Measures of Dispersion

Measures of Dispersion

- It is the measure of extent to which an individual items vary
- Synonym for variability
- Often called "spread" or "scatter"
- Indicator of consistency among a data set
- Indicates how close data are clustered about a measure of central tendency

Compare the following distributions

- Distribution *A*

  200 200 200 200 200

- Distribution B

  200 205 202 203 190

- Distribution *C*

  1 989 2 3 5

Arithmetic mean is same for all the series but distribution differ widely from one another.

Objectives of Measuring Variation

- To gauge the reliability of an average i.e. dispersion is small, means more reliable.
- To serve as a basis for the control of variability.
- To compare two or more series with regard to their variability.

The Range

- Difference between largest value and smallest value in a set of data
- Indicates how spread out the data are
- Dependent on two extreme values.
- Simple, easy and less time consuming.

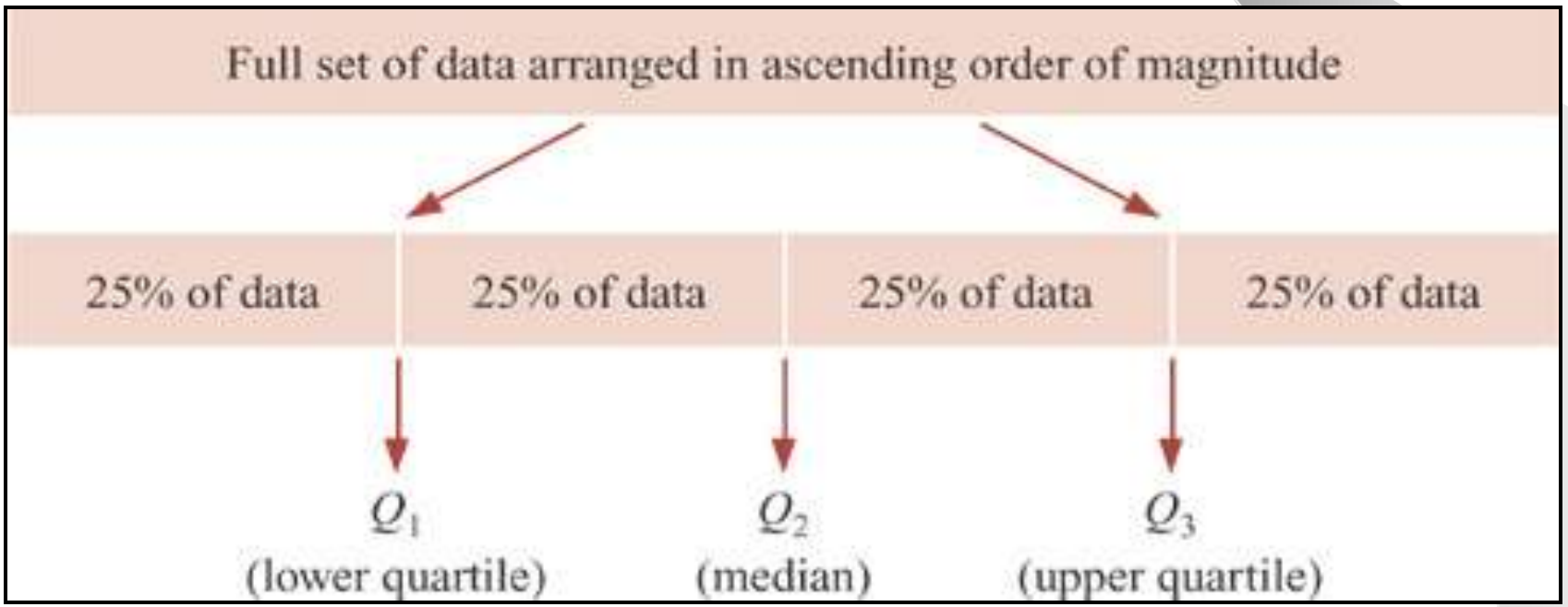
- Calculation:
  - Find largest and smallest number in data set
  - Range = Largest - Smallest

Uses of Range

- Quality Control
- Weather Forecasting
- Fluctuations in Share/Gold prices

Quartile Deviation/Inter-Quartile Range

The **quartiles** divide the data into four parts. There are a total of three quartiles which are usually denoted by Q_1, Q_2 and Q_3.



The **inter-quartile range** is defined as the difference between the upper quartile and the lower quartile of a set of data.

Inter-quartile range = $Q_3 - Q_1$

Quartile Deviation/Semi Inter Quartile Range = $Q_3-Q_1 / 2$

Example

- Calculate Quartile Deviation-

Wages in Rs. per week	No. of wages earners
Less than 35	14
35-37	62
38-40	99
41-43	18
Over 43	7

Deviations from the mean

- Useful for interval or ratio level data
- An examination of deviation from the mean can reveal information about the variability of the data
  - Deviations are used mostly as a tool to compute other measures of variability
- However, the sum of deviations from the arithmetic mean is always zero:

$$\text{Sum}\ (X - \mu) = 0$$

- There are two ways to solve this conundrum…

Mean Absolute Deviation (MAD)

- One solution is to take the absolute value of each deviation around the mean. This is called the Mean Absolute Deviation

X	X-m	|X-m|
5	-8	8
9	-4	4
16	3	3
17	4	4
18	5	5

$$MAD = \frac{\sum |X - \mu|}{n} = \frac{24}{5} = 8.4$$

Note that while the MAD is intuitively simple, it is rarely used in practice

Standard Deviation

- It is most important and widely used measure of studying variation.
- It is a measure of how much 'spread' or 'variability' is present in sample.
- If numbers are less dispersed or very close to each other then standard deviation tends to zero and if the numbers are well dispersed then standard deviation tends to be very large.

For Ungrouped Data

For a set of ungrouped data $x_1, x_2, \ldots, x_n$,

$$\text{Standard deviation } \sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} f_i (x_1 - \bar{x})^2}{n}}$$

where $\bar{x}$ is the mean and n is the total number of data.

