Chapter 10

Measures of Skewness And Kurtosis

Definition of Measures of Skewness (page 184)

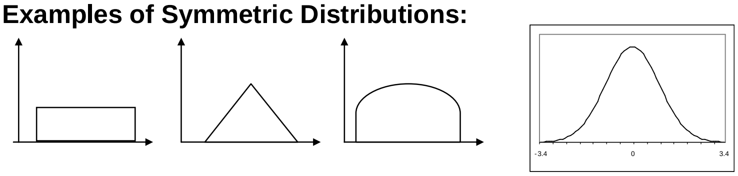
- A **measure of skewness** describes the degree and direction of asymmetry or departure from symmetry of a distribution.

- A measure of skewness equal to 0 indicates that the distribution is symmetric. The farther the measure is from 0, the greater the degree of asymmetry. The sign (positive or negative) of the measure indicates the direction of skewness.

Definition of symmetric distribution (page 181)

A (relative) frequency distribution is said to be **symmetric** if its graph can be folded in the middle along the vertical axis so that the two sides of the graph coincide.

Examples of Symmetric Distributions:



rectangular-shaped triangular-shaped polygon-shaped bell-shaped

Definition of Asymmetry (page 182)

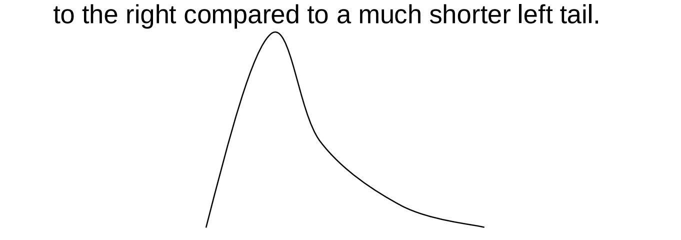
If the graph of the distribution is such that its two sides do not coincide, then the distribution is said to be ***asymmetric***. A distribution that is asymmetric is said to be **skewed**.

Two Types of Skewness:

1. **Positively Skewed or Skewed to the Right**
2. **Negatively Skewed or Skewed to the Left**

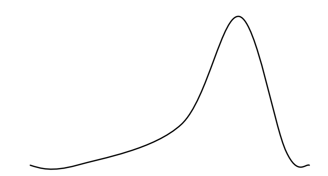
Definition of Skewed to the Right Distribution (page 182)

A ***skewed to the right distribution*** tapers more to the right than to the left. Observations on the left side of the distribution are closer to each other than the observations are on the right side of the distribution. It has a longer tail to the right compared to a much shorter left tail.



Definition of Skewed to the Left Distribution (page 183)

A ***skewed to the left distribution*** tapers more to the left than to the right. Observations on the right side of the distribution are closer to each other than the observations are on the left side of the distribution. It has longer tail to the left compared to a much shorter right tail.



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

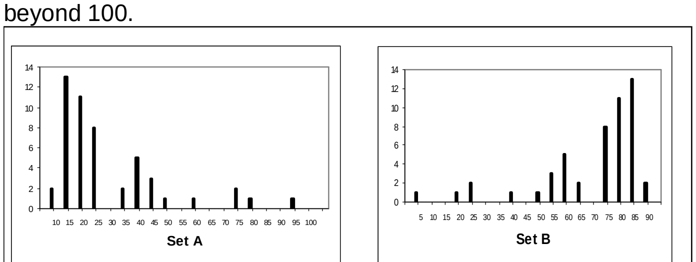
10	10	15	15	15	15	15	15	15	15	15	15	15	15
15	20	20	20	20	20	20	20	20	20	20 X	20	25	25
25	25	25	25	25	25	35	35	40	40	40	40	40	45
45	45	50	60	75	75	80	95						

