Statistics is a set of principles and techniques that are useful in solving real-life problems based on limited data. This course is intended to prepare students to deal with solving problems encountered in research projects, decision making based on data, general life experiences beyond the classroom and university setting. In order to help students stay focused on solving problems, the study of statistics will be approached by a four-step process:

1. Defining the problem
2. Collecting data
3. Summarizing data
4. Analysing data, Interpreting the analysis, and Communicating results.

In fact, unless the problem to be addressed is clearly defined and the data collection carried out properly, the interpretation of the results of the analyses may convey misleading information because the analyses were based on a data set that did not address the problem or that contained improper information. Furthermore, students spend much time in discussing how to analyze data using proper statistical procedures. Especially, the text emphasizes the importance of the assumptions on which statistical procedures are based. Then following the analysis of the data, the results of the analysis must be interpreted and communicated in unambiguous terms to interested people.

Chapter 1

Statistics and the Scientific Method

Let's begin with a definition.

> **Statistics** is the science of collecting, summarizing, analyzing, and interpreting *data* in order to make decisions and to draw conclusions.

Type of Statistics

- **Descriptive statistics** consists of methods for summarizing and displaying data by using tables, graphs, and summary measures.
- **Inferential statistics** consists of methods that use sample results to help make *decisions or prediction* about a population.

Population v.s. Sample, Parameter v.s. Statistic

A **population** consists of *all* elements (individuals, objects, events, measurements) that we are interested in studying. A **parameter** is a numerical description of a population. Each research question refers to a target population.

Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A *portion* of the population selected for study is referred to as a **sample**. A **statistic** is a numerical description of a sample. We use a sample statistic to draw conclusions about the population parameter.

A **variable** is the characteristic of the individual to be measured or observed.

Example 1.1 Identify the population and sample. Describe a population parameter or a sample statistic.

(a) In the 2006 election in California, an exit poll sampled 2705 of the 7 million people who voted. The poll stated that 56.5% reported voting for a Republican candidate. The poll predicted that the Republican candidate would win. In fact, of all 7 million voters, 55.9% voted for the Republican candidate.

(b) Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. **The population consists of all videos on YouTube, and the sample consists of the 1000 randomly selected videos. Because 2% is based on a subset of the population, it is a sample statistic. The population parameter (the percentage of videos on YouTube that are cat videos) is still unknown.**

Example 1.2 True or False?

(a) Based on a survey of 400 students in a university in which 20 percent indicated that they were business majors. The university student newspaper reported that "20 percent of all the students at the university are business majors." This report is an example of descriptive statistics.

(b) A pharmaceutical company conducts a study where 50 patients are given a drug. They find that 10 percent of patients experience nausea as a side effect. This 10 percent is an example of a parameter.

Question How is a sample related to a population? Why is a sample used more often than a population?

It is generally impractical (too time-consuming and too expensive) to obtain *all* measures from an entire population of interest. Thus, true population parameters are almost never known to us. Samples are usually used instead of an entire population. Sample statistics vary from sample to sample, but we are almost always (if not always) interested in generalizing from a smaller sample to a larger population of interest.

Chapter 2

Collecting Data Using Surveys and Scientific Studies

To infer validly that the results of a study are applicable to a larger group than just the participants in the study, we must select the most appropriate method to collect the data. Data can be obtained from **surveys**, **designed experiments**, **observational studies** or **existing sources** such as previous records, censuses, and previous studies. We will learn, in **STAT U516 Statistical methods II**, some standard designs of experiments and methods for analyzing the data obtained from the experiment. In this chapter, we will consider some sampling designs for surveys.

Sampling designs

To collect unbiased data, a researcher must ensure that a sample is representative of the population, so the sample accurately reflects the population as a whole.

1. A **random sample** is one in which every member of the population has an equal chance of being selected. There are several commonly used sampling techniques. Each has advantages and disadvantages.

   (a) Simple random sample

   A **simple random sample** is a sample in which every possible sample of the same size n has the same chance of being selected. One way to collect a simple random sample is to assign a different number to each member of the population, and then use a random number table or generate random numbers from calculators or computer software programs.

   (b) Stratified sample

   A **stratified sample** is obtained by dividing the population into nonoverlapping groups called *strata*, and then obtaining a simple random sample from each *stratum*. The individuals within each stratum should be homogeneous in some way such as age, gender, or political preference, etc.

   (c) Systematic sample

   A **systematic sample** is obtained by selecting every kth individual from the population, where k is some whole number. The first individual selected is a random number between 1 and k. This is useful technique when you can't obtain a list of the individuals in the population.

   (d) Cluster sample

   Cluster sampling is a sampling technique where the entire population is divided into groups (or *clusters*), and a random sample of these clusters are selected. All observations in the selected clusters are included in the sample. In using a cluster sample, care must be taken to ensure that all clusters have similar characteristics.

2. A **convenience sample** is a sample where the patients are selected at the convenience of the researcher. The classic example of a convenience sample is standing at a shopping mall and selecting shoppers as they walk by to fill out a survey. If the individuals included in the survey are selected based on convenience alone, there may be biases in the sample survey, which prevent the survey from accurately reflecting the population as a whole.

Stratified v.s. Cluster

- In **stratified** sampling, the population is divided into strata according to some variables that are thought to be **related** to the variables that we are interested in. The elements within a stratum should be as homogeneous as possible, but the elements in different strata should be as heterogeneous as possible. Then a sample is taken from **every** stratum.
- In **cluster** sampling the elements within a cluster should be as heterogeneous as possible, but clusters themselves should be as homogeneous as possible. Ideally, each cluster should be a small-scale representation of the population. Only **some** of the clusters are taken.

Example 2.1 Identify each of the following samples by naming the sampling technique used.

(a) Every tenth customer entering a health club is asked to select his or her preferred method of exercise.

(b) Divide the subscribers of a magazine into three different income categories and then select a random sample from each category to survey about their favorite feature.

(c) A farmer divides his orchard into 10 subsections, randomly selects 4 subsections and samples all of the trees within the 4 subsections in order to approximate the yield of his orchard.

(d) Use a random number table to select a sample of books and determine the number of pages in each book.

(e) You are researching an average annual salary for nurses and obtain the annual salary of each of the nurses that are on duty at the time you chose to interview at the hospital.

(f) After a hurricane, a disaster area is divided into 50 equal grids. Ten of the grids are selected, and every occupied household in the grid is interviewed to help focus relief efforts on what residents require the most regardless of geographical location.

Example 2.2 Suggest a sampling strategy for carrying out this study.

(a) A large college class has 200 students. All 200 students attend the lectures together, but the students are divided into 4 groups, each of 50 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course and he plans to select 20 students, but he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

(b) Suppose a group of researchers are interested in estimating the malaria rate in a densely tropical portion of a country in Southeast Asia. They learn that there are 30 villages in that part of the country's jungle, each more or less similar to the next. Their goal is to test 200 individuals for malaria.

Data collection techniques

Having chosen a particular sample design, how do we *actually* collect the data? The most commonly used methods are

- personal interviews
- telephone interviews
- self-administered questionnaire

Observational vs Experimental study

We draw a distinction between observational and experimental studies in terms of the inferences (conclusions) that can be drawn from the sample data. Differences found between treatment groups from an observational study are said to be *associated with* the use of the treatments; on the other hand, differences found between treatments in an experimental study are said to be *due to* the treatments.

