Lec 1

# MATH 3070 Applied Statistics I

Nan Shen



## 1 Overview and Descriptive Statistics

**What is Statistics?**

- 

- 

### 1.1 Populations, Samples, and Processes

An investigation will typically focus on _____

_____.

There are two basic methods for studying a population:

- 

- 

**What is a variable?**

A variable is _____

Note:

- 
- 

Data results from making observations either on a single variable or simultaneously on two or more variables.

A **univariate** data set _____

- e.g.

A **bivariate** data set _____

- e.g.

A **multivariate** data set _____

- e.g.

**Types of variables:**

- Categorical: _____

  _____

- Quantitative: _____

– Discrete: _____

_____

– Continuous: _____

_____

Summary:

**Branches of Statistics:**

1. **Descriptive Statistics**, _____

A. Some of these methods are _____ in nature; e.g. _____

_____

B. Other descriptive methods involve calculation of _____ summary measures, e.g. _____

2. Having obtained a sample from a population, an investigator would frequently like to

_____

Techniques for generalizing from a sample to a population are gathered within the branch of our discipline called _____

Lec 2

## 1.2 Pictorial and Tabular Methods in Descriptive Statistics

### Histograms

Histograms are _____.

To construct a histogram, the first step is to "bin" the range of values - that is,

_____

and then _____

The bins are usually specified as _____

The bins must be _____ and are often (but not required to be) of _____

If the bins are of equal size, a _____ is erected over the bin with _____

proportional to the _____ - the number of cases in each bin.

Consider data consisting of observations on a discrete variable $x$. The **frequency** of any particular $x$ value is _____

The **relative frequency** of a value is _____

relative frequency of a value = _____

### Relative Frequency Histograms

A histogram may also be normalized to display _____.

It then shows the proportion of cases that fall into each of several categories, with _____

_____ So relative frequency histograms are

bar charts of the _____.

**Example 1.** A website gives information on 50 charities. Here is a sample of 20 charities and the amount (in thousands) they spend on fundraisers.

20, 10, 5, 1, 2, 19, 18, 2, 6, 29, 35, 11, 23, 13, 31, 32, 35, 25, 26, 22

Find the histogram and relative frequency histogram

**Describing Histogram Shapes**

A **unimodal** histogram is one that _____

A **bimodal** histogram _____

A histogram _____ is said to be **multimodal**.

And a histogram with _____ is said to be **uniform**.

A histogram is **symmetric** if _____

A unimodal histogram is _____ if the right or upper tail is stretched out compared with the left or lower tail,

and _____ if the stretching is to the left.

**Example 2.** Let's take a look at the problem on HW1

1. [7.5 points] A small survey was conducted in which each respondent was asked how many times, in the previous two-week period, they had eaten at a fast food restaurant. The data appear below.

   0, 2, 1, 5, 2, 2, 3, 4, 1, 2, 7, 1, 3, 4, 1, 0, 1, 4, 2, 1, 3, 3, 2, 1, 9, 1

   (a) Construct a frequency histogram. The histogram should be neat, accurate, and well-labeled. [3.5 points]
   (b) How would you describe the shape of the distribution? [1 point]

## 1.3 Measures of Location

Suppose, that our data set is of the form $x_1, x_2, \cdots, x_n$, where each $x_i$ is a number. One important characteristic of such a set of numbers is its _____, and in particular its _____.

### 1.3.1 Sample Mean

For a given set of numbers $x_1, x_2, \cdots, x_n$, the most familiar and useful measure of the center is the _____, or _____ of the set.

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \cdots, x_n$ is given by

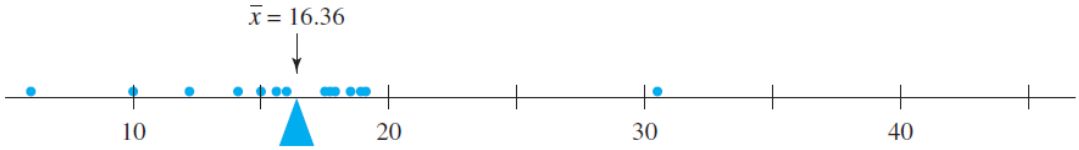
A physical interpretation of the sample mean demonstrates how it assesses the center of a sample. Think of each dot in the dotplot below representing a 1-lb weight. Then a fulcrum placed with its tip on the horizontal axis will _____ precisely when it is located at $\bar{x}$. So the sample mean can be regarded as the _____ of the distribution of observations.

**EXAMPLE 1.14** Here are the 24-hour water-absorption percentages for the specimens:

| | | | | |
|---|---|---|---|---|
| $x_1 = 16.0$ | $x_2 = 30.5$ | $x_3 = 17.7$ | $x_4 = 17.5$ | $x_5 = 14.1$ |
| $x_6 = 10.0$ | $x_7 = 15.6$ | $x_8 = 15.0$ | $x_9 = 19.1$ | $x_{10} = 17.9$ |
| $x_{11} = 18.9$ | $x_{12} = 18.5$ | $x_{13} = 12.2$ | $x_{14} = 6.0$ | |

With $\Sigma x_i = 229.0$, the sample mean is

$$\bar{x} = \frac{229.0}{14} = 16.36$$




Dotplot of the data from Example 1.14

### 1.3.2 Sample Median

The **sample median** is the _____

The sample median _____ is obtained by _____.

Then,

**Example 3.**

(1) Suppose we have a data set as: 1.6, 3.0, 1.9, 0.6, 3.8. What are the mean and median of this sample?

(a)

(b)

(2) Suppose we have a data set as: 6.9, 16.3, 32.8, 41, 47.7, 48.9. What is the median of this sample?

### 1.3.3 Population Mean

The average of all values in the population is called _____ and

is denoted by _____. When there are $N$ values in the population, then $\mu =$ _____

One of our first tasks in statistical inference will be to present methods based on the

_____ for drawing conclusions about a _____

For example,

### 1.3.4 Population Median

Analogous to $\widetilde{x}$ as the middle value in the sample, the **population median** is _____

denoted by _____. As with $\bar{x}$ and $\mu$, we can think of using _____,

to make an inference about _____.

**Summary:**

Lec 3

### 1.3.5 Other Measures of Location

**Maximum**

**Minimum**

An **outlier** is _____

Sometimes _____ are outliers in the data set.

**Quartiles and Percentiles**

Quartiles_____

_____

with the observations above the third quartile constituting the upper quarter of the data set, the second quartile being identical to the _____, and the first quartile separating the lower quarter from the upper three-quarters.


If the quantiles divide the data into _____, then they're called _____.

### 1.3.6 The effect of skewness on the mean and median

Think about the graph in extreme cases.

Suppose we have 11 data in a set of students grades, in which ten of them are 60s and one of them is 100.

On the other hand, suppose we have 11 data in a set of students grades, in which ten of them are 100s and one of them is 60.

**Note.**

**Example 4.** Suppose we have 10 people in a room, the mean salary $\bar{x}$ is $\$105,000$, and the median salary $\widetilde{x}$ is $\$65,000$. What can we say about the distribution?

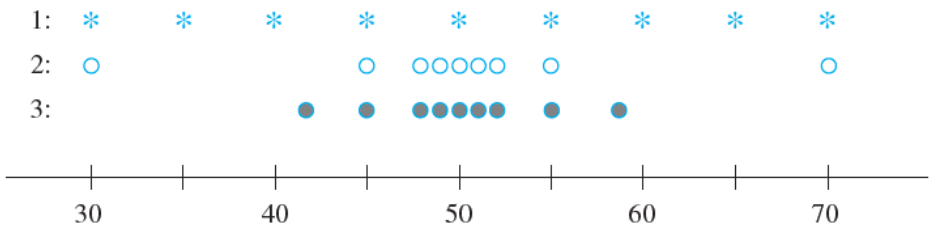
*Ans* :

**What is the more appropriate measure of center when there are extreme valus in the data set and why?**

*Ans* :

## 1.4 Measures of Variability

Figure below shows dotplots of three samples with the same mean and median, but the extent of spread about the center is different for all three samples. The first sample has the _____, the third has the _____, and the second is _____




Samples with identical measures of center but different amounts of variability

**Measures of Variability for Sample data**

The simplest measure of variability in a sample is the _____, which is _____

_____

- 

The value of the range for sample 1 in Figure above is _____ than sample 3, reflecting _____ in the first sample than in the third. A defect of the range, though, is _____

Samples 1 and 2 in Figure above have _____, yet when the observations between the two extremes are taken into account, there is _____ in the second sample than in the first.

Our primary measures of variability involve _____

A deviation will be **positive** (+) if _____ and **negative** (−) if _____

If all the deviations are _____, then all $x_i$'s are _____ _____ and there is _____. Alternatively, if some of the deviations are _____, then some $x_i$'s _____, suggesting a _____

A simple way to combine the deviations into a _____ is to _____.

Unfortunately, this is a bad idea:

sum of deviations =

so that the average deviation is_____.

To prevent _____from counteracting one another when they are combined, we consider instead _____

To get the _____, for several reasons we will not cover here, we divide the sum of squared deviations by _____ instead of _____.

**Definition 1.** The sample **variance**, denoted by _____, is given by

$$s^2 =$$

The sample **standard deviation(SD)**, denoted by _____, is the _____ of the variance:

$$s =$$

The _____ is our preferred measure of variability, because

_____

A shortcut formula for $s$,

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

This is because we can write $s^2$ as

**Example 5.** Suppose we have a data set $x_1, x_2, \cdots, x_5$ where $\sum_{i=1}^{5} x_i = 10.9$ and $\sum_{i=1}^{5} x_i^2 = 29.97$. What is the SD?

*Ans* :

**Properties of the mean and SD:**

Let $x_1, x_2, \cdots, x_n$ be a sample and $c$ be any nonzero constant.

1. If $y_1 = x_1 + c,\ y_2 = x_2 + c, \cdots, y_n = x_n + c,$
   then mean _____
   the sample variance _____
   the SD _____
2. If $y_1 = cx_1,\ y_2 = cx_2, \cdots, y_n = cx_n,$
   then mean _____
   the sample variance _____
   the SD _____

Lec 4

# 2 Probability

The term **probability** refers to _____

With a small number of observations, outcomes of random phenomena may look _____ from what you expect. As we make more observations, the _____ of times that a particular outcome occurs gets closer and closer to a certain number we would expect.

With any random phenomena, the probability of a particular outcome is _____

_____

**NOTE:** A random phenomenon has the characteristic that is _____

## 2.1 Sample Spaces and Events

An **experiment** is _____

Experiments that may be of interest include:

- tossing a coin once or several times,
- selecting a card or cards from a deck,
- weighing a loaf of bread.

**Definition 2.** The **sample space** of an experiment, denoted by $S$, _____

_____

**Example 6.**

- Experiment: Observing the tosses of two fair coins.

  $S =$

- Experiment: Flip a fair coin until a tail appears for the first time

  $S =$

- Experiment: Flip a fair coin until the first tail and record the number of heads that have occurred.

  $S =$

- Experiment: Observe the highest temperature for today:

  $S =$

- Experiment: Randomly select an American household and record the number of TV sets.

  $S =$

In our study of probability, we will be interested not only in the individual outcomes of $S$

but also in _____.

**Definition 3.** An **event** is _____

An event is _____ if it consists of exactly one outcome and _____

if it consists of more than one outcome.

When an experiment is performed, a particular event $A$ is said to occur if _____

_____

**Example 7.** Experiment: Tossing a coin 3 times. The sample space is

$S =$

Suppose our event

$$A = \text{ First toss gives head.}$$

Then $A$ occurs only if the resulting experimental outcome is contained in the set

$$\{ \qquad\qquad\qquad\qquad\qquad\qquad \}$$

**Some relations from Set theory** An event is just a set, so relationships and results from elementary set theory can be used to study events.

**Definition 4.**

- The **complement** of an event $A$, denoted by _____, is the set of

  _____

  Venn diagram:

- The **union** of two events $A$ and $B$, denoted by _____, is the event _____

  _____

  So the union includes outcomes for which _____ $A$ and $B$ occur as well as outcomes

  for which _____, that is, all outcomes in _____.

  Venn diagram:

- The **intersection** of two events $A$ and $B$, denoted by _____, is the event consisting

  of _____.

  Venn diagram:

- Two events are **mutually exclusive or disjoint**, if _____;

  i.e. both events can not happen at the same time

If events $A$ and $B$ are mutually exclusive, then

Venn diagram:

**Example 8.** For the experiment in which the number of pumps in use at a single six-pump gas station is observed, let $A = \{0,\ 1,\ 2,\ 3,\ 4\}, B = \{3,\ 4,\ 5,\ 6\}$, and $C = \{1,\ 3,\ 5\}$. Then

$A^c =$

$A \cup B =$

$A \cup C =$

$A \cap B =$

$A \cap C =$

$(A \cap C)^c =$

## 2.2 Axioms, interpretations, and Properties of Probability

### Interpreting Probability

Consider an experiment that can be _____

and let $A$ be an event consisting of a set of outcomes of the experiment. For examples,

the coin-tossing experiment previously discussed. If the experiment is performed _____,

let _____ denote the number times on which $A$ occurs. Then the ratio _____

is called _____ of the event $A$ in the sequence of $n$ replications.

Given an experiment, the objective of probability is to _____,

called the probability of the event $A$, which will give a precise measure of the chance that $A$ will occur.

Relative frequency will stabilize as the number of replications $n$ increases. That is, as

$n$ gets arbitrarily large, _____

_____.

The objective interpretation of probability identifies this limiting relative frequency with $P(A)$, i.e. as $n \to \infty$

$$P(A) = \text{\_\_\_\_\_}$$

**NOTE:** $0 \leq P(A) \leq 1$. Why?

The assignment of probabilities should satisfy the following axioms of probability.

**Probability Axioms**

1.

2.

3. If $A_1$, $A_2$, ... is an infinite collection of disjoint events, then

Lec 5

**More Probability Properties**

(1)

(2)

(3)

**Example 9.** Suppose we flip a fair coin until a head appears for the first time. What is the probability that more than one flip of the coin is required?

*Solution.* The sample space $S$ is

$$S =$$

So the event $A$=more than one flip of the coin is required contains the outcomes

$$A =$$

Since each of the outcomes in $A$ are can not occur simultaneously, _____

---

Therefore,

$$P(A) =$$

However, if we use property (3), we can solve the problem very quick by noticing that

(4) For any two events $A$ and $B$, probability of either event $A$ **or** event $B$ occurring is

- Special case:

**Example 10.** (Example 2.14 from textbook page 62) In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company. If a household is randomly selected,

**Question:**
1. what is the probability that it gets at least one of these two services from the company?

2. what is the probability that it gets exactly one of these services from the company?

*Solution.*

**Example 11.** (Exercise 11 from textbook page 64) A mutual fund company offers its customers a variety of funds: a money-market fund, three different bond funds (short, intermediate, and long-term), two stock funds (moderate and high-risk), and a balanced fund.

Among customers who own shares in just one fund, the percentages of customers in the different funds are as follows:

| | | |
|---|---|---|
| Money-market 20% | High-risk stock 18% | |
| Short bond 15% | Moderate-risk stock 25% | |
| Intermediate bond 10% | Long bond 5% | Balanced 7% |

A customer who owns shares in just one fund is randomly selected.

**Question:**

**a.** What is the probability that the selected individual owns shares in the balanced fund?

**b.** What is the probability that the individual owns shares in a bond fund?

**c.** What is the probability that the selected individual does not own shares in a stock fund?

*Solution.*

**Example 12.** (Exercise 14 from textbook page 65) Suppose that 55% of all adults regularly consume coffee, 45% regularly consume carbonated soda, and 70% regularly consume at least one of these two products.

**Question:**

**a.** What is the probability that a randomly selected adult regularly consumes both coffee and soda?

**b.** What is the probability that a randomly selected adult does not regularly consume at least one of these two products?

*Solution.*

## 2.4 Conditional Probability

In this section, we examine _____

_____ We will use the notation _____

to represent _____

**Definition 5.** For any two events $A$ and $B$ with $P(B) > 0$, the **conditional probability of $A$ given that $B$ has occurred** is defined by

**Example 13.** (Example 2.25 from textbook page 76) Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and battery.

Consider randomly selecting a buyer and let $A$ = memory card purchased and $B$ = battery purchased.

Move it earlier in 2.3 " More probability properties part"
(5)

**DeMorgan's Laws**:

1. $(A \cap B)^c =$
2. $(A \cup B)^c =$

Thus,

$$P(A^c \cap B^c) =$$

(6) By distributive law we have

$$A =$$

Since _____ and _____ are disjoint, by Probability Axiom 3, we have

$$P(A) =$$

This gives

$$P(A \cap B^c) =$$

Venn diagram:

End

Lec 6

The definition of conditional probability yields the following result.

