

Chapter 3

Descriptive Statistics: Numerical Measures

- q Measures of Location
- q Measures of Variability

Measures of Location

- ▶ q Mean
- q Median
- q Mode
- q Percentiles
- q Quartiles

▶ If the measures are computed for data from a sample, they are called sample statistics.

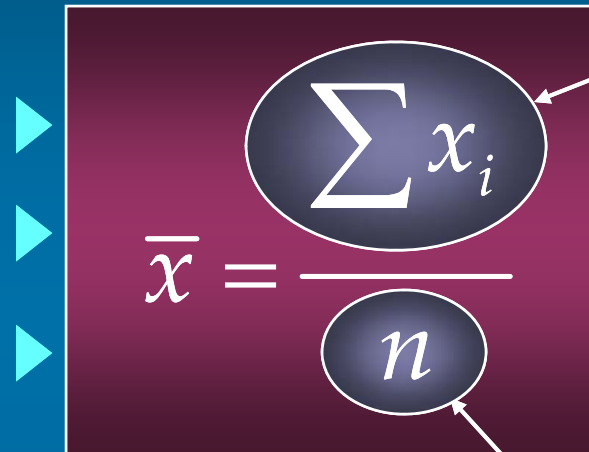
▶ If the measures are computed for data from a population, they are called population parameters.

▶ A sample statistic is referred to as the point estimator of the corresponding population parameter.

Mean

- q The mean of a data set is the average of all the data values.
- q The sample mean \bar{x} is the point estimator of the population mean μ .

Sample Mean \bar{x}



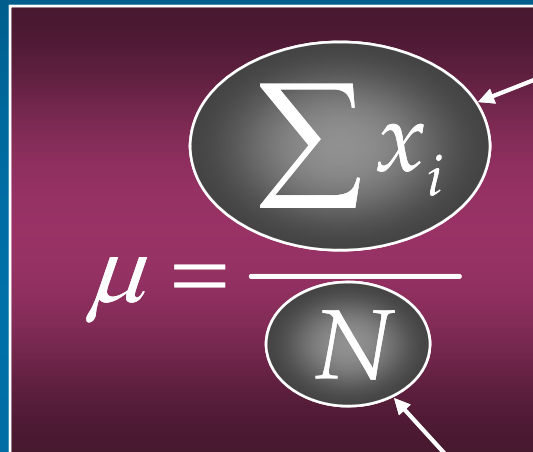
The diagram shows the formula for the sample mean, $\bar{x} = \frac{\sum x_i}{n}$, centered in a dark red square. The numerator, $\sum x_i$, is enclosed in a dark blue oval, and the denominator, n , is enclosed in a dark blue circle. Three light blue triangles point from the left towards the square. Two white callout boxes with arrows point to the numerator and denominator respectively, providing their definitions.

$$\bar{x} = \frac{\sum x_i}{n}$$

Sum of the values
of the n observations

Number of
observations
in the sample

Population Mean μ



The diagram shows the formula for the population mean μ inside a dark red square. The formula is $\mu = \frac{\sum x_i}{N}$. The numerator $\sum x_i$ is enclosed in a dark grey oval, and the denominator N is also enclosed in a dark grey oval. Two white arrows point from callout boxes to these ovals. On the left side of the square, there are three light blue right-pointing triangles.

$$\mu = \frac{\sum x_i}{N}$$

Sum of the values
of the N observations

Number of
observations in
the population

Sample Mean

- Example: Apartment Rents
- ▶ Seventy efficiency apartments were randomly sampled in a small college town. The monthly rent prices for these apartments are listed in ascending order on the next slide.



Sample Mean



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

Sample Mean



$$\blacktriangleright \bar{x} = \frac{\sum x_i}{n} = \frac{34,356}{70} = 490.80$$

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

Median

- ▶ ■ The median of a data set is the value in the middle when the data items are arranged in ascending order.
- ▶ ■ Whenever a data set has extreme values, the median is the preferred measure of central location.
- ▶ ■ The median is the measure of location most often reported for annual income and property value data.
- ▶ ■ A few extremely large incomes or property values can inflate the mean.

