

MCA

SEMESTER - II

**PROBABILITY &
STATISTICS**

PROBABILITY

INTRODUCTION TO PROBABILITY

Managers need to cope with uncertainty in many decision making situations. For example, you as a manager may assume that the volume of sales in the successive year is known exactly to you. This is not true because you know roughly what the next year sales will be. But you cannot give the exact number. There is some uncertainty. Concepts of probability will help you to measure uncertainty and perform associated analyses. This unit provides the conceptual framework of probability and the various probability rules that are essential in business decisions.

Learning objectives:

After reading this unit, you will be able to:

- Appreciate the use of probability in decision making
- Explain the types of probability
- Define and use the various rules of probability depending on the problem situation.
- Make use of the expected values for decision-making.

Probability

Sets and Subsets

The lesson introduces the important topic of sets, a simple idea that recurs throughout the study of probability and statistics.

Set Definitions

- A **set** is a well-defined collection of objects.
- Each object in a set is called an **element** of the set.
- Two sets are **equal** if they have exactly the same elements in them.
- A set that contains no elements is called a **null set** or an **empty set**.
- If every element in Set *A* is also in Set *B*, then Set *A* is a **subset** of Set *B*.

Set Notation

- A set is usually denoted by a capital letter, such as *A*, *B*, or *C*.
- An element of a set is usually denoted by a small letter, such as *x*, *y*, or *z*.
- A set may be described by listing all of its elements enclosed in braces. For example, if Set *A* consists of the numbers 2, 4, 6, and 8, we may say: $A = \{2, 4, 6, 8\}$.
- The null set is denoted by $\{\emptyset\}$.

- Sets may also be described by stating a rule. We could describe Set A from the previous example by stating: Set A consists of all the even single-digit positive integers.

Set Operations

Suppose we have four sets - W , X , Y , and Z . Let these sets be defined as follows: $W = \{2\}$; $X = \{1, 2\}$; $Y = \{2, 3, 4\}$; and $Z = \{1, 2, 3, 4\}$.

- The **union** of two sets is the set of elements that belong to one or both of the two sets. Thus, set Z is the union of sets X and Y .
- Symbolically, the union of X and Y is denoted by $X \cup Y$.
- The **intersection** of two sets is the set of elements that are common to both sets. Thus, set W is the intersection of sets X and Y .
- Symbolically, the intersection of X and Y is denoted by $X \cap Y$.

Sample Problems

1. Describe the set of vowels.

If A is the set of vowels, then A could be described as $A = \{a, e, i, o, u\}$.

2. Describe the set of positive integers.

Since it would be impossible to list *all* of the positive integers, we need to use a rule to describe this set. We might say A consists of all integers greater than zero.

3. Set $A = \{1, 2, 3\}$ and Set $B = \{3, 2, 1\}$. Is Set A equal to Set B ?

Yes. Two sets are equal if they have the same elements. The order in which the elements are listed does not matter.

4. What is the set of men with four arms?

Since all men have two arms at most, the set of men with four arms contains no elements. It is the null set (or empty set).

5. Set $A = \{1, 2, 3\}$ and Set $B = \{1, 2, 4, 5, 6\}$. Is Set A a subset of Set B ?

Set A would be a subset of Set B if every element from Set A were also in Set B . However, this is not the case. The number 3 is in Set A , but not in Set B . Therefore, Set A is not a subset of Set B .

Statistical Experiments

All **statistical experiments** have three things in common:

- The experiment can have more than one possible outcome.
- Each possible outcome can be specified in advance.

- The outcome of the experiment depends on chance.

A coin toss has all the attributes of a statistical experiment. There is more than one possible outcome. We can specify each possible outcome (i.e., heads or tails) in advance. And there is an element of chance, since the outcome is uncertain.

The Sample Space

- A **sample space** is a set of elements that represents all possible outcomes of a statistical experiment.
- A **sample point** is an element of a sample space.
- An **event** is a subset of a sample space - one or more sample points.
-

Types of events

- Two events are **mutually exclusive** if they have no sample points in common.
- Two events are **independent** when the occurrence of one does not affect the probability of the occurrence of the other.

Sample Problems

1. Suppose I roll a die. Is that a statistical experiment?

Yes. Like a coin toss, rolling dice is a statistical experiment. There is more than one possible outcome. We can specify each possible outcome in advance. And there is an element of chance.

2. When you roll a single die, what is the sample space.

The sample space is all of the possible outcomes - an integer between 1 and 6.

3. Which of the following are sample points when you roll a die - 3, 6, and 9?

The numbers 3 and 6 are sample points, because they are in the sample space. The number 9 is not a sample point, since it is outside the sample space; with one die, the largest number that you can roll is 6.

4. Which of the following sets represent an event when you roll a die?

- A. {1}
- B. {2, 4,}
- C. {2, 4, 6}
- D. All of the above

The correct answer is D. Remember that an event is a subset of a sample space. The sample space is any integer from 1 to 6.

Each of the sets shown above is a subset of the sample space, so each represents an event.

5. Consider the events listed below. Which are mutually exclusive?

- A. {1}
- B. {2, 4,}
- C. {2, 4, 6}

Two events are mutually exclusive, if they have no sample points in common. Events A and B are mutually exclusive, and Events A and C are mutually exclusive; since they have no points in common. Events B and C have common sample points, so they are not mutually exclusive.

6. Suppose you roll a die two times. Is each roll of the die an independent event?

Yes. Two events are independent when the occurrence of one has no effect on the probability of the occurrence of the other. Neither roll of the die affects the outcome of the other roll; so each roll of the die is independent.

Basic Probability

The **probability** of a sample point is a measure of the likelihood that the sample point will occur.

Probability of a Sample Point

By convention, statisticians have agreed on the following rules.

- The probability of any sample point can range from 0 to 1.
- The sum of probabilities of all sample points in a sample space is equal to 1.
-

Example 1

Suppose we conduct a simple statistical experiment. We flip a coin one time. The coin flip can have one of two outcomes - heads or tails. Together, these outcomes represent the sample space of our experiment. Individually, each outcome represents a sample point in the sample space. What is the probability of each sample point?

Solution: The sum of probabilities of all the sample points must equal 1. And the probability of getting a head is equal to the probability of getting a tail. Therefore, the probability of each sample point (heads or tails) must be equal to $1/2$.

Example 2

Let's repeat the experiment of Example 1, with a die instead of a coin. If we toss a fair die, what is the probability of each sample point?

Solution: For this experiment, the sample space consists of six sample points: {1, 2, 3, 4, 5, 6}. Each sample point has equal probability. And the sum of probabilities of all the sample points

must equal 1. Therefore, the probability of each sample point must be equal to $1/6$.

Probability of an Event

The probability of an event is a measure of the likelihood that the event will occur. By convention, statisticians have agreed on the following rules.

- The probability of any event can range from 0 to 1.
- The probability of event A is the sum of the probabilities of all the sample points in event A.
- The probability of event A is denoted by $P(A)$.

Thus, if event A were very unlikely to occur, then $P(A)$ would be close to 0. And if event A were very likely to occur, then $P(A)$ would be close to 1.

Example

1

Suppose we draw a card from a deck of playing cards. What is the probability that we draw a spade?

Solution: The sample space of this experiment consists of 52 cards, and the probability of each sample point is $1/52$. Since there are 13 spades in the deck, the probability of drawing a spade is $P(\text{Spade}) = (13)(1/52) = 1/4$

Example

2

Suppose a coin is flipped 3 times. What is the probability of getting two tails and one head?

Solution: For this experiment, the sample space consists of 8 sample points.

$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$

Each sample point is equally likely to occur, so the probability of getting any particular sample point is $1/8$. The event "getting two tails and one head" consists of the following subset of the sample space.

$A = \{TTH, THT, HTT\}$

The probability of Event A is the sum of the probabilities of the sample points in A. Therefore,

$$P(A) = 1/8 + 1/8 + 1/8 = 3/8$$

Working With Probability

The **probability** of an event refers to the likelihood that the event will occur.

How to Interpret Probability

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1. Notationally, the probability of event A is represented by $P(A)$.

- If $P(A)$ equals zero, there is no chance that the event A will occur.

- If $P(A)$ is close to zero, there is little likelihood that event A will occur.
- If $P(A)$ is close to one, there is a strong chance that event A will occur
- If $P(A)$ equals one, event A will definitely occur.

The sum of all possible outcomes in a statistical experiment is equal to one. This means, for example, that if an experiment can have three possible outcomes (A , B , and C), then $P(A) + P(B) + P(C) = 1$.

How to Compute Probability: Equally Likely Outcomes

Sometimes, a statistical experiment can have n possible outcomes, each of which is equally likely. Suppose a subset of r outcomes are classified as "successful" outcomes.

The probability that the experiment results in a successful outcome (S) is:

$$P(S) = (\text{Number of successful outcomes}) / (\text{Total number of equally likely outcomes}) = r / n$$

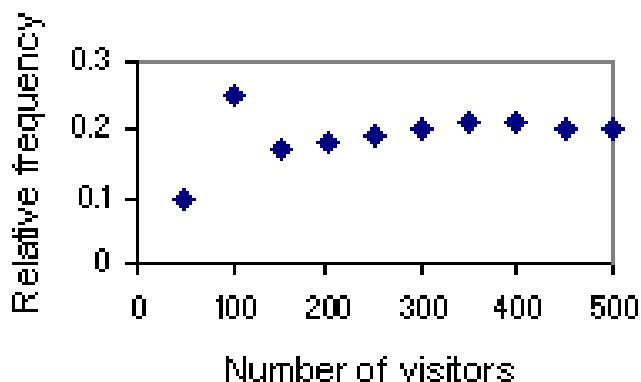
Consider the following experiment. An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is $3/10$ or 0.30 .

How to Compute Probability: Law of Large Numbers

One can also think about the probability of an event in terms of its *long-run* relative frequency. The relative frequency of an event is the number of times an event occurs, divided by the total number of trials.

$$P(A) = (\text{Frequency of Event } A) / (\text{Number of Trials})$$



For example, a merchant notices one day that 5 out of 50 visitors to her store make a purchase. The next day, 20 out of 50 visitors make a purchase. The two relative frequencies ($5/50$ or 0.10 and $20/50$ or 0.40) differ. However, summing results over many visitors,

she might find that the probability that a visitor makes a purchase gets closer and closer 0.20.

The scatterplot (above right) shows the relative frequency as the number of trials (in this case, the number of visitors) increases. Over many trials, the relative frequency converges toward a stable value (0.20), which can be interpreted as the probability that a visitor to the store will make a purchase.

The idea that the relative frequency of an event will converge on the probability of the event, as the number of trials increases, is called the **law of large numbers**.

Test Your Understanding of This Lesson

Problem

A coin is tossed three times. What is the probability that it lands on heads *exactly* one time?

- | | |
|-----|-------|
| (A) | 0.125 |
| (B) | 0.250 |
| (C) | 0.333 |
| (D) | 0.375 |
| (E) | 0.500 |

Solution

The correct answer is (D). If you toss a coin three times, there are a total of eight possible outcomes. They are: HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT. Of the eight possible outcomes, three have exactly one head. They are: HTT, THT, and TTH. Therefore, the probability that three flips of a coin will produce *exactly* one head is $\frac{3}{8}$ or 0.375.

Rules of Probability

Often, we want to compute the probability of an event from the known probabilities of other events. This lesson covers some important rules that simplify those computations.

Definitions and Notation

Before discussing the rules of probability, we state the following definitions:

- Two events are **mutually exclusive** or **disjoint** if they cannot occur at the same time.
- The probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A, given Event B, is denoted by the symbol $P(A|B)$.
- The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by $P(A')$.
- The probability that Events A and B *both* occur is the probability of the **intersection** of A and B. The probability of

the intersection of Events A and B is denoted by $P(A \cap B)$. If Events A and B are mutually exclusive, $P(A \cap B) = 0$.

- The probability that Events A or B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by $P(A \cup B)$.
- If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are **independent**.

Probability Calculator

Use the Probability Calculator to compute the probability of an event from the known probabilities of other events. The Probability Calculator is free and easy to use. It can be found under the Tools menu item, which appears in the header of every Stat Trek web page.

Probability Calculator

Rule of Subtraction

In a previous lesson, we learned two important properties of probability:

- The probability of an event ranges from 0 to 1.
- The sum of probabilities of all possible events equals 1.

The rule of subtraction follows directly from these properties.

Rule of Subtraction The probability that event A will occur is equal to 1 minus the probability that event A will not occur.

$$P(A) = 1 - P(A')$$

Suppose, for example, the probability that Bill will graduate from college is 0.80. What is the probability that Bill will not graduate from college? Based on the rule of subtraction, the probability that Bill will not graduate is $1.00 - 0.80$ or 0.20 .

Rule of Multiplication

The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events (Event A and Event B) both occur.

Rule of Multiplication The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.

$$P(A \cap B) = P(A) P(B|A)$$

Example

An urn contains 6 red marbles and 4 black marbles. Two marbles

are drawn *without replacement* from the urn. What is the probability that both of the marbles are black?

Solution: Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, $P(A) = 4/10$.
- After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, $P(B|A) = 3/9$.

Therefore, based on the rule of multiplication:

$$P(A \cap B) = P(A) \cdot P(B|A)$$

$$P(A \cap B) = (4/10) \cdot (3/9) = 12/90 = 2/15$$

Rule of Addition

The rule of addition applies to the following situation. We have two events, and we want to know the probability that either event occurs.

Rule of Addition The probability that Event A and/or Event B occur is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Note: Invoking the fact that $P(A \cap B) = P(A)P(B|A)$, the Addition Rule can also be expressed as

$$P(A \cup B) = P(A) + P(B) - P(A)P(B|A)$$

Example

A student goes to the library. The probability that she checks out (a) a work of fiction is 0.40, (b) a work of non-fiction is 0.30, and (c) both fiction and non-fiction is 0.20. What is the probability that the student checks out a work of fiction, non-fiction, or both?

Solution: Let F = the event that the student checks out fiction; and let N = the event that the student checks out non-fiction. Then, based on the rule of addition:

$$P(F \cup N) = P(F) + P(N) - P(F \cap N)$$

$$P(F \cup N) = 0.40 + 0.30 - 0.20 = 0.50$$

Test Your Understanding of This Lesson

Problem 1

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *with replacement* from the urn. What is the probability that both of the marbles are black?

- | | |
|-----|------|
| (A) | 0.16 |
| (B) | 0.32 |
| (C) | 0.36 |
| (D) | 0.40 |
| (E) | 0.60 |

Solution

The correct answer is A. Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, $P(A) = 4/10$.
- After the first selection, we replace the selected marble; so there are still 10 marbles in the urn, 4 of which are black. Therefore, $P(B|A) = 4/10$.

Therefore, based on the rule of multiplication:

$$P(A \cap B) = P(A) \cdot P(B|A)$$

$$P(A \cap B) = (4/10) \cdot (4/10) = 16/100 = 0.16$$

Problem 2

A card is drawn randomly from a deck of ordinary playing cards. You win \$10 if the card is a spade or an ace. What is the probability that you will win the game?

- | | |
|-----|--------------------|
| (A) | 1/13 |
| (B) | 13/52 |
| (C) | 4/13 |
| (D) | 17/52 |
| (E) | None of the above. |

Solution

The correct answer is C. Let S = the event that the card is a spade; and let A = the event that the card is an ace. We know the following:

- There are 52 cards in the deck.
- There are 13 spades, so $P(S) = 13/52$.
- There are 4 aces, so $P(A) = 4/52$.
- There is 1 ace that is also a spade, so $P(S \cap A) = 1/52$.

Therefore, based on the rule of addition:

$$P(S \cup A) = P(S) + P(A) - P(S \cap A)$$

$$P(S \cup A) = 13/52 + 4/52 - 1/52 = 16/52 = 4/13$$

Bayes' Theorem (aka, Bayes' Rule)

Bayes' theorem (also known as Bayes' rule) is a useful tool for calculating conditional probabilities. Bayes' theorem can be stated as follows:

Bayes' theorem. Let A_1, A_2, \dots, A_n be a set of mutually exclusive events that together form the sample space S . Let B be any event from the same sample space, such that $P(B) > 0$. Then,

$$P(A_k | B) = \frac{P(A_k \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B)}$$

Note: Invoking the fact that $P(A_k \cap B) = P(A_k)P(B | A_k)$, Bayes' theorem can also be expressed as

$$P(A_k | B) = \frac{P(A_k) P(B | A_k)}{P(A_1) P(B | A_1) + P(A_2) P(B | A_2) + \dots + P(A_n) P(B | A_n)}$$

Unless you are a world-class statistician, Bayes' theorem (as expressed above) can be intimidating. However, it really is easy to use. The remainder of this lesson covers material that can help you understand when and how to apply Bayes' theorem effectively.

When to Apply Bayes' Theorem

Part of the challenge in applying Bayes' theorem involves recognizing the types of problems that warrant its use. You should consider Bayes' theorem when the following conditions exist.

- The sample space is partitioned into a set of mutually exclusive events $\{A_1, A_2, \dots, A_n\}$.
- Within the sample space, there exists an event B , for which $P(B) > 0$.
- The analytical goal is to compute a conditional probability of the form: $P(A_k | B)$.
- You know at least one of the two sets of probabilities described below.
 - $P(A_k \cap B)$ for each A_k
 - $P(A_k)$ and $P(B | A_k)$ for each A_k

Bayes Rule Calculator

Use the Bayes Rule Calculator to compute conditional probability, when Bayes' theorem can be applied. The calculator is free, and it is easy to use. It can be found under the Tools menu item, which appears in the header of every Stat Trek web page.

Bayes Rule Calculator

Sample Problem

Bayes' theorem can be best understood through an example. This section presents an example that demonstrates how Bayes' theorem can be applied effectively to solve statistical problems.

Example**1**

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

Solution: The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. Notation for these events appears below.

- Event A_1 . It rains on Marie's wedding.
- Event A_2 . It does not rain on Marie's wedding
- Event B . The weatherman predicts rain.

In terms of probabilities, we know the following:

- $P(A_1) = 5/365 = 0.0136985$ [It rains 5 days out of the year.]
- $P(A_2) = 360/365 = 0.9863014$ [It does not rain 360 days out of the year.]
- $P(B | A_1) = 0.9$ [When it rains, the weatherman predicts rain 90% of the time.]
- $P(B | A_2) = 0.1$ [When it does not rain, the weatherman predicts rain 10% of the time.]

We want to know $P(A_1 | B)$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes' theorem, as shown below.

$$P(A_1 | B) = \frac{P(A_1) P(B | A_1)}{P(A_1) P(B | A_1) + P(A_2) P(B | A_2)}$$

$$P(A_1 | B) = (0.014)(0.9) / [(0.014)(0.9) + (0.986)(0.1)]$$

$$P(A_1 | B) = 0.111$$

Note the somewhat unintuitive result. When the weatherman predicts rain, it actually rains only about 11% of the time. Despite the weatherman's gloomy prediction, there is a good chance that Marie will not get rained on at her wedding.

This is an example of something called the false positive paradox. It illustrates the value of using Bayes theorem to calculate conditional probabilities.

Probability

For an experiment we define an *event* to be any collection of possible outcomes.

A *simple event* is an event that consists of exactly one outcome.

or: means the union i.e. either can occur

and: means intersection i.e. both must occur

Two events are *mutually exclusive* if they cannot occur simultaneously.

For a Venn diagram, we can tell that two events are mutually exclusive if their regions do not intersect

We define *Probability* of an event E to be to be

$$P(E) = \frac{\text{number of simple events within E}}{\text{total number of possible outcomes}}$$

We have the following:

1. $P(E)$ is always between 0 and 1.
2. The sum of the probabilities of all simple events must be 1.
3. $P(E) + P(\text{not } E) = 1$
4. If E and F are mutually exclusive then

$$P(E \text{ or } F) = P(E) + P(F)$$

The Difference Between And and Or

If E and F are events then we use the terminology

E and F

to mean all outcomes that belong to both E and F

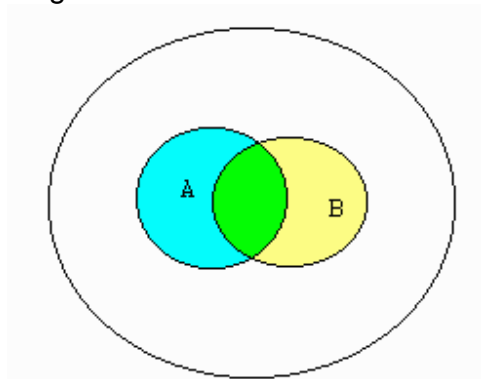
We use the terminology

E Or F

to mean all outcomes that belong to either E or F.

Example

Below is an example of two sets, A and B, graphed in a Venn diagram.



The green area represents A and B while all areas with color represent A or B

Example

Our Women's Volleyball team is recruiting for new members. Suppose that a person inquires about the team.

Let E be the event that the person is female
 Let F be the event that the person is a student
 then E And F represents the qualifications for being a member of the team. Note that E Or F is not enough.
 We define

Definition of Conditional Probability

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

We read the left hand side as
 "The probability of event E *given* event F"
 We call two events *independent* if

For Independent Events

$$P(E|F) = P(E)$$

Equivalently, we can say that E and F are independent if

For Independent Events

$$P(E \text{ and } F) = P(E)P(F)$$

Example

Consider rolling two dice. Let

E be the event that the first die is a 3.

F be the event that the sum of the dice is an 8.

Then E and F means that we rolled a three and then we rolled a 5

This probability is 1/36 since there are 36 possible pairs and only one of them is (3,5)

We have

$$P(E) = 1/6$$

And note that (2,6),(3,5),(4,4),(5,3), and (6,2) give F

Hence

$$P(F) = 5/36$$

We have

$$P(E) P(F) = (1/6) (5/36)$$

which is not 1/36.

We can conclude that E and F are not independent.

Exercise

Test the following two events for independence:

E the event that the first die is a 1.

F the event that the sum is a 7.

A Counting Rule

For two events, E and F, we always have

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

Example

Find the probability of selecting either a heart or a face card from a 52 card deck.

Solution

We let

E = the event that a heart is selected

F = the event that a face card is selected

then

$P(E) = 1/4$ and $P(F) = 3/13$ (Jack, Queen, or King out of 13 choices)

$$P(E \text{ and } F) = 3/52$$

The formula gives

$$P(E \text{ or } F) = 1/4 + 3/13 - 3/52 = 22/52 = 42\%$$

Trees and Counting**Using Trees**

We have seen that probability is defined by

$$P(E) = \frac{\text{Number in } E}{\text{Number in the Sample Space}}$$

Although this formula appears simple, counting the number in each can prove to be a challenge. Visual aids will help us immensely.

Example

A native flowering plant has several varieties. The color of the flower can be red, yellow, or white. The stems can be long or short and the leaves can be thorny, smooth, or velvety. Show all varieties.

Solution

We use a *tree diagram*. A tree diagram is a diagram that branches out and ends in leaves that correspond to the final variety. The picture below shows this.

Outcome

An outcome is the result of an experiment or other situation involving uncertainty.

The set of all possible outcomes of a probability experiment is called a sample space.

Sample Space

The sample space is an exhaustive list of all the possible outcomes of an experiment. Each possible result of such a study is represented by one and only one point in the sample space, which is usually denoted by S .

Examples

Experiment Rolling a die once:

Sample space $S = \{1,2,3,4,5,6\}$

Experiment Tossing a coin:

Sample space $S = \{\text{Heads}, \text{Tails}\}$

Experiment Measuring the height (cms) of a girl on her first day at school:

Sample space S = the set of all possible real numbers

Event

An event is any collection of outcomes of an experiment.

Formally, any subset of the sample space is an event.

Any event which consists of a single outcome in the sample space is called an elementary or simple event. Events which consist of more than one outcome are called compound events.

Set theory is used to represent relationships among events. In general, if A and B are two events in the sample space S , then

$A \cup B$ (A union B) = 'either A or B occurs or both occur'

$A \cap B$ (A intersection B) = 'both A and B occur'

$A \subseteq B$ (A is a subset of B) = 'if A occurs, so does B '

A or \bar{A} = 'event A does not occur'

ϕ (the empty set) = an impossible event

S (the sample space) = an event that is certain to occur

Example

Experiment: rolling a dice once -

Sample space $S = \{1,2,3,4,5,6\}$

Events $A = \text{'score'} < 4 = \{1,2,3\}$

$B = \text{'score is even'} = \{2,4,6\}$

$C = \text{'score is 7'} = \phi$

$A \cup B = \text{'the score is } < 4 \text{ or even or both'} = \{1,2,3,4,6\}$

$A \cap B = \text{'the score is } < 4 \text{ and even'} = \{2\}$

A or \bar{A} = 'event A does not occur' = $\{4,5,6\}$

Relative Frequency

Relative frequency is another term for proportion; it is the value calculated by dividing the number of times an event occurs by the total number of times an experiment is carried out. The probability of an event can be thought of as its long-run relative frequency when the experiment is carried out many times.

If an experiment is repeated n times, and event E occurs r times, then the relative frequency of the event E is defined to be

$$r/n(E) = r/n$$

Example

Experiment: Tossing a fair coin 50 times ($n = 50$)

Event $E = \text{'heads'}$

Result: 30 heads, 20 tails, so $r = 30$

Relative frequency: $\text{rfn}(E) = r/n = 30/50 = 3/5 = 0.6$

If an experiment is repeated many, many times without changing the experimental conditions, the relative frequency of any particular event will settle down to some value. The probability of the event can be defined as the limiting value of the relative frequency:

$$P(E) = \lim_{n \rightarrow \infty} \text{rfn}(E)$$

For example, in the above experiment, the relative frequency of the event 'heads' will settle down to a value of approximately 0.5 if the experiment is repeated many more times.

Probability

A probability provides a quantitative description of the likely occurrence of a particular event. Probability is conventionally expressed on a scale from 0 to 1; a rare event has a probability close to 0, a very common event has a probability close to 1.

The probability of an event has been defined as its long-run relative frequency. It has also been thought of as a personal degree of belief that a particular event will occur (subjective probability).

In some experiments, all outcomes are equally likely. For example if you were to choose one winner in a raffle from a hat, all raffle ticket holders are equally likely to win, that is, they have the same probability of their ticket being chosen. This is the equally-likely outcomes model and is defined to be:

$$P(E) = \frac{\text{number of outcomes corresponding to event } E}{\text{total number of outcomes}}$$

Examples

1. The probability of drawing a spade from a pack of 52 well-shuffled playing cards is $13/52 = 1/4 = 0.25$ since event $E = \text{'a spade is drawn'}$; the number of outcomes corresponding to $E = 13$ (spades); the total number of outcomes = 52 (cards).
2. When tossing a coin, we assume that the results 'heads' or 'tails' each have equal probabilities of 0.5.

Subjective Probability

A subjective probability describes an individual's personal judgement about how likely a particular event is to occur. It is not based on any precise computation but is often a reasonable assessment by a knowledgeable person.

Like all probabilities, a subjective probability is conventionally expressed on a scale from 0 to 1; a rare event has a subjective

probability close to 0, a very common event has a subjective probability close to 1.

A person's subjective probability of an event describes his/her degree of belief in the event.

Example

A Rangers supporter might say, "I believe that Rangers have probability of 0.9 of winning the Scottish Premier Division this year since they have been playing really well."

Independent Events

Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other.

In probability theory we say that two events, A and B, are independent if the probability that they both occur is equal to the product of the probabilities of the two individual events, i.e.

$$P(A \cap B) = P(A) \cdot P(B)$$

The idea of independence can be extended to more than two events. For example, A, B and C are independent if:

- a. A and B are independent; A and C are independent and B and C are independent (pairwise independence);
- b. $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

If two events are independent then they cannot be mutually exclusive (disjoint) and vice versa.

Example

Suppose that a man and a woman each have a pack of 52 playing cards. Each draws a card from his/her pack. Find the probability that they each draw the ace of clubs.

We define the events:

A = probability that man draws ace of clubs = $1/52$

B = probability that woman draws ace of clubs = $1/52$

Clearly events A and B are independent so:

$$P(A \cap B) = P(A) \cdot P(B) = 1/52 \cdot 1/52 = 0.00037$$

That is, there is a very small chance that the man and the woman will both draw the ace of clubs.

Conditional Probability

In many situations, once more information becomes available, we are able to revise our estimates for the probability of further outcomes or events happening. For example, suppose you go out for lunch at the same place and time every Friday and you are served lunch within 15 minutes with probability 0.9. However, given that you notice that the restaurant is exceptionally busy, the probability of being served lunch within 15 minutes may reduce to

0.7. This is the conditional probability of being served lunch within 15 minutes given that the restaurant is exceptionally busy.

The usual notation for "event A occurs given that event B has occurred" is " $A | B$ " (A given B). The symbol $|$ is a vertical line and does not imply division. $P(A | B)$ denotes the probability that event A will occur given that event B has occurred already.

A rule that can be used to determine a conditional probability from unconditional probabilities is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where:

$P(A | B)$ = the (conditional) probability that event A will occur given that event B has occurred already

$P(A \cap B)$ = the (unconditional) probability that event A and event B both occur

$P(B)$ = the (unconditional) probability that event B occurs

Example:

When a fair dice is tossed, the conditional probability of getting '1', given that an odd number has been obtained, is equal to $1/3$ as explained below:

$S = \{1,2,3,4,5,6\}$; $A = \{1,3,5\}$; $B = \{1\}$; $A \cap B = \{1\}$

$P(B/A) = 1/6 / 1/2 = 1/3$

Multiplication rule for dependent events:

The probability of simultaneous occurrence of two events A and B is equal to the product of the probability of one of the events by the conditional probability of the other, given that the first one has already occurred.

Example:

From a pack of cards, 2 cards are drawn in succession one after the other. After every draw, the selected card is not replaced. What is the probability that in both the draws you will get spades?

Solution:

Let A = getting spade in the first draw

Let B = getting spade in the second draw.

The cards are not replaced.

This situation requires the use of conditional probability.

$P(A) = 13/52$

$P(B/A) = 12/51$

Mutually Exclusive Events

Two events are mutually exclusive (or disjoint) if it is impossible for them to occur together.

Formally, two events A and B are mutually exclusive if and only if

$$A \cap B = \emptyset$$

If two events are mutually exclusive, they cannot be independent and vice versa.

Examples

1. Experiment: Rolling a die once
Sample space $S = \{1, 2, 3, 4, 5, 6\}$
Events $A = \text{'observe an odd number'} = \{1, 3, 5\}$
 $B = \text{'observe an even number'} = \{2, 4, 6\}$
 $A \cap B = \emptyset$ = the empty set, so A and B are mutually exclusive.
2. A subject in a study cannot be both male and female, nor can they be aged 20 and 30. A subject could however be both male and 20, or both female and 30.

Addition Rule

The addition rule is a result used to determine the probability that event A or event B occurs or both occur.

The result is often written as follows, using set notation:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where:

$P(A)$ = probability that event A occurs

$P(B)$ = probability that event B occurs

$P(A \cup B)$ = probability that event A or event B occurs

$P(A \cap B)$ = probability that event A and event B both occur

For mutually exclusive events, that is events which cannot occur together:

$$P(A \cap B) = 0$$

The addition rule therefore reduces to

$$P(A \cup B) = P(A) + P(B)$$

For independent events, that is events which have no influence on each other:

$$P(A \cap B) = P(A) \cdot P(B)$$

The addition rule therefore reduces to

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

Example

Suppose we wish to find the probability of drawing either a king or a spade in a single draw from a pack of 52 playing cards.

We define the events $A = \text{'draw a king'}$ and $B = \text{'draw a spade'}$

Since there are 4 kings in the pack and 13 spades, but 1 card is both a king and a spade, we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/52 + 13/52 - 1/52 = 16/52$$

So, the probability of drawing either a king or a spade is $16/52$ (= $4/13$).

See also multiplication rule.

Multiplication Rule

The multiplication rule is a result used to determine the probability that two events, A and B, both occur.

The multiplication rule follows from the definition of conditional probability.