Set B

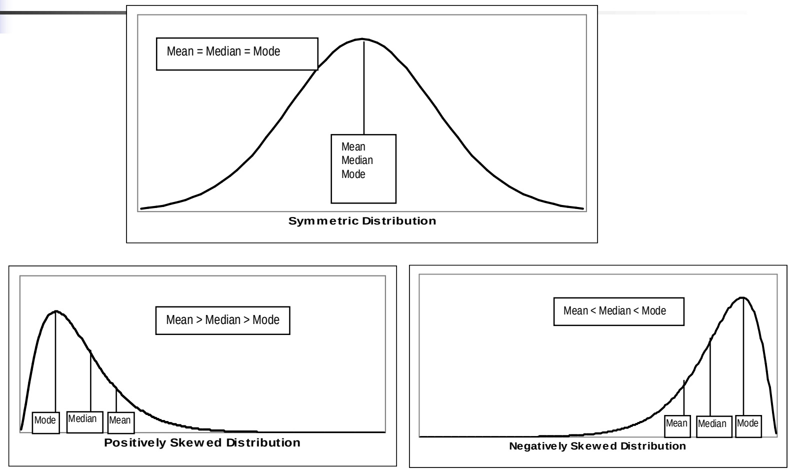
5	20	25	25	40	50	55	55	55	60	60	60	60	60
65	65	75	75	75	75	75	75	75	75	80 X	80	80	80
80	80	80	80	80	80	80	85	85	85	85	85	85	85
85	85	85	85	85	85	90	90						

Example (cont'd)

The distribution of test scores in Set A is positively skewed while that of Set B is negatively skewed. The few relatively high scores in Set A stretched the tail to the right while the few relatively low scores in Set B stretched the tail to the left. It is impossible for the tail of the distribution of Set A to be as long at the left side because the scores cannot be negative. Correspondingly, it is impossible for the tail of the distribution of Set B to be as long at the right side because the scores cannot go beyond 100.



Relationship of the Three Measures of Central Tendency for Unimodal Distributions



Importance of Detecting Skewness

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.

Pearson’s First and Second Coefficient of Skewness (page 184)

Pearson’s First Coefficient of Skewness

$$SK = \frac{\overline{X} - Mo}{s}$$

Pearson’s Second Coefficient of Skewness

$$SK = \frac{3(\overline{X} - Md)}{s}$$

where $\overline{X} = \text{mean}$, Mo=mode, Md = median,
s = standard deviation

Example (page 184)

Two public wet markets in Quezon City, Nepa Q Mart and Kamuning Market, gave the following estimated sales figures for fish sold (in kilos) for one week. Compute for Pearson’s second coefficient of skewness.

Kilos of Fish	
Nepa Q Mart	Kamuning Market
2600	2000
2600	2000
2800	2000
3200	2200
3200	2600
3200	3200
3400	8400

Solution (page 185)

	Nepa Q Mart	Kamuning Market
Mean, μ	3000	3200
Median	3200	2200
Mode	3200	2000
Standard Deviation, σ	302.37	2162.01

Nepa Q Mart: $Sk_2 = \frac{3(3000 - 3200)}{302.37} = -1.98$ (skewed to the left)

Kamuning Market: $Sk_2 = \frac{3(3200 - 2200)}{2162.01} = 1.39$ (skewed to the right)

Based on Pearson's second coefficient of skewness, the degree of skewness is stronger for the distribution of sales (in kilos) in Nepa Q Mart than in Kamuning.

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 that in moderately skewed distributions of a continuous variable, the median tends to fall about 2/3 of the distance from the mode toward the mean.

Preliminary Discussion: Definition of r^{th} Central Moment About the Mean (page 187)

The **r^{th} central moment about the mean** of a finite population $\{X_1, X_2, \ldots, X_N\}$, denoted by μ_r, is defined by:

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

The **r^{th} central moment about the mean** of a sample, denoted by m_r, is defined by:

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

Remarks: (page 187)