Median

- For an odd number of observations:

26	18	27	12	14	27	19
----	----	----	----	----	----	----

7 observations

▶

12	14	18	19	26	27	27
----	----	----	----	----	----	----

in ascending order

the median is the middle value.

$$\text{Median} = 19$$

Median

- For an even number of observations:

26 | 18 | 27 | 12 | 14 | 27 | 30 | 19 8 observations

▶ 12 | 14 | 18 | 19 | 26 | 27 | 27 | 30 in ascending order

the median is the average of the middle two values.

$$\text{Median} = (19 + 26) / 2 = 22.5$$

Median



▶ Averaging the 35th and 36th data values:

▶ Median = $(475 + 475)/2 = 475$

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

Mode

- ▶■ The mode of a data set is the value that occurs with greatest frequency.
- ▶■ The greatest frequency can occur at two or more different values.
- ▶■ If the data have exactly two modes, the data are bimodal.
- ▶■ If the data have more than two modes, the data are multimodal.

Mode



▶ 450 occurred most frequently (7 times)

Mode = 450

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

Percentiles

- ▶■ A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- ▶■ Admission test scores for colleges and universities are frequently reported in terms of percentiles.

Percentiles

- q The p th percentile of a data set is a value such that at least p percent of the items take on this value or less and at least $(100 - p)$ percent of the items take on this value or more.

Percentiles

▶ Arrange the data in ascending order.

▶ Compute index i , the position of the p th percentile.

$$i = (p/100)n$$

▶ If i is not an integer, round up. The p th percentile is the value in the i th position.

▶ If i is an integer, the p th percentile is the average of the values in positions i and $i+1$.

90th Percentile



- ▶ $i = (p/100)n = (90/100)70 = 63$
- ▶ Averaging the 63rd and 64th data values:
- ▶ 90th Percentile = $(580 + 590)/2 = 585$

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

90th Percentile



▶ “At least 90% of the items take on a value of 585 or less.”

▶ “At least 10% of the items take on a value of 585 or more.”

▶ $63/70 = .9$ or 90%

▶ $7/70 = .1$ or 10%

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

Quartiles

- ▶ ■ Quartiles are specific percentiles.
- ▶ ■ First Quartile = 25th Percentile
- ▶ ■ Second Quartile = 50th Percentile = Median
- ▶ ■ Third Quartile = 75th Percentile

Third Quartile



- ▶ Third quartile = 75th percentile
- ▶ $i = (p/100)n = (75/100)70 = 52.5 = 53$
Third quartile = 525

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

Measures of Variability

- ▶■ It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- ▶■ For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Measures of Variability

- ▶ q Range
- ▶ q Interquartile Range
- ▶ q Variance
- ▶ q Standard Deviation
- ▶ q Coefficient of Variation

Range

- ▶■ The range of a data set is the difference between the largest and smallest data values.
- ▶■ It is the simplest measure of variability.
- ▶■ It is very sensitive to the smallest and largest data values.

Range



▶ Range = largest value - smallest value

$$\text{Range} = 615 - 425 = 190$$

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

Interquartile Range

- ▶ ■ The interquartile range of a data set is the difference between the third quartile and the first quartile.
- ▶ ■ It is the range for the middle 50% of the data.
- ▶ ■ It overcomes the sensitivity to extreme data values.

Interquartile Range



- ▶ 3rd Quartile ($Q3$) = 525
- ▶ 1st Quartile ($Q1$) = 445
- ▶ Interquartile Range = $Q3 - Q1 = 525 - 445 = 80$

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

Variance

▶ The variance is a measure of variability that utilizes all the data.

▶ It is based on the difference between the value of each observation (x_i) and the mean (\bar{x} for a sample, μ for a population).

Variance

▶ The variance is the average of the squared differences between each data value and the mean.

▶ The variance is computed as follows:

▶
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

for a
sample

◀
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a
population

Standard Deviation

- ▶ The standard deviation of a data set is the positive square root of the variance.
- ▶ It is measured in the same units as the data, making it more easily interpreted than the variance.

Standard Deviation

- ▶ The standard deviation is computed as follows:

- ▶ $s = \sqrt{s^2}$

for a
sample

- $\sigma = \sqrt{\sigma^2}$ ◀

for a
population

Coefficient of Variation

▶ The coefficient of variation indicates how large the standard deviation is in relation to the mean.

