Chapter 2

Descriptive Statistics

§ 2.1

Frequency Distributions and Their Graphs

Frequency Distributions

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.

Class	Frequency, *f*
1 – 4	4
5 – 8	5
9 – 12	3
13 – 16	4
17 – 20	2

Lower Class Limits

Upper Class Limits

Frequencies

Frequency Distributions

The **class width** is the distance between lower (or upper) limits of consecutive classes.

Class	Frequency, *f*
1 – 4	4
5 – 8	5
9 – 12	3
13 – 16	4
17 – 20	2

5 – 1 = 4
9 – 5 = 4
13 – 9 = 4
17 – 13 = 4

The class width is 4.

The **range** is the difference between the maximum and minimum data entries.

Constructing a Frequency Distribution

Guidelines

1. Decide on the number of classes to include. The number of classes should be between 5 and 20; otherwise, it may be difficult to detect any patterns.
2. Find the class width as follows. Determine the range of the data, divide the range by the number of classes, and *round up to the next convenient number*.
3. Find the class limits. You can use the minimum entry as the lower limit of the first class. To find the remaining lower limits, add the class width to the lower limit of the preceding class. Then find the upper class limits.
4. Make a tally mark for each data entry in the row of the appropriate class.
5. Count the tally marks to find the total frequency *f* for each class.

Constructing a Frequency Distribution

Example:

The following data represents the ages of 30 students in a statistics class. Construct a frequency distribution that has five classes.

Ages of Students

18	20	21	27	29	20
19	30	32	19	34	19
24	29	18	37	38	22
30	39	32	44	33	46
54	49	18	51	21	21

Continued.

Constructing a Frequency Distribution

Example continued:

1. The number of classes (5) is stated in the problem.

2. The minimum data entry is 18 and maximum entry is 54, so the range is 36. Divide the range by the number of classes to find the class width.

Class width = $\frac{36}{5} = 7.2$ Round up to 8.

Continued.

Constructing a Frequency Distribution

Example continued:

3. The minimum data entry of 18 may be used for the lower limit of the first class. To find the lower class limits of the remaining classes, add the width (8) to each lower limit.

The lower class limits are 18, 26, 34, 42, and 50.
The upper class limits are 25, 33, 41, 49, and 57.

4. Make a tally mark for each data entry in the appropriate class.

5. The number of tally marks for a class is the frequency for that class.

Continued.

Constructing a Frequency Distribution

Example continued:

Ages → Class; Number of students → Frequency, f

Ages of Students

Class	Tally	Frequency, f
18 – 25	~~||||~~ ~~||||~~ |||	13
26 – 33	~~||||~~ |||	8
34 – 41	||||	4
42 – 49	|||	3
50 – 57	||	2
		$\Sigma f = 30$

Check that the sum equals the number in the sample.

Midpoint

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*.

$$\text{Midpoint} = \frac{(\text{Lower class limit}) + (\text{Upper class limit})}{2}$$

Class	Frequency, f	Midpoint
1 – 4	4	2.5

$$\text{Midpoint} = \frac{1+4}{2} = \frac{5}{2} = 2.5$$

Midpoint

Example:
Find the midpoints for the "Ages of Students" frequency distribution.

Ages of Students

Class	Frequency, f	Midpoint
18 – 25	13	21.5
26 – 33	8	29.5
34 – 41	4	37.5
42 – 49	3	45.5
50 – 57	2	53.5
	$\sum f = 30$	

18 + 25 = 43
43 ÷ 2 = 21.5

Relative Frequency

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.

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

Class	Frequency, f	Relative Frequency
1 – 4	4	0.222

$\sum f = 18$

$$\text{Relative frequency} = \frac{f}{n} = \frac{4}{18} \approx 0.222$$

Relative Frequency

Example:
Find the relative frequencies for the "Ages of Students" frequency distribution.

