

1 – Chapter 2.1 Frequency Distribution	2 - Chapter 2.1 Class width
3 - Chapter 2.1 Midpoint	4 - Chapter 2.1 Relative frequency
5 - Chapter 2.1 cumulative frequency	6 - Chapter 2.1 frequency histogram
7 - Chapter 2.1 Class boundaries	8 - Chapter 2.1 Frequency polygon

The **class width** is the distance between lower (or upper) limits of consecutive classes. Round up to the next convenient number.

$$\frac{\text{range}}{\# \text{ of classes}} = \text{class width (round up)}$$

A **frequency distribution** is a table that shows **classes** or **intervals** of data with a count of the number in each class. The frequency f of a class is the number of data points in the class.

The **relative frequency** of a class is the portion or percentage of the data that falls in that class. To find the relative frequency of a class, divide the frequency f by the sample size n .

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Sample size}} = \frac{f}{n}$$

The **midpoint** of a class is the sum of the lower and upper limits of the class divided by two. The midpoint is sometimes called the *class mark*.

$$\text{midpoint} = \frac{(\text{lower class limit}) + (\text{upper class limit})}{2}$$

A **frequency histogram** is a bar graph that represents the frequency distribution of a data set.

1. The horizontal scale is quantitative and measures the data values.
2. The vertical scale measures the frequencies of the classes.
3. Consecutive bars must touch.

The **cumulative frequency** of a class is the sum of the frequency for that class and all the previous classes.

A **frequency polygon** is a line graph that emphasizes the continuous change in frequencies.

Class boundaries are the numbers that separate the classes without forming gaps between them.

The horizontal scale of a histogram can be marked with either the class boundaries or the midpoints

9 - Chapter 2.1

relative frequency histogram

10 - Chapter 2.1

cumulative frequency graph

11 - Chapter 2

12 - Chapter 2

13 - Chapter 2

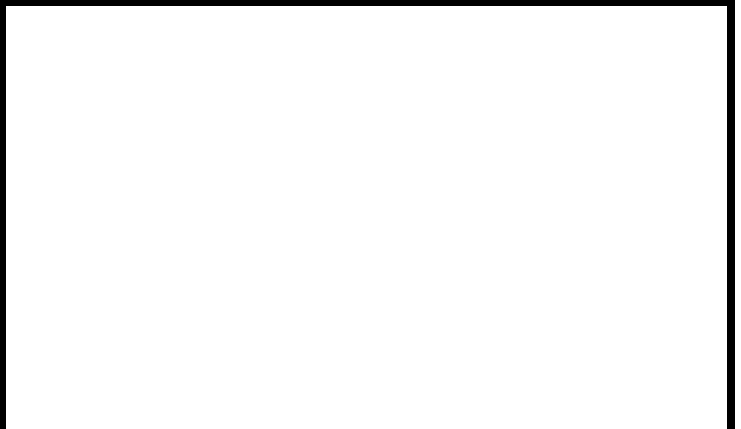
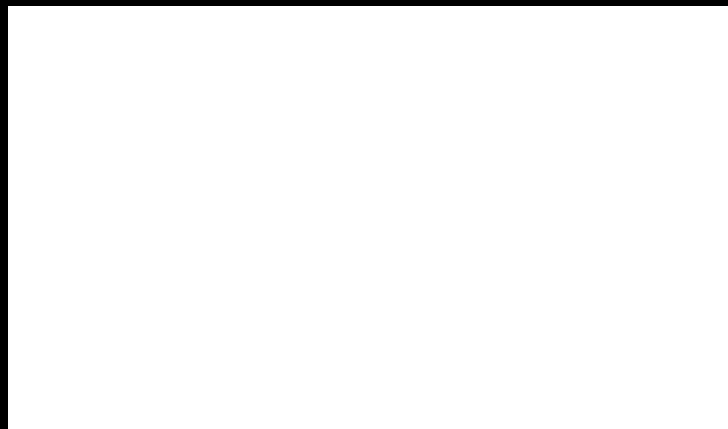
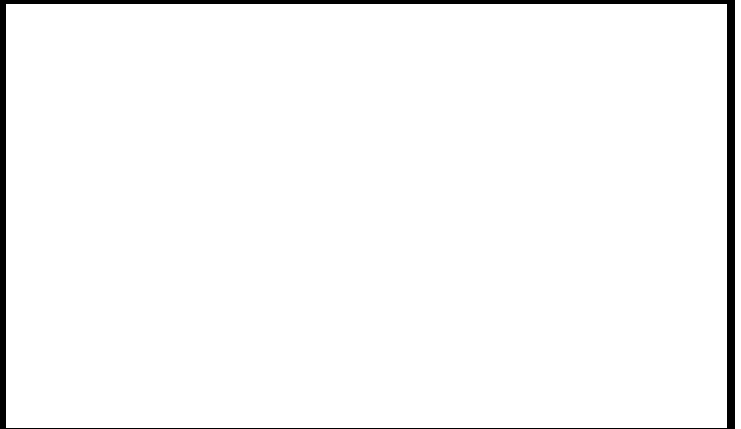
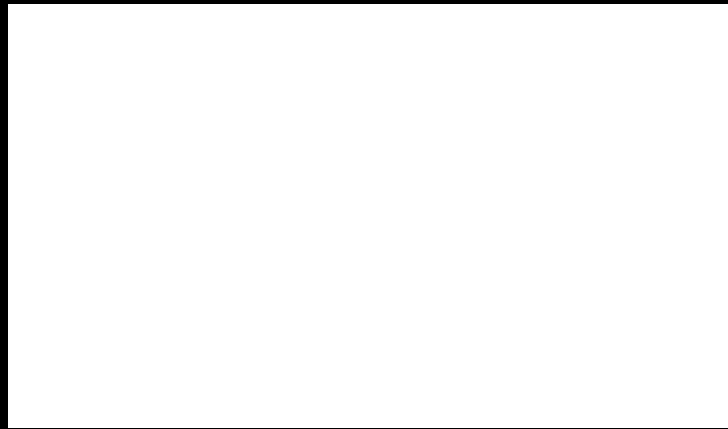
14 - Chapter 2