▶ The coefficient of variation is computed as follows:

$$\left(\frac{s}{\bar{x}} \times 100 \right) \% \quad \left(\frac{\sigma}{\mu} \times 100 \right) \%$$

for a
sample

for a
population

Descriptive Statistics: Numerical Measures

- ▶ q Measures of Distribution Shape, Relative Location, and Detecting Outliers

Measures of Distribution Shape, Relative Location, and Detecting Outliers

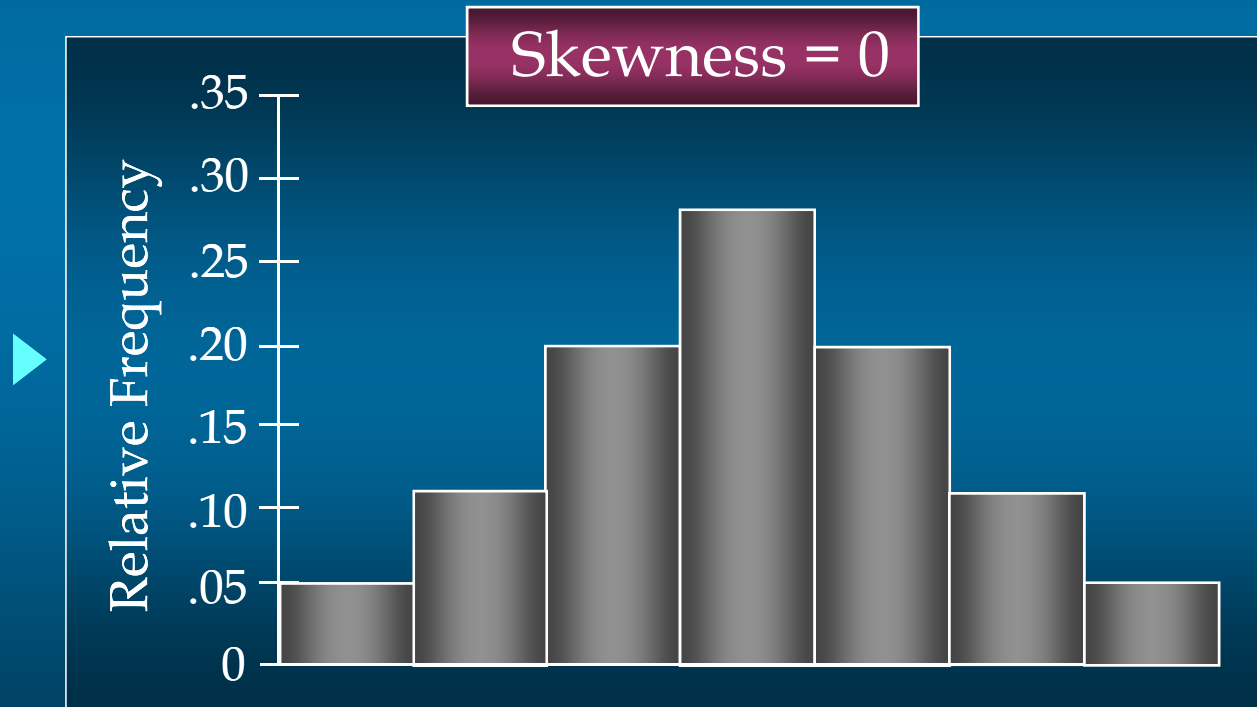
- q Distribution Shape
- q z-Scores
- q Detecting Outliers

Distribution Shape: Skewness

- ▶ q An important measure of the shape of a distribution is called skewness.
- ▶ q The formula for computing skewness for a data set is somewhat complex.
- ▶ q Skewness can be easily computed using statistical software.