The Multiplication rule

**Example 14.** (Example 2.29 from textbook page 78) An electronics store sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?
2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?
3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player?

*Solution.*

Lec 7

Problem 2 in Example 14 is an example of the Law of Total Probability.

**The Law of Total Probability**

NOTE: A set of events is jointly **exhaustive** if at least one of the events must occur.

**Example 15.** (Example 2.30 from textbook page 81) An individual has 3 different email accounts. Most of her messages, in fact 70%, come into account # 1, whereas 20% come into account #2 and the remaining 10% into account #3. Of the messages into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5%, respectively.

**Question:** What is the probability that a randomly selected message is spam?

*Solution.*

**Bayes' theorem**
Let $A_1, A_2, \cdots, A_k$ be a collection of $k$ _____

_____ Then for any other

event $B$ for which $P(B) > 0$, the posterior probability of $A_j$ given that $B$ has occurred is

The transition from the second to the third expression rests on using the multiplication rule in the numerator and the law of total probability in the denominator.

**Example 16.** (Example 2.31 from textbook page 81) Incidence of a rare disease. Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time (the sensitivity of this test is 99% and the specificity is 98%; in contrast, the Sept. 22, 2012 issue of The Lancet reports that the first at-home HIV test has a sensitivity of only 92% and a specificity of 99.98%).

**Question:**

If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

*Solution.*

**Example 17.** (Exercise 59 from textbook page 84) At a certain gas station, 40% of the customers use regular gas ($A_1$), 35% use plus gas ($A_2$), and 25% use premium gas($A_3$). Of those customers using regular gas, only 30% fill their tanks (event B). Of those customers using plus, 60% fill their tanks, whereas of those using premium, 50% fill their tanks.

**Question:**

**a.** What is the probability that the next customer will request plus gas and fill the tank ($A_2 \cap B$)?

**b.** What is the probability that the next customer fills the tank?

**c.** If the next customer fills the tank, what is the probability that regular gas is requested? Plus? Premium?

*Solution.*

**Example 18.** (Exercise 60 from textbook page 84) Seventy percent of the light aircraft that disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 60% have an emergency locator, whereas 90% of the aircraft not discovered do not have such a locator. Suppose a light aircraft has disappeared.

**Question:**

**a.** If it has an emergency locator, what is the probability that it will not be discovered?

**b.** If it does not have an emergency locator, what is the probability that it will be discovered?

*Solution.*

## 2.5 Independence

We say $A$ and $B$ are independent events, meaning that _____

_____

**Definition 6.** Two events $A$ and $B$ are **independent** if

and are dependent otherwise.

**NOTE:** This definition implies if $A$ and $B$ are independent

- 
- 

It is also straightforward to show that if $A$ and $B$ are independent, then so are the following pairs of events:

(1)

(2)

(3)

Lec 8

**Example 19.** Consider an experiment where the next car sold from a dealership is observed. Let $C$ = a car has a CD player, $M$ = a car has a manual transmission.

Given: $P(C) = 0.75, \quad P(M) = 0.15, \quad P(M \cup C) = 0.85.$

**Question:**
**1.** Are $C$ and $M$ mutually exclusive/disjoint events?
**2.** Are $C$ and $M$ independent events?

*Solution.*

**Example 20.** (Exercise 74 from textbook page 89) The proportions of blood phenotypes in the U.S. population are as follows:

| $A$ | $B$ | $AB$ | $O$ |
|---|---|---|---|
| 0.40 | 0.11 | 0.04 | 0.45 |

Assuming that the phenotypes of two **randomly selected** individuals are **independent** of one another.

**NOTE:** randomly selected $\Longrightarrow$ independent

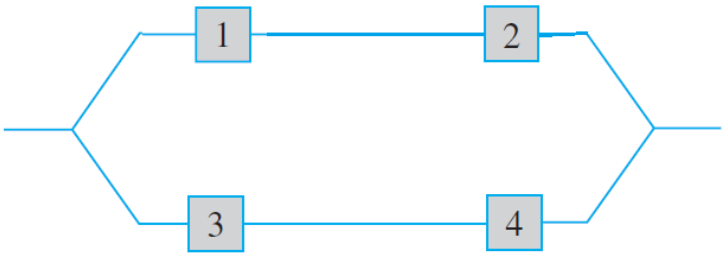
**Question:**

**1.** What is the probability that both phenotypes are O?

**2.** What is the probability that the phenotypes of two randomly selected individuals match?

*Solution.*

**Example 21.** The entire system will work if either the top two components both work or the bottom two components both work. If components work independently of one another and $P(\text{component } i \text{ fails}) = 0.4$ for $i = 1, 2, 3, 4$.




**Question:**

**1.** What is the probability that all four components fails?

**2.** What is the probability that exactly one of the components fails?

**3.** What is the probability that at least one of the components fails?

**4.** What is the probability that at most one of the components fails?

**5.** What is the probability that the system works?

*Solution.*

Lec 9

# 3 Discrete Random Variables and Probability Distributions

## 3.1 Random Variables

**Definition 7.** A function, $f$, is a rule that takes an input value and returns an output.

For example, $y = x + 2$. Then if $x = 1$ then we have $y = 3$.

**Definition 8.** A random variable is

**Example 22.** Suppose our experiment is tossing a coin two times. Then the sample space is

$$S =$$

If we define the random variable

$$X = \text{the number of heads you get in one experiment}$$

Then we can see that $X$ can take values _____.

**NOTE:** Random variables are usually denoted by uppercase letters, such as _____.

We will use lowercase letters to represent _____ of the

corresponding random variable. The notation $X(\omega) = x$ means that _____

_____.

We sometimes consider several different random variables from the same sample space.

**Example 23.** Two gas stations are located at a certain intersection. Each one has six gas pumps. Consider the experiment in which the number of pumps in use at a particular time of day is determined for each of the stations.

Define random variables $X$, $Y$, and $U$ by

$X =$ the total number of pumps in use at the two stations
$Y =$ the difference between the number of pumps in use at station 1 and station 2
$U =$ the maximum of the numbers of pumps in use at the two stations

If this experiment is performed and results $\omega = (2, 3)$, then

$$X =$$

$$Y =$$

$$U =$$

**Example 24.** When a student calls a university help desk for technical support, he/she will either immediately be able to speak to someone ($S$, for success) or will be placed on hold ($F$, for failure). With $S = \{S, F\}$, define an random variable $X$ by

$$X(S) = 1 \text{ and } X(F) = 0$$

The random variable $X$ indicates _____

**Example 25.** Suppose a location in the United States is selected. Define the random variable $Y$ by

$$Y = \text{ the height above sea level at the selected location}$$

Then the largest possible value of $Y$ is 14,494 (Mt. Whitney), and the smallest possible value is 2282 (Death Valley). The set of all possible values of $Y$ is the set of all numbers in the interval between 2282 and 14,494, that is,

and there are _____ in this interval.

**Two Types of Random Variables**

- _____. If the possible outcomes of a random variable can be listed out using a finite (or countably infinite) set of single numbers (Example 22, 23, 24), then the random variable is discrete.
- _____. If the possible outcomes of a random variable can only be described using an interval or union of intervals of real numbers (Example 25), then the random variable is continuous.

## 3.2 Probability Distributions for Discrete Random Variables

### 3.2.1 The Probability Mass Function

The **probability distribution** of $X$ says how the total probability of 1 is distributed among the various possible $X$ values. The probability distribution of $X$ lists all possible values of $X$ and their corresponding probabilities.

**Definition 9.** For discrete random variables, the probability list of $X$ is called _____, which is defined for every number $x$ by $p(x) = P(X = x) = P(\text{all } \omega \in S : X(\omega) = x)$.

The pmf returns the probability that the random variable $X$ is equal to the value $x$.

To be a valid pmf, we need:

(1)

(2)

**Example 26.** Suppose we toss a fair coin three times, and define the random variable $X$ to be the number of heads that appear. Find the pmf of $X$.

*Solution.* The sample space is

$$S =$$

So the possible values for the random variable $X$ are in the set $\{0, 1, 2, 3\}$. The pmf tells us all possible values of $X$ and their corresponding probabilities, i.e. $p(x) = P(X = x)$. Since we have a fair coin, so the _____ is assumed here. Therefore

**NOTE:** This is NOT a proper format of writing a pmf. Write it in a proper way, $X$ should define on _____. So The pmf of $X$ is given by

**Question:** Is this a valid pmf?

Check:

The graph pf the pmf of $X$ could be

**Example 27.** (Exercise #13 on page 107 of the textbook) A mail-order computer business has six telephone lines. Let $X$ denote the number of lines in use at a specified time. Suppose the pmf of $X$ is as given in the accompanying table.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $p(x)$ | .10 | .15 | .20 | .25 | .20 | .06 | .04 |

Calculate the probability of each of the following events.

a. {at most three lines are in use}

b. {fewer than three lines are in use}

c. {at least three lines are in use}

d. {between two and five lines, inclusive, are in use }

e. {between two and four lines, inclusive, are not in use}

f. {at least four lines are not in use}

*Solution.* Before we calculate the probabilities, let's check whether it is a valid pmf.

Lec 10

**Example 28.** (Example 3.8 on textbook page 100) Six boxes of components are ready to be shipped by a certain supplier. The number of defective components in each box is as follows:

| Box | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Number of defectives | 0 | 2 | 0 | 1 | 2 | 0 |

One of these boxes is to be randomly selected for shipment to a particular customer. Let $X$ be the number of defectives in the selected box.

### 3.2.2 The Cumulative Distribution Function

**Definition 10.** The **cumulative distribution function (cdf)**

**Example 29.** Experiment: rolling an unfair die

Define the random variable $X$ as the number on the upper face. Then the pmf of $X$ is give in the table

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $p(x)$ | 0.2 | 0.3 | 0.1 | 0.1 | 0.1 | 0.2 |

Then some of the probability we are interested in are

**Example 30.** Continue with our example of tossing a fair coin three times in Example 26. Find the CDF of $X$.

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $p(x)$ | 0.125 | 0.375 | 0.375 | 0.125 |

*Solution.*

Lec 11

**Question:** Given a CDF, how do we convert to pmf?

In Example 30,

**Example 31.** (Example 3.13 from textbook page 104) A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of $Y$ = the amount of memory in a purchased drive:

| $y$ | 1 | 2 | 4 | 8 | 16 |
|---|---|---|---|---|---|
| $p(y)$ | 0.05 | 0.1 | 0.35 | 0.4 | 0.1 |

Lec 12

## 3.3 Expected Values

As with a sample, there are descriptive statistics that can be used to describe the population.

**Definition 11.** Let $X$ be a discrete random variable with set of possible values $D$ and pmf $p(x)$. The **expected value** or **mean value** of $X$, denoted by _____, is

Recall that, previously we use the **mean** as a measure of the center of the data set, i.e. the arithmetic average. But now, the **mean** refers to the center of the population.

**Example 32.** Consider a university having 15,000 students and let $X$ = the number of courses for which a randomly selected student is registered. The pmf of $X$ follows.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $p(x)$ | 0.01 | 0.03 | 0.13 | 0.25 | 0.39 | 0.17 | 0.02 |

Calculate the expected value of $X$, i.e $E(X)$.

*Solution.*

Notice that $\mu$ here is not 4, the ordinary average of $1, \cdots, 7$, because the distribution puts more weight on 4, 5, and 6 than on other $X$ values.

**Example 33.** Let $X = 1$ if a randomly selected vehicle passes an emissions test and $X = 0$ otherwise. Then $X$ is a _____ with pmf _____, from which $E(X) =$ _____. That is, the expected value of $X$ is _____.

**The Expected Value of a Function**

If the random variable $X$ has a set of possible values $D$ and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$ is computed by

**Example 34.** The cost of a certain vehicle diagnostic test depends on the number of cylinders $X$ in the vehicle's engine. Suppose the cost function is given by $Y = h(X) = 20+3X+0.5X^2$. Calculate the expected value of $Y$. The pmf of $X$ is as follows:

| $x$ | 4 | 6 | 8 |
|---|---|---|---|
| $p(x)$ | 0.5 | 0.3 | 0.2 |

*Solution.*

**Example 35.** A computer store has purchased three computers of a certain type at \$500 apiece. It will sell them for \$1000 apiece. The manufacturer has agreed to repurchase any computers still unsold after a specified period at \$200 apiece. Let $X$ denote the number of computers sold, and suppose that $p(0) = 0.1,\ p(1) = 0.2,\ p(2) = 0.3$, and $p(3) = 0.4$. With $h(X)$ denoting the profit associated with selling $X$ units, the given information implies that $h(X) = \text{revenue} - \text{cost} = 1000X + 200(3 - X) - 1500 = 800X - 900$.

Find the expected profit.

*Solution.*

### Expected Value of a Linear Function

The $h(X)$ function of interest is quite frequently a linear function $aX + b$. In this case, $E[h(X)]$ is easily computed from $E(X)$ without the need for additional summation.

$$E(aX + b) =$$

*Proof.*

□

### The Variance of $X$

We will use the variance of $X$ to assess the amount of variability in (the distribution of) $X$, just as $s^2$ was used in Chapter 1 to measure variability in a sample.

**Definition 12.** Let $X$ be a discrete random variable with pmf $p(x)$ and expected value $\mu$. Then the **variance** of $X$, denoted by _____, is

The **standard deviation** (SD) of $X$ is

**Example 36.** A library has an upper limit of 6 on the number of DVDs that can be checked out to an individual at one time. Consider only those who currently have DVDs checked out, and let $X$ denote the number of DVDs checked out to a randomly selected individual. The pmf of X is as follows:

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $p(x)$ | 0.3 | 0.25 | 0.15 | 0.05 | 0.1 | 0.15 |

Find the variance of $X$.

*Solution.*

**A Shortcut Formula for $\sigma^2$**

$$\mathrm{Var}(X) =$$

*Proof.*

□

**Example 37.** (Example 36 continued)

**Variance of a Linear Function**

The variance of $aX + b$ is

The standard deviation of $aX + b$ is

**Example 38.** In Example 35,

**Example 39.** An dealer sells three different models of freezers having 13.5, 15.9, and 19.1 cubic feet of storage space, respectively. Let $X$ = the amount of storage space purchased by the next customer. Suppose that $X$ has pmf

| $x$ | 13.5 | 15.9 | 19.1 |
|---|---|---|---|
| $p(x)$ | 0.2 | 0.5 | 0.3 |

**a.** Compute $E(X)$, $E(X^2)$, $\text{Var}(X) = \sigma^2$ and standard deviation $\sigma$.

**b.** If the price of a freezer having capacity $X$ cubic feet is $25X - 8.5$, what is the expected price paid by the next customer to buy a freezer?

**c.** What is the variance and the standard deviation of the price $25X - 8.5$ paid by the next customer?

**d.** Suppose that although the rated capacity of a freezer is $X$, the actual capacity is $h(X) = X - 0.01X^2$. What is the expected actual capacity of the freezer purchased by the next customer?

*Solution.*

Lec 13

## 3.4 The Binomial Probability Distribution

Consider an experiment consisting of $n$ trials, each of which has exactly two outcomes.

**Example 40.** Suppose 2% of all items produced from an assembly line are defective. We randomly sample 50 items and count how many are defective (and how many are not).

**Definition 13.** A **binomial experiment** has the following characteristics:

(1)

(2)

(3)

(4)

Note: The probability of failure is _____.

More examples of binomial experiments:

- We toss a coin $n$ times, let $S$ = we observe a "head" and $F$ = we observe a "tail".
- Each of the next $n$ vehicles undergoing an emissions test, and let $S$ denote a vehicle that passes the test and $F$ denote one that fails to pass.
- Tossing a thumbtack $n$ times, with $S$ = point up and $F$ = point down.
- The gender ($S$ for female and $F$ for male) is determined for each of the next $n$ children born at a particular hospital.

### The Binomial random Variable and distribution

In most binomial experiments, it is _____, rather than knowledge of exactly which trials yielded $S$'s, that is of interest.

**Definition 14.** The binomial random variable $X$ associated with a binomial experiment consisting of $n$ trials is defined as

$X =$

Suppose, for example, that $n = 3$. Then there are eight possible outcomes for the experiment:

From the definition of $X$, we have that $X(SSF) =$ , $X(SFF) =$ , and so on.

Possible values for $X$ in an $n$-trial experiment are $x =$ _____. We will often write _____ to indicate that $X$ is a binomial random variable based on _____ trials with success probability _____.

Because the pmf of a binomial random variable $X$ depends on the two parameters $n$ and $p$, we denote the pmf by _____, where $n$ and $p$ are known before the experiment starts.