The result is often written as follows, using set notation:

$$P(A \cap B) = P(A|B) \cdot P(B) \text{ or } P(A \cap B) = P(B|A) \cdot P(A)$$

where:

$P(A)$ = probability that event A occurs

$P(B)$ = probability that event B occurs

$P(A \cap B)$ = probability that event A and event B occur

$P(A | B)$ = the conditional probability that event A occurs given that event B has occurred already

$P(B | A)$ = the conditional probability that event B occurs given that event A has occurred already

For independent events, that is events which have no influence on one another, the rule simplifies to:

$$P(A \cap B) = P(A) \cdot P(B)$$

That is, the probability of the joint events A and B is equal to the product of the individual probabilities for the two events.

Multiplication rule for independent events:

Example:

The probability that you will get an A grade in Quantitative methods is 0.7. The probability that you will get an A grade in Marketing is 0.5. Assuming these two courses are independent, compute the probability that you will get an A grade in both these subjects.

Solution:

Let A = getting A grade in quantitative methods

Let B = getting A grade in Marketing

It is given that A and B are independent.

Applying the formula,

We get, $P(A \text{ and } B) = P(A) \cdot P(B) = .7 \cdot .5 = .35$

Conditional Probability

In many situations, once more information becomes available, we are able to revise our estimates for the probability of further outcomes or events happening. For example, suppose you go out for lunch at the same place and time every Friday and you are served lunch within 15 minutes with probability 0.9. However, given that you notice that the restaurant is exceptionally busy, the probability of being served lunch within 15 minutes may reduce to 0.7. This is the conditional probability of being served lunch within 15 minutes given that the restaurant is exceptionally busy.

The usual notation for "event A occurs given that event B has occurred" is " $A | B$ " (A given B). The symbol $|$ is a vertical line and does not imply division. $P(A | B)$ denotes the probability that event A will occur given that event B has occurred already.

A rule that can be used to determine a conditional probability from unconditional probabilities is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where:

$P(A | B)$ = the (conditional) probability that event A will occur given that event B has occurred already

$P(A \cap B)$ = the (unconditional) probability that event A and event B both occur

$P(B)$ = the (unconditional) probability that event B occurs

Law of Total Probability

The result is often written as follows, using set notation:

$$P(A) = P(A \cap B) + P(A \cap B')$$

where:

$P(A)$ = probability that event A occurs

$P(A \cap B)$ = probability that event A and event B both occur

$P(A \cap B')$ = probability that event A and event B' both occur, i.e. A occurs and B does not.

Using the multiplication rule, this can be expressed as

$$P(A) = P(A | B).P(B) + P(A | B').P(B')$$

Bayes' Theorem

Bayes' Theorem is a result that allows new information to be used to update the conditional probability of an event.

Using the multiplication rule, gives Bayes' Theorem in its simplest form:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A).P(A)}{P(B)}$$

Using the Law of Total Probability:

$$P(A | B) = \frac{P(B | A).P(A)}{P(B | A).P(A) + P(B | A').P(A')}$$

where:

$P(A)$ = probability that event A occurs

$P(B)$ = probability that event B occurs

$P(A')$ = probability that event A does not occur

$P(A | B)$ = probability that event A occurs given that event B has occurred already

$P(B | A)$ = probability that event B occurs given that event A has occurred already

$P(B | A')$ = probability that event B occurs given that event A has not occurred already

Example:

A manufacturing firm is engaged in the production of steel pipes in its three plants with a daily production of 1000, 1500 and 2500 units respectively. According to the past experience, it is known that the fractions of defective pipes produced by the three plants are respectively 0.04, 0.09 and

0.07.If a pipe is selected from a day's total production and found to be defective, find out a) What is the probability of the defective pipes) What is the probability that it has come from the second plant?

Solution:

Let the probabilities of the possible events be

Probability that a pipe is manufactured in plant A = $P(E1)=1000/(1000+1500+2500)=0.2$

Probability that a pipe is manufactured in plant B= $P(E2)=1500/(1000+1500+2500)0.3$

Probability that a pipe is manufactured in plant C = $P(E3)=2500/(1000+1500+2500)=0.5$

Let $P(D)$ be the probability that a defective pipe is drawn. Given that the proportions of the defective pipes coming from the three plants are 0.04,0.09 and 0.07 respectively, these are, in fact, the conditional probabilities ($D/E1$)=0.04; $P(D/E2)$ =0.09 AND $P(D/E3)$ =0.07

Now we can multiply prior probabilities and conditional probabilities in order to obtain the joint probabilities.

Joint probabilities are

Plant A = $.04*.2 =.008$

Plant B = $.09*.3 =.027$

Plant C = $.07*.5 = .035$

Now we can obtain posterior probabilities by the following calculations:

Plant A = $P(E1/D) = .008/0.008+0.027+0.035 = .114$

Plant B = $P(E2/D)= 0.027/0.008+0.027+0.035 =.386$

Plant C = $P(E3/D) = 0.035/0.008+0.027+0.035 =.500$

Computation of posterior probabilities

Event	prior $P(Ei)$	Conditional $P(D/Ei)$	Joint probability	Posterior $P(Ei/D)$
E1	0.2	0.04	$0.04 \times 0.2 = 0.008$	$0.08/0.07 = .11$
E2	0.3	0.09	$0.09 \times 0.3 = 0.027$	$0.027/0.07 = .39$
E3	0.5	0.07	$0.07 \times 0.5 = 0.035$	$0.035/0.07 = .50$
TOTAL	1.0		$P(D)=0.07$	1.00

On the basis of these calculations we can say that a)most probably the defective pipe has come from plant c

b)the probability that the defective pipe has come from the second plant is 0.39

Prior probability vs. posterior probability

We have seen in the foregone table that as any additional information becomes available, it can be used to revise the prior probability. The revised probability is called the posterior probability. Management should know how to use the additional information; it should also assess the utility or worth of the additional information. It may, at times find that

the cost of obtaining the additional information is more than its actual worth. In such cases, obviously it is not advisable to go in for any additional information and management should be satisfied with the prior probabilities.

Example 1:

The Monty Hall problem

We are presented with three doors - red, green, and blue - one of which has a prize. We choose the red door, which is not opened until the presenter performs an action. The presenter *who knows what door the prize is behind, and who must open a door, but is not permitted to open the door we have picked or the door with the prize*, opens the *green* door and reveals that there is no prize behind it and subsequently asks if we wish to change our mind about our initial selection of red. What is the probability that the prize is behind the blue and red doors?

Let us call the situation that the prize is behind a given door A_r , A_g , and A_b .

To start with, , and to make things simpler we shall assume that we have already picked the red door.

Let us call B "the presenter opens the green door". Without any prior knowledge, we would assign this a probability of 50%

- In the situation where the prize is behind the red door, the host is free to pick between the green or the blue door at random. Thus, $P(B | A_r) = 1 / 2$
- In the situation where the prize is behind the green door, the host must pick the blue door. Thus, $P(B | A_g) = 0$
- In the situation where the prize is behind the blue door, the host must pick the green door. Thus, $P(B | A_b) = 1$

Thus,

Note how this depends on the value of $P(B)$.

SOLVE:

1. In a software test environment holding software developed on j2ee specification a down time analysis was done. Based on the 100 earlier records it was found that there is about 5% downtime per day. A study on the components involved in the environment shows that a problem in web sphere cause errors out of which 25% led to downtime. If there are issues in the operating system, 40% of the issues lead to a down time and again 20% of the problems in network led to a downtime. Given that there is a downtime find the probability that each of the above reason could have contributed the downtime between themselves (considering just these 3 reasons)

Solutions:

Let the occurrence of a downtime be D

$P(D) = 5\% = .05$

Let the occurrence of a web sphere error be W

Probability of web sphere error causing downtime $P(D/W) = .25$

Let the occurrence of a OPERATING SYSTEM ERROR BE o

Probability of OS error causing downtime $P(D/O) = .4$

Let the occurrence of network error be N

Probability of N Causing downtime $P(D/M) = .2$

$P(W/D) = ?$

$P(O/D) = ?$

$P(N/D) = ?$

2. In a bolt factory machines A1, A2, A3 manufactures respectively 25%, 35%, 40% of the output of these 5, 4, 2 percent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by machine A2?

3. Suppose there is a chance for a newly constructed building to collapse whether the design is faulty or not. The chance that the design faulty is 10%. the chance that the building collapse is 95% if the design is faulty and otherwise it is 45%. it is seen that the building collapsed. What is the probability that it is due to faulty design?
issued by suburbs?

Summary:

This unit provides a conceptual framework on probability concepts with examples. Specifically this unit is focused on:

- The meaning and definition of the term probability interwoven with other associated terms-event, experiment and sample space.
- The three types of probability –classical probability, statistical probability and subjective probability.
- The concept of mutually exclusive events and independent events.
- The rules for calculating probability which include addition rule for mutually exclusive events and non mutually exclusive events, multiplication rule for independent events and dependent events and conditional probability.
- The application of Bayes's theorem in management.

Have you understood?

1. An urn contains 75 marbles: 35 are blue and 25 of these blue marbles are swirled. Rest of them are red and 30 of red ones are swirled (not swirled are clear ones). What is the probability of drawing?
 - a) blue marble
 - b) clear marble
 - c) blue swirled
 - d) red clear

- e) swirled marble
 - f)
2. Two persons X and Y appear in an interview for two vacancies in the same post, the probability of X's selection is $\frac{1}{5}$ and that of Y's selection is $\frac{1}{3}$. What is the probability that: 1) both X and Y will be selected? 2) Only one of them will be selected? 3) none of them will be selected?
 3. A company is to appoint a person as its managing director who must be an M.Tech and M.B.A and C.A. The probability of which are one in twenty five, one in forty and one in fifty respectively. Find the probability of getting such a person to be appointed by the company.
 4. A sample of 500 respondents was selected in a large metropolitan area to determine various information concerning consumer behaviour. Among the questions asked was, "Do you enjoy shopping for clothing?" of 240 males, 136 answered yes. Of 260 females, 224 answered yes.
 - a) Set up a 2X2 table to evaluate the probabilities
 - b) Give an example of a simple event
 - c) Give an example of a joint event
 - d) What is the complement of "Enjoy shopping for clothing"? What is the probability that a respondent chosen at random
 - e) Is a male?
 - f) enjoys shopping for clothing?
 - g) is a female and enjoys shopping for clothing?
 - h) is a male and does not enjoy shopping for clothing?
 - i) is a female or enjoys shopping for clothing?
 5. A municipal bond service has three rating categories (A, B and C). Suppose that in the past year, of the municipal bonds issued throughout the United States, 70% were rated A, 20% were rated B, and 10% were rated C. Of the municipal bonds rated A, 50% were issued by cities, 40% by suburbs, and 20% by rural areas. Of the municipal bonds rated C, 90% were issued by cities, 5% by suburbs, and 5% by rural areas.
 - a) If a new municipal bond is to be issued by a city, What is the probability it will receive A rating?
 - b) What proportion of municipal bonds is issued by cities?
 - c) What proportion of municipal bonds is issued by suburbs?
 6. An advertising executive is studying television viewing habits of married men and women during prime time hours. On the basis of past viewing records, the executive has determined that during prime time, husbands are watching television 60% of the time. It has also been determined that when the husband is watching television 40% of the time the wife is also watching. When the

husbands are watching television 30% of the wives are watching .Find the probability that

a)if the wife is watching television ,the husbands are also watching television.

b)the wives are watching television in prime time.

7)In the past several years, credit card companies have made an aggressive effort to solicit new accounts from college students. Suppose that a sample of 200 students at your college indicated the following information as to whether the student possessed a bank credit card and/or a travel and entertainment credit card.

8)A software company develops banking software where performance variable depends on the number of customer accounts.The list given below provides the statistics of 150 of the clients using the software along with the performance variable group.

Tot customer accounts no of clientsusing	performance variable group	
0-50	A	7
25-100	B	14
75-200	C	28
200-300	D	60
300-400	E	25
400-500	F	16

a)find the probability that a client has <200 user accounts.

b)find the probability that the performance variable B is used.

c)A client chosen will fit in both variable A and B,if the clients who can use A is 9 and the clients who can use B is 19.

9)there are two men aged 30 and 36 years.The probability to live 35 years more is .67 for the 30 year old [person and .60 for the 36 year old person.Find the probability that atleast one of these persons will be alive 35 years.

10)the probability that a contractor will get a plumbing contract is $\frac{2}{3}$.and the probability that he will not an electric contract is $\frac{5}{9}$.If the probability of getting at least one contract is $\frac{4}{5}$.what is the probability that he will get from both the contracts?



2

PROBABILITY DISTRIBUTION

INTRODUCTION

In unit1, we encountered some experiments where the outcomes were categorical. We found that an experiment results in a number of possible outcomes and discussed how the probability of the occurrence of an outcome can be determined. In this unit, we shall extend our discussion of probability theory. Our focus is on the probability distribution, which describes how probability is spread over the possible numerical values associated with the outcomes.

LEARNING OBJECTIVES

After reading this unit, you will be able to

- Define random variables
- Appreciate what is probability distribution
- Explain and use the binomial distribution
- Explain and use the poison distribution
- Explain and use the uniform distribution
- Explain and use the normal distribution

RANDOM VARIABLE

A **random variable** is an abstraction of the intuitive concept of chance into the theoretical domains of mathematics, forming the foundations of probability theory and mathematical statistics.

The theory and language of random variables were formalized over the last few centuries alongside ideas of probability. Full familiarity with all the properties of random variables requires a strong background in the more recently developed concepts of measure theory, but random variables can be understood intuitively at various levels of mathematical fluency; set theory and calculus are fundamentals.

Broadly, a random variable is defined as a quantity whose values are random and to which a probability distribution is assigned. More

formally, a random variable is a measurable function from a sample space to the measurable space of possible values of the variable. The formal definition of random variables places experiments involving real-valued outcomes firmly within the measure-theoretic framework and allows us to construct distribution functions of real-valued random variables.

WHAT IS A RANDOM VARIABLE?

A *Random Variable* is a function, which assigns unique numerical values to all possible outcomes of a random experiment under fixed conditions (Ali 2000). A random variable is not a variable but rather a function that maps events to numbers (Wikipedia 2006).

Example 1

This example is extracted from (Ali 2000). Suppose that a coin is tossed three times and the sequence of heads and tails is noted. The sample space for this experiment evaluates to: $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Now let the random variable X be the number of heads in three coin tosses. X assigns each outcome in S a number from the set $S_x = \{0, 1, 2, 3\}$. The table below lists the eight outcomes of S and the corresponding values of X .

Outcome	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
X	3	2	2	2	1	1	1	0

X is then a random variable taking on values in the set $S_X = \{0, 1, 2, 3\}$.

Mathematically, a random variable is defined as a measurable function from a probability space to some measurable space (Wikipedia 2006). This measurable space is the space of possible values of the variable, and it is usually taken to be the real numbers (Wikipedia 2006).

The condition for a function to be a random variable is that the random variable cannot be multivalued (Ali 2000).

There are three types of random variables:

- A *Continuous Random Variable* is one that takes an infinite number of possible values (Ali 2000). Example: Duration of a call in a telephone exchange.
- A *Discrete Random Variable* is one that takes a finite distinct values (Ali 2000). Example: A number of students who fail a test.
- A *Mixed Random Variable* is one for which some of its values are continuous and some are discrete (Ali 2000).

Problem 1

Can measurements of power (in dB) received from an antenna be considered a random variable?

Solution 1

Yes. Specifically it should be considered as a continuous random variable as the power of any signal attenuates through a transmission line. The attenuation factors associated with each transmission line are only approximate. Thus the power received from the antenna can take any value.

In contrast, if a random variable represents a measurement on a continuous scale so that all values in an interval are possible, it is called a continuous random variable. In other words, a continuous random variable is a random variable, which can take any value within some interval of real number. Examples of a continuous random variable are price of a car and daily consumption of milk. Measurement of the height and weight of the respondents is an example of a continuous random variable. Similarly, voltage, pressure, and temperature are examples of continuous random variable.

Measures of location

Location

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

Definition of location

The first step is to define what we mean by a typical value. For univariate data, there are three common definitions:

1. mean - the mean is the sum of the data points divided by the number of data points. That is,

$$\bar{Y} = \sum_{i=1}^N Y_i / N$$

The mean is that value that is most commonly referred to as the average. We will use the term average as a synonym for the mean and the term typical value to refer generically to measures of location.

2. median - the median is the value of the point which has half the data smaller than that point and half the data larger than that point. That is, if X_1, X_2, \dots, X_N is a random sample sorted

from smallest value to largest value, then the median is defined as:

$$\tilde{Y} = Y_{(N+1)/2} \quad \text{if } N \text{ is odd}$$

$$\tilde{Y} = (Y_{N/2} + Y_{(N/2)+1})/2 \quad \text{if } N \text{ is even}$$

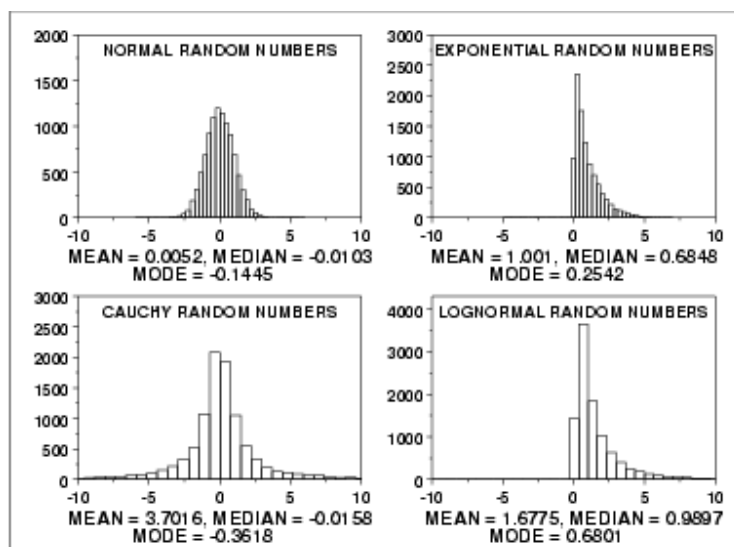
3. mode - the mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique. The mode is typically used in a qualitative fashion. For example, there may be a single dominant hump in the data perhaps two or more smaller humps in the data. This is usually evident from a histogram of the data.

When taking samples from continuous populations, we need to be somewhat careful in how we define the mode. That is, any specific value may not occur more than once if the data are continuous. What may be a more meaningful, if less exact measure, is the midpoint of the class interval of the histogram with the highest peak.

Why different measures?

A natural question is why we have more than one measure of the typical value. The following example helps to explain why these alternative definitions are useful and necessary.

This plot shows histograms for 10,000 random numbers generated from a normal, an exponential, a Cauchy, and a lognormal distribution.



Normal distribution

The first histogram is a sample from a normal distribution. The mean is 0.005, the median is -0.010, and the mode is -0.144 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The normal distribution is a symmetric distribution with well-behaved tails and a single peak at the center of the distribution. By symmetric, we mean that the distribution can be folded about an axis so that the 2 sides coincide. That is, it behaves the same to the left and right of some center point. For a normal distribution, the mean, median, and mode are actually equivalent. The histogram above generates similar estimates for the mean, median, and mode. Therefore, if a histogram or normal probability plot indicates that your data are approximated well by a normal distribution, then it is reasonable to use the mean as the location estimator.

Exponential distribution

The second histogram is a sample from an exponential distribution. The mean is 1.001, the median is 0.684, and the mode is 0.254 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The exponential distribution is a skewed, i. e., not symmetric, distribution. For skewed distributions, the mean and median are not the same. The mean will be pulled in the direction of the skewness. That is, if the right tail is heavier than the left tail, the mean will be greater than the median. Likewise, if the left tail is heavier than the right tail, the mean will be less than the median.

For skewed distributions, it is not at all obvious whether the mean, the median, or the mode is the more meaningful measure of the typical value. In this case, all three measures are useful.

Cauchy distribution

The third histogram is a sample from a Cauchy distribution. The mean is 3.70, the median is -0.016, and the mode is -0.362 (the mode is computed as the midpoint of the histogram interval with the highest peak).

For better visual comparison with the other data sets, we restricted the histogram of the Cauchy distribution to values between -10 and 10. The full Cauchy data set in fact has a minimum of approximately -29,000 and a maximum of approximately 89,000.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. The Cauchy distribution has the interesting property that collecting more data does not provide a more accurate estimate of the mean. That is, the sampling distribution of the mean is equivalent to the sampling

distribution of the original data. This means that for the Cauchy distribution the mean is useless as a measure of the typical value. For this histogram, the mean of 3.7 is well above the vast majority of the data. This is caused by a few very extreme values in the tail. However, the median does provide a useful measure for the typical value.

Although the Cauchy distribution is an extreme case, it does illustrate the importance of heavy tails in measuring the mean. Extreme values in the tails distort the mean. However, these extreme values do not distort the median since the median is based on ranks. In general, for data with extreme values in the tails, the median provides a better estimate of location than does the mean.

Lognormal distribution

The fourth histogram is a sample from a lognormal distribution. The mean is 1.677, the median is 0.989, and the mode is 0.680 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The lognormal is also a skewed distribution. Therefore the mean and median do not provide similar estimates for the location. As with the exponential distribution, there is no obvious answer to the question of which is the more meaningful measure of location.

Robustness

There are various alternatives to the mean and median for measuring location. These alternatives were developed to address non-normal data since the mean is an optimal estimator if in fact your data are normal.

Tukey and Mosteller defined two types of robustness where robustness is a lack of susceptibility to the effects of nonnormality.

1. Robustness of validity means that the confidence intervals for the population location have a 95% chance of covering the population location regardless of what the underlying distribution is.
2. Robustness of efficiency refers to high effectiveness in the face of non-normal tails. That is, confidence intervals for the population location tend to be almost as narrow as the best that could be done if we knew the true shape of the distribution.

The mean is an example of an estimator that is the best we can do if the underlying distribution is normal. However, it lacks robustness of validity. That is, confidence intervals based on the mean tend not to be precise if the underlying distribution is in fact not normal.

The median is an example of an estimator that tends to have robustness of validity but not robustness of efficiency.

The alternative measures of location try to balance these two concepts of robustness. That is, the confidence intervals for the case when the data are normal should be almost as narrow as the confidence intervals based on the mean. However, they should maintain their validity even if the underlying data are not normal. In particular, these alternatives address the problem of heavy-tailed distributions.

Alternative measures of location

A few of the more common alternative location measures are:

1. Mid-Mean - computes a mean using the data between the 25th and 75th percentiles.
2. Trimmed Mean - similar to the mid-mean except different percentile values are used. A common choice is to trim 5% of the points in both the lower and upper tails, i.e., calculate the mean for data between the 5th and 95th percentiles.
3. Winsorized Mean - similar to the trimmed mean. However, instead of trimming the points, they are set to the lowest (or highest) value. For example, all data below the 5th percentile are set equal to the value of the 5th percentile and all data greater than the 95th percentile are set equal to the 95th percentile.
4. Mid-range = $(\text{smallest} + \text{largest})/2$.

The first three alternative location estimators defined above have the advantage of the median in the sense that they are not unduly affected by extremes in the tails. However, they generate estimates that are closer to the mean for data that are normal (or nearly so).

The mid-range, since it is based on the two most extreme points, is not robust. Its use is typically restricted to situations in which the behavior at the extreme points is relevant.

Measures of Skew ness and Kurtosis

Skew ness and Kurtosis

A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

Definition of skewness

For univariate data Y_1, Y_2, \dots, Y_N , the formula for skewness is:

$$\text{skewness} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3}{(N-1)s^3}$$

where \bar{Y} is the mean, s is the standard deviation, and N is the number of data points. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

Definition of kurtosis

For univariate data Y_1, Y_2, \dots, Y_N , the formula for kurtosis is:

$$\text{kurtosis} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N-1)s^4}$$

where \bar{Y} is the mean, s is the standard deviation, and N is the number of data points.

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis:

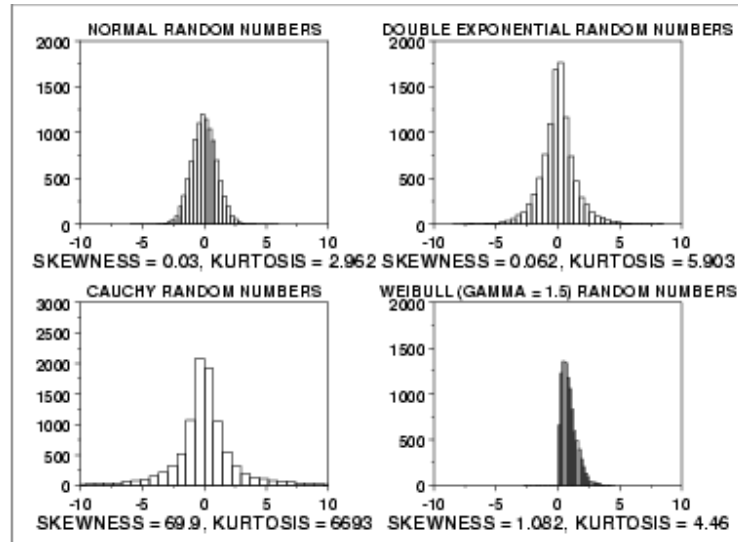
$$\text{kurtosis} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N-1)s^4} - 3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

Which definition of kurtosis is used is a matter of convention. When using software to compute the sample kurtosis, you need to be aware of which convention is being followed.

Examples

The following example shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Weibull distribution.



Probability distribution:

A probability distribution is a total listing of the various values. The random variable can come along with the corresponding probability for each value. A real life example would be the pattern of distribution of the machine breakdowns in a manufacturing unit. The random variable in this example would be the various values the machine breakdown could assume. The probability corresponding to each value of the breakdown is the relative frequency of occurrence of the breakdown. The probability distribution for this example is constructed by the actual breakdown pattern observed over a period of time.

1. A multinational bank is concerned about the waiting time of its customers for using their ATMs. A study of a random sample of 500 customers reveals the following probability distribution:

x(waiting time /customer in minutes):		0	1	2	3	4
	5	6	7	8		
p(x):		.09	.08	.04	.03	.20
						.18
						.16
						.12
						.10

- What is the probability that a customer will have to wait for more than 5 minutes?
- What is the probability that a customer need not wait?
- What is the probability that a customer will have to wait for less than 4 minutes?

Solution:

$$a) p(x > 5) = p(6) + p(8) = .08 + .04 + .03 = .15$$

$$b) p(x=0) = .20$$

$$c) p(x < 4) = p(0) + p(1) + p(2) + p(3) = .20 + .18 + .16 + .12 = .66$$

Types of probability distribution:

There are two types of probability distribution: They are

1. Discrete probability distribution
2. Continuous probability distribution

Discrete probability distribution

We have taken the above examples to explain the concept of a

1. Discrete Distributions

Discrete Densities

Suppose that we have a random experiment with sample space R , and probability measure P . A random variable X for the experiment that takes values in a countable set S is said to have a **discrete distribution**. The (discrete) **probability density function** of X is the function f from S to \mathbf{R} defined by

$$f(x) = P(X = x) \text{ for } x \text{ in } S.$$

 1. Show that f satisfies the following properties:

- a. $f(x) \geq 0$ for x in S .
- b. $\sum_{x \text{ in } S} f(x) = 1$
- c. $\sum_{x \text{ in } A} f(x) = P(X \in A)$ for $A \subseteq S$.

Property (c) is particularly important since it shows that the probability distribution of a discrete random variable is completely determined by its density function. Conversely, any function that satisfies properties (a) and (b) is a (discrete) density, and then property (c) can be used to construct a discrete probability distribution on S . Technically, f is the density of X relative to counting measure on S .

Typically, S is a countable subset of some larger set, such as \mathbf{R}^n for some n . We can always extend f , if we want, to the larger set by defining $f(x) = 0$ for x not in S . Sometimes this extension simplifies formulas and notation.

An element x in S that maximizes the density f is called a **mode** of the distribution. When there is only one mode, it is sometimes used as a measure of the *center* of the distribution.

Interpretation


A discrete probability distribution is equivalent to a **discrete mass distribution**, with total mass 1. In this analogy, S is the (countable) set of point masses, and $f(\mathbf{x})$ is the mass of the point at \mathbf{x} in S . Property (c) in Exercise 1 simply means that the mass of a set A can be found by adding the masses of the points in A .

For a probabilistic interpretation, suppose that we create a new, compound experiment by repeating the original experiment indefinitely. In the compound experiment, we have independent random variables X_1, X_2, \dots , each with the same distribution as X (these are "independent copies" of X). For each \mathbf{x} in S , let


$$f_n(\mathbf{x}) = \#\{i \in \{1, 2, \dots, n\} : X_i = \mathbf{x}\} / n,$$

the relative frequency of \mathbf{x} in the first n runs (the number of times that \mathbf{x} occurred, divided by n). Note that for each \mathbf{x} , $f_n(\mathbf{x})$ is a random variable for the compound experiment. By the law of large numbers, $f_n(\mathbf{x})$ should converge to $f(\mathbf{x})$ as n increases. The function f_n is called the empirical density function; these functions are displayed in most of the simulation applets that deal with discrete variables.

Examples

 2. Suppose that two fair dice are tossed and the sequence of scores (X_1, X_2) recorded. Find the density function of

- (X_1, X_2)
- $Y = X_1 + X_2$, the sum of the scores
- $U = \min\{X_1, X_2\}$, the minimum score
- $V = \max\{X_1, X_2\}$, the maximum score
- (U, V)

 3. In the dice experiment, select $n = 2$ fair dice. Select the following random variables and note the shape and location of the density function. Run the experiment 1000 times, updating every 10 runs. For each variables, note the apparent convergence of the empirical density function to the density function.


- Sum of the scores.
- Minimum score.
- Maximum score.

 4. An element X is chosen at random from a finite set S .


- Show that X has probability density function $f(x) = 1 / \#(S)$ for x in S .

- b. Show that $P(X \in A) = \#(A) / \#(S)$ for $A \subseteq S$.


The distribution in the last exercise is called the **discrete uniform distribution** on S . Many random variables that arise in sampling or combinatorial experiments are transformations of uniformly distributed variables.

 5. Suppose that n elements are chosen at random, without replacement from a set D with N elements. Let \mathbf{X} denote the ordered sequence of elements chosen. Argue that \mathbf{X} is uniformly distributed on the set S of permutations of size n chosen from D :

$$P(\mathbf{X} = \mathbf{x}) = 1 / (N)_n \text{ for each } \mathbf{x} \text{ in } S.$$


 6. Suppose that n elements are chosen at random, without replacement, from a set D with N elements. Let \mathbf{W} denote the unordered set of elements chosen. Show that \mathbf{W} is uniformly distributed on the set T of combinations of size n chosen from D :

$$P(\mathbf{W} = \mathbf{w}) = 1 / C(N, n) \text{ for } \mathbf{w} \text{ in } T.$$


 7. An urn contains N balls; R are red and $N - R$ are green. A sample of n balls is chosen at random (without replacement). Let Y denote the number of red balls in the sample. Show that Y has probability density function.


$$P(Y = k) = C(R, k) C(N - R, n - k) / C(N, n) \text{ for } k = 0, 1, \dots, n.$$

The distribution defined by the density function in the last exercise is the **hypergeometric distribution** with parameters N , R , and n . The hypergeometric distribution is studied in detail in the chapter on Finite Sampling Models, which contains a rich variety of distributions that are based on discrete uniform distributions.


 8. An urn contains 30 red and 20 green balls. A sample of 5 balls is selected at random. Let Y denote the number of red balls in the sample.

- Compute the density function of Y explicitly.
- Graph the density function and identify the mode(s).
- Find $P(Y > 3)$.

 9. In the ball and urn experiment, select sampling without replacement. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the empirical density function of Y to the theoretical density function.


 10. A coin with probability of heads p is tossed n times. For $j = 1, \dots, n$, let $I_j = 1$ if the toss j is heads and $I_j = 0$ if toss j is tails. Show that (I_1, I_2, \dots, I_n) has probability density function

$$f(i_1, i_2, \dots, i_n) = p^k (1 - p)^{n-k} \text{ for } i_j \in \{0, 1\} \text{ for each } j, \text{ where } k = i_1 + i_2 + \dots + i_n.$$


 11. A coin with probability of heads p is tossed n times. Let X denote the number of heads. Show that X has probability density function


$$P(X = k) = C(n, k) p^k (1 - p)^{n-k} \text{ for } k = 0, 1, \dots, n.$$

The distribution defined by the density in the previous exercise is called the **binomial distribution** with parameters n and p . The binomial distribution is studied in detail in the chapter on Bernoulli Trials.

