College of Engineering
Academic Accreditation Committee

CS

UNIVERSITY OF ANBAR
COLLEGE OF ENGINEERING

College of Engineering

COURSE SYLLABUS

Course Title:

ENGINEERING STATISTICS

Course Code: ME3201

2020-2021

Dr.Sattar Abed Mutlag

Lecture -1
Chapter 1

Introduction to Statistics An Overview of Statistics

Dr. Sattar Abed Mutlag

Data and Statistics

Data consists of information coming from observations, counts, measurements, or responses.

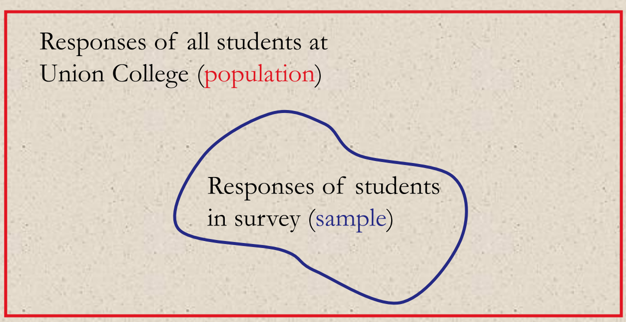
Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

A **population** is the collection of *all* outcomes, responses, measurement, or counts that are of interest.

A **sample** is a subset of a population.

Populations & Samples

Example:
In a recent survey, 250 college students at Union College were asked if they smoked cigarettes regularly. 35 of the students said yes. Identify the population and the sample.



Parameters & Statistics

A **parameter** is a numerical description of a *population* characteristic.

A **statistic** is a numerical description of a *sample* characteristic.

Parameter → **Population**

Statistic → **Sample**

Parameters & Statistics

Example:
Decide whether the numerical value describes a population parameter or a sample statistic.

a.) A recent survey of a sample of 450 college students reported that the average weekly income for students is $325.

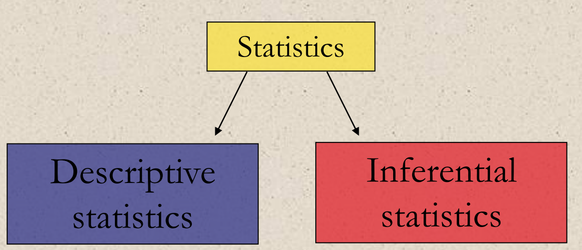
Because the average of $325 is based on a sample, this is a sample statistic.

b.) The average weekly income for all students is $405.

Because the average of $405 is based on a population, this is a population parameter.

Branches of Statistics

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics**.



Involves the organization, summarization, and display of data.

Involves using a sample to draw conclusions about a population.

Descriptive and Inferential Statistics

Example:

In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep. Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics.

The statement "four times more likely to answer incorrectly" is a descriptive statistic. An inference drawn from the sample is that all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.

§ 1.2

Data Classification

Types of Data

Data sets can consist of two types of data: **qualitative data** and **quantitative data**.

Data

Qualitative Data

Consists of attributes, labels, or nonnumerical entries.

Quantitative Data

Consists of numerical measurements or counts.

Qualitative and Quantitative Data

Example:

The grade point averages of five students are listed in the table. Which data are qualitative data and which are quantitative data?

Student	GPA
Sally	3.22
Bob	3.98
Cindy	2.75
Mark	2.24
Kathy	3.84

Qualitative data

Quantitative data

Levels of Measurement

The level of measurement determines which statistical calculations are meaningful. The four levels of measurement are: **nominal**, **ordinal**, **interval**, and **ratio**.

Levels of Measurement

- **Nominal**
- **Ordinal**
- **Interval**
- **Ratio**

Lowest to highest

Nominal Level of Measurement

Data at the **nominal level of measurement** are qualitative only.

Levels of Measurement

Nominal

Calculated using names, labels, or qualities. No mathematical computations can be made at this level.

Colors in the US flag	Names of students in your class	Textbooks you are using this semester

Ordinal Level of Measurement

Data at the **ordinal level of measurement** are qualitative or quantitative.

Levels of Measurement → **Ordinal**

Arranged in order, but differences between data entries are not meaningful.

- Class standings: freshman, sophomore, junior, senior
- Numbers on the back of each player's shirt
- Top 50 songs played on the radio

Interval Level of Measurement

Data at the **interval level of measurement** are quantitative. A zero entry simply represents a position on a scale; the entry is not an inherent zero.

Levels of Measurement → **Interval**

Arranged in order, the differences between data entries can be calculated.

- Temperatures
- Years on a timeline
- Atlanta Braves World Series victories

Ratio Level of Measurement

Data at the **ratio level of measurement** are similar to the interval level, but a zero entry is meaningful.

Levels of Measurement → **Ratio**

A ratio of two data values can be formed so one data value can be expressed as a ratio.

Ages

Grade point averages

Weights

Summary of Levels of Measurement

Level of measurement	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Lcture-2

Experimental Design

Designing a Statistical Study

GUIDELINES

1. Identify the variable(s) of interest (the focus) and the population of the study.
2. Develop a detailed plan for collecting data. If you use a sample, make sure the sample is representative of the population.
3. Collect the data.
4. Describe the data.
5. Interpret the data and make decisions about the population using inferential statistics.
6. Identify any possible errors.

Methods of Data Collection

In an **observational study**, a researcher observes and measures characteristics of interest of part of a population.

In an **experiment**, a treatment is applied to part of a population, and responses are observed.

A **simulation** is the use of a mathematical or physical model to reproduce the conditions of a situation or process.

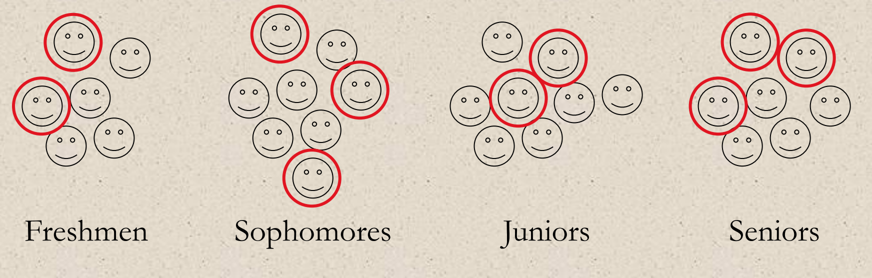
A **survey** is an investigation of one or more characteristics of a population.

→ A **census** is a measurement of an *entire* population.

→ A **sampling** is a measurement of *part* of a population.

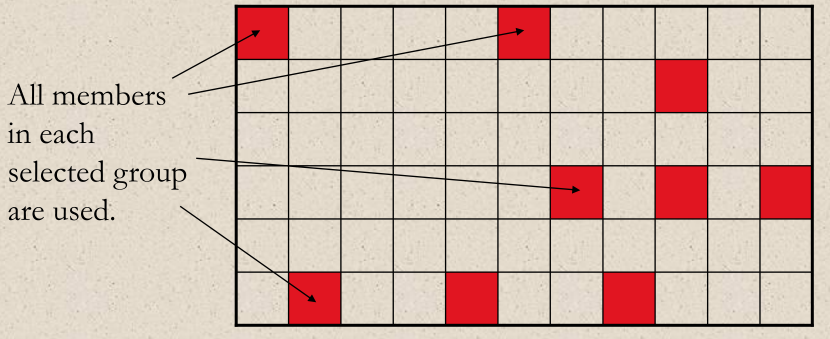
Stratified Samples

A **stratified sample** has members from each segment of a population. This ensures that each segment from the population is represented.



Cluster Samples

A **cluster sample** has all members from randomly selected segments of a population. This is used when the population falls into naturally occurring subgroups.



The city of Clarksville divided into city blocks.

Systematic Samples

A **systematic sample** is a sample in which each member of the population is assigned a number. A starting number is randomly selected and sample members are selected at regular intervals.

Every fourth member is chosen.

Convenience Samples

A **convenience sample** consists only of available members of the population.

Example:

You are doing a study to determine the number of years of education each teacher at your college has. Identify the sampling technique used if you select the samples listed.

1.) You randomly select two different departments and survey each teacher in those departments.

2.) You select only the teachers you currently have this semester.

3.) You divide the teachers up according to their department and then choose and survey some teachers in each department.

Continued.

Identifying the Sampling Technique

Example continued:

You are doing a study to determine the number of years of education each teacher at your college has. Identify the sampling technique used if you select the samples listed.

1.) This is a cluster sample because each department is a naturally occurring subdivision.

2.) This is a convenience sample because you are using the teachers that are readily available to you.

3.) This is a stratified sample because the teachers are divided by department and some from each department are randomly selected.

Lcture-3
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.

Lower Class Limits

Upper Class Limits

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

Frequencies

Frequency Distributions

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

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

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

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.

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:

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:

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

Ages of Students

Ages

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

Number of students

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
	$\sum f = 30$	

$18 + 25 = 43$
$43 \div 2 = 21.5$

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
	$\sum f = 30$	$\sum \frac{f}{n} = 1$

Portion of students

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

$$\approx 0.433$$

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
	$\sum f = 30$	

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.

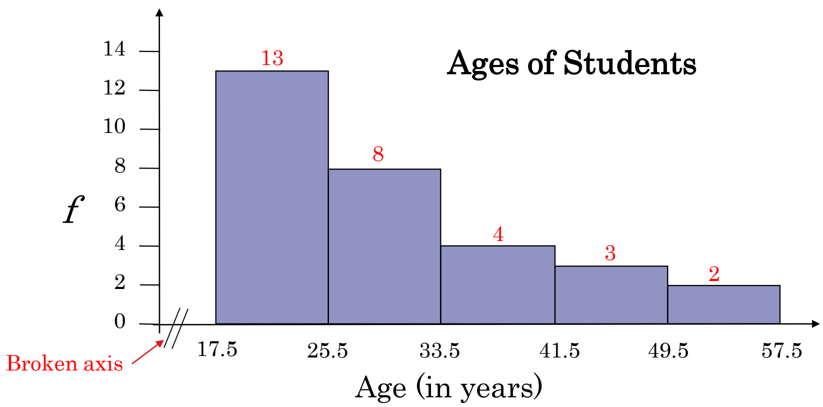
Half this distance is 0.5.

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.

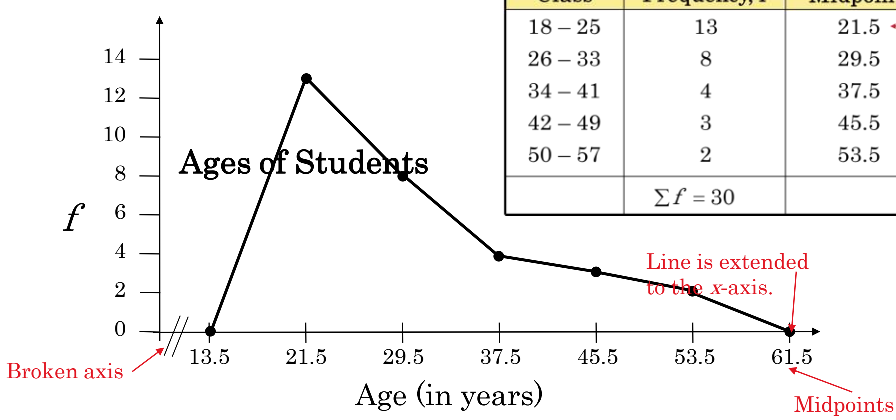


Frequency Polygon

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

Ages of Students

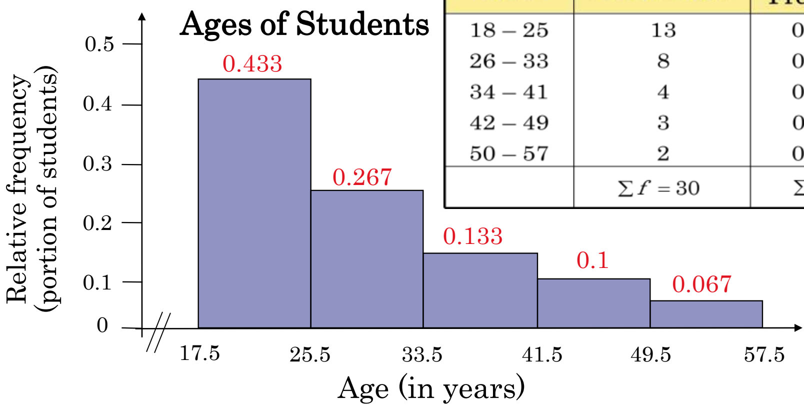
Class	Frequency, f	Midpoint
18 – 25	13	21.5
26 – 33	8	29.5
34 – 41	4	37.5
42 – 49	3	45.5
50 – 57	2	53.5
	$\sum f = 30$	



Relative Frequency Histogram

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

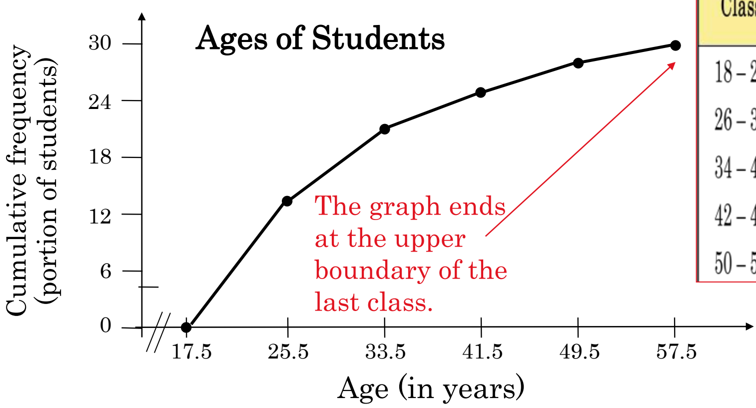
Class	Frequency, f	Relative Frequency
18 – 25	13	0.433
26 – 33	8	0.267
34 – 41	4	0.133
42 – 49	3	0.1
50 – 57	2	0.067
	$\Sigma f = 30$	$\Sigma \frac{f}{n} = 1$



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.

Class	Frequency, f
18 – 25	13
26 – 33	8
34 – 41	4
42 – 49	3
50 – 57	2



§ 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

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

Ages of Students

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.

Key: 1 | 8 = 18

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.


Consider the data in the table describing battery life for 40 similar car batteries in years

[illegible]	4.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.2	[illegible]

How many batteries have fewer than 3 years?

Dot Plot

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

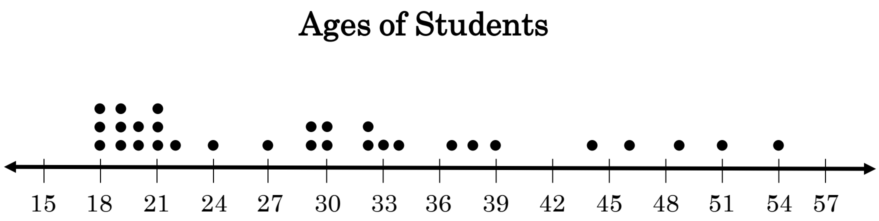
Example:
Use a dot plot to display the ages of the 30 students in the statistics class.

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



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.

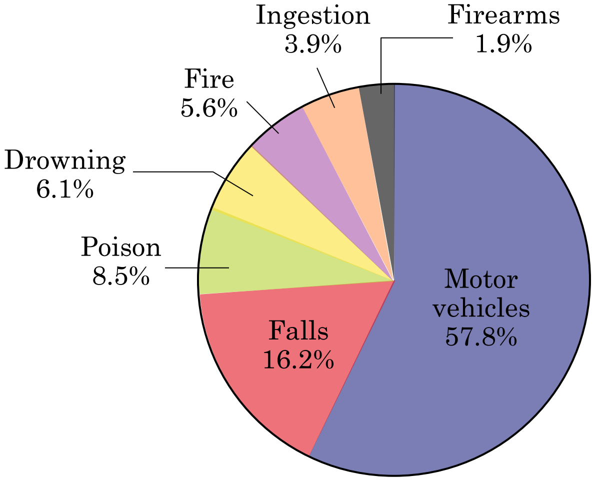
Pie Chart

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.

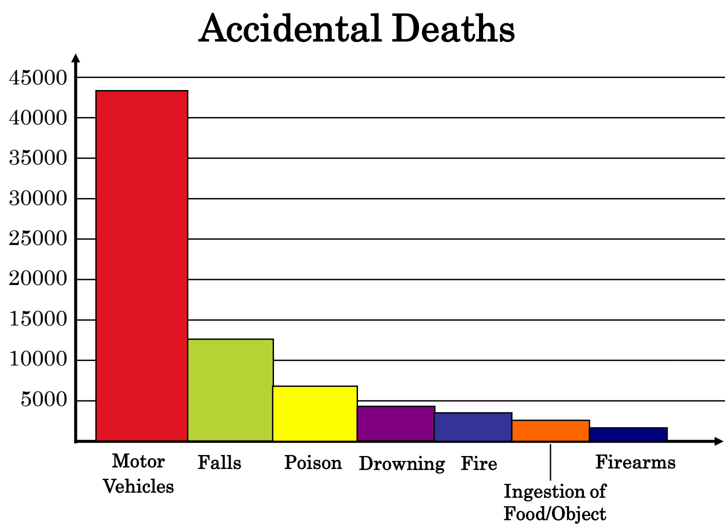
Accidental Deaths in the USA in 2002

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

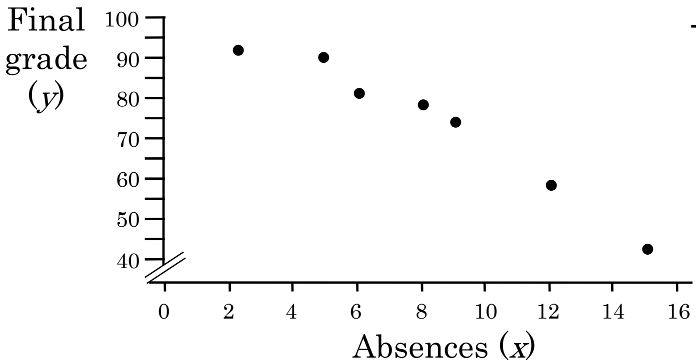
When each entry in one data set corresponds to an entry in another data set, the sets are called **paired data sets**.

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.

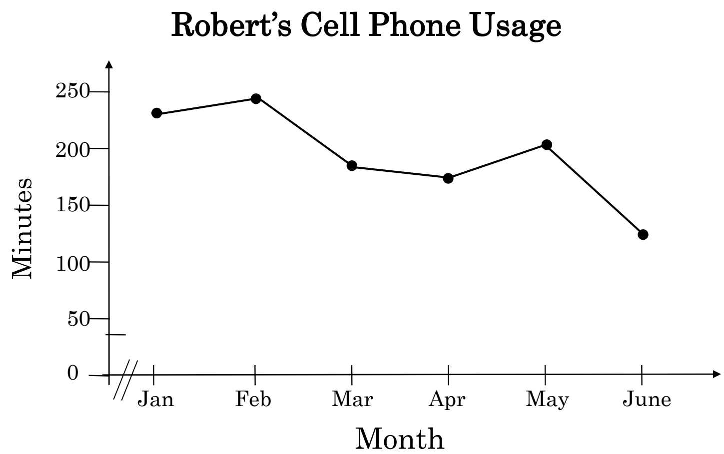
Example:
The following table lists the number of minutes Robert used on his cell phone for the last six months.

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

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

Continued.

Times Series Chart



Lcture-4

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: $\mu = \frac{\sum x}{N}$ Sample mean: $\bar{x} = \frac{\sum x}{n}$

"mu" "*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.

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

Add the ages and divide by 7.

$$= 49 \text{ years}$$

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 $\boxed{53}$ 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 $\boxed{57}$ 39 44 $\boxed{57}$

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

The median and mode are not influenced by extreme values.

Mode = 57

Weighted Mean

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

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

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

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

The student's current grade is 87%.

Mean of a Frequency Distribution

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

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

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

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
		$n = 30$	$\Sigma = 909.0$

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

The mean age of the students is 30.3 years.

Shapes of Distributions

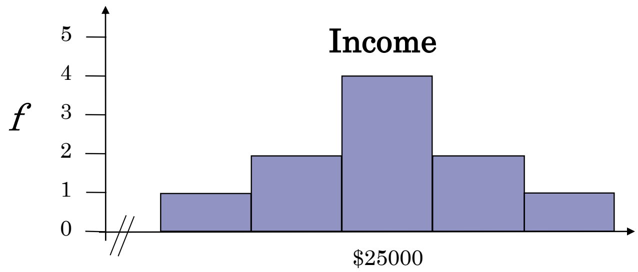
A frequency distribution is **symmetric** when a vertical line can be drawn through the middle of a graph of the distribution and the resulting halves are approximately the mirror images.

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

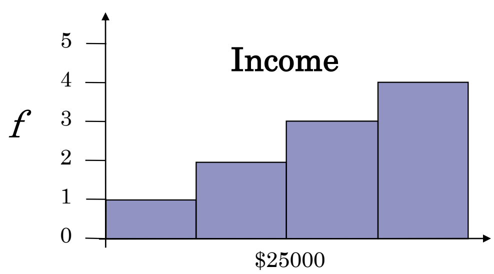
10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000



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

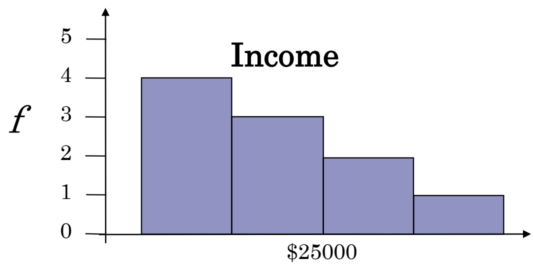


mean = $23,500
median = mode = $25,000

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

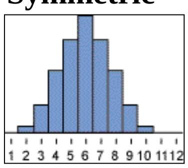


mean = $121,500
median = mode = $25,000

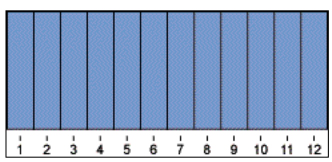
Mean > Median

Summary of Shapes of Distributions

Symmetric

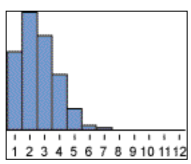


Uniform



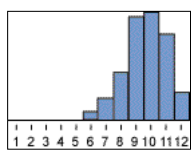
Mean = Median

Skewed right



Mean > Median

Skewed left



Mean < Median

Lcture-5

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

Finding the Sample Standard Deviation

Guidelines

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

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
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$	$\Sigma(x - \mu)^2 = 74$

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

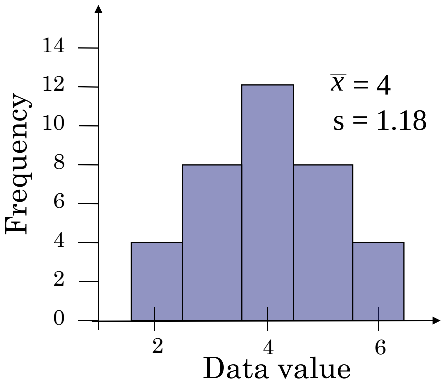
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{74}{5} = 14.8$$

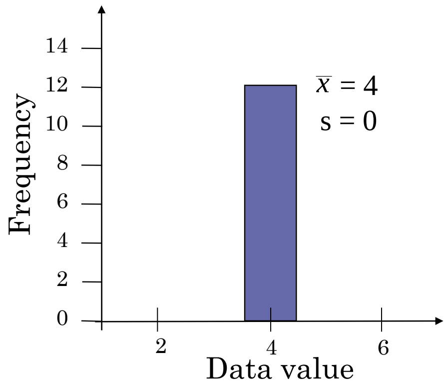
$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{14.8} \approx 3.85$$

$\sigma \approx \$3.85$

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.