To derive the expression of the pmf of a binomial random variable, consider the case $n = 3$ for which each outcome, its probability, and corresponding $x$ value are displayed in the table below.

If $X \sim Bin(n, p)$, then the pmf of $X$ is given by

**NOTE:** $\binom{n}{x}$ reads "$n$ choose $x$", this tells us _____

$$\binom{n}{x} =$$

For example,

$$\binom{5}{2} =$$

**Example 41.** Each of six randomly selected cola drinkers is given a glass containing cola $S$ and one containing cola $F$. Suppose there is no tendency among cola drinkers to prefer one cola to the other. Then $p = P(\text{a selected individual prefers } S) = 0.5$, so with $X =$ the number among the six who prefer $S$,

**Using Binomial tables**

Even for a relatively small value of $n$, the computation of binomial probabilities can be tedious. Appendix Table A.1 tabulates the cdf _____ for $n = 5, 10, 15, 20, 25$ in combination with selected values of $p$ corresponding to different columns of the table.

**Notation:** For $X \sim Bin(n, p)$, the cdf will be denoted by

**Example 42.** Suppose that 20% of all copies of a textbook fail a certain binding strength test. Let $X$ denote the number among 15 randomly selected copies that fail the test. Then

$X$ has _____

**1.** The probability that at most 8 fail the test is

**2.** The probability that exactly 8 fail is

**3.** The probability that at least 8 fail is

**4.** The probability that between 4 and 7, inclusive, fail is



## Table A.1 Cumulative Binomial Probabilities

$$B(x; n, p) = \sum_{y=0}^{x} b(y; n, p)$$

**c.** $n = 15$

| | | $p$ | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.01** | **0.05** | **0.10** | **0.20** | **0.25** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.75** | **0.80** | **0.90** | **0.95** | **0.99** |
| | 0 | .860 | .463 | .206 | .035 | .013 | .005 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | 1 | .990 | .829 | .549 | .167 | .080 | .035 | .005 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | 2 | 1.000 | .964 | .816 | .398 | .236 | .127 | .027 | .004 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | 3 | 1.000 | .995 | .944 | .648 | .461 | .297 | .091 | .018 | .002 | .000 | .000 | .000 | .000 | .000 | .000 |
| | 4 | 1.000 | .999 | .987 | .836 | .686 | .515 | .217 | .059 | .009 | .001 | .000 | .000 | .000 | .000 | .000 |
| | 5 | 1.000 | 1.000 | .998 | .939 | .852 | .722 | .403 | .151 | .034 | .004 | .001 | .000 | .000 | .000 | .000 |
| | 6 | 1.000 | 1.000 | 1.000 | .982 | .943 | .869 | .610 | .304 | .095 | .015 | .004 | .001 | .000 | .000 | .000 |
| $x$ | 7 | 1.000 | 1.000 | 1.000 | .996 | .983 | .950 | .787 | .500 | .213 | .050 | .017 | .004 | .000 | .000 | .000 |
| | 8 | 1.000 | 1.000 | 1.000 | .999 | .996 | .985 | .905 | .696 | .390 | .131 | .057 | .018 | .000 | .000 | .000 |
| | 9 | 1.000 | 1.000 | 1.000 | 1.000 | .999 | .996 | .966 | .849 | .597 | .278 | .148 | .061 | .002 | .000 | .000 |
| | 10 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .999 | .991 | .941 | .783 | .485 | .314 | .164 | .013 | .001 | .000 |
| | 11 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .982 | .909 | .703 | .539 | .352 | .056 | .005 | .000 |
| | 12 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .996 | .973 | .873 | .764 | .602 | .184 | .036 | .000 |

Lec 14

**The Mean and Variance of X**

If $X \sim Bin(n, p)$, then

$$E(X) =$$

$$\text{Var}(X) =$$

$$\sigma_X =$$

where $q = 1 - p$.

**Example 43.** If 75% of all purchases at a certain store are made with a credit card and $X$ is the number among ten randomly selected purchases made with a credit card, then

**Example 44.** (Exercise 46 on Page 123) Compute the following binomial probabilities directly from the formula for $b(x; n, p)$:

**a.** $b(3; 8, 0.35)$

**b.** $b(5; 8, 0.6)$

**c.** $P(3 \leq X \leq 5)$ when $n = 7$ and $p = 0.6$

**d.** $P(1 \leq X)$ when $n = 9$ and $p = 0.1$

*Solution.*

**Example 45.** (Exercise 47 on Page 123) The article "Should You Report That Fender - Bender? " reported that 7 in 10 auto accidents involve a single vehicle. Suppose 15 accidents are randomly selected. Use Appendix Table A.1 to answer each of the following questions.

a. What is the probability that at most 4 involve a single vehicle?

b. What is the probability that exactly 4 involve a single vehicle?

c. What is the probability that exactly 6 involve multiple vehicles?

d. What is the probability that between 2 and 4, inclusive, involve a single vehicle?

e. What is the probability that at least 2 involve a single vehicle?

f. Find the mean and standard deviation of $X$.

*Solution.*

Lec 15

## 3.6 The Poisson Probability Distribution

**Definition 15.** A discrete random variable $X$ is said to have a _____

_____if the pmf of $X$ is

**NOTE:** The Poisson distribution spreads probability over all non-negative integers (in contrast to the binomial distribution), an infinite number of possibilities.

Appendix Table A.2 contains the Poisson cdf $F(x;\mu)$ for $\mu = 0.1, 0.2, \cdots, 1, 2, \cdots, 10$, 15, and 20. Alternatively, many software packages will provide $F(x;\mu)$ and $p(x;\mu)$ upon request.

**Example 46.** Let $X$ denote the number of traps in a particular type of metal transistor, and suppose it has a Poisson distribution with $\mu = 2$.

The probability that there are exactly three traps is

and the probability that there are at most three traps is

This latter cumulative probability is found at the intersection of the _____ and the _____ of Appendix Table A.2, whereas $p(3;2) = F(3;2) - F(2;2) = 0.857 - 0.677 = 0.180$, the difference between two consecutive entries in the $\mu = 2$ column of the cumulative Poisson table.

**Table A.2** Cumulative Poisson Probabilities *(cont.)*

$$F(x;\mu)=\sum_{y=0}^{x}\frac{e^{-\mu}\mu^{y}}{y!}$$

| | $\mu$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **2.0** | **3.0** | **4.0** | **5.0** | **6.0** | **7.0** | **8.0** | **9.0** | **10.0** | **15.0** | **20.0** |
| **0** | .135 | .050 | .018 | .007 | .002 | .001 | .000 | .000 | .000 | .000 | .000 |
| **1** | .406 | .199 | .092 | .040 | .017 | .007 | .003 | .001 | .000 | .000 | .000 |
| **2** | .677 | .423 | .238 | .125 | .062 | .030 | .014 | .006 | .003 | .000 | .000 |
| **3** | .857 | .647 | .433 | .265 | .151 | .082 | .042 | .021 | .010 | .000 | .000 |
| **4** | .947 | .815 | .629 | .440 | .285 | .173 | .100 | .055 | .029 | .001 | .000 |
| **5** | .983 | .916 | .785 | .616 | .446 | .301 | .191 | .116 | .067 | .003 | .000 |
| **6** | .995 | .966 | .889 | .762 | .606 | .450 | .313 | .207 | .130 | .008 | .000 |
| **7** | .999 | .988 | .949 | .867 | .744 | .599 | .453 | .324 | .220 | .018 | .001 |
| **8** | 1.000 | .996 | .979 | .932 | .847 | .729 | .593 | .456 | .333 | .037 | .002 |
| **9** | | .999 | .992 | .968 | .916 | .830 | .717 | .587 | .458 | .070 | .005 |
| **10** | | 1.000 | .997 | .986 | .957 | .901 | .816 | .706 | .583 | .118 | .011 |
| **11** | | | .999 | .995 | .980 | .947 | .888 | .803 | .697 | .185 | .021 |
| **12** | | | 1.000 | .998 | .991 | .973 | .936 | .876 | .792 | .268 | .039 |
| **13** | | | | .999 | .996 | .987 | .966 | .926 | .864 | .363 | .066 |
| **14** | | | | 1.000 | .999 | .994 | .983 | .959 | .917 | .466 | .105 |
| **15** | | | | | .999 | .998 | .992 | .978 | .951 | .568 | .157 |

If $X$ has a Poisson distribution with parameter $\mu$, then

$$E(X) =$$
$$\text{Var}(X) =$$

**Example 47.** (Example 46 continued)

**Example 48.** Suppose the number of accidents per month at an industrial plant has a Poisson distribution with mean 2.6. If we denote $Y$ = the number of accidents per month,

then _____

(a) Find the probability that there will be 4 accidents in the next month.

(b) Find the probability of two or more accidents in the

(c) Find the probability of having between 3 and 6 accidents.

**Example 49.** Let $X$ be the number of material anomalies occurring in a particular region of an aircraft disk. Some article proposes a Poisson distribution for $X$. Suppose that $\mu = 4$.

a. Compute both $P(X \leq 4)$ and $P(X < 4)$.

b. Compute $P(4 \leq X \leq 8)$.

c. Compute $P(8 \leq X)$.

d. What is the probability that the number of anomalies exceeds its mean value by no more than one standard deviation?

*Solution.*

**Example 50.** Some article proposed using the Poisson distribution to model the number of failures in pipelines of various types. Suppose that for cast-iron pipe of a particular length, the expected number of failures is 1. Then $X$, the number of failures, has a Poisson distribution with $\mu = 1$.

a. Obtain $P(X \leq 5)$.

b. Determine $P(X = 2)$.

c. Determine $P(2 \leq X \leq 4)$.

d. What is the probability that $X$ exceeds its mean value by more than one standard deviation?

*Solution.*

## Table A.2 Cumulative Poisson Probabilities

$$F(x;\mu) = \sum_{y=0}^{x} \frac{e^{-\mu}\mu^{y}}{y!}$$

| | | $\mu$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **.1** | **.2** | **.3** | **.4** | **.5** | **.6** | **.7** | **.8** | **.9** | **1.0** |
| | **0** | .905 | .819 | .741 | .670 | .607 | .549 | .497 | .449 | .407 | .368 |
| | **1** | .995 | .982 | .963 | .938 | .910 | .878 | .844 | .809 | .772 | .736 |
| | **2** | 1.000 | .999 | .996 | .992 | .986 | .977 | .966 | .953 | .937 | .920 |
| $x$ | **3** | | 1.000 | 1.000 | .999 | .998 | .997 | .994 | .991 | .987 | .981 |
| | **4** | | | | 1.000 | 1.000 | 1.000 | .999 | .999 | .998 | .996 |
| | **5** | | | | | | | 1.000 | 1.000 | 1.000 | .999 |
| | **6** | | | | | | | | | | 1.000 |

Lec 16

# 4 Continuous Random Variables and Probability Distributions

Chapter 3 concentrated on the development of probability distributions for discrete random variables. In this chapter, we consider the second general type of random variable that arises

in many applied problems: _____

## 4.1 Probability Density Functions

**Recall:** A discrete random variable is one whose possible values either constitute a finite set or can be listed in an infinite sequence.

A random variable $X$ is continuous if

(1) possible values comprise either _____ on the number line (for

some $a < b$, any number $x$ between $a$ and $b$ is a possible value) or _____

_____, and

(2) _____ for any number $c$ that is a possible value of $X$.

Examples:

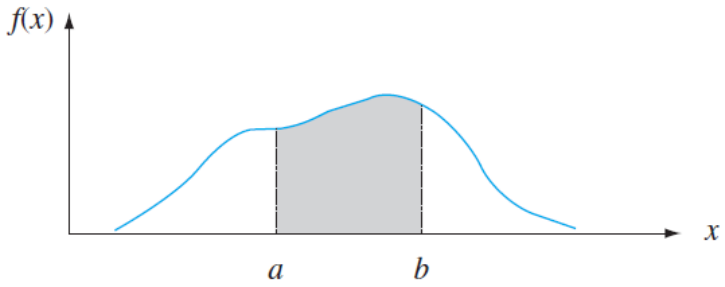
- A chemical compound is randomly selected and its pH $X$ is determined, then $X$ is a continuous random variable because any pH value between 0 and 14 is possible.
- A location in the United States is selected and the height above sea level , $Y$, is observed. Then the set of all possible values of $Y$ is the set of all numbers in the interval between 2282 and 14,494.
- The highest temperature of the day, $Z$, is observed and theoretically speaking, $Z$ could be any numbers in $\mathbb{R}$.

**Probability Distributions for Continuous Variables**

**Definition 16.** Let $X$ be a continuous random variable. Then the _____

_____of $X$ is a function $f(x)$ such that for any two numbers $a$ and $b$ with $a \leq b$,

That is, the probability that $X$ takes on a value in the interval $[a, b]$ is the _____ above this interval and under the graph of the density function.




$P(a \leq X \leq b)$ = the area under the density curve between $a$ and $b$

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

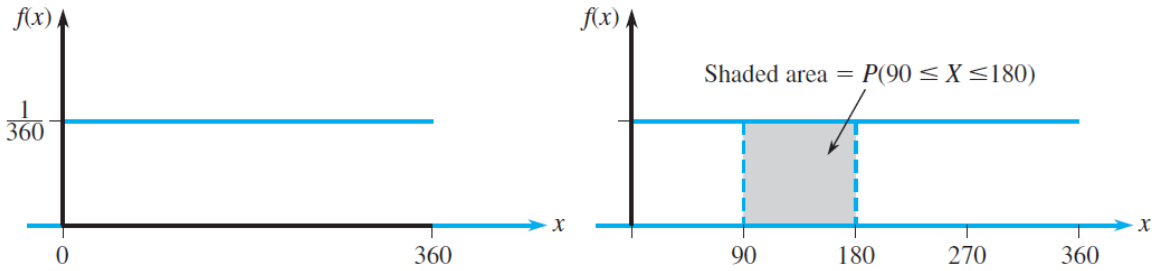
1.

2.

**Example 51.** Consider the reference line connecting the valve stem on a tire to the center point, and let $X$ be the angle measured clockwise to the location of an imperfection. One possible pdf for $X$ is

$$f(x) = \begin{cases} \frac{1}{360} & 0 \leq x \leq 360 \\ 0 & \text{otherwise} \end{cases}$$

*Solution.*




**NOTE:** Because whenever $0 \le a \le b \le 360$ in Example 51, $P(a \le X \le b)$ depends only on the width $b - a$ of the interval, $X$ is said to have a uniform distribution.

**Definition 17.** A continuous random variable $X$ is said to have a _____ on the interval $[A, B]$ if the pdf of $X$ is

The graph of any uniform pdf looks like the graph in Figure above except that the interval of positive density is $[A, B]$ rather than $[0, 360]$.

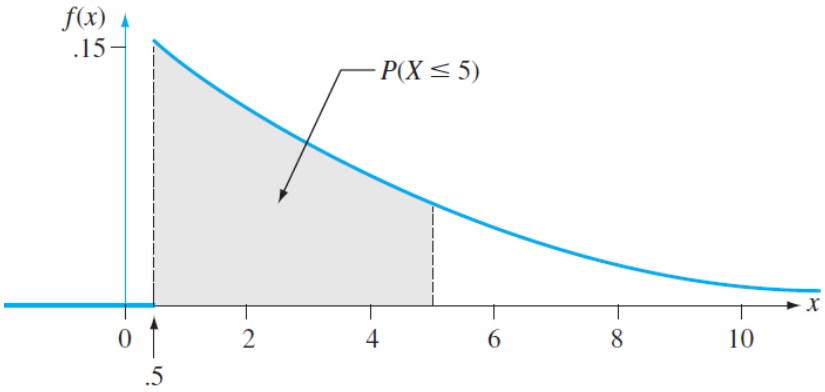
Recall: When $X$ is a discrete random variable, each possible value is assigned a _____ probability. This is not true of a continuous random variable because the area under a density curve that lies above any single value is zero:

The fact that $P(X = c) = 0$ when $X$ is continuous has an important practical consequence: The probability that $X$ lies in some interval between $a$ and $b$ does not depend on whether the lower limit $a$ or the upper limit $b$ is included in the probability calculation:

**Example 52.** "Time headway" in traffic flow is the elapsed time between the time that one car finishes passing a fixed point and the instant that the next car begins to pass that point. Let $X$ = the time headway for two randomly chosen consecutive cars on a freeway during a

period of heavy flow. The pdf of $X$ is

$$f(x) = \begin{cases} 0.15e^{-0.15(x-5)} & x \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$




The density curve for time headway

**Example 53.** (Exercise 1 on textbook page 146) The current in a certain circuit is a continuous random variable $X$ with the following density function:

$$f(x) = \begin{cases} 0.075x + 0.2 & 3 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

a. Graph the pdf and verify that the total area under the density curve is indeed 1.

b. Calculate $P(X \leq 4)$. How does this probability compare to $P(X < 4)$?

c. Calculate $P(3.5 \leq X \leq 4.5)$ and also $P(4.5 < X)$.

