Frequency Distributions and Their Graphs

The raw data

The following represents the ages of the 50 richest people in the world in 2009.

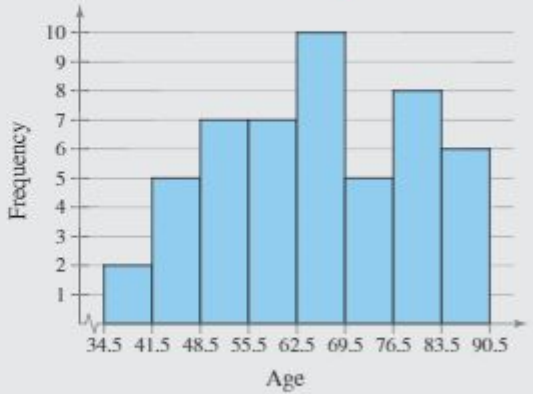
89, 89, 87, 86, 86, 85, 83, 83, 82, 81, 80, 78, 78, 77,76, 73, 73, 73, 72, 69, 69, 68, 67, 66, 66, 65, 65, 64,63, 61, 61, 60, 59, 58, 57, 56, 54, 54, 53, 53, 51, 51,49, 47, 46, 44, 43, 42, 36, 35

Where are we going to do?

Make a frequency distribution table.

Class	Frequency, f
35–41	2
42–48	5
49–55	7
56–62	7
63–69	10
70–76	5
77–83	8
84–90	6

Draw a histogram.



Average = 65.26 years old

Range = 54 years

2.1 Frequency Distributions and Their Graphs

DEFINITION

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

Lower class limit - the least number that can belong to the class.
Upper class limit - the greatest number that can belong to the class.
Class width - the distance between lower (or upper) limits of consecutive classes.
Range - the difference between the **maximum** and **minimum** data entries

DEFINITION

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.

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

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.

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

The **cumulative frequency** of a class is the sum of the frequencies of that class and all previous classes. The cumulative frequency of the last class is equal to the sample size n.

Constructing a Frequency Distribution

1. Decide on the **number of classes** to include in the frequency distribution.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** data 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 limit** of the first class. Remember that classes cannot overlap. Find the remaining 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.

EXAMPLE 1

Construct a frequency distribution using the ages of the 50 richest people dataset listed in the Chapter Opener. Use eight classes.

a. the number of classes.
b. Find the minimum and maximum values and the class width.
c. Find the class limits.
d. Tally the data entries.
e. Write the frequency f of each class.

Frequency Distribution the ages of the 50 richest people in 2009

Class	Tally	Frequency, f	Midpoint	Relative frequency	Cumulative frequency

GRAPHS OF FREQUENCY DISTRIBUTIONS

DEFINITION

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

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.

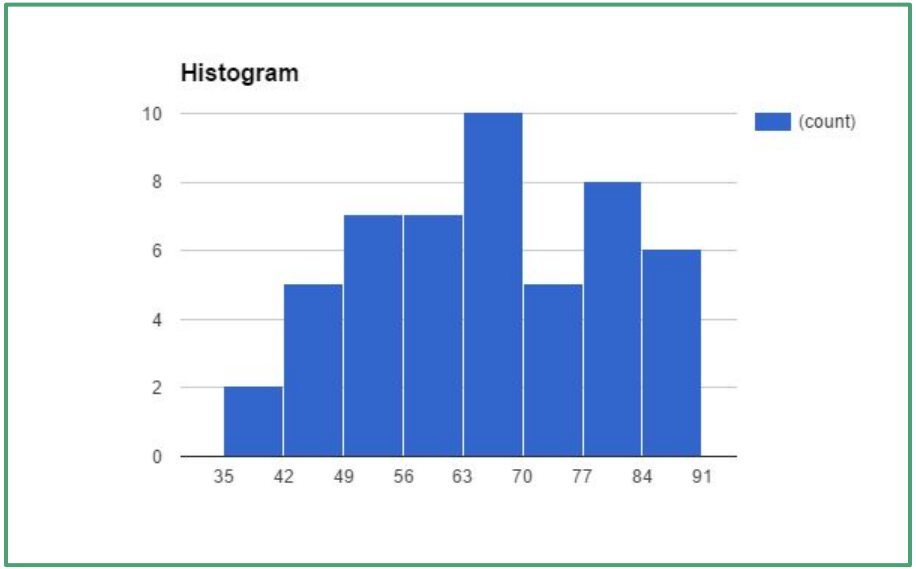
Class boundaries are the numbers that separate classes without forming gaps between them. If data entries are integers, subtract 0.5 from each lower limit to find the lower class boundaries. To find the upper class boundaries, add 0.5 to each upper limit. The upper boundary of a class will equal the lower boundary of the next higher class.