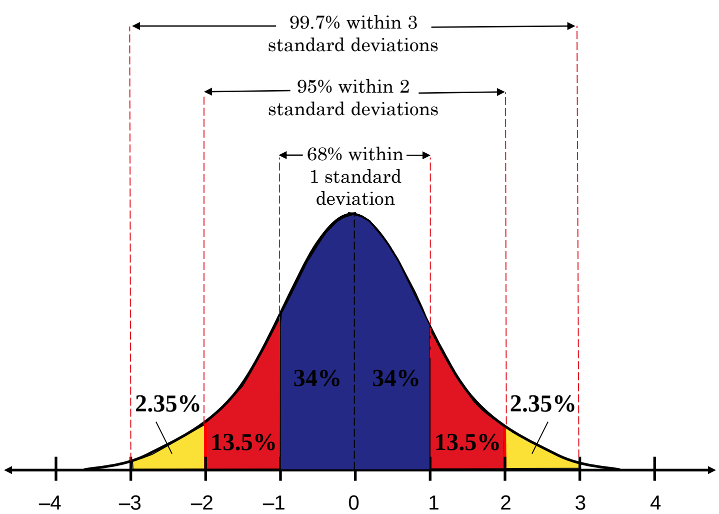
Empirical Rule (68-95-99.7%)

Empirical Rule

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

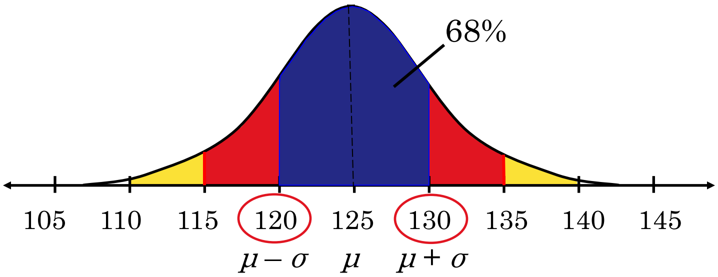
1. About 68% of the data lie within one standard deviation of the mean.
2. About 95% of the data lie within two standard deviations of the mean.
3. About 99.7% of the data lie within three standard deviation of the mean.

Empirical Rule (68-95-99.7%)



Using the Empirical Rule

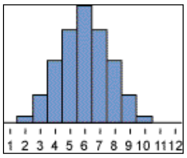
Example:
The mean value of homes on a street is \$125 thousand with a standard deviation of \$5 thousand. The data set has a bell shaped distribution. Estimate the percent of homes between \$120 and \$130 thousand.



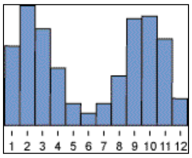
68% of the houses have a value between \$120 and \$130 thousand.

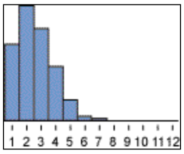
Chebychev's Theorem

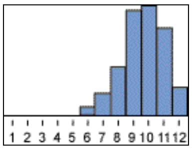
The Empirical Rule is only used for **symmetric distributions.**



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

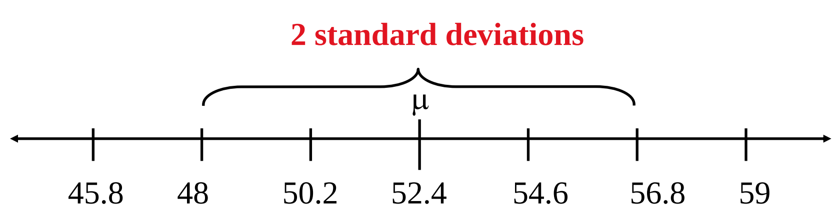
$$1 - \frac{1}{k^2}.$$

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 = $s = \sqrt{\dfrac{\Sigma(x-\bar{x})^2 f}{n-1}}$

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
		$n = 30$			$\Sigma = 2988.80$

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

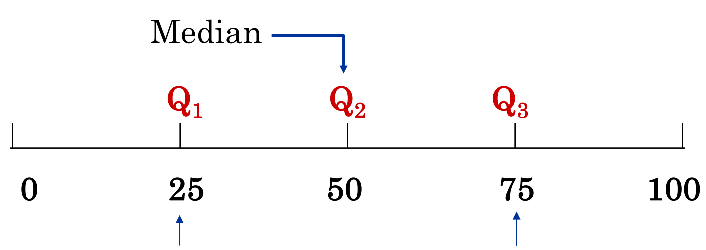
The standard deviation of the ages is 10.2 years.

§ 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.

$$\underbrace{28\ 30\ 33\ 37\ 37\ 38\ 42}_{\text{Lower half}}\ 43\ \underbrace{43\ 44\ 45\ 48\ 48\ 51\ 55}_{\text{Upper half}}$$

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

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.

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

$$\begin{aligned}(IQR) &= Q_3 - Q_1\\ &= 48 - 37\\ &= 11\end{aligned}$$

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.

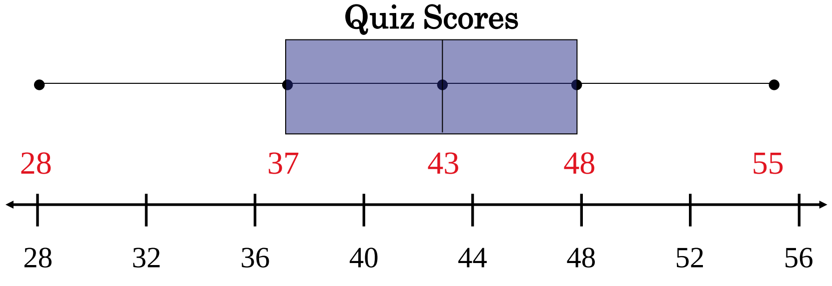
28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

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.

Lecture-6

Chapter 2
Summarizing and Graphing Data

2-1 Review and Preview

2-2 Frequency Distributions

2-3 Histograms

2-4 Statistical Graphics

2-5 Critical Thinking: Bad Graphs

Preview
Important Characteristics of Data

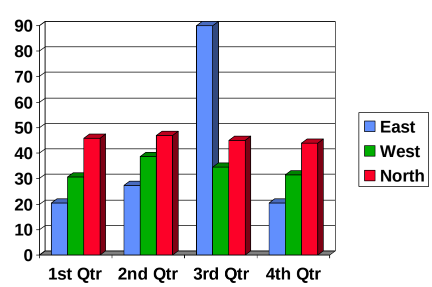
1. **Center**: A representative or average value that indicates where the middle of the data set is located.

2. **Variation**: A measure of the amount that the data values vary.

3. **Distribution**: The nature or shape of the spread of data over the range of values (such as bell-shaped, uniform, or skewed).

4. **Outliers**: Sample values that lie very far away from the vast majority of other sample values.

5. **Time**: Changing characteristics of the data over time.



Chapter 2
Summarizing and Graphing Data

2

Definition

❖ <u>**Frequency Distribution**</u>

shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories.

Heights (inches) of 25 Women

67	64	65	65	64
59	67	67	72	65
64	62	66	67	66
60	70	68	61	64
60	68	65	66	62

Frequency Distribution

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

The *frequency* for a particular class is the number of original values that fall into that class.

Frequency Distributions

Definitions

4

Lower Class Limits

are the smallest numbers that belong to the different classes

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

Lower class limits are red numbers.

Upper Class Limits

are the largest numbers that belong to the different classes

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

Upper class limits are red numbers.

Class Width

is the difference between two consecutive lower class limits or two consecutive upper class boundaries

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

Class width is 2 since: $61 - 59 = 2$

Class Boundaries

are the numbers midway between the numbers that separate classes

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

Class boundaries are:

58.5, 60.5, 62.5, 64.5, 66.5, 68.5, 70.5, 72.5

NOTE: the difference between consecutive class boundaries is equal to the class width and lowest/highest class boundaries are below/above lowest/highest class limits

Calculating Class Boundaries

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

$$\frac{60+61}{2}=60.5$$

Calculating Class Boundaries

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

$$\frac{62+63}{2}=62.5$$

Class Midpoints

are the values in the <u>middle</u> of the classes and can be found by adding the lower class limit to the upper class limit and dividing by 2 or averaging the lower class limit and the upper class limit

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

Class midpoints are:

59.5, 61.5, 63.5, 65.5, 67.5, 69.5, 71.5

Calculating Class Midpoints

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

$$\frac{59+60}{2}=59.5$$

Calculating Class Midpoints

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

$$\frac{61+62}{2}=61.5$$

Relative Frequency Distribution

includes the same class limits as a frequency distribution, but the frequency of a class is replaced with a relative frequencies (a proportion) or a percentage frequency (a percent)

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Relative Frequency Distribution

HEIGHT	RELATIVE FREQUENCY
59-60	3/25=0.12
61-62	3/25=0.12
63-64	4/25=0.16
65-66	7/25=0.28
67-68	6/25=0.24
69-70	1/25=0.04
71-72	1/25=0.04

Note: sum of relative frequencies is 1

Percent Frequency Distribution

HEIGHT	PERCENT FREQUENCY
59-60	12%
61-62	12%
63-64	16%
65-66	28%
67-68	24%
69-70	4%
71-72	4%

Note: sum of percent frequencies is 100%

Cumulative Frequency Distribution

is the sum of the frequency for that class and the previous all previous classes

HEIGHT	FREQUENCY
59-60	3
61-62	3
63-64	4
65-66	7
67-68	6
69-70	1
71-72	1

HEIGHT	CUMULATIVE FREQUENCY
59-60	3
61-62	6
63-64	10
65-66	17
67-68	23
69-70	24
71-72	25

Reasons for Constructing Frequency Distributions

1. Large data sets can be summarized.
2. We can analyze the nature of data.
3. We have a basis for constructing important graphs.

Constructing A Frequency Distribution

1. Determine the number of classes (should be between 5 and 20).
2. Calculate the class width (round up).

$$\text{class width} \approx \frac{\text{(maximum value)} - \text{(minimum value)}}{\text{number of classes}}$$

3. Starting point: Choose the minimum data value or a convenient value below it as the first lower class limit.
4. Using the first lower class limit and class width, proceed to list the other lower class limits.
5. List the lower class limits in a vertical column and proceed to enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.

Example: page 54, problem 18

Amounts of Strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Philadelphia residents born after 1979
Note: this data is related to Three Mile Island nuclear power plant Accident in 1979.

155	145
142	116
149	136
130	158
151	114
163	165
151	169
142	145
156	150
133	150
138	150
161	158
128	151
144	145
172	152
137	140
151	170
166	129
147	188
163	156

DIRECTIONS:

Construct a frequency distribution with 8 classes. Begin with a lower class limit of 110, and use a class width of 10. Cite a reason why such data are important

Example: page 54, problem 18

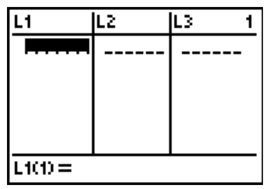
Data can be sorted using the graphing calculator.

The website below has useful graphing calculator tips:

www.mathbits.com/MathBits/TISection/Openpage.htm www.mathbits.com

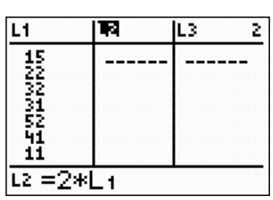
Entering Data:

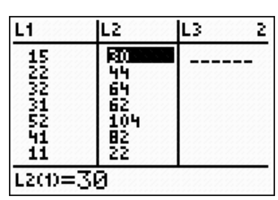
Data is stored in ***Lists*** on the calculator. Locate and press the **STAT** button on the calculator. Choose **EDIT**. The calculator will display the first three of six lists (columns) for entering data. Simply type your data and press **ENTER.** Use your arrow keys to move between lists.



www.mathbits.com

Data can be entered in a second list based upon the information in a previous list. In the example below, we will double all of our data values in **L1** and store them in **L2.** If you arrow up ONTO **L2**, you can enter a formula for generating **L2.** The formula will appear at the bottom of the screen. Press **ENTER** and the new list is created.





www.mathbits.com

To clear all data from a list:

Press **STAT**. From the **EDIT** menu, move the cursor up **ONTO** the name of the list (**L1**). Press **CLEAR**. Move the cursor down.

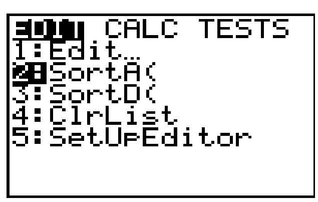
NOTE: The list entries will not disappear until the cursor is moved down. (**Avoid** pressing **DEL** as it will delete the entire column. If this happens, you can reinstate the column by pressing **STAT #5 SetUpEditor**.)

You may also clear a list by choosing option **#4** under the **EDIT** menu, **ClrList**. **ClrList** will appear on the home screen waiting for you to enter which list to clear. Enter the name of a list by pressing the **2nd** button and the yellow **L1** (above the **1**).

www.mathbits.com

Sorting Data: (helpful when finding the mode)

• Locate and press the **STAT** button. Choose option **#2, SortA(.**



• Specify the list you wish to sort by pressing the **2nd** button and the yellow **L1** list name.
• Press **ENTER** and the list will be put in ascending order (lowest to highest). **SortD** will put the list in descending order.

Example: page 54, problem 18

Data has been sorted.

114	150
116	151
128	151
129	151
130	151
133	152
136	155
137	156
138	156
140	158
142	158
142	161
144	163
145	163
145	165
145	166
147	169
149	170
150	172
150	188

Example: page 54, problem 18

List lower class limits:
(start with 110 and use class widths of 10)

Strontium-90 Level	Frequency
110-	
120-	
130-	
140-	
150-	
160-	
170-	
180-	

Example: page 54, problem 18

List upper class limits:

Strontium-90 Level	Frequency
110-119	
120-129	
130-139	
140-149	
150-159	
160-169	
170-179	
180-189	

Example: page 54, problem 18

Use sorted data and count the number of data values in each class:

Strontium-90 Level	Frequency
110-119	2
120-129	2
130-139	5
140-149	9
150-159	13
160-169	6
170-179	2
180-189	1

Interpreting Frequency Distributions

A frequency distribution is a normal distribution if it has a “bell” shape.

- The frequencies start low, then increase to one or two high frequencies, then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

Example: page 54, problem 18

Strontium-90 Level	Frequency
110-119	2
120-129	2
130-139	5
140-149	9
150-159	13
160-169	6
170-179	2
180-189	1

This frequency distribution is (approximately) a normal distribution.

Example: page 54, problem 19

Frequency distribution for the amounts of nicotine in nonfiltered king-sized cigarettes (from Data set 4 in Appendix B):

Nicotine (mg)	Frequency
1.0-1.1	14
1.2-1.3	4
1.4-1.5	3
1.6-1.7	3
1.8-1.9	1

This frequency distribution is not a normal distribution.

Gaps

- **Gaps**
  The presence of gaps can show that we have data from two or more different populations. **However, the converse is not true, because data from different populations do not necessarily result in gaps.**

See example 4 on page 51.

Example: page 55, problem 30

An analysis of 50 train derailment incidents identified the main causes as:

- *T = bad track*
- *H = human error*
- *O = other causes*

The categorical data that was collected is summarized below:

T T T E E H H H H H O O H H H E E T T T E T H O T
T T T T T T H T T H E E T T E E T T T H T T O O O

Construct a table summarizing the frequency distribution of this data.

Example: page 55, problem 30

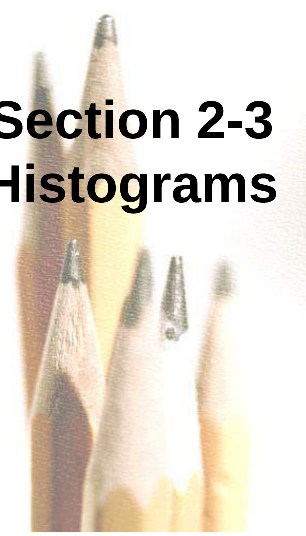
Causes	Frequency
Bad Track (T)	23
Faulty Equipment (E)	9
Human Error (H)	12
Other (O)	6

Recap

In this Section we have discussed

- **Important characteristics of data**
- **Frequency distributions**
- **Procedures for constructing frequency distributions**
- **Relative frequency distributions**
- **Cumulative frequency distributions**
- **Normal frequency distributions**

Lecture-7



Section 2-3
Histograms

Key Concept

A histogram is a graph of the frequency distribution.

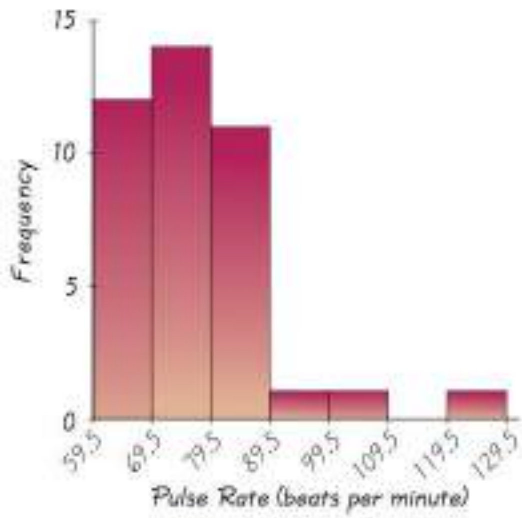
Histogram

A graph consisting of bars of equal width drawn adjacent to each other (without gaps). The horizontal scale represents the classes of quantitative data values and the vertical scale represents the frequencies. The heights of the bars correspond to the frequency values.

Histogram

Table 2-2 Pulse Rates of Females

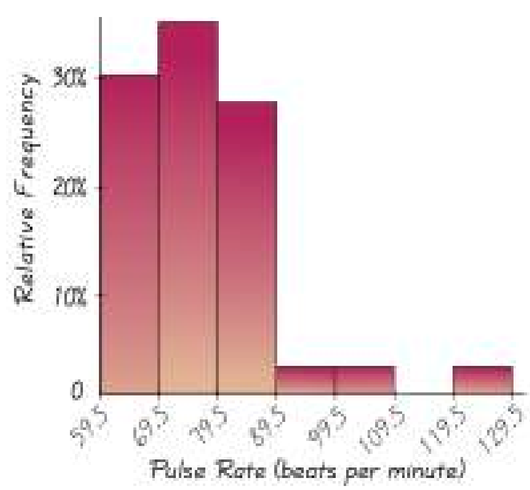
Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1



Relative Frequency Histogram

Table 2-3 Relative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Relative Frequency
60-69	30%
70-79	35%
80-89	27.5%
90-99	2.5%
100-109	2.5%
110-119	0
120-129	2.5%



Histogram

Horizontal Scale for Histogram: Use class boundaries or class midpoints.

Vertical Scale for Histogram: Use the class frequencies or relative frequencies.

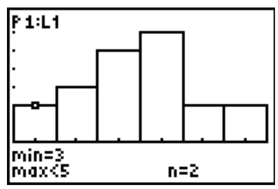
Histogram with Graphing Calculator

www.mathbits.com

•Press **2nd STATPLOT** and choose **#1 PLOT 1.** You should see the screen below. Be sure the plot is **ON**, the histogram icon is highlighted, and that the list you will be using is indicated next to **Xlist. Freq: 1** means that each piece of data will be counted one time.

Histogram with Graphing Calculator

• To see the histogram, press **ZOOM** and **#9 ZoomStat.** (**ZoomStat** automatically sets the window to an appropriate size to view all of the data.)
• Press the **TRACE** key to see on-screen data about the histogram. The spider will jump from bar to bar showing the range of values contained within each bar and the number of entries from the list (**n**) that fall within that range.



Histogram with Graphing Calculator

Example: problem 10 from page 58 which is a histogram of the frequency distribution from the previous example problem 18 from page 54

• NOTE: choosing **ZoomStat** automatically adjusts **Xmin, Xmax, Ymin, Ymax,** and **Xscl.**

Histogram with Graphing Calculator

- To *manually* adjust the histogram:
  - Under your **WINDOW** button, the **Xscl value controls the width of each bar** beginning with **Xmin and ending with Xmax**. (If you wish to see EACH piece of data as a separate interval, set the **Xscl** to 1)
  - Select **GRAPH (not ZoomStat)**
  - NOTE: If you wish to adjust your own viewing window, **(Xmax-Xmin)/Xscl** must be less than or equal to 47 for the histogram to be seen in the viewing window.

Histogram with Graphing Calculator

Example: problem 10 from page 58 again but choose:

Xmin=110
Xmax=190
Xscl=10

This histogram matches the histogram in solutions manual.

Note: this histogram is approximately a normal distribution.

Histogram with Graphing Calculator

Example: problems 11 from page 58 (frequency distribution below)

Nicotine (mg)	Frequency
1.0-1.1	14
1.2-1.3	4
1.4-1.5	3
1.6-1.7	3
1.8-1.9	1

Plot the histogram, it is not a normal distribution.

Critical Thinking Interpreting Histograms

Objective is not simply to construct a histogram, but rather to *understand* something about the data.

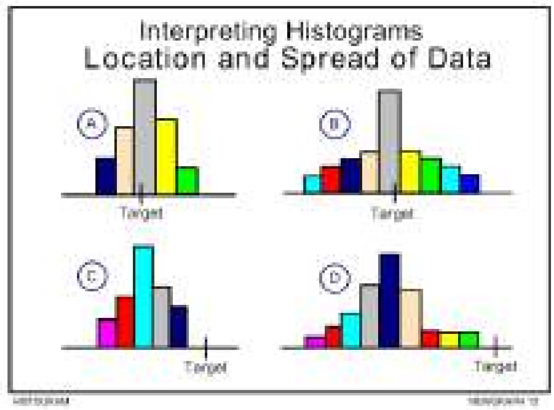
Special case: a normal distribution has a “bell” shape. Characteristic of the bell shape are

(1) The frequencies increase to a maximum, and then decrease, and

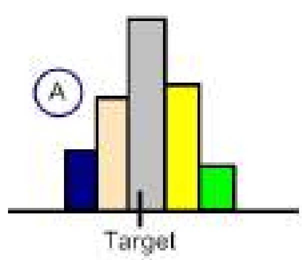
(2) symmetry, with the left half of the graph roughly a mirror image of the right half.

Interpreting Histograms

Compare these histograms from the webpage:
www.saferpak.com/histogram_articles/

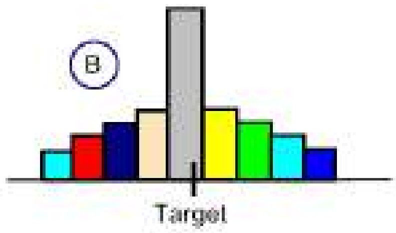


Interpreting Histograms



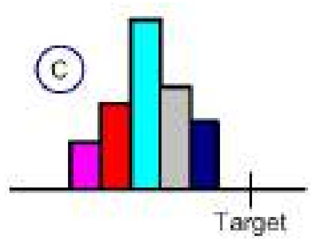
Most of the data were on target with small variation.

Interpreting Histograms



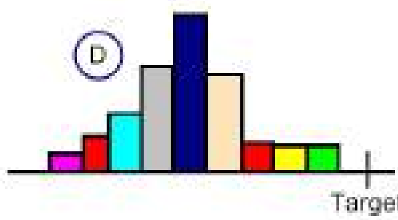
Some of the data were on target, but data shows large variation.

Interpreting Histograms



Data is below target with small variation.

Interpreting Histograms



Data is below target with large variation.

Interpreting Histograms

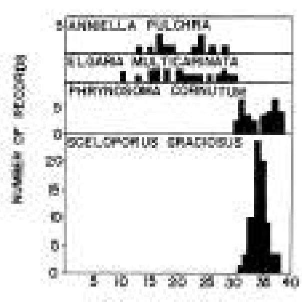


Fig. 8.—Histograms of body temperatures of active individuals of four genera of lizards, showing the absence of a preferred body temperature in *Anniella* and *Elgaria*, a broad range of preferred temperatures in *Phrynosoma*, and a relatively sharply defined preferred temperature in *Sceloporus*.

Published in:

Body Temperatures of Reptiles
Author(s): Bayard H. Brattstrom
Source: *American Midland Naturalist*, Vol. 73, No. 2 (Apr., 1965), pp. 376-422

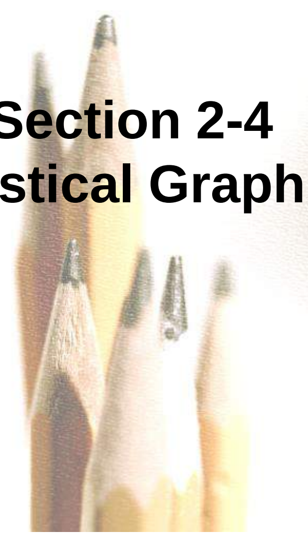
Recap

In this Section we have discussed

❖ **Histograms**

❖ **Relative Frequency Histograms**

Lecture-8



Section 2-4
Statistical Graphics

Key Concept

This section discusses other types of statistical graphs.

Our objective is to identify a suitable graph for representing the data set. The graph should be effective in revealing the important characteristics of the data.

Frequency Polygon

Uses line segments connected to points directly above class midpoint values

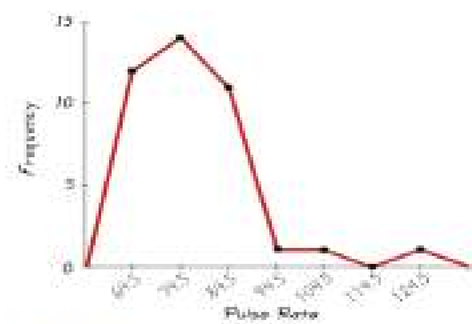


Figure 2-5 Frequency Polygon: Pulse Rates of Women

NOTE: this graph is generated from the frequency distribution in Table 2-2 on page 47

Relative Frequency Polygon

Uses relative frequencies (proportions or percentages) for the vertical scale.