Chapter 3

Summarizing Data

In this text, we will deal with a random sample and methods for summarizing and analyzing data collected in such a way. Good descriptive statistics enable use to make sense of the data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

Calculators can be great aids in performing some of calculations in this chapter, especially for small data sets. For larger data sets, a computer can help in these situations. Many statistical software packages are available for use on computers. The more commonly used systems are Minitab, R, SAS, and SPSS. The ability of such packages to perform complicated analyses on large amounts of data more than repays the initial investment of time and irritation.

3.1 Graphical Methods

Types of Data

- **Quantitative** data are measurements that can be measured on a *numerical* scale.
- **Qualitative** data are measurements that cannot be measured on a natural numerical scale but can be classified into two or more non-numerical categories (e.g. gender, color of car).

Frequency distribution and Histograms

A histogram is a graphical method that groups numeric data into bins (also called segments or classes).

Example 3.1 Professor H, a college nursing instructor, randomly selects 30 students from the college to obtain the pulse rates (measured in heart beats per minute while resting) of the students at the university.

52	74	76	70	75	46	70	50	76	65	73	78	74	73	66
59	70	62	65	72	55	70	60	72	76	66	73	76	68	79

Construct frequency distributions. Use 4 classes. Describe the shape of the distribution.

The Shape of Distributions

- A distribution is **symmetric** if the right and left sides of the curve are approximately mirror images of each other.
- A distribution is **uniform** when all entries have equal frequencies.
- A distribution is **skewed right** if the *tail* of the graph extends to the right.
- A distribution is **skewed left** if the *tail* of the graph extends to the left.

Constructing a Frequency Distribution from a Data Set

1. 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.
2. Find the class limits. You can use the minimum data entry as the lower limit of the first class. To find the remaining lower limits, add the class width to the lower limit of the preceding class. Then find the upper limit of the first class. Remember that classes cannot overlap. Find the remaining upper class limits.
3. Count the tally marks to find the frequency f for each class.

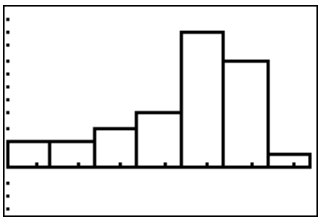


Figure 3.1: A histogram from a TI-83/84 using automatically determined number of classes.

Stem-and-Leaf Plot

In a stem-and-leaf plot, each number is separated into a stem and a leaf.

Example 3.2 The following are the times (in minutes) taken to commute from home to work for 19 workers.

10	22	61	33	48	5	11	23	39	26
26	32	17	7	15	19	29	43	21	

Question What is an advantage of using a stem-and-leaf plot instead of a histogram? What is a disadvantage?

One of the differences between a stem-and-leaf plot and a histogram is that even for variables involving a large number of different values, the stem-and-leaf plot shows the individual data values whereas the histogram requires you to group the data and lose the individual values.

Pie Chart (Circle graph)

A circle is divided into portion that represents the relative frequencies of data belonging to different categories.

Example 3.3 The data on the status of 30 students.

Jr	F	S	Sr	Jr	S	Jr	S	Jr	S
Sr	Jr	S	Jr	S	Sr	F	F	Jr	S
S	Sr	S	Sr	S	Sr	S	F	Jr	Sr

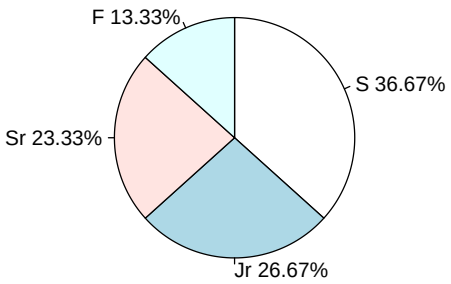
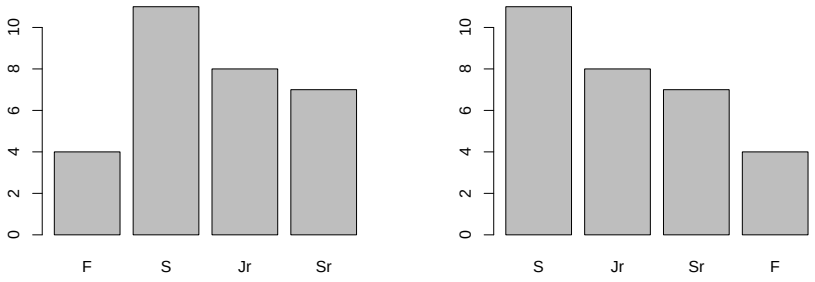


Figure 3.2: Pie chart for the data on the status of 30 students.

Pareto Chart (Bar Graph)

A Pareto chart is a vertical bar graph in which the height of each bar represents frequency or relative frequency. The bars are positioned in order of decreasing height.



(a) Bar graph (b) Pareto chart

Figure 3.3: Bar graph and Pareto chart for the data on the status of 30 students, respectively.

Guidelines for categorical data

Choose a small number of categories for the variable because too many make the charts difficult to interpret. Whenever possible, construct the charts so that percentages are in either ascending or descending order.

Time Series Chart

A data set that is composed of quantitative entries taken at regular intervals over a period of time is time series.

Example 3.4 Garbage that is not recycled is buried in landfills. Here are data that emphasize the need for recycling: the number of landfills operating in the United States in the years 1988 to 2004.

Year	Landfills	Year	Landfills	Year	Landfills
1988	7924	1994	3558	2000	1967
1990	6326	1996	3091	2002	1767
1992	5386	1998	2314	2004	1605

Make a time plot of these data and describe the trend that your plot shows. Why does the trend emphasize the need for recycling?

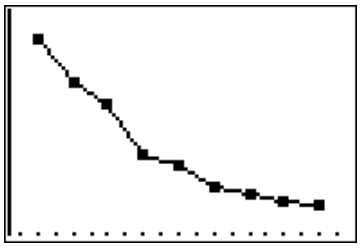


Figure 3.4: Time series plot on a TI-83/84.

Intensity Map

When we encounter geographic data, we should map it using an intensity map, where colors are used to show higher and lower values of a variable.

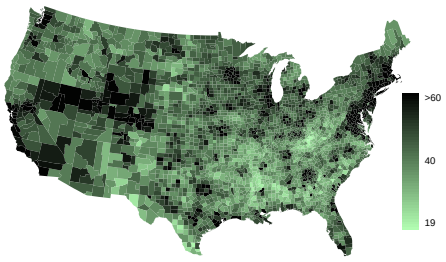


Figure 3.5: Median household income ($1000s).

3.2 Numerical Methods

We often represent a data set by numerical summary measures.

Measures of Central Tendency

A measure of central tendency gives the tendency of the data to cluster or center. That is, we seek to describe the center of the distribution. We will learn how to find the mean, median, and mode of a population and of a sample.

- **Mean** is obtained by dividing the sum of all entries by the number of entries in the data set.
- **Median** is the value of the *middle* entry in the data set that has been ranked in *increasing* order.
- **Mode** is the value that occurs with the *highest frequency* in a data set.

Example 3.5 Consider the following sample data.