Lec 17

## 4.2 Cumulative Distribution Functions and Expected Values

### 4.2.1 The Cumulative Distribution Function

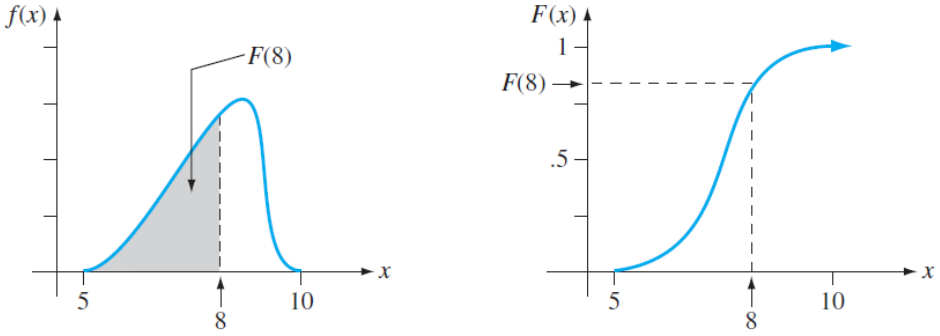
**Recall:** The cumulative distribution function (cdf) $F(x)$ of a discrete random variable $X$ with pmf $p(x)$ is defined for every number $x$ in $\mathbb{R}$ by

$$F(x) = P(X \le x) = \sum_{y:\ y \le x} p(y)$$

The cdf of a continuous random variable gives the same probabilities $P(X \le x)$ and is obtained by replacing summation by integration.

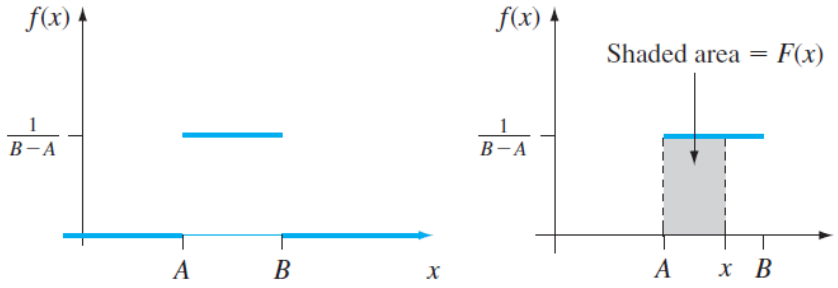
The **cumulative distribution function** $F(x)$ for a continuous random variable $X$ is defined for every number $x$ by

For each $x$, $F(x)$ is the area under the density curve to the left of $x$. This is illustrated in Figure below, where $F(x)$ increases smoothly as $x$ increases.




A pdf and associated cdf

**Example 54.** Let $X$, the thickness of a certain metal sheet, have a uniform distribution on $[A, B]$. The density function is shown in Figure below.




The pdf for a uniform distribution

For $x < A$, $F(x) = 0$, since there is no area under the graph of the density function to the left of such an $x$. For $x \geq B, F(x) = 1$, since all the area is accumulated to the left of such an $x$. Finally, for $A \leq x \leq B$,

The entire cdf is

The graph of this cdf is

**Using $F(x)$ to Compute Probabilities**

As for discrete random variables, probabilities of various intervals can be computed from a formula or table of $F(x)$.

Let $X$ be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. Then for any number $a$,

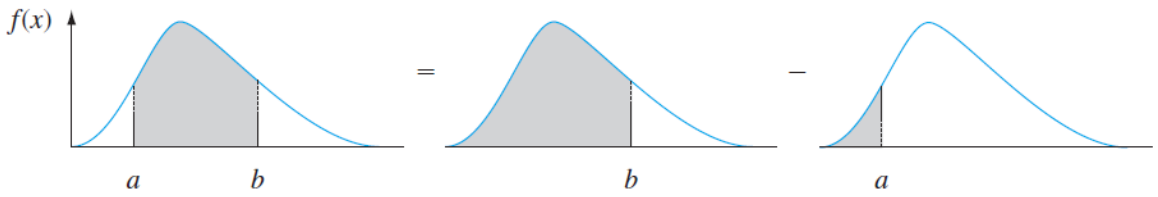
$$P(X > a) =$$

and for any two numbers $a$ and $b$ with $a < b$,

$$P(a \leq X \leq b) =$$

Figure below illustrates the second part of this proposition; the desired probability is the shaded area under the density curve between $a$ and $b$, and it equals the difference between the two shaded cumulative areas.

**NOTE:** This is different from a discrete random variable (e.g., binomial or Poisson): $P(a \leq X \leq b) = F(b) - F(a-1)$ when $a$ and $b$ are integers.




Computing $P(a \leq X \leq b)$ from cumulative probabilities

**Example 55.** Suppose the pdf of the magnitude $X$ of a dynamic load on a bridge (in newtons) is given by

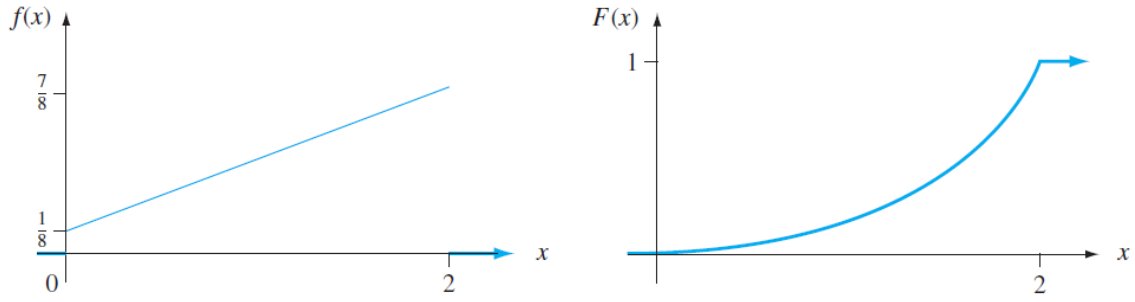
$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

For any number $x$ between 0 and 2,

$$F(x) =$$

Thus in summary

The graphs of $f(x)$ and $F(x)$ are shown in Figure below.




The pdf and cdf for Example 4.7

The probability that the load is between 1 and 1.5 is

The probability that the load exceeds 1 is

**Obtaining $f(x)$ from $F(x)$**

**Recall:** For $X$ discrete, the pmf is obtained from the cdf by taking the difference between two $F(x)$ values.

> If $X$ is a continuous random variable with pdf $f(x)$ and cdf $F(x)$, then at every $x$ at which the derivative $F'(x)$ exists,

This result is a consequence of the Fundamental Theorem of Calculus.

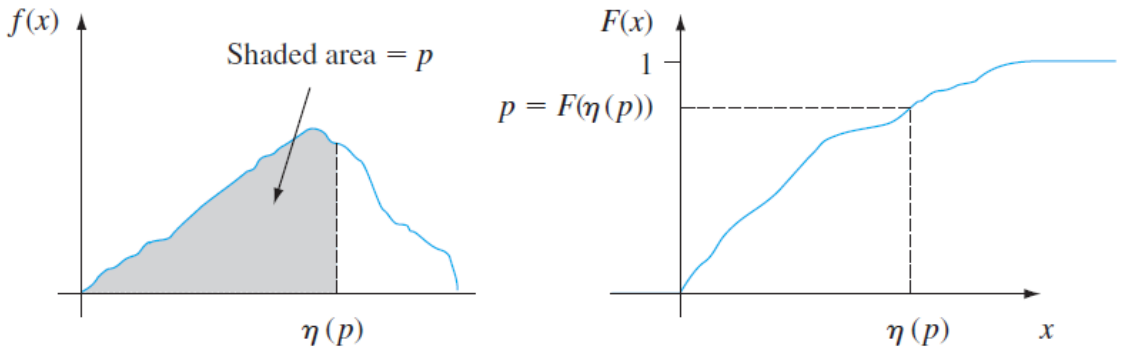
**Example 56.** (Example 54 continued) When $X$ has a uniform distribution, $F(x)$ is differentiable except at $x = A$ and $x = B$, where the graph of $F(x)$ has sharp corners. Since $F(x) = 0$ for $x < A$ and $F(x) = 1$ for $x > B$, $F'(x) = 0 = f(x)$ for such $x$. For $A < x < B$,

**Percentiles of a Continuous Distribution**

When we say that an individual's test score was at the 85th percentile of the population, we mean that 85% of all population scores were below that score and 15% were above.

**Definition 18.** Let $p$ be a number between 0 and 1. The **(100$p$)th percentile** of the distribution of a continuous random variable $X$, denoted by $\eta(p)$, is defined by

According to Definition 18, $\eta(p)$ is that value such that $100p\%$ of the area under the graph of $f(x)$ lies to the left of $\eta(p)$ and $100(1-p)\%$ lies to the right. Thus $\eta(0.75)$, the 75th percentile, is such that the area under the graph of $f(x)$ to the left of $\eta(0.75)$ is 0.75. Figure below illustrates the definition.



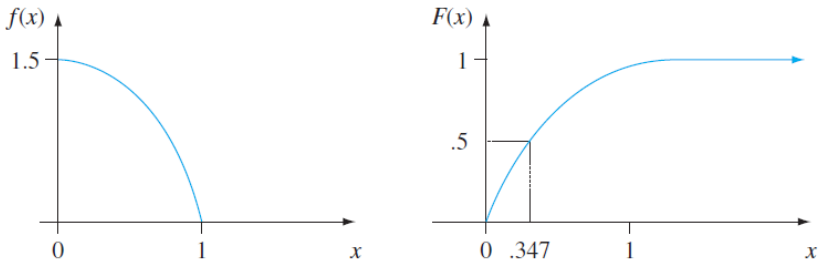

The (100$p$)th percentile of a continuous distribution

**Example 57.** The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous random variable $X$ with pdf

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The cdf of sales for any $x$ between 0 and 1 is

The graphs of both $f(x)$ and $F(x)$ appear in Figure below.




The $(100p)$th percentile of this distribution satisfies the equation

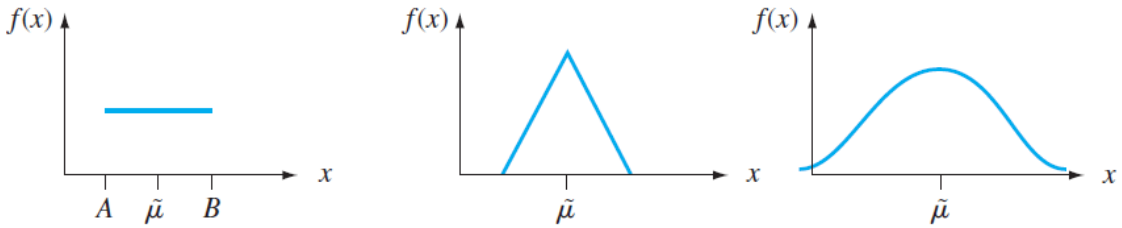
For the 50th percentile, $p = 0.5$, and the equation to be solved is

the solution is

Which means if the distribution remains the same from week to week, then in the long run 50% of all weeks will result in sales of less than 0.347 ton and 50% in more than 0.347 ton.

**Definition 19.** The median of a continuous distribution, denoted by $\widetilde{\mu}$, is the 50th percentile, so $\widetilde{\mu}$, satisfies $0.5 = F(\widetilde{\mu})$. That is, half the area under the density curve is to the left of $\widetilde{\mu}$ and half is to the right of $\widetilde{\mu}$.

A continuous distribution whose pdf is **symmetric** - the graph of the pdf to the left of some point is a mirror image of the graph to the right of that point - has median $\widetilde{\mu}$, equal to the point of symmetry, since half the area under the curve lies to either side of this point. Figure below gives several examples.




Medians of symmetric distributions

Lec 18

### 4.2.2 Expected Values

For a discrete random variable $X$, $E(X)$ was obtained by summing $x \cdot p(x)$ over possible $X$ values. Here we replace summation by integration and the pmf by the pdf to get a continuous weighted average.

**Definition 20.** The expected or mean value of a continuous random variable $X$ with pdf $f(x)$ is

**NOTE:** $E(X)$ is the most frequently used measure of population location or center.

**Example 58.** (Example 57 continued) The pdf of weekly gravel sales $X$ was

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$E(X) =$$

Often we wish to compute the expected value of some function $h(X)$ of the random variable $X$.

If $X$ is a continuous random variable with pdf $f(x)$ and $h(X)$ is any function of $X$, then

$$E[h(X)] =$$

**Definition 21.** The **variance** of a continuous random variable $X$ with pdf $f(x)$ and mean value $\mu$ is

$$\sigma_X^2 = \text{Var}(X) =$$

The **standard deviation** (SD) of $X$ is

$$\sigma_X =$$

The variance and standard deviation give quantitative measures of how much spread there is in the distribution or population of $x$ values.

Shortcut formula

**Example 59.** (Example 57 continued) For $X$ = weekly gravel sales, we computed $E(X) = \frac{3}{8}$. Since

$$E(X^2) =$$

Then

$$\text{Var}(X) =$$

and

$$\sigma_X =$$

When $h(X) = aX + b$, the expected value and variance of $h(X)$ satisfy the same properties as in the discrete case:

$$E[h(X)] =$$

and

$$\text{Var}[h(X)] =$$

## 4.3 The Normal Distribution

The normal distribution is the most important one in all of probability and statistics.

**Definition 22.** A continuous random variable $X$ is said to have a **normal distribution** with parameters $\mu$ and $\sigma^2$, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of $X$ is

$$f(x; \mu, \sigma) =$$

The statement that $X$ is normally distributed with parameters $\mu$ and $\sigma^2$ is often

abbreviated _____

It can be shown that for $X \sim N(\mu, \sigma^2)$

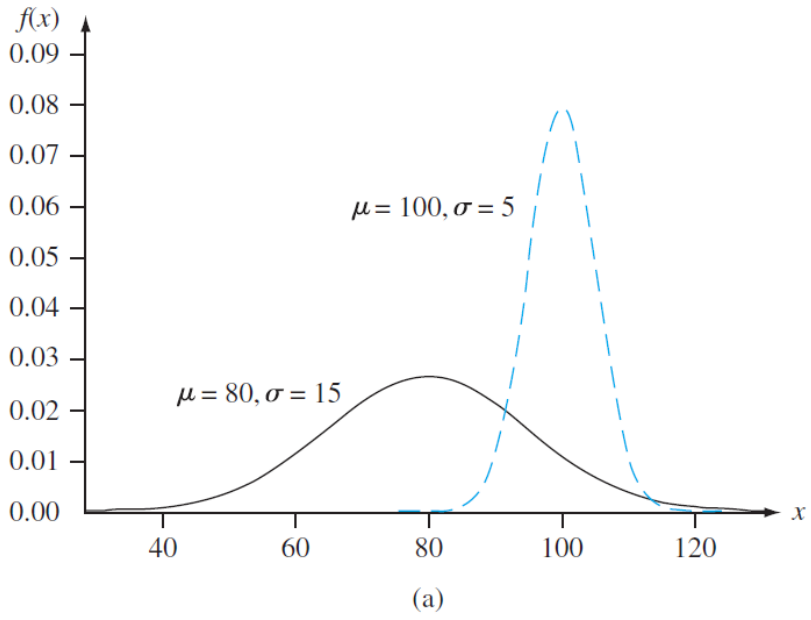
$$E(X) =$$

and

$$\text{Var}(X) =$$

Figure below presents graphs of $f(x; \mu, \sigma^2)$ for several different $(\mu, \sigma^2)$ pairs.

- Each density curve is symmetric about $\mu$ and bell-shaped, so the center of the bell (point of symmetry) is both the mean of the distribution and the median.
- The mean $\mu$ is a location parameter, since changing its value shifts the density curve.
- $\sigma^2$ is referred to as a scale parameter, because changing its value stretches or compresses the curve.
- The inflection points of a normal curve (points at which the curve changes from turning downward to turning upward) occur at $\mu - \sigma$ and $\mu + \sigma$.