Distribution Shape: Skewness

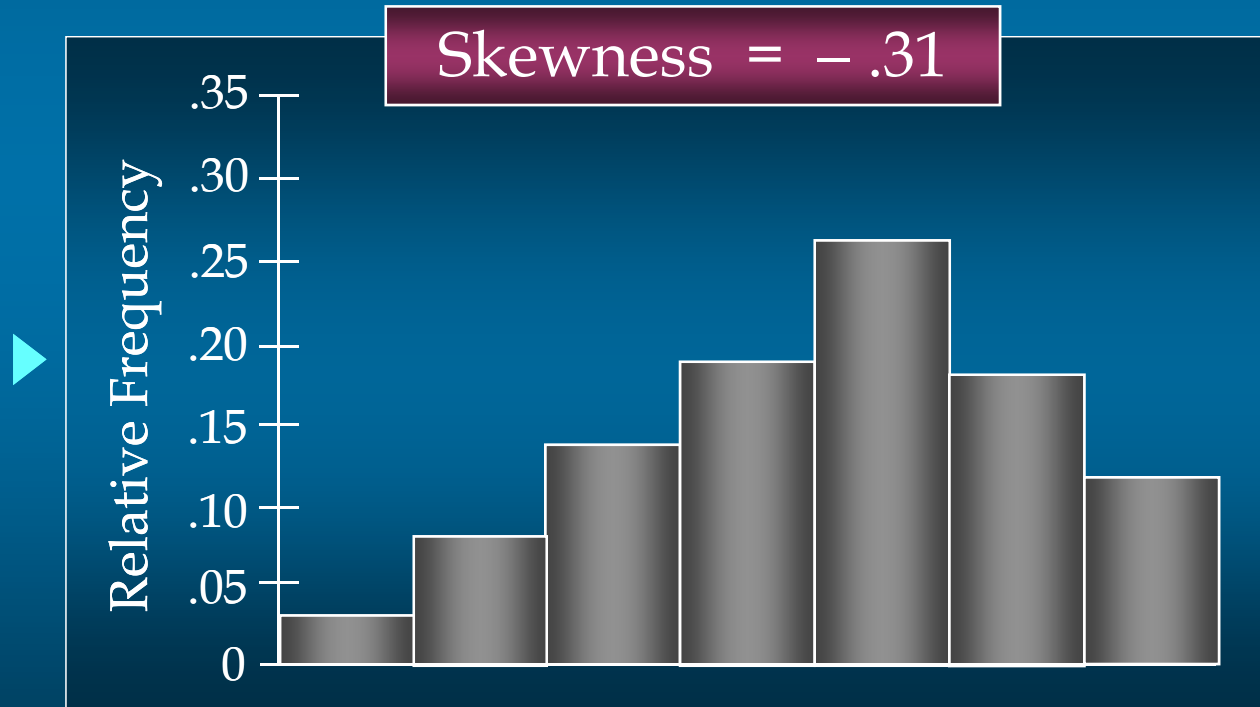
- q Symmetric (not skewed)
 - Skewness is zero.
 - Mean and median are equal.



Distribution Shape: Skewness

q Moderately Skewed Left

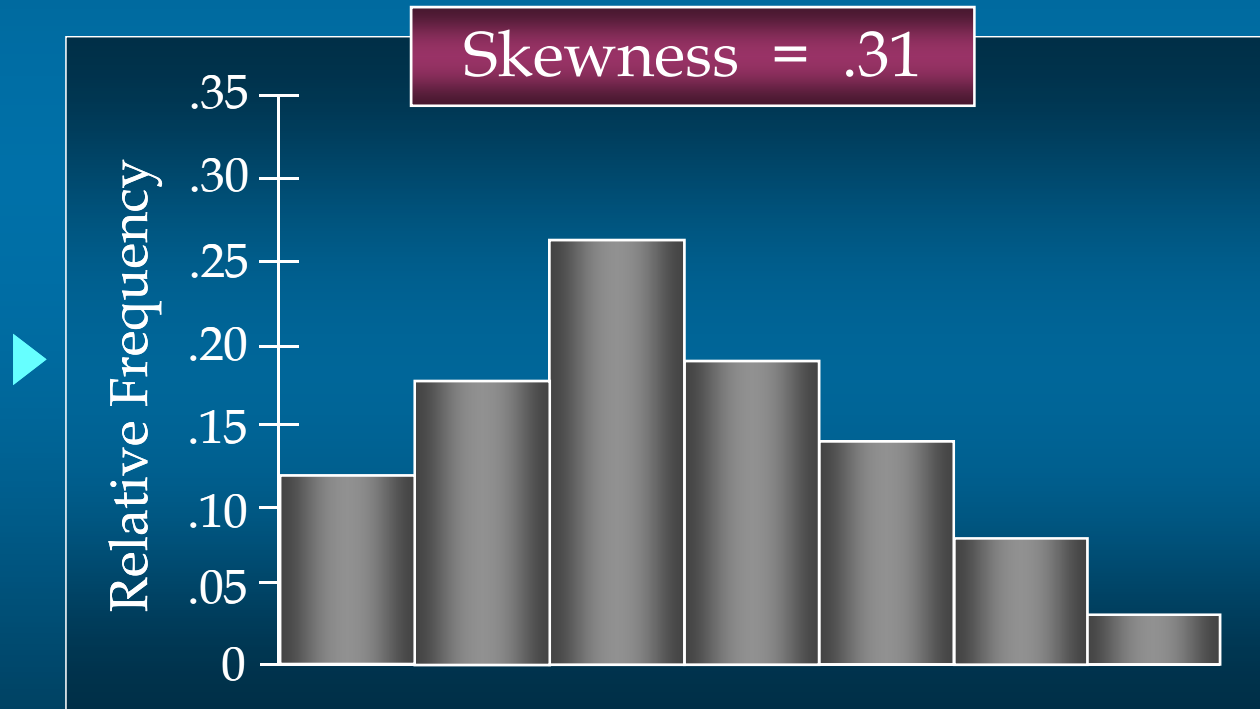
- Skewness is negative.
- Mean will usually be less than the median.