5 7 4 5 10 6 2

Example 3.6 The status of seven students who are members of the student senate at a college are senior, sophomore, senior, junior, sophomore, senior, junior.

Example 3.7 True or False?

(a) The median is the measure of central tendency most likely to be affected by an outlier.

(b) A data set can have the same mean, median, and mode.

(c) When a distribution is skewed left, the mean is to the left of the median.

As distributions go from symmetrical to more skewed, the researcher is more likely to choose the median over the mean.

Measures of Variability

The mean, median, or mode is usually not by itself a sufficient measure to reveal the shape of the distribution of a data set. The measures that help us know about the spread of a data set are called the measures of dispersion such as range, variance and standard deviation.

- **Range**
- **Variance: σ^2 vs. s^2**
- **Standard deviation: σ vs. s**

> The standard deviation is the most used measure of variation. In general, a *lower* value of the standard deviation for a data set means that the values of that data set are spread over a relatively *smaller range* around the mean.

Example 3.8 Consider the following sample data set $\{1, 2, 8, 5, 4\}$ which is drawn from the population data set $\{1, 2, 3, 9, 5, 4, 8, 6\}$. Find the range, variance, and standard deviation.

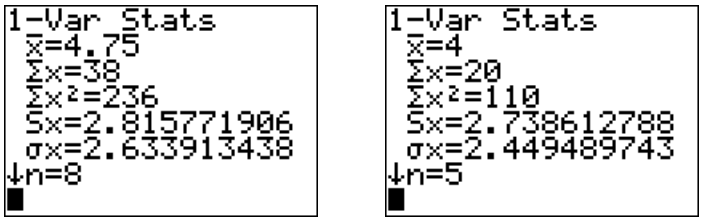


Figure 3.6: Population standard deviation σ vs. sample standard deviation s

Example 3.9 Consider the following sample data set $\{1, 2, 5, 4\}$ which is drawn from the population data set $\{1, 2, 3, 10, 5, 4, \cdots\}$. Find the range, variance, and standard deviation of the sample data set.

Example 3.10 Consider the cholesterol level of a *sample* of 6 female employees in a university:

154 216 171 188 229 203

Example 3.11 A random sample of 10 junior students currently enrolled in school is selected, and their GPAs are listed in the table.

3.7 3.4 3.2 2.9 3.3 3.5 2.5 3.2 3.5 3.8

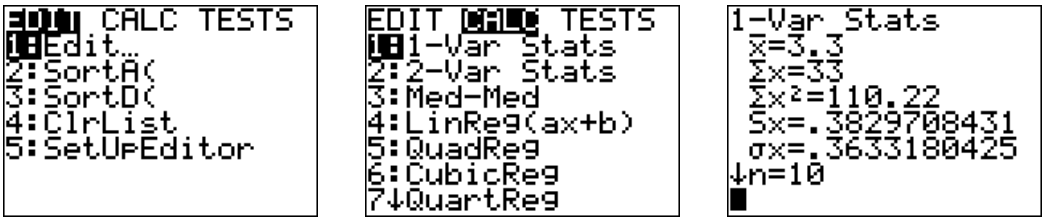


Figure 3.7: ON a TI-83/84, choose the EDIT menu, enter the sample data into a list, and choose 1-Var Stats option on the CALC menu.

Question

(a) Why is the standard deviation used more frequently than the variance?

(b) **Given a data set, how do you know whether to calculate σ or s?** Describe the difference between the calculation of population standard deviation σ and sample standard deviation s. **When given a data set, one would have to determine if it represented the population or was a sample taken from the population. If the data are a population, then σ is calculated. If the data are a sample, then s is calculated.**

In problems requiring statistical inference, we will not be able to calculate values for various population parameters (e.g. μ and σ), but we will be able to compute corresponding sample statistics ($\overline{x}$ and s) from the sample and use these quantities to estimate the unknown population parameters.

Note Students majoring in mathematics are encouraged to try.

1. Show that the average deviation is always zero. i.e. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

2. Let a and b be constants and let $y_i = ax_i + b$ for $i = 1, 2, \ldots, n$. Find $\overline{y}$ in terms of $\overline{x}$. Find the variance of y, S_y^2 in terms of the variance of x, S_x^2.

3. **Why divide by $(\mathbf{n-1})$?** Let a population consist of the values 1, 3, 5. Assume that samples of 2 values are randomly selected with replacement from this population. (That is, a selected value is replaced before the second selection is made.)

   (a) Find the population variance σ^2 of the population $\{1, 3, 5\}$.

   (b) After listing the 9 different possible samples of 2 values selected with replacement, find the sample variance s^2 (which includes division by $n-1$) for each of them, then find the mean of the 9 sample variances s^2.

   (c) Now, for each of the 9 different possible samples of 2 values selected with replacement, find the variance by treating each sample as if it is a population (i.e. using the formula for population variance, which includes division by n), then find the mean of those 9 population variances.

   (d) Which approach results in values that are better estimated of σ^2? The preceding parts show that s^2 is an **unbiased** estimator of σ^2.

Measures of Position

In this section, we will learn how to specify the position of a data entry within a data set.

Quartiles

The three **quartiles** are summary measures that divide an ordered data set into four equal parts, each containing 25% of the measurements.

- first quartile (Q_1) : the middle term among the observations failing below the Q_2 position (not including Q_2).
- second quartile (Q_2) : the same as the median of a data set.
- third quartile (Q_3) : the middle term among the observations failing above the Q_2 position (not including Q_2).

Interquartile Range is the difference between the third and the first quartiles

$$\text{IQR} = Q_3 - Q_1.$$

Five-number Summary

The **five-number summary** consists of the five numbers (Lowest value, Q_1, median, Q_3, highest value).

Box-and-Whisker plot

Another important application of quartiles is to represent data sets using box-and-whisker plot.

Example 3.12 The following data give the weights (in pounds) lost by 11 members of a health club at the end of two months after joining the club.

5 10 8 7 25 12 5 14 11 10 21

A box-and-whisker plot can show whether a data set is roughly symmetric or skewed.

Example 3.13 A random sample of 10 junior students currently enrolled in school is selected, and their GPAs are listed in the table. Find the range, variance, and standard deviation of the sample data set.

3.7 3.4 3.2 2.9 3.3 3.5 2.5 3.2 3.5 3.8

(a) Find the five-number summary of the data set.

$$\boldsymbol{Min = 2.5, Q_1 = 3.2, Q_2 = 3.35, Q_3 = 3.5, Max = 3.8}$$

(b) Find the interquartile range. $\mathbf{IQR = Q_3 - Q_1 = 3.5 - 3.2 = 0.3}$

(c) Draw a box-and-whisker plot, and describe the shape of the distribution. **The data set is skewed to the left.**