EXAMPLE 2

Use the frequency distribution from EXAMPLE 1 to construct a frequency histogram that represents the ages of the 50 richest people. Describe any patterns.

a. Find the class boundaries.
b. Choose appropriate horizontal and vertical scales.
c. Use the frequency distribution to find the height of each bar.
d. Describe any patterns in the data.

EXAMPLE



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

EXAMPLE 3

Draw a relative frequency histogram for the frequency distribution in Example 2

DEFINITION

A **cumulative frequency graph**,or **ogive**, is a line graph that displays the cumulative frequency of each class at its upper class boundary. The upper boundaries are marked on the horizontal axis, and the cumulative frequencies are marked on the vertical axis.

Constructing an Ogive (Cumulative Frequency Graph)

1. Construct a frequency distribution that includes cumulative frequencies as one of the columns.
2. Specify the horizontal and vertical scales. The horizontal scale consists of upper class boundaries, and the vertical scale measures cumulative frequencies.
3. Plot points that represent the upper class boundaries and their corresponding cumulative frequencies.
4. Connect the points in order from left to right.
5. The graph should start at the lower boundary of the first class (cumulative frequency is zero) and should end at the upper boundary of the last class (cumulative frequency is equal to the sample size).

EXAMPLE 4

Construct an ogive that represents the ages of the 50 richest people. Estimate the number of people who are 80 years old or younger.

a. Specify the horizontal and vertical scales.
b. Plot the points given by the upper class boundaries and the cumulative frequencies.
c. Construct the graph.
d. Estimate the number of people who are 80 years old or younger.
e. Interpret the results in the context of the data.

2.2 Measures of Central Tendency

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**.

DEFINITION

The mean of a finite data set is the sum of the data entries divided by the number of entries. To find the mean of a data set, use one of the following formulas.

Population Mean: $\mu = \frac{\sum x}{N}$ Sample Mean: $\overline{x} = \frac{\sum x}{n}$

The lowercase Greek letter μ (pronounced mu) represents the population mean and $\overline{x}$ (read as “xbar”) represents the sample mean. Note that N represents the number of entries in a population and n represents the number of entries in a sample. Recall that the uppercase Greek letter sigma indicates a summation of values.

EXAMPLE 1

The heights (in inches) of the players on the 2009–2010 Cleveland Cavaliers basketball team are listed.

74 78 81 87 81 80 77 8085 78 80 83 75 81 73

What is the mean height?

a. Find the sum of the data entries.
b. Divide the sum by the number of data entries.
c. Interpret the results in the context of the data.

DEFINITION

The **median** of a data set is the value that lies in the middle of the data when the data set is ordered. The median measures the center of an ordered data set by dividing it into two equal parts. 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.

In a data set, there are the same number of data values above the median as there are below the median.

EXAMPLE 2

The ages of a sample of fans at a rock concert are listed. Find the median age.

24 27 19 21 18 23 21 20 19 33 30 29 2118 24 26 38 19
35 34 33 30 21 27 30

a. Order the data entries.
b. Find the middle data entry.
c. Interpret the results in the context of the data.

DEFINITION

The **mode** of a data set is the data entry that occurs with the greatest frequency. A data set can have one mode, more than one mode, or no mode. 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 mode is the only measure of central tendency that can be used to describe data at the nominal level of measurement. But when working with quantitative data, the mode is rarely used.

EXAMPLE 3

In a survey, 1000 U.S. adults were asked if they thought public cellular phone conversations were rude. Of those surveyed, 510 responded “Yes,” 370 responded “No,” and 120 responded “Not sure.”

What is the mode of the responses?

a. Identify the entry that occurs with the greatest frequency.
b. Interpret the results in the context of the data.

WEIGHTED MEAN

Sometimes data sets contain entries that have a greater effect on the mean than do other entries. To find the mean of such a data set, you must find the weighted mean.

DEFINITION

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

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

where w is the weight of each entry x.