Distribution Shape: Skewness

q Moderately Skewed Right

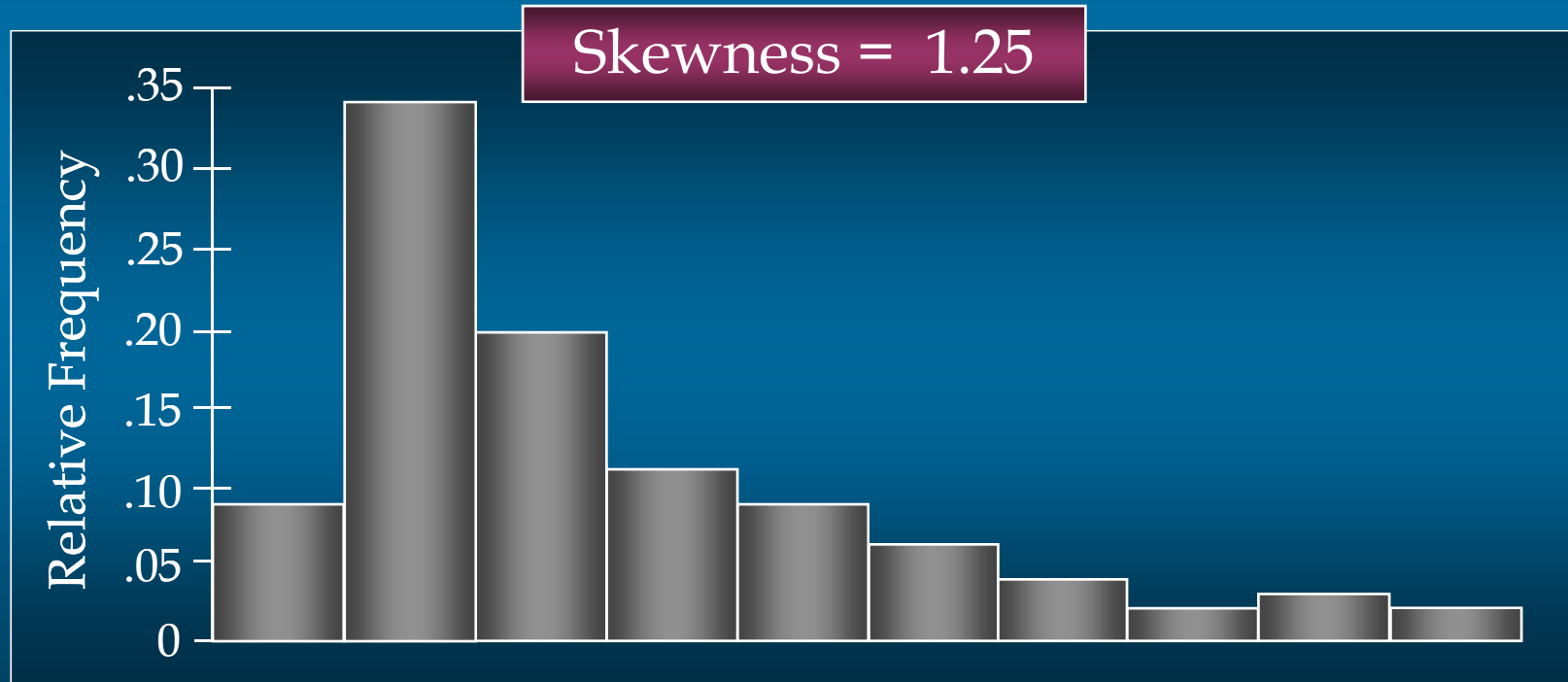
- Skewness is positive.
- Mean will usually be more than the median.