- It is the average of the distances of the observed values from the mean value for a set of data

$$SD = \sqrt{\frac{\text{Sum of squares of individual deviations from arithmetic mean}}{\text{Number of items}}}$$

Example:

Scores	Deviations From Mean	Squares of Deviations
01	-13	169
03	-11	121
05	-09	81
06	-08	64
11	-03	9
12	-02	4
15	+01	1
19	+05	25
34	+20	400
37	+23	529
143		1403

No. of scores = 10

M = 143/10 = 14

$$SD = \sqrt{\frac{1403}{10}} = 11.8$$

For Grouped Data

$$\text{Standard deviation } \sigma = \sqrt{\frac{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + \cdots + f_n(x_n-\bar{x})^2}{f_1 + f_2 + \cdots + f_n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} f_i(x_1-\bar{x})^2}{\sum_{i=1}^{n} f_i}}$$

where f_i is the frequency of the ith group of data, $\bar{x}$ is the mean and n is the total number of data.

Example

- Calculate standard deviation-

Size of item	Frequency	Size of item	Frequency
3.5	3	7.5	85
4.5	7	8.5	32
5.5	22	9.5	8
6.5	60		

Example

- Calculate standard deviation –

Marks	No. of students
0-10	5
10-20	12
20-30	30
30-40	45
40-50	50
50-60	37
60-70	21

Variance

- σ^2 = Variance= (standard deviation)^2
- For comparing the variability of two or more distributions,
- Coefficient of variation = $\frac{\text{s.d.}}{\text{mean}}$ X 100
-

$$C.V. = \frac{\sigma}{\bar{x}} X 100$$

More C.V., More variability

Example

	Organization A	Organization B
Number of employees	100	200
Average wage per employee	5000	8000
Variance of wages per employee	6000	10000

- Which organization is more uniform wages?

Measure of Shape

- Tools that can be used to describe the shape of a distribution of data.
- Two measures of shape-
- Skewness
- Kurtosis

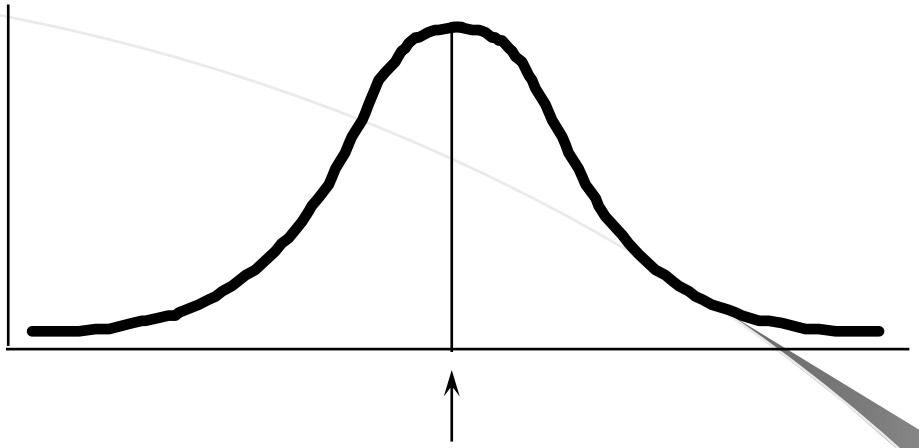
SKEWNESS

- A distribution of data in which the right half is a mirror image of the left half is symmetrical(no skewness)

- Distribution lacks symmetry i.e. asymmetrical i.e skewness

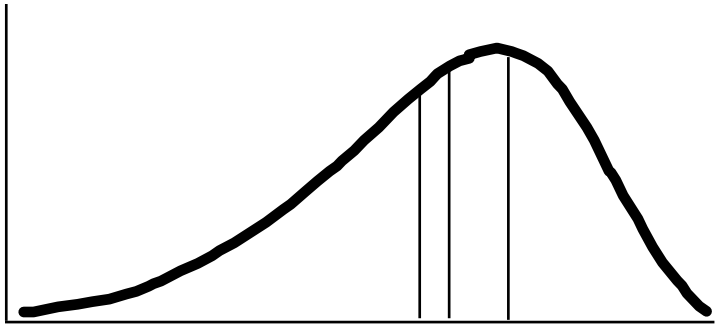
- The skewed portion is long, thin part of the curve.
- Skewed left or negatively skewed
- Skewed right or positively skewed.

- Data sparse at one end and piled up at the other .
- - patients suffering from diabetes



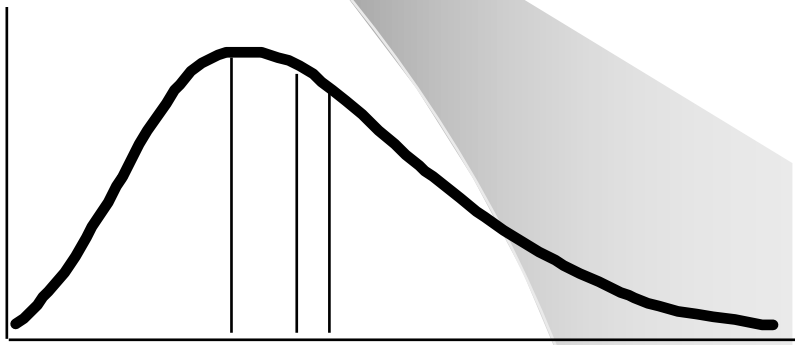
MODE = MEAN = MEDIAN

SYMMETRIC



MEAN MEDIAN MODE

SKEWED LEFT (NEGATIVELY)