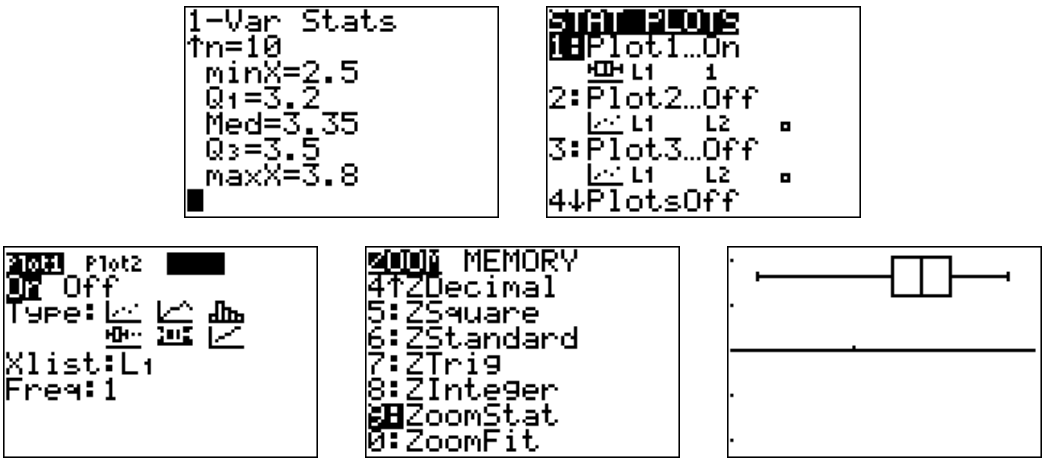


Figure 3.8: Five number summary and a box-and-whisker plot on a TI-83/84.

Chebyshev's Theorem

This theorem applies to *all* distributions.
For any number $k > 1$, the proportion of the data that must lie within k standard deviations on either side of the mean is **at least** $1 - \frac{1}{k^2}$. For example ($k = 2$), **at least 75%** of the observations lies within **two** standard deviation on either side of the mean, $\overline{x} - 2s$ to $\overline{x} + 2s$.

Example 3.14 The mean time taken by all participants to a 400-meter race was 52.4 seconds with a standard deviation of 2.2. Using Chebyshev's Theorem, find what percentage of runners who ran this race between

(a) 48 and 56.8 seconds

(b) 45.8 and 59 seconds

(c) 44.7 and 60.1 seconds

Example 3.15 In a recent year, the SAT verbal scores have the mean of 500 and the standard deviation of 80. Find what percentage of the SAT verbal scores that will be between 380 and 620.

Example 3.16 The U.S. mint produces dimes with an average diameter of 0.5 and a standard deviation of 0.01. Using Chebyshev's theorem, find the number of coins in a lot of 400 coins having diameter between 0.485 and 0.515. **At least 223 coins**

Example 3.17 New car prices of a certain brand in 2013 averaged \$25,500 with standard deviation of \$1500. What can you say about the distribution of new car prices? **Any one of the following statements would be true.**

(a) In 2013 at least 75% of all new cars were priced between \$22500 and \$28500.

(b) In 2013 at least 88.89% of all new cars were priced between \$21000 and \$30000.

(c) In 2013 at most 11.11% of all new cars were priced over \$30000.

Rule of thumb: Empirical Rule (68-95-99.7%)

For a mound-shaped distribution only, the following properties apply.

- **About 68%** of the observations lies within **one** standard deviation of the mean, $\bar{x} - s$ to $\bar{x} + s$.
- **About 95%** of the observations lies within **two** standard deviations of the mean, $\bar{x} - 2s$ to $\bar{x} + 2s$.
- **About 99.7%** of the observations lies within **three** standard deviations of the mean, $\bar{x} - 3s$ to $\bar{x} + 3s$.

Example 3.18 The GPAs of all students enrolled at A university are bell-shaped with a mean of 3.10 and a standard deviation of 0.28.

(a) What percentage of students at the university have a GPA between 2.82 and 3.38?

(b) What percentage of students at the university have a GPA less than 2.26?

(c) What percentage of students at the university have a GPA greater than 3.66?

(d) What percentage of students at the university have a GPA between 2.54 and 3.66?

Example 3.19 The mean life of a certain brand of auto batteries is 44 months with a standard deviation of 3 months. Assume that the lives of all auto batteries of this brand have a mound-shaped distribution. Find the percentage of auto batteries of this brand that have a life of

(a) 41 to 47 months

(b) 41 to 50 months

(c) longer than 47 months

Example 3.20 Consumer Reports Magazine wrote an article stating that monthly charges for cell phone plans in the U.S. has a bell-shaped distribution with a mean of \$62 and a standard deviation of \$18.

(a) What percent of people in the U.S. have a cell phone bill between \$62 and \$80 per month? **About 34%.**

(b) What percent of people in the U.S. have a monthly cell phone bill between \$26 and \$44? **About 13.5%.**

Example 3.21 The scores for all high school seniors taking the verbal section of the Scholastic Aptitude Test (SAT) in a particular year had a mean of 490 and a standard deviation of 100.

(a) What percentage of seniors scored between 290 and 590 on this SAT test? **Because the shape of the distribution of SAT scores is not known or given, the percentage can't be determined.**

(b) The distribution of SAT scores is bell-shaped. What percentage of seniors scored between 290 and 590 on this SAT test? **About 81.5%.**

(c) A rather exclusive university only admits students who were among the highest 2.5% of the scores on this test. The distribution of SAT scores is bell-shaped. What score would a student need on this test to be qualified for admittance to this university? **690 or higher SAT score.**

(d) The distribution of SAT scores is mound-shaped. What percentage of seniors scored between 300 and 690 on this SAT test?

Chapter 4

Probability Distributions

Most management decisions are made in the presence of uncertainty. Probability is the language of uncertainty and probability is the tool that enables us to make an inference. The basic language of probability deals with many different kinds of events associated with quantitative and qualitative variables. A **random variable** is a variable whose value is determined by the outcome of a random experiment.

- Discrete Random Variable : it has a finite or countable number of possible outcomes that can be listed. Ex: Binomial distribution, Poisson distribution, etc.
- Continuous Random variable : it has an uncountable number of possible outcomes. Ex: Normal distribution, t-distribution, χ^2-distribution, F-distribution, etc.

Qeustion Suppose adult IQ scores have a mean of 100 and a standard deviation of 15. What percentage of adults scored between 85 and 115 on IQ test?

4.1 A Useful Continuous Random Variable: Normal Distribution

Properties of a Normal Distribution

One of the most commonly observed continuous random variables which are bell-shaped is a **normal** random variable and its probability distribution is called a **normal distribution**.

The normal distribution has the following properties:

- It is bell-shaped and perfectly symmetric about its mean μ.
- The mean, median, and mode are equal.
- The total area under the curve is 1.
- There are an infinitely large number of normal curves–one for each pair of values for the mean of μ and the standard deviation of σ.
- The function for the normal distribution is

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

Standard Normal Distribution

Especially, the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution** and a random variable with a standard normal distribution is typically denoted by the symbol Z.

The standard normal distribution has the following properties:

- It is bell-shaped and perfectly symmetric about 0.
- The mean, median, and mode are equal.
- The curve has the expression

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

- The cumulative area (proportion, probability) under the standard normal curve to the left of c is

$$P(Z < c) = P(-\infty < Z < c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \, dz,$$

which can be calculated by a calculator, statistical software, or by using a **normal distribution table**. Modern technology has made the tables virtually obsolete, but there is still some insight to be gained from working with the tables so we will discuss it here.