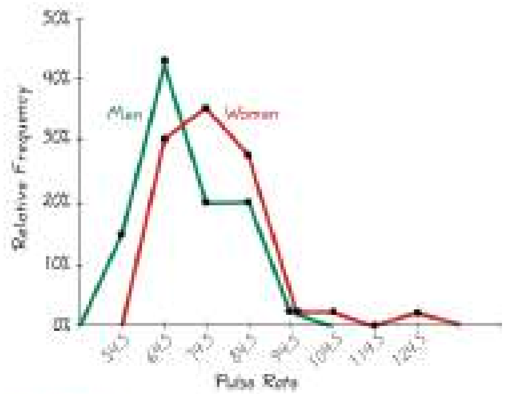


Figure 2-6 Relative Frequency Polygons: Pulse Rates of Women and Men

Ogive ("oh-jive")

A line graph that depicts cumulative frequencies

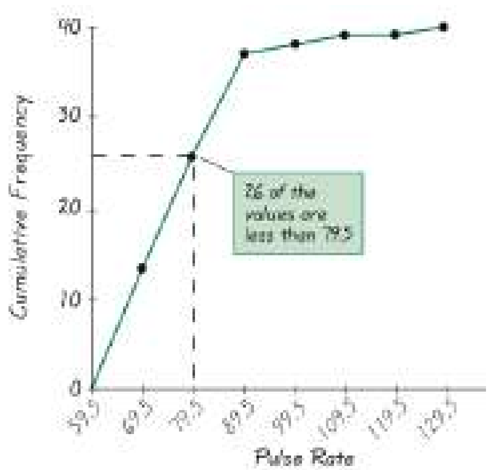
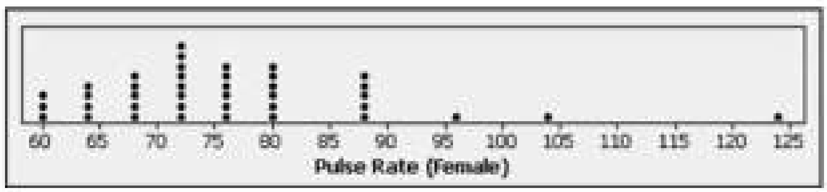


Figure 2-7 Ogive

Dot Plot

Each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.



NOTE: data is from Table 2-1 on page 47

Stemplot (or Stem-and-Leaf Plot)

Represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

Stemplot

Stem (tens)	Leaves (units)	
6	000444488888	← Data values are 60, 60, 60, 64, . . . , 68.
7	2222222266666	
8	00000088888	
9	6	← Data value is 96.
10	4	← Data value is 104.
11		
12	4	

NOTE: data is from Table 2-1 on page 47

Stemplot Example

114	150
116	151
128	151
129	151
130	151
133	152
136	155
137	156
138	156
140	158
142	158
142	161
144	163
145	163
145	165
145	166
147	169
149	170
150	172
150	188

Example: problem 6 from page 68 which is a stemplot of the Strontium-90 data from problem 18 from page 54

Stemplot Example

114	150
116	151
128	151
129	151
130	151
133	152
136	155
137	156
138	156
140	158
142	158
142	161
144	163
145	163
145	165
145	166
147	169
149	170
150	172
150	188

Stem	Leaf
11	46
12	89
13	03678
14	022455579
15	0001111256688
16	133569
17	02
18	8

The stemplot shows the distribution is approximately normal centered around 150.

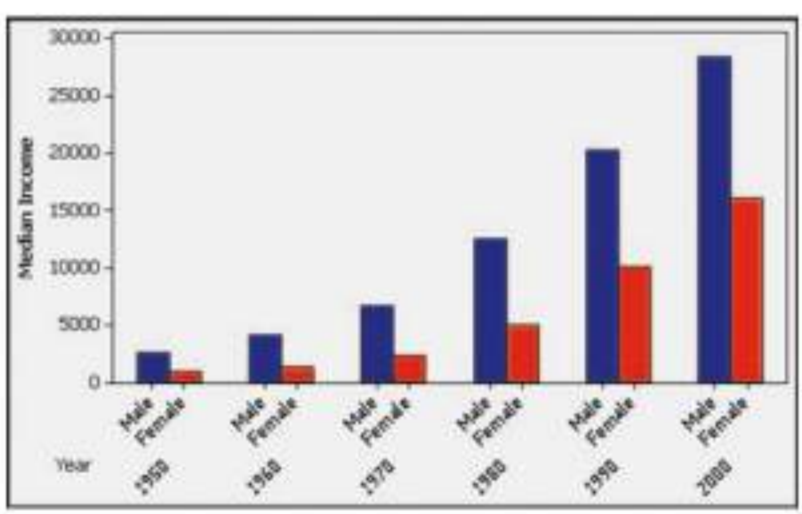
Bar Graph

Uses bars of equal width to show frequencies of categories of <u>qualitative</u> data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data.

A *multiple bar graph* has two or more sets of bars, and is used to compare two or more data sets.

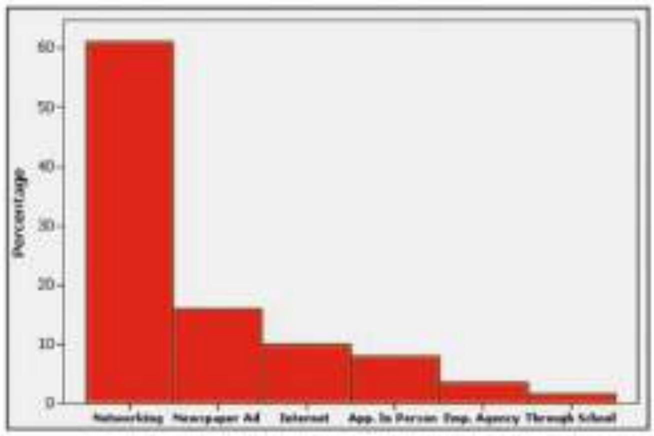
Multiple Bar Graph

Median Income of Males and Females



Pareto Chart

A bar graph for qualitative data, with the bars arranged in descending order according to frequencies

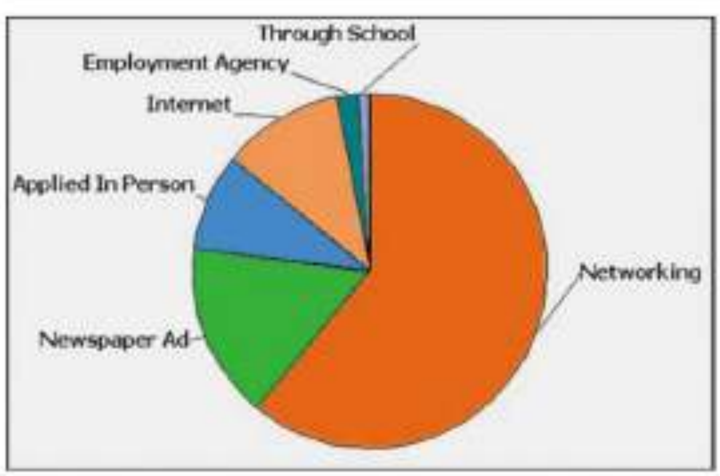


2.1 - 71

36

Pie Chart

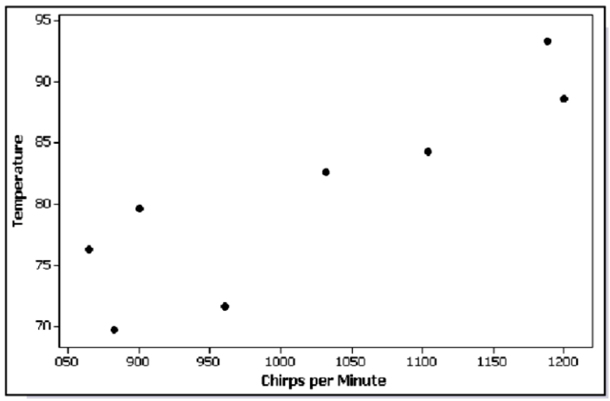
A graph depicting qualitative data as slices of a circle, size of slice is proportional to frequency count



2.1 - 72

Scatter Plot (or Scatter Diagram)

A plot of paired (*x,y*) data with a horizontal *x*-axis and a vertical *y*-axis. Used to determine whether there is a relationship between the two variables



Scatterplot with Graphing Calculator

www.mathbits.com

• Enter the X data values in **L1**. Enter the Y data values in **L2**, being careful that each X data value and its matching Y data value are entered on the same horizontal line.

• Activate the scatter plot. Press **2nd STATPLOT** and choose **#1 PLOT 1.** Be sure the plot is **ON**, the scatter plot icon is highlighted, and that the list of the X data values are next to **Xlist**, and the list of the Y data values are next to **Ylist.** Choose any of the three marks. Press **ZOOM** and **#9 ZoomStat.**

Example of Scatterplot

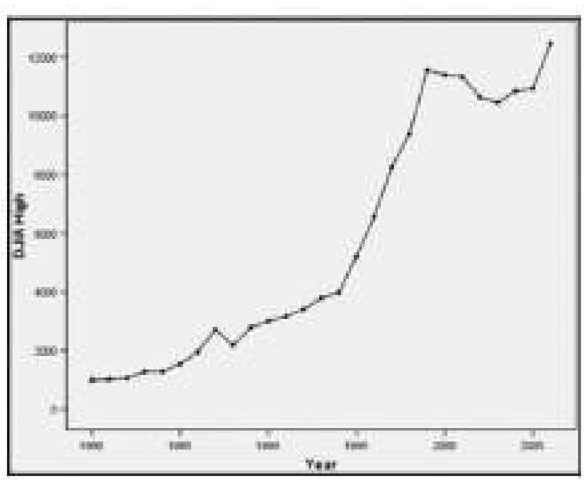
Data was collected from a random sample of 16 students which measured student height (inches) and student armspan (inches). Construct a scatterplot of this data.

Height	Arm Span
152	159
156	155
160	160
163	166
165	163
168	176
168	164
173	171

Height	Arm Span
173	170
173	169
173	176
179	183
180	175
182	181
183	188
193	188

Time-Series Graph

Data that have been collected at different points in time: *time-series data*



Important Principles Suggested by Edward Tufte

For small data sets of 20 values or fewer, use a table instead of a graph.

A graph of data should make the viewer focus on the true nature of the data, not on other elements, such as eye-catching but distracting design features.

Do not distort data, construct a graph to reveal the true nature of the data.

Almost all of the ink in a graph should be used for the data, not the other design elements.

Important Principles Suggested by Edward Tufte

Don’t use screening consisting of features such as slanted lines, dots, cross-hatching, because they create the uncomfortable illusion of movement.

Don’t use area or volumes for data that are actually one-dimensional in nature. (Don’t use drawings of dollar bills to represent budget amounts for different years.)

Never publish pie charts, because they waste ink on nondata components, and they lack appropriate scale.

Recap

In this section we saw that graphs are excellent tools for describing, exploring and comparing data.

Describing data: Histogram - consider distribution, center, variation, and outliers.

Exploring data: features that reveal some useful and/or interesting characteristic of the data set.

Comparing data: Construct similar graphs to compare data sets.

Lecture-9

Section 2-5
Critical Thinking:
Bad Graphs

Key Concept

Some graphs are bad in the sense that they contain errors.

Some are bad because they are technically correct, but misleading.

It is important to develop the ability to recognize bad graphs and identify exactly how they are misleading.

Nonzero Axis

Are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.

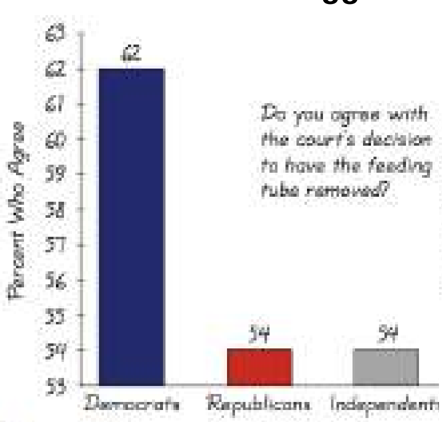


Figure 2-1 Survey Results by Party

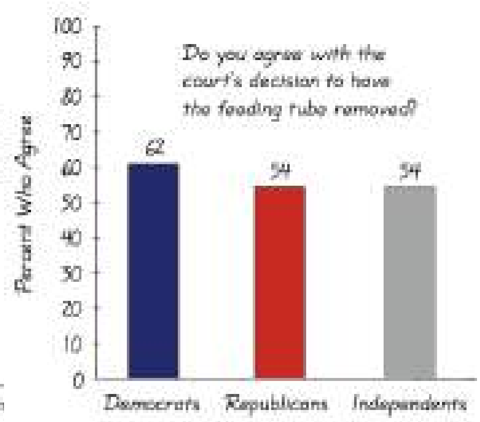


Figure 2-9 Survey Results by Party

Pictographs

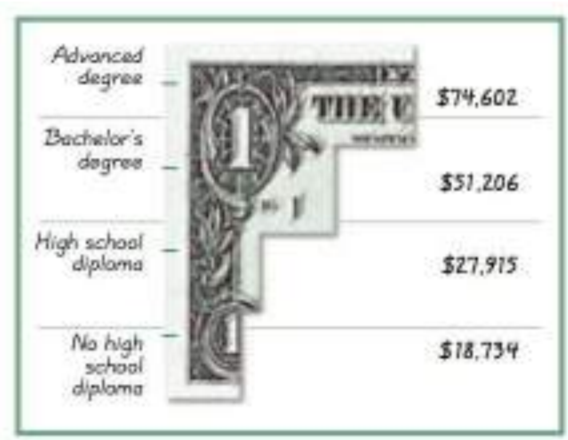
are drawings of objects. Three-dimensional objects - money bags, stacks of coins, army tanks (for army expenditures), people (for population sizes), barrels (for oil production), and houses (for home construction) are commonly used to depict data.

These drawings can create false impressions that distort the data.

If you double each side of a square, the area does not merely double; it increases by a factor of four;if you double each side of a cube, the volume does not merely double; it increases by a factor of eight.

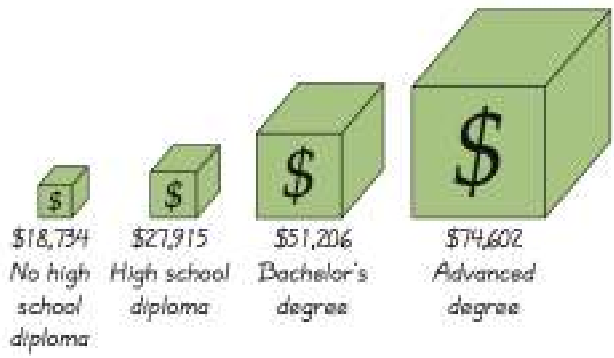
Pictographs using areas and volumes can therefore be very misleading.

Annual Incomes of Groups with Different Education Levels



Bars have same width, too busy, too difficult to understand.

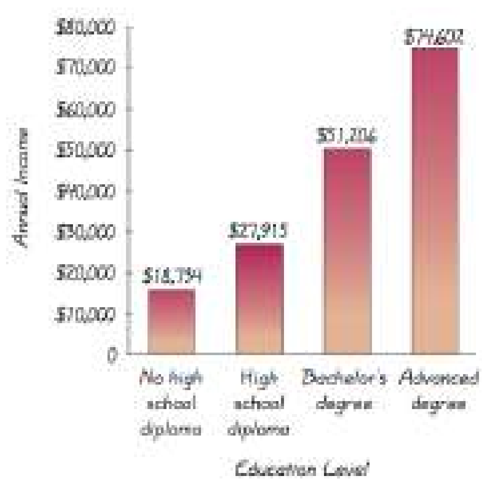
Annual Incomes of Groups with Different Education Levels



Misleading. Depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

43

Annual Incomes of Groups with Different Education Levels



Fair, objective, unencumbered by distracting features.

Lecture-9

Chapter 5

Probability Distributions

5-1 Review and Preview

5-2 Random Variables

5-3 Binomial Probability Distributions

5-4 Mean, Variance and Standard Deviation for the Binomial Distribution

5-5 Poisson Probability Distributions

Section 5-1
Review and Preview

Review and Preview

This chapter combines the methods of descriptive statistics presented in Chapter 2 and 3 and those of probability presented in Chapter 4 to describe and analyze

probability distributions.

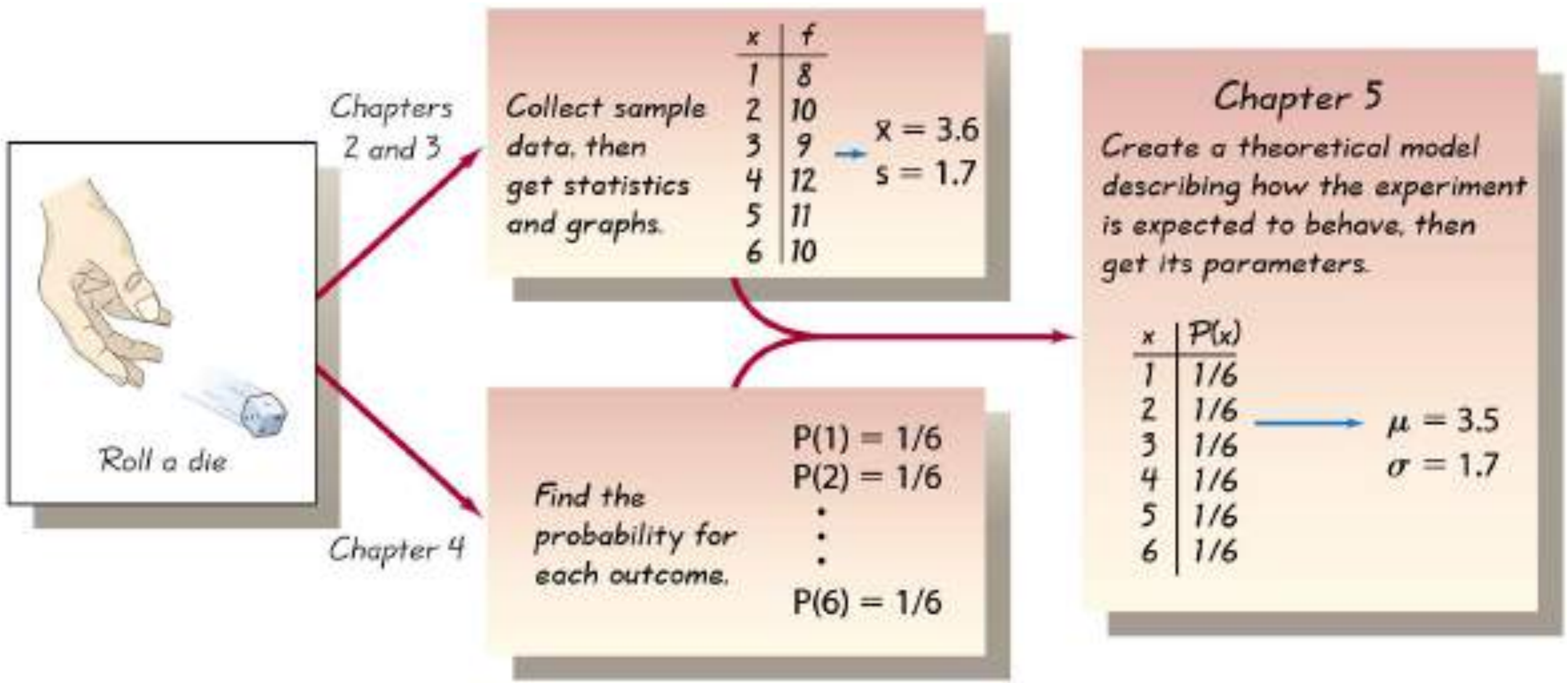
Probability Distributions describe what will probably happen instead of what actually did happen, and they are often given in the format of a graph, table, or formula.

Preview

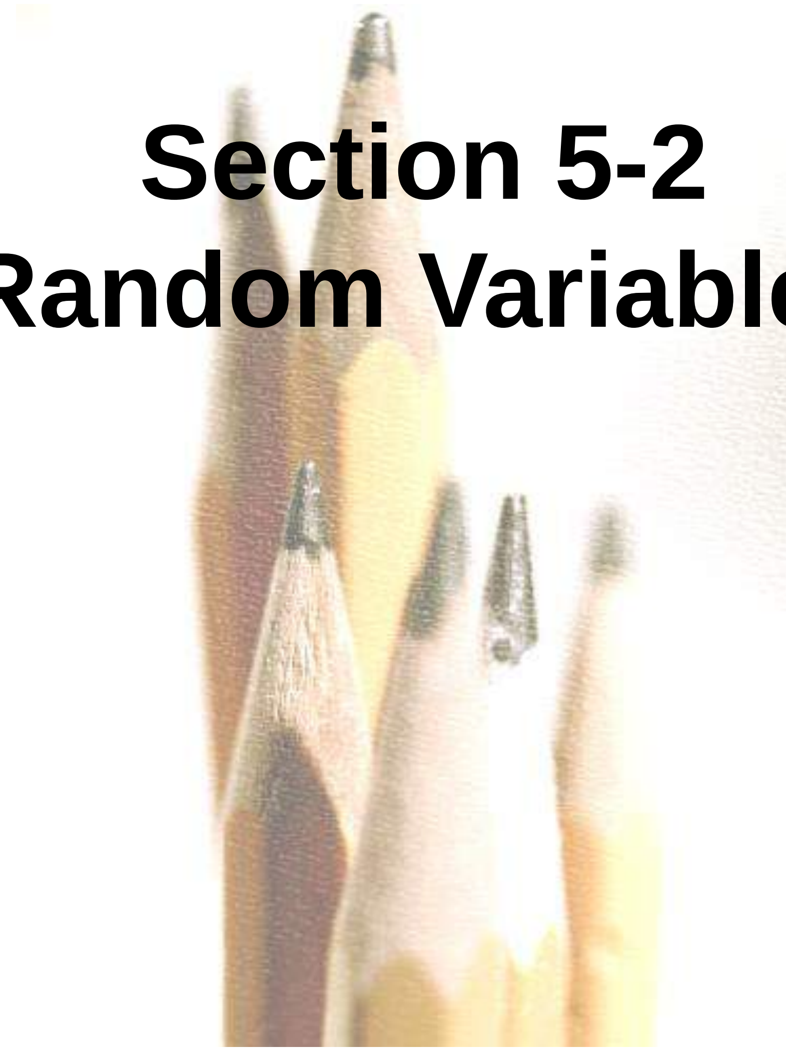
In order to fully understand probability distributions, we must first understand the concept of a random variable, and be able to distinguish between discrete and continuous random variables. In this chapter we focus on discrete probability distributions. In particular, we discuss binomial and Poisson probability distributions.

Combining Descriptive Methods and Probabilities

In this chapter we will construct probability distributions by presenting possible outcomes along with the relative frequencies we expect.



Section 5-2
Random Variables



Key Concept

This section introduces the important concept of a probability distribution, which gives the probability for each value of a variable that is determined by chance.

Give consideration to distinguishing between outcomes that are likely to occur by chance and outcomes that are “unusual” in the sense they are not likely to occur by chance.

Key Concept

- The concept of random variables and how they relate to probability distributions
- Distinguish between discrete random variables and continuous random variables
- Develop formulas for finding the mean, variance, and standard deviation for a probability distribution
- Determine whether outcomes are likely to occur by chance or they are unusual (in the sense that they are not likely to occur by chance)

Random Variable
Probability Distribution

- **Random variable**
  a variable (typically represented by *x*) that has a single numerical value, determined by chance, for each outcome of a procedure
- **Probability distribution**
  a description that gives the probability for each value of the random variable; often expressed in the format of a graph, table, or formula

Discrete and Continuous Random Variables

- **Discrete random variable**
  either a finite number of values or countable number of values, where "countable" refers to the fact that there might be infinitely many values, but they result from a counting process

- **Continuous random variable**
  infinitely many values, and those values can be associated with measurements on a continuous scale without gaps or interruptions

Example

Page 208, problem 6

Identify each as a discrete or continuous random variable.