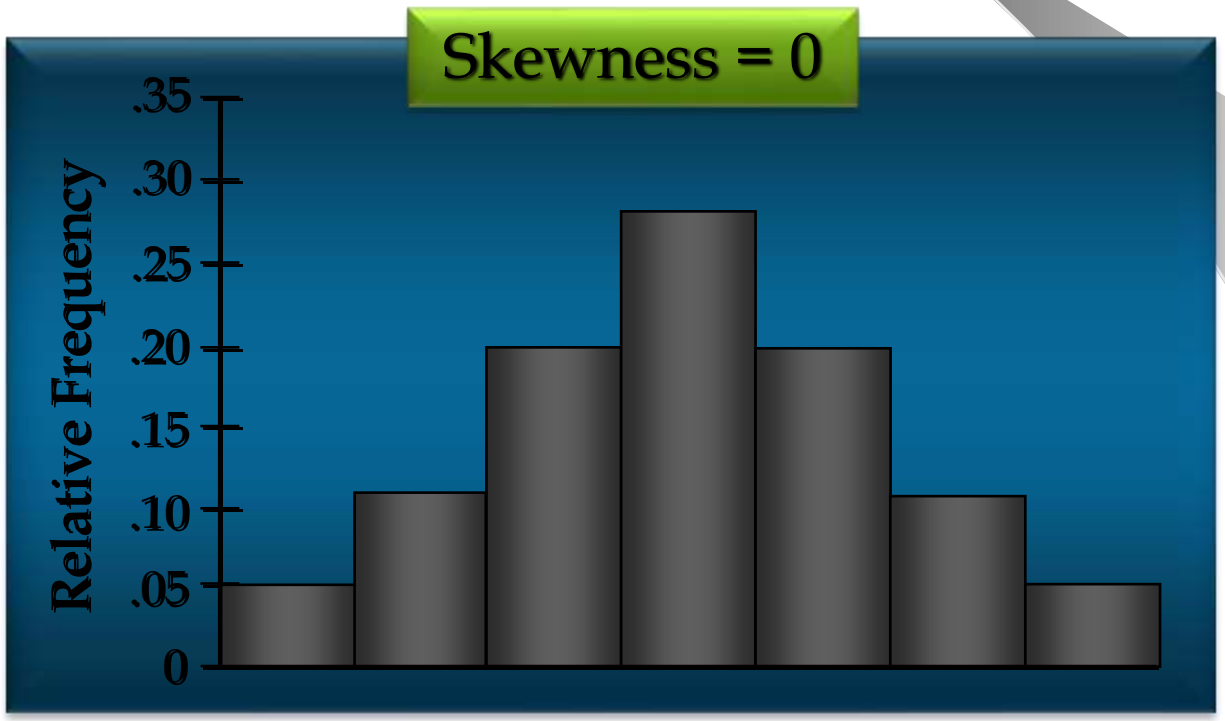


MODE MEDIAN MEAN

SKEWED RIGHT (POSITIVELY)

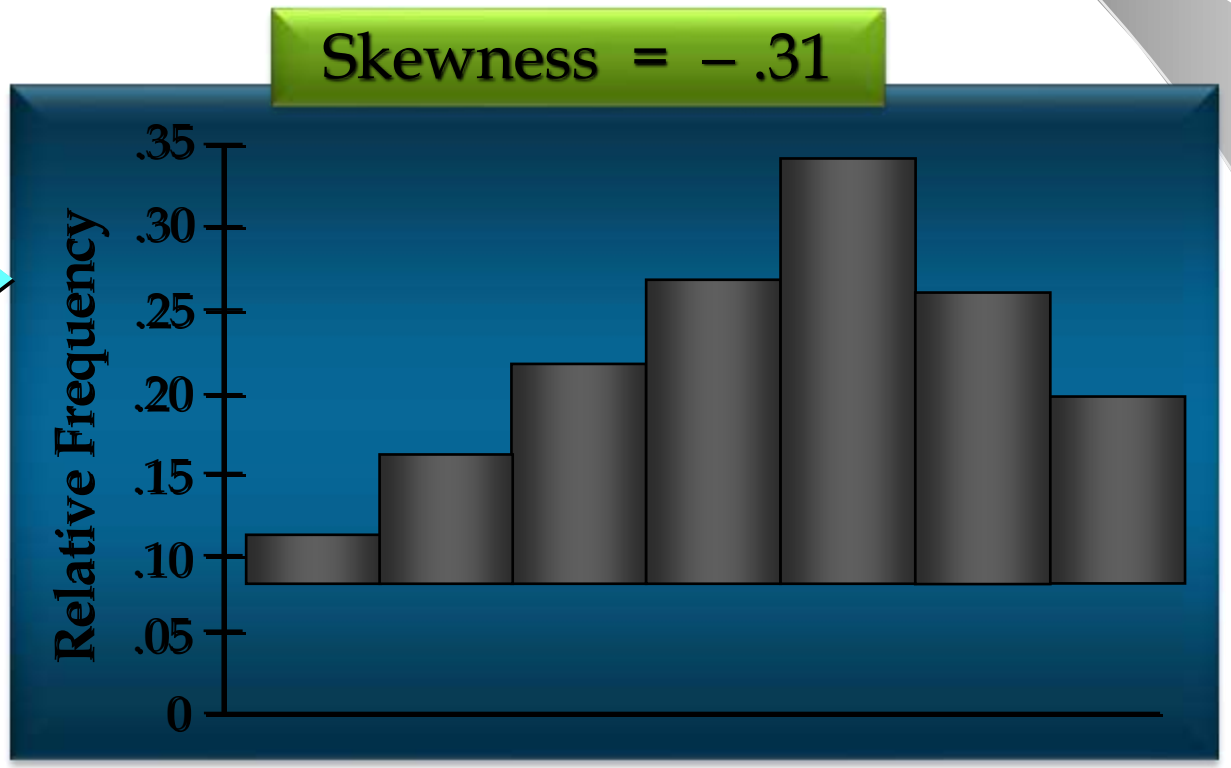
Distribution Shape: Skewness

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.