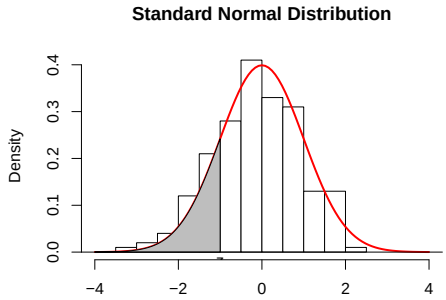


Figure 4.1: Probability under the standard normal distribution

Example 4.1 Find the area under the standard normal curve to the left of $z = -1.24$.

- Our normal probability table gives the area to the left of z, i.e. the area under the standard normal curve to the left of a z-score gives the probability $P(Z < z)$.
- Generally, we round z to two decimals.
- Identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation.

Example 4.2 Use the standard normal table.

(a) Find the area (probability) under the standard normal curve to the left of $z = 1.08$ which is $P(z < 1.08)$.

(b) Find the area (probability) under the standard normal curve to the left of $z = -1.23456$ which is $P(z < -1.23456)$.

(c) Find the area (probability) under the standard normal curve to the right of $z = 1.12$ which is $P(z > 1.12)$.

(d) Find the area (probability) under the standard normal curve to the right of $z = -2.05$ which is $P(z > -2.05)$.

(e) Find the probability $P(-1.32 < z < 0.65)$ which is the area under the standard normal curve between $z = -1.32$ and $z = 0.65$.

(f) Find the probability $P(z < -2.05 \quad \text{or} \quad z > 2.15)$ which is the area under the standard normal curve to the left of $z = -2.05$ or to the right of $z = 2.15$.

(g) Find the probability $P(z < -2.65 \quad \text{or} \quad z > 2.65)$ which is the area under the standard normal curve to the left of $z = -2.65$ or to the right of $z = 2.65$.

(h) Find the probability $P(-1.00 < z < 1.00)$.

(i) Find the probability $P(-2.00 < z < 2.00)$.

(j) Find the probability $P(-3.00 < z < 3.00)$.

Example 4.3 Adult IQ scores are normally distributed with $\mu = 100$ and $\sigma = 15$. Find the percentage of adult IQ scores between 70 and 115.

Probability and Normal Distributions

Because every normal distribution can be transformed to the standard normal distribution, we can use z-scores

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

and the standard normal curve to find areas under any normal curve.

Example 4.4 The scores for all high school seniors taking the verbal section of the Scholastic Aptitude Test (SAT) in a particular year had a mean of 490 and a standard deviation of 100. What percentage of seniors scored between 290 and 690 on this SAT test if the distribution of SAT math scores is normally distributed?

Example 4.5 The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days. Find the probability of a pregnancy lasting 308 days or longer.

Example 4.6 A construction zone on a highway has a posted speed limit of 40 miles per hour. The speeds of vehicles passing through this construction zone are normally distributed with a mean of 46 miles per hour and a standard deviation of 4 miles per hour. Find the percentage of vehicles passing through this zone that are

(a) exceeding the posted speed limit of 40

(b) traveling at speeds between 50 and 55 miles per hour.

Example 4.7 The GPAs of all students enrolled at A university have a normal distribution with a mean of 3.10 and a standard deviation of 0.28.

(a) What percentage of students at the university have a GPA less than 2.5?

(b) What percentage of students at the university have a GPA greater than 3.0?

(c) What percentage of students at the university have a GPA greater than 3.5?

(d) What percentage of students at the university have a GPA between 2.8 and 3.3?

(e) What is the probability that a randomly selected student has a GPA lower than 2.0?

(f) What is the probability that a randomly selected student has a GPA greater than 3.7?

Example 4.8 The SAT scores (verbal + math) of all freshman students currently enrolled at B university are normally distributed with a mean of 976 and a standard deviation of 180.

(a) What percentage of freshman students at the university have a SAT score less than 800? **16.35%** (*Tech*: **16.41%**)

(b) What percentage of freshman students at the university have a SAT score greater than 1100? **24.51%** (*Tech*: **24.54%**)

(c) What percentage of freshman students at the university have a SAT score greater than 1200? **10.75%** (*Tech*: **10.67%**)

(d) What percentage of freshman students at the university have a SAT score between 900 and 1000? **21.45%** (*Tech*: **21.66%**)

(e) What is the probability that a randomly selected freshman student has a SAT score lower than 750? **10.38%** (*Tech*: **10.46%**)

(f) What is the probability that a randomly selected freshman student has a SAT score greater than 1300? **3.59%** (*Tech*: **3.59%**)

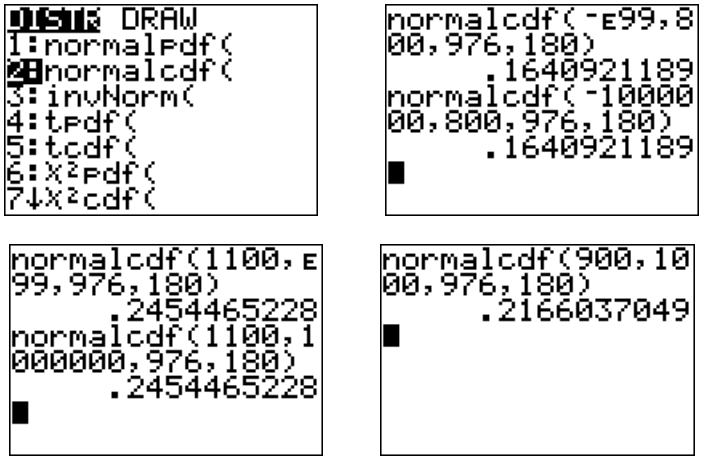


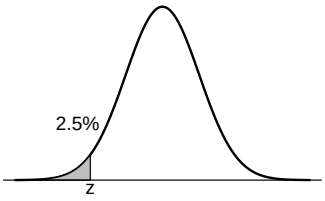
Figure 4.2: You can use a TI-83/84 to find the probability $P(x)$ as well.

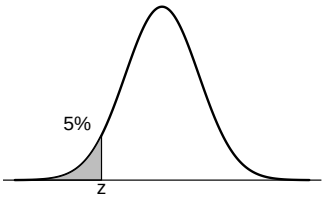
Example 4.9 The Sociology of Sport Journal examined the SAT scores of male student-athletes at NCAA Division I institutions. The SAT score is used to determine whether athletes are eligible to participate in athletics their first year, with the NCAA requiring a minimum score of 700. Suppose that SAT score of athletes on scholarship have an average of 950 and a standard deviation of 200. Assuming the distribution of SAT scores is normal, what percentage of the athletes will not be eligible their first year? **10.56%** (*Tech*: **10.56%**)

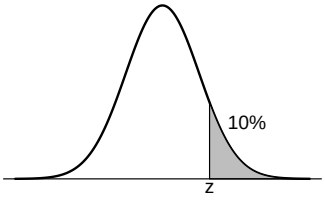
Finding Values from given probabilities

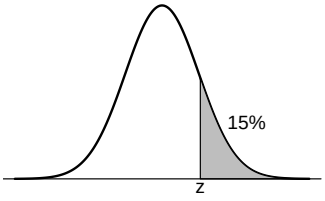
What if we are given a probability and want to find a value? i.e. **Determining the z and x values when an area under the normal curve is known.** An important aspect of the normal distribution is that we can easily find the percentiles of the distribution.