EXAMPLE 4

You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework).

What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?

MEAN OF GROUPED DATA

If data are presented in a frequency distribution, you can approximate the mean as follows.

DEFINITION

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

$$\overline{x} = \frac{\sum(f \cdot x)}{n} = \frac{\sum(f \cdot x)}{\sum f}$$

where x and f are the midpoints and frequencies of a class, respectively.

EXAMPLE 5

Use a frequency distribution to approximate the mean age of the 50 richest people.

a. Find the midpoint of each class.
b. Find the sum of the products of each midpoint and corresponding frequency.
c. Find the sum of the frequencies.
d. Find the mean of the frequency distribution.

Frequency Distribution the ages of the 50 richest people in 2009

Class	Frequency, f	Midpoint	fx

MEDIAN OF GROUPED DATA

The median of a frequency distribution for a sample is approximated by

Step 1: Construct the cumulative frequency distribution.

Step 2: Decide the class that contain the median.

> **Class Median** is the first class with the value of cumulative frequency equal at least n/2.

Step 3: Find the median by using the following formula:

$$Median = L_{me} + \left(\frac{\frac{n}{2} - F}{f_{me}} \right) I$$

where

L_m = the lower boundary of the class median

F = the cumulative frequency before class median

f_m = the frequency of the class median

n = the total frequency

i = the class width

EXAMPLE 6

Use a frequency distribution to approximate the median age of the 50 richest people.

Class	Frequency, f	

MODE OF GROUPED DATA

The mode of a frequency distribution for a sample is approximated by

Step 1: Identify the modal class that contains highest frequency.

Step 3: Find the mode by using the following formula:

$$Mode = L_{mo} + \left(\frac{y_1 - y_0}{(y_1 - y_0) + (y_1 - y_2)} \right) I$$

Where L_{mo} = the lower boundary of the modal class

y_0, y_1, y_2 = the frequencies of three consecutive classes, where y_1 is the frequency of the modal class

i = the class width

EXAMPLE 7

Use a frequency distribution to approximate the mode of age of the 50 richest people.

Class	Frequency, f	

2.3 Measures of Variation

RANGE

DEFINITION

The range of a data set is the difference between the maximum and minimum data entries in the set. To find the range, the data must be quantitative.

Range = Maximum data entry - Minimum data entry

VARIANCE AND STANDARD DEVIATION

DEFINITION

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

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

The symbol σ is the lowercase Greek letter sigma

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

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

VARIANCE AND STANDARD DEVIATION

DEFINITION

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

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

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

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

Symbols in Variance and Standard Deviation Formulas

	Population	Sample
Variance	σ^2	s^2
Standard deviation	σ	s
Mean	μ	$\bar{x}$
Number of entries	N	n
Deviation	$x - \mu$	$x - \bar{x}$
Sum of squares	$\Sigma(x - \mu)^2$	$\Sigma(x - \bar{x})^2$

GUIDELINES

Finding the Population Variance and Standard Deviation

IN WORDS	IN SYMBOLS
1. Find the mean of the population data set.	$\mu = \dfrac{\Sigma x}{N}$
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.**	$\sigma^2 = \dfrac{\Sigma (x - \mu)^2}{N}$
6. Find the square root of the variance to get the **population standard deviation.**	$\sigma = \sqrt{\dfrac{\Sigma (x - \mu)^2}{N}}$

EXAMPLE 1

Two corporations each hired 10 graduates. The starting salaries for each graduate are shown.

Starting Salaries for Corporation A (1000s of dollars)

41	38	39	45	47	41	41	44	37	42

Starting Salaries for Corporation B (1000s of dollars)

40	23	41	50	49	32	41	29	52	58

a. Find the range of the starting salaries for Corporation A.
b. Find the population standard deviation of the starting salaries for Corporation A.
c. Find the sample variance of the starting salaries for the Chicago branch of Corporation B.

STANDARD DEVIATION FOR GROUPED DATA

The formula for the sample standard deviation for a frequency distribution is

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

where n is the number of entries in the data set.

EXAMPLE 2

Find the sample standard deviation from the following frequency distribution.