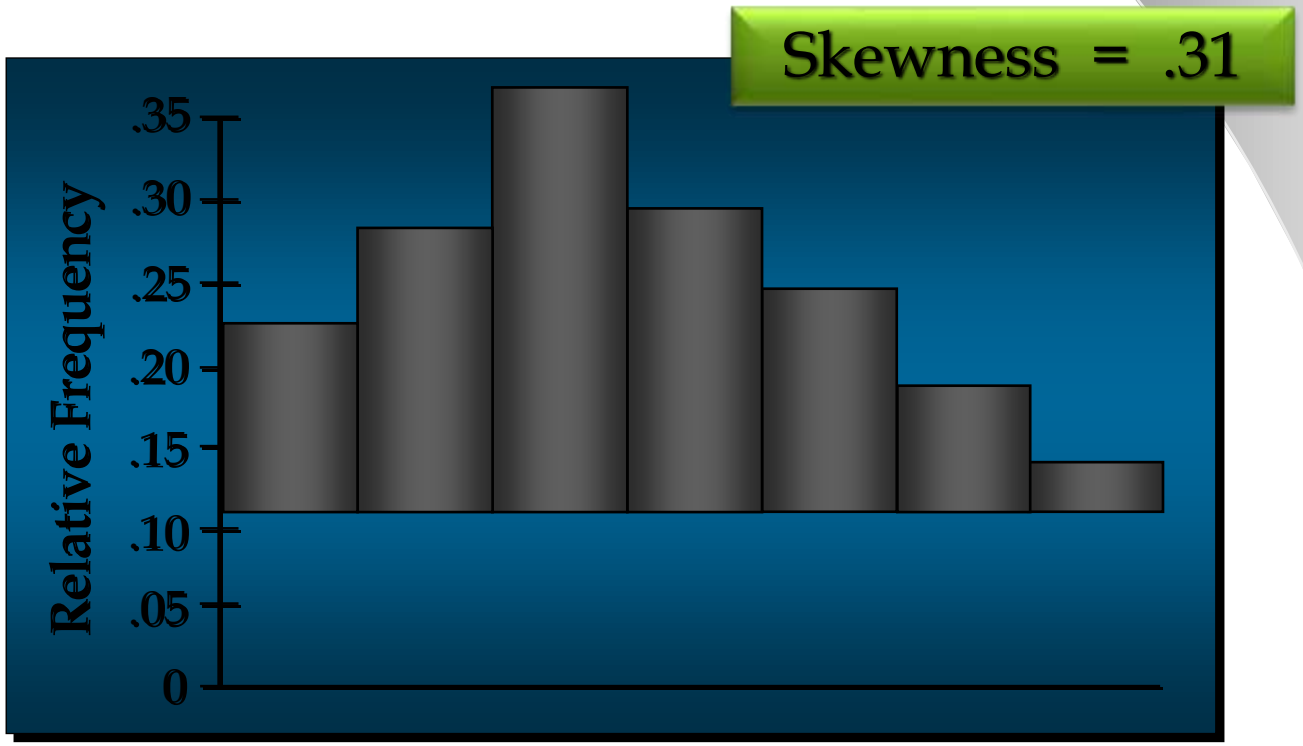
Distribution Shape: Skewness

- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.