Example 4.10 Find the z-score that the cumulative area under the standard normal curve is (a) 0.2743, (b) 0.4567, (c) 0.8888, and the z-score that corresponds to the shaded area.









In most cases, the given area will not be found in the table, so use the entry closest to it. If the given area is halfway between two area entries (e.g. 5%), use the z-score halfway between the corresponding z-scores. For example, use $z = -1.645$ for 5th percentile and $z = 1.645$ for 95th percentile.

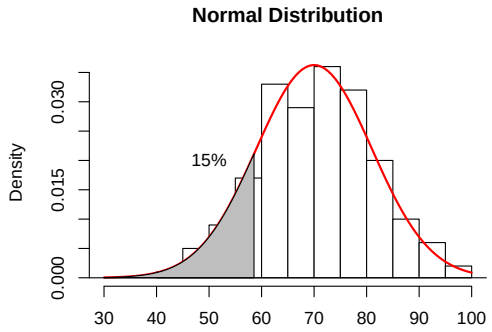
To transform a z-score to a data value in a given normal population with known values of μ and σ, the x value is calculated as

$$x = \mu + z\sigma$$

Example 4.11 The scores on a university examination are normally distributed with a mean of 70 and a standard deviation of 11.

(a) If the bottom 15% of the students will fail the course, what is the lowest mark that a student must have in order to pass the course?

(b) What is the lowest score that a student can be in the top 5% of students?



Example 4.12 Almost all high school students who intend to go to college take the SAT test. Matt is planning to take this test soon. Suppose the SAT scores of all students who take this test with Matt will have a normal distribution with a mean of 1020 and a standard deviation of 153. What should his score be on this test so that only 15% of all examinees score *higher* than he does?

Example 4.13 A local community college decides to award scholarships based upon the results of a SAT. If this year's 5000 test results are normal with $\mu = 1240$ and $\sigma = 350$, determine the minimum test grade needed to win a scholarship if the top 300 students receive scholarships.

Example 4.14 The monthly utility bills in a certain city are normally distributed, with a mean of \$100 and a standard deviation of \$12. A utility bill is randomly selected. Find the monthly utility bills that represents Q_3, the 75% percentile.

Example 4.15 The SAT scores (verbal + math) of all freshman students currently enrolled at B university are normally distributed with a mean of 976 and a standard deviation of 160.

(a) Find the 40th percentile of the SAT scores. **936 SAT score (*Tech*: 935.46)**

(b) If students with a SAT score in the top 5% will be offered a scholarship, then what is the minimum SAT score required to receive the scholarship? **1239.2 SAT score or better (*Tech*: 1239.18)**

(c) What GPA is at the 85th percentile? **1142.4 SAT score (*Tech*: 1141.83)**

(d) What is the cutoff for the 75th percentile? **1083.2 SAT score (*Tech*: 1083.92)**

(e) Find the interquartile range (IQR) of the GPAs. **IQR=$Q_3 - Q_1$ = 214.4 SAT score (*Tech*: 215.84)**

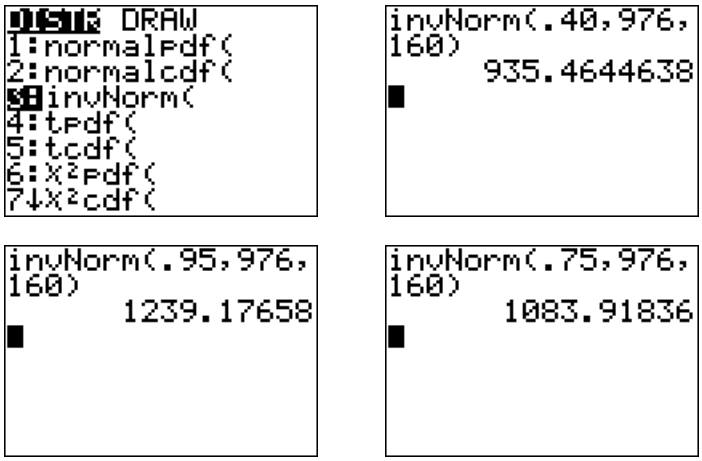


Figure 4.3: You can use a TI-83/84 to find a specific value x.

Example 4.16 The GPAs of all students enrolled at B university are normally distributed with a mean of 3.20 and a standard deviation of 0.31.

(a) The Dean of Students wants to place students with GPAs in the bottom 10% of the distribution on probation. What is the GPA cutoff for being placed on probation? **2.803 GPA**

(b) Find the 70th percentile of the GPAs. **3.361 GPA**

(c) If students with a GPA in the top 10% will be offered a scholarship, then what is the minimum GPA required to receive the scholarship? **3.597 GPA or better**

(d) What GPA is at the 95th percentile? **3.710 GPA**

(e) What is the cutoff for the 75th percentile? **3.408 GPA**

(f) Find the interquartile range (IQR) of the GPAs. **IQR=$Q_3 - Q_1$ = 0.415 GPA**

Example 4.17 Suppose that you are in charge of ordering caps and gowns for senior graduation in May. You know that the heights of the population of students at your college have a mean of 68 inches and a standard deviation of 3 inches.

(a) If the distribution can be considered mound-shaped and symmetric, what percentage of students are between 62 and 71 inches tall? What about the percentage between 60 and 70 inches tall?

Suppose you know that the distribution is normally distributed.

(b) What percentage of students are between 60 and 70 inches tall?

(c) The gown manufacturer wants to know how many students will need to special order their gowns because they are very tall. Find the proportion of students who are above 74 inches tall.

(d) Suppose that you wanted only the tallest 1% of the students to have to special-order gowns. What is the height at which tall students will have to special-order their gowns?

Suppose that a variable of interest x is normally distributed.

- To find the area (or probability) under a normal distribution for a given value x, we first convert x to a z-score $z = \dfrac{x - \mu}{\sigma}$ and then use the standard normal distribution table to find the area.
- But if we are given a probability and want to find a value x, we first find a z-score that corresponds to the given probability (or percentile) and then convert the z-score to a value x by $x = \mu + z\sigma$.

Central Limit Theorem

Sampling Distribution of Sample Means

A population distribution is the distribution of the population data. **A sampling distribution of sample means** $\bar{x}$ is the distribution of a sample mean that is formed when samples of size n are repeatedly taken from a population.

Example 4.18 Consider the population $\{1, 3, 5, 7\}$. A random sample of $n = 2$ is selected from the population. The followings are the all possible samples and the means of each sample.

possible samples	$\bar{x}$	possible samples	$\bar{x}$	possible samples	$\bar{x}$	possible samples	$\bar{x}$
1,1	1	3,1	2	5,1	3	7,1	4
1,3	2	3,3	3	5,3	4	7,3	5
1,5	3	3,5	4	5,5	5	7,5	6
1,7	4	3,7	5	5,7	6	7,7	7

(a) Find the mean of the distribution of all possible sample means.

(b) Find the standard deviation of the distribution of all possible sample means.

Properties of Sampling Distribution of Sample Means $\bar{x}$

The mean and standard deviation of the sampling distribution of $\bar{x}$ are denoted by $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, respectively.

- The mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population. Thus
$$\mu_{\bar{x}} = \mu$$
- The standard deviation of the sampling distribution of $\bar{x}$ is
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
where σ is the standard deviation of the population and n is the sample size.