(a) **Total amount in ounces of soft drinks you consumed in the past year.**

(b) **The number of cans of soft drinks that you consumed in the past year.**

(c) **The number of movies currently playing in U.S. theaters.**

Example

Page 208, problem 6

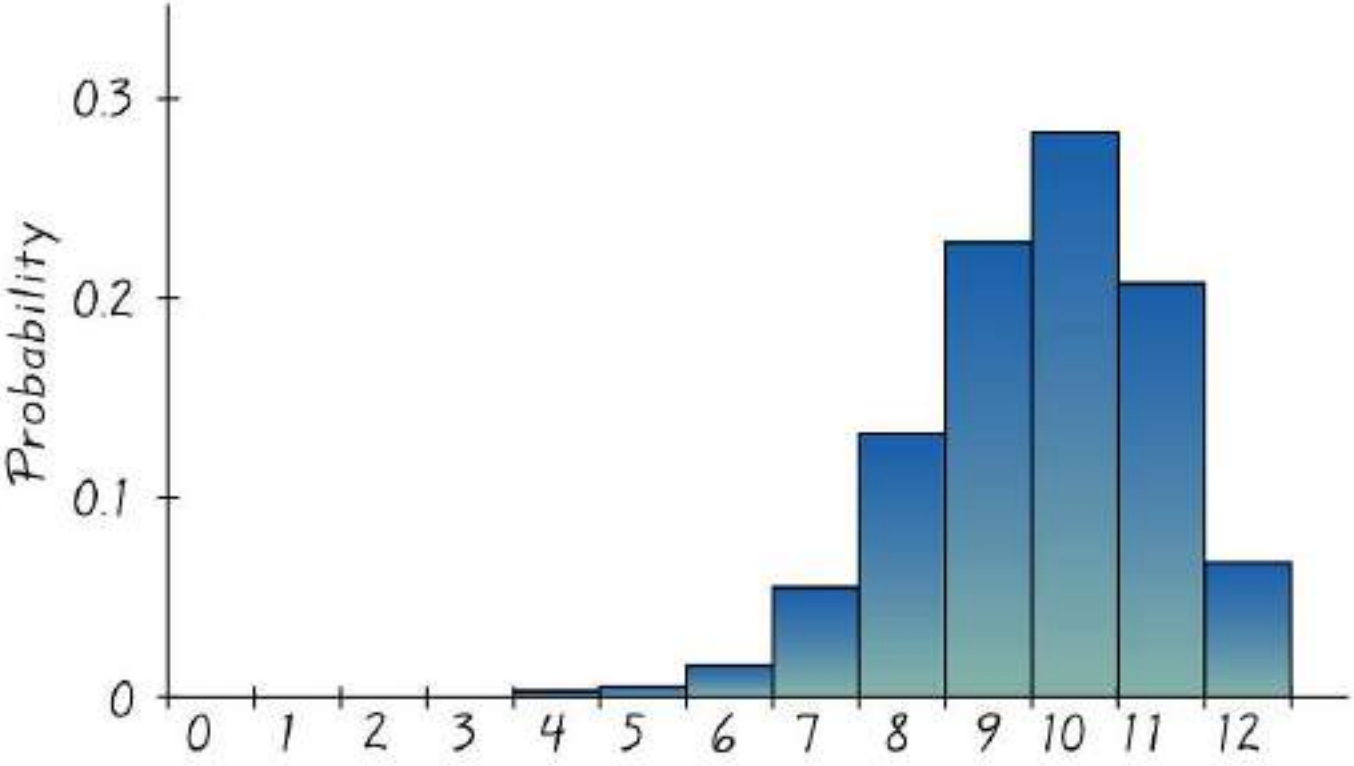
Identify each as a discrete or continuous random variable.

(d) The running time of a randomly selected movie

(e) The cost of making a randomly selected movie.

Graphs

The probability histogram is very similar to a relative frequency histogram, but the vertical scale shows probabilities.

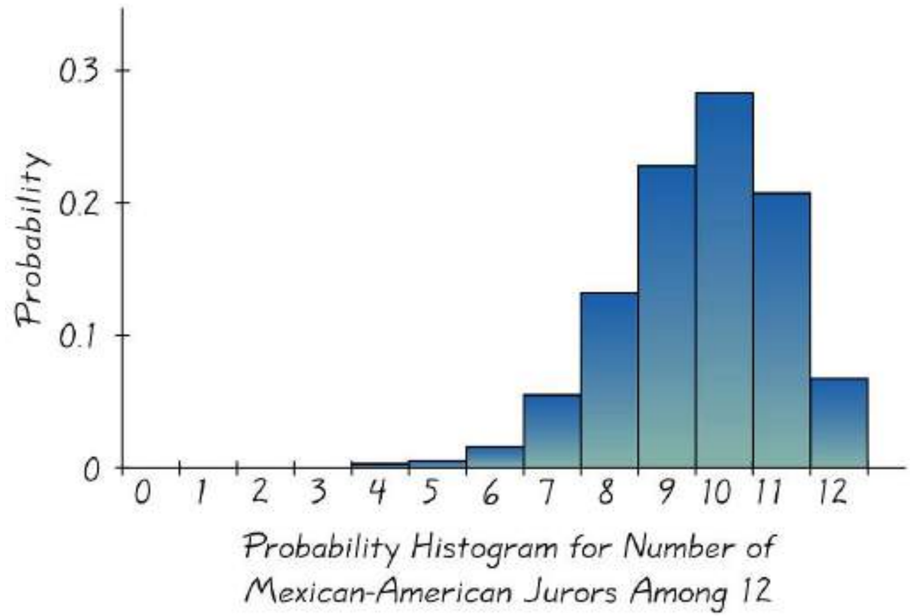


Probability Histogram for Number of Mexican-American Jurors Among 12

Visual Representation of Probability Distributions

Probability distributions can be represented by tables and graphs.

Number of Mex. Am. Jurors (*x*)	***P(x)***
4	0.005
5	0.010
6	0.030
7	0.045
8	0.130
9	0.230
10	0.290
11	0.210
12	0.050



Requirements for Probability Distribution

$$\sum P(x) = 1$$

where *x* assumes all possible values.

$$0 \leq P(x) \leq 1$$

for every individual value of *x*.

Example

Page 208, problem 8

The variable x represents the number of cups or cans of caffeinated beverages consumed by Americans each day.

x	*P(x)*
0	0.22
1	0.16
2	0.21
3	0.16

Is this a probability distribution?

Example

Page 208, problem 8

x	$P(x)$
0	0.22
1	0.16
2	0.21
3	0.16

Total last column:

$\sum P(x) =$ **0.22+0.16+0.21+0.16=0.75**

This is <u>not</u> a probability distribution.

Example

Page 210, problem 18

Based on information from the MRINetwork, some job applicants are required to have several interviews before a decision is made. The number of required interviews and the corresponding probabilities are 1 (0.09); 2 (0.31); 3 (0.37); 4 (0.12); 5 (0.05); 6 (0.05)

(a) Does this information describe a probability distribution?

Example

If x is the number of required interviews. This is a probability distribution.

x	$P(x)$
1	0.09
2	0.31
3	0.37
4	0.12
5	0.05
6	0.05

$\sum P(x) =$ **0.09+0.31+0.37+0.12+0.05+0.05=0.99**

and each *P(x)* is between 0 and 1.

Example

Page 210, problem 18

- **(b) If this is a probability distribution, find the mean and standard deviation.**

Mean, Variance and Standard Deviation of a Probability Distribution

$$\mu = \Sigma\,[x \cdot P(x)]$$

Mean

$$\sigma^2 = \Sigma\,[(x - \mu)^2 \cdot P(x)]$$

Variance

$$\sigma^2 = \Sigma\,[x^2 \cdot P(x)] - \mu^2$$

Variance (shortcut)

$$\sigma = \sqrt{\Sigma\,[x^2 \cdot P(x)] - \mu^2}$$

Standard Deviation

Roundoff Rule for μ, σ, and σ^2

Round results by carrying one more decimal place than the number of decimal places used for the random variable x.

If the values of x are integers, round μ, σ, and σ^2 to one decimal place.

Example

Page 210, problem 18

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
1	0.09	0.09	1	0.09
2	0.31	0.62	4	1.24
3	0.37	1.11	9	3.33
4	0.12	0.48	16	1.92
5	0.05	0.25	25	1.25
6	0.05	0.30	36	1.80

$$\mu = \sum xP(x) = 2.85 \approx 2.9$$

Example

Page 210, problem 18

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
1	0.09	0.09	1	0.09
2	0.31	0.62	4	1.24
3	0.37	1.11	9	3.33
4	0.12	0.48	16	1.92
5	0.05	0.25	25	1.25
6	0.05	0.30	36	1.80

$$\sigma^2 = \sum x^2 \cdot P(x) - \mu^2 = 9.63 - (2.85)^2 = 1.5075$$

$$\sigma = \sqrt{1.5075} \approx 1.2$$

Example

Page 210, problem 18

- **(c) Use the range rule of thumb to identify the range of values for usual numbers of interviews.**

Identifying *Unusual* Results
Range Rule of Thumb

According to the **range rule of thumb**, most values should lie within 2 standard deviations of the mean.

We can therefore identify "unusual" values by determining if they lie outside these limits:

Maximum usual value = $\mu + 2\sigma$

Minimum usual value = $\mu - 2\sigma$

Example

ANSWER:

Minimum usual value:

$$\mu - 2\sigma = 2.9 - 2(1.2) = 0.5$$

Maximum usual value:

$$\mu + 2\sigma = 2.9 - 2(1.2) = 5.3$$

Usual numbers of interviews are between 0.5 and 5.3.

(d) It is not unusual to have a decision after one interview since 1 is between 0.5 and 5.3

Identifying Unusual Results Probabilities

Rare Event Rule for Inferential Statistics

If, under a given assumption (such as the assumption that a coin is fair), the probability of a particular observed event (such as 992 heads in 1000 tosses of a coin) is extremely small, we conclude that the assumption is probably not correct.

Identifying Unusual Results Probabilities

Using Probabilities to Determine When Results Are Unusual

- **Unusually high**: x successes among n trials is an **unusually high** number of successes if $P(x \text{ or more}) \leq 0.05$.

- **Unusually low**: x successes among n trials is an **unusually low** number of successes if $P(x \text{ or fewer}) \leq 0.05$.

Example

Page 210, problem 22

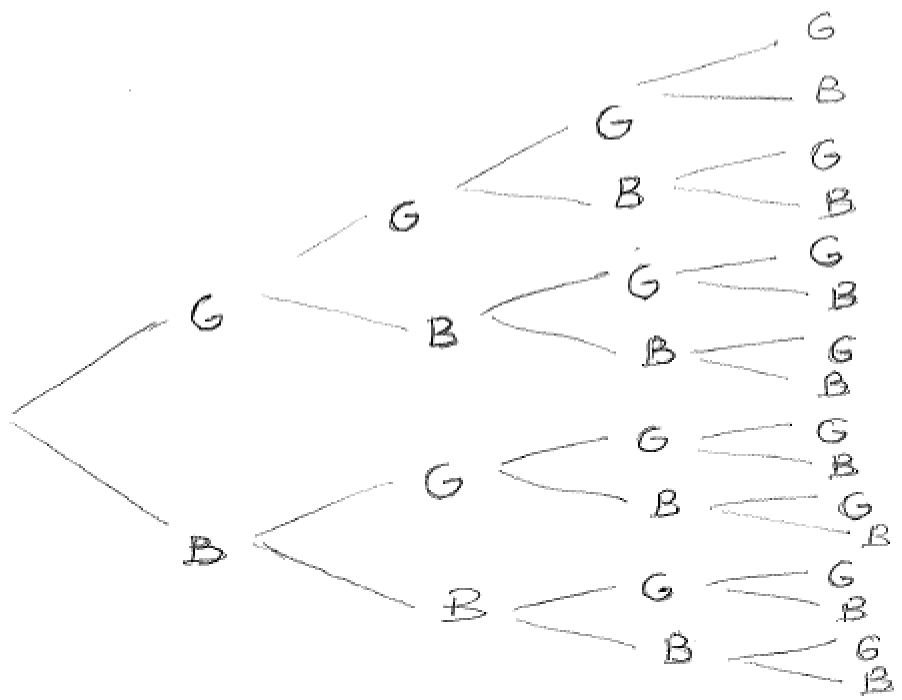
Let the random variable *x* represent the number of girls in a family of four children. Construct a table describing the probability distribution, then find the mean and standard deviation.

(NOTE: unlike previous example, we must compute the probabilities here)

Example

Page 210, problem 22

Determine the outcomes with a tree diagram:



Example

Page 210, problem 22

Determine the outcomes with a tree diagram.

- **Total number of outcomes is 16**
- **Total number of ways to have 0 girls is 1**

$$P(0\text{ girls}) = 1/16 = 0.0625$$

- **Total number of ways to have 1 girl is 4**

$$P(1\text{ girl}) = 4/16 = 0.2500$$

- **Total number of ways to have 2 girls is 6**

$$P(2\text{ girls}) = 6/16 = 0.375$$

Example

Page 210, problem 22

Determine the outcomes with a tree diagram.

- **Total number of ways to have 3 girls is 4**

$$P(3\text{ girls}) = 4/16 = 0.2500$$

- **Total number of ways to have 4 girls is 1**

$$P(4\text{ girls}) = 1/16 = 0.0625$$

Example

Page 210, problem 22

Distribution is:

x	$P(x)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625

NOTE: $$\sum P(x) = 1$$

Example

Page 210, problem 22

Determine the outcomes with counting formulas.

- **Total number of outcomes is**

$$2 \cdot 2 \cdot 2 \cdot 2 = 2^4 = 16$$

Now use permutations when some items may be identical (formula on page 181).

- **Total number of ways to have 0 girls (select 4 from from 4 boys)**

$$\frac{4!}{0! \cdot 4!} = 1$$

Example

Page 210, problem 22

Determine the outcomes with counting formulas.

- **Total number of ways to have 1 girl (select 4 from from 3 boys and one girl)**

$$\frac{4!}{1!\cdot 3!}=4$$

- **Total number of ways to have 2 girls (select 4 from from 2 boys and two girls)**

$$\frac{4!}{2!\cdot 2!}=6$$

- **Etc.**

Example

Page 210, problem 22

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
0	0.0625	0	0	0
1	0.2500	0.25	1	0.2500
2	0.3750	0.75	4	1.5000
3	0.2500	0.75	9	2.2500
4	0.0625	0.25	16	1.0000

$$\mu = \sum xP(x) = 2.0$$

Example

Page 210, problem 22

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
0	0.0625	0	0	0
1	0.2500	0.25	1	0.2500
2	0.3750	0.75	4	1.5000
3	0.2500	0.75	9	2.2500
4	0.0625	0.25	16	1.0000

$$\sigma^2 = \sum x^2 \cdot P(x) - \mu^2 = 5.0000 - 4.0000 = 1.0000$$

$$\sigma = \sqrt{1.0000} = 1.0$$

Expected Value

The expected value of a discrete random variable is denoted by E, and it represents the mean value of the outcomes. It is obtained by finding the value of $\Sigma\,[x \cdot P(x)]$.

$$E = \sum [x \cdot P(x)]$$

Example

Page 210, problem 26

> **In New Jersey's pick 4 lottery game, you pay 50 cents to select a sequence of four digits, such as 1332. If you select the same sequence of four digits that are drawn, you win and collect $2788.**

(a) How many selections are possible?

(b) What is the probability of winning?

Example

ANSWER:

(a) Each of the four positions can be filled with 10 numbers 0,1,2,…,9 to get

$$10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = \boxed{10{,}000}$$

(b) There is only one winning sequence

$$P(W) = 1/10{,}000 = 0.0001$$

Example

Page 210, problem 26

(c) What is the net profit if you win?

Example

ANSWER:

(c) Net profit is payoff minus the original bet:

$$\$2788.00 - \$0.50 = \boxed{\$2787.50}$$

(c) There is only one winning sequence

$$P(W) = 1/10{,}000 = 0.0001$$

Example

Page 210, problem 26

(d) Find the expected value

Example

Summarize in a table (again):

	x	$P(x)$	$x \cdot P(x)$
lose $\rightarrow$	-0.50	0.9999	-0.49995
win $\rightarrow$	2787.50	0.0001	0.27875

$$E = \sum xP(x) = -0.22120 \approx -0.221$$

Expected value is -22.1 cents.

Example

Page 210, problem 26

(e) If you bet 50 cents in the Illinois Pick 4 game, the expected value is -25 cents. Which bet is better: a 50 cent bet in the Illinois Pick 4 or 50 cent bet in the New Jersey Pick 4? Explain

Example

Page 210, problem 26

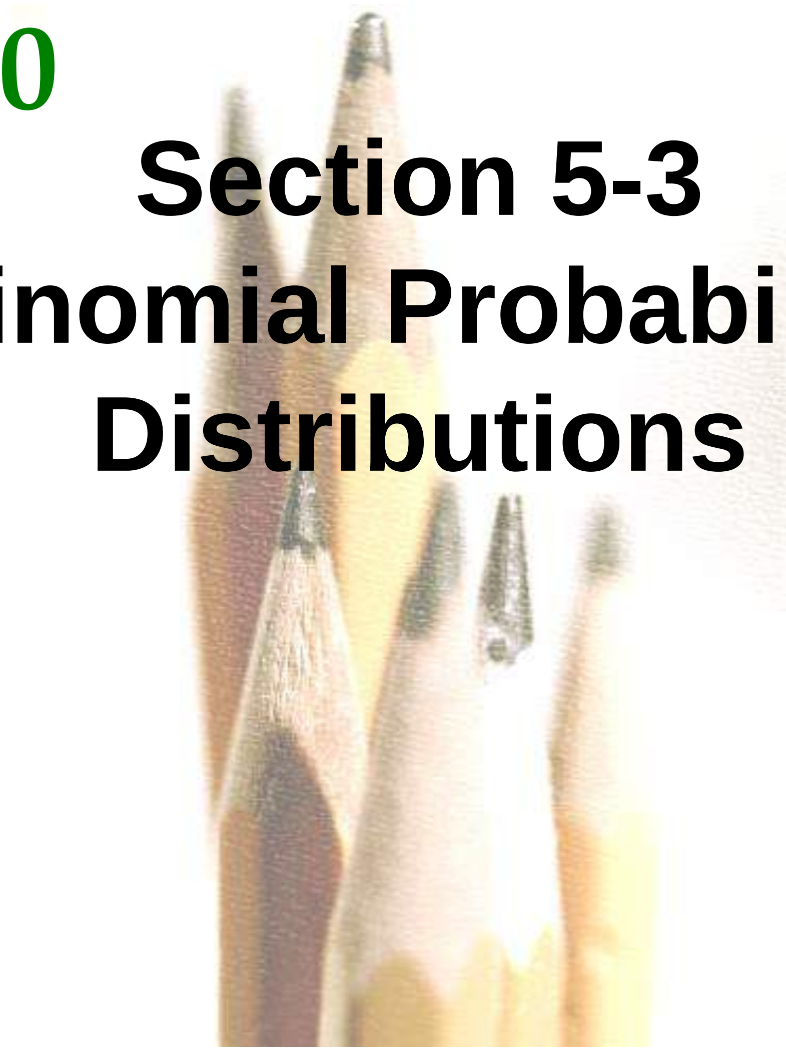
(e) Since -22.1 is larger than -25, New Jersey has a better Pick 4 (on average, you can expect to lose less money!)

Recap

In this section we have discussed:

- **Combining methods of descriptive statistics with probability.**
- **Random variables and probability distributions.**
- **Probability histograms.**
- **Requirements for a probability distribution.**
- **Mean, variance and standard deviation of a probability distribution.**
- **Identifying unusual results.**
- **Expected value.**

Lecture-10



Section 5-3
Binomial Probability Distributions

Key Concept

This section presents a basic definition of a binomial distribution along with notation, and methods for finding probability values.

Binomial probability distributions allow us to deal with circumstances in which the outcomes belong to two relevant categories such as acceptable/defective or survived/died.

Motivational Example

Genetics

- **In mice an allele A for agouti (gray-brown, grizzled fur) is <u>dominant</u> over the allele a, which determines a non-agouti color. Suppose each parent has the genotype Aa and 4 offspring are produced. What is the probability that <u>exactly</u> 3 of these have agouti fur?**

Motivational Example

- *A single offspring has genotypes:*

	A	a
A	AA	Aa
a	aA	aa

Sample Space

$$\{AA, Aa, aA, aa\}$$

Motivational Example

- *Agouti genotype is <u>dominant</u>*
  - *Event that offspring is agouti:*

$$\{AA, Aa, aA\}$$

- *Therefore:*

$$P(\text{agouti genotype}) = 3/4$$

$$P(\text{not agouti genotype}) = 1/4$$

Motivational Example

- *Let G represent an agouti offspring and N represent non-agouti*
- *<u>Exactly three</u> agouti offspring may occur in four different ways (in order of birth):*

NGGG, GNGG, GGNG, GGGN

Motivational Example

- *Events (birth of a mouse) are independent and using multiplication rule:*

$$P(NGGG) = P(N)\cdot P(G)\cdot P(G)\cdot P(G) = \frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} = \frac{27}{256}$$

$$P(GNGG) = P(G)\cdot P(N)\cdot P(G)\cdot P(G) = \frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} = \frac{27}{256}$$

$$P(GGNG) = P(G)\cdot P(G)\cdot P(N)\cdot P(G) = \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4} = \frac{27}{256}$$

$$P(GGGN) = P(G)\cdot P(G)\cdot P(G)\cdot P(N) = \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} = \frac{27}{256}$$

Motivational Example

- *P(exactly 3 offspring has agouti fur)*

$$P(NGGG \cup GNGG \cup GGNG \cup GGGN)$$

$$= P(NGGG) + P(GNGG) + P(GGNG) + P(GGGN)$$

$$= \frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}$$

$$= 4\cdot\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right) = 4\cdot\frac{27}{256} \approx \boxed{0.422}$$

Binomial Probability Distribution

A **binomial probability distribution** results from a procedure that meets all the following requirements:

1. The procedure has a **fixed number of trials**.
2. The trials must be **independent**. (The outcome of any individual trial doesn't affect the probabilities in the other trials.)
3. Each trial must have all outcomes classified into **two categories** (commonly referred to as **success** and **failure**).
4. The probability of a success remains the same in all trials.

Previous example is binomial distribution

1. number of trials is 4 in all cases
2. trials are independent
3. each trial results in success (agouti fur) and failure (non-agouti fur).
4. probability of a success is always ¾

Example

Page 219, problem 6

Treat 863 subjects with Lipitor and ask each subject how their heads feel.

Does this result in a binomial distribution?

Example

Page 219, problem 6

ANSWER:

no, there are more than two possible outcomes when asked how your head feels.

Example

Page 219, problem 8

Treat 152 couples with YSORT gender selection method (developed by the Genetics and IVF Institute) and record the gender of each of the 152 babies that are born.

Does this result in a binomial distribution?

Example

Page 219, problem 8

ANSWER:

yes, all four requirements are met.

Example

Page 219, problem 12

Two hundred statistics students are randomly selected and each is asked if he or she owns a TI-84 Plus calculator

5% Rule

- **When sampling without replacement, consider events to be independent if $n < 0.05N$ where n is the number of items sampled and N is the total number of data items in the sample space**
- ***This is the same as:*** $$\frac{n}{N} < 0.05 = 5\%$$

Example

Page 219, problem 8

ANSWER:

yes, all four requirements are met if we use the 5% rule for independence:

$$\frac{200}{N} < 0.05 = 5\%$$

where *N* is the number of *all* statistics students which is assumed to be much larger than 200.

Notation for Binomial Probability Distributions

n denotes the fixed number of trials.

x denotes a specific number of successes in *n* trials, so *x* can be any whole number between 0 and *n*, inclusive.

p denotes the probability of success in *one* of the *n* trials.

q denotes the probability of failure in one of the *n* trials.

P(*x*) denotes the probability of getting exactly *x* successes among the *n* trials.

Notation: Agouti Fur Genotype Example

n=4 **denotes the fixed number of four trials**

x=3 **denotes 3 successes in 4 trials**

p=3/4 **the probability of success in *one* of the *4* trials is 3/4**

q=1/4 **the probability of failure in one of the *four* trials is 1/4**

P(x) **denotes the probability of getting exactly 3 successes among the 4 trials.**

The Binomial Probability Formula

$$P(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$$

for $x = 0, 1, 2, \ldots, n$

where

n = number of trials

x = number of successes among n trials

p = probability of success in any one trial

q = probability of failure in any one trial ($q = 1 - p$)

Agouti Fur Genotype Example

$$P(x) = \frac{4!}{(4-3)! \cdot 3!} \cdot \left(\frac{3}{4}\right)^3 \cdot \left(\frac{1}{4}\right)^1$$

$$= 4 \cdot \left(\frac{3}{4}\right)^3 \cdot \left(\frac{1}{4}\right)^1$$

$$= 4 \cdot \left(\frac{27}{64}\right) \cdot \left(\frac{1}{4}\right) = 0.422$$

Rationale for the Binomial Probability Formula

$$P(x) = \underbrace{\frac{n!}{(n-x)!\,x!}}_{\text{The number of outcomes with exactly } x \text{ successes among } n \text{ trials}} \cdot p^x \cdot q^{n-x}$$

Binomial Probability Formula

$$P(x) = \underbrace{\frac{n!}{(n-x)!\,x!}}_{\text{Number of outcomes with exactly } x \text{ successes among } n \text{ trials}} \cdot \underbrace{p^x \cdot q^{n-x}}_{\text{The probability of } x \text{ successes among } n \text{ trials for any one particular order}}$$

Number of outcomes with exactly *x* successes among *n* trials

The probability of *x* successes among *n* trials for any one particular order

Binomial Probability Formula

Compare:

$$\frac{n!}{(n-x)!\cdot x!}$$

With counting formula for permutations when some items are identical to others (page 181, 4-6)

$$\frac{n!}{n_1!\cdot n_2!}$$

Lecture-11

Methods for Finding Probabilities

We will now discuss three methods for finding the probabilities corresponding to the random variable *x* in a binomial distribution.

