be used to explain the type of variability of a distribution (few observations that highly deviate from the mode as opposed to many observations that moderately deviate from the mode).
- It is used to detect nonnormality since many classical statistical procedures assume normality.

Coefficient of Kurtosis

Population Coefficient of Kurtosis Based on the Fourth Moment

$$K = \frac{\sum_{i=1}^{N}(X_i - \mu)^4 / N}{\sigma^4}$$

Interpretation:

$K - 3 < 0 \rightarrow$ platykurtic, $K - 3 > 0 \rightarrow$ leptokurtic, and $K - 3 = 0 \rightarrow$ mesokurtic

The value $\boldsymbol{K - 3}$ is called the "**excess of kurtosis**"

$$kurt_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\left(s\sqrt{\frac{n-1}{n}}\right)^4} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\left(s^2 \cdot \frac{n-1}{n}\right)^2}$$ **sample counterpart of K**

$$kurt_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left(kurt_1 - \frac{3(n-1)}{n+1}\right)$$ **unbiased estimator of K-3**