 12. Suppose that a coin with probability of heads $p = 0.4$ is tossed 5 times. Let X denote the number of heads.


- Compute the density function of X explicitly.
- Graph the density function and identify the mode.
- Find $P(X > 3)$.

 13. In the coin experiment, set $n = 5$ and $p = 0.4$. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the empirical density function of X to the density function.

 14. Let $f_t(n) = \exp(-t) t^n / n!$ for $n = 0, 1, 2, \dots$ where $t > 0$ is a parameter.


- Show that f_t is a probability density function for each $t > 0$.
- Show that $f_t(n) > f_t(n - 1)$ if and only if $n < t$.
- Show that the mode occurs at $\text{floor}(t)$ if t is not an integer, and at $t - 1$ and t if t is an integer.


The distribution defined by the density in the previous exercise is the **Poisson distribution** with parameter t , named after Simeon Poisson. The Poisson distribution is studied in detail in the Chapter on Poisson Processes, and is used to model the number of "random points" in a region of time or space. The parameter t is proportional to the size of the region of time or space.

 15. Suppose that the number of misprints N on a web page has the Poisson distribution with parameter 2.5.


- Find the mode.

b. Find $P(N > 4)$.

 16. In the Poisson process, select parameter 2.5. Run the simulation 1000 times updating every 10 runs. Note the apparent convergence of the empirical density function to the true density function.

 17. In the die-coin experiment, a fair die is rolled and then a fair coin is tossed the number of times shown on the die. Let I denote the sequence of coin results (0 for tails, 1 for heads). Find the density of I (note that I takes values in a set of sequences of varying lengths).


Constructing Densities

 18. Suppose that g is a nonnegative function defined on a countable set S and that


$$c = \sum_{x \in S} g(x).$$

Show that if c is positive and finite, then $f(x) = g(x) / c$ for x in S defines a discrete density function on S .

The constant c in the last exercise is sometimes called the **normalizing constant**. This result is useful for constructing density functions with desired functional properties (domain, shape, symmetry, and so on).


 19. Let $g(x) = x^2$ for x in $\{-2, -1, 0, 1, 2\}$.

- Find the probability density function f that is proportional to g .
- Graph the density function and identify the modes.
- Find $P(X \in \{-1, 1, 2\})$ where X is a random variable with the density in (a)..


 20. Let $g(n) = q^n$ for $n = 0, 1, 2, \dots$ where q is a parameter in $(0, 1)$.

- Find the probability density function f that is proportional to g .
- Find $P(X < 2)$ where X is a random variable with the density in (a).
- Find the probability that X is even.

The distribution constructed in the last exercise is a version of the geometric distribution, and is studied in detail in the chapter on Bernoulli Trials.

 21. Let $g(x, y) = x + y$ for $(x, y) \in \{0, 1, 2\}^2$.

- a. Find the probability density function f that is proportional to g .
- b. Find the mode of the distribution.
- c. Find $P(X > Y)$ where (X, Y) is a random vector with the density in (a).

 22. Let $g(x, y) = xy$ for $(x, y) \in \{(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)\}$.


- a. Find the probability density function f that is proportional to g .
- b. Find the mode of the distribution.
- c. Find $P[(X, Y) \in \{(1, 2), (1, 3), (2, 2), (2, 3)\}]$ where (X, Y) is a random vector with the density in (a).

Conditional Densities


The density function of a random variable X is based, of course, on the underlying probability measure P on the sample space S for the experiment. This measure could be a conditional probability measure, conditioned on a given event E (with $P(E) > 0$). The usual notation is

$$f(x | E) = P(X = x | E) \text{ for } x \text{ in } S.$$


The following exercise shows that, except for notation, no new concepts are involved. Therefore, all results that hold for densities in general have analogues for conditional densities.


 23. Show that as a function of x for fixed E , $f(x | E)$ is a discrete density function. That is, show that it satisfies properties (a) and (b) of Exercise 1, and show that property (c) becomes


$$P(X \in A | E) = \sum_{x \in A} f(x | E) \text{ for } A \subseteq S.$$


 24. Suppose that $B \subseteq S$ and $P(X \in B) > 0$. Show that the conditional density of X given $X \in B$ is

- a. $f(x | X \in B) = f(x) / P(X \in B)$ for $x \in B$.
- b. $f(x | X \in B) = 0$ if $x \in B^c$.

 25. Suppose that X is uniformly distributed on a finite set S and that B is a nonempty subset of S . Show that the conditional distribution of X given $X \in B$ is uniform on B .

 26. Suppose that X has probability density function $f(x) = x^2 / 10$ for $x = -2, -1, 0, 1, 2$. Find the conditional density of X given that $X > 0$.

 27. A pair of fair dice are rolled. Let Y denote the sum of the scores and U the minimum score. Find the conditional density of U given $Y = 8$.

 28. Run the dice experiment 200 times, updating after every run. Compute the empirical conditional density of U given $Y = 8$ and compare with the conditional density in the last exercise.

Law of Total Probability and Bayes' Theorem

Suppose that \mathbf{X} is a discrete random variable taking values in a countable set S , and that B be an event in the experiment (that is, a subset of the underlying sample space R).

 29. Prove the **law of total probability**:


$$P(B) = \sum_{\mathbf{x} \in S} P(\mathbf{X} = \mathbf{x}) P(B \mid \mathbf{X} = \mathbf{x}).$$

This result is useful, naturally, when the distribution of \mathbf{X} and the conditional probability of B given the values of \mathbf{X} are known. We sometimes say that we are **conditioning** on \mathbf{X} .


 30. Prove **Bayes' Theorem**, named after Thomas Bayes:

$$P(\mathbf{X} = \mathbf{x} \mid B) = P(\mathbf{X} = \mathbf{x}) P(B \mid \mathbf{X} = \mathbf{x}) / \sum_{\mathbf{y} \in S} P(\mathbf{X} = \mathbf{y}) P(B \mid \mathbf{X} = \mathbf{y}) \text{ for } \mathbf{x} \text{ in } S.$$

Bayes' theorem is a formula for the conditional density of \mathbf{X} given B . As with the law of total probability, it is useful, when the quantities on the right are known. The (unconditional) distribution of \mathbf{X} is referred to as the **prior** distribution and the conditional density as the **posterior** density.


 31. In the die-coin experiment, a fair die is rolled and then a fair coin is tossed the number of times showing on the die.

- Find the probability that there will be exactly two heads.
- Given that there were 2 heads, find the conditional density of the die score.

 32. Run the die-coin experiment 200 times, updating after each run.


- Compute the empirical probability of exactly two heads and compare with the probability in the last exercise.


- b. Compute the empirical conditional density of the die score given exactly two heads and compare with the theoretical conditional density in the last exercise..


 33. Suppose that a bag contains 12 coins: 5 are fair, 4 are biased with probability of heads $1/3$; and 3 are two-headed. A coin is chosen at random from the bag and tossed twice.

- a. Find the probability that there will be exactly 2 heads.
- b. Given that there were 2 heads, find the conditional density of the type of coin

Compare Exercises 31 and 33. In Exercise 31, we toss a coin with a *fixed* probability of heads a *random* number of times. In Exercise 33, we effectively toss a coin with a *random* probability of heads a *fixed* number of times.


 34. In the coin-die experiment, a fair coin is tossed. If the coin lands tails, a fair die is rolled. If the coin lands heads, an ace-six flat die is tossed (1 and 6 have probability $1/4$ each, while 2, 3, 4, 5 have probability $1/8$ each). Find the density function of the die score.

 35. Run the coin-die experiment 1000 times, updating every 10 runs. Compare the empirical density of the die score with the theoretical density in the last exercise.

 36. A plant has 3 assembly lines that produces memory chips. Line 1 produces 50% of the chips and has a defective rate of 4%; line 2 has produces 30% of the chips and has a defective rate of 5%; line 3 produces 20% of the chips and has a defective rate of 1%. A chip is chosen at random from the plant.

- a. Find the probability that the chip is defective.
- b. Given that the chip is defective, find the conditional density of the line that produced the chip.

Data Analysis Exercises

 37. In the M&M data, let R denote the number of red candies and N the total number of candies. Compute and graph the empirical density of

- a. R
- b. N
- c. R given $N > 57$.

38. In the Cicada data, let G denotes gender, S denotes species type, and W denotes body weight (in grams). Compute the empirical density of

- G
- S
- (G, S)
- G given $W > 0.20$ grams.

Discrete probability distribution

Introduction:

In lecture number two, we said a **Random Variable** is a quantity resulting from a random experiment that, by chance, can assume different values. Such as, number of defective light bulbs produced during a week. Also, we said a **Discrete Random Variable** is a variable which can assume only integer values, such as, 7, 9, and so on. In other words, a discrete random variable cannot take fractions as value. Things such as people, cars, or defectives are things we can count and are discrete items. In this lecture note, we would like to discuss three types of **Discrete Probability Distribution**: *Binomial Distribution*, *Poisson Distribution*, and *Hypergeometric Distribution*.

Probability Distribution:

A probability distribution is similar to the frequency distribution of a quantitative population because both provide a long-run frequency for outcomes. In other words, a probability distribution is listing of all the possible values that a random variable can take along with their probabilities. for example, suppose we want to find out the probability distribution for the number of heads on three tosses of a coin:

First	toss.....	T	T	T	T	H	H	H	H
Second	toss.....	T	T	H	H	T	T	H	H
Third	toss.....	T	H	T	H	T	H	T	H

the probability distribution of the above experiment is as follows (columns 1, and 2 in the following table).

(Column 1).....	(Column 2).....	(Column 3)
Number of heads.....	Probability.....	(1)(2)

X.....	P(X).....	(X)P(X)
0.....	1/8.....	0.0
1.....	3/8.....	0.375
2.....	3/8.....	0.75

3.....	1/8.....	0.375		
Total.....	1.5	=	E(X)	

Mean, and Variance of Discrete Random Variables:

The equation for computing the **mean**, or **expected value** of discrete random variables is as follows:

Mean = $E(X)$ = Summation[$X \cdot P(X)$]
 where: $E(X)$ = expected value, X = an event, and $P(X)$ = probability of the event

Note that in the above equation, the probability of each event is used as the weight. For example, going back to the problem of tossing a coin three times, the expected value is: $E(X) = [0(1/8)+1(3/8)+2(3/8)+3(1/8)] = 1.5$ (column 3 in the above table). Thus, on the average, the number of heads showing face up in a large number of tossing a coin is 1.5. The expected value has many uses in gambling, for example, it tells us what our long-run average losses per play will be.

The equations for computing the **expected value**, **variance**, and **standard deviation** of discrete random variables are as follows:

Mean (Expected) Value of a Discrete Distribution

$$\mu = E(X) = \sum [X \cdot P(X)]$$

Variance of a Discrete Distribution

$$\sigma^2 = \sum [(X - \mu)^2 \cdot P(X)]$$

Standard Deviation of a Discrete Distribution

$$\sigma = \sqrt{\sum [(X - \mu)^2 \cdot P(X)]}$$

Example:

Suppose a charity organization is mailing printed return-address stickers to over one million homes in the U.S. Each recipient is asked to donate either \$1, \$2, \$5, \$10, \$15, or \$20. Based on past experience, the amount a person donates is believed to follow the following probability distribution:

X:.....	\$1.....	\$2.....	\$5.....	\$10.....	\$15.....	\$20
P(X).....	0.1.....	0.2.....	0.3.....	0.2.....	0.15.....	0.05

The question is, what is expected that an average donor to contribute, and what is the standard deviation. The solution is as follows.

(1).....	(2).....	(3).....	(4).....	(5).....	(6)
X.....	P(X).....	X.P(X).....	X - mean.....	[(X - mean)]squared.....	(5)x(2)
1.....	0.1.....	0.1.....	- 6.25.....	39.06..	
.....	3.906	
2.....	0.2.....	0.4.....	- 5.25.....	27.56	
.....	5.512	
5.....	0.3.....	1.5.....	- 2.25.....	5.06.	
.....	1.518	
10.....	0.2.....	2.0.....	2.75.....	7.56..	
.....	1.512	
15.....	0.15.....	2.25.....	7.75.....	60.06.....	
.....	9.009	
20.....	0.05.....	1.0.....	12.75.....	162.56.....	
.....	8.125	
Total.....	7.25 =				
E(X).....					29.585

Thus, the expected value is \$7.25, and standard deviation is the square root of \$29.585, which is equal to \$5.55. In other words, an average donor is expected to donate \$7.25 with a standard deviation of \$5.55.

Binomial Distribution:

One of the most widely known of all discrete probability distributions is the binomial distribution. Several characteristics underlie the use of the binomial distribution.

Characteristics of the Binomial Distribution:

1. The experiment consists of n identical trials.
2. Each trial has only one of the two possible mutually exclusive outcomes, success or a failure.
3. The probability of each outcome does not change from trial to trial, and
4. The trials are independent, thus we must sample with replacement.

Note that if the sample size, n , is less than 5% of the population, the independence assumption is not of great concern. Therefore

the acceptable sample size for using the binomial distribution with samples taken without replacement is $[n < 5\% N]$ where n is equal to the sample size, and N stands for the size of the population. The birth of children (male or female), true-false or multiple-choice questions (correct or incorrect answers) are some examples of the binomial distribution.

Binomial Equation:

When using the binomial formula to solve problems, all that is necessary is that we be able to identify three things: the number of trials (n), the probability of a success on any one trial (p), and the number of successes desired (X). The formulas used to compute the probability, the mean, and the standard deviation of a binomial distribution are as follows.

Binomial Formula

$$C_n^X \cdot p^X \cdot q^{n-X} = \frac{n!}{X!(n-X)!} \cdot p^X \cdot q^{n-X}$$

Mean of a Binomial Distribution

$$\mu = n \cdot p$$

Standard Deviation of a Binomial Distribution

$$\sigma = \sqrt{n \cdot p \cdot q}$$

where: n = the sample size or the number of trials, X = the number of successes desired, p = probability of getting a success in one trial, and $q = (1 - p)$ = the probability of getting a failure in one trial.

Example:

Let's go back to lecture number four and solve the probability problem of defective TVs by applying the binomial equation once again. We said, suppose that 4% of all TVs made by W&B Company in 1995 are defective. If eight of these TVs are randomly selected from across the country and tested, what is the probability that *exactly* three of them are defective? Assume that each TV is made independently of the others.

In this problem, $n=8$, $X=3$, $p=0.04$, and $q=(1-p)=0.96$. Plugging these numbers into the binomial formula (see the above equation) we get: $P(X) = P(3) = 0.0003$ or 0.03% which is the same answer as in lecture number four. The mean is equal to $(n) \times (p) = (8)(0.04)=0.32$, the variance is equal to $np(1 - p) = (0.32)(0.96) = 0.31$, and the standard deviation is the square root of 0.31, which is equal to 0.6.

The Binomial Table:Mathematicians constructed a set of binomial tables containing presolved probabilities. Binomial distributions are a family of distributions. In other words, every different value of n and/or every different value of p gives a different binomial distribution. Tables are available for different combinations of n and p values. For the tables, refer to the text. Each table is headed by a value of n , and values of p are presented in the top row of each table of size n . In the column below each value of p is the binomial distribution for that value of n and p . The binomial tables are easy to use. Simply look up n and p , then find X (located in the first column of each table), and read the corresponding probability. The following table is the binomial probabilities for $n = 6$. Note that the probabilities in each column of the binomial table must add up to 1.0.

Binomial (n	Probability										Distribution	Table 6)

Probability												
X.....	0.1.....	0.2.....	0.3.....	0.4.....	0.5.....	0.6.....	0.7.....	0.8.....	0.9.....			

0.....	0.531.....	0.118.....								0.000		
1.....	0.354.....	0.303.....								0.000		
2.....	0.098.....	0.324.....								0.001		
3.....	0.015.....	0.185.....								0.015		
4.....	0.001.....	0.060.....								0.098		
5.....	0.000.....	0.010.....								0.354		
6.....	0.000.....	0.001.....								0.531		

Example:

Suppose that an examination consists of six true and false questions, and assume that a student has no knowledge of the subject matter. The probability that the student will guess the correct answer to the first question is 30%. Likewise, the probability of guessing each of the remaining questions correctly is also 30%. What is the probability of getting more than three correct answers? For the above problem, $n = 6$, $p = 0.30$, and $X > 3$. In the above table, search along the row of p values for 0.30. The problem is to locate the $P(X > 3)$. Thus, the answer involves summing the

probabilities for $X = 4, 5$, and 6 . These values appear in the X column at the intersection of each X value and $p = 0.30$, as follows:
 $P(X > 3) = \text{Summation of } \{P(X=4) + P(X=5) + P(X=6)\} = (0.060) + (0.010) + (0.001) = 0.071 \text{ or } 7.1\%$
 Thus, we may conclude that if 30% of the exam questions are answered by guessing, the probability is 0.071 (or 7.1%) that more than four of the questions are answered correctly by the student.

Graphing the Binomial Distribution:

The graph of a binomial distribution can be constructed by using all the possible X values of a distribution and their associated probabilities. The X values are graphed along the X axis, and the probabilities are graphed along the Y axis. Note that the graph of the binomial distribution has three shapes: *If $p < 0.5$, the graph is positively skewed, if $p > 0.5$, the graph is negatively skewed, and if $p = 0.5$, the graph is symmetrical.* The skewness is eliminated as n gets large. In other words, if n remains constant but p becomes larger and larger up to 0.50, the shape of the binomial probability distribution becomes more symmetrical. If p remains the same but n becomes larger and larger, the shape of the binomial probability distribution becomes more symmetrical.

Example:

In a large consignment of electric bulb 10% are defective. A random sample of 20 is taken for inspection. Find the probability that a) all are good bulbs b) there are almost 3 defective bulbs c) there are exactly 3 defective bulbs

Solution:

Here $n = 20$: $p = 10/100 = 0.1$: $q = 0.9$;

By binomial distribution, the probability of getting x defective bulbs

a) probability of getting all good bulbs = probability of getting zero defective bulbs.

$$= p(x=0)$$

$$= (0.9)^{20}$$

$$b) p(x \leq 3) = p(x=0) + p(x=1) + p(x=2) + p(x=3)$$

$$= .8671$$

$$c) p(x=3) = .1901$$

Binomial Distribution

To understand binomial distributions and binomial probability, it helps to understand binomial experiments and some associated notation; so we cover those topics first.

Binomial Experiment

A **binomial experiment** (also known as a **Bernoulli trial**) is a [statistical experiment](#) that has the following properties:

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by P , is the same on every trial.
- The trials are [independent](#); that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation

The following notation is helpful, when we talk about binomial probability.

- x : The number of successes that result from the binomial experiment.
- n : The number of trials in the binomial experiment.
- P : The probability of success on an individual trial.
- Q : The probability of failure on an individual trial. (This is equal to $1 - P$.)
- $b(x, n, P)$: Binomial probability - the probability that an n -trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is P .

- ${}_nC_r$: The number of [combinations](#) of n things, taken r at a time.

Binomial Distribution

A **binomial random variable** is the number of successes x in n repeated trials of a binomial experiment. The [probability distribution](#) of a binomial random variable is called a **binomial distribution** (also known as a **Bernoulli distribution**).

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The binomial distribution has the following properties:

- The mean of the distribution (μ_x) is equal to $n * P$.
- The [variance](#) (σ^2_x) is $n * P * (1 - P)$.
- The [standard deviation](#) (σ_x) is $\sqrt{n * P * (1 - P)}$.

Binomial Probability

The **binomial probability** refers to the probability that a binomial experiment results in exactly x successes. For example, in the above table, we see that the binomial probability of getting exactly one head in two coin flips is 0.50.

Given x , n , and P , we can compute the binomial probability based on the following formula:

Binomial Formula. Suppose a binomial experiment consists of n trials and results in x successes. If the probability of success on an individual trial is P , then the binomial probability is:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$$

Example**1**

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is $1/6$ or about 0.167. Therefore, the binomial probability is:

$$b(2; 5, 0.167) = {}_5C_2 * (0.167)^2 * (0.833)^3$$

$$b(2; 5, 0.167) = 0.161$$

Cumulative Binomial Probability

A **cumulative binomial probability** refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

For example, we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

$$b(x < 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \dots + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5)$$

Binomial Calculator

As you may have noticed, the binomial formula requires many time-consuming computations. The Binomial Calculator can do this work for you - quickly, easily, and error-free. Use the Binomial Calculator to compute binomial probabilities and cumulative binomial probabilities. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

[Binomial
Calculator](#)

Example 1

What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?

Solution: To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$$\begin{aligned} b(x < 45; 100, 0.5) &= b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \dots + \\ &= b(x = 45; 100, 0.5) \\ b(x < 45; 100, 0.5) &= 0.184 \end{aligned}$$

Example 2

The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

Solution: To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$$\begin{aligned} b(x < 2; 5, 0.3) &= b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3) \\ b(x < 2; 5, 0.3) &= 0.1681 + 0.3601 + 0.3087 \\ b(x < 2; 5, 0.3) &= 0.8369 \end{aligned}$$

Example 3

What is the probability that the world series will last 4 games? 5 games? 6 games? 7 games? Assume that the teams are evenly matched.

Solution: This is a very tricky application of the binomial distribution. If you can follow the logic of this solution, you have a good understanding of the material covered in the tutorial, to this point.

In the world series, there are two baseball teams. The series ends when the winning team wins 4 games. Therefore, we define a success as a win by the team that ultimately becomes the world series champion.

For the purpose of this analysis, we assume that the teams are evenly matched. Therefore, the probability that a particular team wins a particular game is 0.5.

Let's look first at the simplest case. What is the probability that the series lasts only 4 games. This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

$$b(4; 4, 0.5) = {}_4C_4 * (0.5)^4 * (0.5)^0 = 0.0625$$

Similarly, when we compute the probability of the American League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be $0.0625 + 0.0625 = 0.125$; since the series would end if either the American or National League team won 4 games in a row.

Now let's tackle the question of finding probability that the world series ends in 5 games. The trick in finding this solution is to recognize that the series can only end in 5 games, if one team has won 3 out of the first 4 games. So let's first find the probability that the American League team wins exactly 3 of the first 4 games.

$$b(3; 4, 0.5) = {}_4C_3 * (0.5)^3 * (0.5)^1 = 0.25$$

Okay, here comes some more tricky stuff, so listen up. Given that the American League team has won 3 of the first 4 games, the American League team has a 50/50 chance of winning the fifth game to end the series. Therefore, the probability of the American League team winning the series in 5 games is $0.25 * 0.50 = 0.125$. Since the National League team could also win the series in 5 games, the probability that the series ends in 5 games would be $0.125 + 0.125 = 0.25$.

The rest of the problem would be solved in the same way. You should find that the probability of the series ending in 6 games is 0.3125; and the probability of the series ending in 7 games is also 0.3125.

While this is statistically correct in theory, over the years the actual world series has turned out differently, with more series than expected lasting 7 games. For an interesting discussion of why world series reality differs from theory, see Ben Stein's explanation of [why 7-game world series are more common than expected](#).

Negative Binomial and Geometric Distributions

In this lesson, we cover the negative binomial distribution and the geometric distribution. As we will see, the geometric distribution is a special case of the negative binomial distribution.

Negative Binomial Experiment

A **negative binomial experiment** is a [statistical experiment](#) that has the following properties:

- The experiment consists of x repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.

- The probability of success, denoted by P , is the same on every trial.
- The trials are [independent](#); that is, the outcome on one trial does not affect the outcome on other trials.
- The experiment continues until r successes are observed, where r is specified in advance.

Consider the following statistical experiment. You flip a coin repeatedly and count the number of times the coin lands on heads. You continue flipping the coin until it has landed 5 times on heads. This is a negative binomial experiment because:

- The experiment consists of repeated trials. We flip a coin repeatedly until it has landed 5 times on heads.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.
- The experiment continues until a fixed number of successes have occurred; in this case, 5 heads.

Notation

The following notation is helpful, when we talk about negative binomial probability.

- x : The number of trials required to produce r successes in a negative binomial experiment.
- r : The number of successes in the negative binomial experiment.
- P : The probability of success on an individual trial.
- Q : The probability of failure on an individual trial. (This is equal to $1 - P$.)
- $b^*(x, r, P)$: Negative binomial probability - the probability that an x -trial negative binomial experiment results in the r th success on the x th trial, when the probability of success on an individual trial is P .
- ${}_nC_r$: The number of [combinations](#) of n things, taken r at a time.

Negative Binomial Distribution

A **negative binomial random variable** is the number X of repeated trials to produce r successes in a negative binomial experiment. The [probability distribution](#) of a negative binomial random variable is called a **negative binomial distribution**. The

negative binomial distribution is also known as the **Pascal distribution**.

Suppose we flip a coin repeatedly and count the number of heads (successes). If we continue flipping the coin until it has landed 2 times on heads, we are conducting a negative binomial experiment. The negative binomial random variable is the number of coin flips required to achieve 2 heads. In this example, the number of coin flips is a random variable that can take on any integer value between 2 and plus infinity. The negative binomial probability distribution for this example is presented below.

Number of coin flips	Probability
2	0.25
3	0.25
4	0.1875
5	0.125
6	0.078125
7 or more	0.109375

Negative Binomial Probability

The **negative binomial probability** refers to the probability that a negative binomial experiment results in $r - 1$ successes after trial $x - 1$ and r successes after trial x . For example, in the above table, we see that the negative binomial probability of getting the second head on the sixth flip of the coin is 0.078125.

Given x , r , and P , we can compute the negative binomial probability based on the following formula:

Negative Binomial Formula. Suppose a negative binomial experiment consists of x trials and results in r successes. If the

probability of success on an individual trial is P , then the negative binomial probability is:

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * (1 - P)^{x-r}$$

The negative binomial distribution has the following properties:

- The mean of the distribution is: $\mu = rQ / P$.
- The [variance](#) is: $\sigma^2 = rQ / P^2$.

Geometric Distribution

The **geometric distribution** is a special case of the negative binomial distribution. It deals with the number of trials required for a single success. Thus, the geometric distribution is negative binomial distribution where the number of successes (r) is equal to 1.

An example of a geometric distribution would be tossing a coin until it lands on heads. We might ask: What is the probability that the first head occurs on the third flip? That probability is referred to as a **geometric probability** and is denoted by $g(x; P)$. The formula for geometric probability is given below.

Geometric Probability Formula. Suppose a negative binomial experiment consists of x trials and results in one success. If the probability of success on an individual trial is P , then the geometric probability is:

$$g(x; P) = P * Q^{x-1}$$

The geometric distribution has the following properties:

- The mean of the distribution is: $\mu = Q / P$.
- The [variance](#) is: $\sigma^2 = Q / P^2$.

Sample Problems

The problems below show how to apply your new-found knowledge of the negative binomial distribution (see Example 1) and the geometric distribution (see Example 2).

Negative Binomial Calculator

As you may have noticed, the negative binomial formula requires some potentially time-consuming computations. The Negative Binomial Calculator can do this work for you - quickly, easily, and error-free. Use the Negative Binomial Calculator to compute negative binomial probabilities and geometric probabilities. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

[Negative Binomial Calculator](#)

Example

1

Bob is a high school basketball player. He is a 70% free throw shooter. That means his probability of making a free throw is 0.70. During the season, what is the probability that Bob makes his third free throw on his fifth shot?

Solution: This is an example of a negative binomial experiment. The probability of success (P) is 0.70, the number of trials (x) is 5, and the number of successes (r) is 3.

To solve this problem, we enter these values into the negative binomial formula.

$$b^*(x; r, P) = {}^{x-1}C_{r-1} * P^r * Q^{x-r}$$

$$b^*(5; 3, 0.7) = {}^4C_2 * 0.7^3 * 0.3^2$$

$$b^*(5; 3, 0.7) = 6 * 0.343 * 0.09 = 0.18522$$

Thus, the probability that Bob will make his third successful free throw on his fifth shot is 0.18522.

Example

2

Let's reconsider the above problem from Example 1. This time, we'll ask a slightly different question: What is the probability that Bob makes his first free throw on his fifth shot?

Solution: This is an example of a geometric distribution, which is a special case of a negative binomial distribution. Therefore, this problem can be solved using the negative binomial formula or the geometric formula. We demonstrate each approach below, beginning with the negative binomial formula.

The probability of success (P) is 0.70, the number of trials (x) is 5, and the number of successes (r) is 1. We enter these values into the negative binomial formula.

$$\begin{aligned}
 b^*(x; r, P) &= {}^{x-1}C_{r-1} * P^r * Q^{n-x} \\
 b^*(5; 1, 0.7) &= {}^4C_0 * 0.7^1 * 0.3^4 \\
 b^*(5; 3, 0.7) &= 0.00567
 \end{aligned}$$

Now, we demonstrate a solution based on the geometric formula.

$$\begin{aligned}
 g(x; P) &= P * Q^{x-1} \\
 g(5; 0.7) &= 0.7 * 0.3^4 = 0.00567
 \end{aligned}$$

Notice that each approach yields the same answer.

Hypergeometric Distribution

This lesson covers hypergeometric experiments, hypergeometric distributions, and hypergeometric probability.

Hypergeometric Experiments

A **hypergeometric experiment** is a [statistical experiment](#) that has the following properties:

- A [sample](#) of size n is randomly selected [without replacement](#) from a [population](#) of N items.
- In the population, k items can be classified as successes, and $N - k$ items can be classified as failures.

Consider the following statistical experiment. You have an urn of 10 marbles - 5 red and 5 green. You randomly select 2 marbles without replacement and count the number of red marbles you have selected. This would be a hypergeometric experiment.

Note that it would not be a [binomial experiment](#). A binomial experiment requires that the probability of success be constant on every trial. With the above experiment, the probability of a success changes on every trial. In the beginning, the probability of selecting a red marble is 5/10. If you select a red marble on the first trial, the probability of selecting a red marble on the second trial is 4/9. And if you select a green marble on the first trial, the probability of selecting a red marble on the second trial is 5/9.

Note further that if you selected the marbles with replacement, the probability of success would not change. It would be 5/10 on every trial. Then, this would be a binomial experiment.

Notation

The following notation is helpful, when we talk about hypergeometric distributions and hypergeometric probability.

- N : The number of items in the [population](#).
- k : The number of items in the population that are classified as successes.
- n : The number of items in the [sample](#).
- x : The number of items in the sample that are classified as successes.
- ${}_kC_x$: The number of [combinations](#) of k things, taken x at a time.
- $h(x; N, n, k)$: **hypergeometric probability** - the probability that an n -trial hypergeometric experiment results in exactly x successes, when the population consists of N items, k of which are classified as successes.

Hypergeometric Distribution

A **hypergeometric random variable** is the number of successes that result from a hypergeometric experiment. The [probability distribution](#) of a hypergeometric random variable is called a **hypergeometric distribution**.

Given x , N , n , and k , we can compute the hypergeometric probability based on the following formula:

Hypergeometric Formula. Suppose a population consists of N items, k of which are successes. And a random sample drawn from that population consists on n items, x of which are successes. Then the hypergeometric probability is:

$$h(x; N, n, k) = [{}_kC_x] [{}_{N-k}C_{n-x}] / [{}_NC_n]$$

The hypergeometric distribution has the following properties:

- The mean of the distribution is equal to $n * k / N$.
- The [variance](#) is $n * k * (N - k) * (N - n) / [N^2 * (N - 1)]$.

Example

1

Suppose we randomly select 5 cards without replacement from an ordinary deck of playing cards. What is the probability of getting exactly 2 red cards (i.e., hearts or diamonds)?

Solution: This is a hypergeometric experiment in which we know the following:

- $N = 52$; since there are 52 cards in a deck.
- $k = 26$; since there are 26 red cards in a deck.
- $n = 5$; since we randomly select 5 cards from the deck.
- $x = 2$; since 2 of the cards we select are red.