- First central moment about the mean is always 0. That is, $\mu_1=0$ and $m_1=0$.
- The second central moment about the mean of a finite population is the population variance. That is, $\mu_2=\sigma^2$. The second central moment about the mean of a sample is $m_2=(n-1)s^2/n$
- The third central moment about the mean will be used as a measure of skewness. Rationale: Because of the cubing operation, large deviations, $(X_i-\mu)$, tend to dominate the sum in the numerator of μ_3. If the large deviations are predominantly positive, μ_3 will be positive because $(X_i- \mu)^3$ has the same sign as $(X_i - \mu)$. Likewise, if the large deviations are predominantly negative, μ_3 will be negative. Since large deviations are associated with the long tail of a distribution, μ_3 will be positive or negative according to whether the direction of skewness is positive or negative. If the distribution is symmetric, the third central moment will be zero. Even if there are large deviations, $(X_i-\mu)$, we are assured that these large deviations will occur on both tail ends because of symmetry. Thus, the positive $(X_i-\mu)^3$ will simply cancel out with the negative $(X_i-\mu)^3$.

Definition of Coefficient of Skewness Based on the Third Moment

The ***population coefficient of skewness based on the third moment*** is:

$$Sk_3 = \frac{\mu_3}{\sigma^3} = \frac{\sum_{i=1}^{N}(X_i - \mu)^3 / N}{\sigma^3}$$

where σ is the population standard deviation.

The **sample coefficient of skewness based on the third moment** is:

$$Sk_3 = \frac{m_3}{\left(\sqrt{m_2}\right)^3} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3 / n}{\left(s\sqrt{(n-1)/n}\right)^3}$$

The ***adjusted sample coefficient of skewness based on the third moment*** is:

$$Sk_3^* = \frac{\sqrt{n(n-1)}}{n-2} Sk_3$$

Example (page 186)

X_i	$(X_i - 66)^3$
64	-8
59	-343
67	1
69	27
65	-1
70	64
68	8
$\sum_{i=1}^{7}(X_i - \bar{X})^3 =$	-252

Verify $\bar{X} = 66$ and s = 3.741657. Thus,

$$Sk_3 = \frac{m_3}{\left(\sqrt{m_2}\right)^3} = \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(s\sqrt{(n-1)}\right)^3} = \frac{\sqrt{7}(-252)}{((3.741657)(\sqrt{6}))^3} = -0.86603$$

$$Sk_3^* = \frac{\sqrt{n(n-1)}}{n-2} Sk_3 = \frac{\sqrt{(7)(6)}}{5}(-0.86603) = -1.12$$

Computational Formula of the Third Central Moment About the Mean

for the population: $$\mu_3 = \frac{\sum_{i=1}^{N} X_i^3}{N} - 3\mu \frac{\sum_{i=1}^{N} X_i^2}{N} + 2\mu^3$$

for the sample: $$m_2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n^2}$$

$$m_3 = \frac{\sum_{i=1}^{n} X_i^3}{n} - 3\bar{X} \frac{\sum_{i=1}^{n} X_i^2}{n} + 2\bar{X}^3$$

Proof: (Exercise)

Example: Computing m_3 using computational formula

	X_i	X_i^2	X_i^3	$(X_i - \bar{X})^3$
	64	4096	262144	-8
	59	3481	205379	-343
	67	4489	300763	1
	69	4761	328509	27
	65	4225	274625	-1
	70	4900	343000	64
	68	4624	314432	8
Total	462	30576	2028852	-252

Definitional formula: $$\frac{\sum_{i=1}^{7}(X_i - \bar{X})^3}{7} = \frac{-252}{7} = -36$$

$$m_3 = \frac{\sum_{i=1}^{n} X_i^3}{n} - 3\bar{X} \frac{\sum_{i=1}^{n} X_i^2}{n} + 2\bar{X}^3 = \frac{2028852}{7} - (3)(66)\frac{30576}{7} + (2)(66)^3 = -36$$

Definition of Coefficient of Skewness based on the Quartiles

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

Remarks:

- 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.
- If the distribution is symmetric then the distance between Q_1 and the median must be the same as the distance between the median and Q_3. On the other hand, most of the values will cluster at the left-end of the distribution for positively skewed distributions. As a result, the median will be closer to Q_1 than to Q_3. Correspondingly, most of the values will cluster at the right-end of the distribution for negatively skewed distributions so that the median, this time, will be closer to Q_3 than to Q_1.