Method 1: Using the Binomial Probability Formula

$$P(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$$

for $x = 0, 1, 2, \ldots, n$

where

n = number of trials

x = number of successes among n trials

p = probability of success in any one trial

q = probability of failure in any one trial ($q = 1 - p$)

Example

Page 220, problem 22

Use the binomial probability formula to find the probability of 2 successes (x=2) in 9 trials (n=9) given the probability of success is 0.35 (p=0.35)

Example

Page 220, problem 22

ANSWER:

$$P(2) = \frac{9!}{7! \cdot 2!}(0.35)^2(0.65)^7$$

$$= 36(0.35)^2(0.65)^7$$

$$= \boxed{0.216}$$

Method 2: Using Technology

STATDISK, Minitab, Excel, SPSS, SAS and the TI-83/84 Plus calculator can be used to find binomial probabilities.

STATDISK

Binomial Probability

Num Trials, n: 5

Success Prob, p: 0.75

Evaluate

Mean:	3.7500
St Dev:	0.9682
Variance:	0.9375

x	P(x)	P(x or fewer)	P(x or greater)
0	0.0009766	0.0009766	1.0000000
1	0.0146484	0.0156250	0.9990234
2	0.0878906	0.1035156	0.9843750
3	0.2636719	0.3671875	0.8964844
4	0.3955078	0.7626953	0.6328125
5	0.2373047	1.0000000	0.2373047

Clear Copy

Help ?

MINITAB

x	P(x)
0	0.000977
1	0.014648
2	0.087891
3	0.263672
4	0.395508
5	0.237305

Method 2: Using Technology

STATDISK, Minitab, Excel and the TI-83 Plus calculator can all be used to find binomial probabilities.

EXCEL

	A	B
1	0	0.000977
2	1	0.014648
3	2	0.087891
4	3	0.263672
5	4	0.395508
6	5	0.237305

TI-83 PLUS Calculator

L1	L2	L3 2
0	9.8E-4	------
1	.01465	
2	.08789	
3	.26367	
4	.39551	
5	.2373	

L2(7) =

Example

Page 220, problem 22 (using TI-84+)

ANSWER:

2ND VARS ⇨ A:binompdf(⇨ 9, .35, 2)

n*, *p*, *x

Then Enter gives the result 0.216

Method 3: Using Table A-1 in Appendix A

TABLE A-1 Binomial Probabilities

							p								
n	*x*	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	*x*
2	0	.980	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002	0+	0
	1	.020	.095	.180	.320	.420	.480	.500	.480	.420	.320	.180	.095	.020	1
	2	0+	.002	.010	.040	.090	.160	.250	.360	.490	.640	.810	.902	.980	2
3	0	.970	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	0+	0+	0
	1	.029	.135	.243	.384	.441	.432	.375	.288	.189	.096	.027	.007	0+	1
	2	0+	.007	.027	.096	.189	.288	.375	.432	.441	.384	.243	.135	.029	2
	3	0+	0+	.001	.008	.027	.064	.125	.216	.343	.512	.729	.857	.970	3
4	0	.961	.815	.656	.410	.240	.130	.062	.026	.008	.002	0+	0+	0+	0
	1	.039	.171	.292	.410	.412	.346	.250	.154	.076	.026	.004	0+	0+	1
	2	.001	.014	.049	.154	.265	.346	.375	.346	.265	.154	.049	.014	.001	2
	3	0+	0+	.004	.026	.076	.154	.250	.346	.412	.410	.292	.171	.039	3
	4	0+	0+	0+	.002	.008	.026	.062	.130	.240	.410	.656	.815	.961	4
5	0	.951	.774	.590	.328	.168	.078	.031	.010	.002	0+	0+	0+	0+	0
	1	.048	.204	.328	.410	.360	.259	.156	.077	.028	.006	0+	0+	0+	1
	2	.001	.021	.073	.205	.309	.346	.312	.230	.132	.051	.008	.001	0+	2
	3	0+	.001	.008	.051	.132	.230	.312	.346	.309	.205	.073	.021	.001	3
	4	0+	0+	0+	.006	.028	.077	.156	.259	.360	.410	.328	.204	.048	4
	5	0+	0+	0+	0+	.002	.010	.031	.078	.168	.328	.590	.774	.951	5

Example

Page 220, problem 16 use table A-1 in appendix

$n=5, \quad x=1, \quad p=0.95$

Example

Page 220, problem 16 use table A-1 in appendix

ANSWER: $\boxed{\textbf{0+}}$

NOTE:

0+ means positive but "close to" zero

Calculator answer is:

2.96875E-5 = 0.0000296875 (almost zero)

Example

Page 220, problem 30

The brand name of McDonald's has a 95% recognition rate. If a McDonald's executive wants to verify this rate by beginning with a small sample of 15 randomly selected consumers, find the probability that exactly 13 of the 15 consumers recognize the McDonald's brand name. Also find the probability that the number who recognize the brand name is not 13.

Example

ANSWER:

x = number of consumers who recognize McDonald's brand name

Probability a consumer recognizes McDonald's brand name is **95%=0.95**

(a) Probability x is exactly 13?

Use binomial distribution with ***n=15, p=0.95, q=0.05, x=13***

$$P(x=13)=\frac{15!}{13!\cdot 2!}(0.95)^{13}\cdot(0.05)^{2}=0.135$$

Example

(b) Probability *x* is <u>not</u> 13?

$$P(x \neq 13) = 1 - P(x = 13) = 1 - 0.135 = 0.865$$

Example

Page 221, problem 36

The author purchased a slot machine configured so that there is a 1/2000 probability of winning the jackpot on any individual trial.

(a) Find the probability of exactly 2 jackpots in 5 trials.

(b) Find the probability of at least 2 jackpots in 5 trials

(c) If a guest claims that she played the slot machine 5 times and hit the jackpot twice, is this claim valid? Explain.

Example

ANSWER:

x **= number of Jackpots hit**

Probability a jackpot is hit **is 1/2000 = 0.0005**

(a) Probability *x* is exactly 2?

Use binomial distribution with *n=5, p=0.0005, x=2*

$$P(x=2)=\frac{5!}{3!\cdot 2!}(0.0005)^2\cdot(0.9995)^3=0.000002496$$

Example

(b) Probability of at least 2 jackpots?

At least 2 jackpots means 2 or more which means x=2 or x=3 or x=4 or x=5.

It will be easier to compute the complement of at least 2 jackpots which means less than 2 which means x=0 or x=1 then use the complement rule for probabilities:

$$P(\text{at least } 2) = 1 - P(x=0 \text{ OR } x=1)$$

Example

(b)

$$P(0\ \text{or}\ 1) = P(x=0) + P(x=1)$$

$$= \frac{5!}{0!\cdot 5!}(0.0005)^0 \cdot (0.9995)^5 + \frac{5!}{1!\cdot 4!}(0.0005)^1 \cdot (0.9995)^4$$

$$= 0.9975025 + 0.0024950$$

$$= 0.9999975$$

$$1 - P(0\ \text{or}\ 1) = 1 - 0.9999975 = \boxed{0.00000250}$$

Example

(c) **If a guest claims that she played the slot machine 5 times and hit the jackpot twice, is this claim valid? Explain.**

It could happen, but since 0.00000250<0.05 this would be considered a rare event.

Recap

In this section we have discussed:

- **The definition of the binomial probability distribution.**
- **Notation.**
- **Important hints.**
- **Three computational methods.**
- **Rationale for the formula.**

Lecture-12

Section 5-4
Mean, Variance, and Standard Deviation for the Binomial Distribution

For Any Discrete Probability Distribution: Formulas

Mean $\mu = \sum[x \bullet P(x)]$

Variance $\sigma^2 = [\sum x^2 \bullet P(x)] - \mu^2$

Std. Dev $\sigma = \sqrt{[\sum x^2 \bullet P(x)] - \mu^2}$

Binomial Distribution: Formulas

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Std. Dev. $\sigma = \sqrt{n \cdot p \cdot q}$

Where

n = number of fixed trials

p = probability of success in one of the n trials

q = probability of failure in one of the n trials

Example

Page 226, problem 6

In an analysis of test results from the YSORT gender selection method, 152 babies are born and it is assumed that boys and girls are equally likely, so n=152 and p=0.5 Find the mean and standard deviation.

$$\mu = np = (152)(0.5) = 76.0$$

Example

ANSWER

Mean:

$$\mu = np = (152)(0.5) = 76.0$$

Variance:

$$\sigma^2 = npq = (152)(0.5)(0.5) = 38.0$$

Standard deviation:

$$\sqrt{\sigma} = \sqrt{38.0} \approx 6.2$$

Interpretation of Results

The range rule of thumb suggests that values are unusual if they lie outside of these limits:

$$\textbf{Maximum usual values} = \mu + 2\,\sigma$$

$$\textbf{Minimum usual values} = \mu - 2\,\sigma$$

Example

Page 226, problem 6 (continued)

Maximum usual values

$$\mu + 2\sigma = 76.0 + 2(6.2) = 88.4$$

Minimum usual values

$$\mu - 2\sigma = 76.0 - 2(6.2) = 63.6$$

Example

Page 226, problem 14

In a test of the YSORT method of gender selection 152 babies are born to couples trying to have baby boys, and 127 of those babies are boys.

(a) If the gender selection method has no effect and boys and girls are equally likely, find the mean and standard deviation for 152 babies.

(b) Is the result of 127 boys unusual? Does it suggest that the gender selection method appears to be effective?

Example

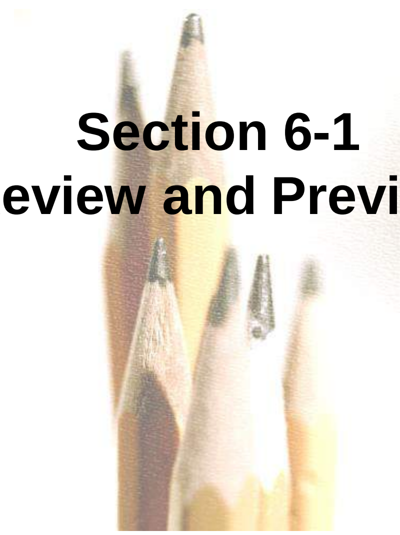
ANSWER

(a) **We did this part in problem 6**

(b) **Since 127 is not within the limits 63.6 and 88.4 of usual values we found in problem 6, it would be unusual to have 127 boys in 152 births. This suggests that the gender selection method is effective.**

Lecture-13 Chapter 6
Normal Probability Distributions

1



Section 6-1
Review and Preview

Review

- Chapter 2: Distribution of data
- Chapter 3: Measures of data sets, including measures of center and variation
- Chapter 4: Principles of probability
- Chapter 5: Discrete probability distributions

2

Preview

Chapter focus is on:

- Continuous random variables
- Normal distributions

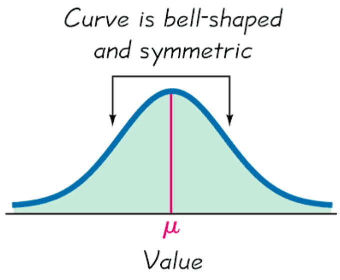
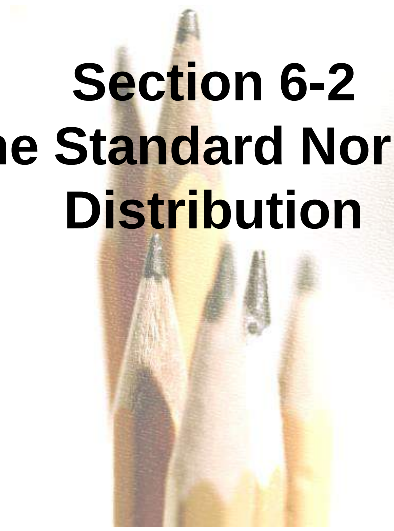


Figure 6-1

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Formula 6-1

Distribution determined by fixed values of mean and standard deviation



Section 6-2
The Standard Normal Distribution

3

Density Curve

A **density curve** is the graph of a continuous probability distribution. It must satisfy the following properties:

1. The total area under the curve must equal 1.
2. Every point on the curve must have a vertical height that is 0 or greater. (That is, the curve cannot fall below the *x*-axis.)

Area and Probability

Because the total area under the density curve is equal to 1, there is a correspondence between *area* and *probability*.

Uniform Distribution

A continuous random variable has a uniform distribution if its values are spread evenly over the range of probabilities. The graph of a uniform distribution results in a rectangular shape.

Notation

$P(x > a)$

denotes the probability that the *x* is greater than *a.*

$P(x < a)$

denotes the probability that the *x* is less than *a.*

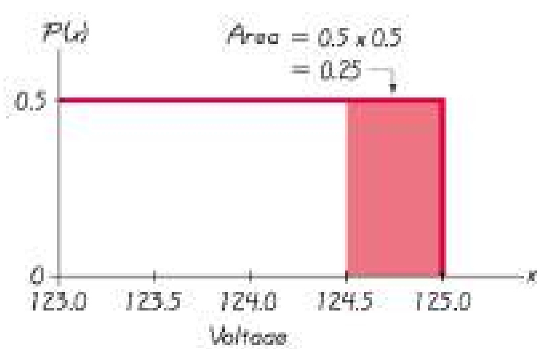
$P(a < x < b)$

denotes the probability that the *x* is between *a* and *b.*

5

Using Area to Find Probability

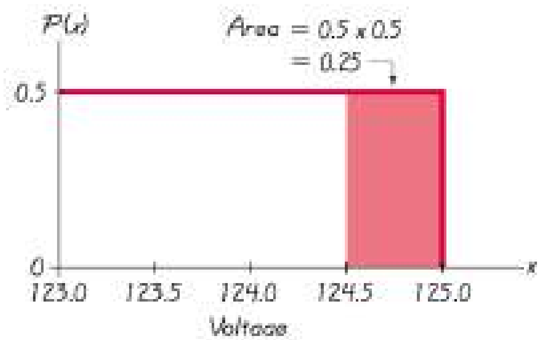
Given the uniform distribution illustrated, find the probability that a randomly selected voltage level (*x*) is greater than 124.5 volts.



$P(x>124.5) = ?$

Using Area to Find Probability

***ANSWER: P(x>124.5) =* 0.25**



Shaded area represents voltage levels greater than 124.5 volts. Correspondence between area and probability:
***P(x>124.5) =* 0.25**

6

Example

Page 249 problem 6

$$
\begin{aligned}
P(x < 123.5) &= (width)\cdot(height) \\
&= (123.5 - 123.0)(0.5) \\
&= (0.5)(0.5) = \boxed{0.25}
\end{aligned}
$$

Example

Page 249 problem 8

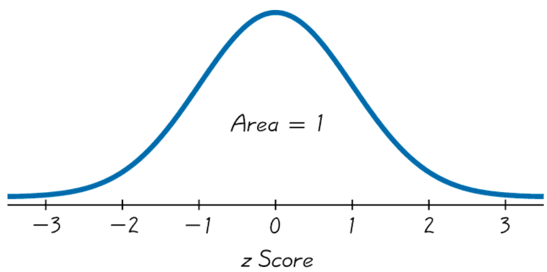
$$
\begin{aligned}
P(124.1 < x < 124.5) &= (width)\cdot(height) \\
&= (124.5 - 124.1)(0.5) \\
&= (0.4)(0.5) = \boxed{0.20}
\end{aligned}
$$

Standard Normal Distribution

The standard normal distribution is a normal probability distribution (bell-shaped graph) with μ = 0 and σ = 1. The total area under its density curve is equal to 1.

The horizontal axis is the *z-score*

Standard Normal Distribution



8

Finding Probabilities When Given *z*-scores

- **Table A-2 (in Appendix A)**
- **Gives the probability that *z* is less than some value which is the <u>cumulative area from the left</u> for the standard normal distribution curve.**

Finding Probabilities – Other Methods

- STATDISK
- Minitab
- Excel
- TI-83/84 Plus

9

Methods for Finding Normal Distribution Areas

Table A-2, STATDISK, Minitab, Excel

Gives the cumulative area from the left up to a vertical line above a specific value of *z*.

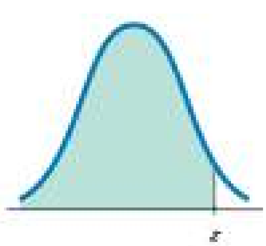


Table A-2 The procedure for using Table A-2 is described in the text.

STATDISK Select **Analysis, Probability Distributions, Normal Distribution.** Enter the *z* value, then click on **Evaluate.**

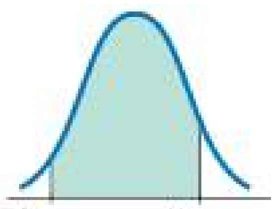
MINITAB Select **Calc, Probability Distributions, Normal.** In the dialog box, select **Cumulative Probability, Input Constant.**

EXCEL Select **fx, Statistical, NORMDIST.** In the dialog box, enter the value and mean, the standard deviation, and "true."

Methods for Finding Normal Distribution Areas

TI-83/84 Plus Calculator

Gives area bounded on the left and bounded on the right by vertical lines above any specific values.



TI-83/84 Press 2ND VARS **[2: normal cdf ( ],** then enter the two z scores separated by a comma, as in (left z score, right z score).

10

Table A-2

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

Using Table A-2

1. It is designed only for the *standard* normal distribution, which has a mean of 0 and a standard deviation of 1.
2. It is on two pages, with one page for *negative* *z*-scores and the other page for *positive* *z*-scores.
3. Each value in the body of the table is a *cumulative area from the left* up to a vertical boundary above a specific *z*-score.

Using Table A-2

4. When working with a graph, avoid confusion between *z*-scores and areas.

   ***z* Score**

   Distance along horizontal scale of the standard normal distribution; refer to the leftmost column and top row of Table A-2.

   Area

   Region under the curve; refer to the values in the body of Table A-2.

5. The part of the z-score denoting hundredths is found across the top.

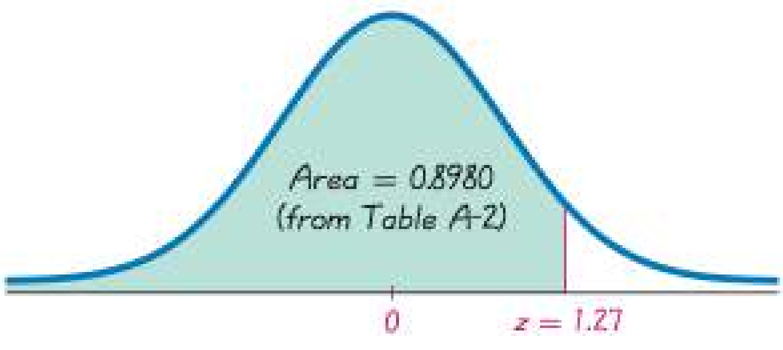
Example - Thermometers

The Precision Scientific Instrument Company manufactures thermometers that are supposed to give readings of 0ºC at the freezing point of water. Tests on a large sample of these instruments reveal that at the freezing point of water, some thermometers give readings below 0º (denoted by negative numbers) and some give readings above 0º (denoted by positive numbers). Assume that the mean reading is 0ºC and the standard deviation of the readings is 1.00ºC. Also assume that the readings are normally distributed. If one thermometer is randomly selected, find the probability that, at the freezing point of water, the reading is less than 1.27º.

12

Example - (Continued)

P(z < 1.27) =



Look at Table A-2

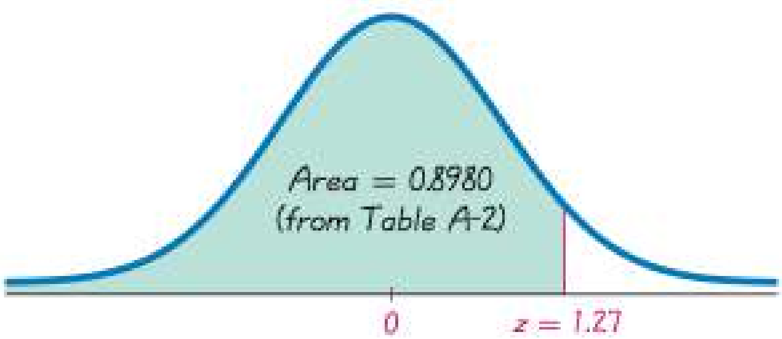
TABLE A-2 *(continued)* Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292

13

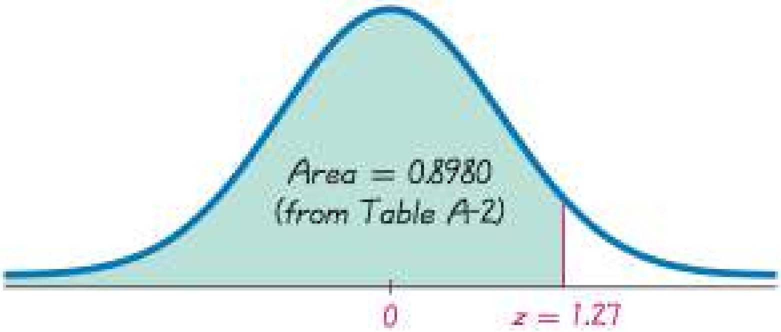
Example - cont

P (z < 1.27) = 0.8980



Example - cont

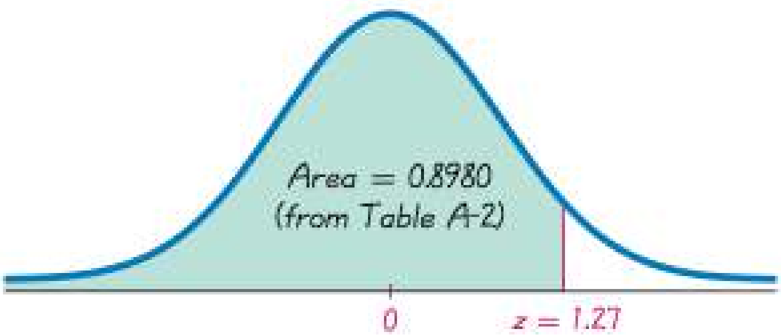
$P(z < 1.27) = 0.8980$



The *probability* of randomly selecting a thermometer with a reading less than 1.27º is 0.8980.

Example - cont

$P(z < 1.27) = 0.8980$



Or 89.80% will have readings below 1.27º.

A Sum Rule for Normal Probability Distribution

Because the events *z<a* and *z>a* are complements (if we ignore the *z=a* case):

$$P(z < a) + P(z > a) = 1$$

$$\boxed{P(z > a) = 1 - P(z < a)}$$

Example - Thermometers Again

If thermometers have an average (mean) reading of 0 degrees and a standard deviation of 1 degree for freezing water, and if one thermometer is randomly selected, find the probability that it reads (at the freezing point of water) above –1.23 degrees.

Example - Thermometers Again

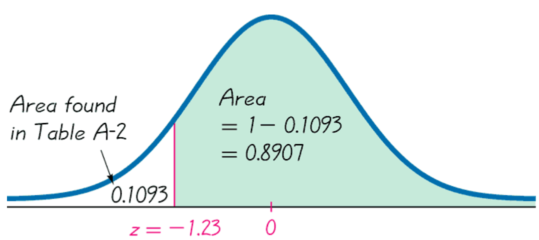
ANSWER:

find from Table A-2

$$P(z > -1.23) = 1 - P(z < -1.23)$$
$$= 1 - 0.1093 = \boxed{0.8907}$$

16

Example - cont



89.07% of the thermometers have readings above –1.23 degrees.

A Difference Rule for Normal Probability Distribution

Using area under the normal curve shows that:

$$\boxed{P(a < z < b) = P(z < b) - P(z < a)}$$

17

Example - Thermometers III

A thermometer is randomly selected. Find the probability that it reads (at the freezing point of water) <u>between</u> –2.00 and 1.50 degrees.