We plug these values into the hypergeometric formula as follows:

$$h(x; N, n, k) = \frac{[{}_k C_x] [{}_{N-k} C_{n-x}]}{[{}_N C_n]}$$

$$h(2; 52, 5, 26) = \frac{[{}_{26} C_2] [{}_{26} C_3]}{[{}_{52} C_5]}$$

$$h(2; 52, 5, 26) = \frac{[325] [2600]}{[2,598,960]} = 0.32513$$

Thus, the probability of randomly selecting 2 red cards is 0.32513.

Hypergeometric Calculator

As you surely noticed, the hypergeometric formula requires many time-consuming computations. The Stat Trek Hypergeometric Calculator can do this work for you - quickly, easily, and error-free. Use the Hypergeometric Calculator to compute hypergeometric probabilities and cumulative hypergeometric probabilities. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

[Hypergeometric Calculator](#)

Cumulative Hypergeometric Probability

A **cumulative hypergeometric probability** refers to the probability that the hypergeometric random variable is greater than or equal to some specified lower limit and less than or equal to some specified upper limit.

For example, suppose we randomly select five cards from an ordinary deck of playing cards. We might be interested in the cumulative hypergeometric probability of obtaining 2 or fewer hearts. This would be the probability of obtaining 0 hearts plus the probability of obtaining 1 heart plus the probability of obtaining 2 hearts, as shown in the example below.

Example

1

Suppose we select 5 cards from an ordinary deck of playing cards. What is the probability of obtaining 2 or fewer hearts?

Solution: This is a hypergeometric experiment in which we know the following:

- $N = 52$; since there are 52 cards in a deck.
- $k = 13$; since there are 13 hearts in a deck.
- $n = 5$; since we randomly select 5 cards from the deck.
- $x = 0$ to 2 ; since our selection includes 0, 1, or 2 hearts.

We plug these values into the hypergeometric formula as follows:

$$\begin{aligned}
 h(x < 2; N, n, k) &= h(x < 2; 52, 5, 13) \\
 h(x < 2; 52, 5, 13) &= h(x = 0; 52, 5, 13) + h(x = 1; 52, 5, 13) + h(x = 2; 52, 5, 13) \\
 h(x < 2; 52, 5, 13) &= \left[\frac{{}_{13}C_0 {}_{39}C_5}{{}_{52}C_5} \right] + \left[\frac{{}_{13}C_1 {}_{39}C_4}{{}_{52}C_5} \right] \\
 &+ \left[\frac{{}_{13}C_2 {}_{39}C_3}{{}_{52}C_5} \right] \\
 h(x < 2; 52, 5, 13) &= \left[\frac{(1)(575,757)}{(2,598,960)} \right] + \left[\frac{(13)(82,251)}{(270,725)} \right] + \left[\frac{(78)(9139)}{(22,100)} \right] \\
 h(x < 2; 52, 5, 13) &= [0.2215] + [0.4114] + [0.2743] \\
 h(x < 2; 52, 5, 13) &= 0.9072
 \end{aligned}$$

Thus, the probability of randomly selecting at most 2 hearts is 0.9072.

Multinomial Distribution

Multinomial Experiment

A **multinomial experiment** is a [statistical experiment](#) that has the following properties:

- The experiment consists of n repeated trials.
- Each trial has a discrete number of possible outcomes.
- On any given trial, the probability that a particular outcome will occur is constant.
- The trials are [independent](#); that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You toss two dice three times, and record the outcome on each toss. This is a multinomial experiment because:

- The experiment consists of repeated trials. We toss the dice three times.
- Each trial can result in a discrete number of outcomes - 2 through 12.
- The probability of any outcome is constant; it does not change from one toss to the next.

- The trials are independent; that is, getting a particular outcome on one trial does not affect the outcome on other trials.

Note: A [binomial experiment](#) is a special case of a multinomial experiment. Here is the main difference. With a binomial experiment, each trial can result in two - and only two - possible outcomes. With a multinomial experiment, each trial can have two *or more* possible outcomes.

Multinomial Distribution

A **multinomial distribution** is the [probability distribution](#) of the outcomes from a multinomial experiment. The multinomial formula defines the probability of any outcome from a multinomial experiment.

Multinomial Formula. Suppose a multinomial experiment consists of n trials, and each trial can result in any of k possible outcomes: E_1, E_2, \dots, E_k . Suppose, further, that each possible outcome can occur with probabilities p_1, p_2, \dots, p_k . Then, the probability (P) that E_1 occurs n_1 times, E_2 occurs n_2 times, \dots , and E_k occurs n_k times is

$$P = [n! / (n_1! * n_2! * \dots * n_k!)] * (p_1^{n_1} * p_2^{n_2} * \dots * p_k^{n_k})$$

where $n = n_1 + n_2 + \dots + n_k$.

The examples below illustrate how to use the multinomial formula to compute the probability of an outcome from a multinomial experiment.

Multinomial Calculator

As you may have noticed, the multinomial formula requires many time-consuming computations. The Multinomial Calculator can do this work for you - quickly, easily, and error-free. Use the Multinomial Calculator to compute the probability of outcomes from multinomial experiments. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

[Multinomial
Calculator](#)

Multinomial Probability: Sample Problems

Example

1

Suppose a card is drawn randomly from an ordinary deck of playing cards, and then put back in the deck. This exercise is repeated five times. What is the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs?

Solution: To solve this problem, we apply the multinomial formula. We know the following:

- The experiment consists of 5 trials, so $n = 5$.
- The 5 trials produce 1 spade, 1 heart, 1 diamond, and 2 clubs; so $n_1 = 1$, $n_2 = 1$, $n_3 = 1$, and $n_4 = 2$.
- On any particular trial, the probability of drawing a spade, heart, diamond, or club is 0.25, 0.25, 0.25, and 0.25, respectively. Thus, $p_1 = 0.25$, $p_2 = 0.25$, $p_3 = 0.25$, and $p_4 = 0.25$.

We plug these inputs into the multinomial formula, as shown below:

$$P = [n! / (n_1! * n_2! * \dots * n_k!)] * (p_1^{n_1} * p_2^{n_2} * \dots * p_k^{n_k})$$

$$P = [5! / (1! * 1! * 1! * 2!)] * [(0.25)^1 * (0.25)^1 * (0.25)^1 * (0.25)^2]$$

$$P = 0.05859$$

Thus, if we draw five cards [with replacement](#) from an ordinary deck of playing cards, the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs is 0.05859.

Example

2

Suppose we have a bowl with 10 marbles - 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select 4 marbles from the bowl, [with replacement](#). What is the probability of selecting 2 green marbles and 2 blue marbles?

Solution: To solve this problem, we apply the multinomial formula. We know the following:

- The experiment consists of 4 trials, so $n = 4$.
- The 4 trials produce 0 red marbles, 2 green marbles, and 2 blue marbles; so $n_{\text{red}} = 0$, $n_{\text{green}} = 2$, and $n_{\text{blue}} = 2$.

- On any particular trial, the probability of drawing a red, green, or blue marble is 0.2, 0.3, and 0.5, respectively. Thus, $p_{\text{red}} = 0.2$, $p_{\text{green}} = 0.3$, and $p_{\text{blue}} = 0.5$

We plug these inputs into the multinomial formula, as shown below:

$$P = \left[\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!} \right] \cdot (p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_k^{n_k})$$

$$P = \left[\frac{4!}{0! \cdot 2! \cdot 2!} \right] \cdot [(0.2)^0 \cdot (0.3)^2 \cdot (0.5)^2]$$

$$P = 0.135$$

Thus, if we draw 4 marbles [with replacement](#) from the bowl, the probability of drawing 0 red marbles, 2 green marbles, and 2 blue marbles is 0.135.

The Poisson Distribution:

The poisson distribution is another discrete probability distribution. It is named after Simeon-Denis Poisson (1781-1840), a French mathematician. The poisson distribution depends *only* on the average number of occurrences per unit time of space. There is no n , and no p . The poisson probability distribution provides a close approximation to the binomial probability distribution when n is large and p is quite small or quite large. In other words, if $n > 20$ and $np \leq 5$ [or $n(1-p) \leq 5$], then we may use poisson distribution as an approximation to binomial distribution. for detail discussion of the poisson probability distribution, refer to the text.

Poisson Distribution

Attributes of a Poisson Experiment

A **Poisson experiment** is a [statistical experiment](#) that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- e : A constant equal to approximately 2.71828. (Actually, e is the base of the natural logarithm system.)
- μ : The mean number of successes that occur in a specified region.
- x : The actual number of successes that occur in a specified region.
- $P(x; \mu)$: The **Poisson probability** that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ .

Poisson Distribution

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The [probability distribution](#) of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ . Then, the Poisson probability is:

$$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ .
- The [variance](#) is also equal to μ .

Example

1

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 2$; since 2 homes are sold per day, on average.
- $x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.
- $e = 2.71828$; since e is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$\begin{array}{rclclcl}
 P(x; \mu) & = & (e^{-\mu}) & (\mu^x) & / & x! \\
 P(3; 2) & = & (2.71828^{-2}) & (2^3) & / & 3! \\
 P(3; 2) & = & (0.13534) & (8) & / & 6 \\
 P(3; 2) & = & 0.180 & & &
 \end{array}$$

Thus, the probability of selling 3 homes tomorrow is 0.180 .

Poisson Calculator

Clearly, the Poisson formula requires many time-consuming computations. The Stat Trek Poisson Calculator can do this work for you - quickly, easily, and error-free. Use the Poisson Calculator to compute Poisson probabilities and cumulative Poisson probabilities. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

[Poisson
Calculator](#)

Cumulative Poisson Probability

A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

Example

1

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 5$; since 5 lions are seen per safari, on average.
- $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- $e = 2.71828$; since e is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: $P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$. To compute this sum, we use the Poisson formula:

$$\begin{aligned}
 P(x < 3, 5) &= P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5) \\
 P(x < 3, 5) &= [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!] \\
 P(x < 3, 5) &= [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6] \\
 P(x < 3, 5) &= [0.0067] + [0.03369] + [0.084224] + [0.140375] \\
 P(x < 3, 5) &= 0.2650
 \end{aligned}$$

Thus, the probability of seeing at no more than 3 lions is 0.2650.

Standard Normal Distribution

The **standard normal distribution** is a special case of the [normal distribution](#). It is the distribution that occurs when a [normal random variable](#) has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a **standard score** or a **z-score**. Every normal random variable X can be transformed into a z score via the following equation:

$$z = (X - \mu) / \sigma$$

where X is a normal random variable, μ is the mean mean of X , and σ is the standard deviation of X .

Standard Normal Distribution Table

A **standard normal distribution table** shows a [cumulative probability](#) associated with a particular z -score. Table rows show the whole number and tenths place of the z -score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z -score) appears in the cell of the table.

For example, a section of the standard normal table is reproduced below. To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, $P(Z < -1.31) = 0.0951$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
...
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
...
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Of course, you may not be interested in the probability that a standard normal random variable falls between minus infinity and a given value. You may want to know the probability that it lies between a given value and plus infinity. Or you may want to know the probability that a standard normal random variable lies between two given values. These probabilities are easy to compute from a normal distribution table. Here's how.

- Find $P(Z > a)$. The probability that a standard normal random variable (z) is greater than a given value (a) is easy to find. The table shows the $P(Z < a)$. The $P(Z > a) = 1 - P(Z < a)$.

Suppose, for example, that we want to know the probability that a z-score will be greater than 3.00. From the table (see above), we find that $P(Z < 3.00) = 0.9987$. Therefore, $P(Z > 3.00) = 1 - P(Z < 3.00) = 1 - 0.9987 = 0.0013$.

- Find $P(a < Z < b)$. The probability that a standard normal random variables lies between two values is also easy to find. The $P(a < Z < b) = P(Z < b) - P(Z < a)$.

For example, suppose we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table (see above), we find that $P(Z < -1.20) = 0.1151$; and $P(Z < -1.40) = 0.0808$. Therefore, $P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) = 0.1151 - 0.0808 = 0.0343$.

In school or on the Advanced Placement Statistics Exam, you may be called upon to use or interpret standard normal distribution tables. Standard normal tables are commonly found in appendices of most statistics texts.

The Normal Distribution as a Model for Measurements

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

- Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier: $z = (X - \mu) / \sigma$.
- Find probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek's free [normal distribution calculator](#)), or handheld [graphing calculators](#) to find probabilities associated with the z-scores.

The problem in the next section demonstrates the use of the normal distribution as a model for measurement.

Test Your Understanding of This Lesson

Problem 1

Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed.)

- (A) 0.10
- (B) 0.18
- (C) 0.50
- (D) 0.82
- (E) 0.90

Solution

The correct answer is B. As part of the solution to this problem, we assume that test scores are normally distributed. In this way, we use the [normal distribution](#) as a model for measurement. Given an assumption of normality, the solution involves three steps.

- First, we transform Molly's test score into a [z-score](#), using the z-score transformation equation.

$$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$$

- Then, using an online calculator (e.g., Stat Trek's free [normal distribution calculator](#)), a handheld [graphing calculator](#), or the standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find $P(Z < 0.90) = 0.8159$.
- Therefore, the $P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$.

Thus, we estimate that 18.41 percent of the students tested had a higher score than Molly.

The Hypergeometric Distribution:

Another discrete probability distribution is the hypergeometric distribution. The binomial probability distribution assumes that the population from which the sample is selected is very large. For this reason, the probability of success does not change with each trial. The hypergeometric distribution is used to determine the probability of a specified number of successes and/or failures when (1) a sample is selected from a finite population without replacement and/or (2) when the sample size, n , is greater than or equal to 5% of the population size, N , i.e., $[n \geq 5\% N]$. Note that by *finite population* we mean a population which consist of a fixed number of known individuals, objects, or measurments. For example, there were 489 applications for the nursing school at Clayton State College in 1994. For detail discussion of the hypergeometric probability distribution, refer to the text.

Introduction:

In lecture number four we said that a continuous random variable is a variable which can take on any value over a given interval. Continuous variables are measured, not counted. Items such as height, weight and time are continuous and can take on fractional values. For example, a basketball player may be 6.8432 feet tall. There are many continuous probability distributions, such as, uniform distribution, normal distribution, the t distribution, the chi-square distribution, exponential distribution, and F distribution. In

this lecture note, we will concentrate on the uniform distribution, and normal distribution.

Uniform (or Rectangular) Distribution:

Among the continuous probability distribution, the uniform distribution is the simplest one of all. The following figure shows an example of a uniform distribution. In a uniform distribution, the area under the curve is equal to the product of the length and the height of the rectangle and equals to one.

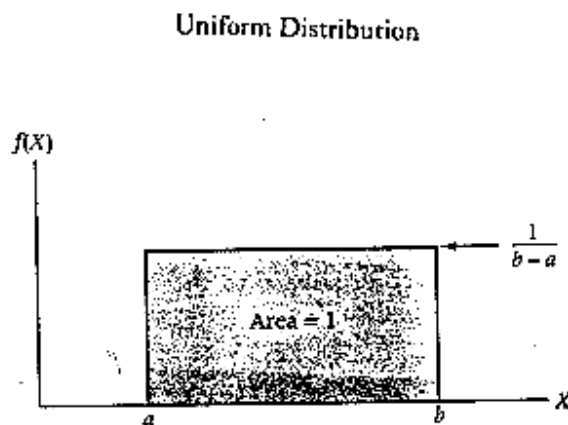


Figure 1

where: a=lower limit of the range or interval, and b=upper limit of the range or interval.

Note that in the above graph, since area of the rectangle = (length)(height) = 1, and since length = (b - a), thus we can write: (b - a)(height) = 1 or height = $f(X) = 1/(b - a)$. The following equations are used to find the mean and standard deviation of a uniform distribution:

$$\text{Mean} = \mu = \frac{a + b}{2}$$

$$\text{Standard deviation} = \sigma = \frac{b - a}{\sqrt{12}}$$

$$f(X) = \text{height} = \frac{1}{(b - a)}$$

Example:

There are many cases in which we may be able to apply the uniform distribution. As an example, suppose that the research department of a steel factory believes that one of the company's rolling machines is producing sheets of steel of different thickness. The thickness is a uniform random variable with values between 150 and 200 millimeters. Any sheets less than 160 millimeters thick must be scrapped because they are unacceptable to the buyers. We want to calculate the mean and the standard deviation of the X (the thickness of the sheet produced by this machine), and the fraction of steel sheet produced by this machine that have to be scrapped. The following figure displays the uniform distribution for this example.

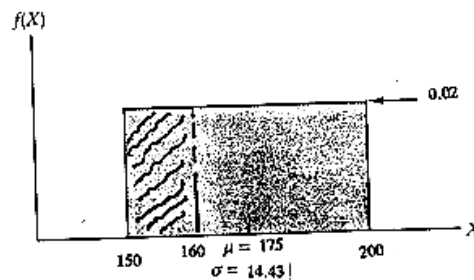


Figure 2

Note that for continuous distribution, probability is calculated by finding the area under the function over a specific interval. In other words, for continuous distributions, there is no probability at any one point. The probability of $X \geq b$ or of $X \leq a$ is zero because there is no area above b or below a , and area between a and b is equal to one, see figure 1.

The probability of the variables falling between any two points, such as c and d in figure 2, are calculated as follows:
 $P(c \leq x \leq d) = (d - c) / (b - a) = ?$
 In this example $c = a = 150$, $d = 160$, and $b = 200$, therefore:
 Mean = $(a + b) / 2 = (150 + 200) / 2 = 175$ millimeters, standard deviation is the square root of 208.3, which is equal to 14.43 millimeters, and $P(c \leq x \leq d) = (160 - 150) / (200 - 150) = 1/5$ thus, of all the sheets made by this machine, 20% of the production must be scrapped.)=....

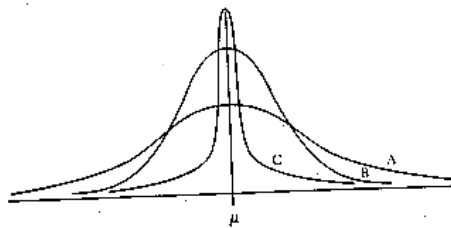
Normal Distribution or Normal Curve:

Normal distribution is probably one of the most important and widely used continuous distribution. It is known as a normal random variable, and its probability distribution is called a normal distribution. The following are the characteristics of the normal distribution:

Characteristics of the Normal Distribution:

1. It is bell shaped and is symmetrical about its mean.
2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.
3. It is a continuous distribution.
4. It is a family of curves, i.e., every unique pair of mean and standard deviation defines a different normal distribution. Thus, the normal distribution is completely described by two parameters: mean and standard deviation. See the following figure.
5. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the mean is 0.5.
6. It is unimodal, i.e., values mound up only in the center of the curve.
7. The probability that a random variable will have a value between any two points is equal to the area under the curve between those points.

Normal Curves with the Same Mean but Different Standard Deviations



Normal Curves with Different Means but the Same Standard Deviation

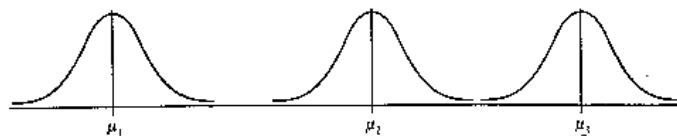


Figure 3

Note that the integral calculus is used to find the area under the normal distribution curve. However, this can be avoided by transforming all normal distribution to fit the standard normal distribution. This conversion is done by rescaling the normal distribution axis from its true units (time, weight, dollars, and...) to a standard measure called Z score or Z value. A Z score is the number of standard deviations that a value, X, is away from the mean. If the value of X is greater than the mean, the Z score is positive; if the value of X is less than the mean, the Z score is negative. The Z score or equation is as follows:

$$Z = (X - \text{Mean}) / \text{Standard deviation}$$

A standard Z table can be used to find probabilities for any normal curve problem that has been converted to Z scores. For the table, refer to the text. The Z distribution is a normal distribution with a mean of 0 and a standard deviation of 1. The following steps are helpful when working with the normal curve problems:

1. Graph the normal distribution, and shade the area related to the probability you want to find.
2. Convert the boundaries of the shaded area from X values to the standard normal random variable Z values using the Z formula above.
3. Use the standard Z table to find the probabilities or the areas related to the Z values in step 2.

Example

One:

Graduate Management Aptitude Test (GMAT) scores are widely used by graduate schools of business as an entrance requirement. Suppose that in one particular year, the mean score for the GMAT was 476, with a standard deviation of 107. Assuming that the GMAT scores are normally distributed, answer the following questions:

Question 1. What is the probability that a randomly selected score from this GMAT falls between 476 and 650? $476 \leq x \leq 650$ the following figure shows a graphic representation of this problem.

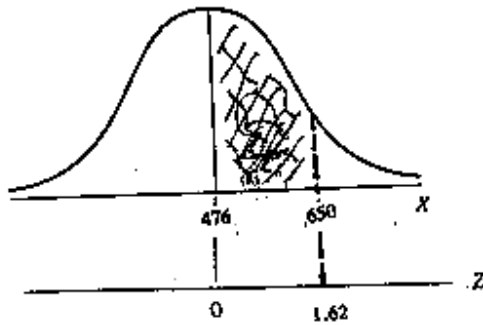


Figure 4

Applying the Z equation, we get: $Z = (650 - 476)/107 = 1.62$. The Z value of 1.62 indicates that the GMAT score of 650 is 1.62 standard deviation above the mean. The standard normal table gives the probability of value falling between 650 and the mean. The whole number and tenths place portion of the Z score appear in the first column of the table. Across the top of the table are the values of the hundredths place portion of the Z score. Thus the answer is that 0.4474 or 44.74% of the scores on the GMAT fall between a score of 650 and 476.

Question 2. What is the probability of receiving a score greater than 750 on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., $P(X \geq 750) = ?$. This problem is asking for determining the area of the upper tail of the distribution. The Z score is: $Z = (750 - 476)/107 = 2.56$. From the table, the probability for this Z score is 0.4948. This is the probability of a GMAT with a score between 476 and 750. The rule is that when we want to find the probability in either tail, we must subtract the table value from 0.50. Thus, the answer to this problem is: $0.5 - 0.4948 = 0.0052$ or 0.52%. Note that $P(X \geq 750)$ is the same as $P(X > 750)$, because, in continuous distribution, the area under an exact number such as $X = 750$ is zero. The following figure shows a graphic representation of this problem.

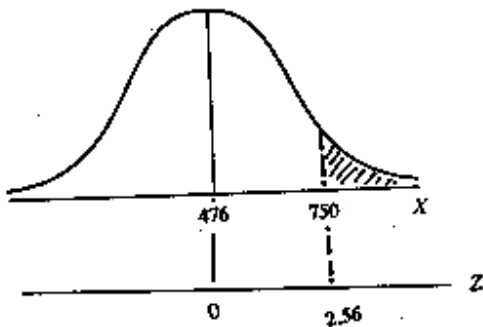


Figure 5

Question 3. What is the probability of receiving a score of 540 or less on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., $P(X \leq 540) = ?$. We are asked to determine the area under the curve for all values less than or equal to 540. The z score is: $z = (540 - 476)/107 = 0.6$. From the table, the probability for this z score is 0.2257 which is the probability of getting a score between the mean (476) and 540. The rule is that when we want to find the probability between two values of x on either side of the mean, we just add the two areas together. Thus, the answer to this problem is: $0.5 + 0.2257 = 0.73$ or 73%. The following figure shows a graphic representation of this problem.

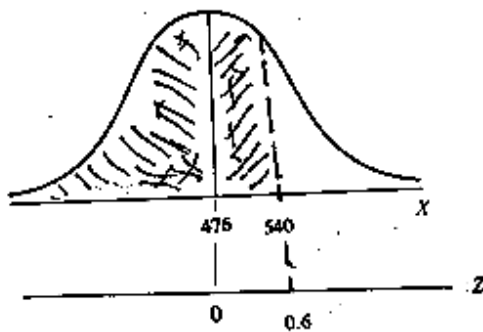


Figure 6

Question 4. What is the probability of receiving a score between 440 and 330 on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., $P(330 < X < 440)$

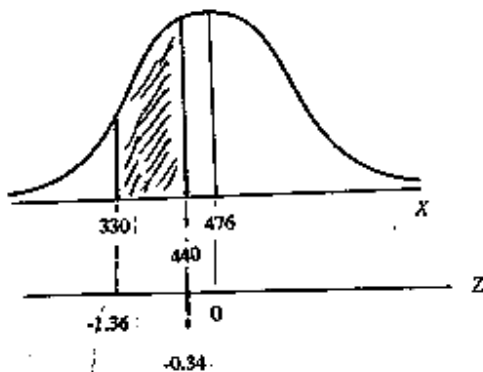


Figure 7

In this problem, the two values fall on the same side of the mean. The Z scores are: $Z_1 = (330 - 476)/107 = -1.36$, and $Z_2 = (440 - 476)/107 = -0.34$. The probability associated with $Z = -1.36$ is 0.4131, and the probability associated with $Z = -0.34$ is 0.1331. The rule is that when we want to find the probability between two values of X on one side of the mean, we just subtract the smaller area

from the larger area to get the probability between the two values. Thus, the answer to this problem is: $0.4131 - 0.1331 = 0.28$ or 28%.

Example

Two:

Suppose that a tire factory wants to set a mileage guarantee on its new model called LA 50 tire. Life tests indicated that the mean mileage is 47,900, and standard deviation of the normally distributed distribution of mileage is 2,050 miles. The factory wants to set the guaranteed mileage so that no more than 5% of the tires will have to be replaced. What guaranteed mileage should the factory announce? i.e., $P(X \leq ?) = 5\%$. In this problem, the mean and standard deviation are given, but X and Z are unknown. The problem is to solve for an X value that has 5% or 0.05 of the X values less than that value. If 0.05 of the values are less than X , then 0.45 lie between X and the mean ($0.5 - 0.05$), see the following graph.

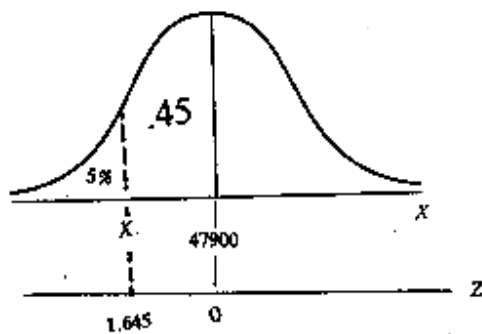


Figure 8

Refer to the standard normal distribution table and search the body of the table for 0.45. Since the exact number is not found in the table, search for the closest number to 0.45. There are two values equidistant from 0.45-- 0.4505 and 0.4495. Move to the left from these values, and read the Z scores in the margin, which are: 1.65 and 1.64. Take the average of these two Z scores, i.e., $(1.65 + 1.64)/2 = 1.645$. Plug this number and the values of the mean and the standard deviation into the Z equation, you get: $Z = (X - \text{mean})/\text{standard deviation}$ or $-1.645 = (X - 47,900)/2,050 =$

44,528 miles.
Thus, the factory should set the guaranteed mileage at 44,528 miles if the objective is not to replace more than 5% of the tires.

The Normal Approximation to the Binomial Distribution:

In lecture note number 5 we talked about the binomial probability distribution, which is a discrete distribution. You remember that we

said as sample sizes get larger, binomial distribution approach the normal distribution in shape regardless of the value of p (probability of success). For large sample values, the binomial distribution is cumbersome to analyze without a computer. Fortunately, the normal distribution is a good approximation for binomial distribution problems for large values of n . The commonly accepted guidelines for using the normal approximation to the binomial probability distribution is when $(n \times p)$ and $[n(1 - p)]$ are both greater than 5.

Example:

Suppose that the management of a restaurant claimed that 70% of their customers returned for another meal. In a week in which 80 new (first-time) customers dined at the restaurant, what is the probability that 60 or more of the customers will return for another meal?, ie., $P(X \geq 60) = ?$.

The solution to this problem can be illustrated as follows: First, the two guidelines that $(n \times p)$ and $[n(1 - p)]$ should be greater than 5 are satisfied: $(n \times p) = (80 \times 0.70) = 56 > 5$, and $[n(1 - p)] = 80(1 - 0.70) = 24 > 5$. Second, we need to find the mean and the standard deviation of the binomial distribution. The mean is equal to $(n \times p) = (80 \times 0.70) = 56$ and standard deviation is square root of $[(n \times p)(1 - p)]$, i.e., square root of 16.8, which is equal to 4.0988. Using the Z equation we get, $Z = (X - \text{mean})/\text{standard deviation} = (59.5 - 56)/4.0988 = 0.85$. From the table, the probability for this Z score is 0.3023 which is the probability between the mean (56) and 60. We must subtract this table value 0.3023 from 0.5 in order to get the answer, i.e., $P(X \geq 60) = 0.5 - 0.3023 = 0.1977$. Therefore, the probability is 19.77% that 60 or more of the 80 first-time customers will return to the restaurant for another meal. See the following graph.

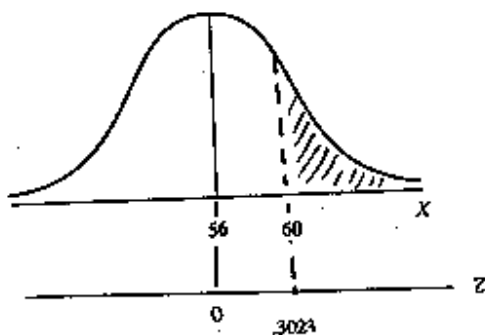


Figure 9

Correction Factor:

The value 0.5 is added or subtracted, depending on the problem, to the value of X when a binomial probability distribution is being approximated by a normal distribution. This correction ensures that most of the binomial problem's information is correctly transferred to the normal curve analysis. This correction is called the correction for continuity. The decision as to how to correct for continuity depends on the equality sign and the direction of the desired outcomes of the binomial distribution. The following table shows some rules of thumb that can help in the application of the correction for continuity, see the above example.

Value Being Determined.....Correction

$X > \dots\dots\dots +0.50$

$X \geq \dots\dots\dots -0.50$

$X < \dots\dots\dots -0.50$

$X \leq \dots\dots\dots +0.50$

$\leq X \leq \dots\dots\dots -0.50" \& +0.50$

$X = \dots\dots\dots -0.50 \& +0.50$

Expectation value

The expectation value of a function $f(x)$ in a variable x is denoted $\langle f(x) \rangle$ or $E\{f(x)\}$. For a single discrete variable, it is defined by

$$\langle f(x) \rangle = \sum_x f(x) P(x),$$

where $P(x)$ is the [probability function](#).

For a single continuous variable it is defined by,

$$\langle f(x) \rangle = \int f(x) P(x) dx.$$

The expectation value satisfies

$$\begin{aligned} \langle ax + by \rangle &= a \langle x \rangle + b \langle y \rangle \\ \langle a \rangle &= a \\ \langle \sum x \rangle &= \sum \langle x \rangle. \end{aligned}$$

For multiple discrete variables

$$\langle f(x_1, \dots, x_n) \rangle = \sum_{x_1, \dots, x_n} f(x_1, \dots, x_n) P(x_1, \dots, x_n).$$

For multiple continuous variables

$$\langle f(x_1, \dots, x_n) \rangle = \int f(x_1, \dots, x_n) P(x_1, \dots, x_n) dx_1 \dots dx_n.$$

The (multiple) expectation value satisfies

$$\begin{aligned} \langle (x - \mu_x)(y - \mu_y) \rangle &= \langle xy - \mu_x y - \mu_y x + \mu_x \mu_y \rangle \\ &= \langle xy \rangle - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\ &= \langle xy \rangle - \langle x \rangle \langle y \rangle, \end{aligned}$$

where μ_i is the [mean](#) for the variable i .

Uniform distribution

Trains arrive at a station at 15 minutes intervals starting at 4am. If a passenger arrives at the station at a time that is uniformly distributed between 9 and 9:30. Find the probability that he has to wait for the train for

- a) less than 6 minutes
- b) more than 10 minutes

Let X be the random variable representing the number of minutes past 9 that the passenger arrives at the station.

- a) he has to wait for less than 6 minutes if he arrives between 9:09 and 9:15 or between 9:24 and 9:30.