Class	Frequency, f	Relative Frequency
18 – 25	13	0.433
26 – 33	8	0.267
34 – 41	4	0.133
42 – 49	3	0.1
50 – 57	2	0.067
	$\Sigma f = 30$	$\Sigma \frac{f}{n} = 1$

Portion of students

$\frac{f}{n} = \frac{13}{30}$

≈ 0.433

Cumulative Frequency

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

Ages of Students

Class	Frequency, f	Cumulative Frequency
18 – 25	13	13
26 – 33	+ 8	21
34 – 41	+ 4	25
42 – 49	+ 3	28
50 – 57	+ 2	30
	$\Sigma f = 30$	

Total number of students

Frequency Histogram

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.

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.

Class Boundaries

Example:
Find the class boundaries for the "Ages of Students" frequency distribution.

The distance from the upper limit of the first class to the lower limit of the second class is 1.

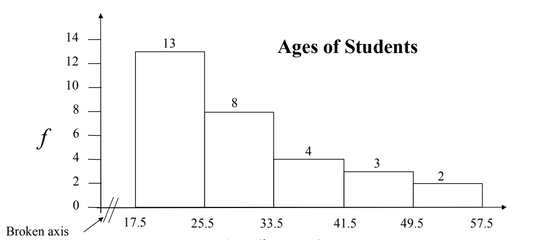
Half this distance is 0.5.

Ages of Students

Class	Frequency, f	Class Boundaries
18 – 25	13	17.5 – 25.5
26 – 33	8	25.5 – 33.5
34 – 41	4	33.5 – 41.5
42 – 49	3	41.5 – 49.5
50 – 57	2	49.5 – 57.5
	$\sum f = 30$	

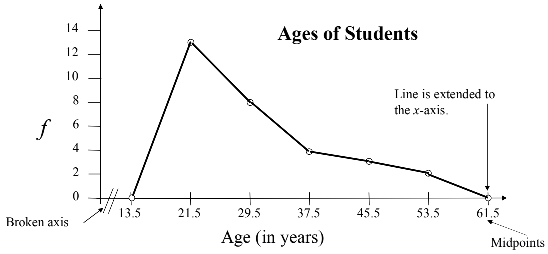
Frequency Histogram

Example:
Draw a frequency histogram for the "Ages of Students" frequency distribution. Use the class boundaries.



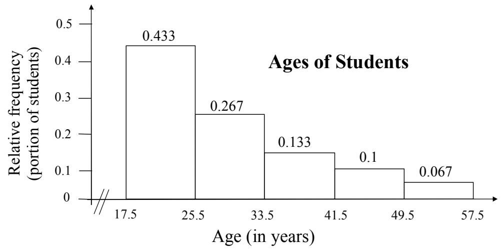
Frequency Polygon

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



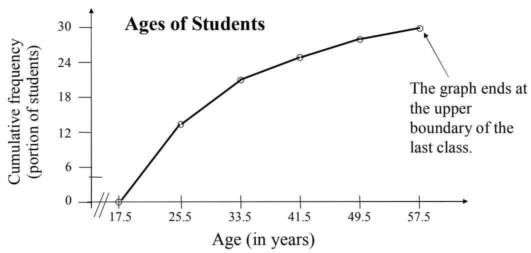
Relative Frequency Histogram

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



Cumulative Frequency Graph

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



§ 2.2

More Graphs and Displays

Stem-and-Leaf Plot

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

Example:
The following data represents the ages of 30 students in a statistics class. Display the data in a stem-and-leaf plot.

Ages of Students

18	20	21	27	29	20
19	30	32	19	34	19
24	29	18	37	38	22
30	39	32	44	33	46
54	49	18	51	21	21

Continued.

Stem-and-Leaf Plot

Ages of Students

Key: 1|8 = 18

1	8 8 8 9 9 9
2	0 0 1 1 1 2 4 7 9 9
3	0 0 2 2 3 4 7 8 9
4	4 6 9
5	1 4

Most of the values lie between 20 and 39.