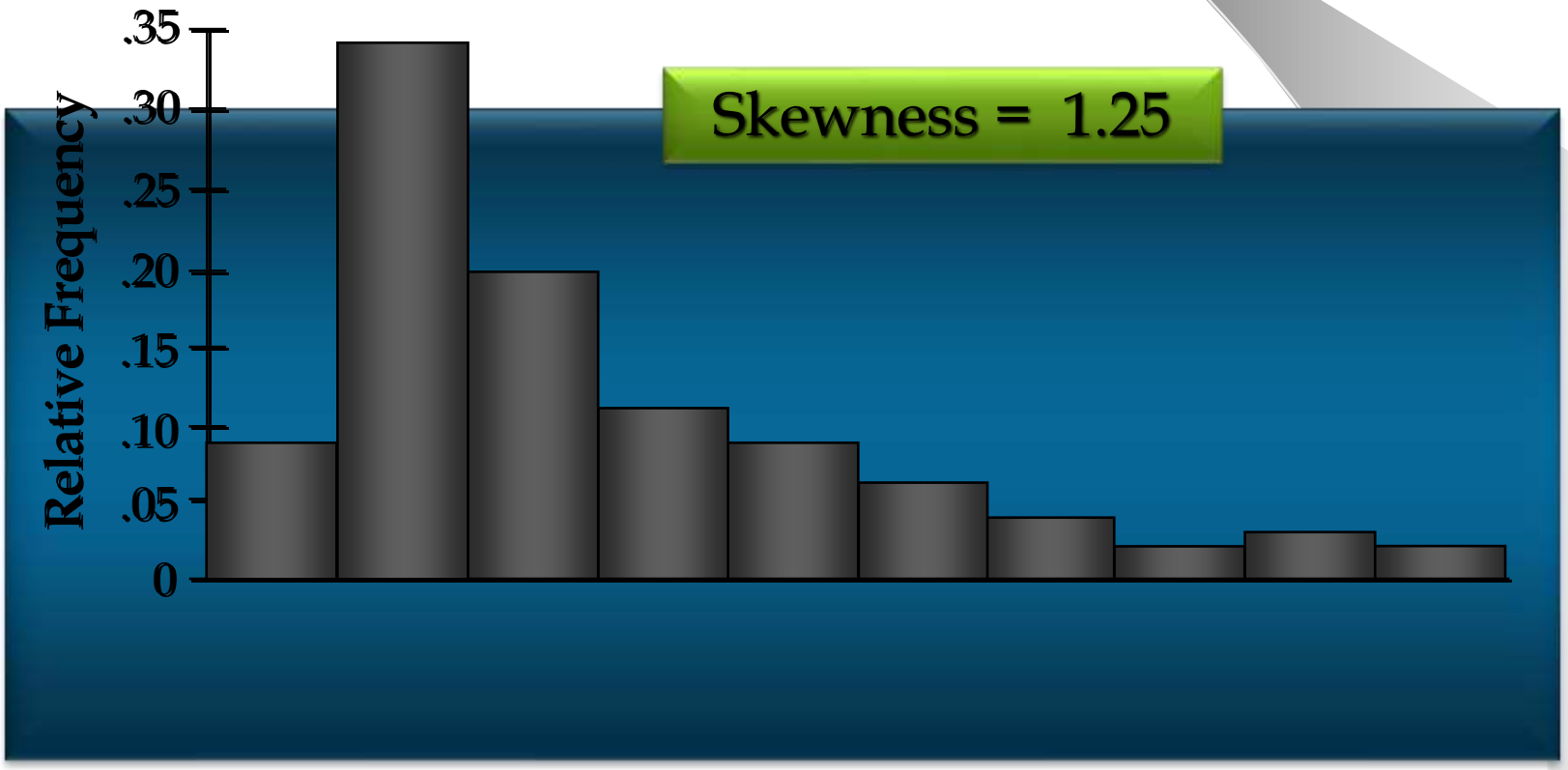
Distribution Shape: Skewness

- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



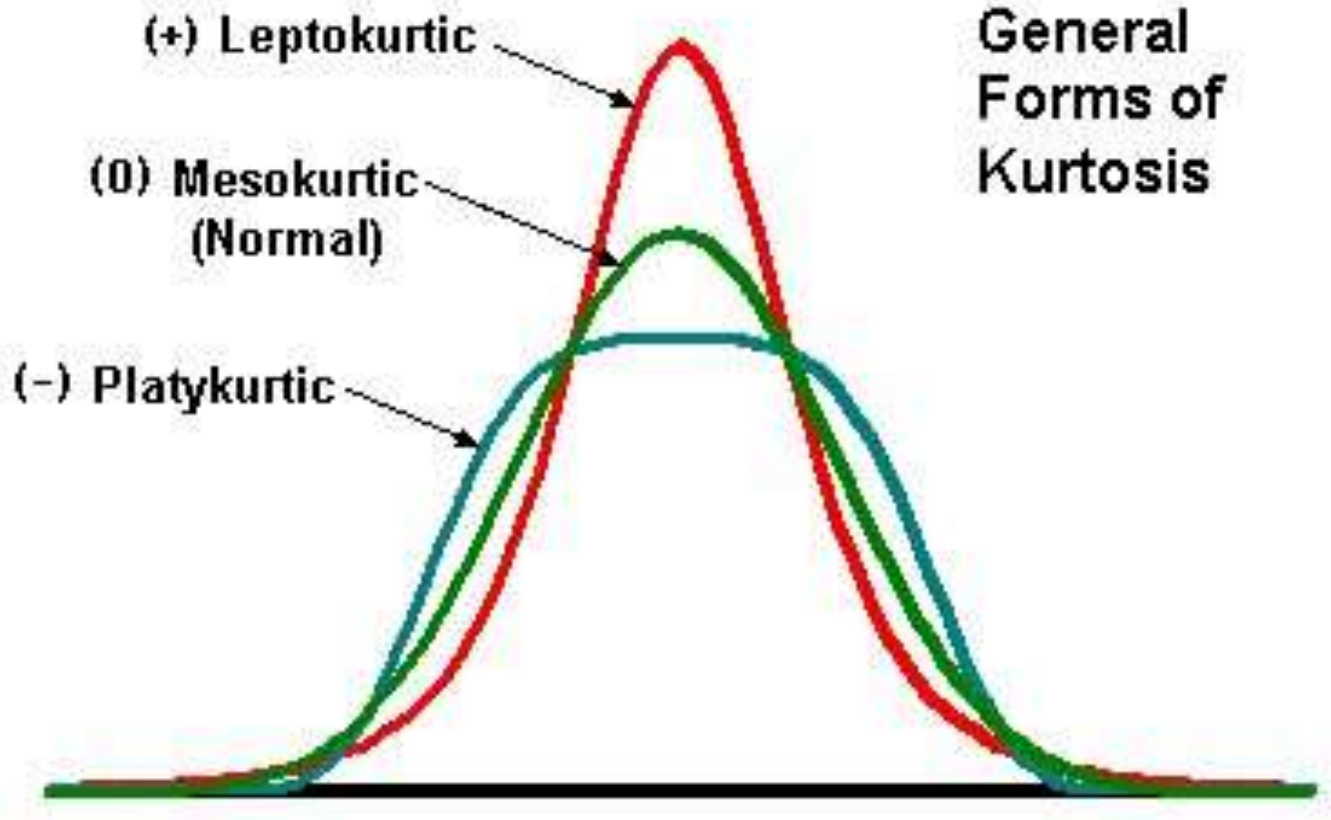
Distribution Shape: Skewness

- Highly Skewed Right
  - Skewness is positive (often above 1.0).
  - Mean will usually be more than the median.



Kurtosis

- Describes the amount of peakedness of a distribution.
- Lepto-kurtic: more peaked than normal curve
  Platy-kurtic: more flat-tapped than normal curve
  Meso-kurtic: a normal shape in a frequency distribution


(+) Leptokurtic
(0) Mesokurtic
(Normal)
(-) Platykurtic
General
Forms of
Kurtosis

z-Scores

- The z-score is often called the standardized value.
- It denotes the number of standard deviations a data value x_i is from the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Excel's STANDARDIZE function can be used to compute the z-score.

z-Scores

- An observation's z-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

z-Scores

- **Example: Apartment Rents**
  - z-Score of Smallest Value (425)

$$z = \frac{x_i - \bar{x}}{s} = \frac{425 - 490.80}{54.74} = -1.20$$

Standardized Values for Apartment Rents

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

Chebyshev's Theorem

- At least $(1 - 1/z^2)$ of the items in any data set will be within z standard deviations of the mean, where z is any value greater than 1.
- Chebyshev's theorem requires $z > 1$, but z need not be an integer.