So the required probability = $p(9 < X < 15) + p(24 < X < 30)$

$$= 2/5$$

- b) he has to wait for more than 10 minutes if he arrives between 9:00 and 9:05 or between 9:15 and 9:20

hence required probability = $p(0 < X < 5) + p(15 < x < 20) = 1/3$

Have you understood?

1. A production process manufactures computer chips on an average of 2% non-conforming. Every day a random sample size of 50 is taken from the process. If the sample contains more than two non-conforming chips, the process will be stopped. Determine the probability that the process is stopped by the sampling scheme.

Solution:

Here $n=50$; Bernoulli trials $p=.02$

Let X be the total number of non conformal chips

$$\begin{aligned}
 P(x > 2) &= 1 - p(x \leq 2) \\
 &= 1 - \{p(x=0) + p(x=1) + p(x=2)\} \\
 &= 1 - [(.98) + 50(.02)(.98) + 1225(.02)(.98)] \\
 &= 1 - .922 = .078
 \end{aligned}$$

thus the probability that the process is stopped on any day, based on the symphony process is approximately 0.078.

2. When a computer terminal repair person is beeped each time there is a call for service. The number of beeps per hour is known to occur in accordance with a poisson distribution with a mean of $\alpha=2$ per hour. Determine the probability of the two or more beeps in a 1-hour period.

3. A bus arrives every 20 minutes at a specified stop beginning at 6:40am and continues until 8:40am. A certain passenger does not know the schedule, but arrives randomly (uniformly distributed) between 7:00am and 7:30am every morning. What is the probability that the passenger waits for more than 5 minutes for a bus?

4. Harley Davidson, Director Quality control for the Kyoto motor company, is conducting his monthly spot check of automatic transmission. In this procedure, 10 transmissions are removed from the pool of components and are checked for manufacturing defects. Historically only 2% of the transmissions have such flaws. (assume that flaws occur independently in different transmissions)

- a) What is the probability that Harley's sample contains more than two transmissions with manufacturing flaws?
- b) What is the probability that none of the selected transmissions has any manufacturing flaws?

5. The customer accounts of a certain departmental store have an average balance of the Rs120/- and a standard deviation of Rs40/-. Assuming that the account balances are normally distributed, find out

- a) What proportion of accounts is more than Rs150/-?
- b) What proportion of account is between Rs100/- and Rs150/-?
- c) What proportion of account is between Rs60/- and Rs90/-?

6. A book contains 100 misprints distributed at random throughout its 100 pages. What is the probability that a page observed at random contains at least two misprints? (Assume Poisson distribution)

7. the technical team says that on an average, 3 hits of 10 million hits made by the software fails. The marketing department requires that a service level agreement on the Q.S that the probability of occurrence of failure of 4 request hits failing amidst 10 million requests is less than 15.

A) Can the agreement be signed?

b) A technical up gradation at a higher cost can bring down the hit failure rate from a mean of $3/10$ million to $1/10$ million. Is it required?

8. Server crash, brought a down time of 19 minutes in a particular environment. A user does an operation in it that gives a request, once in 3 minutes. Find the probability that the number of requests that fail is greater than 3, assuming that the problem is uniformly distributed?

9. The response time for an application to send a request to another application and get back a response in an enterprise application interconnection was monitored by a tool for 3 months continuously. The mean response time was found to be 600 milli seconds with a standard deviation of 200 milli seconds for the normally distributed random variable.

a) A response time of greater than 1.0 second is flagged as a severity. Find the probability of occurrence of a severity.

b) Find the probability of a response time 800ms.

10) suppose a person who logs onto a particular site in a shopping mall on the world wide web purchases an item is .20. If the site has 10 people accessing it in the next minute, what is the probability that

a) none of the individuals purchase an item?

b) exactly 2 individuals purchases an item?

c) at least 2 individuals purchase an item?

d) at most 2 individuals purchase an item?

Summary

This unit is extremely important from the point of view of many fascinating aspects of statistical inference that would follow in the subsequent units. Certainly, it is expected from you that you master the nitty-gritty of this unit. This unit specifically focuses on

- The definition, meaning and concepts of a probability distribution.
- The related terms-discrete random variable and continuous random variable.
- Discrete probability distribution and continuous probability distribution.
- The binomial distribution and its role in business problems.
- The Poisson distribution and its uses
- The normal distribution and its role in statistical inference.
- The concept of the standard normal distribution and its role.



3

JOINT PROBABILITY

A joint probability table is a table in which all possible events (or outcomes) for one variable are listed as row headings, all possible events for a second variable are listed as column headings, and the value entered in each cell of the table is the probability of each joint occurrence. Often the probabilities in such a table are based on observed frequencies of occurrence for the various joint events, rather than being a priori in nature. The table of joint occurrence frequencies which can serve as the basis for constructing a joint probability table is called a contingency table.

Table 1 is a contingency table which describes 200 people who entered a clothing store according to sex and age, while table 1b is the associated joint probability table. The frequency reported in each cell of the contingency table is converted into a probability value by dividing by the total number of observations, in this case, 200

1a contingency table for clothing store customers

Age	Sex		Total
	Male	Female	
Under 30	60	50	110
30 and over	80	10	90
Total	140	60	200

1b joint probability table for clothing store customers

Age	Sex		Total
	Male	Female	
under 30	0.3	0.25	0.55
30 and over	0.4	0.05	0.45
Marginal probability	0.7	0.3	1

In the context of joint probability tables, a marginal probability is so named because it is a marginal total of a row or a column. Where the probability values in the cells are probabilities of joint occurrence, the marginal probabilities are the unconditional or simple probabilities of particular events.

Table 2a is the contingency table which presents voter reactions to a new property tax plan according to party affiliation. a) prepare the joint probability table for these data. b) Determine the marginal probabilities and indicate what they mean.

Contingency table for voter reactions to a new property tax plan

party affiliation	reaction in			total
	favour	neutral	opposed	
Democratic(d)	120	20	20	160
Republican®	50	30	60	140
Independent(i)	50	10	40	100
total	220	60	120	400

See table 2b

Joint probability table for voter reactions to a new property tax plan

party affiliation	reaction in			Marginal probability
	favour	neutral	opposed	
Democratic(d)	.30	.05	.05	.40
Republican®	.125	.075	.15	.35
Independent(i)	.125	.025	.10	.25
total	.55	.15	.30	1.00

b) Each marginal probability value indicates the unconditional probability of the event identified as the column or row heading. For example, if a person is chosen randomly from this group of 400 voters, the probability that the person will be in favor of the tax plan is $p(f) = .55$. If a voter is chosen randomly, the probability that the voter is a republican is $p(r) = .35$

referring to the table, determine the following probabilities:

a) $p(o)$

b) $p(r \text{ and } o)$

c) $P(i)$

d) $p(l \text{ and } f)$

e) $p(o/r), (f)p(r/o)$

g) $p(r \text{ or } d)$

h) $p(d \text{ or } f)$

solution

a) = .30 (the marginal probability)

b) = .15 (joint probability)

c) = .25 (marginal probability)

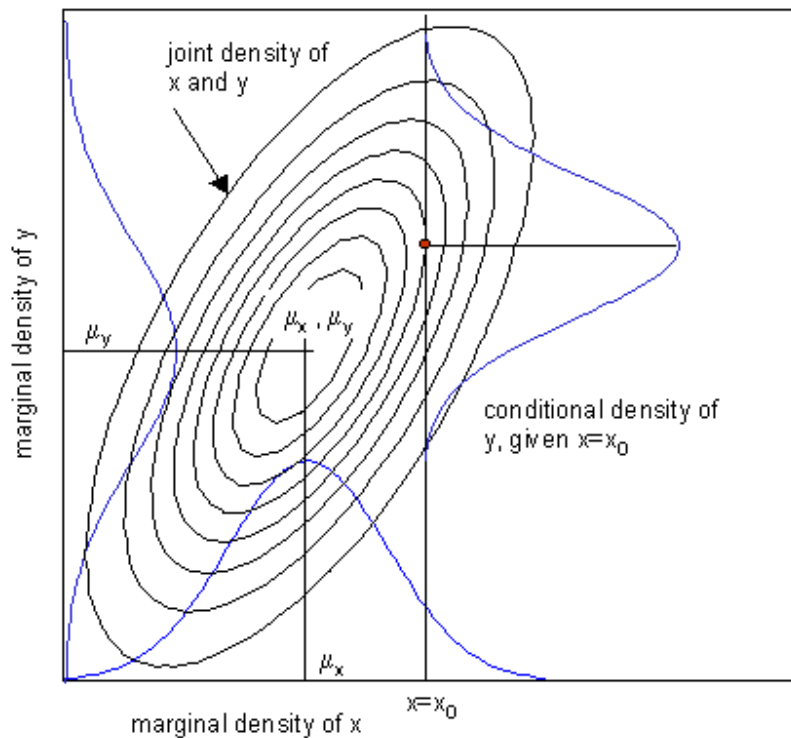
Joint probability is the probability of two or more things happening together. $f(x, y | q)$ where f is the probability of x and y together as a pair, given the distribution parameters, q . Often these events are not independent, and sadly this is often ignored. Furthermore, the correlation coefficient itself does NOT adequately describe these interrelationships.

Consider first the idea of a probability density or distribution: $f(x | q)$ where f is the probability density of x , given the distribution parameters, q . For a normal distribution, $q = (m, s^2)^T$ where m is the mean, and s is the standard deviation. This is sometimes called a pdf, probability density function. The integral of a pdf, the area under the curve (corresponding to the probability) between specified values of x , is a cdf, cumulative distribution function, $F(x | q)$. For discrete f , F is the corresponding summation.

A joint probability density two or more variables is called a multivariate distribution. It is often summarized by a vector of parameters, which may or may not be sufficient to characterize the distribution completely. Example, the normal is summarized (sufficiently) by a mean vector and covariance matrix.

marginal probability: $f(x | q)$ where f is the probability density of x , for all possible values of y , given the distribution parameters, q . The marginal probability is determined from the joint distribution of x and y by integrating over all values of y , called "integrating out" the variable y . In applications of Bayes's Theorem, y is often a matrix of possible parameter values. The figure illustrates Joint, marginal, and conditional probability.

- Schematic showing joint, marginal, and conditional densities



conditional probability: $f(x | y; q)$ where f is the probability of x by itself, given specific value of variable y , and the distribution parameters, q . (See Figure) If x and y represent events A and B , then $P(A|B) = n_{AB}/n_B$, where n_{AB} is the number of times both A and B occur, and n_B is the number of times B occurs. $P(A|B) = P(AB)/P(B)$, since $P(AB) = n_{AB}/N$ and $P(B) = n_B/N$ so that

$$P(A|B) = \frac{n_{AB}/N}{n_B/N} = n_{AB}/n_B$$

Note that in general the conditional probability of A given B is not the same as B given A . The probability of both A and B together is $P(AB)$, and $P(A|B) \times P(B) = P(AB) = P(B|A) \times P(A)$, if both $P(A)$ and $P(B)$ are non-zero. This leads to a statement of Bayes's Theorem:

$P(B|A) = P(A|B) \times P(B)/P(A)$. Conditional probability is also the basis for statistical dependence and independence.

independence: Two variables, A and B , are independent if their conditional probability is equal to their unconditional probability. In other words, A and B are independent if, and only if, $P(A|B)=P(A)$, and $P(B|A)=P(B)$. In engineering terms, A and B are independent if knowing something about one tells nothing about the other. This is the origin of the familiar, but often misused, formula $P(AB) = P(A) \times P(B)$, which is true only when A and B are independent.

conditional independence: A and B are conditionally independent, given C, if $\text{Prob}(A=a, B=b \mid C=c) = \text{Prob}(A=a \mid C=c) \times \text{Prob}(B=b \mid C=c)$ whenever $\text{Prob}(C=c) > 0$. So the joint probability of ABC, when A and B are conditionally independent, given C, is then $\text{Prob}(C) \times \text{Prob}(A \mid C) \times \text{Prob}(B \mid C)$. A directed graph illustrating this conditional independence is $A \rightarrow C \rightarrow B$.

Correlation, Regression

Introduction

At this point, you know the basics, how to look at data, compute and interpret probabilities draw a random sample, and to do statistical inference. Now it's a question of applying these concepts to see the relationships hidden within the more complex situations of real life. This unit shows you how statistics can summarize the relationships between two factors based on a bivariate data set with two columns of numbers. The correlation will tell you how strong the relationship is, and regression will help you predict one factor from the other.

Learning objectives

After reading this unit, you will be able to:

Define correlation coefficient with its properties

Calculate correlation coefficient and interpret

Appreciate the role of regression

Formulate the regression equation and use it for estimation and prediction.

Correlation analysis

Pearson's product-moment coefficient

Mathematical properties

The correlation coefficient $\rho_{X, Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

where E is the expected value operator and cov means covariance. Since $\mu_X = E(X)$, $\sigma_X^2 = E(X^2) - E^2(X)$ and likewise for Y , we may also write

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value.

The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1 , the stronger the correlation between the variables.

If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Here is an example: Suppose the random variable X is uniformly distributed on the interval from -1 to 1 , and $Y = X^2$. Then Y is completely determined by X , so that X and Y are dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

A correlation between two variables is diluted in the presence of measurement error around estimates of one or both variables, in which case disattenuation provides a more accurate coefficient.

The sample correlation

If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \dots, n$, then the Pearson product-moment correlation coefficient can be used to estimate the correlation of X and Y . The Pearson coefficient is also known as the "sample correlation coefficient". The Pearson correlation coefficient is then the best estimate of the correlation of X and Y . The Pearson correlation coefficient is written:

where \bar{x} and \bar{y} are the sample means of X and Y , s_x and s_y are the sample standard deviations of X and Y and the sum is from $i = 1$ to n . As with the population correlation, we may rewrite this as

Again, as is true with the population correlation, the absolute value of the sample correlation must be less than or equal to 1. Though the above formula conveniently suggests a single-pass algorithm for calculating sample correlations, it is notorious for its numerical instability (see below for something more accurate).

The square of the sample correlation coefficient, which is also known as the coefficient of determination, is the fraction of the variance in y_i that is accounted for by a linear fit of x_i to y_i . This is written

where $s_{y|x}^2$ is the square of the error of a linear regression of x_i on y_i by the equation $y = a + bx$:

and s_y^2 is just the variance of y :

Note that since the sample correlation coefficient is symmetric in x_i and y_i , we will get the same value for a fit of y_i to x_i :

This equation also gives an intuitive idea of the correlation coefficient for higher dimensions. Just as the above described sample correlation coefficient is the fraction of variance accounted for by the fit of a 1-dimensional linear submanifold to a set of 2-dimensional vectors (x_i, y_i) , so we can define a correlation coefficient for a fit of an m -dimensional linear submanifold to a set of n -dimensional vectors. For example, if we fit a plane $z = a + bx + cy$ to a set of data (x_i, y_i, z_i) then the correlation coefficient of z to x and y is

The distribution of the correlation coefficient has been examined by R. A. Fisher^{[1][2]} and A. K. Gayen.^[3]

Geometric Interpretation of correlation

The correlation coefficient can also be viewed as the cosine of the angle between the two vectors of samples drawn from the two random variables.

Caution: This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero. Some practitioners prefer an uncentered (non-Pearson-compliant) correlation coefficient. See the example below for a comparison.

As an example, suppose five countries are found to have gross national products of 1, 2, 3, 5, and 8 billion dollars, respectively. Suppose these same five countries (in the same order) are found to have 11%, 12%, 13%, 15%, and 18% poverty. Then let \mathbf{x} and \mathbf{y} be ordered 5-element vectors containing the above data: $\mathbf{x} = (1, 2, 3, 5, 8)$ and $\mathbf{y} = (0.11, 0.12, 0.13, 0.15, 0.18)$.

By the usual procedure for finding the angle between two vectors (see dot product), the *uncentered* correlation coefficient is:

Note that the above data were deliberately chosen to be perfectly correlated: $y = 0.10 + 0.01 x$. The Pearson correlation coefficient must therefore be exactly one. Centering the data (shifting \mathbf{x} by $E(\mathbf{x}) = 3.8$ and \mathbf{y} by $E(\mathbf{y}) = 0.138$) yields $\mathbf{x} = (-2.8, -1.8, -0.8, 1.2, 4.2)$ and $\mathbf{y} = (-0.028, -0.018, -0.008, 0.012, 0.042)$, from which

as expected.

Motivation for the form of the coefficient of correlation

Another motivation for correlation comes from inspecting the method of simple linear regression. As above, X is the vector of independent variables, x_i , and Y of the dependent variables, y_i , and a simple linear relationship between X and Y is sought, through a least-squares method on the estimate of Y :

Then, the equation of the least-squares line can be derived to be of the form:

which can be rearranged in the form:

where r has the familiar form mentioned above :

Interpretation of the size of a correlation

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988),^[4] for example, has suggested the following interpretations for correlations in psychological research, in the table on the right.

Correlation	Negative	Positive
Small	-0.29 to -0.10	0.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Large	-1.00 to -0.50	0.50 to 1.00

As Cohen himself has observed, however, all such criteria are in some ways arbitrary and should not be observed too strictly. This is because the interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

Along this vein, it is important to remember that "large" and "small" should not be taken as synonyms for "good" and "bad" in terms of determining that a correlation is of a certain size. For example, a correlation of 1.00 or -1.00 indicates that the two variables analyzed are equivalent modulo scaling. Scientifically, this more frequently indicates a trivial result than an earth-shattering one. For example, consider discovering a correlation of 1.00 between how many feet tall a group of people are and the number of inches from the bottom of their feet to the top of their heads.

Non-parametric correlation coefficients

Pearson's correlation coefficient is a parametric statistic and when distributions are not normal it may be less useful than non-parametric correlation methods, such as Chi-square, Point biserial correlation, Spearman's ρ and Kendall's τ . They are a little less

powerful than parametric methods if the assumptions underlying the latter are met, but are less likely to give distorted results when the assumptions fail.

Other measures of dependence among random variables

To get a measure for more general dependencies in the data (also nonlinear) it is better to use the correlation ratio which is able to detect almost any functional dependency, or mutual information/total correlation which is capable of detecting even more general dependencies.

The polychoric correlation is another correlation applied to ordinal data that aims to estimate the correlation between theorised latent variables.

Copulas and correlation

The information given by a correlation coefficient is not enough to define the dependence structure between random variables; to fully capture it we must consider a copula between them. The correlation coefficient completely defines the dependence structure only in very particular cases, for example when the cumulative distribution functions are the multivariate normal distributions. In the case of elliptic distributions it characterizes the (hyper-)ellipses of equal density, however, it does not completely characterize the dependence structure (for example, the a multivariate t-distribution's degrees of freedom determine the level of tail dependence).

Correlation matrices

The correlation matrix of n random variables X_1, \dots, X_n is the $n \times n$ matrix whose i,j entry is $\text{corr}(X_i, X_j)$. If the measures of correlation used are product-moment coefficients, the correlation matrix is the same as the covariance matrix of the standardized random variables $X_i / \text{SD}(X_i)$ for $i = 1, \dots, n$. Consequently it is necessarily a positive-semidefinite matrix.

The correlation matrix is symmetric because the correlation between X_i and X_j is the same as the correlation between X_j and X_i .

Removing correlation

It is always possible to remove the correlation between zero-mean random variables with a linear transform, even if the relationship between the variables is nonlinear. Suppose a vector of n random variables is sampled m times. Let X be a matrix where $X_{i,j}$ is the j th variable of sample i . Let $Z_{r,c}$ be an r by c matrix with every element

1. Then D is the data transformed so every random variable has zero mean, and T is the data transformed so all variables have zero mean, unit variance, and zero correlation with all other variables. The transformed variables will be uncorrelated, even though they may not be independent.

where an exponent of $-1/2$ represents the matrix square root of the inverse of a matrix. The covariance matrix of T will be the identity matrix. If a new data sample x is a row vector of n elements, then the same transform can be applied to x to get the transformed vectors d and t .

Common misconceptions about correlation

Correlation and causality

The conventional dictum that "correlation does not imply causation" means that correlation cannot be validly used to infer a causal relationship between the variables. This dictum should not be taken to mean that correlations cannot indicate causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).

Here is a simple example: hot weather may cause both crime and ice-cream purchases. Therefore crime is correlated with ice-cream purchases. But crime does not cause ice-cream purchases and ice-cream purchases do not cause crime.

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health? Or does good health lead to good mood? Or does some other factor underlie both? Or is it pure coincidence? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

Correlation and linearity

Four sets of data with the same correlation of 0.81

While Pearson correlation indicates the strength of a linear relationship between two variables, its value alone may not be sufficient to evaluate this relationship, especially in the case where the assumption of normality is incorrect.

The image on the right shows scatterplots of Anscombe's quartet, a set of four different pairs of variables created by Francis Anscombe.^[5] The four y variables have the same mean (7.5), standard deviation (4.12), correlation (0.81) and regression line ($y = 3 + 0.5x$). However, as can be seen on the plots, the distribution of the variables is very different. The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality. The second one (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant. In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.81. Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

These examples indicate that the correlation coefficient, as a summary statistic, cannot replace the individual examination of the data.

Computing correlation accurately in a single pass

The following algorithm (in pseudocode) will estimate correlation with good numerical stability

```

sum_sq_x = 0
sum_sq_y = 0
sum_coproduct = 0
mean_x = x[1]
mean_y = y[1]
for i in 2 to N:
    sweep = (i - 1.0) / i
    delta_x = x[i] - mean_x
    delta_y = y[i] - mean_y
    sum_sq_x += delta_x * delta_x * sweep
    sum_sq_y += delta_y * delta_y * sweep
    sum_coproduct += delta_x * delta_y * sweep
    mean_x += delta_x / i
    mean_y += delta_y / i
pop_sd_x = sqrt( sum_sq_x / N )
pop_sd_y = sqrt( sum_sq_y / N )
cov_x_y = sum_coproduct / N

```


$\text{correlation} = \text{cov_x_y} / (\text{pop_sd_x} * \text{pop_sd_y})$

For an enlightening experiment, check the correlation of $\{900,000,000 + i \text{ for } i=1\dots 100\}$ with $\{900,000,000 - i \text{ for } i=1\dots 100\}$, perhaps with a few values modified. Poor algorithms will fail.

Autocorrelation

A plot showing 100 random numbers with a "hidden" sine function, and an autocorrelation (correlogram) of the series on the bottom.

Autocorrelation is a mathematical tool used frequently in signal processing for analyzing functions or series of values, such as time domain signals. Informally, it is the strength of a relationship between observations as a function of the time separation between them. More precisely, it is the cross-correlation of a signal with itself. Autocorrelation is useful for finding repeating patterns in a signal, such as determining the presence of a periodic signal which has been buried under noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.

Uses of correlation

The uses of correlation are as follows:

Economic theory and business studies show relationship between variable like price and quantity demanded, advertising expenditure and sales promotion measures etc.

Correlation analysis helps in deriving precisely the degree and the direction of such relationships.

The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.

Regression analysis

In regression analysis using time series data, autocorrelation of the residuals ("error terms", in econometrics) is a problem.

Autocorrelation violates the OLS assumption that the error terms are uncorrelated. While it does not bias the OLS coefficient estimates, the standard errors tend to be underestimated (and the t-scores overestimated).

The traditional test for the presence of first-order autocorrelation is the Durbin–Watson statistic or, if the explanatory variables include a lagged dependent variable, Durbin's h statistic. A more flexible test, covering autocorrelation of higher orders and applicable whether or not the regressors include lags of the dependent variable, is the Breusch–Godfrey test. This involves an auxiliary regression, wherein the residuals obtained from estimating the model of interest are regressed on (a) the original regressors and (b) k lags of the residuals, where k is the order of the test. The simplest version of the test statistic from this auxiliary regression is TR^2 , where T is the sample size and R^2 is the coefficient of determination. Under the null hypothesis of no autocorrelation, this statistic is asymptotically distributed as χ^2 with k degrees of freedom.

Responses to nonzero autocorrelation include generalized least squares and Newey–West standard errors.

Applications

- One application of autocorrelation is the measurement of optical spectra and the measurement of very-short-duration light pulses produced by lasers, both using optical autocorrelators.
- In optics, normalized autocorrelations and cross-correlations give the degree of coherence of an electromagnetic field.
- In signal processing, autocorrelation can give information about repeating events like musical beats or pulsar frequencies, though it cannot tell the position in time of the beat. It can also be used to estimate the pitch of a musical tone.

Correlation Coefficient

How well does your regression equation truly represent your set of data?

One of the ways to determine the answer to this question is to exam the *correlation coefficient* and the *coefficient of determination*.

```

LinReg
y=ax+b
a=1.690909091
b=.2727272727
r²=.9701626472
r=.9849683483

```

The correlation coefficient, r , and the coefficient of determination, r^2 , will appear on the screen that shows the regression equation information (be sure the Diagnostics are turned on - -- 2nd Catalog (above 0), arrow down to DiagnosticOn, press ENTER twice.)

In addition to appearing with the regression information, the values r and r^2 can be found under VARS, #5 Statistics → EQ #7 r and #8 r^2 .

Correlation Coefficient, r :

- ✦ The quantity r , called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson.

- ✦ The mathematical formula for computing r is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data.

(Aren't you glad you have a graphing calculator that computes this formula?)

- ✦ The value of r is such that $-1 < r < +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

- ✦ **Positive correlation:** If x and y have a strong positive linear correlation, r is close

to +1. An r value of exactly +1 indicates a perfect positive fit.

Positive values

indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

- ✦ **Negative correlation:** If x and y have a strong negative linear correlation, r is close

to -1. An r value of exactly -1 indicates a perfect negative fit.

Negative values

indicate a relationship between x and y such that as values for x increase, values for y decrease.

- ✦ **No correlation:** If there is no linear correlation or a weak linear correlation, r is

close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

✦ Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

✦ A **perfect correlation** of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.

✦ A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination, r^2 or R^2 :

✦ The *coefficient of determination*, r^2 , is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable.

It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

✦ The *coefficient of determination* is the ratio of the explained variation to the total variation.

✦ The *coefficient of determination* is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y .

✦ The *coefficient of determination* **represents the percent of the data that is the closest to the line of best fit.** For example, if $r = 0.922$, then $r^2 = 0.850$, which means that

85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

✦ The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The

further the line is
away from the points, the less it is able to explain.

Correlation Coefficient

A correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value, the other has a low (high) value. A correlation coefficient of 0 means that there is no linear relationship between the variables.

There are a number of different correlation coefficients that might be appropriate depending on the kinds of variables being studied.

Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient, usually denoted by r , is one example of a correlation coefficient. It is a measure of the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height in inches and weight in pounds. However, it can be misleadingly small when there is a relationship between the variables but it is a non-linear one.

There are procedures, based on r , for making inferences about the population correlation coefficient. However, these make the implicit assumption that the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman Rank Correlation Coefficient might be more appropriate.

Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient is one example of a correlation coefficient. It is usually calculated on occasions when it is not convenient, economic, or even possible to give actual values to variables, but only to assign a rank order to instances of each variable. It may also be a better indicator that a relationship exists between two variables when the relationship is non-linear.

Commonly used procedures, based on the Pearson's Product Moment Correlation Coefficient, for making inferences about the population correlation coefficient make the implicit assumption that

the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman Rank Correlation Coefficient might be more appropriate.

have outliers, Pearson's correlation coefficient will be greatly affected. Also, Pearson's correlation coefficient only measures linear relationships between variables. There is another alternative. *Spearman's rank correlation coefficient* r_s does not use the actual observed data, but the *ranks* of the data, to compute a correlation coefficient. That is, replace the smallest X value with a 1, the next smallest with a 2, and so on. Repeat the same procedure for the Y values. Then instead of having our data (X, Y) in the form

$$(12.3, 2.7) \quad (10.4, 3.2) \quad (13.2, 3.0),$$

they will be as follows

$$(2, 1) \quad (1, 3) \quad (3, 2).$$

The formula for r_s is as follows,

$$r_s = \frac{\sum (\text{rank}_x - (n+1)/2) (\text{rank}_y - (n+1)/2)}{n(n-1)(n+1)/12}.$$

Actually, this is just Pearson's formula applied to the ranks.

Correlation Coefficients: Examples

Spearman's Rank Correlation Coefficient

In calculating this coefficient, we use the Greek letter 'rho' or ρ . The formula used to calculate this coefficient is:

$$r = 1 - (6 \sum d^2) / (n(n^2 - 1))$$

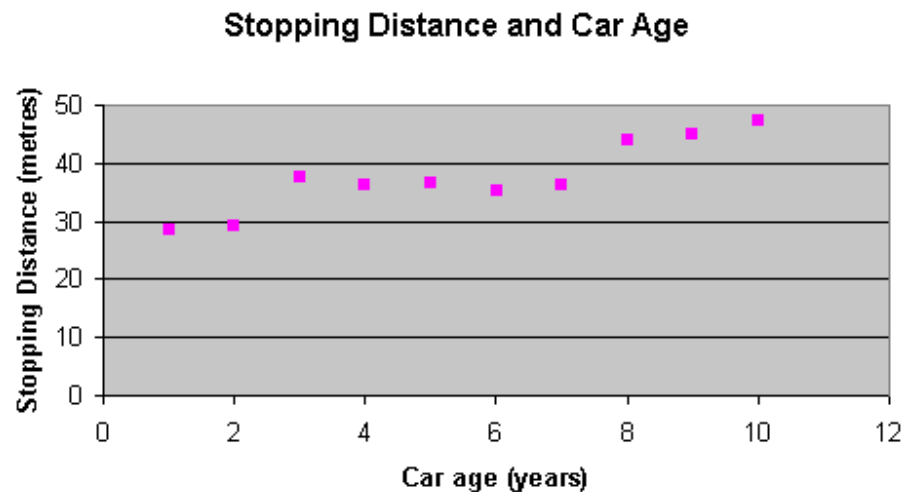
To illustrate this, consider the following worked example: Researchers at the European Centre for Road Safety Testing are trying to find out how the age of cars affects their braking capability. They test a group of ten cars of differing ages and find out the minimum stopping distances that the cars can achieve. The results are set out in the table below:

Table 1: Car ages and stopping distances

Car	Age (months)	Minimum Stopping at 40 kph (metres)
A	9	28.4

B	15	29.3
C	24	37.6
D	30	36.2
E	38	36.5
F	46	35.3
G	53	36.2
H	60	44.1
I	64	44.8
J	76	47.2

These figures form the basis for the scatter diagram, below, which shows a reasonably strong positive correlation - the older the car, the longer the stopping distance.



Graph 1: Car age and Stopping distance (data from Table 1 above)

To process this information we must, firstly, place the ten pieces of data into order, or rank them according to their age and ability to stop. It is then possible to process these ranks.

Table 2: Ranked data from Table 1 above

Car	Age (months)	Minimum Stopping at 40 kph (metres)	Age rank	Stopping rank
A	9	28.4	1	1
B	15	29.3	2	2
C	24	37.6	3	7

D	30	36.2	4	4.5
E	38	36.5	5	6
F	46	35.3	6	3
G	53	36.2	7	4.5
H	60	44.1	8	8
I	64	44.8	9	9
J	76	47.2	10	10

Notice that the ranking is done here in such a way that the youngest car and the best stopping performance are rated top and vice versa. There is no strict rule here other than the need to be consistent in your rankings. Notice also that there were two values the same in terms of the stopping performance of the cars tested. They occupy 'tied ranks' and must share, in this case, ranks 4 and 5. This means they are each ranked as 4.5, which is the mean average of the two ranking places. It is important to remember that this works despite the number of items sharing tied ranks. For instance, if five items shared ranks 5, 6, 7, 8 and 9, then they would each be ranked 7 - the mean of the tied ranks.