This graph allows us to see the shape of the data as well as the actual values.

Stem-and-Leaf Plot

Example:
Construct a stem-and-leaf plot that has two lines for each stem.

Ages of Students

Key: 1|8 = 18

1	
1	8 8 8 9 9 9
2	0 0 1 1 1 2 4
2	7 9 9
3	0 0 2 2 3 4
3	7 8 9
4	4
4	6 9
5	1 4
5	

From this graph, we can conclude that more than 50% of the data lie between 20 and 34.

Dot Plot

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

Example:
Use a dot plot to display the ages of the 30 students in the statistics class.

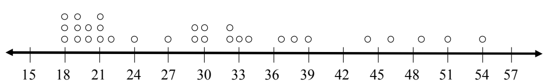
Ages of Students

18	20	21	27	29	20
19	30	32	19	34	19
24	29	18	37	38	22
30	39	32	44	33	46
54	49	18	51	21	21

Continued.

Dot Plot

Ages of Students



From this graph, we can conclude that most of the values lie between 18 and 32.

Pie Chart

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.

Accidental Deaths in the USA in 2002

Type	Frequency
Motor Vehicle	43,500
Falls	12,200
Poison	6,400
Drowning	4,600
Fire	4,200
Ingestion of Food/Object	2,900
Firearms	1,400

(Source: US Dept. of Transportation)

Continued.

Pie Chart

To create a pie chart for the data, find the relative frequency (percent) of each category.

Type	Frequency	Relative Frequency
Motor Vehicle	43,500	0.578
Falls	12,200	0.162
Poison	6,400	0.085
Drowning	4,600	0.061
Fire	4,200	0.056
Ingestion of Food/Object	2,900	0.039
Firearms	1,400	0.019

$n = 75{,}200$

Continued.

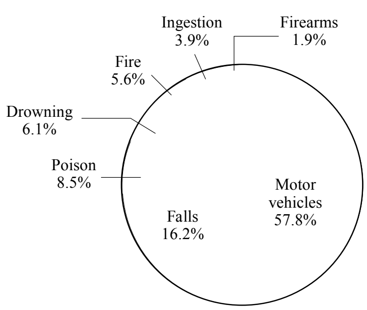
Pie Chart

Next, find the central angle. To find the central angle, multiply the relative frequency by 360°.

Type	Frequency	Relative Frequency	Angle
Motor Vehicle	43,500	0.578	208.2°
Falls	12,200	0.162	58.4°
Poison	6,400	0.085	30.6°
Drowning	4,600	0.061	22.0°
Fire	4,200	0.056	20.1°
Ingestion of Food/Object	2,900	0.039	13.9°
Firearms	1,400	0.019	6.7°

Continued.

Pie Chart



Pareto Chart

A **Pareto chart** is a vertical bar graph is 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.

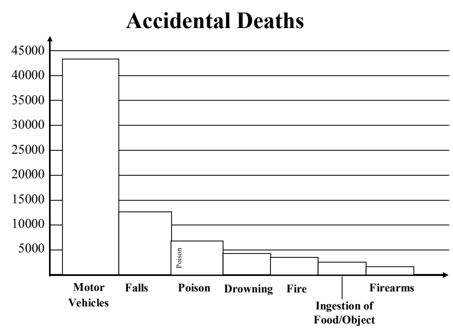
Accidental Deaths in the USA in 2002

Type	Frequency
Motor Vehicle	43,500
Falls	12,200
Poison	6,400
Drowning	4,600
Fire	4,200
Ingestion of Food/Object	2,900
Firearms	1,400

(Source: US Dept. of Transportation)

Continued.