Example

Kilos of Fish	
Nepa Q Mart	Kamuning Market
2600	2000
2600	2000
2800	2000
3200	2200
3200	2600
3200	3200
3400	8400

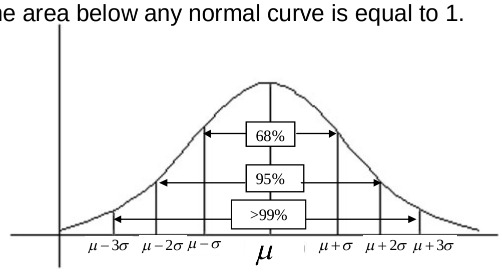
$Q_1 = 2600$ $\quad$ $Q_1 = 2000$
$Q_2 = 3200$ $\quad$ $Q_2 = 2200$
$Q_3 = 3200$ $\quad$ $Q_3 = 3200$

Nepa Q Mart: $$Sk_4 = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1} = \frac{2600 + 3200 - 2(3200)}{3200 - 2600} = -1$$

Kamuning: $$Sk_4 = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1} = \frac{2000 + 3200 - 2(2200)}{3200 - 2000} = 0.667$$

Preliminary Discussion: 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.

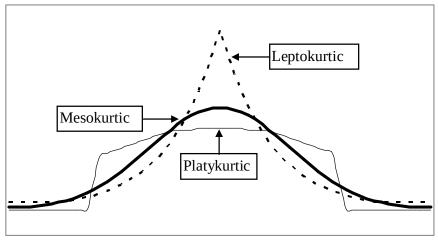


Types of Kurtosis: (page 188)

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: (page 189)



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 of Describing Kurtosis

- It can be used to explain the type of variability of a distribution (few observations that highly deviate from the mode or many observations that moderately deviate from the mode).
- It is used to detect nonnormality.

Population Coefficient of Kurtosis Based on the Fourth Moment

$$K = \frac{\mu_4}{\sigma^4} = \frac{\sum_{i=1}^{N}(X_i - \mu)^4 \big/ N}{\sigma^4}$$

Interpretation:

In general, $\mu_4/\sigma^4 - 3 < 0 \rightarrow$ platykurtic

$\mu_4/\sigma^4 - 3 > 0 \rightarrow$ leptokurtic

$\mu_4/\sigma^4 - 3 = 0 \rightarrow$ mesokurtic

Note: $\mu_4/\sigma^4 - 3$ is called "excess of kurtosis".

Sample Coefficient of Kurtosis Based on the Fourth Moment

$$k_1 = \frac{m_4}{m_2^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\left(s^2(n-1)/n\right)^2}$$

adjusted sample coefficient of kurtosis based on the fourth moment:

$$k_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left(k_1 - \frac{3(n-1)}{n+1}\right)$$

Computational formula of fourth central moment:

Population: $$\mu_4 = \frac{\sum_{i=1}^{N} X_i^4}{N} - 4\mu\frac{\sum_{i=1}^{N} X_i^3}{N} + 6\mu^2\frac{\sum_{i=1}^{N} X_i^2}{N} - 3\mu^4$$

Sample: $$m_4 = \frac{\sum_{i=1}^{n} X_i^4}{n} - 4\bar{X}\frac{\sum_{i=1}^{n} X_i^3}{n} + 6\bar{X}^2\frac{\sum_{i=1}^{n} X_i^2}{n} - 3\bar{X}^4$$

Assignment

Compute the following summary measures using the data on length of life of a sample of flies in page 131, Exercise no. 5. Interpret the results.

1. Adjusted coefficient of skewness based on the third moment
2. Adjusted coefficient of kurtosis based on the fourth moment

Note: Use the computational formula to compute for the third and fourth central moments. Show your solution.