Example 4.19 The GPAs of all students enrolled at A university have a normal distribution with a mean of 3.10 and a standard deviation of 0.28. Let $\overline{x}$ be the sample mean GPA for a random sample of 25 students selected from the university. Find the mean and standard deviation of the sampling distribution of $\overline{x}$.

SHAPE OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

1. If the population itself is normally distributed, the sampling distribution of sample means is **normally** distributed for *any* sample size n.
2. **(Central Limit Theorem)** For a large sample size ($n \geq 30$), the sampling distribution of $\bar{x}$ is **approximately normal**, irrespective of the shape of the population distribution.
3. In either case, the mean and standard deviation of the sampling distribution of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The actual mathematical derivation of sampling distributions is one of the basic problems of mathematical statistics, **STAT U512**.

Example 4.20 The GPAs of all students enrolled at A university have a normal distribution with a mean of 3.10 and a standard deviation of 0.28. Let $\overline{x}$ be the sample mean GPA for a random sample of 25 students selected from the university. Find the mean and standard deviation of the sampling distribution of $\overline{x}$ and comment on the shape of its sampling distribution.

Example 4.21 The weight of all people living in a town have a distribution that is skewed to the right with a mean of 133 pounds and a standard deviation of 24 pounds. Let $\bar{x}$ be the mean weight of a random sample of 45 persons selected from the town. Find the mean and standard deviation of $\bar{x}$ and comment on the shape of its sampling distribution.

Example 4.22 Phone bills for residents of a certain city have a mean of \$64 and a standard deviation of \$9.

(a) Random samples of 16 phone bills are drawn from this population, and the mean of each sample is determined. Find the mean and standard deviation of the mean phone bills of the sampling distribution and comment on the shape of its sampling distribution. **$\mu_{\overline{x}} =$ \$64, $\sigma_{\overline{x}} =$ \$2.25. Because of the small sample size $n = 16 < 30$ from unknown shape of the population, the shape of the sampling distribution of $\overline{x}$ may not be determined.**

(b) This time random samples of 36 phone bills are drawn from this population, and the mean of each sample is determined. Find the mean and standard deviation of the mean phone bills of the sampling distribution and comment on the shape of its sampling distribution. **$\mu_{\overline{x}} =$ \$64, $\sigma_{\overline{x}} =$ \$1.50. Because of the large sample size $n = 36 > 30$, the sampling distribution of $\overline{x}$ is approximately normal.**

Applications of the Sampling Distributions of $\bar{x}$

To transform $\bar{x}$ to a z-score, one can use the formula

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Example 4.23 The GPAs of all students enrolled at A university have a normal distribution with a mean of 3.10 and a standard deviation of 0.28.

(a) Find the percentage of students at the university who have a GPA greater than 3.15.

(b) If a sample of 16 students is taken, what is the probability that the **sample mean** GPA will be greater than 3.15?

(c) If a sample of 36 students is taken, what is the probability that the sample mean GPA will be greater than 3.15?

Example 4.24 The GPAs of all students enrolled at B university are normally distributed with a mean of 2.98 and a standard deviation of 0.36.

(a) Find the percentage of students at the university who have a GPA greater than 3.15. **31.92%**

(b) If a sample of 16 students is taken, what is the probability that the sample mean GPA will be greater than 3.15? **2.94%**

(c) If a sample of 36 students is taken, what is the probability that the sample mean GPA will be between 2.90 and 3.15? **87.88%**

Example 4.25 The population of the ages of all U.S. college students is skewed to the right with a mean age of 26.4 years and a standard deviation of 4.8 years. Determine the probability that a random sample of 49 students selected from the population will have a sample mean age within one year of the population mean age?

Example 4.26 The time that college students spend studying per week has a distribution of 8.4 hours and a standard deviation of 2.7 hours. Find the probability that the mean time spent studying per week for a random sample of 25 students would be less than 8 hours. **Because of the small sample size $n = 25 < 30$ from unknown shape of the population, the sampling distribution of $\overline{x}$ may not be approximated by a normal distribution. So the probability of $P(\overline{x} < 8)$ can't be determined.**

Example 4.27 A brake pad manufacturer claims its brake pads will have a mean life of 38,000 miles and a standard deviation of 1000 miles. You work for a consumer protection agency and you are testing this manufacturer's brake pads. You randomly select 50 brake pads.

In your tests, the mean life of the brake pads is 37,650 miles. Assuming the manufacturer's claim is correct, what is the probability the mean of the sample is 37,650 miles or less?

Example 4.28 The quality-control manager of a restaurant analyzed the length of time that a car spends at the drive-through window waiting for an order. The distribution of time spent at the window is skewed right with the mean time of 59.3 seconds and the standard deviation of 13.1 seconds. The manager wishes to use a new delivery system designed to get cars through the drive-through system faster. A random sample of 40 cars results in a sample mean time spent at the window of 56.8 seconds. What is the probability of obtaining a sample mean of 56.8 seconds or less, assuming that the population mean is 59.3 seconds? Do you think that the new system is effective?

Example 4.29 The GPAs of all students enrolled at E university are skewed to the right with a mean of 2.90 and a standard deviation of 0.39.

(a) If a sample of 9 students is taken, what is the probability that the sample mean GPA will be greater than 3.00? **Because of the small sample size $n = 9 < 30$ from the skewed-right population, the sampling distribution of $\overline{x}$ may not be approximated by a normal distribution. So the probability of $P(\overline{x} > 3.0)$ can't be determined.**

(b) If a sample of 40 students is taken, what is the probability that the sample mean GPA will be between 2.80 and 3.00? **89.48%**

Finding probabilities for x and $\overline{x}$

Example 4.30 The SAT scores (verbal + math) of all freshman students currently enrolled at C university have a **right-skewed** distribution with a mean of 1005 and a standard deviation of 140.

(a) Find the percentage of freshman students at the university who have a SAT score greater than 1050.

(b) If a random sample of 49 freshman students are drawn from this university, what is the probability that the sample mean SAT score will be greater than 1050?

Example 4.31 The SAT scores (verbal + math) of all freshman students currently enrolled at C university are **normally** distributed with a mean of 1005 and a standard deviation of 140.

(a) Find the percentage of freshman students at the university who have a SAT score greater than 1050.

(b) If a random sample of 49 freshman students are drawn from this university, what is the probability that the sample mean SAT score will be greater than 1050?

Example 4.32 The monthly utility bills in a certain city are distributed, with mean of \$100 and a standard deviation of \$12.

(a) Find the probability that a randomly selected utility bill is greater than \$105. **Because the shape of the distribution of the utility bills is not provided, the probability of $P(x > \$105)$ can't be determined.**

(b) You randomly select 36 utility bills. What is the probability that their mean utility bill is greater than \$105? **0.62%**

Example 4.33 The monthly utility bills in a certain city are **normally** distributed with mean of \$100 and a standard deviation of \$12.