Now we can start to process these ranks to produce the following table:

Table 3: Differential analysis of data from Table 2

Car	Age (mths)	Stopping distance	Age rank	Stopping rank	d	d ²
A	9	28.4	1	1	0	0
B	15	29.3	2	2	0	0
C	24	37.6	3	7	4	16
D	30	36.2	4	4.5	0.5	0.25
E	38	36.5	5	6	1	1
F	46	35.3	6	3	-3	9
G	53	36.2	7	4.5	-2.5	6.25
H	60	44.1	8	8	0	0
I	64	44.8	9	9	0	0
J	76	47.2	10	10	0	0
					$\sum d^2$	32.5

Note that the two extra columns introduced into the new table are Column 6, 'd', the difference between stopping distance rank and

squared figures are summed at the foot of Column 7. Calculation of Spearman Rank Correlation Coefficient (r) is:

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Number in sample (n) = 10

$$r = 1 - \frac{6 \times 32.5}{10(10 \times 10 - 1)}$$

$$r = 1 - \frac{195}{10 \times 99}$$

$$r = 1 - 0.197$$

$$r = 0.803$$

What does this tell us? When interpreting the Spearman Rank Correlation Coefficient, it is usually enough to say that:

- for values of r of 0.9 to 1, the correlation is very strong.
- for values between 0.7 and 0.9, correlation is strong.
- and for values between 0.5 and 0.7, correlation is moderate.

This is the case whether r is positive or negative. In our case of car ages and stopping distance performance, we can say that there is a strong correlation between the two variables.

Pearson's or Product-Moment Correlation Coefficient

The Pearson Correlation Coefficient is denoted by the symbol r. Its formula is based on the standard deviations of the x-values and the y-values:

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Going back to the original data we recorded from the European Centre for Road Safety Testing, the calculation needed for us to work out the Product-Moment Correlation Coefficient is best set out as in the table that follows.

Note that in the table below,
 x = age of car
 y = stopping distance

From this, the other notation should be obvious.

	x	y	x ²	y ²	xy
	9	28.4	81	806.56	255.6
	15	29.3	225	858.49	439.5
	24	37.6	576	1413.76	902.4

	30	36.2	900	1310.44	1086
	38	36.5	1444	1332.25	1387
	46	35.3	2116	1246.09	1623.8
	53	36.2	2809	1310.44	1918.6
	60	44.1	3600	1944.81	2646
	64	44.8	4096	2007.04	2867.2
	76	47.2	5776	2227.84	3587.2
Totals	415	375.6	21623	14457.72	16713.3

$$\begin{aligned} \bar{x} &= 415/10 = 41.5 \\ \bar{y} &= 375.6/10 = 37.7 \end{aligned}$$

$$\begin{aligned} r &= \frac{10 \times 16713.3 - 415 \times 375.6}{\sqrt{(10 \times 21623 - 415^2)(10 \times 14457.72 - 375.6^2)}} \\ r &= \frac{11259}{11259} \times \frac{3501.84}{124.14} \\ r &= 0.91 \end{aligned}$$

What does this tell us?

To interpret the value of r you need to follow these guidelines:

- r always lies in the range -1 to +1. If it lies close to either of these two values, then the dispersion of the scattergram points is small and therefore a strong correlation exists between the two variables.
- For r to equal exactly -1 or +1 must mean that correlation is perfect and all the points on the scattergram lie on the line of best fit (otherwise known as the regression line.) If r is close to 0, the dispersion is large and the variables are uncorrelated. The positive or negative sign on the value of r indicates positive or negative correlation.

So in the above case, there is evidence of strong positive correlation between stopping distance and age of car; in other words, the older the car, the longer the distance we could expect it to take to stop.

Illustration:

Let's say that we want to track the progress of a group of new employees of a large service organisation. We think we can judge the effectiveness of our induction and initial training scheme by analysing employee competence in weeks one, four and at the end of the six months.

Let's say that Human Resource managers in their organisation have

been urging the company to commit more resources to induction and basic training. The company now wishes to know which of the two assessments - the new employee's skills on entry or after week four - provides a better guide to the employee's performance after six months. Although there is a small sample here, let's assume that it is accurate.

The raw data is given in the table below:

Name	Skills on entry % score	Skills at week 4 % score	Skills at 6 mths % score
ab	75	75	75
bc	72	69	76
cd	82	76	83
de	78	77	65
ef	86	79	85
fg	76	65	79
gh	86	82	65
hi	89	78	75
ij	83	70	80
jk	65	71	70

Copy this information onto a fresh Excel worksheet, putting the names in Column A, the entry test results in Column B, the week four test Marks in Column D, and the six month test scores in Column F.

When you have entered the information, select the three number columns (do not include any cells with words in them). Go to the Data Analysis option on the Tools menu, select from that Data Analysis menu the item Correlation (note that if the Data Analysis option is not on the Tools menu you have to add it in).

When you get the Correlation menu, enter in the first Input Range box the column of cells containing the dependent variables you wish to analyze (D3 to D12 if your spreadsheet looks like TimeWeb's). Next, enter into the second input box the column of cells that contain the independent variables (B3 to B12, again if your sheet resembles TimeWeb's).

Then click the mouse pointer in the circle to the left of the Output Range label (unless there is a black dot in it already), and click the left mouse button in the Output Range box. Then enter the name of cell where you want the top left corner of the correlation table to appear (e.g., \$A\$14). Then click OK.

After a second or two, the Correlation Table should appear giving you the correlation between all the different pairs of data. We are interested in the correlation between Column B (the first column in the Table) and Column D (the third column in the table). The correlation between Column C (the second column in the Table) and Column D, can be approached in the same way.

Which of these two is the better predictor of success according to this study. How reliable is it?

Expected Answer:

The correlation between the Entry Mark and the Final Mark is 0.23; the correlation between the four week test and the Final Mark is 0.28. Thus, both of the tests have a positive correlation to the Final (6 month) Test; the entry test has a slightly weaker positive correlation with the Final Mark, than the Four Week Test. However, both figures are so low, that the correlation is minimal. The skills measured by the Entry test account for about 5 per cent of the skills measured by the Six Month Mark. This figure is obtained by using the R-Squared result and expressing it as a percentage.

Beware!

It's vital to remember that a correlation, even a very strong one, does not mean we can make a conclusion about causation. If, for example, we find a very high correlation between the weight of a baby at birth and educational achievement at age 25, we may make some predictions about the numbers of people staying on at university to study for post-graduate qualifications. Or we may urge mothers-to-be to take steps to boost the weight of the unborn baby, because the heavier their baby the higher their baby's educational potential, but we should be aware that the correlation, in itself, is no proof of these assertions.

This is a really important principle: correlation is not necessarily proof of causation. It indicates a relationship which may be based on cause and effect, but then again, it may not be. If weight at birth is a major cause of academic achievement, then we can expect that variations in birth weight will cause changes in achievement. The reverse, however is not necessarily true. If any two variables are correlated, we cannot automatically assume that one is the cause of the other.

The point of causation is best illustrated perhaps, using the example of AIDS.

A very high correlation exists between HIV infection and cases of

AIDS. This has caused many researchers to believe that HIV is the principal cause of AIDS. This belief has led to most of the money for AIDS research going into investigating HIV.

But the cause of AIDS is still not clear. Some people (especially, not surprisingly, those suffering from AIDS) have argued vehemently that investigating HIV instead of AIDS is a mistake. They say that something else is the real cause. This is the area, they argue, that requires greater research funding. More money should be going into AIDS research rather than studies into HIV.

Correlation and Linearity

Correlation coefficients measure the strength of association between two variables. The most common correlation coefficient, called the **Pearson product-moment correlation coefficient**, measures the strength of the *linear association* between variables.

In this tutorial, when we speak simply of a correlation coefficient, we are referring to the Pearson product-moment correlation.

How to Calculate a Correlation Coefficient

A formula for computing a sample correlation coefficient (r) is given below.

Sample correlation coefficient. The correlation r between two variables is:

$$r = [1 / (n - 1)] * \Sigma \{ [(x_i - \bar{x}) / s_x] * [(y_i - \bar{y}) / s_y] \}$$

where n is the number of observations in the sample, Σ is the summation symbol, x_i is the x value for observation i , \bar{x} is the mean x value, y_i is the y value for observation i , \bar{y} is the mean y value, s_x is the sample standard deviation of x , and s_y is the sample standard deviation of y .

A formula for computing a population correlation coefficient (ρ) is given below.

Population correlation coefficient. The correlation ρ between two variables is:

$$\rho = [1 / N] * \Sigma \{ [(X_i - \mu_x) / \sigma_x] * [(Y_i - \mu_y) / \sigma_y] \}$$

where N is the number of observations in the population, Σ is the summation symbol, X_i is the X value for observation i , μ_x is the

population mean for variable X , Y_i is the Y value for observation i , μ_Y is the population mean for variable Y , σ_x is the standard deviation of X , and σ_y is the standard deviation of Y .

Fortunately, you will rarely have to compute a correlation coefficient by hand. Many software packages (e.g., Excel) and most graphing calculators have a correlation function that will do the job for you.

Note: Sometimes, it is not clear whether a software package or a graphing calculator uses a population correlation coefficient or a sample correlation coefficient. For example, a casual user might not realize that Microsoft uses a population correlation coefficient (ρ) for the `Pearson()` function in its Excel software.

How to Interpret a Correlation Coefficient

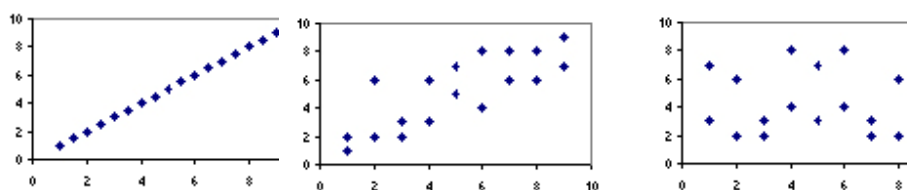
The sign and the absolute value of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.

- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of a correlation coefficient, the stronger the *linear* relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
- The weakest linear relationship is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

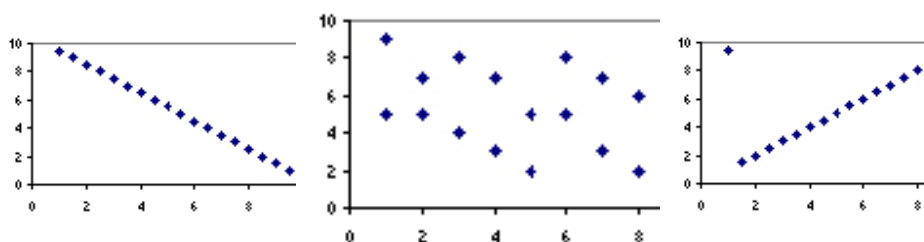
Keep in mind that the Pearson product-moment correlation coefficient only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero *linear* relationship. (It is possible for two variables to have zero linear relationship and a strong curvilinear relationship at the same time.)

Scatterplots and Correlation Coefficients

The scatterplots below show how different patterns of data produce different degrees of correlation.



Maximum positive correlation (r = 1.0) **Strong positive correlation (r = 0.80)** **Zero correlation (r = 0)**



Minimum negative correlation (r = -1.0) **Moderate negative correlation (r = -0.43)** **Strong negative correlation with outlier (r = 0.71)**

Several points are evident from the scatterplots.

- When the slope of the line in the plot is negative, the correlation is negative; and vice versa.
- The strongest correlations ($r = 1.0$ and $r = -1.0$) occur when data points fall *exactly* on a straight line.
- The correlation becomes weaker as the data points become more scattered.
- If the data points fall in a random pattern, the correlation is equal to zero.
- Correlation is affected by outliers. Compare the first scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

Test Your Understanding of This Lesson

Problem 1

A national consumer magazine reported the following correlations.

- The correlation between car weight and car reliability is - 0.30.
- The correlation between car weight and annual maintenance cost is 0.20.

Which of the following statements are true?

- I. Heavier cars tend to be less reliable.
- II. Heavier cars tend to cost more to maintain.
- III. Car weight is related more strongly to reliability than to maintenance cost.

- (A) I only
- (B) II only
- (C) III only
- (D) I and II only
- (E) I, II, and III

Solution

The correct answer is (E). The correlation between car weight and reliability is negative. This means that reliability tends to decrease as car weight increases. The correlation between car weight and maintenance cost is positive. This means that maintenance costs tend to increase as car weight increases.

The strength of a relationship between two variables is indicated by the absolute value of the correlation coefficient. The correlation between car weight and reliability has an absolute value of 0.30. The correlation between car weight and maintenance cost has an absolute value of 0.20. Therefore, the relationship between car weight and reliability is stronger than the relationship between car weight and maintenance cost.

Least Squares

The method of least squares is a criterion for fitting a specified model to observed data. For example, it is the most commonly used method of defining a straight line through a set of points on a scatterplot.

Least Squares Line

Recall the equation of a line from algebra:

$$Y = b_0 + b_1X.$$

(You may have seen $Y = mX + b$, we are going to change notation slightly.) Above, b_1 is called the *slope* of the line and b_0 is the *y-intercept*. The slope measures the amount Y increases when X increases by one unit. The Y -intercept is the value of Y when $X = 0$.

Our objective is to fit a straight line to points on a scatterplot that do not lie along a straight line (see the figure above). So we want to find b_0 and b_1 such that the line $Y = b_0 + b_1X$ fits the data as well as

possible. First, we need to define what we mean by a "best" fit. We want a line that is in some sense closest to all of the data points simultaneously. In statistics, we define a *residual*, e_i , as the vertical distance between a point and the line,

$$e_i = Y_i - (b_0 + b_1 X_i).$$

(see the vertical line in the figure) Since residuals can be positive or negative, we will square them to remove the sign. By adding up all of the squared residuals, we get a measure of how far away from the data our line is. Thus, the "best" line will be one which has the minimum sum of squared residuals, i.e., $\min (\sum_i e_i^2)$. This method of finding a line is called *least squares*.

The formulas for the slope and intercept of the least squares line are

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Using algebra, we can express the slope b_1 as

$$b_1 = r \left(\frac{s_y}{s_x} \right).$$

Least Squares Linear Regression

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable Y , based on the value of an independent variable X .

In this tutorial, we focus on the case where there is only one independent variable. This is called simple regression (as opposed to multiple regression, which handles two or more independent variables).

Tip: The next lesson presents a simple regression example that shows how to apply the material covered in this lesson. Since this lesson is a little dense, you may benefit by also reading the next lesson.

Prediction

Given a least squares line, we can use it for prediction. The equation for prediction is simply the equation for a line with b_0 and b_1 replaced by their estimates. The predicted value of y is traditionally

denoted \hat{y} ("y-hat"). Thus, suppose we are given the least squares equation

$$\hat{y} = 64.93 + 0.635x,$$

where x is the age of a child in months and y is the height of that child, and let's further assume that the range of x is from 1 to 24 months. To predict the height of an 18 month old child, we just plug in to get

$$\begin{aligned}\hat{y} &= 64.93 + 0.635(18) \\ &= 76.36.\end{aligned}$$

What if we wanted to know the height of a child at age 32 months? From our least squares equation, we could get a prediction. However, we're predicting outside of the range of our x values. This is called *extrapolation* and is not "legal" in good statistics unless you are very sure that the line is valid. When we predict within the range of our x values, this is known as *interpolation*; this is the way we want to predict.

Prerequisites for Regression

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable Y has a linear relationship to the independent variable X . To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern.
- For each value of X , the probability distribution of Y has the same standard deviation σ . When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X , which is easily checked in a residual plot.
- For any given value of X ,
 - The Y values are independent, as indicated by a random pattern on the residual plot.
 - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is:

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and \hat{y} is the *predicted* value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator - to find b_0 and b_1 . You enter the X and Y values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sum (x_i - \bar{x})^2} \quad b_1 = r \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y , x_i is the X value of observation i , y_i is the Y value of observation i , \bar{x} is the mean of X , \bar{y} is the mean of Y , s_x is the standard deviation of X , and s_y is the standard deviation of Y .

Properties of the Regression Line

When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties.

- The line minimizes the sum of squared differences between observed values (the y values) and predicted values (the \hat{y} values computed from the regression equation).
- The regression line passes through the mean of the X values (\bar{x}) and the mean of the Y values (\bar{y}).
- The regression constant (b_0) is equal to the y intercept of the regression line.
- The regression coefficient (b_1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X ; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \left\{ \left(\frac{1}{N} \right) \sum [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \right\}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i , \bar{x} is the mean x value, y_i is the y value for observation i , \bar{y} is the mean y value, σ_x is the standard deviation of x , and σ_y is the standard deviation of y .

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Test Your Understanding of This Lesson

Problem 1

A researcher uses a regression equation to predict home heating bills (dollar cost), based on home size (square feet). The correlation between predicted bills and home size is 0.70. What is the correct interpretation of this finding?

- (A) 70% of the variability in home heating bills can be explained by home size.
- (B) 49% of the variability in home heating bills can be explained by home size.
- (C) For each added square foot of home size, heating bills increased by 70 cents.
- (D) For each added square foot of home size, heating bills increased by 49 cents.
- (E) None of the above.

Solution

The correct answer is (B). The coefficient of determination measures the proportion of variation in the dependent variable that is predictable from the independent variable. The coefficient of determination is equal to R^2 ; in this case, $(0.70)^2$ or 0.49. Therefore, 49% of the variability in heating bills can be explained by home size.

Regression Equation

A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

A linear regression equation is usually written

$$Y = a + bX + e$$

where

Y	is	the	dependent	variable
a	is		the	intercept
b	is	the	slope	or regression coefficient

X is the independent variable (or covariate)
e is the error term

The equation will specify the average magnitude of the expected change in Y given a change in X.

The regression equation is often represented on a scatterplot by a regression line.

A Simple Regression Example

In this lesson, we show how to apply regression analysis to some fictitious data, and we show how to interpret the results of our analysis.

Note: Regression computations are usually handled by a software package or a graphing calculator. For this example, however, we will do the computations "manually", since the gory details have educational value.

Problem Statement

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

How to Find the Regression Equation

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
---------	-------	-------	-------------------	-------------------	---------------------	---------------------	----------------------------------

1	95	85	17	8	289	64	136
2	85	95	7	18	49	324	126
3	80	70	2	-7	4	49	-14
4	70	65	-8	-12	64	144	96
5	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1 . Computations are shown below.

$b_1 = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sum [(x_i - \bar{x})^2]}$ $b_1 = 470/730 = 0.644$	$b_0 = \bar{y} - b_1 * \bar{x}$ $b_0 = 77 - (0.644)(78) = 26.768$
--	---

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

How to Use the Regression Equation

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (x), perform the computation, and you have an estimated value (\hat{y}) for the dependent variable.

In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade would be:

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80 = 26.768 + 51.52 = 78.288$$

Warning: When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimates.

In this example, the aptitude test scores used to create the regression equation ranged from 60 to 95. Therefore, only use values inside that range to estimate statistics grades. Using values outside that range (less than 60 or greater than 95) is problematic.

How to Find the Coefficient of Determination

Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the coefficient of determination, which can be computed from the following formula.

$$R^2 = \{ (1 / N) * \sum [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, \sum is the summation symbol, x_i is the x value for observation i, \bar{x} is the mean x value, y_i is the y value for observation i, \bar{y} is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y. Computations for the sample problem of this lesson are shown below.

$\sigma_x = \text{sqrt} [\sum (x_i - \bar{x})^2 / N]$ $\sigma_x = \text{sqrt}(730/5) = \text{sqrt}(146) = 12.083$	$\sigma_y = \text{sqrt} [\sum (y_i - \bar{y})^2 / N]$ $\sigma_y = \text{sqrt}(630/5) = \text{sqrt}(126) = 11.225$
$R^2 = \{ (1 / N) * \sum [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \}^2$ $R^2 = [(1/5) * 470 / (12.083 * 11.225)]^2 = (94 / 135.632)^2 = (0.693)^2 = 0.48$	

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Regression Line

A regression line is a line drawn through the points on a scatterplot to summarise the relationship between the variables being studied. When it slopes down (from top left to bottom right), this indicates a negative or inverse relationship between the variables; when it slopes up (from bottom right to top left), a positive or direct relationship is indicated.

The regression line often represents the regression equation on a scatterplot.

Simple Linear Regression

Simple linear regression aims to find a linear relationship between a response variable and a possible predictor variable by the method of least squares.

Multiple Regression

Multiple linear regression aims to find a linear relationship between a response variable and several possible predictor variables.

Nonlinear Regression

Nonlinear regression aims to describe the relationship between a response variable and one or more explanatory variables in a non-linear fashion.

Residual

Residual (or error) represents unexplained (or residual) variation after fitting a regression model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model.

Residuals

Earlier, we defined the residuals as the vertical distance between the fitted regression line and the data points. Another way to look at this is, the residual is the difference between the predicted value and the observed value,

$$e = y - \hat{y}.$$

Note that the sum of residuals, $\sum_i e_i$, always equals zero. However, the quantity we minimized to obtain our least squares equation was the *residual sum of squares*, SSRes,

$$\begin{aligned} \text{SSRes} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n e_i^2. \end{aligned}$$

An alternative formula for SSRes is

$$\text{SSRes} = S_{yy} - b_1 S_{xy}.$$

Notice that the residual sum of squares is never negative. The larger SSRes is, the further away from the line data points fall.

- A data point which does not follow the patterns of the majority of the data..
- Point A point whose removal makes a large difference in the equation of the regression line.
- Variable A variable which affects the response but is not one of the explanatory variables.
- Plot A plot of the residuals versus x (or the fitted value). We do not want any pattern in this plot.

Multiple Regression Correlation Coefficient

The multiple regression correlation coefficient, R^2 , is a measure of the proportion of variability explained by, or due to the regression (linear relationship) in a sample of paired data. It is a number between zero and one and a value close to zero suggests a poor model.

A very high value of R^2 can arise even though the relationship between the two variables is non-linear. The fit of a model should never simply be judged from the R^2 value.

Stepwise Regression

A 'best' regression model is sometimes developed in stages. A list of several potential explanatory variables are available and this list is repeatedly searched for variables which should be included in the model. The best explanatory variable is used first, then the second best, and so on. This procedure is known as stepwise regression.

Dummy Variable (in regression)

In regression analysis we sometimes need to modify the form of non-numeric variables, for example sex, or marital status, to allow their effects to be included in the regression model. This can be done through the creation of dummy variables whose role it is to identify each level of the original variables separately.

Transformation to Linearity

Transformations allow us to change all the values of a variable by using some mathematical operation, for example, we can change a number, group of numbers, or an equation by multiplying or dividing by a constant or taking the square root. A transformation to linearity is a transformation of a response variable, or independent variable, or both, which produces an approximate linear relationship between the variables.

$$x - 14 \qquad x - 14$$

Residuals, Outliers, and Influential Points

A linear regression model is not always appropriate for the data. You can assess the appropriateness of the model by examining residuals, outliers, and influential points.

Residuals

The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the **residual** (e). Each data point has one residual.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

$$e = y - \hat{y}$$

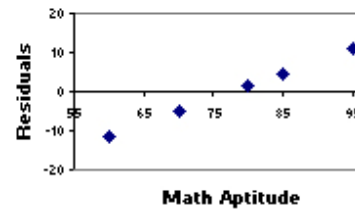
Both the sum and the mean of the residuals are equal to zero. That is, $\sum e = 0$ and $\bar{e} = 0$.

Residual Plots

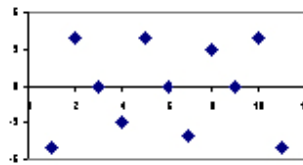
A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below the table on the left summarizes regression results from the example presented in a previous lesson, and the chart on the right displays those results as a residual plot. The residual plot shows a non-random pattern - negative residuals on the low end of the X axis and positive residuals on the high end. This indicates that a non-linear model will provide a much better fit to the data. Or it may be possible to "transform" the data to allow us to use a linear model. We discuss linear transformations in the next lesson.

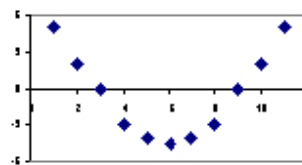
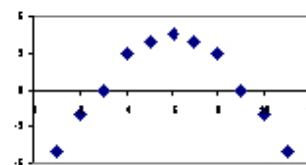
x	95	85	80	70	60
y	85	95	70	65	70
\hat{y}	87.95	81.51	78.29	71.84	65.41
e	10.95	4.51	1.29	5.159	11.59



Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.



Random pattern

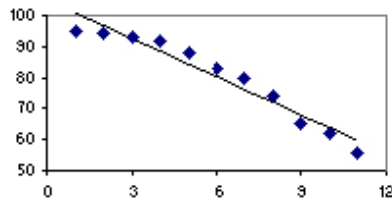
Non-random:
shaped curveU- Non-random:
Inverted U

Outliers

Data points that diverge from the overall pattern and have large residuals are called outliers.

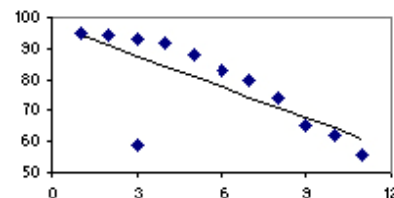
Outliers limit the fit of the regression equation to the data. This is illustrated in the scatterplots below. The coefficient of determination is bigger when the outlier is not present.

Without Outlier



Regression equation: $\hat{y} = 104.78 - 4.10x$
Coefficient of determination: $R^2 = 0.94$

With Outlier



Regression equation: $\hat{y} = 97.51 - 3.32x$
Coefficient of determination: $R^2 = 0.55$

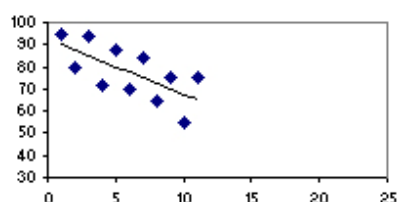
Influential Points

Influential points are data points with extreme values that greatly affect the slope of the regression line.

The charts below compare regression statistics for a data set with and without an influential point. The chart on the right has a single influential point, located at the high end of the X axis (where $x = 24$). As a result of that single influential point, the slope of the regression line increases dramatically, from -2.5 to -1.6.

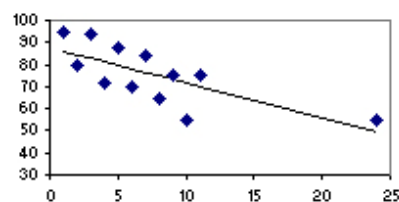
Note that this influential point, unlike the outliers discussed above, did not reduce the coefficient of determination. In fact, the coefficient of determination was bigger when the influential point was present.

Without Influential Point



Regression equation: $\hat{y} = 92.54 - 2.5x$
 Slope: $b_0 = -2.5$
 Coefficient of determination: $R^2 = 0.46$

With Influential Point



Regression equation: $\hat{y} = 87.59 - 1.6x$
 Slope: $b_0 = -1.6$
 Coefficient of determination: $R^2 = 0.52$

Test Your Understanding of This Lesson

In the context of regression analysis, which of the following statements are true?

- I. When the sum of the residuals is greater than zero, the model is nonlinear.
- II. Outliers reduce the coefficient of determination.
- III. Influential points reduce the correlation coefficient.

- (A) I only
- (B) II only
- (C) III only
- (D) I and II only
- (E) I, II, and III

Solution

The correct answer is (B). Outliers reduce the ability of a regression model to fit the data, and thus reduce the coefficient of determination. The sum of the residuals is always zero, whether the

regression model is linear or nonlinear. And influential points often increase the correlation coefficient.

Transformations to Achieve Linearity

When a residual plot reveals a data set to be nonlinear, it is often possible to "transform" the raw data to make it linear. This allows us to use linear regression techniques appropriately with nonlinear data.

What is a Transformation to Achieve Linearity?

Transforming a variable involves using a mathematical operation to change its measurement scale. Broadly speaking, there are two kinds of transformations.

- Linear transformation. A linear transformation preserves linear relationships between variables. Therefore, the correlation between x and y would be unchanged after a linear transformation. Examples of a linear transformation to variable x would be multiplying x by a constant, dividing x by a constant, or adding a constant to x .
- Nonlinear transformation. A nonlinear transformation changes (increases or decreases) linear relationships between variables and, thus, changes the correlation between variables. Examples of a nonlinear transformation of variable x would be taking the square root of x or the reciprocal of x .

In regression, a transformation to achieve linearity is a special kind of nonlinear transformation. It is a nonlinear transformation that *increases* the linear relationship between two variables.

Methods of Transforming Variables to Achieve Linearity

There are many ways to transform variables to achieve linearity for regression analysis. Some common methods are summarized below.

Method	Transformation(s)	Regression equation	Predicted value (\hat{y})
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = $\log(y)$	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = \sqrt{y}	$\sqrt{y} = b_0 + b_1x$	$\hat{y} = (\hat{y} = b_0 + b_1x)^2$

Reciprocal model	Dependent variable = $1/y$	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = $\log(x)$	$y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = $\log(y)$ Independent variable = $\log(x)$	$\log(y) = b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Each row shows a different nonlinear transformation method. The second column shows the specific transformation applied to dependent and/or independent variables. The third column shows the regression equation used in the analysis. And the last column shows the "back transformation" equation used to restore the dependent variable to its original, non-transformed measurement scale.

In practice, these methods need to be tested on the data to which they are applied to be sure that they *increase* rather than *decrease* the linearity of the relationship. Testing the effect of a transformation method involves looking at residual plots and correlation coefficients, as described in the following sections.

Note: The logarithmic model and the power model require the ability to work with logarithms. Use a graphic calculator to obtain the log of a number or to transform back from the logarithm to the original number. If you need it, the Stat Trek glossary has a brief refresher on logarithms.

How to Perform a Transformation to Achieve Linearity

Transforming a data set to achieve linearity is a multi-step, trial-and-error process.

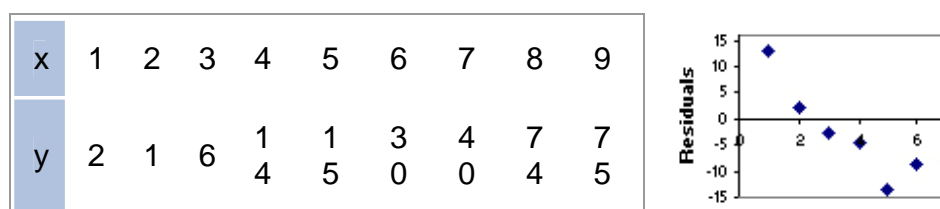
- Choose a transformation method (see above table).
- Transform the independent variable, dependent variable, or both.
- Plot the independent variable against the dependent variable, using the transformed data.
 - If the scatterplot is linear, proceed to the next step.
 - If the plot is not linear, return to Step 1 and try a different approach. Choose a different transformation method and/or transform a different variable.
- Conduct a regression analysis, using the transformed variables.
- Create a residual plot, based on regression results.

- If the residual plot shows a linear pattern, the transformation was successful. Congratulations!
- If the plot pattern is nonlinear, return to Step 1 and try a different approach.

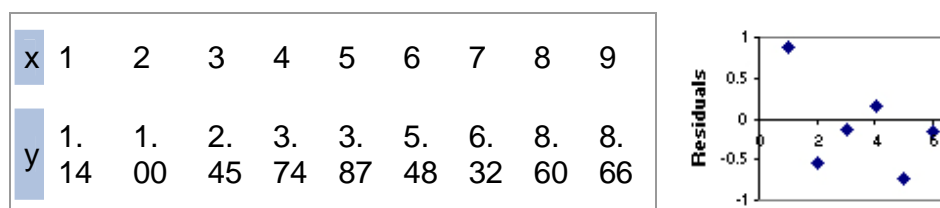
The best transformation method (exponential model, quadratic model, reciprocal model, etc.) will depend on nature of the original data. The only way to determine which method is best is to try each and compare the result (i.e., residual plots, correlation coefficients).

A Transformation Example

Below, the table on the left shows data for independent and dependent variables - x and y , respectively. When we apply a linear regression to the raw data, the residual plot shows a non-random pattern (a U-shaped curve), which suggests that the data are nonlinear.



Suppose we repeat the analysis, using a quadratic model to transform the dependent variable. For a quadratic model, we use the square root of y , rather than y , as the dependent variable. The table below shows the data we analyzed.



The residual plot (above right) suggests that the transformation to achieve linearity was successful. The pattern of residuals is random, suggesting that the relationship between the independent variable (x) and the transformed dependent variable (square root of y) is linear. And the coefficient of determination was 0.96 with the transformed data versus only 0.88 with the raw data. The transformed data resulted in a better model.