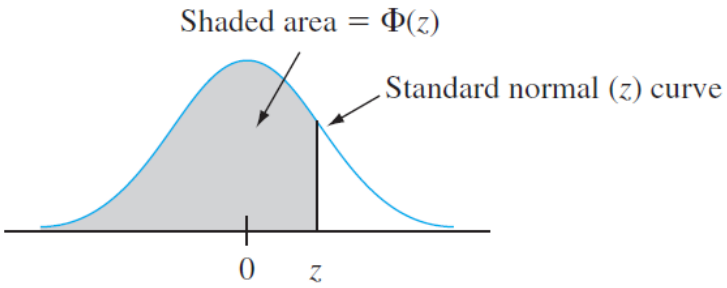

Two different normal density curves

### 4.3.1 The Standard Normal Distribution

**Definition 23.** The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**. A random variable having a standard normal distribution is called a **standard normal random variable** and will be denoted by $Z$. The pdf of $Z$ is

$$f(z; 0, 1) =$$

Appendix Table A.3 gives _____, the area under the standard normal density curve to the left of $z$ for selected $z$'s. Figure below illustrates the type of cumulative area (probability) tabulated in Table A.3.




Standard normal cumulative areas tabulated in Appendix Table A.3

**Example 60.** Let's determine the following standard normal probabilities:

(a) $P(Z \leq 1.25)$;

(b) $P(Z > 1.25)$;

(c) $P(Z \leq -1.25)$;

(d) $P(-0.38 \leq Z \leq 1.25)$;

(e) $P(Z \leq 5)$

*Solution.*

**Table A.3** Standard Normal Curve Areas *(cont.)* $\Phi(z) = P(Z \leq z)$

| $z$ | .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 |
| **0.6** | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| **0.7** | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| **0.8** | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| **0.9** | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| **1.0** | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| **1.1** | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| **1.2** | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| **1.3** | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| **1.4** | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9278 | .9292 | .9306 | .9319 |

Lec 19

### 4.3.2 Percentiles of the Standard Normal Distribution

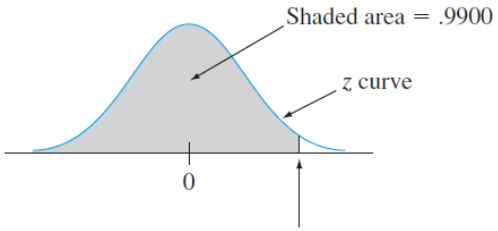
For any $p$ between 0 and 1, Appendix Table A.3 can be used to obtain the $(100p)$th percentile of the standard normal distribution.

In general, the $(100p)$th percentile is identified by the row and column of Appendix Table A.3 in which the entry $p$ is found (e.g., the 67th percentile is obtained by finding .6700 in the body of the table, which gives $z = 0.44$).

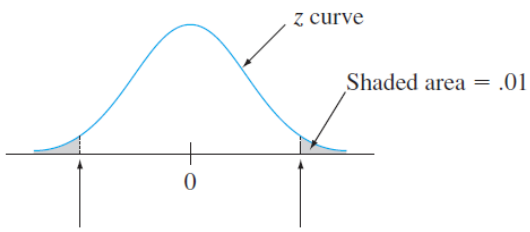
If $p$ does not appear, the number closest to it is typically used, although linear interpolation gives a more accurate answer. For example, to find the 95th percentile, look for .9500 inside the table. Although it does not appear, both .9495 and .9505 do, corresponding to $z = 1.64$ and 1.65, respectively. Since .9500 is halfway between the two probabilities that do appear, we will use 1.645 as the 95th percentile.

**Example 61.** Find the 99th percentile of the standard normal distribution.

*Solution.*



Finding the 99th percentile



The relationship between the 1st and 99th percentiles

**Table A.3** Standard Normal Curve Areas *(cont.)* $\Phi(z) = P(Z \le z)$

| $z$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1.5** | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| **1.6** | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| **1.7** | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| **1.8** | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| **1.9** | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| **2.0** | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| **2.1** | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| **2.2** | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| **2.3** | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| **2.4** | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |

**Example 62.** Determine the value of the constant $c$ that makes the probability statement correct.

(a) $\Phi(c) = 0.9838$

(b) $P(0 \leq Z \leq c) = 0.291$

(c) $P(c \leq Z) = 0.121$

(d) $P(-c \leq Z \leq c) = 0.668$

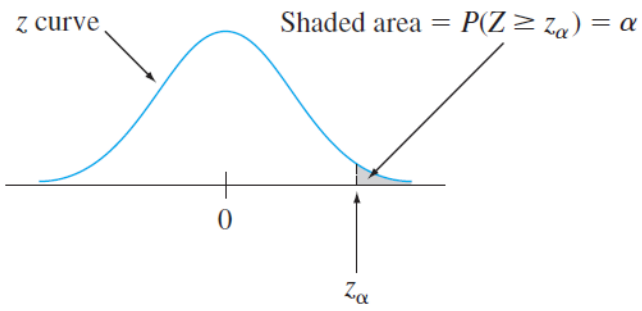
(e) $P(c \leq |Z|) = 0.016$

(f) Find the 6th percentile for the standard normal curve

### 4.3.3 $z_\alpha$ Notation for $z$ Critical Values

**Notation:** $z_\alpha$ will denote the value on the $z$ axis for which $\alpha$ of the area under the $z$ curve lies to the right of $z_\alpha$. (See Figure below.)

For example, $z_{.10}$ captures upper-tail area .10, and $z_{.01}$ captures upper-tail area .01.




$z_\alpha$ notation Illustrated

Since $\alpha$ of the area under the $z$ curve lies to the right of $z_\alpha$, $1 - \alpha$ of the area lies to its left. Thus $z_\alpha$ is the $100(1 - \alpha)$th percentile of the standard normal distribution. By symmetry the area under the standard normal curve to the left of $-z_\alpha$ is also $\alpha$. The $z_\alpha$'s are usually referred to as $z$ critical values.

**Example 63.**

(a) Find $z_{0.0055}$

(b) Find $z_{0.6630}$

### 4.3.4 Nonstandard Normal Distributions

When $X \sim N(\mu, \sigma^2)$, probabilities involving $X$ are computed by "standardizing."

If $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, then

has a standard normal distribution. Thus

By standardizing, any probability involving $X$ can be expressed as a probability involving a standard normal random variable $Z$, so that Appendix Table A.3 can be used.

**Example 64.** (Exercise 33 on textbook page 167) $X$ = maximum speed of a car. A normal distribution with mean value 46.8 km/h and standard deviation 1.75 km/h is postulated. Consider randomly selecting a single such car.

a. What is the probability that maximum speed is at most 50 km/h?

b. What is the probability that maximum speed is at least 48 km/h?

c. What is the probability that maximum speed differs from the mean value by at most 1.5 standard deviations?

d. Find the 75th percentile for the max speed.

*Solution.*

Part (c) in Example 64 can be answered without knowing either $\mu$ or $\sigma$, as long as the distribution is known to be normal, the answer is the same for any normal distribution:

**Example 65.** The breakdown voltage of a randomly chosen diode of a particular type is known to be normally distributed. What is the probability that a diode's breakdown voltage is within 1 standard deviation of its mean value?

*Solution.*

If the population distribution of a variable is (approximately) normal, then

1. Roughly _____ of the values are within 1 SD of the mean.
2. Roughly _____ of the values are within 2 SDs of the mean.
3. Roughly _____ of the values are within 3 SDs of the mean.

### 4.3.5 Normal Approximation to the Binomial Distribution

- Let $X$ be a binomial random variable based on $n$ trials with success probability $p$. So $X \sim b(n, p)$.
- If the binomial probability histogram is not too skewed, and both $np$ and $n(1-p)$ are $\geq 10$.

Then $X$ has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. Then

**Note:** 0.5 is the correction for continuity. Let's see why we want this correction in the next example.

**Example 66.** If $X \sim b(25, 0.6)$, We can approximate $X$ with

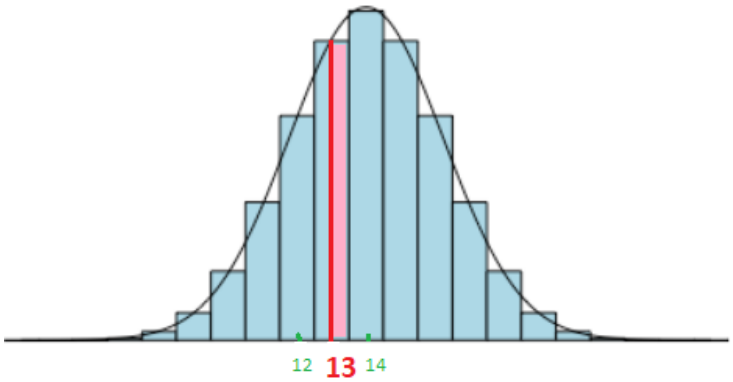
Therefore,

$$P(X \leq 13)$$

while the exact binomial calculation gives:

$$P(X \leq 13)$$

The approximation is good! But still can be improved.




**Normal approximation with the continuity correction.**

Figure above shows that when we use $P(Y \leq 13)$ to approximate $P(X \leq 13)$, the normal approximation is _____ than the exact binomial value. The area of the bars to the left of 13.5 gives $P(X \leq 13)$; the area under the curve to the left of 13 gives $P(Y \leq 13)$.

**Correction:**

$$P(X \leq 13)$$

The result is improved greatly!

**Summary:**

$$P(X \leq x) \approx P(Y \leq x + 0.5)$$
$$P(X \geq x) \approx P(Y \geq x - 0.5)$$

**Example 67.** (Exercise 55 on textbook page 169) Suppose only 75% of all drivers in a certain state regularly wear a seat belt. A random sample of 500 drivers is selected. What is the probability that

a. Between 360 and 400 (inclusive) of the drivers in the sample regularly wear a seat belt?

b. Fewer than 400 of those in the sample regularly wear a seat belt?

*Solution.* Let $X$ = the number of drivers regularly wear a seat belt.

Lec 20

**Example 68.** (Exercise 54 on textbook page 169) Suppose that 10% of all steel shafts produced by a certain process are nonconforming but can be reworked. Consider a random sample of 200 shafts, and let $X$ denote the number among these that are nonconforming and can be reworked. What is the (approximate) probability that $X$ is

a. At most 30?

b. Less than 30?

c. Between 15 and 25 (inclusive)?

*Solution.* $X=$ the number of nonconforming and can be reworked shafts,

# 5 Joint Probability Distributions and Random Samples

## 5.1 Jointly Distributed Random Variables

There are many experimental situations in which more than one random variables will be of interest to an investigator. For example, $X$ and $Y$ might be the height and weight, respectively, of a randomly selected individual.

### 5.1.1 Two Discrete Random Variables

**Recall:** The probability mass function (pmf) of a single discrete random variable $X$ specifies how much probability mass is placed on each possible $X$ value.

The joint pmf of two discrete random variables $X$ and $Y$ describes how much probability

mass is placed _____.

**Definition 24.** Let $X$ and $Y$ be two discrete random variables defined on the sample space $S$. The joint probability mass function $p(x, y)$ is defined for each pair of numbers $(x, y)$ by

To make it a valid pmf:

1.

2.

Let $A$ be any particular set consisting of pairs of $(x, y)$ values. Then the probability $P[(X, Y) \in A]$ is

**Example 69.** The joint pmf is given as

| $p(x, y)$ | $y = 0$ | $y = 1$ | $y = 2$ |
|---|---|---|---|
| $x = 0$ | 0.1 | 0.04 | 0.02 |
| $x = 1$ | 0.08 | 0.2 | 0.06 |
| $x = 2$ | 0.06 | 0.14 | 0.3 |

**Question:**

a. Is it a valid pmf?

b. Find $P(X = 1 \text{ and } Y = 1)$.

c. Find $P(X \leq 1 \text{ and } Y \leq 1)$.

d. Find $P(X \neq 0 \text{ and } Y \neq 0)$.

**Definition 25.** The **marginal probability mass function** of $X$, denoted by $p_X(x)$, is given by

for each possible value $x$. Similarly, the marginal probability mass function of $Y$ is

for each possible value $y$.

The marginal pmf of $X$ gives the distribution of $X$ if it is observed without $Y$, and the marginal pmf of $Y$ gives the distribution of $Y$ if it is observed without $X$,

**Example 70.** (Example 69 continued)

e. Find the marginal pmf of $X$.

f. Find $P(X \leq 1)$.

g. Find the marginal pmf of $Y$.

**Definition 26.** Two discrete random variables are independent if **for every pair** of $(x, y)$, we have

If $p(x, y) \neq p_X(x) \cdot p_Y(y)$ for **at least one pair** of $(x, y)$, then $X$ and $Y$ are dependent.

**Example 71.** (Example 69 continued)

h. Are $X$ and $Y$ independent?

i. Find $P(X < Y)$.

j. Find $P(5X + Y \leq 7)$.

k. Find $E(X)$.

l. Find $E(Y)$.

## 5.2 Expected Values, Covariance, and Correlation

Let $X$ and $Y$ be jointly distributed random variables with pmf $p(x, y)$. Let $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be any function. Then the expected value of $h(X, Y)$, denoted by $E[h(X, Y)]$ is given by

For example, $E(XY) =$

### 5.2.1 Covariance

When two random variables $X$ and $Y$ are not independent, it is frequently of interest to assess if the two random variables are linearly dependent and if so, how strongly they are related to one another.

**Definition 27.** The **covariance** is a measure of the strength between two random variables $X$ and $Y$ of _____, which is defined as

**Example 72.** The joint and marginal pmf's for $X$ = automobile policy deductible amount and $Y$ = homeowner policy deductible amount are

| | | $y$ | | |
|---|---|---|---|---|
| | $p(x, y)$ | 500 | 1000 | 5000 |
| | 100 | 0.3 | 0.05 | 0 |
| $x$ | 500 | 0.15 | 0.20 | 0.05 |
| | 1000 | 0.10 | 0.10 | 0.05 |

| $x$ | 100 | 500 | 1000 |
|---|---|---|---|
| $p_X(x)$ | 0.35 | 0.4 | 0.25 |

| $y$ | 500 | 1000 | 5000 |
|---|---|---|---|
| $p_Y(y)$ | 0.55 | 0.35 | 0.10 |

Find the $\text{Cov}(X, Y)$.

*Solution.*

Lec 21

### 5.2.2 Correlation

Note: The unit of the covariance are squared and thus it is difficult to interpret.

The correlation coefficient, $\rho$, divides the covariance by its maximum value to give a measure of linear strength between the values of $-1$ and 1.

> The correlation coefficient of $X$ and $Y$, denoted by $\text{corr}(X, Y)$, $\rho_{X,Y}$, or just $\rho$, is defined by

**Example 73.** (Example 72 continued) It is easily verified that

**Properties of covariance and correlation**

1. Covariance and correlation both measure the strength of the _____ between $X$ and $Y$.

2. If the covariance and correlation are both $> 0$, then $X$ and $Y$ have a _____, i.e., as $X$ increases, $Y$ increases as well.

3. If the covariance and correlation are both $< 0$, then $X$ and $Y$ have a _____, i.e., as $X$ increases, $Y$ decreases.

Figure:

4. If the covariance and correlation are $\approx 0$, then there is _____ between $X$ and $Y$.

5. For any two random variables $X$ and $Y$, $-1 \leq \rho \leq 1$.

6. If $a$ and $c$ are either both positive or both negative,

7. If $X$ and $Y$ are independent or uncorrelated, then _____, but $\rho = 0$ does not imply independence. It just means that there is no **linear** association between $X$ and $Y$. But it can also mean that $X$ and $Y$ may have a non-linear association.

8. $\rho = 1$ or $-1$ if and only if _____ for some numbers $a$ and $b$ with $a \neq 0$.

## 5.3 Statistics and Their Distributions

**Definition 28.** The random variables $X_1, X_2, \cdots, X_n$ are said to form a (simple) **random sample** of size $n$ if

1.

2.

Consider taking a random sample from a population, and compute the sample mean, $\overline{y}$, for the observations.

- Because the sample is _____, the observations will also be _____.

  Hence, $\overline{y}$ will also be _____.
- Because $\overline{y}$ is random, it has a _____ associated with it. This

  distribution plays an important role in drawing conclusion about the population, This

  is what we called _____

**Definition 29.** A **statistic** is any quantity whose value can be calculated from

_____. Prior to obtaining data, there is uncertainty as to

what value of any particular statistic will result. Therefore, a statistic is a

_____.

Examples: the sample mean $\overline{y}$, the sample median $y$, the sample standard deviation $s$, etc. are all statistics.

**Definition 30.** The distribution of the statistics is called the _____ of the statistics.

## 5.4 The Distribution of the Sample Mean

The importance of the sample mean $\overline{X}$ arises from its use in drawing conclusions about

_____.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean $\mu$ and standard deviation $\sigma$. Then

- $E(\overline{X})$
- $\text{Var}(\overline{X})$

- The distribution of $\overline{X}$ becomes more concentrated about $\mu$ as _____, i.e. averaging moves probability in toward the middle.