Pareto Chart



Scatter Plot

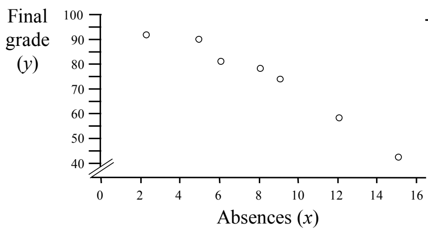
When each entry in one data set corresponds to an entry in another data set, the sets are called **paired data sets**.

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.

The following scatter plot represents the relationship between the number of absences from a class during the semester and the final grade.

Continued.

Scatter Plot



Absences x	Grade y
8	78
2	92
5	90
12	58
15	43
9	74
6	81

From the scatter plot, you can see that as the number of absences increases, the final grade tends to decrease.

Times Series Chart

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.

Example:
The following table lists the number of minutes Robert used on his cell phone for the last six months.

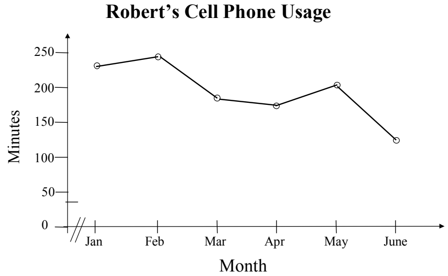
Month	Minutes
January	236
February	242
March	188
April	175
May	199
June	135

Construct a time series chart for the number of minutes used.

Continued.

Times Series Chart

Robert's Cell Phone Usage



§ 2.3

Measures of Central Tendency

Mean

A measure of central tendency is a value that represents a typical, or central, entry of a data set. The three most commonly used measures of central tendency are the mean, the median, and the mode.

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

Population mean: $\mu = \frac{\sum x}{N}$ ("mu")   Sample mean: $\bar{x} = \frac{\sum x}{n}$ ("*x*-bar")

Mean

Example:
The following are the ages of all seven employees of a small company:

53 32 61 57 39 44 57

Calculate the population mean.

$$\mu = \frac{\sum x}{N} = \frac{343}{7} = 49 \text{ years}$$

Add the ages and divide by 7.

The mean age of the employees is 49 years.

Median

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.

Example:
Calculate the median age of the seven employees.

53 32 61 57 39 44 57

To find the median, sort the data.

32 39 44 [53] 57 57 61

The median age of the employees is 53 years.

Mode

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**.

Example:
Find the mode of the ages of the seven employees.

53 32 61 [57] 39 44 [57]

The mode is 57 because it occurs the most times.

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

Comparing the Mean, Median and Mode

Example:
A 29-year-old employee joins the company and the ages of the employees are now:

53 32 61 57 39 44 57 29

Recalculate the mean, the median, and the mode. Which measure of central tendency was affected when this new age was added?

Mean = 46.5 — The mean takes every value into account, but is affected by the outlier.

Median = 48.5

Mode = 57 — The median and mode are not influenced by extreme values.

Weighted Mean

A **weighted mean** is the mean of a data set whose entries have varying weights. A weighted mean is given by

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

where w is the weight of each entry x.

Example:
Grades in a statistics class are weighted as follows:
Tests are worth 50% of the grade, homework is worth 30% of the grade and the final is worth 20% of the grade. A student receives a total of 80 points on tests, 100 points on homework, and 85 points on his final. What is his current grade?

Continued.

Weighted Mean

Begin by organizing the data in a table.

Source	Score, *x*	Weight, *w*	*xw*
Tests	80	0.50	40
Homework	100	0.30	30
Final	85	0.20	17

$$\bar{x} = \frac{\Sigma(x \cdot w)}{\Sigma w} = \frac{87}{100} = 0.87$$

The student's current grade is 87%.

Mean of a Frequency Distribution

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

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

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

Example:
The following frequency distribution represents the ages of 30 students in a statistics class. Find the mean of the frequency distribution.

Continued.

Mean of a Frequency Distribution

Class midpoint → x

Class	x	f	$(x \cdot f)$
18 – 25	21.5	13	279.5
26 – 33	29.5	8	236.0
34 – 41	37.5	4	150.0
42 – 49	45.5	3	136.5
50 – 57	53.5	2	107.0
		n = 30	Σ = 909.0

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{909}{30} = 30.3$$

The mean age of the students is 30.3 years.