Example - Thermometers III

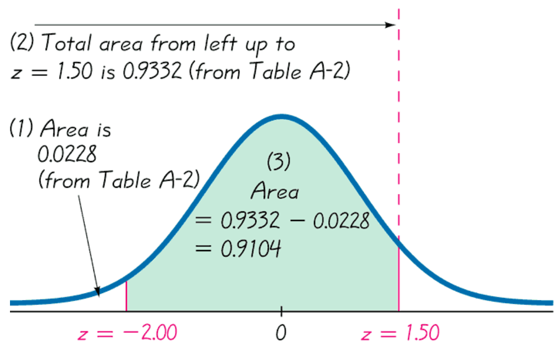
ANSWER:

find from Table A-2

$$P(-2.00 < z < 1.50) = P(z < 1.50) - P(z < -2.00)$$
$$= 0.9332 - 0.0228$$
$$= \boxed{0.9104}$$

18

Example - Thermometers III



The probability that the chosen thermometer has a reading between – 2.00 and 1.50 degrees is 0.9104.

If many thermometers are selected and tested at the freezing point of water, then 91.04% of them will read between –2.00 and 1.50 degrees.

Example

Page 250, problem 9

Example

Page 250, problem 9

ANSWER

$$P(z < 0.75) = 0.7734$$

Example

Page 250, problem 12

Example

ANSWER:

$$
\begin{aligned}
P(-.90 < z < 1.60) &= P(z < 1.60) - P(z < -0.90) \\
&= 0.9452 - 0.1841 \\
&= \boxed{0.7611}
\end{aligned}
$$

Example

Page 250, problem 18

$$P(z < -2.75) = \boxed{0.0030}$$

Example

ANSWER:

$$P(z < -2.75) = \boxed{0.0030}$$

Example

Page 250, problem 22

Example

ANSWER:

$$\begin{aligned} P(z > 2.33) &= 1 - P(z < 2.33) \\ &= 1 - 0.9901 = \boxed{0.0099} \end{aligned}$$

Example

Page 250, problem 26

23

Example

ANSWER:

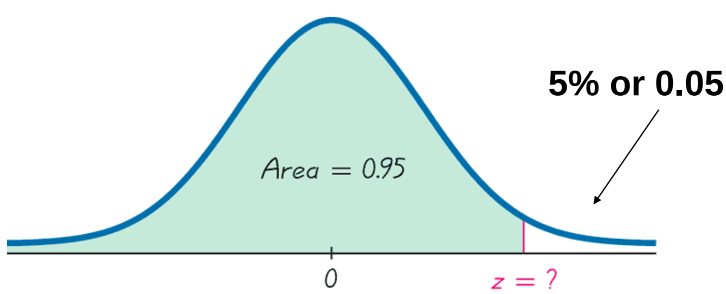
$$
\begin{aligned}
P(1.00 < z < 3.00) &= P(z < 3.00) - P(z < 1.00) \\
&= 0.9987 - 0.8413 \\
&= \boxed{0.1574}
\end{aligned}
$$

Finding a *z* Score When Given a Probability Using Table A-2

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given probability. If that region is not a cumulative region from the left, work instead with a known region that is a cumulative region from the left.

2. Using the cumulative area from the left, locate the closest probability in the body of Table A-2 and identify the corresponding *z* score.

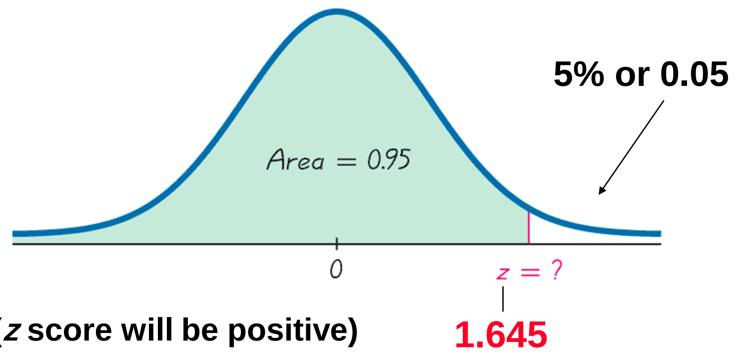
Finding *z* Scores When Given Probabilities



(*z* score will be positive)

Finding the 95th Percentile

Finding z Scores When Given Probabilities - cont

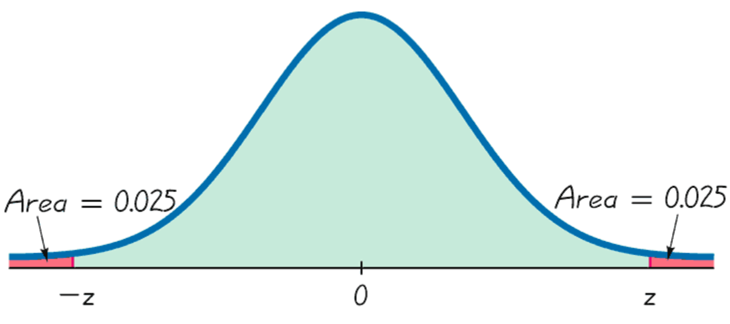


(*z* score will be positive)

Finding the 95th Percentile

25

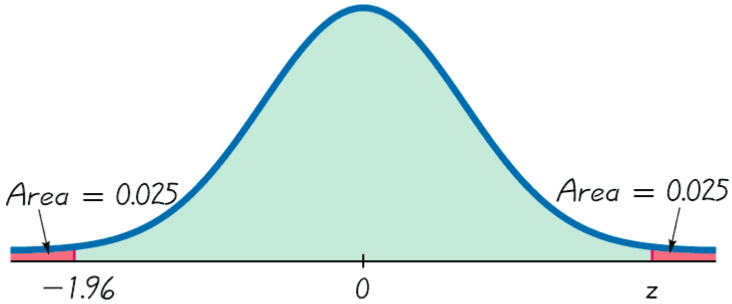
Finding z Scores When Given Probabilities - cont



(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

Finding *z* Scores When Given Probabilities - cont

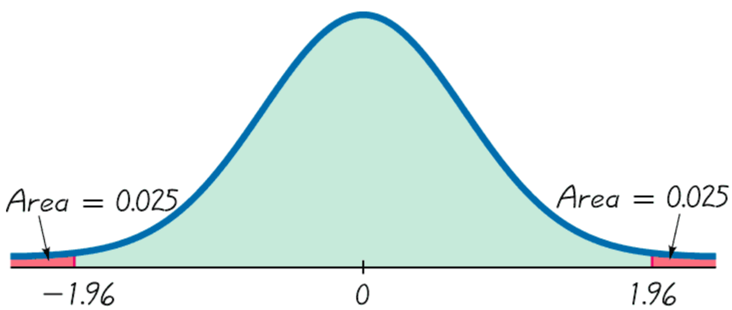


(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

26

Finding *z* Scores When Given Probabilities - cont



(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

Example

Page 250, problem 14

Example

ANSWER:

Area to the left of *z* is 0.2456 and from Table A-2 we get that

$$\boxed{z = -0.66}$$

Example

Page 250, problem 15

28

Example

ANSWER:

Area to the right of *z* is 0.1075

Area to the left of *z* is 1-0.1075=0.8925

From Table A-2 we get that

$$\boxed{z = 1.24}$$

Example

Page 251, problem 40

Use the standard normal distribution to answer this question:

About ______% of the area is between

z=-3.5 and z=3.5

(or what percent of the area is within 3.5 standard deviations of the mean?)

29

Example

Page 251, problem 40

$$
\begin{aligned}
P(-3.5 < z < 3.5) &= P(z < 3.5) - P(z < -3.5) \\
&= 0.9999 - 0.001 \\
&= 0.9998
\end{aligned}
$$

ANSWER: $\boxed{\textbf{99.98\%}}$

Example

Page 251, problem 42

- **Uses the notation:** Z_α **is the <u>critical value</u>**
- **Which means the z-score with an area of "alpha" (α) to its <u>right</u>**
- **To find** Z_α **find the z-score that corresponds to an area of** $1-\alpha$

Example

Page 251, problem 42

ANSWER: $Z_{0.01}$

The cumulative area from the left is:

1-0.01=0.9900

Table A-2 (next slide)

TABLE A-2 (continued) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495 *	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961

Closest value in the <u>body</u> of the table to 0.9900 is 0.9901 and the Corresponding z value is Z=2.33

31

Example

Page 251, problem 42

ANSWER:

$$z_{0.01} = \boxed{2.33}$$

Example

Page 251, problem 48

Use the standard normal distribution to find:

$$P(z < -1.96 \text{ or } z > 1.96)$$

32

Example

Page 251, problem 48

ANSWER: use the addition (“or”) rule for independent events

$$\begin{aligned} &P(z < -1.96 \text{ or } z > 1.96) \\ &\qquad = P(z < -1.96) + \mathrm{P}(z > 1.96) \\ &\qquad = 0.0250 + (1 - 0.9750) \\ &\qquad = \boxed{0.050} \end{aligned}$$

Example

Page 251, problem 50

Find the 1st percentile (P_1) separating the bottom 1% from the top 99% using the standard normal distribution.

Example

Page 251, problem 50

ANSWER:

This is the z-value whose cumulative area (area to the left of z) is 0.01

Directly from the table, the z-value that has cumulative area from the left closest to 0.01 is:

$$z = \boxed{-2.33}$$

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08
−3.50 and lower	.0001								
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051 *	.0049
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146

Closest value in the body of the table to 0.01 is 0.0099
Corresponding z value is -2.33

Example

Page 251, problem 54

If a continuous uniform distribution has mean 0 and standard deviation 1, then the minimum is $-\sqrt{3}$ and the maximum is $\sqrt{3}$

a) For this distribution find

$$P(-1 < x < 1)$$

NOTE: we use x for the random variable instead of z here

Example

Page 251, problem 54

ANSWER:

We first need to find the <u>height</u> of the uniform distribution which (recall) has a rectangular shape.

35

Example

Page 251, problem 54

ANSWER:

- **Fact: total area under the curve of a continuous probability distribution must equal 1 and the "curve" for the uniform distribution is a horizontal line so that the shape is rectangular**
- **We are told that this uniform distribution is a rectangle of width:**

$$\sqrt{3} - (-\sqrt{3}) = 2\sqrt{3}$$

Example

Page 251, problem 54

ANSWER:

- **rectangle area = (width)(height)**

$$1 = \left(2\sqrt{3}\right)(\text{height})$$

- **Solve for height:**

$$\text{height} = 1 \div 2\sqrt{3} \approx 0.2887$$

Example

Page 251, problem 54

If a continuous uniform distribution has mean 0 and standard deviation 1, then the minimum is $-\sqrt{3}$ **and the maximum is** $\sqrt{3}$

a)

$$P(-1 < x < 1) = (\text{width})(\text{height})$$
$$= (2)(0.2887)$$
$$= \boxed{0.5774}$$

Example

Page 251, problem 54

b) Find $P(-1 < x < 1)$ **if you incorrectly assume that the distribution is <u>normal</u> (not uniform).**

37

Example

Page 251, problem 54

b) ANSWER: here we use Table A-2 to get the answer

$$P(-1 < z < 1) = P(z < 1) - P(z < -1)$$
$$= 0.8413 - 0.1587$$
$$= \boxed{0.6826}$$

Example

Page 251, problem 54

c) Compare the results from parts (a) and (b). Does the distribution affect the results very much?

Example

Page 251, problem 54

c) Compare the results from parts (a) and (b). Does the distribution affect the results very much?

ANSWER:

Yes since 0.6826-0.5774 = 0.1052 is a 10.52% difference in the probability predictions.

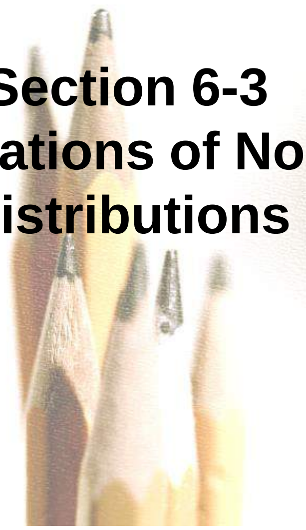
Recap

In this section we have discussed:

- ❖ Density curves.
- ❖ Relationship between area and probability.
- ❖ Standard normal distribution.
- ❖ Using Table A-2.

Lecture-14

Section 6-3 Applications of Normal Distributions



Key Concept

This section presents methods for working with normal distributions that are not standard. That is, the mean is not 0 or the standard deviation is not 1, or both.

The key concept is that we can use a simple conversion that allows us to standardize any normal distribution so that the same methods of the previous section can be used.

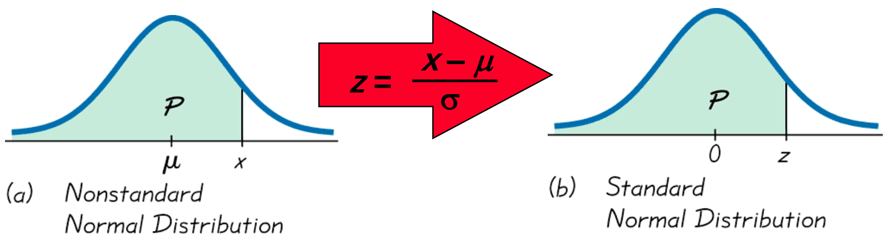
40

Conversion Formula 6-2

$$z = \frac{x - \mu}{\sigma}$$

Round *z* scores to 2 decimal places

Converting to a Standard Normal Distribution



41

Example – Weights of Water Taxi Passengers

In the Chapter Problem, we noted that the safe load for a water taxi was found to be 3500 pounds. We also noted that the mean weight of a passenger was assumed to be 140 pounds. Assume the worst case that all passengers are men. Assume also that the weights of the men are normally distributed with a mean of 172 pounds and standard deviation of 29 pounds. If one man is randomly selected, what is the probability he weighs less than 174 pounds?

Example - cont

Use the given mean and standard deviation values:

$$\mu = 172 \qquad \sigma = 29$$

to compute a z-score:

$$z = \frac{174 - 172}{29} = 0.07$$

Use this z-score and Table A-2 to find the answer:

$$\boxed{0.5279}$$

Example - cont

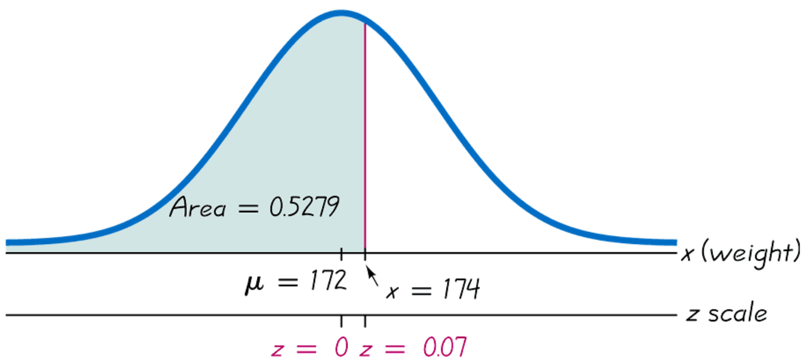
Cumulative area to the left of 0.07 is 0.5279

TABLE A-2 (continued) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224

Example - cont

PICTURE: **P (*x* < 174 lb.) = P(*z* < 0.07) = 0.5279**



43

Helpful Hints

1. **Don't confuse *z* scores and areas**. ***z* scores are distances along the horizontal scale, but areas are regions under the normal curve. Table A-2 lists *z* scores in the left column and across the top row, but areas are found in the body of the table.**
2. **Choose the correct (right/left) side of the graph.**
3. **A z score must be negative whenever it is located in the left half of the normal distribution.**
4. **Areas (or probabilities) are positive or zero values, but they are never negative.**

Using Formula 6-2

❖ **Sometimes we need to find the value of x that corresponds to a given value of z in the z-score formula 6-2. This can be accomplished with a small bit of algebra:**

$$z = \frac{x-\mu}{\sigma}$$

$$\Rightarrow z \cdot \sigma = x - \mu \qquad \text{(multiply by } \sigma\text{)}$$

$$\Rightarrow \boxed{x = \mu + z \cdot \sigma} \qquad \text{(add } \mu\text{)}$$

Procedure for Finding Values Using Table A-2 and Formula 6-2

1. **Sketch a normal distribution curve, enter the given probability or percentage in the appropriate region of the graph, and identify the *x* value(s) being sought.**

2. **Use Table A-2 to find the *z* score corresponding to the cumulative left area bounded by *x*. Refer to the body of Table A-2 to find the closest area, then identify the corresponding *z* score.**

3. **Using Formula 6-2, enter the values for μ, σ, and the *z* score found in step 2, then solve for *x*.**

   $x = \mu + (z \bullet \sigma)$ **(Another form of Formula 6-2)**

   (If *z* is located to the left of the mean, be sure that it is a negative number.)

4. **Refer to the sketch of the curve to verify that the solution makes sense in the context of the graph and the context of the problem.**

Example – Lightest and Heaviest

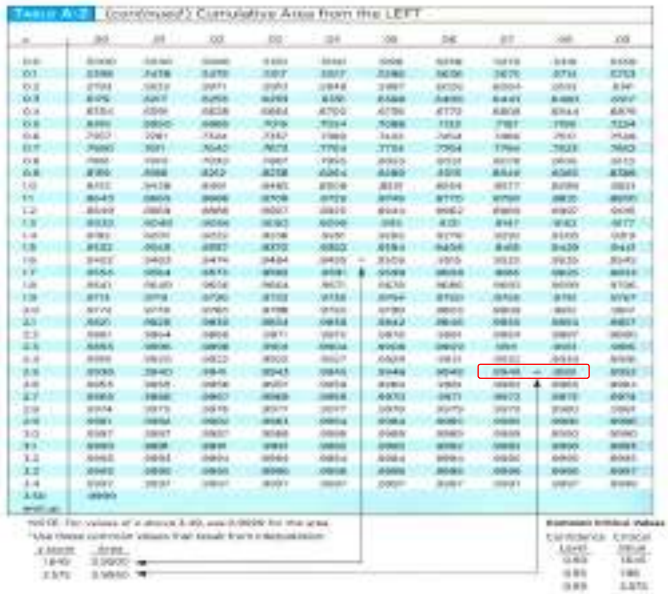
Use the data from the previous example to determine what weight separates the lightest 99.5% from the heaviest 0.5%?

Example – Lightest and Heaviest

Here the z-score corresponds to a cumulative area of 0.9950 to the left of z.

That is, 99.5% of the area is to the left of this z-score in the standard normal distribution.

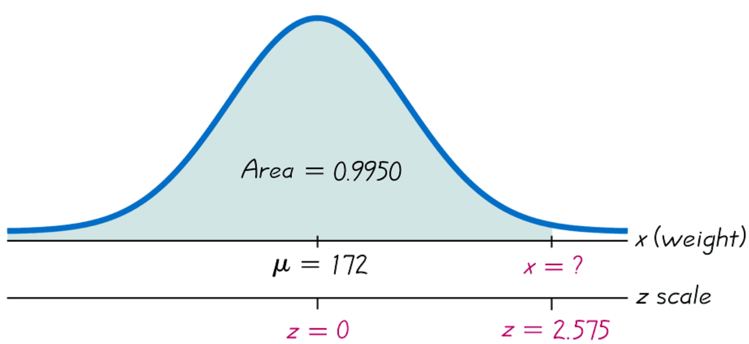
Use Table A-2 to get a z-score of 2.575 (see next slide)



NOTE: in the body of the table, 0.9950 is midway between 0.9949 and 0.9951

46

Example –
Lightest and Heaviest - cont



Example – Lightest and Heaviest

Now compute the x value using the previous example with values for mean (172 pounds) and standard deviation (29 pounds) and the z-score that we found 2.575

$$2.575 = \frac{x - 172}{29}$$

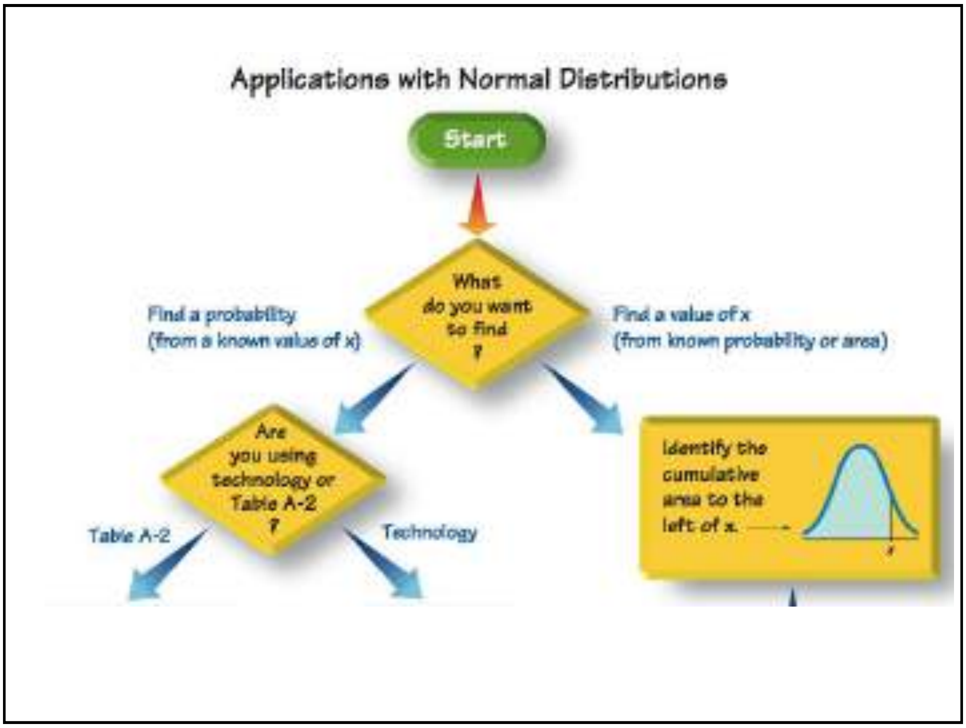
$$\Rightarrow 74.675 = x - 172 \qquad \text{(multiply by 29)}$$

$$\Rightarrow x = 74.675 + 172 \qquad \text{(add 172)}$$

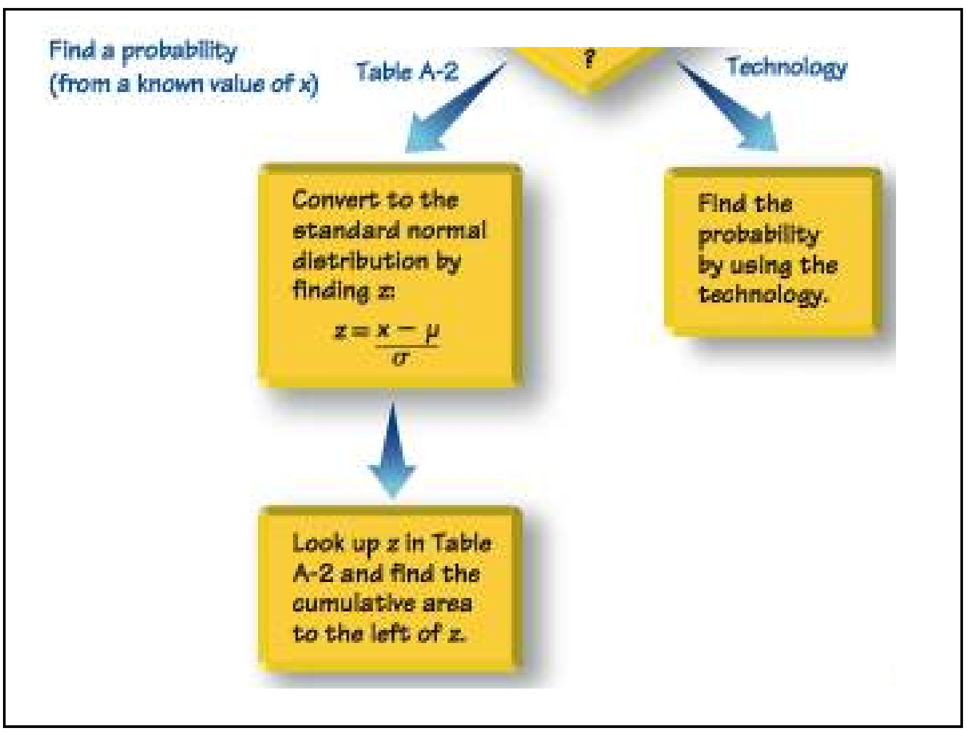
$$\Rightarrow x = 246.675 \approx \boxed{247 \text{ pounds}}$$

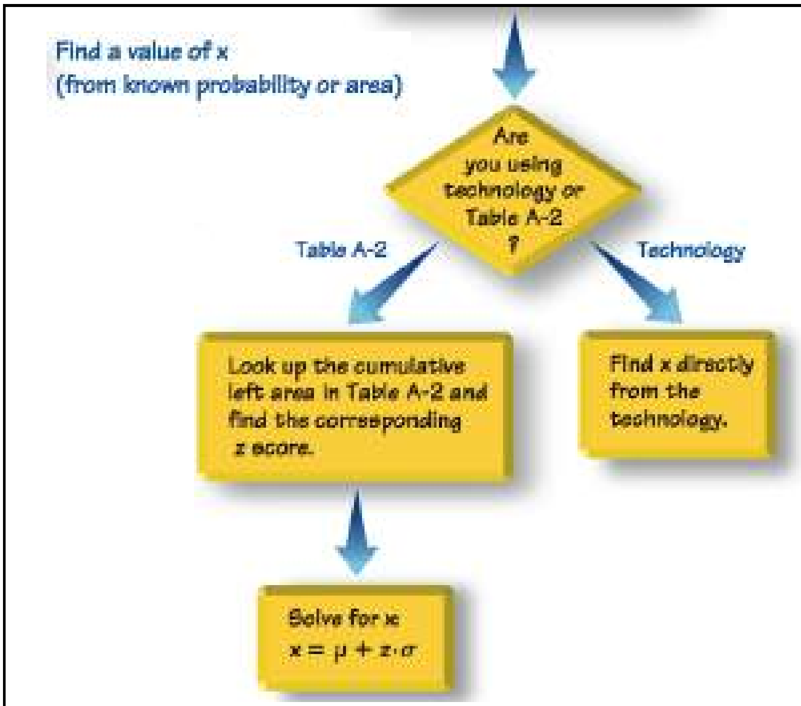
Example – Lightest and Heaviest - cont

The weight of 247 pounds separates the lightest 99.5% from the heaviest 0.5%


Applications with Normal Distributions
Start
What do you want to find ?
Find a probability (from a known value of x)
Find a value of x (from known probability or area)
Are you using technology or Table A-2 ?
Table A-2
Technology
Identify the cumulative area to the left of x.