Chebyshev's Theorem

- At least **75%** of the data values must be within **$z = 2$ standard deviations** of the mean.
- At least **89%** of the data values must be within **$z = 3$ standard deviations** of the mean.
- At least **94%** of the data values must be within **$z = 4$ standard deviations** of the mean.

Chebyshev's Theorem

- Example: Apartment Rents
  - Let $z = 1.5$ with $\bar{x} = 490.80$ and $s = 54.74$

  - At least $(1 - 1/(1.5)^2) = 1 - 0.44 = 0.56$ or 56% of the rent values must be between
  - $\bar{x} - z(s) = 490.80 - 1.5(54.74) =$ 409

and

$\bar{x} + z(s) = 490.80 + 1.5(54.74) =$ 573

(Actually, 86% of the rent values are between 409 and 573.)

Empirical Rule

When the data are believed to approximate a bell-shaped distribution …

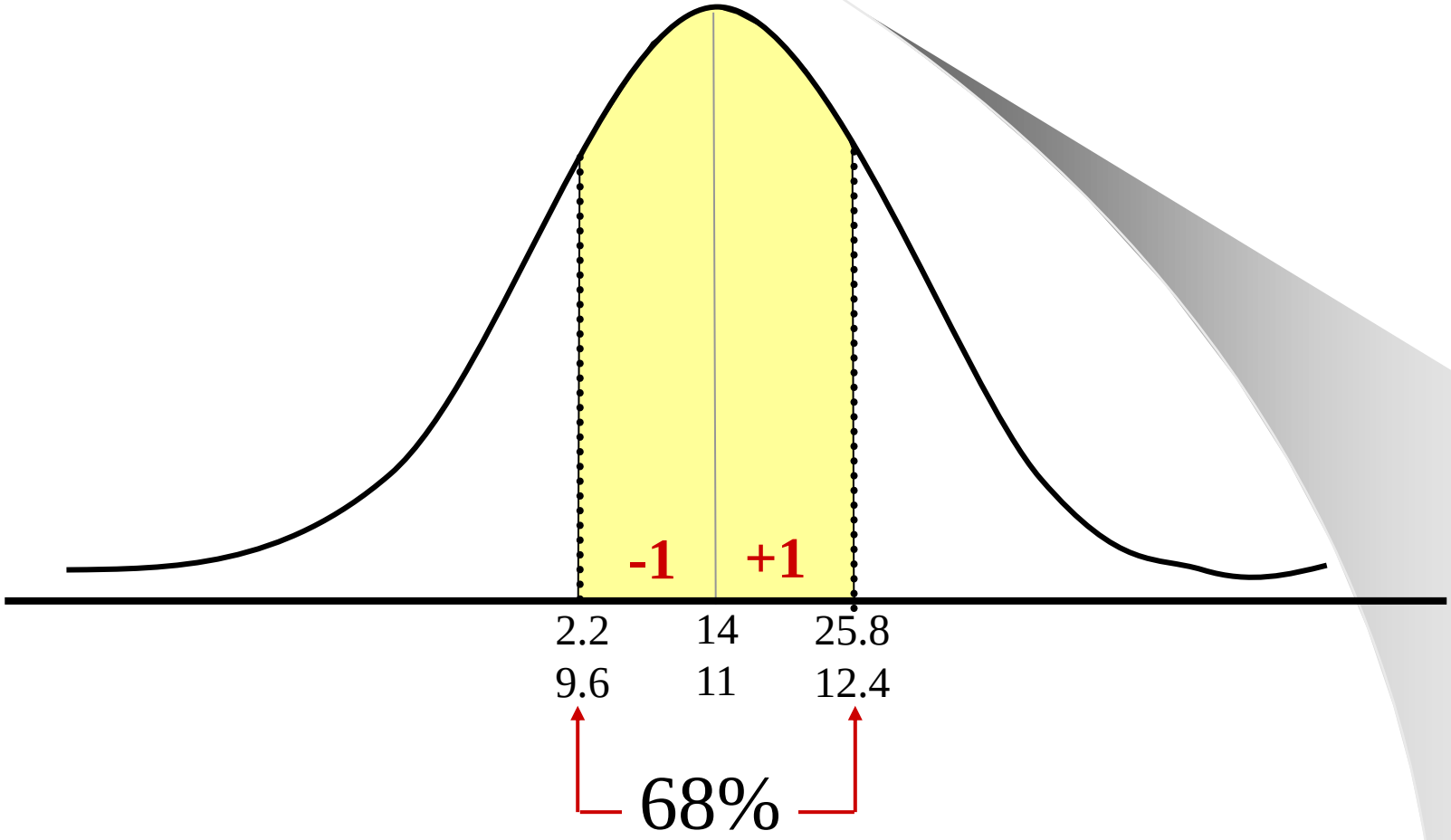
- The empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.
- The empirical rule is based on the normal distribution