Shapes of Distributions

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.

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

A frequency distribution is skewed if the "tail" of the graph 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.

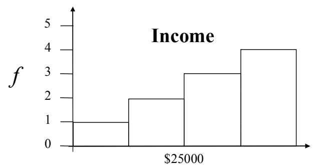
Symmetric Distribution

10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000

mean = median = mode = $25,000

Skewed Left Distribution

10 Annual Incomes
0
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000



mean = $23,500
median = mode = $25,000

Mean < Median

Skewed Right Distribution

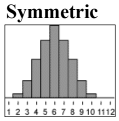
10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
1,000,000

mean = $121,500
median = mode = $25,000

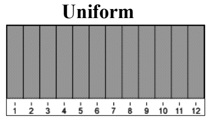
Mean > Median

Summary of Shapes of Distributions

Symmetric

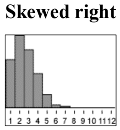


Uniform



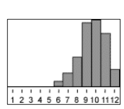
Mean = Median

Skewed right



Mean > Median

Skewed left



Mean < Median

§ 2.4

Measures of Variation

Range

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

Example:
The following data are the closing prices for a certain stock on ten successive Fridays. Find the range.

Stock	56	56	57	58	61	63	63	67	67	67

The range is 67 – 56 = 11.

Deviation

The **deviation** of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

Deviation of $x = x - \mu$

Example:
The following data are the closing prices for a certain stock on five successive Fridays. Find the deviation of each price.

The mean stock price is
$\mu = 305/5 = 61$.

Stock x	Deviation $x - \mu$
56	56 – 61 = – 5
58	58 – 61 = – 3
61	61 – 61 = 0
63	63 – 61 = 2
67	67 – 61 = 6
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$

Variance and Standard Deviation

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

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x-\mu)^2}{N}.$$

"sigma squared"

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

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}.$$

"sigma"

Finding the Population Standard Deviation

Guidelines

In Words	***In Symbols***
1. Find the mean of the population data set.	$\mu = \frac{\Sigma x}{N}$
2. Find the deviation of each entry.	$x - \mu$
3. Square each deviation.	$(x-\mu)^2$
4. Add to get the **sum of squares**.	$SS_x = \Sigma(x-\mu)^2$
5. Divide by N to get the **population variance**.	$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$
6. Find the square root of the variance to get the **population standard deviation**.	$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Finding the Sample Standard Deviation

Guidelines

In Words	***In Symbols***
1. Find the mean of the sample data set.	$\bar{x} = \frac{\Sigma x}{n}$
2. Find the deviation of each entry.	$x - \bar{x}$
3. Square each deviation.	$(x-\bar{x})^2$
4. Add to get the **sum of squares**.	$SS_x = \Sigma(x-\bar{x})^2$
5. Divide by *n – 1* to get the **sample variance**.	$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
6. Find the square root of the variance to get the **sample standard deviation**.	$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

Finding the Population Standard Deviation

Example:
The following data are the closing prices for a certain stock on five successive Fridays. The population mean is 61. Find the population standard deviation.

Always positive!

Stock x	Deviation $x - \mu$	Squared $(x - \mu)^2$
56	– 5	25
58	– 3	9
61	0	0
63	2	4
67	6	36
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$	$\Sigma(x - \mu)^2 = 74$

$SS_2 = \Sigma(x - \mu)^2 = 74$

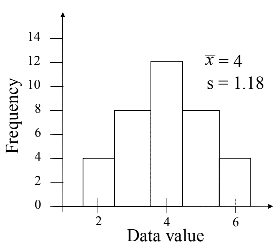
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{74}{5} = 14.8$$

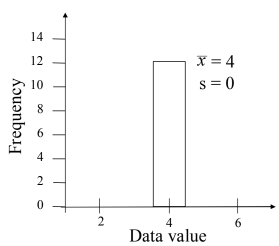
$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{14.8} \approx 3.85$$

$\sigma \approx \$3.85$

Interpreting Standard Deviation

When interpreting standard deviation, remember that is a measure of the typical amount an entry deviates from the mean. The more the entries are spread out, the greater the standard deviation.