Test Your Understanding of This Lesson

Problem

In the context of regression analysis, which of the following statements are true?

- I. A linear transformation increases the linear relationship between variables.
- II. A logarithmic model is the most effective transformation method.
- III. A residual plot reveals departures from linearity.

- (A) I only
- (B) II only
- (C) III only
- (D) I and II only
- (E) I, II, and III

Solution

The correct answer is (C). A linear transformation neither increases nor decreases the linear relationship between variables; it preserves the relationship. A *nonlinear* transformation is used to increase the relationship between variables. The most effective transformation method depends on the data being transformed. In some cases, a logarithmic model may be more effective than other methods; but in other cases it may be less effective. Non-random patterns in a residual plot suggest a departure from linearity in the data being plotted.

Estimate Regression Slope

This lesson describes how to construct a confidence interval to estimate the slope of a regression line

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the slope (also called the regression coefficient), x is the value of the independent variable, and \hat{y} is the *predicted* value of the dependent variable.

Estimation Requirements

The approach described in this lesson is valid whenever the standard requirements for simple linear regression are met.

- The dependent variable Y has a linear relationship to the independent variable X .
- For each value of X , the probability distribution of Y has the same standard deviation σ .
- For any given value of X ,

- The Y values are independent.
- The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large.

Previously, we described how to verify that regression requirements are met.

The Variability of the Slope Estimate

To construct a confidence interval for the slope of the regression line, we need to know the standard error of the sampling distribution of the slope. Many statistical software packages and some graphing calculators provide the standard error of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: $y = 76 + 35x$.

Predictor	Coef	SE Coef	T	P
Constant	76	30	2.53	0.01
X	35	20	1.75	0.04

In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as "SE Coeff". However, other software packages might use a different label for the standard error. It might be "StDev", "SE", "Std Dev", or something else.

If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

$$SE = s_{b1} = \sqrt{ \sum (y_i - \hat{y}_i)^2 / (n - 2) } / \sqrt{ \sum (x_i - \bar{x})^2 }$$

where y_i is the value of the dependent variable for observation i , \hat{y}_i is estimated value of the dependent variable for observation i , x_i is the observed value of the independent variable for observation i , \bar{x} is the mean of the independent variable, and n is the number of observations.

How to Find the Confidence Interval for the Slope of a Regression Line

Previously, we described how to construct confidence intervals. The confidence interval for the slope uses the same general approach. Note, however, that the critical value is based on a t score with $n - 2$ degrees of freedom.

- Identify a sample statistic. The sample statistic is the regression slope b_1 calculated from sample data. In the table above, the regression slope is 35.
- Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
- Find the margin of error. Previously, we showed how to compute the margin of error, based on the critical value and standard error. When calculating the margin of error for a regression slope, use a t score for the critical value, with degrees of freedom (DF) equal to $n - 2$.
- Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic + margin of error*. And the uncertainty is denoted by the confidence level.

In the next section, we work through a problem that shows how to use this approach to construct a confidence interval for the slope of a regression line.

Test Your Understanding of This Lesson

Problem 1

The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

Regression equation: Annual bill = $0.55 * \text{Home size} + 15$

Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

What is the 99% confidence interval for the slope of the regression line?

- | | | | |
|-----|-------|----|------|
| (A) | 0.25 | to | 0.85 |
| (B) | 0.02 | to | 1.08 |
| (C) | -0.08 | to | 1.18 |
| (D) | 0.20 | to | 1.30 |
| (e) | 0.30 | to | 1.40 |

Solution

The correct answer is (C). Use the following four-step approach to construct a confidence interval.

- Identify a sample statistic. Since we are trying to estimate the slope of the true regression line, we use the regression coefficient for home size (i.e., the sample estimate of slope) as the sample statistic. From the regression output, we see that the slope coefficient is 0.55.
- Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 99% confidence level.
- Find the margin of error. Elsewhere on this site, we show how to compute the margin of error. The key steps applied to this problem are shown below.
 - Find standard deviation or standard error. The standard error is given in the regression output. It is 0.24.
 - Find critical value. The critical value is a factor used to compute the margin of error. With simple linear regression, to compute a confidence interval for the slope, the critical value is a t score with degrees of freedom equal to $n - 2$. To find the critical value, we take these steps.
 - Compute alpha (α): $\alpha = 1 - (\text{confidence level} / 100) = 1 - 99/100 = 0.01$
 - Find the critical probability (p^*): $p^* = 1 - \alpha/2 = 1 - 0.01/2 = 0.995$
 - Find the degrees of freedom (df): $df = n - 2 = 101 - 2 = 99$.
 - The critical value is the t score having 99 degrees of freedom and a cumulative probability equal to 0.995. From the t Distribution Calculator, we find that the critical value is 2.63.
 - Compute margin of error (ME): $ME = \text{critical value} * \text{standard error} = 2.63 * 0.24 = 0.63$
- Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic + margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 99% confidence interval is -0.08 to 1.18. That is, we are 99% confident that the true slope of the regression line is in the range defined by 0.55 ± 0.63 .

Hypothesis Test for Slope of Regression Line

This lesson describes how to conduct a hypothesis test to determine whether there is a significant linear relationship between an independent variable X and a dependent variable Y . The test focuses on the slope of the regression line

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the slope (also called the regression coefficient), X is the value of the independent variable, and Y is the value of the dependent variable.

Test Requirements

The approach described in this lesson is valid whenever the standard requirements for simple linear regression are met.

- The dependent variable Y has a linear relationship to the independent variable X .
- For each value of X , the probability distribution of Y has the same standard deviation σ .
- For any given value of X ,
 - The Y values are independent.
 - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large.

Previously, we described how to verify that regression requirements are met.

The test procedure consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

If there is a significant linear relationship between the independent variable X and the dependent variable Y , the slope will *not* equal zero.

$$\begin{array}{llll} H_0: & B_1 & = & 0 \\ H_a: & B_1 \neq 0 & & \end{array}$$

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use a linear regression t-test (described in the next section) to determine whether the slope of the regression line differs significantly from zero.

Analyze Sample Data

Using sample data, find the standard error of the slope, the slope of the regression line, the degrees of freedom, the test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- Standard error. Many statistical software packages and some graphing calculators provide the standard error of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: $y = 76 + 35x$.

Predictor	Coef	SE Coef	T	P
Constant	76	30	2.53	0.01
X	35	20	1.75	0.04

- In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as "SE Coeff". However, other software packages might use a different label for the standard error. It might be "StDev", "SE", "Std Dev", or something else.

If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

- $SE = s_{b1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n - 2)}} / \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$
- where y_i is the value of the dependent variable for observation i , \hat{y}_i is estimated value of the dependent variable for observation i , x_i is the observed value of the independent variable for observation i , \bar{x} is the mean of the independent variable, and n is the number of observations.
- Slope. Like the standard error, the slope of the regression line will be provided by most statistics software packages. In the hypothetical output above, the slope is equal to 35.

- Degrees of freedom. The degrees of freedom (DF) is equal to:

$$DF = n - 2$$

where n is the number of observations in the sample.

- Test statistic. The test statistic is a t-score (t) defined by the following equation.

$$t = b_1 / SE$$

where b_1 is the slope of the sample regression line, and SE is the standard error of the slope.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding of This Lesson

Problem

The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

Regression equation: Annual bill = 0.55 * Home size + 15

Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

Is there a significant linear relationship between annual bill and home size? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

H_0 : The slope of the regression line is equal to zero.

H_a : The slope of the regression line is *not* equal to zero.

If the relationship between home size and electric bill is significant, the slope will *not* equal zero.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.
- **Analyze sample data.** To apply the linear regression t-test to sample data, we require the standard error of the slope, the slope of the regression line, the degrees of freedom, the t-score test statistic, and the P-value of the test statistic.

We get the slope (b_1) and the standard error (SE) from the regression output.

$$b_1 = 0.55 \quad SE = 0.24$$

We compute the degrees of freedom and the t-score test statistic, using the following equations.

$$DF = n - 2 = 101 - 2 = 99$$

$$t = b_1/SE = 0.55/0.24 = 2.29$$

where DF is the degrees of freedom, n is the number of observations in the sample, b_1 is the slope of the regression line, and SE is the standard error of the slope.

Based on the t-score test statistic and the degrees of freedom, we determine the P-value. The P-value is the probability that a t-score having 99 degrees of freedom is more extreme than 2.29. Since this is a two-tailed test, "more extreme" means greater than 2.29 or less than -2.29. We use the t Distribution Calculator to find $P(t > 2.29) = 0.0121$ and $P(t < -2.29) = 0.0121$. Therefore, the P-value is $0.0121 + 0.0121$ or 0.0242.

- **Interpret results.** Since the P-value (0.0242) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention that this approach is only appropriate when the standard requirements for simple linear regression are satisfied.

Hypothesis Test for Slope of Regression Line

This lesson describes how to conduct a hypothesis test to determine whether there is a significant linear relationship between an independent variable X and a dependent variable Y . The test focuses on the slope of the regression line

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the slope (also called the regression coefficient), X is the value of the independent variable, and Y is the value of the dependent variable.

Test Requirements

The approach described in this lesson is valid whenever the standard requirements for simple linear regression are met.

- The dependent variable Y has a linear relationship to the independent variable X .
- For each value of X , the probability distribution of Y has the same standard deviation σ .
- For any given value of X ,
 - The Y values are independent.
 - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large.

Previously, we described how to verify that regression requirements are met.

The test procedure consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

If there is a significant linear relationship between the independent variable X and the dependent variable Y , the slope will *not* equal zero.

$$H_0: B_1 = 0$$

$$H_a: B_1 \neq 0$$

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use a linear regression t-test (described in the next section) to determine whether the slope of the regression line differs significantly from zero.

Analyze Sample Data

Using sample data, find the standard error of the slope, the slope of the regression line, the degrees of freedom, the test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- Standard error. Many statistical software packages and some graphing calculators provide the standard error of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: $y = 76 + 35x$.

Predictor	Coef	SE Coef	T	P
Constant	76	30	2.53	0.01
X	35	20	1.75	0.04

- In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as "SE Coeff". However, other software packages might use a different label for the standard error. It might be "StDev", "SE", "Std Dev", or something else.

If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

- $SE = s_{b1} = \sqrt{\sum (y_i - \hat{y}_i)^2 / (n - 2)} / \sqrt{\sum (x_i - \bar{x})^2}$

- where y_i is the value of the dependent variable for observation i , \hat{y}_i is estimated value of the dependent variable for observation i , x_i is the observed value of the independent variable for observation i , \bar{x} is the mean of the independent variable, and n is the number of observations.
- Slope. Like the standard error, the slope of the regression line will be provided by most statistics software packages. In the hypothetical output above, the slope is equal to 35.
- Degrees of freedom. The degrees of freedom (DF) is equal to:

$$DF = n - 2$$

where n is the number of observations in the sample.

- Test statistic. The test statistic is a t-score (t) defined by the following equation.

$$t = b_1 / SE$$

where b_1 is the slope of the sample regression line, and SE is the standard error of the slope.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding of This Lesson

Problem

The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

Regression equation: Annual bill = $0.55 * \text{Home size} + 15$

Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

Is there a significant linear relationship between annual bill and home size? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

H_0 : The slope of the regression line is equal to zero.

H_a : The slope of the regression line is *not* equal to zero.

If the relationship between home size and electric bill is significant, the slope will *not* equal zero.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.
- **Analyze sample data.** To apply the linear regression t-test to sample data, we require the standard error of the slope, the slope of the regression line, the degrees of freedom, the t-score test statistic, and the P-value of the test statistic.

We get the slope (b_1) and the standard error (SE) from the regression output.

$$b_1 = 0.55 \quad SE = 0.24$$

We compute the degrees of freedom and the t-score test statistic, using the following equations.

$$DF = n - 2 = 101 - 2 = 99$$

$$t = b_1/SE = 0.55/0.24 = 2.29$$

where DF is the degrees of freedom, n is the number of observations in the sample, b_1 is the slope of the regression line, and SE is the standard error of the slope.

Based on the t-score test statistic and the degrees of freedom, we determine the P-value. The P-value is the probability that a t-score having 99 degrees of freedom is more extreme than 2.29. Since this is a two-tailed test, "more extreme" means greater than 2.29 or less than -2.29. We use the t Distribution Calculator to find $P(t > 2.29) = 0.0121$ and $P(t < -2.29) = 0.0121$. Therefore, the P-value is $0.0121 + 0.0121$ or 0.0242.

- **Interpret results.** Since the P-value (0.0242) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention that this approach is only appropriate when the standard requirements for simple linear regression are satisfied.

Coefficient of Determination

A statistic that is widely used to determine how well a regression fits is the coefficient of determination (or multiple correlation coefficient), R^2 . R^2 represents the fraction of variability in y that can be explained by the variability in x . In other words, R^2 explains how much of the variability in the y 's can be explained by the fact that they are related to x , i.e., how close the points are to the line. The equation for R^2 is

$$R^2 = \frac{SSTotal - SSRes}{SSTotal} = 1 - \frac{SSRes}{SSTotal},$$

where SSTotal is the total sums of squares of the data.

NOTE: In the simple linear regression case, R^2 is simply the square of the correlation coefficient.

From simple regression to multiple regression

What happens if we have more than two independent variables? In most cases, we can't draw graphs to illustrate the relationship between them all. But we can still represent the relationship by an equation. This is what multiple regression does. It's a straightforward extension of simple regression. If there are n independent variables, we call them x_1, x_2, x_3 and so on up to x_n . Multiple regression then finds values of a, b_1, b_2, b_3 and so on up to b_n which give the best fitting equation of the form

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

b_1 is called the **coefficient** of x_1 , b_2 is the coefficient of x_2 , and so forth. The equation is exactly like the one for simple regression, except that it is very laborious to work out the values of a , b_1 etc by hand. Minitab, however, does it with exactly the same command as for simple regression. What do the regression coefficients mean? The coefficient of each independent variable tells us what relation that variable has with y , the dependent variable, with all the other independent variables held constant. So, if b_1 is high and positive, that means that if x_2 , x_3 and so on up to x_n do not change, then increases in x_1 will correspond to large increases in y .

Goodness of fit in multiple regression

In multiple regression, as in simple regression, we can work out a value for R^2 . However, every time we add another independent variable, we necessarily increase the value of R^2 (you can get an idea of why this happens if you compare Fig 3 with Fig 1 in the handout on "The idea of a regression equation"). Therefore, in assessing the goodness of fit of a regression equation, we usually work in terms of a slightly different statistic, called R^2 or R^2_{adj} . This is calculated as

$$R^2_{\text{adj}} = 1 - (1 - R^2)(N - n - 1) / (N - 1)$$

where N is the number of observations in the data set (usually the number of people) and n the number of independent variables or **regressors**. This allows for the extra regressors. Check that you can see from the formula that R^2_{adj} will always be lower than R^2 if there is more than one regressor. There is also another way of assessing goodness of fit in multiple regression, using the F statistic which we will meet in a moment.

The main questions multiple regression answers

Multiple regression enables us to answer five main questions about a set of data, in which n independent variables (regressors), x_1 to x_n , are being used to explain the variation in a single dependent variable, y .

1. How well do the regressors, taken together, explain the variation in the dependent variable? This is assessed by the value of R^2_{adj} . As a very rough guide, in psychological applications we would usually reckon an R^2_{adj} of above 75% as very good; 50% to 75% as good; 25% to 50% as poor but acceptable; and below 25% as very poor and perhaps unacceptable. Alas, an R^2_{adj} value above 90% is very rare in psychology, and should make you wonder whether there is some artefact in your data.

2. Are the regressors, taken together, significantly associated with the dependent variable? This is assessed by the statistic F in the "Analysis of Variance" or anova part of the regression output. F is like some other statistics (e.g. t , χ^2) in that its significance depends on its **degrees of freedom**, which in turn depend on sample sizes and/or the nature of the test used. Unlike t , though, F has two degrees of freedom associated with it. In general they are referred to as the **numerator** and **denominator** degrees of freedom (because F is actually a ratio). In regression, the numerator degrees of freedom are associated with the regression, and the denominator degrees of freedom with the **residual** or **error**; you can find them in the Regression and Error rows of the anova table in the Minitab output. If you were finding the significance of an F value by looking it up in a book of tables, you would need the degrees of freedom to do it. Minitab works out significances for you, and you will find them in the anova table next to the F value; but you need to use the degrees of freedom when reporting the results (see below). Note that the higher the value of F , the more significant it will be for given degrees of freedom.
3. What relationship does each regressor have with the dependent variable when all other regressors are held constant? This is answered by looking at the regression coefficients. Minitab reports these twice, once in the regression equation and again (to an extra decimal place) in the table of regression coefficients and associated statistics. Note that regression coefficients have units. So if the dependent variable is score on a psychometric test of depression, and one of the regressors is monthly income, the coefficient for that regressor would have units of (scale points) per (income per month). That means that if we changed the units of one of the variables, the regression coefficient would change but the relationship it is describing, and what it is saying about it, would not. So the size of a regression coefficient doesn't tell us anything about the strength of the relationship it describes until we have taken the units into account. The fact that regression coefficients have units also means that we can give a precise interpretation to each coefficient. So, staying with depression score and income, a coefficient of -0.0934 (as in the worked example on the next sheet) would mean that, with all other variables held constant, increasing someone's income by 1 per month is associated with a decrease of depression score of 0.0934 points (we might want to make this more meaningful by saying that an increase in income of 100 per month would be associated with a decrease in depression score of $100 * 0.0934 = 9.34$ scale units). As in this example, negative coefficients mean that when the regressor

increases, the dependent variable decreases. If the regressor is a **dichotomous** variable (e.g. gender), the size of the coefficient tells us the size of the difference between the two classes of individual (again, with all other variables held constant). So a gender coefficient of 3.3, with men coded 0 and women coded 1, would mean that with all other variables held constant, women's dependent variable scores would average 3.3 units higher than men's.

4. Which regressor has most effect on the dependent variable? It is not possible to give a fully satisfactory answer to this question, for a number of reasons. The chief one is that we are always looking at the effect of each variable in the presence of all the others; since the dependent variable need not be independent, it is hard to be sure which one is contributing to a joint relationship (or even to be sure that that means anything). However, the usual way of addressing the question is to look at the **standardised regression coefficients** or **beta weights** for each variable; these are the regression coefficients we would get if we converted all variables (independent and dependent) to **z-scores** before doing the regression. Minitab, unfortunately, does not report beta weights for the independent variable in its regression output, though it is possible to calculate them; SPSS, which you will learn about later in the course, does give them directly.
5. Are the relationships of each regressor with the dependent variable statistically significant, with all other regressors taken into account? This is answered by looking at the *t* values in the table of regression coefficients. The degrees of freedom for *t* are those for the residual in the anova table, but Minitab works out significances for us, so we need to know the degrees of freedom only when it comes to reporting results. Note that if a regression coefficient is negative, Minitab will report the corresponding *t* value as negative, but if you were looking it up in tables, you would use the **absolute** (unsigned) value.

Further questions to ask

Either the nature of the data, or the regression results, may suggest further questions. For example, you may want to obtain means and standard deviations or histograms of variables to check on their distributions; or plot one variable against another, or obtain a matrix of correlations, to check on first order relationships. Minitab does some checking for you automatically, and reports if it finds "unusual observations". If there are unusual observations, PLOT or HISTOGRAM may tell you what the possible problems are. The usual kinds of unusual observations are "**outliers**" points which lie far from the main distributions or the main trends of one or more

variables. Serious outliers should be dealt with as follows:

1. temporarily remove the observations from the data set. In Minitab, this can be done by using the LET command to set the outlier value to "**missing**", indicated by an asterisk instead of a numerical value. For example, if item 37 in the variable held in C1 looks like an outlier, we could type:
LET C1(37)='*'note the single quotes round the asterisk

2. repeat the regression and see whether the same qualitative results are obtained (the quantitative results will inevitably be different).

3. if the same general results are obtained, we can conclude that the outliers are not distorting the results. Report the results of the original regression, adding a note that removal of outliers did not greatly affect them.

4. if different general results are obtained, accurate interpretation will require more data to be collected. Report the results of both regressions, and note that the interpretation of the data is uncertain. The outliers may be due to errors of observation, data coding, etc, and in this case they should be corrected or discarded. However, they may also represent a subpopulation for which the effects of interest are different from those in the main population. If they are not due to error, the group of data contributing to outliers will need to be identified, and if possible a reasonably sized sample collected from it so that it can be compared with the main population. This is a scientific rather than a statistical problem.

Reporting regression results

Research articles sometimes report the results of several different regressions done on a single data set. In this case, it is best to present the results in a table. Where a single regression is done, however, that is unnecessary, and the results can be reported in text. The wording should be something like the following this is for the depression vs age, income and gender example used in the class:

The data were analysed by multiple regression, using as regressors age, income and gender. The regression was a rather poor fit ($R^2_{\text{adj}} = 40\%$), but the overall relationship was significant ($F_{3,12} = 4.32$, $p < 0.05$). With other variables held constant, depression scores were negatively related to age and income, decreasing by 0.16 for every extra year of age, and by 0.09 for every extra pound per week income. Women tended to have higher scores than men, by 3.3 units. Only the effect of income was significant ($t_{12} = 3.18$, $p < 0.01$).

Note the following:

- The above brief paragraph does not exhaust what you can say about a set of regression results. There may be features of the data you should look at "Unusual observations", for example. Normally you will need to go on to discuss the meaning of the trends you have described.
- Always report what happened before moving on to its significance so R^2_{adj} values before F values, regression coefficients before t values. Remember, descriptive statistics are more important than significance tests.
- Although Minitab will give a negative t value if the corresponding regression coefficient is negative, you should drop the negative sign when reporting the results.
- Degrees of freedom for both F and t values must be given. Usually they are written as subscripts. For F the numerator degrees of freedom are given first. You can also put degrees of freedom in parentheses, or report them explicitly, e.g.: " $F(3,12) = 4.32$ " or " $F = 4.32$, d. of f. = 3, 12".
- Significance levels can either be reported exactly (e.g. $p = 0.032$) or in terms of conventional levels (e.g. $p < 0.05$). There are arguments in favour of either, so it doesn't much matter which you do. But you should be consistent in any one report.
- Beware of highly significant F or t values, whose significance levels will be reported by statistics packages as, for example, 0.0000. It is an act of statistical illiteracy to write $p = 0.0000$; significance levels can never be exactly zero there is always some probability that the observed data could arise if the **null hypothesis** was true. What the package means is that this probability is so low it can't be represented with the number of columns available. We should write it as $p < 0.00005$.
- Beware of **spurious precision**, i.e. reporting coefficients etc to huge numbers of **significant figures** when, on the basis of the sample you have, you couldn't possibly expect them to replicate to anything like that degree of precision if someone repeated the study. F and t values are conventionally reported to two decimal places, and R^2_{adj} values to the nearest percentage point (sometimes to one additional decimal place). For coefficients, you should be guided by the sample size: with a sample size of 16, as in the example used above, two significant figures is plenty, but even with more realistic samples, in the range of 100 to 1000, three significant figures is usually as far as you should go. This means that you will usually have to round off the numbers that Minitab will give you.

Worked example of an elementary multiple regression

```

MTB > set c1
DATA> 74 82 15 23 35 54 12 28 66 43 55 31 83 29 53 32
DATA> end
MTB > set c2
DATA> 120 55 350 210 185 110 730 150 61 175 121 225 45 325
171 103
DATA> end
MTB > set c3
DATA> 0 0 1 0 0 1 1 0 1 1 1 0 1 0 0 1
DATA> end
MTB > set c4
DATA> 33 28 47 55 32 63 59 68 27 32 42 51 47 33 51 20
DATA> end
MTB > name c1 'depress'
MTB > name c2 'income'
MTB > name c3 'm0f1'
MTB > name c4 'age'

```

```
MTB > regress c1 3 c2-c4
```

The regression equation is

depress = 68.3 0.0934 income + 3.31 m0f1 - 0.162 age

Predictor	Coef	Stdev	t-ratio	p
Constant	68.28	15.44	4.42	0.001
income	-0.09336	0.02937	-3.18	0.008
m0f1	3.306	8.942	0.37	0.718
age	-0.1617	0.3436	-0.47	0.646

s = 17.70 R = 52.0% R = 39.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	4065.4	1355.1	4.32	0.028
Error	12	3760.0	313.3		
Total	15	7825.4			

SOURCE	DF	SEQ SS
income	1	3940.5
m0f1	1	55.5
age	1	69.4

Continue? y

Unusual Observations

Obs.	income	depress	Fit	Stdev.Fit	Residual	St.Resid
7	730	12.00	-6.10	15.57	18.10	2.15RX

R denotes an obs. with a large st. resid.

X denotes an obs. whose X value gives it large influence.

LINEAR REGRESSION

1. Introduction

- very often when 2 (or more) variables are observed, relationship between them can be visualized
- predictions are always required in economics or physical science from existing and historical data
- regression analysis is used to help formulate these predictions and relationships
- linear regression is a special kind of regression analysis in which 2 variables are studied and a straight-line relationship is assumed
- linear regression is important because
 1. there exist many relationships that are of this form
 2. it provides close approximations to complicated relationships which would otherwise be difficult to describe
- the 2 variables are divided into (i) independent variable and (ii) dependent variable
- Dependent Variable is the variable that we want to forecast
- Independent Variable is the variable that we use to make the forecast
- e.g. Time vs. GNP (time is independent, GNP is dependent)
- scatter diagrams are used to graphically presenting the relationship between the 2 variables
- usually the independent variable is drawn on the horizontal axis (X) and the dependent variable on vertical axis (Y)
- the regression line is also called the regression line of Y on X

2. Assumptions

- there is a linear relationship as determined (observed) from the scatter diagram
- the dependent values (Y) are independent of each other, i.e. if we obtain a large value of Y on the first observation, the result of the second and subsequent observations will not necessarily provide a large value. In simple term, there should not be auto-correlation
- for each value of X the corresponding Y values are normally distributed
- the standard deviations of the Y values for each value of X are the same, i.e. *homoscedasticity*

3. Process

- observe and note what is happening in a systematic way
- form some kind of theory about the observed facts
- draw a scatter diagram to visualize relationship
- generate the relationship by mathematical formula
- make use of the mathematical formula to predict

4. Method of Least Squares

- from a scatter diagram, there is virtually no limit as to the number of lines that can be drawn to make a linear relationship between the 2 variables
- the objective is to create a BEST FIT line to the data concerned
- the criterion is the called the method of least squares
- i.e. the *sum of squares* of the *vertical deviations* from the points to the line be a minimum (based on the fact that the dependent variable is drawn on the vertical axis)
- the linear relationship between the dependent variable (Y) and the independent variable can be written as $Y = a + bX$, where a and b are parameters describing the vertical intercept and the slope of the regression line respectively

5. Calculating a and b

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

○

$$a = \bar{Y} - b\bar{X}$$

- where X and Y are the raw values of the 2 variables
- \bar{X} and \bar{Y} are means of the 2 variables

6. Correlation

- when the value of one variable is related to the value of another, they are said to be correlated
- there are 3 types of correlation: (i) perfectly correlated; (ii) partially correlated; (iii) uncorrelated
- Coefficient of Correlation (r) measures such a relationship

$$r = \frac{\sqrt{\sum(\hat{Y} - \bar{Y})^2}}{\sqrt{\sum(Y - \bar{Y})^2}}$$

$$= \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} \times \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

- the value of r ranges from -1 (perfectly correlated in the negative direction) to +1 (perfectly correlated in the positive direction)
- when $r = 0$, the 2 variables are not correlated

7. Coefficient of Determination

- this calculates the proportion of the variation in the actual values which can be predicted by changes in the values of the independent variable
- denoted by r^2 , the square of the coefficient of correlation

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

- r^2 ranges from 0 to 1 (r ranges from -1 to +1)
- expressed as a percentage, it represents the proportion that can be predicted by the regression line
- the value $1 - r^2$ is therefore the proportion contributed by other factors

8. Standard Error of Estimate (SEE)

- a measure of the variability of the regression line, i.e. the dispersion around the regression line
- it tells how much variation there is in the dependent variable between the raw value and the expected value in the regression

$$S_{se} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

- this SEE allows us to generate the confidence interval on the regression line as we did in the estimation of means

9. Confidence interval for the regression line (estimating the expected value)

- estimating the *mean value* of \hat{Y} for a given value of X is a very important practical problem
- e.g. if a corporation's profit Y is linearly related to its advertising expenditures X , the corporation may want to estimate the *mean profit* for a given expenditure X
- this is given by the formula

$$\hat{Y} \pm t \cdot S_{se} \cdot \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

- at n-2 degrees of freedom for the t-distribution
- 10. Confidence interval for individual prediction**
- for technical reason, the above formula must be amended and is given by

$$\hat{Y} \pm t \cdot S_{\text{est}} \cdot \sqrt{\frac{n+1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

An Example

	Accounting X	Statistics Y	X ²	Y ²	XY
1	74.00	81.00	5476.00	6561.00	5994.00
2	93.00	86.00	8649.00	7396.00	7998.00
3	55.00	67.00	3025.00	4489.00	3685.00
4	41.00	35.00	1681.00	1225.00	1435.00
5	23.00	30.00	529.00	900.00	690.00
6	92.00	100.00	8464.00	10000.00	9200.00
7	64.00	55.00	4096.00	3025.00	3520.00
8	40.00	52.00	1600.00	2704.00	2080.00
9	71.00	76.00	5041.00	5776.00	5396.00
10	33.00	24.00	1089.00	576.00	792.00
11	30.00	48.00	900.00	2304.00	1440.00
12	71.00	87.00	5041.00	7569.00	6177.00
Sum	687.00	741.00	45591.00	52525.00	48407.00
Mean	57.25	61.75	3799.25	4377.08	4033.92

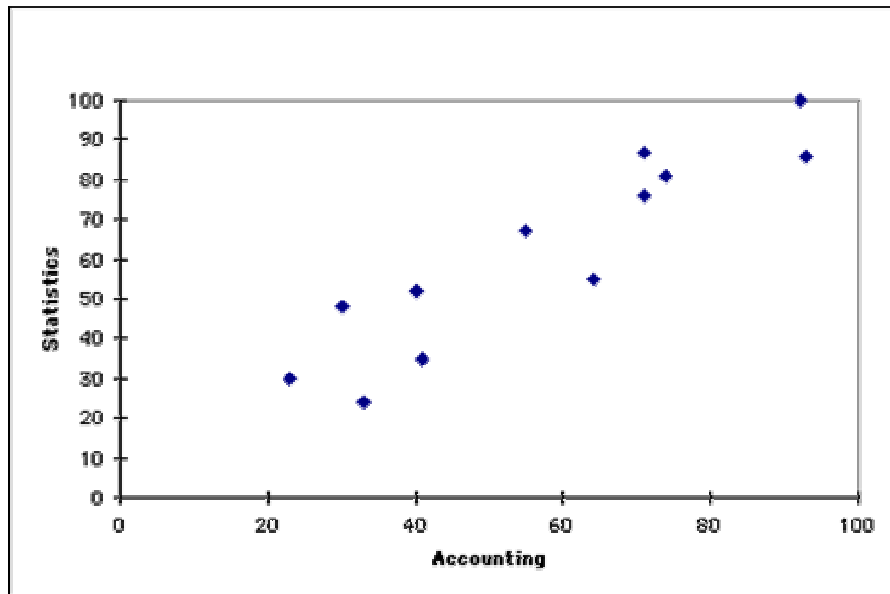


Figure 1: Scatter Diagram of Raw Data

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$b = \frac{48407 - 12 \times 57.25 \times 61.75}{45591 - 12 \times (57.25)^2}$$

$$b = 0.9560$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 61.75 - 0.9560 \times 57.25$$

$$a = 7.0194$$

$$\hat{Y} = 7.0194 + 0.9560X$$

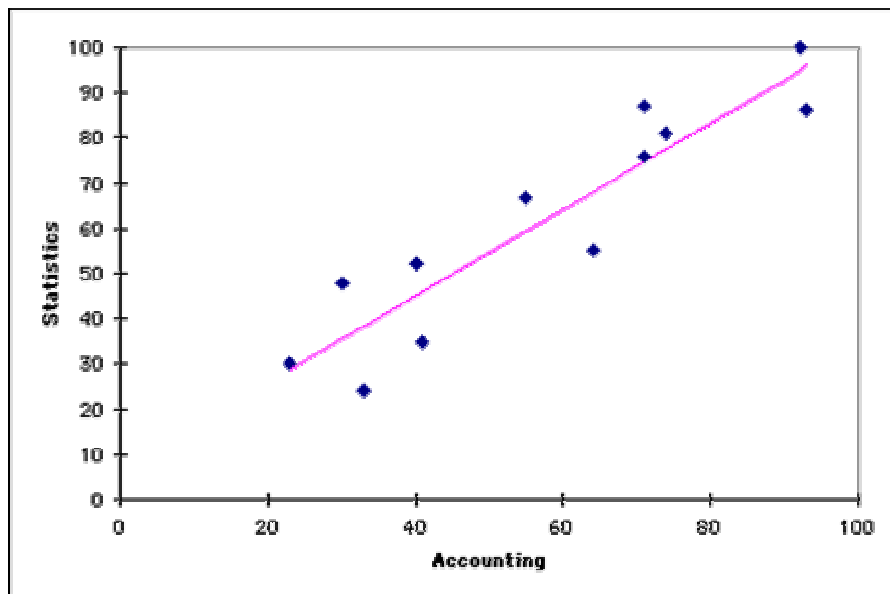


Figure 2: Scatter Diagram and Regression Line

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \times \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{12 \times 48407 - 687 \times 741}{\sqrt{12 \times 45591 - 687^2} \times \sqrt{12 \times 52525 - 741^2}}$$

$$r = \mathbf{0.9194}$$

$$r^2 = \mathbf{0.8453}$$

Interpretation/Conclusion

There is a linear relation between the results of Accounting and Statistics as shown from the scatter diagram in Figure 1. A linear regression analysis was done using the least-square method. The resultant regression line is represented by $\hat{Y} = 7.0194 + 0.9560X$ in which X represents the results of Accounting and Y that of Statistics. Figure 2 shows the regression line. In this example, the choice of dependent and independent variables is arbitrary. It can be said that the results of Statistics are correlated to that of Accounting or vice versa.