**NOTE:** The standard deviation $\sigma_{\overline{X}}$ is often called the **standard error of the mean**; it describes the magnitude of a typical or representative deviation of the sample mean from the population mean.

**NOTE:** These formulas are true for any distribution.

**Example 74.** The inside diameter of a randomly selected piston ring is a random variable with mean value 12cm and standard deviation 0.04cm.

a. If $\overline{X}$ is the sample mean diameter for a random sample of $n = 16$ rings, where is the sampling distribution of $\overline{X}$ centered, and what is the standard deviation of the $\overline{X}$ distribution?

b. Answer the questions posted in part (a) for a sample size of $n = 64$ rings.

c. For which of the two random samples, the one of part (a) or the one of part (b), is $\overline{X}$ more likely to be within 0.01cm of 12cm? Explain your reasoning.

*Solution.*

### 5.4.1 Samples from Normal Distribution

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with mean $\mu$ and standard deviation $\sigma$. Then for any $n$, $\overline{X}$ is

**Example 75.** (Example 74 continued) For the random sample in part (a), suppose $X_i$'s are normally distributed, what is the probability that $\overline{X}$ is within one standard error of the mean?

*Solution.*

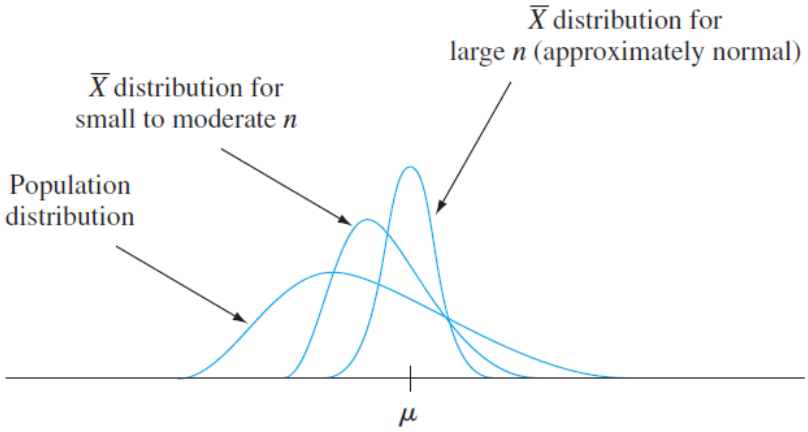
### 5.4.2 The Central Limit Theorem

When the $X_i$'s are normally distributed, so is $\overline{X}$ for any sample size $n$. Even when the population distribution is highly non-normal, if $n$ is large, a normal curve will approximate the actual distribution of $\overline{X}$.

**The Central Limit Theorem (CLT)**
Let $X_1, X_2, \cdots, X_n$ be a random sample from **any** distribution with mean $\mu$ and variance $\sigma^2$. Then if $n$ is sufficiently large,

- 
- 
- 

In summary,




The Central Limit Theorem illustrated

Figure above illustrates the Central Limit Theorem. According to the CLT, when $n$ is large and we wish to calculate a probability such as $P(a \leq \overline{X} \leq b)$, we need only "pretend" that $\overline{X}$ is normal, standardize it, and use the normal table.

**Example 76.** Suppose that a certain type of cable strength is normally distributed with mean $\mu = 450$lb and sd=$\sigma = 50$lb.

(a) Find the probability that the strength of the cable is greater than 536lbs.

(b) Let $\overline{X}$ = the mean strength for a sample of 9 cables. Find the shape, mean and sd of $\overline{X}$.

(c) Find the probability that the sample mean of 9 cables will be between 423lbs and 480 lbs.

(d) Suppose we have 35 cables, find $P(\overline{X} < 428)$.

(e) Suppose that the population is not normal (or unknown)

- Can we still solve part(a)?
- Can we still solve part (c)?
- Can we still solve part (d)?

## 5.5 The Distribution of a Linear Combination

**Definition 31.** Given a collection of $n$ random variables $X_1, \cdots, X_n$ and $n$ numerical constants $a_1, \cdots, a_n$, the random variable

is called a **linear combination** of the $X_i$'s.

Taking $a_1 = a_2 = \cdots = a_n = 1$ gives

Taking $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$ gives

Let $X_1, X_2, \cdots, X_n$ have mean values $\mu_1$, $\mu_2$, $\cdots$, $\mu_n$, respectively and variances $\sigma_1^2$, $\sigma_2^2$, $\cdots$, $\sigma_n^2$, respectively.

1. Whether or not the $X_i$'s are independent,

2. If $X_1, \cdots, X_n$ are independent,

**The Case of Normal Random Variables**

If $X_1, X_2, \cdots, X_n$ are independent normally distributed random variables (with possibly different means and variances), then any linear combination of $X_i$'s is _____ with mean and variance as given earlier.

**Example 77.** (From homework)

I have two errands to take care of on campus. Let $X_1$ and $X_2$ represent the times that it takes for the first and second errands, respectively. Let $X_3$ = the total time in minutes that I spend walking to and from my office and between the errands. Suppose that $X_1$, $X_2$, $X_3$ are independent and normally distributed with $\mu_1$= 15, $\sigma_1 = 4$, $\mu_2 = 5$, $\sigma_2 = 1$, $\mu_3$= 12, and $\sigma_3$=3.

(a) Find the chance that the total time I am away from my office is less than 45 minutes, i.e. find $P(X_1 + X_2 + X_3 < 45)$.

(b) Find the probability of the average amount of time it takes less than 12 minutes.

(c) Find the probability $P(X_1 - X_3 > 0)$.

Lec 22

# 7 Statistical Intervals Based on a Single Sample

Consider a population with an unknown parameter $\mu$, the mean.

- _____ for $\mu$ is a single value that can be considered as a sensible estimate for $\mu$.
- The point estimate is obtained by taking a representative sample and use the corresponding statistics, $\overline{x}$.
- Because of sampling variability, it is virtually never the case where $\overline{x} = \mu$.
- The statistics $\overline{x}$ does not give any information about how close $\overline{x}$ is to $\mu$. Thus, we need to consider the _____ around $\mu$.
- An alternative way to a point estimate for $\mu$ is to report _____ of plausible values, called _____.
- A confidence interval reports a range of values where $\mu$ is likely to fall.
- A confidence interval depends on $\alpha$, where $(1-\alpha)100\%$ is_____, which is a measure of the degree of reliability of the interval. A confidence level of 95% implies that 95% of all samples would give an interval that includes $\mu$.

## 7.1 Basic Properties of Confidence Intervals

Suppose $X_1, \cdots, X_n$ is a random sample of size $n$. Suppose that the unknown parameter of interest is the population mean $\mu$.

1.

2.

Then,

Note: $\sigma$ and $n$ are known, but $\overline{X}$ is unknown since we did not take the sample yet.

This interval is random because the two endpoints of the interval involve a random variable. The interval's width is $2 \cdot (1.96) \cdot \sigma/\sqrt{n}$, a fixed number; only the location of the interval (its midpoint $\overline{X}$) is random.

In general, for all CI:

$$\text{estimate } \pm \text{critical value } \cdot \sigma \text{ estimate}$$

**Example 78.** (Exercises 1 on textbook page 284) Consider a normal population distribution with the value of $\sigma$ known.

**Question:** What is the confidence level for the interval $\overline{x} \pm 2.81\sigma/\sqrt{n}$?

**Example 79.** Consider a normal population distribution with the value of $\sigma = 3$.

**Question:** What is the 95% for the population mean, $\mu$, when $n = 25$ and $\overline{x} = 58.3$?

**Example 80.** A sample of $n = 31$ trained typists was selected, and the preferred keyboard height was determined for each typist. The resulting sample average preferred height was $\overline{x} = 80.0$ cm. Assuming that the preferred height is normally distributed with $\sigma = 2.0$ cm, obtain the 95% confidence interval for $\mu$, the true average preferred height for the population of all experienced typists.

*Solution.*

**NOTE:** How to interpret a CI?

**Choosing a level of confidence:**

- The precision of the CI refers to the width, $w$, of the interval. The more _____ a CI is, the _____ its width. Because the smaller width implies the interval identifies fewer value $\mu$.
- The reliability of a CI refers to its CL. The more _____ the CI is, the _____ you are to the population mean.
- As the CL _____, the width of the interval also _____. So less precision implies more reliable.
- As the CL _____, the width of the interval also _____. So more precision implies less reliable.

**Question:** How do we balance precision and reliability?

Answer: Specify both CL and interval width, then determine the _____ as needed. Suppose we have a normal population where $\sigma$ is known. Then

**Example 81.** A new operating system has been installed, and we wish to estimate the true average response time $\mu$ for the new environment. Assuming that response time is normally distributed with $\sigma = 25$ millisec, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10?

*Solution.*

## 7.2 Large-Sample Confidence Intervals for a Population Mean and Proportion

The CI for $\mu$ given in the previous section assumed that the population distribution is _____ with the value of _____. We now present a _____ whose validity does not require these assumptions.

### 7.2.1 A Large-Sample Interval for $\mu$

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population having a mean $\mu$ and standard deviation $\sigma$. Provided that $n$ is sufficiently large, the _____ implies that $\overline{X}$ has approximately a _____ distribution whatever the nature of the population distribution. It then follows that _____ has approximately a standard normal distribution, so that

A practical difficulty with this development is that computation of the CI requires the value of _____, which will rarely be known. Consider replacing the population standard deviation $\sigma$ by the _____, which gives

If $n$ is sufficiently large, the standardized variable

has approximately a standard normal distribution. This implies that

is a large-sample confidence interval for $\mu$ with confidence level approximately $100(1-\alpha)\%$. This formula is valid regardless of the shape of the population distribution.
**NOTE:** _____ will be sufficient to justify the use of this interval.

**Example 82.** A random sample of 110 lightning flashes in a certain region resulted in a sample average radar echo duration of 0.81 sec and a sample standard deviation 0.34 sec. Calculate a 99% (two-sided) confidence interval for the true average echo duration $\mu$, and interpret the resulting interval.

*Solution.*

Lec 23

### 7.2.2 A confidence Interval for a Population Proportion

Consider a population whose members can be divided into 2 separate groups. Let $p$ be the population proportion or the true proportion, which is unknown. To estimate $p$, we use

_____ where
$X$ = the number of people in the sample who have a given characteristic, $n$ = sample size.

- $X$ follows a _____.
- Furthermore, if both $np \geq 10$ and $n(1-p) \geq 10$, $X$ has approximately a _____ distribution.
- Since $\hat{p}$ is just $X$ multiplied by the constant $\frac{1}{n}$, $\hat{p}$ also has approximately a _____ distribution with mean

  $$E(\hat{p})$$

  and variance

  $$\text{Var}(\hat{p})$$

- If $n > 40$, then a CI for the proportion $p$ is

  Since in general, for all CI:

  $$\text{estimate} \pm \text{critical value} \cdot \sigma \text{ estimate}$$

**Example 83.** (Exercise 21 on textbook page 294) In a sample of 1000 randomly selected consumers who had opportunities to send in a rebate claim form after purchasing a product, 250 of these people said they never did so. Calculate an upper confidence bound at the 95% confidence level for the true proportion of such consumers who never apply for a rebate. Based on this bound, is there compelling evidence that the true proportion of such consumers is smaller than 1/3? Explain your reasoning.

*Solution.*

### 7.2.3 Sample Size Consideration

Again, if we specify our confidence level (i.e. the reliability) and the width $w$ of the CI (i.e. the precision), then the "smallest" sample size can be found as

**Example 84.** (Exercise 25 on textbook page 294) A state legislator wishes to survey residents of her district to see what proportion of the electorate is aware of her position on using state funds to pay for abortions.

a. What sample size is necessary if the 95% CI for $p$ is to have a width of at most 0.1 irrespective of $p$?

b. If the legislator has strong reason to believe that at least 2/3 of the electorate know of her position, how large a sample size would you recommend?

*Solution.*

## 7.3 Intervals Based on a Normal Population Distribution

We know how to estimate $\mu$ when $n$ is large. But what should we do when $n$ is small?

**Assumptions:**

(1) The population of interest has to be normal.

(2) Population mean, $\mu$, is unknown.

(3) Population SD, $\sigma$, is unknown.

**Theorem 7.1.** When $\overline{X}$ is the mean of a random sample of size $n$ from a _____ distribution with mean $\mu$, then the random variable

has a probability distribution called _____

_____

**Properties of $t$ distribution:**

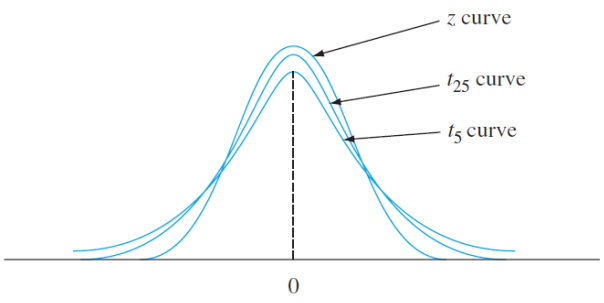
Let $t_\nu$ denote the $t$ distribution with $\nu$ df.

1. Each $t_\nu$ curve is _____.

2. Each $t_\nu$ curve is more spread out than the _____.

3. As $\nu$ _____, the spread of the corresponding $t_\nu$ curve _____.

4. As $\nu \to \infty$, the sequence of $t_\nu$ curves approaches _____.
Usually when $n > 30$, then we can use the $z$ curve.




$t_\nu$ and $z$ curves

**Notation:**
Let $t_{\alpha,\nu}$ = the number on the $x-$axix for which the area under the $t_\nu$ curve to the _____ of $t_{\alpha,\nu}$ is $\alpha$. $t_{\alpha,\nu}$ is called a _____.

**Example 85.** (HW question 8) Given the sample size $n = 15$ and the confidence level is 90%. Assume $\sigma$ is unknown. Find the $t_{\alpha/2}$ critical value for the confidence interval $\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$.

*Solution.*

Let $\overline{x}$ and $s$ be the sample mean and sample standard deviation computed from the results of a random sample from a _____ population with mean $\mu$. Then a $100(1-\alpha)\%$ confidence interval for $\mu$ is

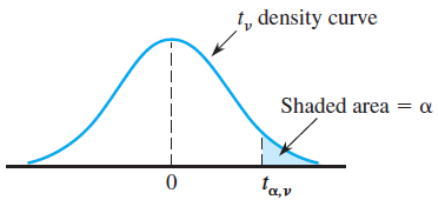
**Example 86.** (HW question 9) A random sample of 10 brands of vanilla yogurt was selected and the calorie count per serving was recorded, resulting in the following data:

$$130, 160, 150, 120, 120, 110, 170, 160, 110, 90$$

Calculate a 90% confidence interval to estimate the true mean calorie count.

*Solution.*

**Table A.5** Critical Values for $t$ Distributions




| | $\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $\nu$ | .10 | .05 | .025 | .01 | .005 | .001 | .0005 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.31 | 636.62 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.326 | 31.598 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.213 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |

**Summary:**

Lec 24

# 8 Tests of Hypotheses Based on a Single Sample

A parameter can be estimated from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval). Frequently, however, the objective of an investigation is not to estimate a parameter but to decide which of two contradictory claims about the parameter is correct. Methods for accomplishing this is called _____.

## 8.1 Hypotheses and Test Procedures

A hypothesis is _____

Examples:

- The claim $\mu = 0.75$, where $\mu$ is the true average inside diameter of a certain type of PVC pipe.
- The statement $p = 0.1$, where $p$ is the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer.

In any hypothesis-testing problem, there are _____ under consideration.

- One hypothesis might be the claim $\mu = 0.75$ and the other $\mu \neq 0.75$.
- One hypothesis might be the claim $p = 0.1$ and the other $p < 0.1$.

### 8.1.1 Test Procedures

The objective is to decide, based on sample information, which of the two hypotheses is correct.

_____ is the claim that is initially assumed to be true.

_____ is the assertion that is contradictory to $H_0$.
The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that $H_0$ is false. If the sample does not strongly contradict $H_0$, we will continue to believe in the plausibility of the null hypothesis.
The two possible conclusions from a hypothesis-testing analysis are_____

_____.

_____ is a method for using sample data to decide whether the null hypothesis should be rejected.

**Step 1:** _____

The null hypothesis, $H_0$, (in STAT 300)

-

-

The alternative hypothesis, $H_A$,

-

-

For example, suppose the true average time to pain relief for the current best-selling pain reliever is known to be 15 minutes. A new formulation has been developed that it is hoped will reduce this time. The relevant hypotheses are $H_0 : \mu = 15$ versus $H_A : \mu < 15$, where $\mu$ is the true average time to relief using the new formulation.