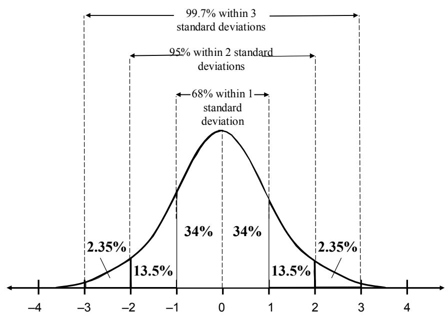


Empirical Rule (68-95-99.7%)

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.

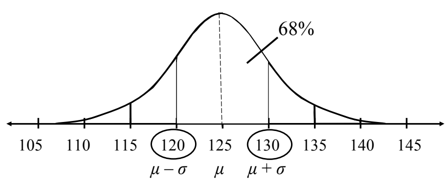
Empirical Rule (68-95-99.7%)



Using the Empirical Rule

Example:

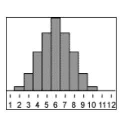
The mean value of homes on a street is \$125 thousand with a standard deviation of \$5 thousand. The data set has a bell shaped distribution. Estimate the percent of homes between \$120 and \$130 thousand.



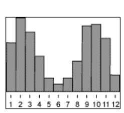
68% of the houses have a value between \$120 and \$130 thousand.

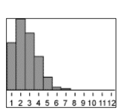
Chebychev's Theorem

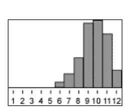
The Empirical Rule is only used for **symmetric distributions.**



Chebychev's Theorem can be used for **any distribution**, regardless of the shape.







Chebychev's Theorem

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

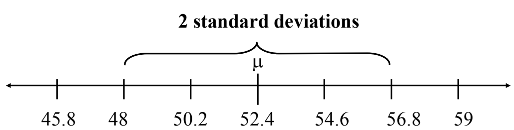
$$1 - \frac{1}{k^2}.$$

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.

Using Chebychev's Theorem

Example:
The mean time in a women's 400-meter dash is 52.4 seconds with a standard deviation of 2.2 sec. At least 75% of the women's times will fall between what two values?



At least 75% of the women's 400-meter dash times will fall between 48 and 56.8 seconds.

Standard Deviation for Grouped Data

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

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

Example:
The following frequency distribution represents the ages of 30 students in a statistics class. The mean age of the students is 30.3 years. Find the standard deviation of the frequency distribution.

Continued.

Standard Deviation for Grouped Data

The mean age of the students is 30.3 years.

Class	x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
18 – 25	21.5	13	– 8.8	77.44	1006.72
26 – 33	29.5	8	– 0.8	0.64	5.12
34 – 41	37.5	4	7.2	51.84	207.36
42 – 49	45.5	3	15.2	231.04	693.12
50 – 57	53.5	2	23.2	538.24	1076.48
		$n = 30$			$\Sigma = 2988.80$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 f}{n-1}} = \sqrt{\frac{2988.8}{29}} = \sqrt{103.06} = 10.2$$

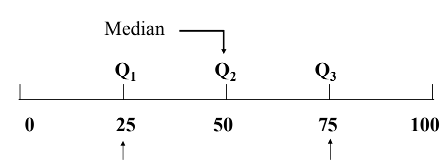
The standard deviation of the ages is 10.2 years.

§ 2.5

Measures of Position

Quartiles

The three **quartiles**, Q_1, Q_2, and Q_3, approximately divide an ordered data set into four equal parts.