(a) Find the probability that a randomly selected utility bill is greater than \$105. **33.72%** (*Tech*: **33.85%**)

(b) You randomly select 25 utility bills. What is the probability that their mean utility bill is greater than \$105? **1.88%** (*Tech*: **1.86%**)

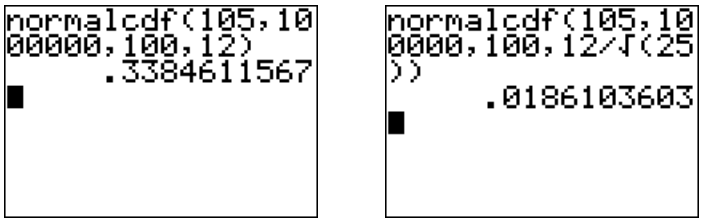


Figure 4.4: You can use a TI-83/84 to find probabilities $P(x)$ and $P(\overline{x})$.

Example 4.34 Air travel from JFK New York to Sarasota, Florida takes an average 150 minutes with a standard deviation 38 minutes. What is the probability that a random sample of 64 flights from JFK to Sarasota will take more than 160 minutes on average? **1.74%** (*Tech*: **1.76%**)

Example 4.35 The population mean weight of newborn babies for a western suburb is 7.4 lbs with a standard deviation of 0.8 lbs. What is the probability that a sample of 64 newborns selected at random will have a mean weight greater than 7.5 lbs? **15.87%** (*Tech*: **15.87%**)

- To find probabilities for individual members of a population with a normally distributed x, use the formula $z = \dfrac{x - \mu}{\sigma}$ for $P(x)$.
- To find probabilities for the mean $\overline{x}$ of a sample size n, use the formula $z = \dfrac{\overline{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ for $P(\overline{x})$.

4.2 A Useful Discrete Random Variable: Binomial Distribution

PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE
lists *all the possible values* that the random variable can assume and their *corresponding probabilities.*

1. $0 \leq P(x) \leq 1$ for each value of x.
2. $\sum P(x) = 1$

Example 4.36 A random sample of graduating seniors was surveyed just before graduation. One question that was asked is: How many times did you change majors? The results are displayed in a probability distribution.

x	0	1	2	3	4	5
$P(x)$	0.28	0.37	0.23	0.09	0.02	0.01

Example 4.37 The following table lists the probability distribution of the number of television sets owned by certain families.

x	0	1	2	3
$P(x)$	.10	.05	.45	

(a) Find the probability that the number of television sets is

- three
- at least 1

(b) Find the probability $P(1 \leq x < 3)$.

(c) Find the mean number of television sets.

Although we omit the derivations, the mean and standard deviation of a discrete population probability distribution are

$$\mu = \sum xP(x),$$
$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)}.$$

μ is also called **the expected value** of x, $E(x)$.

Binomial Distribution

Characteristics of a Binomial Random Variable:

1. The experiment consists of n identical trials.
2. There are only two possible outcomes on each trial. We will denote one outcome a success and the other a failure.
3. The probability of success remains the same from trial to trial. This probability is denoted by π.
4. The trials are independent.
5. The binomial random variable x is the number of successes in n trial.

The probability of observing x successes in n trials of a binomial experiment is

$$P(x) = {}_nC_x\pi^x(1-\pi)^{n-x} = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} \qquad (x = 0, 1, 2, \cdots, n)$$

$$\begin{aligned} n &= \text{total number of trials} \\ \pi &= \text{probability of success} \\ x &= \text{number of successes in } n \text{ trials} \\ n-x &= \text{number of failures in } n \text{ trials} \end{aligned}$$

FACTORIALS

The symbol $n!$, read as "n factorial", represents the product of all the integer from n to 1. In other words,

$$n! = n(n-1)(n-2)(n-3)\cdots 3 \cdot 2 \cdot 1$$

By definition, $0! = 1$

Example 4.38 Draw 4 balls with replacement from a box that contains 10 balls, 6 of which are red and 4 are blue, and observe the colors of the drawn balls.

(a) What is the probability that exactly *two* of red balls are drawn?

(b) Find the probability distribution for the number of red ball drawn.

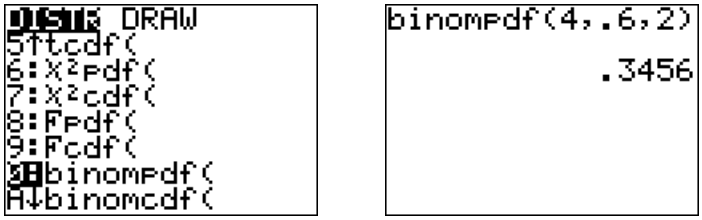


Figure 4.5: You can use a TI-83/84 to find $P(X = 2)$.

Example 4.39 Determine whether the given procedure results in a binomial distribution.

(a) Rolling a single die 19 times, keeping track of the numbers that are rolled.

(b) Rolling a single die 19 times, keeping track of the "fives" rolled.

(c) Choosing 4 marbles from a box of 40 marbles (20 purple, 12 red, and 8 green) one at a time with replacement, keeping track of the number of red marbles chosen.

(d) Choosing 5 marbles from a box of 40 marbles (20 purple, 12 red, and 8 green) one at a time with replacement, keeping track of the colors of the marbles chosen.

Example 4.40 A professional basketball player makes 80% of the free throws he tries. Assuming this percentage will hold true for future attempts, find the probability that in the next ten tries the number of free throws he will make is

(a) exactly 6

(b) at least 6

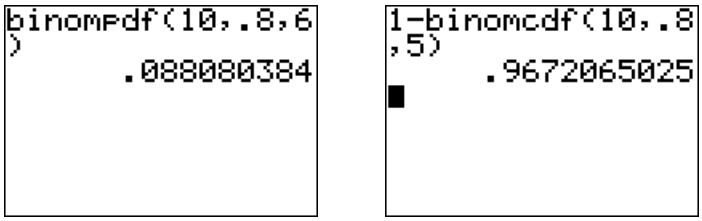


Figure 4.6: You can use a TI-83/84 to find $P(X = 6)$ and $P(X \geq 6)$.

Example 4.41 Sixty percent of all small businesses in the United States have a website. If you randomly select 50 small businesses, what is the mean number (or expected number) of small businesses that will have a website.

Although we omit the derivations, we give the formulas for the binomial distribution. The mean and standard deviation of a binomial probability distribution are obtained by

$$\begin{aligned} \text{Mean } \mu &= n\pi \\ \text{Standard deviation } \sigma &= \sqrt{n\pi(1-\pi)} \end{aligned}$$

Normal Approximation to the Binomial Distribution

Finding probabilities for a binomial random variable becomes more difficult when n gets large. Although we omit the derivations,

- For large n and π with $n\pi \geq 5$ and $n(1-\pi) \geq 5$, the Binomial distribution can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$.
- Thus, the normal distribution can be used to approximate the binomial distribution when n is large.
- A **continuity correction** will improve the quality of the approximation. The general idea of the continuity correction is to add or subtract 0.5 from a binomial value before using normal probabilities.

Example 4.42 A product is manufactured in batches of 100 and the overall rate of defects is 5%. Estimate the probability that a randomly selected batch contains more than 5 defects.

Example 4.43 Thirty percent of people in the U.S. say they are confident that passenger trips to the moon will occur in their lifetime.

(a) You randomly select 200 people in the U.S. and ask each if he or she thinks passenger trips to the moon will occur in his or her lifetime. What is the probability that at least 50 will say yes?

(b) You randomly select 1000 people in the U.S. and ask each if he or she thinks passenger trips to the moon will occur in his or her lifetime. What is the probability that at least 310 will say yes?