Distribution Shape: Skewness

q Highly Skewed Right

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



Distribution Shape: Skewness

q Example: Apartment Rents

- ▶ Seventy efficiency apartments were randomly sampled in a small college town. The monthly rent prices for these apartments are listed in ascending order on the next slide.

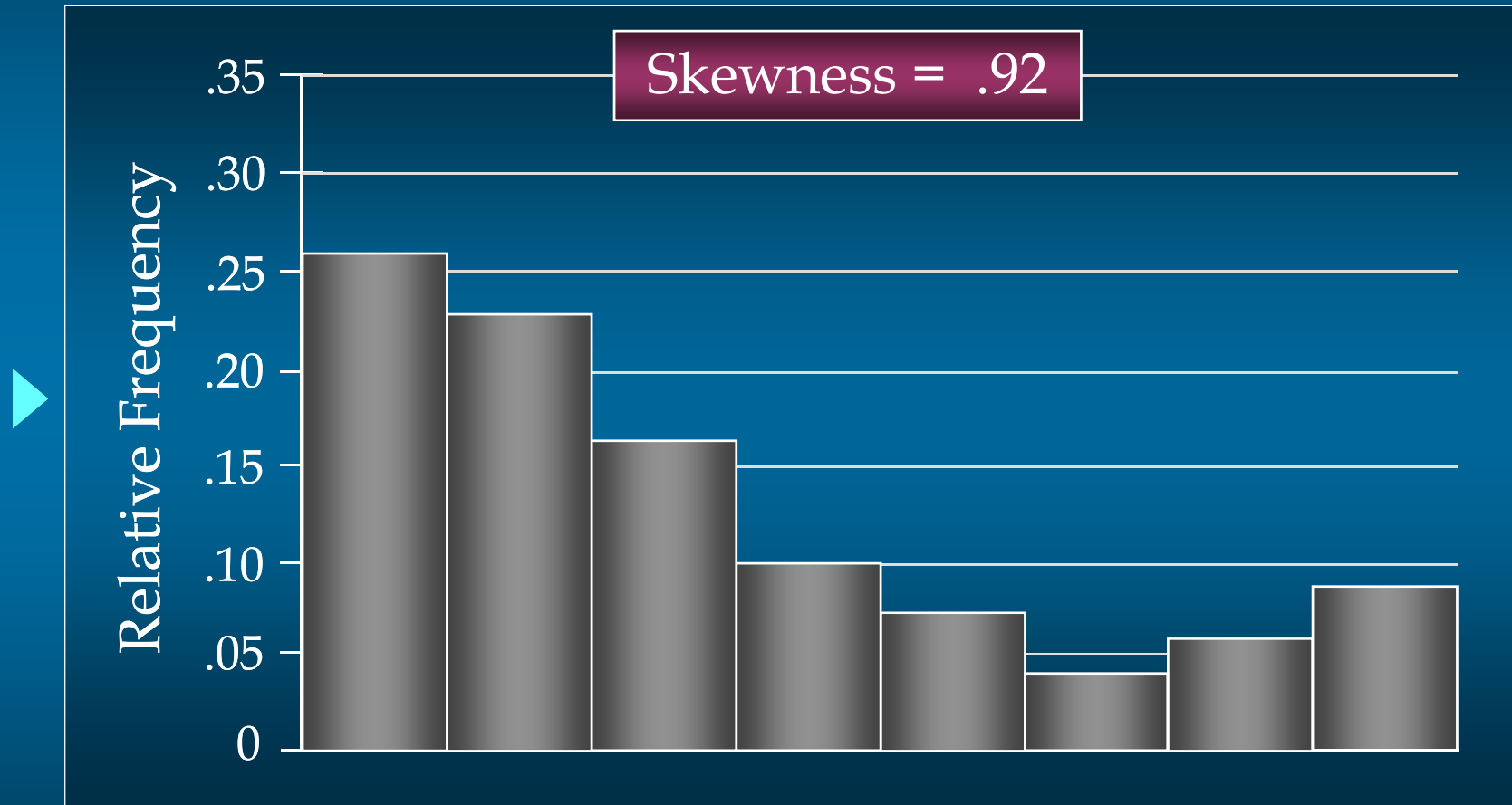


Distribution Shape: Skewness



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

Distribution Shape: Skewness



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.