Q_1 is the median of the data below Q_2.

Q_3 is the median of the data above Q_2.

Finding Quartiles

Example:
The quiz scores for 15 students is listed below. Find the first, second and third quartiles of the scores.
28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

Order the data.

Lower half / Upper half

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Q_1 = 37, Q_2 = 43, Q_3 = 48

About one fourth of the students scores 37 or less; about one half score 43 or less; and about three fourths score 48 or less.

Interquartile Range

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

Interquartile range (IQR) = $Q_3 - Q_1$.

Example:
The quartiles for 15 quiz scores are listed below. Find the interquartile range.

$Q_1 = 37$ $Q_2 = 43$ $Q_3 = 48$

(IQR) = $Q_3 - Q_1$
= 48 – 37
= 11

The quiz scores in the middle portion of the data set vary by at most 11 points.

Box and Whisker Plot

A **box-and-whisker plot** is an exploratory data analysis tool that highlights the important features of a data set.

The **five-number summary** is used to draw the graph.
- The minimum entry
- Q_1
- Q_2 (median)
- Q_3
- The maximum entry

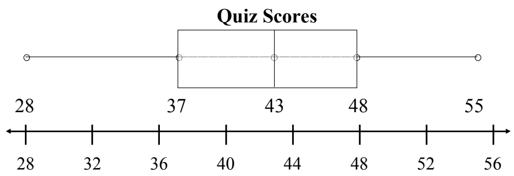
Example:
Use the data from the 15 quiz scores to draw a box-and-whisker plot.
28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Continued.

Box and Whisker Plot

Five-number summary

- The minimum entry 28
- Q_1 37
- Q_2 (median) 43
- Q_3 48
- The maximum entry 55



Percentiles and Deciles

Fractiles are numbers that partition, or divide, an ordered data set.

Percentiles divide an ordered data set into 100 parts. There are 99 percentiles: P_1, P_2, $P_3 \ldots P_{99}$.

Deciles divide an ordered data set into 10 parts. There are 9 deciles: D_1, D_2, $D_3 \ldots D_9$.

A test score at the 80th percentile (P_{80}), indicates that the test score is greater than 80% of all other test scores and less than or equal to 20% of the scores.

Standard Scores

The **standard score** or ***z*-score**, represents the number of standard deviations that a data value, *x*, falls from the mean, μ.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

Example:
The test scores for all statistics finals at Union College have a mean of 78 and standard deviation of 7. Find the *z*-score for

a.) a test score of 85,

b.) a test score of 70,

c.) a test score of 78.

Continued.

Standard Scores

Example continued:

a.) $\mu = 78,\ \sigma = 7,\ x = 85$

$$z = \frac{x - \mu}{\sigma} = \frac{85 - 78}{7} = 1.0$$

This score is 1 standard deviation higher than the mean.

b.) $\mu = 78,\ \sigma = 7,\ x = 70$

$$z = \frac{x - \mu}{\sigma} = \frac{70 - 78}{7} = -1.14$$

This score is 1.14 standard deviations lower than the mean.

c.) $\mu = 78,\ \sigma = 7,\ x = 78$

$$z = \frac{x - \mu}{\sigma} = \frac{78 - 78}{7} = 0$$

This score is the same as the mean.

Relative Z-Scores

Example:
John received a 75 on a test whose class mean was 73.2 with a standard deviation of 4.5. Samantha received a 68.6 on a test whose class mean was 65 with a standard deviation of 3.9. Which student had the better test score?

John's *z*-score

$$z = \frac{x - \mu}{\sigma} = \frac{75 - 73.2}{4.5} = 0.4$$

Samantha's *z*-score

$$z = \frac{x - \mu}{\sigma} = \frac{68.6 - 65}{3.9} = 0.92$$

John's score was 0.4 standard deviations higher than the mean, while Samantha's score was 0.92 standard deviations higher than the mean. Samantha's test score was better than John's.