Class	x	f				
0 - 99		380				
100 - 199		230				
200 - 299		210				
300 - 399		50				
400 - 499		60				
500 - 599		70				

2.4 Measures of Position

QUARTILES

DEFINITION

The three **quartiles**, Q_1 , Q_2 and Q_3 approximately divide an ordered data set into four equal parts. About one quarter of the data fall on or below the first quartile. About one half of the data fall on or below the second quartile(the second quartile is the same as the median of the data set). About three quarters of the data fall on or below the third quartile Q_3.

1. Sort data to increasing order.
2. Find the position of the quartile

   $$\text{Position of } Q_i = i(N+1)/4$$

3. Find the quartile in the sorted data set, the quartile can be the average value between two data points.

EXAMPLE 1

The number of nuclear power plants in the top 15 nuclear power-producing countries in the world are listed. Find the first, second, and third quartiles of the data set. What can you conclude?

7 18 11 6 59 17 18 54 104 20 31 8 10 15 19

a. Order the data set.
b. Find the median.
c. Find the first and third quartiles.
d. Interpret the results in the context of the data.

DEFINITION

The interquartile range (IQR) of a data set is a measure of variation that gives the range of the middle 50% of the data. It is the difference between the third and first quartiles.

Interquartile range (IQR) = $Q_3 - Q_1$

EXAMPLE 2

Find the interquartile range of the data set given in Example 1. What can you conclude from the result?

PERCENTILES AND DECILES

In addition to using quartiles to specify a measure of position, you can also use percentiles and deciles.

Fractiles	**Summary**	**Symbols**
Quartiles	Divide a data set into 4 equal parts	Q_1, Q_2, Q_3
Deciles	Divide a data set into 10 equal parts.	$D_1, D_2, D_3, \ldots, D_9$
Percentiles	Divide a data set into 100 equal parts.	$P_1, P_2, P_3, \ldots, P_{99}$

FRACTILES of GROUPED DATA

1. Find the cumulative frequencies of the frequency distribution.
2. Find the position of the quartile (or decile, percentile)

   Position of $Q_i = i(n)/4$
3. Find the quartile in the class that the cumulative frequency is just greater or equal to the position of Q_i
   - 3.1. If there is a cumulative frequency that is equal to the position of Q_i, then the Q_i equal to the upper boundary of the class.
   - 3.2. If there is no cumulative frequency that is equal to the position of Q_i, the Q_i is in the class that the cumulative frequency is just greater than the position of Q_i. Calculate the Q_i by proportion (the rule of three) or use the formula

$$Q_i = L_q + \left(\frac{\frac{i(n)}{4} - F}{f_q} \right) I$$

EXAMPLE 3

Find the Q_3, D_2, P_{80} from the following frequency distribution.

Class	f				
0 - 99	75				
100 - 199	80				
200 - 299	95				
300 - 399	50				
400 - 499	45				
500 - 599	55				

THE COEFFICIENT OF VARIATION

DEFINITION

The coefficient of variation (C.V.) is a measure of spread that describes the amount of variability relative to the mean.

Coefficient of variation CV = σ/μ

The coefficient of variation is unitless, you can use it instead of the standard deviation to compare the spread of data sets that have different units or different means.

THE STANDARD SCORE

When you know the mean and standard deviation of a data set, you can measure a data value's position in the data set with a standard score, or z-score.

DEFINITION

The **standard score**,or **z-score**, represents the number of standard deviations a given value x falls from the mean μ. To find the z-score for a given value, use the following formula.

$$z = \frac{Value - Mean}{Standard\ deviation} = \frac{x-\mu}{\sigma}$$

EXAMPLE 4

In 2009, Sean Penn won the Oscar for Best Actor at age 48 for his role in the movie Milk. Kate Winslet won the Oscar for Best Actress at age 33 for her role in The Reader. The mean age of all Best Actor winners is 43.7, with a standard deviation of 8.7. The mean age of all Best Actress winners is 35.9, with a standard deviation of 11.4. Find the z-scores that correspond to the ages of Penn and Winslet. Then compare your results.

a. Identify μ and σ for each data set.
b. Transform each value to a z-score.
c. Compare your results.

EXAMPLE 5

The monthly utility bills in city A have a mean of $70 and a standard deviation of $8. While the monthly utility bills in city B have a mean of $60 and a standard deviation of $7.

a. Which city has higher variation in the monthly utility bills?
b. Find the z-scores that correspond to utility bills of $71 in the city A and $65, in the city B.
c. What can you conclude?