-

- 
- 
- 

**NOTE:**

1. The only difference between the null hypothesis $H_0$ and $H_A$ is the sign in the middle.

2. Only one hypothesis can be true in a given situation.

**Example 87.** (From HW) Determine whether or not each of the following is a valid pair of hypotheses.

(a) $H_0 : \overline{x} = 5$ vs $H_A : \overline{x} < 5$

(b) $H_0 : p = 0.7$ vs $H_A : p \neq 0.7$

(c) $H_0 : \mu = 5$ vs $H_A : \mu \geq 5$

(d) $H_0 : p = 0.3$ vs $H_A : p = 0.5$

(e) $H_0 : \mu = 5$ vs $H_A : \mu < 5$

**Step 2:** _____

Suppose we want to test a population mean $\mu$.

**Example 88.** We know that the national average on a standardized math exam is 50 points. The test average for a random sample of 100 second-graders is 54 points with a sd of 10 points.

**Question:** Are the second-graders smarter than the national average?

Step 1:

Step 2:

**Step 3:** _____

In order to find the RR, we need:

(1) _____, (for our example, we need the distribution of $\overline{X}$).

(2) _____, usually given to you. If $\alpha$ is not given, always assume _____.

**Example 88 continued.** Say 1% significant level,

**Step 4:** _____

**Example 88 continued.**

**Step 5:** Decision.

The decision always has to be a statement about $H_0$. We reject $H_0$ or we fail to reject $H_0$ based on where our test statistic falls in relationship to our RR.

**Example 88 continued.**

**Example 89.** (From HW) Determine whether or not each of the following statements is correct.

(a) The value of the test statistic does not lie in the rejection region. Therefore, we accept the null hypothesis.

(b) The value of the test statistic lies in the rejection region. Therefore, there is sufficient evidence to suggest the alternative hypothesis is true.

(c) The value of the test statistic does not lie in the rejection region. Therefore, there is evidence to suggest the null hypothesis is true.

(d) The value of the test statistic does not lie in the rejection region. Therefore, there is insufficient evidence to suggest the alternative hypothesis is true.

### 8.1.2 Errors in Hypothesis Testing

A **type I error**

A **type II error**

**Example 90.**

1) $H_0 : \mu = 100$ (This is true) vs $H_A : \mu > 100$. If our test fails to reject $H_0$,

2) $H_0 : \mu = 100$ (This is true) vs $H_A : \mu > 100$. If our test reject $H_0$,

Usually $\alpha$ is given in our test. If $\alpha$ is not given, we use $\alpha$=0.05.

3) $H_0 : \mu = 100$ vs $H_A : \mu > 100$ (This is true). If our test fails to reject $H_0$,

4) $H_0 : \mu = 100$ vs $H_A : \mu > 100$ (This is true). If our test reject $H_0$,

**Example 91.** Water samples are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most $150°F$, there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above $150°F$, 50 water samples will be taken at randomly selected times and the temperature of each sample recorded. The resulting data will be used to test the hypotheses $H_0 : \mu = 150$ versus $H_A : \mu > 150$. In the context of this situation, describe type I and type II errors.

*Solution.*

**Step 6:** _____ "Based on our evidence, at $\alpha$ significant level, we conclude that ... "

**Example 92.** (Exercise 19 on textbook page 333) The melting point of each of 16 samples of a certain brand of hydrogenated vegetable oil was determined, resulting in $\overline{x} = 94.32$. Assume that the distribution of the melting point is normal with $\sigma = 1.2$.

**Question:** Test $H_0 : \mu = 95$ versus $H_A : \mu \neq 95$ using a two-tailed level 0.01 test.

*Solution.*

Lec 25

## 8.2 $z$ Tests for Hypotheses about a Population Mean

### 8.2.1 A normal Population distribution with Known $\sigma$

Null Hypothesis:
Test statistic:

Significant level: $\alpha$
Alternative Hypothesis:

Assumptions:

**Example 93.** A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F. A sample of $n = 9$ systems yields a sample average activation temperature of 131.08°F. If the distribution of activation times is **normal with standard deviation** 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = 0.01$?

*Solution.*

**Example 94.** The desired percentage of $SiO_2$ in a certain type of aluminous cement is 5.5. To test whether the true average percentage is 5.5 for a particular production facility, 16 independently obtained samples are analyzed. Suppose that the percentage of $SiO_2$ in a sample is **normally distributed** with $\sigma = 0.3$ and that $\overline{x} = 5.25$.

Does this indicate conclusively that the true average percentage differs from 5.5?

*Solution.*

### 8.2.2 Large-Sample Tests

When the sample size is large, the foregoing $z$ tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known $\sigma$.

When we have a large sample $n > 30$, $\overline{X}$ is approximately normal, $\sigma$ is unknown, then we use the test statistic

**Example 95.** (HW #6) The biological dessert in the Gulf of Mexico called the Dead Zone is a region in which there is very little or no oxygen. Most marine life in the Dead Zone dies or leaves the region. The area of this region varies and is affected by agriculture, fertilizer runoff, and weather. The long-term mean area of the Dead Zone is 5960 square miles. As a result of recent flooding in the Midwest and subsequent runoff from the Mississippi River, researchers believe that the Dead Zone area will increase. A random sample of 50 days was obtained and the sample mean area of the Dead Zone was 6759 $mi^2$ with a sample standard deviation of 1850 $mi^2$. Does the sample provide enough evidence to confirm the researchers' belief? Test using $\alpha = 0.025$.

*Solution.*

**Example 96.** Suppose that for a particular application it is required that the true average DCP value for a certain type of pavement be less than 30. The pavement will not be used unless there is conclusive evidence that the specification has been met. A descriptive summary obtained from a sample of $n = 52$ data shows that the sample mean $\overline{x} = 28.76$ and the sample sd $s = 12.2647$. Let's state and test the appropriate hypotheses for the use of the pavement.

*Solution.*

Lec 26

## 8.3 The One-Sample $t$ Test

When $n$ is small, the Central Limit Theorem (CLT) can no longer be invoked to justify the use of a large-sample test. Our approach here is _____ _____ and describing test procedures whose validity rests on this assumption.

Hypothesis test for $\mu$ when:
1. sample size is small
2. $X$ follows normal distribution
3. $\sigma$ is unknown

Then the test statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $n-1$ degrees of freedom (df).
Null hypothesis: $H_0 : \mu = \mu_0$
Significant level: $\alpha$
Alternative Hypothesis:

| | | |
|---|---|---|
| $H_A : \mu > \mu_0$ | RR: $T \geq t_{n-1,\alpha}$ | upper-tail test |
| $H_A : \mu < \mu_0$ | RR: $T \leq -t_{n-1,\alpha}$ | lower-tail test |
| $H_A : \mu \neq \mu_0$ | RR: $\lvert T\rvert \geq t_{n-1,\alpha/2}$ | two-tail test |

**Example 97.** (HW #7) Light bulbs of a certain type are advertised as having an average lifetime of 750 hours. The price of these bulbs is very favorable and so a potential customer has decided to go ahead with the purchase unless it can be conclusively demonstrated that the true average lifetime is smaller than what is advertised. A random sample of 20 bulbs was selected and the lifetime of each was recorded. Suppose that the sample mean was 738.4 hours with a sample standard deviation of 41.2 hours. Does the sample provide evidence that the true mean lifetime is less than 750? Assume lifetimes vary according to a normal distribution and test using $\alpha = 0.10$.

*Solution.*

**Example 98.** After a 24-hour smoking abstinence, each of 20 smokers was asked to estimate how much time had elapsed during a 45-second period. The collected elapsed time gives sample mean $\overline{x} = 59.3$ sec and sample sd $s = 9.84$ sec. Assume the data follows a normal distribution. Let's carry out a test of hypotheses at significance level 0.05 to decide whether true average perceived elapsed time differs from the known time 45 sec.

*Solution.*

## 8.4 Tests Concerning a Population Proportion

Let $p$ denote the proportion of individuals or objects in a population who possess a specified property (e.g., college students who graduate without any debt, or computers that do not need service during the warranty period). If an individual or object with the property is labeled a success ($S$), then $p$ is _____.

$X$ (the number of $S$'s in the sample) has approximately _____. Furthermore, if $n$ is large [$np \geq 10$ and $n(1-p) \geq 10$], both $X$ and the estimator $\hat{p} = \frac{X}{n}$ are approximately _____.

For the estimator $\hat{p} = \frac{X}{n}$, we have

It follows that when $n$ is large and $H_0 : p = p_0$ is true, the test statistic

has approximately _____.

Null hypothesis: $H_0 : p = p_0$
Test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Significance level : $\alpha$
Alternative Hypothesis:

| | | |
|---|---|---|
| $H_A : p > p_0$ | RR: $Z \geq z_\alpha$ | upper-tailed test |
| $H_A : p < p_0$ | RR: $Z \leq -z_\alpha$ | lower-tailed test |
| $H_A : p \neq p_0$ | RR: $\lvert Z \rvert \geq z_{\alpha/2}$ | two-tailed test |

These test procedures are valid provided that

$$np_0 \geq 10 \quad \text{and} \quad n(1-p_0) \geq 10$$

**Example 99.** The use of a phone to text during an exam is a serious breach of conduct. One article reported that 27 of the 267 students in a sample admitted to doing this. Can it be concluded at significance level 0.001 that more than 5% of all students in the population sampled had texted during an exam?

*Solution.*

**Example 100.** A plan for an executive travelers' club has been developed by an airline on the premise that 5% of its current customers would qualify for membership. A random sample of 500 customers yielded 40 who would qualify. Using this data, test at level 0.01 the null hypothesis that the company's premise is correct against the alternative that it is not correct.

*Solution.*

Lec 27

**Section 8.1: $P$-value**

One way to report the result of a hypothesis test is using _____.

**Definition 32.** The $P$-value is _____, calculated _____, of obtaining a value of the test statistic at least _____ as the value calculated from the available sample data.

A conclusion is reached in a hypothesis testing analysis by comparing the $P$-value with the specified significant level $\alpha$.

1. $P$-value $\leq \alpha \Longrightarrow$

2. $P$-value $> \alpha \Longrightarrow$

$P$-value for $z$ Test:

**Example 92 continued :** The melting point of each of 16 samples of a certain brand of hydrogenated vegetable oil was determined, resulting in $\overline{x} = 94.32$. Assume that the distribution of the melting point is normal with $\sigma = 1.2$.

**Question:** Test $H_0 : \mu = 95$ versus $H_A : \mu \neq 95$ using a two-tailed level 0.01 test.

*Solution.* Given $n = 16$, $\overline{x} = 94.32$, $\sigma = 1.2$, $X \sim N(\mu,\ \sigma = 1.2)$.So

$$\overline{X} \sim N\left(\mu,\ \frac{\sigma}{\sqrt{n}} = \frac{1.2}{4} = 0.3\right)$$

Step 1: $H_0 : \mu = 95$ versus $H_A : \mu \neq 95$ Note, this is a two tail test.

Step 2: Assume the null is true, i.e.

$$\overline{X} \sim N\,(95,\ 0.3)$$

Step 3: Since $\alpha = 0.01$, so from the $z$-table we have the critical values are

$$z_{\alpha/2} = 2.57, \quad -z_{\alpha/2} = -2.57$$

Thus, the RR = we will reject $H_0$ is $|z| \geq 2.57$.

Step 4: Test statistic

$$z = \frac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{94.32 - 95}{0.3} = -2.27$$

Step 5: Decision: we fail to reject $H_0$.

Step 6: Conclusion: Bases on our evidence at $\alpha = 0.01$ or at the 1% significance level, we cannot conclude that the average melting point is different from 95°F.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 93 continued :** A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°F. A sample of $n = 9$ systems yields a sample average activation temperature of 131.08°F. If the distribution of activation times is **normal with standard deviation** 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = 0.01$?

*Solution.*

Step 1:

Hypothesis: $H_0 : \mu = 130$ vs $H_A : \mu \neq 130$

Step 2: Assume $H_0$ is true

Step 3: From the $z$-table, we have $\Phi(2.575) = 0.995 = 1 - \alpha/2$. Thus, the critical value is $z_{0.005} = 2.575$, the RR for a two-tail test is: reject $H_0$ if $|Z| > 2.575$.

Step 4: Test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{131.08 - 130}{1.5/\sqrt{9}} = 2.16$$

Step 5: Decision: since $Z = 2.16 < 2.575$, so we fail to reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha$= 0.01, we conclude that the data does not give strong support to the claim that the true average differs from the design value of 130.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 94 continued :** The desired percentage of $SiO_2$ in a certain type of aluminous cement is 5.5. To test whether the true average percentage is 5.5 for a particular production facility, 16 independently obtained samples are analyzed. Suppose that the percentage of $SiO_2$ in a sample is **normally distributed** with $\sigma = 0.3$ and that $\overline{x} = 5.25$.

Does this indicate conclusively that the true average percentage differs from 5.5?

*Solution.*

Step 1: Hypothesis: $H_0 : \mu = 5.5$ vs $H_A : \mu \neq 5.5$

Step 2: Assume $H_0$ is true.

Step 3: Since part (a) does not specify a significant level, so we use $\alpha = 0.05$ for this two-tailed test. From the $z$-table we have $\Phi(1.96) = 0.975$ which give the critical value $z_{0.025} = 1.96$. So RR: we reject $H_0$ if $|Z| \geq 1.96$.

Step 4: Test statistic

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{5.25 - 5.5}{0.3/\sqrt{16}} = -3.33$$

Step 5: Decision: since $|Z| = 3.33 > 1.96$, we reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha$= 0.05, we conclude that the true average percentage differs from 5.5.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 95 continued :** The biological dessert in the Gulf of Mexico called the Dead Zone is a region in which there is very little or no oxygen. Most marine life in the Dead Zone dies or leaves the region. The area of this region varies and is affected by agriculture, fertilizer runoff, and weather. The long-term mean area of the Dead Zone is 5960 square miles. As a result of recent flooding in the Midwest and subsequent runoff from the Mississippi River, researchers believe that the Dead Zone area will increase. A random sample of 50 days was obtained and the sample mean area of the Dead Zone was 6759 $mi^2$ with a sample standard deviation of 1850 $mi^2$. Does the sample provide enough evidence to confirm the researchers' belief? Test using $\alpha = 0.025$.

*Solution.* Given information $\mu = 5960$, $n = 50$, $\overline{x} = 6759$, $s = 1850$, $\alpha = 0.025$.

Step 1: $H_0 : \mu = 5960$ vs $H_A : \mu > 5960$ (right tail test)

Step 2: Assume $H_0$ is true.

Step 3: From $z$-table, we have $\Phi(1.96) = 0.975$, so the critical value $z_{0.025} = 1.96$. We will reject $H_0$ if $Z$ is greater than 1.96. RR: $Z \geq 1.96$

Step 4:

$$z = \frac{\overline{x} - \mu_{\overline{X}}}{s/\sqrt{n}} = 3.05$$

Step 5: Decision: we reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha = 0.025$, we conclude that the area of the Dead zone was increased.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 96 continued :** Suppose that for a particular application it is required that the true average DCP value for a certain type of pavement be less than 30. The pavement will not be used unless there is conclusive evidence that the specification has been met. A descriptive summary obtained from a sample of $n = 52$ data shows that the sample mean $\overline{x} = 28.76$ and the sample sd $s = 12.2647$. Let's state and test the appropriate hypotheses for the use of the pavement.

*Solution.*

Step 1: Hypothesis: $H_0 : \mu = 30$ vs $H_A : \mu < 30$

So the pavement will not be used unless the null hypothesis is rejected.

Step 2: Assume the $H_0$ is true.

Step 3: The significant level is not specified, so we use $\alpha = 0.05$. Since we have a lower-tail test and $\Phi(1.645) = 0.95$, the critical value is $-z_{0.05} = -1.645$. RR: we reject $H_0$ if $Z \leq -1.645$.

Step 4: Test statistics

$$Z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{28.76 - 30}{12.2647/\sqrt{52}} = -0.73$$

Step 5: Decision: since $Z = -0.73 > -1.64$, we fail to reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha$= 0.05, we conclude that the use of the pavement is not justified.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 97 continued :** Light bulbs of a certain type are advertised as having an average lifetime of 750 hours. The price of these bulbs is very favorable and so a potential customer has decided to go ahead with the purchase unless it can be conclusively demonstrated that the true average lifetime is smaller than what is advertised. A random sample of 20 bulbs was selected and the lifetime of each was recorded. Suppose that the sample mean was 738.4 hours with a sample standard deviation of 41.2 hours. Does the sample provide evidence that the true mean lifetime is less than 750? Assume lifetimes vary according to a normal distribution and test using $\alpha = 0.10$.