15 - Chapter 2

16 - Chapter 2

A **cumulative frequency graph** or **ogive**, is a line graph that displays the cumulative frequency of each class at its upper class boundary.

A **relative frequency histogram** has the same shape and the same horizontal scale as the corresponding frequency histogram.



1 – Chapter 2.2

stem-and-leaf plot

2 - Chapter 2.2

dot plot

3 - Chapter 2.2

pie chart

4 - Chapter 2.2

Pareto chart

5 - Chapter 2.2

scatter plot

6 - Chapter 2.2

time series chart

7 - Chapter 2.2

8 - Chapter 2.2

In a **dot plot**, each data entry is plotted, using a point, above a horizontal axis.

In a **stem-and-leaf plot**, each number is separated into a stem (the most significant digits) and a leaf (usually the rightmost digit). This is an example of **exploratory data analysis**.

A **Pareto chart** is a vertical bar graph in which the height of each bar represents the frequency. The bars are placed in order of decreasing height, with the tallest bar to the left.

A **pie chart** is a circle that is divided into sectors that represent categories. The area of each sector is proportional to the frequency of each category.

A data set that is composed of quantitative data entries taken at regular intervals over a period of time is a **time series**. A **time series chart** is used to graph a time series.

When each entry in one data set corresponds to an entry in another data set, the sets are called **paired data sets**.

In a **scatter plot**, the ordered pairs are graphed as points in a coordinate plane. The scatter plot is used to show the relationship between two quantitative variables.

1 - Chapter 2.3 mean	2 - Chapter 2.3 median
3 - Chapter 2.3 mode	4 - Chapter 2.3 outlier
5 - Chapter 2.3 weighted mean	6 - Chapter 2.3 mean of a frequency distribution for a sample is approximated by
7 - Chapter 2.3 symmetric (shape of distribution)	8 - Chapter 2.3 uniform or rectangular (shape of distribution)

The **median** of a data set is the value that lies in the middle of the data when the data set is ordered. If the data set has an odd number of entries, the median is the middle data entry. If the data set has an even number of entries, the median is the mean of the two middle data entries.

The **mean** of a data set is the sum of the data entries divided by the number of entries.

$$\text{population mean: } \mu = \frac{\sum x}{N}$$

$$\text{sample mean: } \bar{x} = \frac{\sum x}{N}$$

An **outlier** is a data entry that is far removed from the other entries in the data set.

The **mode** of a data set is the data entry that occurs with the greatest frequency. If no entry is repeated, the data set has no mode. If two entries occur with the same greatest frequency, each entry is a mode and the data set is called **bimodal**.

The **mean of a frequency distribution** for a sample is approximated by

$$\bar{x} = \frac{\sum(x \cdot f)}{n} \quad \text{Note that } n = \sum f$$

where x and f are the midpoints and frequencies of the classes

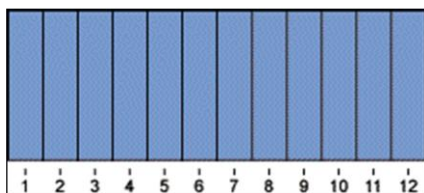
An outlier is a data entry that is far removed from the other entries in the data set. A weighted mean is given by

$$\bar{x} = \frac{\sum(x \cdot w)}{\sum w}$$

where w is the weight of each entry x

A frequency distribution is **uniform** (or **rectangular**) when all entries, or classes, in the distribution have equal frequencies. A uniform distribution is also symmetric.

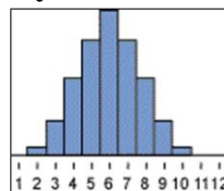
Uniform



Mean = Median

A frequency distribution is **symmetric** when a vertical line can be drawn through the middle of a graph of the distribution and the resulting halves are approximately the mirror images.