48


Find a probability (from a known value of x)
Table A-2
Technology
Convert to the standard normal distribution by finding z:
z = (x − μ)/σ
Find the probability by using the technology.
Look up z in Table A-2 and find the cumulative area to the left of z.



Example

Page 261, problem 20

For problems 13-20, assume that adults have IQ scores that are normally distributed with a mean of 100 and a standard deviation of 15.

Find the IQ score separating the top 37% from the others.

Example

Page 261, problem 20

<u>ANSWER</u>

First find the z-score for the <u>top</u> 37%.

Use Table A-2 to find z so that the cumulative area to the <u>right</u> of z is 37% or 0.37

That means that the cumulative area to the <u>left</u> of z is: 1-0.37=0.63 and

z=0.33

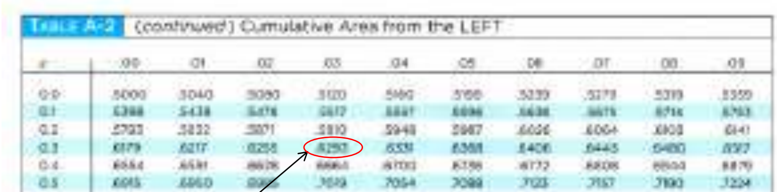


TABLE A-2 (continued) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224

Cumulative area that is closest to 0.63 in the body of the table is 0.6293

Corresponding z-value is 0.33

Example

Page 261, problem 20

ANSWER

Next compute the corresponding x-value using the z-value and given

$$\mu = 100 \quad \text{and} \quad \sigma = 15$$

$$x = \mu + z\sigma = 100 + (0.33)(15) = 104.95 \approx \boxed{105.0}$$

That is, 37% of all IQ scores are 105.0 or higher.

Example

Page 261, problem 23

For problems 21-26, assume:

- **Men's heights are normally distributed with mean 69.0 inches and standard deviation 2.8 inches.**
- **Women's heights are normally distributed with mean 63.6 inches and standard deviation 2.5 inches.**

Example

Page 261, problem 23(a)

Tall Clubs International is a social organization for tall people. It has a requirement that men must be at least 74 inches tall and women must be at least 70.0 inches tall.

(a) What percentage of men meet that requirement?

(NOTE: Students do parts (b) and (c) for homework)

52

Example

Page 261, problem 23(a)

ANSWER:

(a) We must find $P(x > 74\,\text{inches})$

given that $\mu = 69.0\,\text{inches}$ and $\sigma = 2.8\,\text{inches}$

First, convert x to a z-score using the formula 6-2:

$$z = \frac{x - \mu}{\sigma}$$

Example

Page 261, problem 23(a)

ANSWER:

(a) (cont.)

$$z = \frac{74.0 - 69.0}{2.8} = \frac{5.0}{2.8} = 1.7875 \approx 1.79$$

(Note: z will never have dimensions like inches etc.)

And we must find

$$P(x > 74) = P(z > 1.79)$$

53

Example

Page 261, problem 23(a)

ANSWER:

(a) (cont.)

$$\begin{aligned} P(z > 1.79) &= 1 - P(z < 1.79) \\ &= 1 - 0.9633 \\ &= 0.0367 \end{aligned}$$

From body of Table A-2

Therefore, ANSWER is 3.67%

Example

Page 261, problem 26

The US Marine Corps requires that men have heights between 64 inches and 80 inches.

(a) Find the percentage of men who meet the height requirements.

(b) If the height requirements are changed so that all men are eligible except the shortest 3% and the tallest 4%, what are the new height requirements?

Example

Page 261, problem 26

ANSWER

(a) We must find

$$P(64\text{ inches} < x < 80\text{ inches})$$

After using the formula 6-2 to convert x values to z values:

$$P(64 < x < 80) = P(-1.79 < z < 3.92)$$

Example

Page 261, problem 26

ANSWER

(a) (cont.)

$$P(-1.79 < z < 3.92) = P(z < 3.79) - P(z < -1.79)$$
$$= 0.9999 - 0.0367$$
$$= 0.9632 = \boxed{96.32\%}$$

55

Example

Page 261, problem 26

The US Marine Corps requires that men have heights between 64 inches and 80 inches.

(b) If the height requirements are changed so that all men are eligible except the shortest 3% and the tallest 4%, what are the new height requirements?

Example

Page 261, problem 26

ANSWER:

(b) We must find the x-values at which 3% of the area is <u>below</u> x and 4% of the area is <u>above</u> x.

From the body of the table A-2 where cumulative area to the left is 3% = 0.0300 we get

z = -1.88

Example

Page 261, problem 26

ANSWER:

(b) Using the Formula 6-2 for x when

$$\mu = 69.0 \text{ inches} \quad \text{and} \quad \sigma = 2.8 \text{ inches}$$

and z=-1.88 gives

$$x = \mu + z\sigma = 69.0 + (-1.88)(2.8) \approx 63.7 \text{ inches}$$

Example

Page 261, problem 26

ANSWER:

(b) From the body of the table A-2 where cumulative area to the right is 4% = 0.0400 so that the cumulative area to the left is 1-0.04=0.9600 we get

z = 1.75

57

Example

Page 261, problem 26

ANSWER:

(b) Using the Formula 6-2 for x when

$$\mu = 69.0\,\text{inches} \quad \text{and} \quad \sigma = 2.8\,\text{inches}$$

and z=1.75 gives

$$x = \mu + z\sigma = 69.0 + (1.75)(2.8) \approx 73.9\,\text{inches}$$

Example

Page 261, problem 26

ANSWER:

(b) New height requirements are 63.7 inches to 73.9 inches.

Example

Page 262, problem 31

The lengths of pregnancies are normally distributed with a mean of 268 days and standard deviation of 15 days.

(a) (From Dear Abby letter) A wife claimed to have given birth 308 days after a brief visit from her husband, who was serving in the Navy. Find the probability of a pregnancy lasting 308 days or longer. What does the result suggest?

Example

Page 262, problem 31

ANSWER:

(a) We must find $P(x > 308\,\text{days})$

given that $\mu = 268\,\text{days}$ and $\sigma = 15\,\text{days}$

First, convert x to a z-score using the formula 6-2:

$$z = \frac{x - \mu}{\sigma}$$

Example

Page 262, problem 31

ANSWER:

(a) we get $P(x > 308) = P(z > 2.67)$

$$= 1 - P(z < 2.67)$$
$$= 1 - 0.9962$$
$$= 0.0038$$

it would be unusual for this to occur since probability of occurring is small (only 38 out of every 10000 pregnancies last longer than 308 days)

Example

Page 262, problem 31

The lengths of pregnancies are normally distributed with a mean of 268 days and standard deviation of 15 days.

(b) If a baby is premature if the length of the pregnancy is in the lowest 4%, find the length that separates premature babies from those who are not premature.

Example

Page 262, problem 31

ANSWER:

(b) We must find the x-value at which 4% of the area is <u>below</u> x

From the body of the table A-2 where cumulative area to the left is 4% = 0.0400 we get

$$z = -1.75$$

Example

Page 262, problem 31

ANSWER:

(b) Using the Formula 6-2 for x when

$$\mu = 268 \text{ days} \quad \text{and} \quad \sigma = 15 \text{ days}$$

and z=-1.75 gives

$$x = \mu + z\sigma = 268 + (-1.75)(15) \approx \boxed{242 \text{ days}}$$

That is, a baby is considered premature if it is born on or before the 242^{nd} day or 34.6^{th} week of a woman's pregnancy.

Recap

In this section we have discussed:

- **Non-standard normal distribution.**
- **Converting to a standard normal distribution.**
- **Procedures for finding values using Table A-2 and Formula 6-2.**

Lecture-15

Section 6-4
Sampling Distributions and Estimators

62

Key Concept

The main objective of this section is to understand the concept of a **sampling distribution of a statistic**, which is the distribution of all values of that statistic when all possible samples of the same size are taken from the same population.

We will also see that some statistics are better than others for estimating population parameters.

Definition

The **sampling distribution of a statistic** (such as the sample mean or sample proportion) is the distribution of all values of the statistic when all possible samples of the same size *n* are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

63

NOTATION

- Sample Mean is $\bar{x}$
- Sample Standard Deviation is s
- Population Mean is μ
- Population Standard Deviation is σ

Definition

The sampling distribution of the mean is the distribution of sample means, with all samples having the same sample size *n* taken from the same population. (The sampling distribution of the mean is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Properties

- Sample means target the value of the population mean. (That is, the mean of the sample means is the population mean. The expected value of the sample mean is equal to the population mean.)
- The distribution of the sample means tends to be a normal distribution.

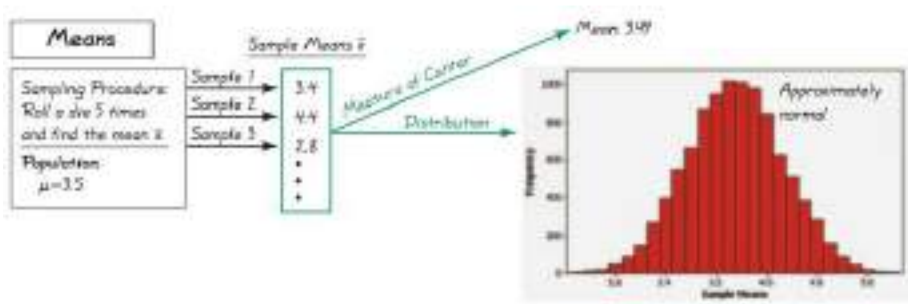
Example - Sampling Distributions

Consider repeating this process: Roll a die 5 times, find the sample mean. Repeat this over and over. What do we know about the behavior of all sample means that are generated as this process continues indefinitely?

65

Example - Sampling Distributions

Specific results from 10,000 trials



All outcomes are equally likely so the population mean is 3.5; the mean of the 10,000 trials is 3.49. If continued indefinitely, the sample mean will be 3.5. Also, notice the distribution is "normal."

Definition

The sampling distribution of the variance is the distribution of sample variances, with all samples having the same sample size *n* taken from the same population. (The sampling distribution of the variance is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Properties

- **Sample variances target the value of the population variance. (That is, the mean of the sample variances is the population variance. The expected value of the sample variance is equal to the population variance.)**
- **The distribution of the sample variances tends to be a distribution skewed to the right.**

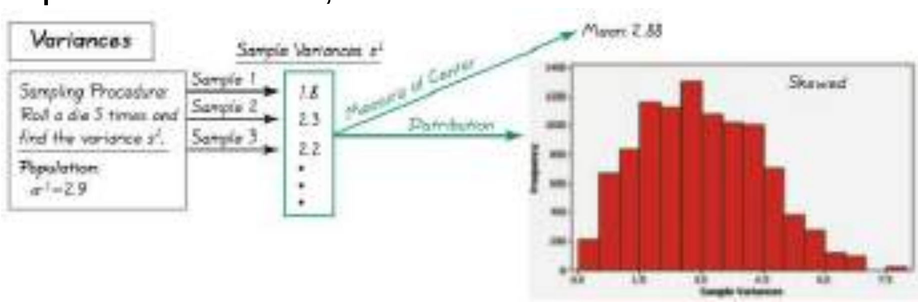
Example - Sampling Distributions

Consider repeating this process: Roll a die 5 times, find the variance. Repeat this over and over. What do we know about the behavior of all sample variances that are generated as this process continues indefinitely?

67

Example - Sampling Distributions

Specific results from 10,000 trials



All outcomes are equally likely so the population variance is 2.9; the mean of the 10,000 trials is 2.88. If continued indefinitely, the sample variance will be 2.9. Also, notice the distribution is “skewed to the right.”

Unbiased Estimators

Sample means and variances are unbiased estimators.

That is, they target the population parameter.

These statistics are good at estimating the population parameter.

Biased Estimators

Sample medians, ranges and standard deviations are biased estimators.

That is they do NOT target the population parameter.

Note: the bias with the standard deviation is relatively small in large samples so *s* is often used to estimate the population standard deviation.

Why Sample with Replacement?

Sampling *without replacement* would have the very practical advantage of avoiding wasteful duplication whenever the same item is selected more than once. However, we are interested in sampling *with replacement* for these two reasons:

1. When selecting a relatively small sample form a large population, it makes no significant difference whether we sample with replacement or without replacement.
2. Sampling with replacement results in independent events that are unaffected by previous outcomes, and independent events are easier to analyze and result in simpler calculations and formulas.

Caution

Many methods of statistics require a *simple random sample*. Some samples, such as voluntary response samples or convenience samples, could easily result in very wrong results.

Example

Page 274, problem 12

In example 4 of the book it was assumed that samples were of size 2,3, and 10 representing the numbers of people in households. Table 6-4 lists the 9 different possible samples (see next slide)

(a) Find the mean of each of the nine samples and summarize the sampling distribution of the means in the format of a table representing the probability distribution. We will assume here that the order of the sample matters (2,3 is different than 3,2)

Example

Example 4 possible samples and means

Sample	Mean of Sample ($\bar{x}$)
2,2	2.0
2,3	2.5
2,10	6.0
3,2	2.5
3,3	3.0
3,10	6.5
10,2	6.0
10,3	6.5
10,10	10.0

$$\bar{x} = \frac{2+2}{2}$$

$$\bar{x} = \frac{2+3}{2}$$

$$\bar{x} = \frac{2+10}{2}$$

etc.

Example

ANSWER to part (a)

<u>probability distribution</u> of means is:

Mean $\bar{x}$	Probability $P(\bar{x})$
2.0	1/9
2.5	2/9
3.0	1/9
6.0	2/9
6.5	2/9
10.0	1/9

Example

Page 274, problem 12

(b) Compare the population mean to the mean of the sample means.

Example

(b) The population mean is:

$$\mu = \frac{2+3+10}{3} = 5.0$$

72

Example

mean of the sample means is also the expected value of the means:

Mean $\bar{x}$	Probability $P(\bar{x})$	$\bar{x} \cdot P(\bar{x})$
2.0	1/9	2/9
2.5	2/9	5/9
3.0	1/9	3/9
6.0	2/9	12/9
6.5	2/9	13/9
10.0	1/9	10/9

Add up last column: $\sum \bar{x} \cdot P(\bar{x}) = 45/9 = 5.0$

which is the mean of the sample means

Example

ANSWER:

(b) The population mean agrees with the mean of the sample means.

Example

Page 274, problem 12

(c) Do the sample means target the value of the population mean? In general, do sample means make good estimators of the population means? Why or why not.

Example

ANSWER:

(c) The sample mean always targets the population mean. For this reason, the sample mean is a good estimator of the population mean.

Example

Page 274, problem 10 and 11

Repeat problem 12 using the variance and standard deviation instead of the means.

(a) Find the variance and standard deviation of each of the nine samples and summarize the sampling distribution of these in the format of a table representing the probability distribution of each.

Sample Variance and Standard Deviation (Formula 3-4 pg. 101)

Variance:

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

Standard Deviation:

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

75

Example

First find the variances (see next slide):

Sample	Variance of the Sample (s^2)
2,2	0
2,3	0.5
2,10	32
3,2	0.5
3,3	0
3,10	24.5
10,2	32
10,3	24.5
10,10	0

Example

- For the first sample, the mean is $\bar{x} = 2.0$

$$s^2 = \frac{(2-2)^2 + (2-2)^2}{2-1} = \frac{0}{1} = 0$$

- For the second sample, the mean is $\bar{x} = 2.5$

$$s^2 = \frac{(2-2.5)^2 + (3-2.5)^2}{2-1} = \frac{0.5}{1} = 0.5$$

- For the third sample, the mean is $\bar{x} = 6.0$

$$s^2 = \frac{(2-6.0)^2 + (10-6.0)^2}{2-1} = \frac{32}{1} = 32.0$$

ETC.

Example

<u>probability distribution</u> of variance is:

Variance (s^2)	Probability $P(s^2)$
0.0	3/9
0.5	2/9
24.5	2/9
32	2/9

Example

Take square root of sample variances to get sample standard deviations:

Sample	Variance (s^2)	Standard Deviation (s)
2,2	0	0
2,3	0.5	0.707
2,10	32	5.657
3,2	0.5	0.707
3,3	0	0
3,10	24.5	4.950
10,2	32	5.657
10,3	24.5	4.950
10,10	0	0

Example

<u>probability distribution</u> of standard deviations

Standard Deviation (s)	Probability $P(s)$
0.0	3/9
0.707	2/9
4.950	2/9
5.657	2/9

Example

Page 274, problem 10, 11

(b) Compare the population variance and standard deviation to the mean of the sample variances and standard deviations.

78

Population Standard Deviation (Formula 3-4 pg. 103)

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

Note: here we use N not N-1 (see bottom of page 266)

Example

Page 274, problem 10, 11

(b) Population variance:

$$\sigma^2 = \frac{(2-5.0)^2 + (3-5.0)^2 + (10-5.0)^2}{3} = \frac{38}{3}$$

Population standard deviation:

$$\sigma = \sqrt{\frac{38.0}{3}} \approx 3.559$$

79

Example

mean of the sample variance is also the expected value of the variances:

Variance (s^2)	Probability $P(s^2)$	$s^2 \cdot P(s^2)$
0.0	3/9	0/9
0.5 = 1/2	2/9	1/9
24.5 = 49/2	2/9	49/9
32	2/9	64/9

Add up last column: $\sum s^2 \cdot P(s^2) = 114/9 = 38/3$

which is the <u>mean of the sample variances</u>

Example

mean of the sample standard deviations is also the expected value of the standard deviations:

Standard Deviation (s)	Probability $P(s)$	$s \cdot P(s)$
0.0	3/9	0.000
0.707	2/9	0.157
4.950	2/9	1.100
5.657	2/9	1.257

Add up last column: $\sum s \cdot P(s) = 2.514$

which is the <u>mean of the sample standard deviations</u>

Example

ANSWER:

(b) The population variance agrees with mean of the sample variances.

The population standard deviation does <u>not</u> agree with mean of the sample standard deviations.

Example

Page 274, problem 10, 11

(c) Do the sample variances and standard deviations target the value of the population variances and standard deviations? In general, do sample variances and standard deviations make good estimators of the population variances and standard deviations ? Why or why not.

Example

ANSWER:

(c) The population variance agrees with the mean of the sample variances. In general, the sample variances target the value of the population variances and the sample variance is a good estimator of the population variance.

Example

ANSWER:

(c) The population standard deviation does <u>not</u> agree with the mean of the sample standard deviations. In general, the sample standard deviations do not target the value of the population standard deviation and the sample standard deviation is <u>not</u> a good estimator of the population standard deviation.

Standard Deviation For Large Samples

The bias with the standard deviation is relatively small in <u>large</u> samples so *s* is often used to estimate the population standard deviation σ when the sample is large.

Definition

The sampling distribution of the proportion is the distribution of sample proportions, with all samples having the same sample size n taken from the same population.

83

Definition

We need to distinguish between a population proportion p and some sample proportion:

p = population proportion

$\hat{p}$ = sample proportion

Properties

- Sample proportions target the value of the population proportion. (That is, the mean of the sample proportions is the population proportion. The expected value of the sample proportion is equal to the population proportion.)
- The distribution of the sample proportion tends to be a normal distribution.

84

Unbiased Estimators

Sample proportions are unbiased estimators.

That is they target the population parameter.

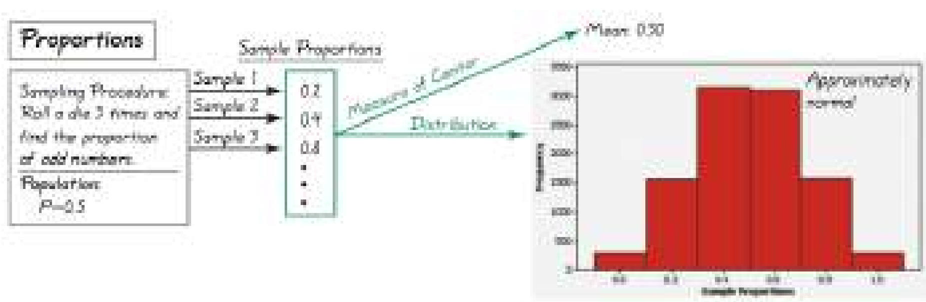
Example - Sampling Distributions

Consider repeating this process: Roll a die 5 times, find the proportion of *odd* numbers of the results. Repeat this over and over. What do we know about the behavior of all sample proportions that are generated as this process continues indefinitely?

85

Example - Sampling Distributions

Specific results from 10,000 trials



All outcomes are equally likely so the population proportion of odd numbers is 0.50; the proportion of the 10,000 trials is 0.50. If continued indefinitely, the mean of sample proportions will be 0.50. Also, notice the distribution is “approximately normal.”

Example

Page 275, problem 20

After constructing a new manufacturing machine, 5 prototype integrated circuit chips are produced and it is found that 2 are defective (D) and 3 are acceptable (A). Assume that two of the chips are randomly selected with replacement from this population.

(a) After identifying the 25 different possible samples, find the proportion of defects in each of them, then use a table to describe the sampling distribution of the proportion of defects.

Example

Page 275, problem 20

NOTE: for this problem we need to identify each chip separately. Denote the two defective chips as x and y and the three acceptable chips as a, b, and c. We consider here that order matters so that a sample of x,y is different than y,x

Sample	Proportion of Defects ($\hat{p}$)
x, x	1.0
x, y	1.0
x, a	0.5
x, b	0.5
x, c	0.5
y, x	1.0
y, y	1.0
y, a	0.5
y, b	0.5
y, c	0.5

Sample	Proportion of Defects ($\hat{p}$)
a, x	0.5
a, y	0.5
a, a	0.0
a, b	0.0
a, c	0.0
b, x	0.5
b, y	0.5
b, a	0.0
b, b	0.0
b, c	0.0
c, x	0.5
c, y	0.5
c, a	0.0
c, b	0.0
c, c	0.0

87

Example

probability distribution of sample proportions

Proportion ($\hat{p}$)	Probability $P(\hat{p})$
0.0	9/25
0.5	12/25
1.0	4/25

Example

Page 275, problem 20

(b) Find the mean of the sampling distribution

Example

Proportion ($\hat{p}$)	Probability $P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$
0.0	9/25	0
0.5	12/25	6/25
1.0	4/25	4/25

Add up last column: $\sum \hat{p} \cdot P(\hat{p}) = 10/25 = 2/5$

which is the <u>mean of the sample proportions</u>

Example

Page 275, problem 20

(c) Is the mean of the sampling distribution from part (b) equal to the population proportion of defects? Does the mean of the sampling distribution of proportions always equal the population proportion?