*Solution.* Given information $\mu = 750$, $n = 20$, $\overline{x} = 738.4$, $s = 41.2$, $\alpha = 0.1$.

Step 1: $H_0 : \mu = 750$ vs $H_A : \mu < 750$ (lower-tail test)

Step 2: Assume $H_0$ is true.

Step 3: From $t$-table, we have $t_{n-1,\alpha} = t_{19,0.1} = 1.328$. Since we have a left tail test, so the critical value is $-t_{19,0.1} = -1.328$. We will reject $H_0$ if $T$ is smaller than -1.328. RR: $T \leq -1.328$

Step 4:

$$t = \frac{\overline{x} - \mu_{\overline{X}}}{s/\sqrt{n}} = -1.259 > -1.328$$

Step 5: Decision: we fail to reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha = 0.1$, we conclude that the sample does not have enough evidence to show the true mean bulb lifetime is less than 750 hours.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 98 continued :** After a 24-hour smoking abstinence, each of 20 smokers was asked to estimate how much time had elapsed during a 45-second period. The collected elapsed time gives sample mean $\overline{x} = 59.3$ sec and sample sd $s = 9.84$ sec. Assume the data follows a normal distribution. Let's carry out a test of hypotheses at significance level 0.05 to decide whether true average perceived elapsed time differs from the known time 45 sec.

*Solution.* $\mu =$ true average perceived elapsed time for all smokers.

Hypothesis: $\mu = 45$ vs $\mu \neq 45$

Test statistic:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{59.3 - 45}{9.84/\sqrt{20}} = 6.5$$

Rejection Region: Since we have a two-tailed test , so we reject $H_0$ if $|T| \geq t_{n-1,\alpha/2} = t_{19,0.025} = 2.093$.

Decision: Since $T = 6.5 \geq 2.093$, we reject $H_0$.

Conclusion: Based on our evidence at $\alpha = 0.05$, we conclude that the true average perceived elapsed time is evidently something other than 45.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 99 continued :** The use of a phone to text during an exam is a serious breach of conduct. One article reported that 27 of the 267 students in a sample admitted to doing this. Can it be concluded at significance level 0.001 that more than 5% of all students in the population sampled had texted during an exam?

*Solution.* The parameter of interest is the proportion $p$ of the sampled population that has texted during an exam.

Step 1: Hypothesis: $H_0 : p = 0.05$ vs $H_A : p > 0.05$

Step 2: Assume $H_0$ is true and check conditions:

$$np_0 = (267)(0.05) = 13.35 \geq 10 \quad \text{and} \quad n(1 - p_0) = (267)(0.95) = 253.65 \geq 10$$

the large-sample $z$ test can be used.

Step 3: Since it is the upper-tail test, from the $z$-table we have the critical value is $z_{0.001} = 3.1$ which give RR : reject $H_0$ if $Z \geq 3.1$.

Step 4: Since $\hat{p} = \frac{27}{267} = 0.1011$, then the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1011 - 0.05}{\sqrt{\frac{(0.05)(0.95)}{267}}} = 3.84$$

Step 5: Decision: we reject $H_0$.

Step 6: Conclusion: Based on our evidence at $\alpha = 0.001$, we conclude that the evidence for concluding that the population percentage of students who text during an exam exceeds 5% is very compelling.

Let's achieve the same conclusion by calculating the $P$-value.

**Example 100 continued :** A plan for an executive travelers' club has been developed by an airline on the premise that 5% of its current customers would qualify for membership. A random sample of 500 customers yielded 40 who would qualify. Using this data, test at level 0.01 the null hypothesis that the company's premise is correct against the alternative that it is not correct.

*Solution.*

Hypothesis: $H_0 : p = 0.05$ vs $H_A : p \neq 0.05$

Test statistic:

$$\hat{p} = \frac{40}{500} = 0.08$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.08 - 0.05}{\sqrt{\frac{(0.05)(0.95)}{500}}} = 3.0779$$

Rejection Region: $\alpha = 0.01$ and we have a two tailed test here, so the critical value is

$$z_{\alpha/2} = z_{0.005} = 2.575$$

So we reject $H_0$ when $|Z| \geq 2.575$.

Decision: since $Z = 3.0779 \geq 2.575$, we reject $H_0$.

Conclusion: Based on our evidence at $\alpha = 0.01$, we conclude that the company's premise is not correct.

Let's achieve the same conclusion by calculating the $P$-value.

Lec 28

# 9 Inferences Based on Two Samples

Chapters 7 and 8 presented confidence intervals (CI's) and hypothesis-testing procedures for a single mean $\mu$, and a single proportion $p$. Chapter 9 extend these methods to situations involving the means, proportions, and variances of _____.

## 9.1 $z$ Tests and Confidence Intervals for a Difference Between Two Population Means

The inferences discussed in this section concern _____
_____.

Basic Assumptions:

- Suppose we have a random sample from population 1 with mean _____, population standard deviation _____, sample size _____, sample mean _____, and sample standard deviation _____.
- Suppose we have a random sample from population 2 with mean _____, population standard deviation _____, sample size _____, sample mean _____, and sample standard deviation _____.
- The two samples are _____ of one another.

The natural estimator of $\mu_1 - \mu_2$ is _____, the difference between the corresponding _____. Inferential procedures are based on _____ this estimator, so we need expressions for the expected value and standard deviation of $\overline{X}_1 - \overline{X}_2$.

Let the random variable $Y = \overline{X}_1 - \overline{X}_2$, then

### 9.1.1 Test Procedures for Normal Populations with Known Variances

Assume that both population distributions are _____ and the values of _____ _____. Thus the difference $\overline{X}_1 - \overline{X}_2$ is also _____ distributed, with expected value _____ and standard deviation _____given previously. Standardizing $\overline{X}_1 - \overline{X}_2$ gives the standard normal variable

Null Hypothesis:
Test statistic:

Significance level:
Alternative Hypothesis:

Assumptions:

- 
- 
- 

$P$-value for $z$ Test:

**Example 101.** Analysis of a random sample consisting of $m = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\overline{x} = 29.8$ ksi. A second random sample of $n = 25$ two-sided galvanized steel specimens gave a sample average strength of $\overline{y} = 34.7$ ksi. Assuming that the two yield-strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths $\mu_1$ and $\mu_2$ are different? Test at significance level $\alpha = 0.01$.

*Solution.*

### 9.1.2 Large-Sample Tests

The assumptions of _____ population distributions and _____ values of $\sigma_1$ and $\sigma_2$ are fortunately unnecessary when both sample sizes are _____. In this case, the Central Limit Theorem guarantees that $\overline{X_1} - \overline{X_2}$ has approximately a _____ distribution regardless of the underlying population distributions.

Null Hypothesis:
Test statistic:

Significance level:
Alternative Hypothesis:

Assumptions:

- 
- 

### 9.1.3 Confidence Intervals for $\mu_1 - \mu_2$

When both population distributions are (at least approximately) normal, standardizing $\overline{X}_1 - \overline{X}_2$ gives a random variable Z with a standard normal distribution. Since the area under the $z$ curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

which implies

This implies that a $100(1-\alpha)\%$ for $\mu_1 - \mu_2$.

Provided that $n_1 > 40$ and $n_2 > 40$, a CI for $\mu_1 - \mu_2$ with a confidence level of $100(1-\alpha)\%$ is

**Example 102.**

Let $\mu_1$ = the true average tread life for a premium brand of P205/65R15 radial tire, and let $\mu_2$ = the true average tread life for an economy brand of the same size. Independent random samples of tires of each brand were obtained and yielded the summaries below.

| | Sample Size | Sample Mean | Sample Std. Dev. |
|---|---|---|---|
| Premium | $n_1 = 45$ | $\bar{x}_1 = 42500$ | $s_1 = 2200$ |
| Economy | $n_2 = 45$ | $\bar{x}_2 = 36800$ | $s_2 = 1500$ |

(a) Do the samples give enough evidence to conclude that $\mu_1$ will exceed $\mu_2$ by more than 5000 miles? Test $H_0 : \mu_1 - \mu_2 = 5000$ versus $H_A : \mu_1 - \mu_2 > 5000$ using the level of significance $\alpha = 0.05$.

(b) By how many miles will $\mu_1$ exceed $\mu_2$ ? Calculate a 90% confidence interval to estimate the true value of $\mu_1 - \mu_2$ .

(c) Do the tread lives of the two brands need to be normally distributed for the test of hypotheses and the confidence interval to be valid? Why or why not?

*Solution.*

Lec 29

## 9.2 The Two-Sample $t$ Test and Confidence Interval

### Pooled $t$ Procedures

- Consider 2 _____ samples from 2 _____ populations.
- _____ for at least one sample.
- Population variance _____.

The pooled test statistic uses a _____ average of the two sample variances:

Null Hypothesis:
Test statistic:

Significant level:
Alternative Hypothesis:

Assumptions:

- 
- 
- 

Provided the above assumptions in a pooled $t$ test, a CI for $\mu_1 - \mu_2$ with a confidence level of $100(1-\alpha)\%$ is

**Example 103.**

Many homeowners use tiki torches for outside decoration and to burn special oil to repel insects. Independent random samples of two types of oil were obtained and the burn time for 3 ounces of each was recorded (in hours). The summary statistics are given in the following table. Assume the underlying populations of burn times are normal.

| Oil | Sample Size | Sample Mean | Sample Variance |
|---|---|---|---|
| Citronella Torch Fuel | 18 | 6.25 | 1.04 |
| Black Flag Mosquito Control | 24 | 5.98 | 0.77 |

Based on the sample variances, do you think the assumption of equal population variances is reasonable? Why or why not?

**Example 104.**

Frequently, patients must wait a long time for elective surgery. Suppose the wait time for patients needing a knee replacement at two hospitals was investigated. Independent random samples of patients were obtained and the wait time for each (in weeks) was recorded. The resulting summary statistics are given in the following table.

| Hospital | Sample Size | Sample Mean | Sample Variance |
|---|---|---|---|
| Hospital 1 | 15 | 17.4 | 34.81 |
| Hospital 2 | 17 | 12.1 | 46.24 |

(a) Do the samples give enough evidence to conclude that the mean wait time at hospital 1 is longer than that of hospital 2? Assume that the unknown population variances are equal and test using $\alpha = 0.05$. [12 points]

(b) Find bounds on the p-value associated with the test. [2 points]

(c) How much is the mean wait time for hospital 1 longer than the mean for hospital 2? Calculate a 90% confidence interval to estimate the true value of $\mu_1 - \mu_2$. [4 points]

(d) Besides the population variances being equal, what else must be true (or what else must we assume) about the populations for parts (a) and (b) to be valid? [2 points]

## 9.3 Analysis of Paired Data

We are interested in the _____ between two observations of each subject. Suppose the data consists of $n$ independently selected pairs $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, with $E(X_i) = \mu$ and $E(Y_i) = \mu_2$. Let _____, so the $D_i$'s are the difference within pairs.

**The Paired $t$ test**

Because different pairs are independent, the $D_i$'s are _____ of one another. Let $D = X - Y$, where $X$ and $Y$ are the first and second observations, respectively, within an arbitrary pair. Then the expected difference is

Null Hypothesis:
Test statistic value:

Significance level:
Alternative Hypothesis:

Assumptions: The $D_i$'s constitute a random sample from a _____ "difference" population.

**Example 105.**

Seven patients who claim to be suffering from job related stress were selected at random. After an initial resting pulse rate (in beats per minute) was obtained, each person participated in a relaxation therapy program. A final resting pulse rate was taken at the end of the program. The data is given in the following table.

| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Initial Pulse Rate | 67 | 71 | 67 | 84 | 71 | 68 | 73 |
| Final Pulse Rate | 61 | 72 | 70 | 76 | 74 | 59 | 61 |
| Difference (Initial – Final) | 6 | –1 | –3 | 8 | –3 | 9 | 12 |

Lec 30

## 9.4 Inferences Concerning a Difference Between Population Proportions

Population 1:

$n_1 = \#$ of observations in sample 1

$X_1 = \#$ of subjects in sample 1 that have a certain characteristic we are interested

$\hat{p}_1 = \dfrac{X_1}{n_1} =$ sample 1 proportion

$p_1 =$ population 1 proportion (unknown)

Population 2:

$n_2 = \#$ of observations in sample 2

$X_2 = \#$ of subjects in sample 2 that have the same characteristic we are interested

$\hat{p}_2 = \dfrac{X_2}{n_2} =$ sample 2 proportion

$p_2 =$ population 2 proportion (unknown)

The natural estimator for $p_1 - p_2$, the difference in population proportions, is the corresponding difference in sample proportions _____. Since we know _____ _____with $X_1$ and $X_2$ are _____ variables. Then

### A Large-Sample Test Procedure

Suppose we want to test _____ or equivalently _____. When $H_0$ is true, let $p$ denote the common value of $p_1$ and $p_2$, i.e. _____. Then we have the standardized variable

has approximately a _____ distribution when $H_0$ is true.

Null Hypothesis:
Test statistic:

where

Significant level:
Alternative Hypothesis:

Assumptions:

- 
- 

Provided the above assumptions satisfied, a CI for $p_1 - p_2$ with a confidence level of $100(1-\alpha)\%$ is

**Example 106.**

A survey of 400 elementary school teachers (Group 1) and 300 high school teachers (Group 2) was conducted. Of the elementary teachers, 224 said they were very satisfied with their jobs; whereas, 141 of the high school teachers were very satisfied with their work.

(a) Do the samples give enough evidence to conclude that a larger proportion of elementary school teachers are more satisfied with their jobs? Test using $\alpha = 0.05$.

(b) (i) What conditions must be verified for the test in part (a) to be valid?
(ii) Check the conditions and indicate whether or not they are satisfied.

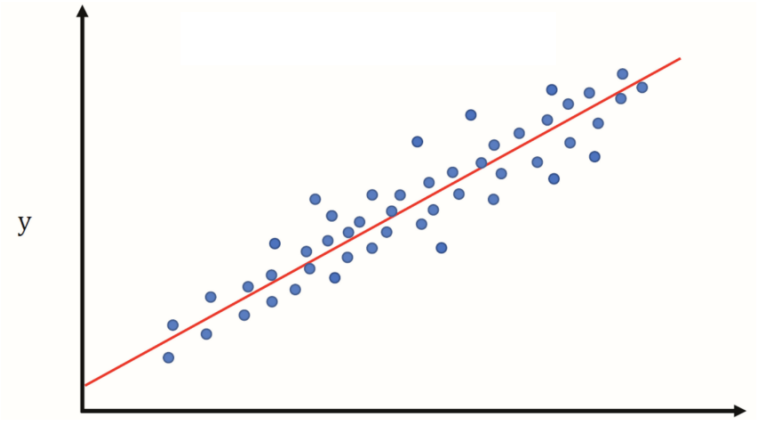
# 12 Simple Linear Regression and Correlation

_____ is the part of statistics that investigates the relationship between two or more variables related in a non-deterministic fashion.

## 12.1 The Simple Linear Regression Model

The simplest deterministic mathematical relationship between two variables $x$ and $y$ is called a _____

The set of pairs $(x, y)$ determines _____ with _____ _____.




_____ gives preliminary impression about the nature of a relationship between $(x_i, y_i)$ in a 2 dimension coordinate system.

## 12.2 Estimating Model Parameters

$\beta_0$ and $\beta_1$ are almost never known. We take a sample and approximate for $\beta_0$ and $\beta_1$.

The _____ of the slope coefficient $\beta_1$ of the true regression line is

Computing formulas for the numerator and denominator of $\hat{\beta}_1$ are

The least squares estimate of the intercept $\beta_0$ of the true regression line is

We define the _____ to be an equation:

Now, for each _____ we obtain a corresponding _____ _____, which can be used to estimate _____.

**Example 107.**

Suppose that a small coastal community is considering the construction of a windmill to generate electricity for the town hall. A study is conducted to measure the windmill's noise level (in dB) at various distances (in meters) from the proposed site. The data is summarized in the table below.

| X = Distance | 10 | 50 | 75 | 120 | 150 | 160 | 200 | 250 | 400 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y = Noise Level | 75 | 110 | 73 | 52 | 58 | 77 | 56 | 57 | 28 | 4 |

(a) (i) Construct a scatterplot of the data.
(ii) Does a simple linear regression model appear to be reasonable in this situation?

(b) Calculate the equation of the least squares regression line for predicting noise level based on the distance from the windmill.

(c) Predict the noise level at a distance of 100 meters.