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

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



q 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

Empirical Rule

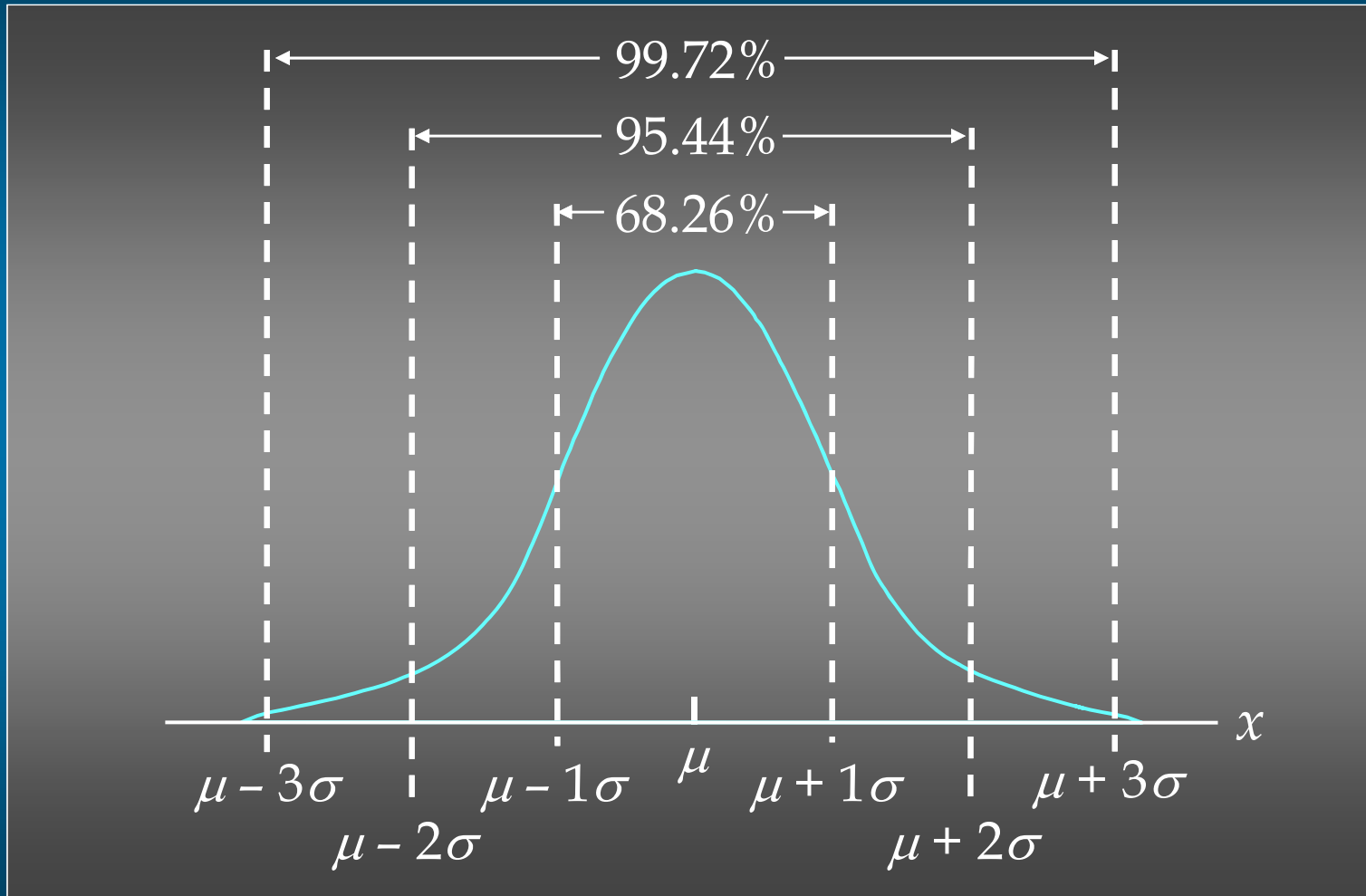
For data having a bell-shaped distribution:

▶ 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.

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



- ▶ ■ 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