89

Example

Page 275, problem 20

(c) The population proportion of defectives is:

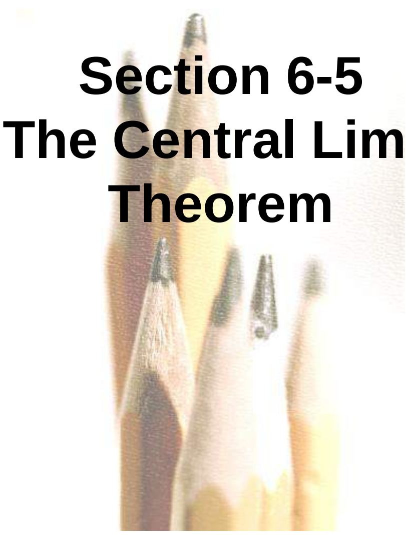
$$p = \frac{2}{5}$$

which is equal to the mean of the sampling distribution of proportions. The mean of the sampling distribution of proportions always <u>targets</u> the population proportion.

Recap

In this section we have discussed:

- **Sampling distribution of a statistic.**
- **Sampling distribution of the mean.**
- **Sampling distribution of the variance.**
- **Sampling distribution of the proportion.**
- **Estimators.**



Section 6-5
The Central Limit Theorem

Key Concept

The *Central Limit Theorem* tells us that for a population with *any* distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

The procedure in this section form the foundation for estimating population parameters and hypothesis testing.

Central Limit Theorem

Given:

1. The random variable x has a distribution (which may or may not be normal) with mean μ and standard deviation σ.
2. Simple random samples all of size n are selected from the population. (The samples are selected so that all possible samples of the same size n have the same chance of being selected.)

Central Limit Theorem – cont.

Conclusions:

1. **The distribution of sample $\bar{x}$ will, as the sample size increases, approach a normal distribution.**
2. **The mean of the sample means is the population mean μ.**
3. **The standard deviation of all sample means is $\sigma/\sqrt{n}$.**

92

Practical Rules Commonly Used

1. **For samples of size *n* larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets closer to a normal distribution as the sample size *n* becomes larger.**
2. **If the original population is *normally distributed*, then for any sample size *n*, the sample means will be normally distributed (not just the values of *n* larger than 30).**

Notation

the mean of the sample means

$$\mu_{\bar{x}} = \mu$$

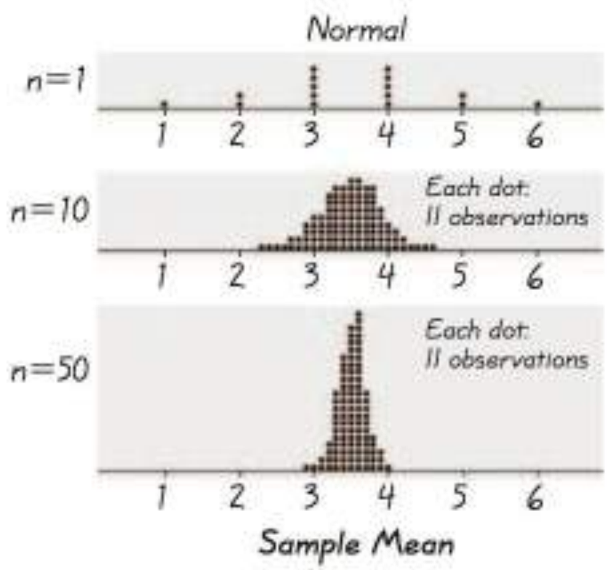
the standard deviation of sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(often called the standard error of the mean)

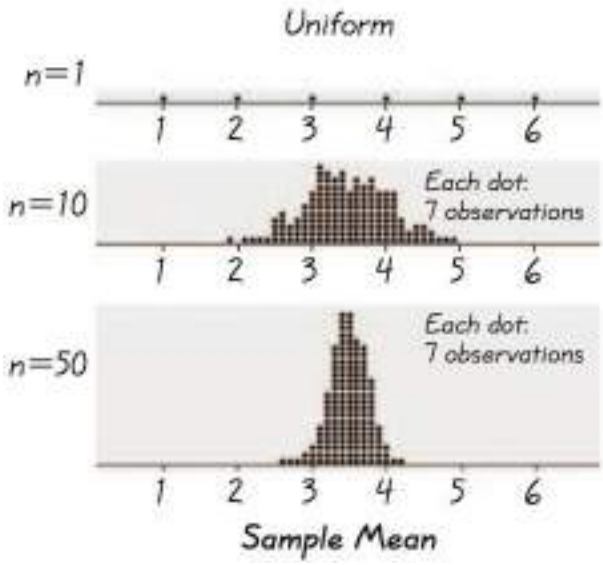
Example - Normal Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



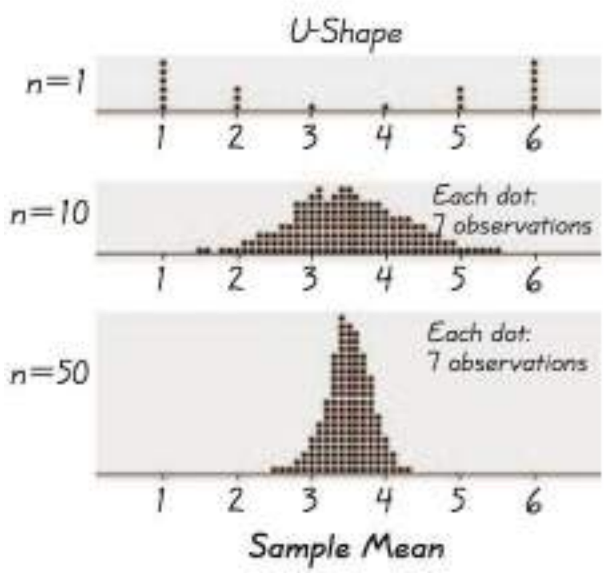
Example - Uniform Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



Example - U-Shaped Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



Important Point

As the sample size increases, the sampling distribution of sample means approaches a normal distribution.

Example – Water Taxi Safety

Use the Chapter Problem. Assume the population of weights of men is normally distributed with a mean of 172 lb and a standard deviation of 29 lb.

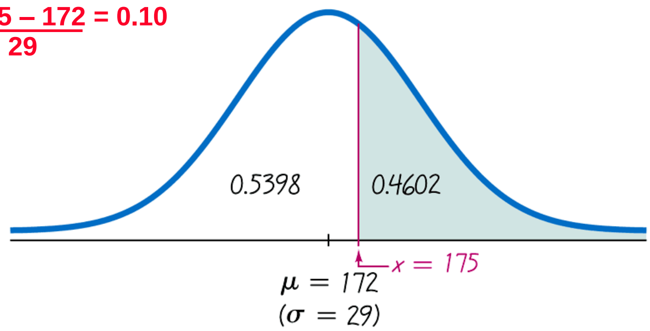
a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

b) b) Find the probability that *20* randomly selected men will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

Example – cont

a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

$$z = \frac{175 - 172}{29} = 0.10$$



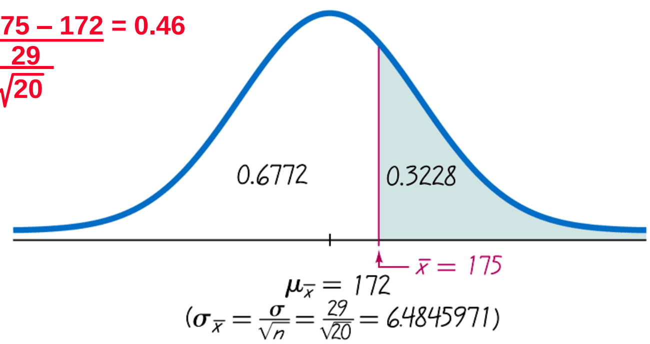
(a)

96

Example – cont

b) Find the probability that *20 randomly selected men* will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

$$z = \frac{175 - 172}{\frac{29}{\sqrt{20}}} = 0.46$$



(b)

Example - cont

a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

$$P(x > 175) = 0.4602$$

b) Find the probability that *20 randomly selected men* will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

$$P(\bar{x} > 175) = 0.3228$$

It is much easier for an individual to deviate from the mean than it is for a group of 20 to deviate from the mean.

Interpretation of Results

Given that the safe capacity of the water taxi is 3500 pounds, there is a fairly good chance (with probability 0.3228) that it will be overloaded with 20 randomly selected men.

Example

Page 284, problem 8

Assume SAT scores are normally distributed with mean $\mu = 1518$ and standard deviation $\sigma = 325$.

(a) If 1 SAT score is randomly selected, find the probability that it is between 1440 and 1480.

Example

ANSWER

(a) If the random variable x represents a randomly selected SAT score, we must find

$$P(1440 < x < 1480)$$

After using the formula 6-2 to convert x values to z values (see next slide):

$$P(1440 < x < 1480) = P(-0.24 < z < -0.12)$$

Example

<u>ANSWER</u>

Next compute the corresponding z-values using the x-values and

$$\mu = 1518 \quad \text{and} \quad \sigma = 325$$

x=1440

$$z = \frac{1440 - 1518}{325} = -0.24$$

x=1480

$$z = \frac{1480 - 1518}{325} = -0.12$$

Example

(a) (cont.) Use Table A-5 to find the answer

$$\begin{aligned} P(-0.24 < z < -0.12) &= P(z < -0.12) - P(z < -0.24) \\ &= 0.4522 - 0.4052 \\ &= \boxed{0.0470} \end{aligned}$$

Example

Page 284, problem 8

Assume SAT scores are normally distributed with mean $\mu = 1518$ and standard deviation $\sigma = 325$.

(b) If 16 SAT scores are randomly selected, find the probability that they have a mean between 1440 and 1480.

Example

ANSWER

(b) If the random variable $\bar{x}$ represents the mean of 16 randomly selected SAT scores, we must find

$$P(1440 < \bar{x} < 1480)$$

After using the formula 6-2 to convert $\bar{x}$ value to z value (see following slides):

$$P(1440 < \bar{x} < 1480) = P(-0.96 < z < -0.47)$$

Example

ANSWER

There are 16 randomly selected SAT scores and the original distribution is a normal distribution, we use the Central Limit Theorem to get

$$\mu_{\bar{x}} = \mu = 1518$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 325 / \sqrt{16} = 81.25$$

Example

ANSWER

Next compute the corresponding z-values using the $\bar{x}$-values and

$$\mu = 1518 \quad \text{and} \quad \sigma = 81.25$$

- ***$\bar{x}$=1440***

$$z = \frac{1440 - 1518}{81.25} = -0.96$$

- ***$\bar{x}$=1480***

$$z = \frac{1480 - 1518}{81.25} = -0.47$$

Example

(a) (cont.) Use Table A-5 to find the answer

$$P(1440 < \bar{x} < 1480) = P(-0.96 < z < -0.47)$$
$$= P(z < -0.47) - P(z < -0.96)$$
$$= 0.3192 - 0.1685$$
$$= \boxed{0.1507}$$

102

Example

Page 284, problem 8

Assume SAT scores are normally distributed with mean $\mu = 1518$ and standard deviation $\sigma = 325$.

(c) Why can the central limit theorem be used in part (b), even though the sample size does not exceed 30?

Example

Page 284, problem 8

(c) ANSWER: the original distribution is a normal distribution

Example

Page 285, problem 12

Assume the lengths of pregnancies are normally distributed with mean 268 days and standard deviation 15 days.

(a) If 1 pregnant woman is randomly selected, find the probability that her length of pregnancy is less than 260 days.

Example

ANSWER

(a) If the random variable x represents a randomly selected pregnancy length, we must find

$$P(x < 260 \text{ days})$$

After using the formula 6-2 to convert x value to z value (see next slide):

$$P(x < 260) = P(z < -0.53)$$

Example

ANSWER

Next compute the corresponding z-value using the x-value and

$$\mu = 268 \text{ days} \quad \text{and} \quad \sigma = 15 \text{ days}$$

- ***x=260 days***

$$z = \frac{260 - 268}{15} = -0.53$$

Example

(a) (cont.) Use Table A-5 to find the answer

$$P(z < -0.53) = \boxed{0.2981}$$

Example

Page 285, problem 12

Assume the lengths of pregnancies are normally distributed with mean 268 days and standard deviation 15 days.

(b) If 25 pregnant woman are randomly selected and put on a special diet just before they become pregnant, find the probability that the lengths of pregnancy have a mean that is less than 260 days (assuming the diet has no effect).

Example

ANSWER

(b) If the random variable $\bar{x}$ represents the mean of 25 randomly selected pregnancy lengths, we must find

$$P(\bar{x} < 260 \text{ days})$$

After using the formula 6-2 to convert $\bar{x}$ value to z value (see following slides):

$$P(\bar{x} < 260) = P(z < -2.67)$$

106

Example

<u>ANSWER</u>

There are 25 randomly selected pregnancies and the original distribution is a normal distribution, we use the Central Limit Theorem with

$$\mu_{\bar{x}} = \mu = 268$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 15 / \sqrt{25} = 3$$

Example

ANSWER

For $\bar{x}$=*260 days* this gives:

$$z = \frac{260 - 268}{3} = -2.67$$

Then:

$$P(\bar{x} < 260) = P(z < -2.67) = \boxed{0.0038}$$

Example

Page 285, problem 12

(c) If the 25 women do have a mean of less than 260 days, does it appear that the diet has an effect on the length of the pregnancy, and should the medical supervisors be concerned?

Example

Page 285, problem 12

(c) ANSWER: yes, it is very unlikely to experience a mean that low (from part (b) since 0.0038<0.05) by chance alone (that is, we assumed the diet had no effect to get the answer in part (b)), and the effects of the diet on the pregnancy should be a matter of concern.

Example

Page 286, problem 20

Assume the population of human body temperatures has a mean of 98.6 deg. F, as is commonly agreed. Assume the population standard deviation is 0.62 deg. F. If a sample size of n=106 is randomly selected, find the probability of getting a mean temperature of 98.2 deg. F or lower. Does that probability suggest that the mean body temp. is not 98.6 deg. F.?

Example

ANSWER

(a) If the random variable $\bar{x}$ represents the mean of 106 randomly selected body temperatures, find:

$$P(\bar{x} < 98.2^o)$$

After using the formula 6-2 to convert $\bar{x}$ value to z value (see next slides):

$$P(\bar{x} < 98.2^o) = P(z < -6.67)$$

109

Example

ANSWER

There are 106 randomly selected body temperatures and we use the Central Limit Theorem because 106>30 (note: we are not told that the original distribution is a normal distribution)

$$\mu_{\bar{x}} = \mu = 98.6^o\,\mathrm{F}$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 0.62 / \sqrt{106} = 0.06^o\,\mathrm{F}$$

Example

ANSWER

For *$\bar{x}$=98.2 deg. F.* this gives:

$$z = \frac{98.2 - 98.6}{0.06} = -6.67$$

Then:

$$P(\bar{x} < 98.2^o) = P(z < -6.67) = \boxed{0.0001}$$

Example

Page 286, problem 20

Does the probability suggest that the mean body temp. is not 98.6 deg. F.?

Example

Page 286, problem 20

ANSWER: yes, if 106 randomly selected body temperatures resulted in a mean of 98.2 deg. F or lower, that would be an extremely rare event based on the probability we computed. The conclusion here is that we used a population mean *μ = 98.6 deg F.* **that was not correct when we computed this probability.**

Correction for a Finite Population

When sampling without replacement and the sample size n is greater than 5% of the finite population of size N (that is, $n > 0.05N$), adjust the standard deviation of sample means by multiplying it by the *finite population correction factor*:

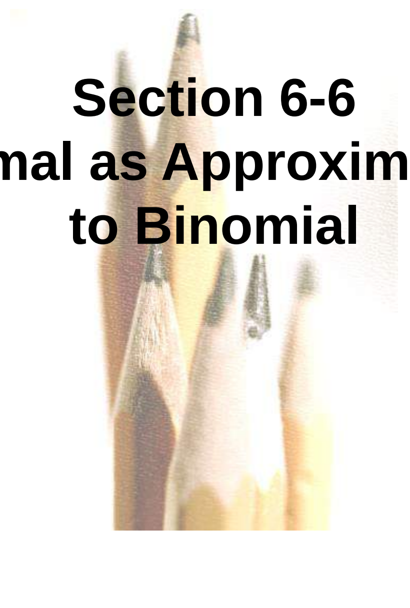
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\text{finite population correction factor}}$$

Recap

In this section we have discussed:

- **Central limit theorem.**
- **Practical rules.**
- **Effects of sample sizes.**
- **Correction for a finite population.**

Section 6-6 Normal as Approximation to Binomial



Review (chapter 5-3)

Binomial Probability Distribution

1. The procedure must have a fixed number of trials.
2. The trials must be independent.
3. Each trial must have all outcomes classified into two categories (commonly, success and failure).
4. The probability of success remains the same in all trials.

Solve by binomial probability formula, Table A-1, or technology.

The Binomial Probability Formula

$$P(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$$

for $x = 0, 1, 2, \ldots, n$

where

n = number of trials

x = number of successes among n trials

p = probability of success in any one trial

q = probability of failure in any one trial ($q = 1 - p$)

Example

Page 221, problem 36 (from chapter 5-3)

The author purchased a slot machine configured so that there is a 1/2000=0.0005 probability of winning the jackpot on any individual trial.

(b) Find the probability of at least 2 jackpots in 5 trials

x = number of jackpots in 5 trials

Example

(b) Probability of at least 2 jackpots?

At least 2 **jackpots means** **<u>2 or more</u>** **which means x=2 or x=3 or x=4 or x=5.**

P(at least 2) = P(x=2 OR x=3 OR x=4 OR x=5)

= P(x=2)+P(x=3)+ P(x=4)+P(x=5)

Example

(b) Probability of at least 2 jackpots?

$$P(2 \text{ or } 3 \text{ or } 4 \text{ or } 5) = P(x=2) + P(x=3) + P(x=4) + P(x=5)$$

$$= \frac{5!}{3! \cdot 2!}(0.0005)^2 \cdot (0.9995)^3 + \frac{5!}{2! \cdot 3!}(0.0005)^3 \cdot (0.9995)^2$$

$$+ \frac{5!}{1! \cdot 4!}(0.0005)^4 \cdot (0.9995)^1 + \frac{5!}{0! \cdot 5!}(0.0005)^5 \cdot (0.9995)^0$$

$$= \boxed{0.00000250}$$

115

Observation

The previous method is not always practical. In particular, if n is very large we may need to consider another method.

Example

Page 294, problems 20

20. Gender Selection The Genetics & IVF Institute developed its YSORT method to increase the probability of conceiving a boy. Among 152 women using that method, 127 had baby boys. Assuming that the method has no effect so that boys and girls are equally likely, find the probability of getting at least 127 boys among 152 babies. Does the result suggest that the YSORT method is effective? Why or why not?

Example

Page 294, problems 20

Here we should use a binomial probability distribution, but it would be impractical to find the answer as in slot machine example:

P(at least 127) =

P(x=127 OR x=128 OR … OR x=151 OR 152)

= P(x=127) + P(x=128) + … +P(x=151)+P(x=152)

Key Concept

This section presents a method for using a normal distribution as an approximation to the binomial probability distribution.

If the conditions of $np \geq 5$ and $nq \geq 5$ are both satisfied, then probabilities from a binomial probability distribution can be approximated well by using a normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$.

117

Approximation of a Binomial Distribution with a Normal Distribution

Validity conditions are that:

$np \geq 5$	$nq \geq 5$

then $\mu = np$ and $\sigma = \sqrt{npq}$

and the random variable has

a  **distribution.**

(normal)

Always Check the Conditions for Approximation Validity

1. For slot machine example,

$$np=5(0.0005)=0.0025<5$$

so we __cannot__ use the normal distribution as an approximation to binomial distribution

2. For YSORT gender selection example:

$$np=152(0.5)=76>5 \quad \text{and } nq=152(0.5)=76>5$$

so we __can__ use the normal distribution as an approximation to binomial distribution

Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

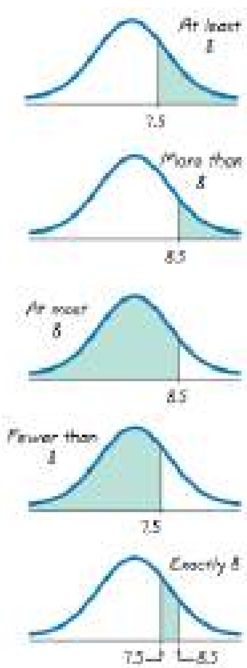
1. **Verify that both $np \geq 5$ and $nq \geq 5$. If not, you must use software, a calculator, a table or calculations using the binomial probability formula.**
2. **Find the values of the parameters μ and σ by calculating $\mu = np$ and $\sigma = \sqrt{npq}$.**

Definition

When we use the normal distribution (which is a continuous probability distribution) as an approximation to the binomial distribution (which is discrete), a continuity correction is made to a discrete whole number x in the binomial distribution by representing the discrete whole number x by the interval from

$$x - 0.5 \text{ to } x + 0.5$$

(that is, adding and subtracting 0.5).



***X* = <u>at least</u> 8**
(includes 8 and above)

***X* = <u>more than</u> 8**
(doesn't include 8)

***X* = <u>at most</u> 8**
(includes 8 and below)

***X* = <u>fewer than</u> 8**
(doesn't include 8)

***X* = <u>exactly</u> 8**

Example

Page 294, problems 5-12

Applying Continuity Correction. *In Exercises 5–12, the given values are discrete. Use the continuity correction and describe the region of the normal distribution that corresponds to the indicated probability. For example, the probability of "more than 20 defective items" corresponds to the area of the normal curve described with this answer: "the area to the right of 20.5."*

Example

Page 294, problem 6

Probability of at least 2 traffic tickets this year.

Example

Page 294, problem 6

Probability of at least 2 traffic tickets this year.

ANSWER:

The area to the right of 1.5

Example

Page 294, problem 8

Probability that the number of students who are absent is exactly 4

Example

Page 294, problem 8

Probability that the number of students who are absent is exactly 4

ANSWER:

The area between 3.5 and 4.5

Example

Page 294, problems 20

20. Gender Selection The Genetics & IVF Institute developed its YSORT method to increase the probability of conceiving a boy. Among 152 women using that method, 127 had baby boys. Assuming that the method has no effect so that boys and girls are equally likely, find the probability of getting at least 127 boys among 152 babies. Does the result suggest that the YSORT method is effective? Why or why not?

Example

Page 294, problems 20

ANSWER: first check that normal approximation can be used:

binomial: $n=152$ and $p=0.50$
normal approximation appropriate since

$$np = 152(0.50) = 76 \geq 5$$
$$nq = 152(0.50) = 76 \geq 5$$

Example

Page 294, problems 20

Next compute mean and standard deviation

$$\mu = np = 152(0.50) = 76$$
$$\sigma = \sqrt{npq} = \sqrt{152(0.50)(0.50)} = 6.164$$

Example

Page 294, problems 20

Compute probability of getting at least 127 boys:

$$P(x \geq 127)$$

using continuity correction and converting to z-score:

$$= P(x>126.5)$$
$$= P(z>8.19)$$
$$= 1 - 0.9999$$
$$= 0.0001$$

124

Example

Page 294, problems 20

The result <u>does</u> suggest that YSORT method is effective sinced the probability of getting 127 boys simply by chance is very small (less than 0.05)

Example

Page 295, problems 26

26. Acceptance Sampling With the procedure called *acceptance sampling*, a sample of items is randomly selected and the entire batch is either rejected or accepted, depending on the results. The Telektronics Company has just manufactured a large batch of backup power supply units for computers, and 7.5% of them are defective. If the acceptance sampling plan is to randomly select 80 units and accept the whole batch if at most 4 units are defective, what is the probability that the entire batch will be accepted? Based on the result, does the Telektronics Company have quality control problems?

125

Example

Page 295, problems 26

ANSWER: first check that normal approximation can be used:

binomial: $n=80$ and $p=0.075$
normal approximation appropriate since

$$np = 80(0.075) = 6 \geq 5$$
$$nq = 80(0.925) = 74 \geq 5$$

Example

Page 295, problems 26

Next compute mean and standard deviation

$$\mu = np = 80(0.075) = 6$$
$$\sigma = \sqrt{npq} = \sqrt{80(0.075)(0.925)} = 2.356$$

126

Example

Page 295, problems 26

Compute probability of getting at most 4 units that are defictive:

$$P(x \leq 4)$$

using continuity correction and converting to z-score:

$$\begin{aligned} &= P(x<4.5) \\ &= P(z<-0.64) \\ &= 0.2611 \end{aligned}$$

Example

Page 295, problems 26

The probability a batch is accepted is 0.2611 or a little more than 26% of batches are accepted. Thus it seems that there is a quality control problem here as the acceptance rate seems low.

Recap

In this section we have discussed:

- **Approximating a binomial distribution with a normal distribution.**
- **Procedures for using a normal distribution to approximate a binomial distribution.**
- **Continuity corrections.**