Symmetric



Mean = Median

9 - Chapter 2.3

skewed (shape of distribution)

10 - Chapter 2.3

11 - Chapter 2.3

12 - Chapter 2.3

13 -

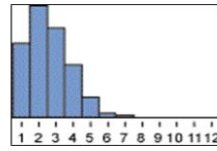
14 -

15 -

16 -

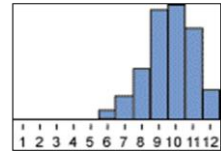
elongates more to one side than to the other. A distribution is **skewed left (negatively skewed)** if its tail extends to the left. A distribution is **skewed right (positively skewed)** if its tail extends to the right.

Skewed right



Mean > Median

Skewed left



Mean < Median

1 - Chapter 2.4 range	2 - Chapter 2.4 deviation
3 - Chapter 2.4 population variance	4 - Chapter 2.4 population standard deviation
5 - Chapter 2.4 Finding the Population Standard Deviation	6 - Chapter 2.4 Finding the Sample Standard Deviation
7 - Chapter 2.4 Empirical Rule	8 - Chapter 2.4 Chebychev's Theorem (slide 1)

The **deviation** of an entry x in a population data set is the set.

Deviation of $x = x$

The **range** of a data set is the difference between the maximum and minimum data entries in the set.
 Range = (Maximum data entry) - (Minimum data entry)

The **population standard deviation** of a population data set of N entries is the square root of the population variance.

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2}$$

$$= \frac{\sqrt{\sum(x - \mu)^2}}{N}$$

$\sigma = \text{sigma}$

The **population variance** of a population data set of N entries is

$$\text{population variance} = \sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

$\sigma = \text{sigma}$

Guidelines

In Words

1. Find the mean of the sample data set.
2. Find the deviation of each entry.
3. Square each deviation.
4. Add to get the **sum of squares**.
5. Divide by $n - 1$ to get the **sample variance**.
6. Find the square root of the variance to get the **sample standard deviation**.

In Symbols

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

$$x - \bar{x}$$

$$(x - \bar{x})^2$$

$$SS_x = \sum (x - \bar{x})^2$$

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

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

Guidelines

In Words

1. Find the mean of the population data set.
2. Find the deviation of each entry.
3. Square each deviation.
4. Add to get the **sum of squares**.
5. Divide by N to get the **population variance**.
6. Find the square root of the variance to get the **population standard deviation**.

In Symbols

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

$$x - \mu$$

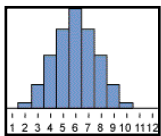
$$(x - \mu)^2$$

$$SS_x = \sum (x - \mu)^2$$

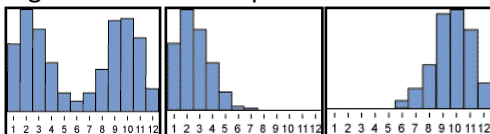
$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

The Empirical Rule is only used for **symmetric distributions**.



any distribution, regardless of the shape.



Empirical Rule

For data with a (symmetric) bell-shaped distribution, the standard deviation has the following characteristics.

1. About 68% of the data lie within one standard deviation of the mean.
2. About 95% of the data lie within two standard deviations of the mean.
3. About 99.7% of the data lie within three standard deviation of the mean.

9 - Chapter 2.4

Chebychev u (slide 2)

10 - Chapter 2.4

Standard Deviation for Grouped Data

11 -

12 -

13 -

14 -

15 -

16 -

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum(x - \bar{x})^2 f}{n - 1}}$$

where $n = \sum f$ is the number of entries in the data set, and x is the data value or the midpoint of an interval.

The portion of any data set lying within k standard deviations ($k > 1$) of the mean is at least

$$1 - \frac{1}{k^2}$$

For $k = 2$: In any data set, at least $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}$, or 75%, of the data lie within 2 standard deviations of the mean.

For $k = 3$: In any data set, at least $1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9}$, or 88.9%, of the data lie within 3 standard deviations of the mean.

1 - Chapter 2.5

Finding Quartiles

2 - Chapter 2.5

Interquartile Range

3 - Chapter 2.5

box-and-whisker plot

4 - Chapter 2.5

percentiles and deciles

5 - Chapter 2.5

standard scores

6 - Chapter 2.5

7 - Chapter 2.5

8 - Chapter 2.5

<p>The interquartile range (IQR) of a data set is the difference between the third and first quartiles. Interquartile range (IQR) = $Q_3 - Q_1$.</p>	<p>The three quartiles, Q_1, Q_2, and Q_3, approximately divide an ordered data set into four equal parts.</p>
<p>Fractiles are numbers that partition, or divide, an ordered data set.</p> <p>Percentiles divide an ordered data set into 100 parts. There are 99 percentiles: $P_1, P_2, P_3, \dots, P_{99}$.</p> <p>Deciles divide an ordered data set into 10 parts. There are 9 deciles: $D_1, D_2, D_3, \dots, D_9$.</p>	<p>A box-and-whisker plot is an exploratory data analysis tool that highlights the important features of a data set.</p> <p>The five-number summary is used to draw the graph.</p> <p>The minimum entry Q_1 Q_2 (median) Q_3 The maximum entry</p>
	<p>The standard score or z-score, represents the number of standard deviations that a data value, x, falls from the mean, μ.</p> $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$