Empirical Rule

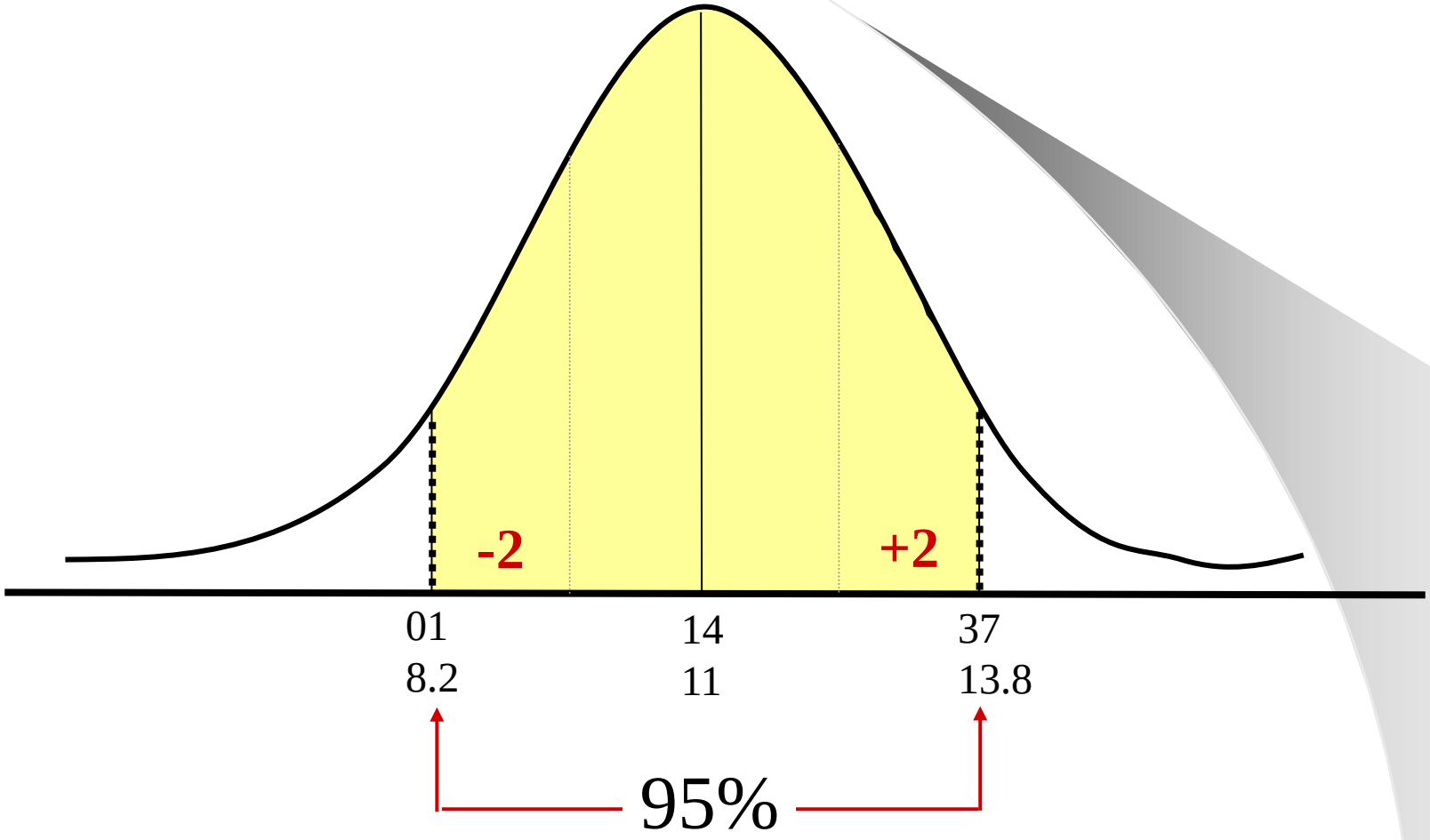
For data having a bell-shaped

- 68.26% of the values of a normal random variable are within +/- 1 standard deviation of its mean.
- 95.44% of the values of a normal random variable are within +/- 2 standard deviations of its mean.
- 99.72% of the values of a normal random variable are within +/- 3 standard deviations of its mean.

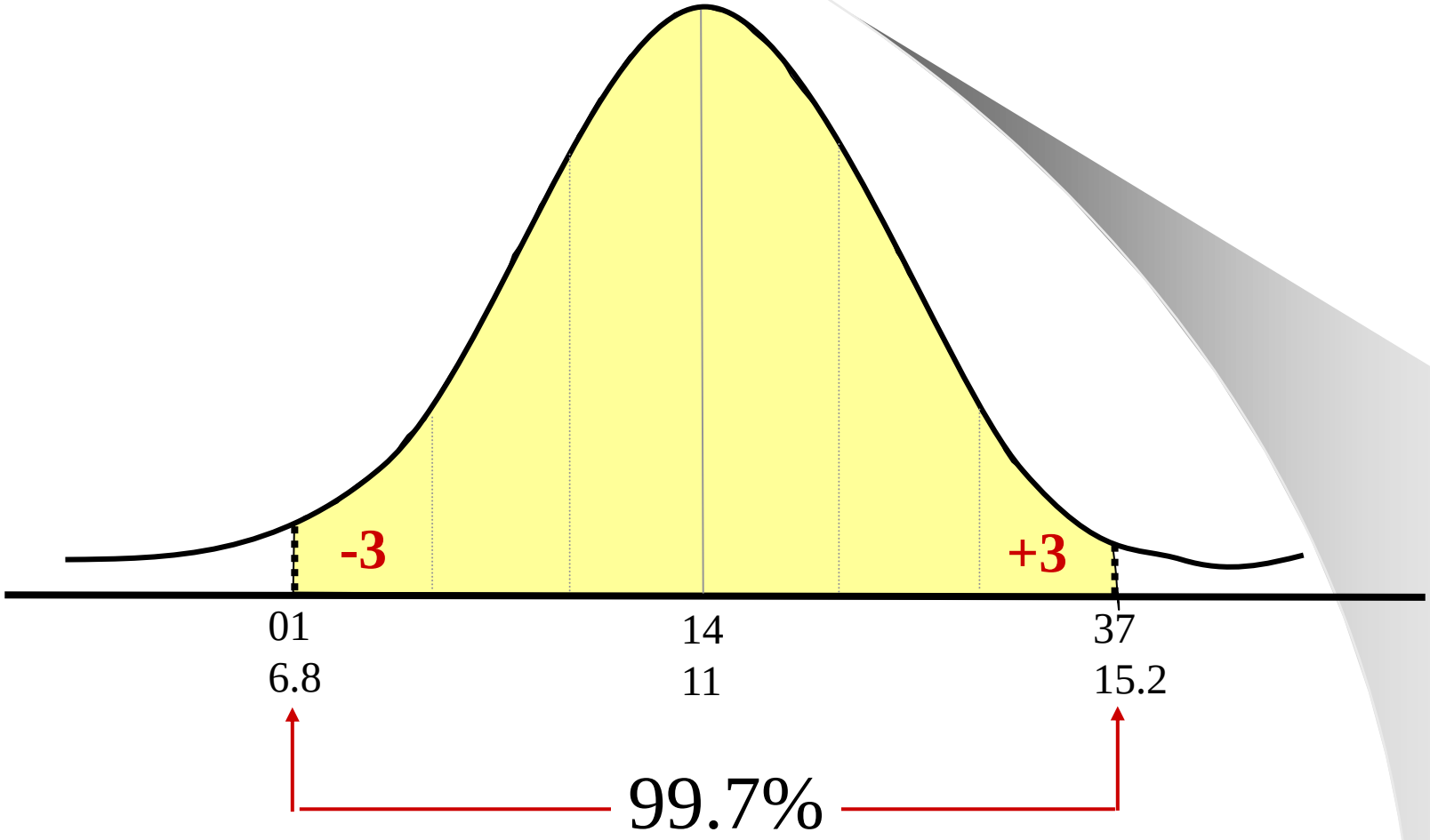
NORMAL DISTRIBUTION CURVE
1 Standard Deviation



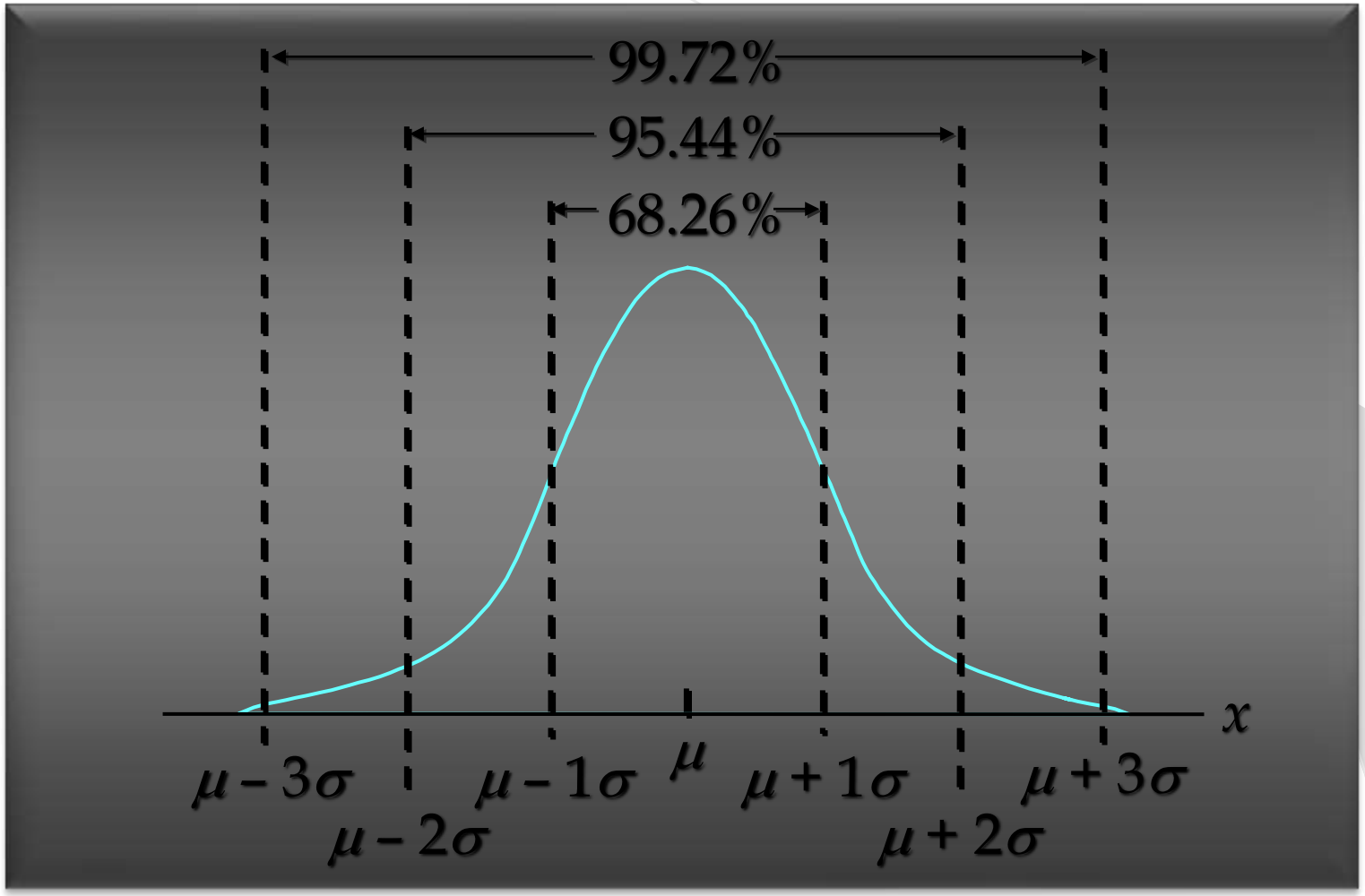
NORMAL DISTRIBUTION CURVE
2 Standard Deviations



NORMAL DISTRIBUTION CURVE
3 Standard Deviations



Empirical Rule



Detecting Outliers

- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded data value that belongs in the data set

Detecting Outliers

- Example: Apartment Rents
  - The most extreme z-scores are -1.20 and 2.27
  - Using $|z| \geq 3$ as the criterion for an outlier, there are no outliers in this data set.

Standardized Values for Apartment Rents

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

Five-Number Summaries and Box Plots

- Summary statistics and easy-to-draw graphs can be used to quickly summarize large quantities of data.
- Two tools that accomplish this are five-number summaries and box plots.

Five-Number Summary

1. Smallest Value
2. First Quartile
3. Median
4. Third Quartile
5. Largest Value

Five-Number Summary

- Example: Apartment Rents

Lowest Value = 425 First Quartile = 445

Median = 475

Third Quartile = 525 Largest Value = 615

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615