The Coefficient of Determination r^2 is 0.8453. This shows that the two variables are correlated. Nearly 85% of the variation in Y is explained by the regression line.

The Coefficient of Correlation (r) has a value of 0.92. This indicates that the two variables are positively correlated (Y increases as X increases).

Method of the Least Square

To a statistician, the line will have a good fit if it minimizes the error between the estimated points on the line and the actual observed points that were used to draw it.

One way we can measure the error of our estimating line is to sum all the individual differences, or errors, between the estimated points. Let \hat{Y} be the individual values of the estimated points and Y be the raw data values.

Figure 1 shows an example.

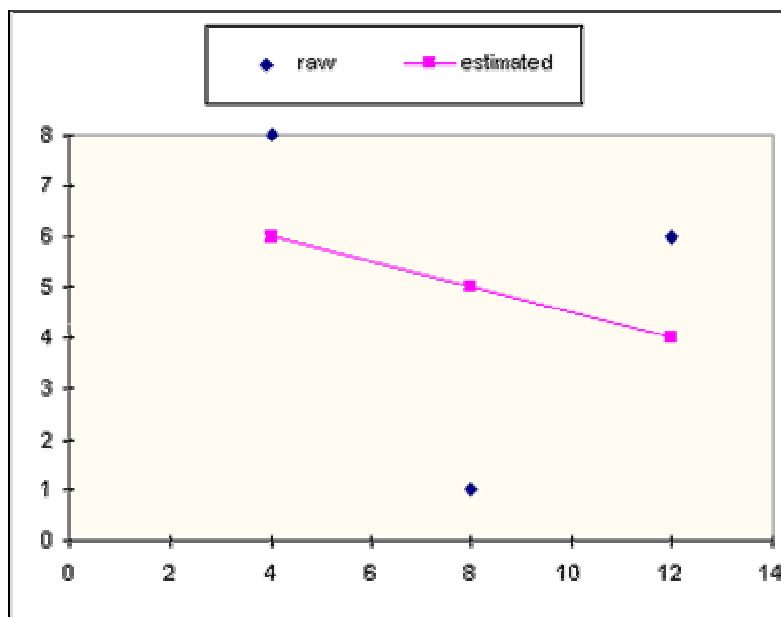


Figure 1

Y	\hat{Y}	diff
8	6	2
1	5	-4
6	4	2
total error		0

Figure 2 shows another example.

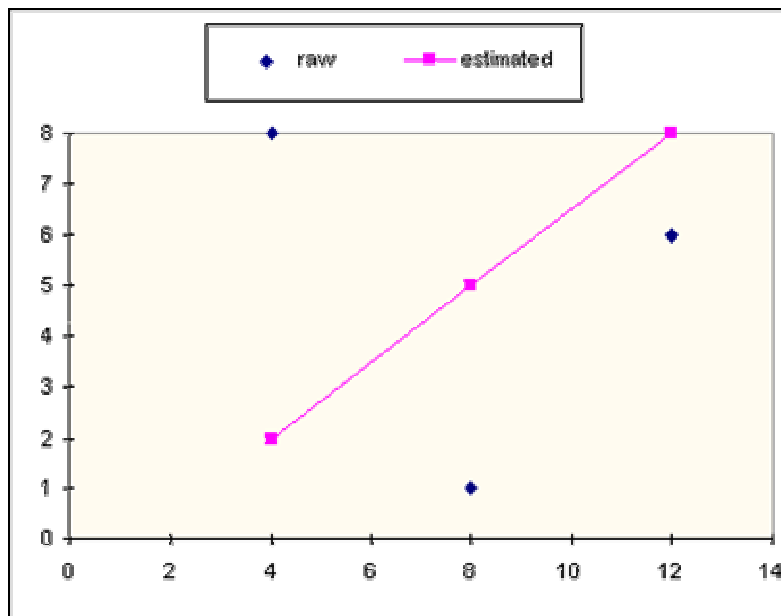


Figure 2

Y	\hat{Y}	diff
8	2	6
1	5	-4
6	8	-2
total error		0

It also has a zero sum of error as shown from the above table.

A visual comparison between the two figures shows that the regression line in Figure 1 fits the three data points better than the line in Figure 2. However the process of summing the individual differences in the above 2 tables indicates that both lines describe the data equally well. Therefore we can conclude that the process of summing individual differences for calculating the error is not a reliable way to judge the goodness of fit of an estimating line.

The problem with adding the individual errors is the canceling effect of the positive and negative values. From this, we might deduce that the proper criterion for judging the goodness of fit would be to add the absolute values of each error. The following table shows a comparison between the absolute values of Figure 1 and Figure 2.

Figure 1			Figure 2		
Y	\hat{Y}	abs. diff	Y	\hat{Y}	abs. diff
8	6	2	8	2	6
1	5	4	1	5	4
6	4	2	6	8	2
total error		8	total error		12

Since the absolute error for Figure 1 is smaller than that for Figure 2, we have confirmed our intuitive impression that the estimating line in Figure 1 is the better fit.

Figure 3 and Figure 4 below show another scenarios.

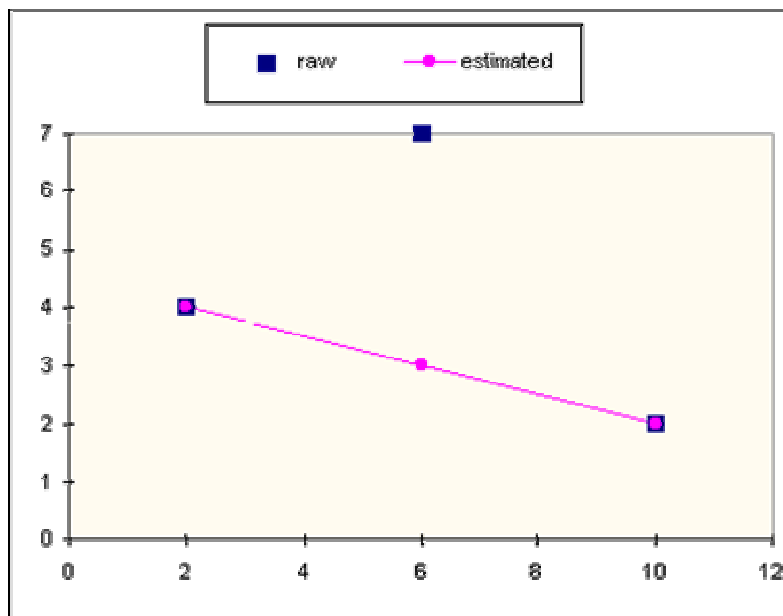


Figure 3

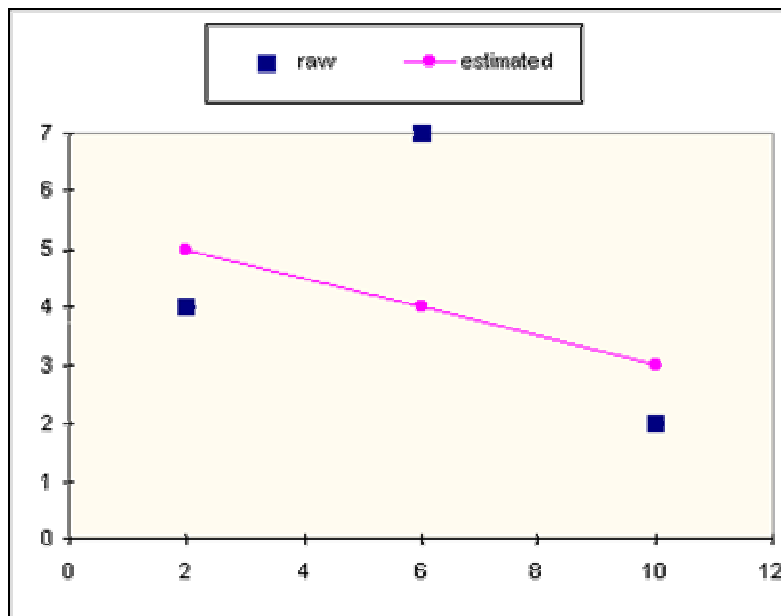


Figure 4

The following table shows the calculations of absolute values of the errors.

Figure 3			Figure 4		
Y	\hat{Y}	abs. diff	Y	\hat{Y}	abs. diff
4	4	0	4	5	1
7	3	4	7	4	3
2	2	0	2	3	1
total error		4	total error		5
proper					

We have added the absolute values of the errors and found that the estimating line in Figure 3 is a better fit than the line in Figure 4. Intuitively, however, it appears that the line in Figure 4 is the better fit line, because it has been moved vertically to take the middle point into consideration. Figure 3, on the other hand, seems to ignore the middle point completely.

Because The sum of the absolute values does not stress the *magnitude* of the error.

In effect, we want to find a way to penalize large absolute errors, so that we can avoid them. We can accomplish this if we square the individual errors before we add them. Squaring each term accomplishes two goals:

It magnifies, or penalizes, the larger errors.

1. It cancels the effect of the positive and negative values (a negative error squared is still positive).

Figure 3				Figure 4			
Y	\hat{Y}	abs diff	square diff	Y	\hat{Y}	abs diff	square diff
4	4	0	0	4	5	1	1
7	3	4	16	7	4	3	9
2	2	0	0	2	3	1	1
sum of squares			16	sum of squares			11

Applying the Least Square Criterion to the Estimating Lines (Fig 3 & 4)

Since we are looking for the estimating line that minimizes the sum of the squares of the errors, we call this the **Least Squares Method**.



STATISTICS - DISPERSION

Frequency distribution

In statistics, a frequency distribution is a list of the values that a variable takes in a sample. It is usually a list, ordered by quantity, showing the number of times each value appears. For example, if 100 people rate a five-point Likert scale assessing their agreement with a statement on a scale on which 1 denotes strong agreement and 5 strong disagreement, the frequency distribution of their responses might look like:

Rank	Degree of agreement	Number
1	Strongly agree	25
2	Agree somewhat	35
3	Not sure	20
4	Disagree somewhat	15
5	Strongly disagree	30

This simple tabulation has two drawbacks. When a variable can take continuous values instead of discrete values or when the number of possible values is too large, the table construction is cumbersome, if it is not impossible. A slightly different tabulation scheme based on the range of values is used in such cases. For example, if we consider the heights of the students in a class, the frequency table might look like below.

Height range	Number of students	Cumulative Number
4.5 -5.0 feet	25	25
5.0-5.5 feet	35	60
5.5-6 feet	20	80
6.0-6.5 feet	20	100

Applications

Managing and operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions. This

assessment involves measures of central tendency or averages, such as the mean and median, and measures of variability or statistical dispersion, such as the standard deviation or variance.

A frequency distribution is said to be skewed when its mean and median are different. The kurtosis of a frequency distribution is the concentration of scores at the mean, or how peaked the distribution appears if depicted graphically—for example, in a histogram. If the distribution is more peaked than the normal distribution it is said to be leptokurtic; if less peaked it is said to be platykurtic.

Frequency distributions are also used in frequency analysis to crack codes and refer to the relative frequency of letters in different languages.

Class Interval

In statistics, the range of each class of data, used when arranging large amounts of raw data into grouped data. To obtain an idea of the distribution, the data are broken down into convenient classes (commonly 6–16), which must be mutually exclusive and are usually equal in width to enable histograms to be drawn. The class boundaries should clearly define the range of each class. When dealing with discrete data, suitable intervals would be, for example, 0–2, 3–5, 6–8, and so on. When dealing with continuous data, suitable intervals might be $170 \leq X < 180$, $180 \leq X < 190$, $190 \leq X < 200$, and so on.

Cross tabulation

A cross tabulation (often abbreviated as cross tab) displays the joint distribution of two or more variables. They are usually presented as a contingency table in a matrix format. Whereas a frequency distribution provides the distribution of one variable, a contingency table describes the distribution of two or more variables simultaneously. Each cell shows the number of respondents that gave a specific combination of responses, that is, each cell contains a single cross tabulation.

The following is a fictitious example of a 3×2 contingency table. The variable “Wikipedia usage” has three categories: heavy user, light user, and non user. These categories are all inclusive so the columns sum to 100%. The other variable “underpants” has two categories: boxers, and briefs. These categories are not all inclusive so the rows need not sum to 100%. Each cell gives the percentage of subjects that share that combination of traits.

	boxers	briefs
heavy Wiki user	70%	5%
light Wiki user	25%	35%
non Wiki user	5%	60%

Cross tabs are frequently used because:

1. They are easy to understand. They appeal to people that do not want to use more sophisticated measures.
2. They can be used with any level of data: nominal, ordinal, interval, or ratio - cross tabs treat all data as if it is nominal
3. A table can provide greater insight than single statistics
4. It solves the problem of empty or sparse cells
5. they are simple to conduct

Statistics related to cross tabulations

The following list is not comprehensive.

Chi-square - This tests the statistical significance of the cross tabulations. Chi-squared should not be calculated for percentages. The cross tabs must be converted back to absolute counts (numbers) before calculating chi-squared. Chi-squared is also problematic when any cell has a joint frequency of less than five. For an in-depth discussion of this issue see Fienberg, S.E. (1980). "The Analysis of Cross-classified Categorical Data." 2nd Edition. M.I.T. Press, Cambridge, MA.

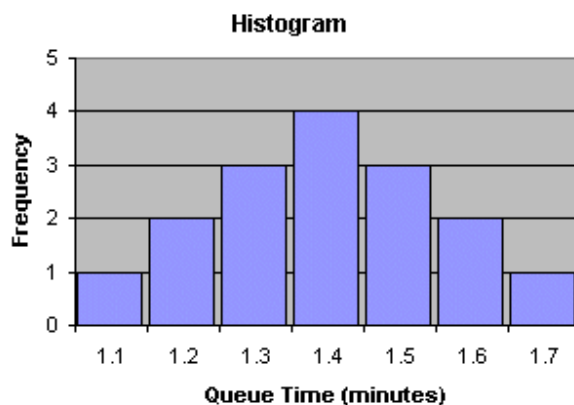
- Contingency Coefficient - This tests the strength of association of the cross tabulations. It is a variant of the phi coefficient that adjusts for statistical significance. Values range from 0 (no association) to 1 (the theoretical maximum possible association).
- Cramer's V - This tests the strength of association of the cross tabulations. It is a variant of the phi coefficient that adjusts for the number of rows and columns. Values range from 0 (no association) to 1 (the theoretical maximum possible association).
- Lambda Coefficient - This tests the strength of association of the cross tabulations when the variables are measured at the nominal level. Values range from 0 (no association) to 1 (the theoretical maximum possible association). Asymmetric lambda measures the percentage improvement in predicting the dependent variable. Symmetric lambda measures the

percentage improvement when prediction is done in both directions.

- phi coefficient - If both variables instead are nominal and dichotomous, phi coefficient is a measure of the degree of association between two binary variables. This measure is similar to the correlation coefficient in its interpretation. Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal.
- Kendall tau:
 - Tau b - This tests the strength of association of the cross tabulations when both variables are measured at the ordinal level. It makes adjustments for ties and is most suitable for square tables. Values range from -1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association.
 - Tau c - This tests the strength of association of the cross tabulations when both variables are measured at the ordinal level. It makes adjustments for ties and is most suitable for rectangular tables. Values range from -1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association.
- Gamma - This tests the strength of association of the cross tabulations when both variables are measured at the ordinal level. It makes no adjustment for either table size or ties. Values range from -1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association.
- Uncertainty coefficient, entropy coefficient or Theil's U

Purpose Of A Histogram A histogram is used to graphically summarize and display the distribution of a process data set.

Sample Bar Chart Depiction



How To Construct A Histogram

A histogram can be constructed by segmenting the range of the data into equal sized bins (also called segments, groups or classes). For example, if your data ranges from 1.1 to 1.8, you could have equal bins of 0.1 consisting of 1 to 1.1, 1.2 to 1.3, 1.3 to 1.4, and so on.

The vertical axis of the histogram is labeled Frequency (the number of counts for each bin), and the horizontal axis of the histogram is labeled with the range of your response variable.

You then determine the number of data points that reside within each bin and construct the histogram. The bins size can be defined by the user, by some common rule, or by software methods (such as Minitab).

What Questions The Histogram Answers

What is the most common system response?

What distribution (center, variation and shape) does the data have?

Does the data look symmetric or is it skewed to the left or right?

Does the data contain outliers?

Geometric mean

The geometric mean, in mathematics, is a type of mean or average, which indicates the central tendency or typical value of a set of numbers. It is similar to the arithmetic mean, which is what most people think of with the word "average," except that instead of adding the set of numbers and then dividing the sum by the count of numbers in the set, n , the numbers are multiplied and then the n th root of the resulting product is taken.

For instance, the geometric mean of two numbers, say 2 and 8, is just the square root (i.e., the second root) of their product, 16, which is 4. As another example, the geometric mean of 1, $\frac{1}{2}$, and $\frac{1}{4}$ is simply the cube root (i.e., the third root) of their product, 0.125, which is $\frac{1}{2}$.

The geometric mean can be understood in terms of geometry. The geometric mean of two numbers, a and b , is simply the side length of the square whose area is equal to that of a rectangle with side lengths a and b . That is, what is n such that $n^2 = a \times b$? Similarly, the geometric mean of three numbers, a , b , and c , is the side length of a cube whose volume is the same as that of a rectangular prism with side lengths equal to the three given numbers. This geometric interpretation of the mean is very likely what gave it its name.

The geometric mean only applies to positive numbers.^[1] It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment. The geometric mean is also one of the three classic Pythagorean means, together with the aforementioned arithmetic mean and the harmonic mean.

Calculation

The geometric mean of a data set $[a_1, a_2, \dots, a_n]$ is given by

$$\left(\prod_{i=1}^n a_i \right)^{1/n} = \sqrt[n]{a_1 \cdot a_2 \cdot \dots \cdot a_n}$$

The geometric mean of a data set is smaller than or equal to the data set's arithmetic mean (the two means are equal if and only if all members of the data set are equal). This allows the definition of the arithmetic-geometric mean, a mixture of the two which always lies in between.

The geometric mean is also the arithmetic-harmonic mean in the sense that if two sequences (a_n) and (h_n) are defined:

$$a_{n+1} = \frac{a_n + h_n}{2}, \quad a_0 = x$$

and

$$h_{n+1} = \frac{2}{\frac{1}{a_n} + \frac{1}{h_n}}, \quad h_0 = y$$

then a_n and h_n will converge to the geometric mean of x and y .

Relationship with arithmetic mean of logarithms

By using logarithmic identities to transform the formula, we can express the multiplications as a sum and the power as a multiplication.

$$\left(\prod_{i=1}^n x_i \right)^{1/n} = \exp \left[\frac{1}{n} \sum_{i=1}^n \ln x_i \right]$$

This is sometimes called the log-average. It is simply computing the arithmetic mean of the logarithm transformed values of x_i (i.e. the arithmetic mean on the log scale) and then using the

exponentiation to return the computation to the original scale. I.e., it is the generalised f-mean with $f(x) = \ln x$.

Therefore the geometric mean is related to the log-normal distribution. The log-normal distribution is a distribution which is normal for the logarithm transformed values. We see that the geometric mean is the exponentiated value of the arithmetic mean of the log transformed values.

Harmonic mean

The harmonic mean $H(x_1, \dots, x_n)$ of n numbers x_i (where $i = 1, \dots, n$) is the number H defined by

$$\frac{1}{H} \equiv \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}.$$

The harmonic mean of a list of numbers may be computed in *Mathematica* using `HarmonicMean[list]`.

The special cases of $n = 2$ and $n = 3$ are therefore given by

$$H(x_1, x_2) = \frac{2 x_1 x_2}{x_1 + x_2}$$

$$H(x_1, x_2, x_3) = \frac{3 x_1 x_2 x_3}{x_1 x_2 + x_1 x_3 + x_2 x_3},$$

and so on.

The harmonic means of the integers from 1 to n for $n = 1, 2, \dots$ are 1, 4/3, 18/11, 48/25, 300/137, 120/49, 980/363, ... (Sloane's A102928 and A001008).

For $n = 2$, the harmonic mean is related to the arithmetic mean A and geometric mean G by

$$H = \frac{G^2}{A}$$

(Havil 2003, p. 120).

The harmonic mean is the special case M_{-1} of the power mean and is one of the Pythagorean means. In older literature, it is sometimes called the subcontrary mean.

The volume-to-surface area ratio for a cylindrical container with height h and radius r and the mean curvature of a general surface are related to the harmonic mean.

Hoehn and Niven (1985) show that

$$H(\alpha_1 + c, \alpha_2 + c, \dots, \alpha_n + c) > c + H(\alpha_1, \alpha_2, \dots, \alpha_n)$$

for any positive constant c .

Measures of Variability - Variance and Standard Deviation

The measures of central tendency discussed in the last section are useful because data tend to cluster around central values. However, as the individual values in a distribution of data differ from each other, central values provide us only an incomplete picture of the features of the distribution.

To obtain a more complete picture of the nature of a distribution, the variability (or dispersion or spread) of the data needs to be considered.

The measures of variability for data that we look at are: the range, the mean deviation and the standard deviation.

Range for a Set of Data

Example 4.4.1

When Arnie started attending Fanshawe College, he was keen to arrive at school on time so he kept a record of his travel times to get to school each day for the first ten days. The number of minutes taken each day was:

55 69 93 59 68 75 62 78 97 83 .

This data can be rearranged in ascending order:

55 59 62 68 69 75 78 83 93 97

The range for a set of data is defined as:

$$\text{range} = \text{maximum} - \text{minimum}$$

Arnie's range of travel times = $97 - 55 = 42$ minutes.

Interquartile Ranges

We defined the interquartile range for a set of data earlier. The quartiles Q1, Q2 and Q3 divide a body of data into four equal parts. Q1 is called the lower quartile and contains the lower 25% of the data.

Q2 is the median

Q3 is called the upper quartile and contains the upper 25% of the data.

The interquartile range is $Q3 - Q1$ and is called the IQR. The IQR contains the middle 50% of the data.

A list of values must be written in order from least to greatest before the quartile values can be determined. If the quartile division comes between two values, the quartile value is the average of the two.

Arnie's travel times in ascending order are:

55 59 62 68 68 75 78 83 83 97

Q3 divide a body of data into four equal parts.

Q1 is called the lower quartile and contains the lower 25% of the data.

Q2 is the median

Q3 is called the upper quartile and contains the upper 25% of the data.

The interquartile range is $Q3 - Q1$ and is called the IQR. The IQR contains the middle 50% of the data.

A list of values must be written in order from least to greatest before the quartile values can be determined. If the quartile division comes between two values, the quartile value is the average of the two.

Arnie's travel times in ascending order are:

55 59 62 68 68 75 78 83 83 97

To find the value of Q1, first find its position. There are 10 data items, so the position for Q1 is $10/4 = 2.5$. Since this is between items 2 and 3 we take the average of items 2 and 3.

$$\frac{59 + 62}{2} = 60.5$$

Therefore, $Q1 = 60.5$.

Similarly the median is the average of 68 and 75. The median is 71.5.

The position of Q3 from the upper end of the data is again 2.5. The average of 83 and 83 is 83.

In summary we have: $Q1 = 60.5$, $Q2 = 71.5$ and $Q3 = 83$.

Arnie's average or mean travel time is 72.8 minutes.

The term deviation refers to the difference between the value of an individual observation in a data set and the mean of the data set. Since the mean is a central value, some of the deviations will be positive and some will be negative. The sum of all of the deviations from the mean is always zero. Therefore, the absolute value of the deviations from the mean are used to calculate the mean deviation for a set of data.

The mean deviation for a set of data is the average of the absolute values of all the deviations.

Example 4.4.2

Compute the mean deviation for the traveling times for Arnie in Example 4.1.1

The mean deviation for Arnie's traveling times is computed in the table below:

Commuting Time, x , in minutes	Deviation from Mean \bar{x}	Absolute Deviation from the Mean $ x - \bar{x} $
55	-17.8	17.8
59	-13.8	13.8
62	-10.8	10.8
68	-4.8	4.8
68	-4.8	4.8
75	2.2	2.2
78	5.2	5.2
83	10.2	10.2
83	10.2	10.2
97	24.2	24.2
	Sum of deviations = 0	Sum of Absolute Deviations = 104
mean = 72.8		

minutes		
<p>Mean Deviation = Sum of Absolute Deviations/n = 104/10 = 10.4</p> <p>On the average, each travel time varies 10.4 minutes from the mean travel time of 72.8 minutes.</p>		

The mean deviation tells us that, on average, Arnie's commuting time is 72.8 minutes and that, on average, each commuting time deviates from the mean by 10.4 minutes.

If classes start at 8:00 a.m., can you suggest a time at which Arnie should leave each day?

The Variance and Standard Deviation

The variance is the sum of the squares of the mean deviations divided by n where n is the number of items in the data.

The standard deviation is the square root of the variance. It is the most widely used measure of variability for a set of data.

The formulas for variance and for standard deviation are given:

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

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

Example 4.4.3

Compute the variance and standard deviation for Arnie's traveling times as given in Example 4.4.1.

Solution:

Commuting Times - Arnie from Home to Fanshawe

Commuting (minutes)	Time	Deviation from Mean	Deviation Squared
55		-17.8	316.84
59		-13.8	190.44
62		-10.8	116.64
68		-4.8	23.04
68		-4.8	23.04
75		2.2	4.84

78	5.2	27.04
83	10.2	104.04
83	10.2	104.04
97	24.2	585.64
mean = 72.8		Sum =1495.6
Sample Variance $s^2 = 1495.6/10 = 149.56$	Sample Standard Deviation $s = \sqrt{149.56}$ $= 12.2$	

problem4.4

The following data gives the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees baseball team from 1920 to 1934:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

The following are the number of home runs that Roger Maris hit in each of the ten years he played in the major leagues from 1957 on [data are already arrayed]:

8 13 14 16 23 26 28 33 39 61

Analyze the data in terms of consistency of performance.

Calculate the mean and standard deviation for each player's data and comment on the consistency of performance of each player.

Mean, Median, Mode, and Range

Mean, median, and mode are three kinds of "averages". There are many "averages" in statistics, but these are, I think, the three most common, and are certainly the three you are most likely to encounter in your pre-statistics courses, if the topic comes up at all.

The "mean" is the "average" you're used to, where you add up all the numbers and then divide by the number of numbers. The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order, so you may have to rewrite your list first. The "mode" is the value that

occurs most often. If no number is repeated, then there is no mode for the list.

The "range" is just the difference between the largest and smallest values.

- Find the mean, median, mode, and range for the following list of values:

13, 18, 13, 14, 13, 16, 14, 21, 13

The mean is the usual average, so:

$$(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) \div 9 = 15$$

Note that the mean isn't a value from the original list. This is a common result. You should not assume that your mean will be one of your original numbers.

The median is the middle value, so I'll have to rewrite the list in order:

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the $(9 + 1) \div 2 = 10 \div 2 = 5$ th number:

13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.

The mode is the number that is repeated more often than any other, so 13 is the mode.

The largest value in the list is 21, and the smallest is 13, so the range is $21 - 13 = 8$.

mean:	15
median:	14
mode:	13
range: 8	

Note: The formula for the place to find the median is " $(\text{[the number of data points]} + 1) \div 2$ ", but you don't have to use this formula. You can just count in from both ends of the list until you meet in the middle, if you prefer. Either way will work.

- Find the mean, median, mode, and range for the following list of values:

1, 2, 4, 7

The mean is the usual average: $(1 + 2 + 4 + 7) \div 4 = 14 \div 4 = 3.5$

The median is the middle number. In this example, the numbers are already listed in numerical order, so I don't have to rewrite the list. But there is no "middle" number, because there are an even number of numbers. In this case, the median is the mean (the usual average) of the middle two values: $(2 + 4) \div 2 = 6 \div 2 = 3$

The mode is the number that is repeated most often, but all the numbers appear only once. Then there is no mode.

The largest value is 7, the smallest is 1, and their difference is 6, so the range is 6.

mean:	3.5
median:	3
mode:	none
range:	6

The list values were whole numbers, but the mean was a decimal value. Getting a decimal value for the mean (or for the median, if you have an even number of data points) is perfectly okay; don't round your answers to try to match the format of the other numbers.

- Find the mean, median, mode, and range for the following list of values:

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

The mean is the usual average:

$$(8 + 9 + 10 + 10 + 10 + 11 + 11 + 11 + 12 + 13) \div 10 = 105 \div 10 = 10.5$$

The median is the middle value. In a list of ten values, that will be the $(10 + 1) \div 2 = 5.5$ th value; that is, I'll need to average the fifth and sixth numbers to find the median:

$$(10 + 11) \div 2 = 21 \div 2 = 10.5$$

The mode is the number repeated most often. This list has two values that are repeated three times.

The largest value is 13 and the smallest is 8, so the range is $13 - 8 = 5$.

mean:			10.5
median:			10.5
modes:	10	and	11
range: 5			

While unusual, it can happen that two of the averages (the mean and the median, in this case) will have the same value.

Note: Depending on your text or your instructor, the above data set may be viewed as having no mode (rather than two modes), since no single solitary number was repeated more often than any other. I've seen books that go either way; there doesn't seem to be a consensus on the "right" definition of "mode" in the above case. So if you're not certain how you should answer the "mode" part of the above example, ask your instructor before the next test.

About the only hard part of finding the mean, median, and mode is keeping straight which "average" is which. Just remember the following:

mean:	regular	meaning	of	"average"
median:		middle		value
mode:	most often			

(In the above, I've used the term "average" rather casually. The technical definition of "average" is the arithmetic mean: adding up the values and then dividing by the number of values. Since you're probably more familiar with the concept of "average" than with "measure of central tendency", I used the more comfortable term.)

- A student has gotten the following grades on his tests: 87, 95, 76, and 88. He wants an 85 or better overall. What is the minimum grade he must get on the last test in order to achieve that average?

The unknown score is "x". Then the desired average is:

$$(87 + 95 + 76 + 88 + x) \div 5 = 85$$

Multiplying through by 5 and simplifying, I get:

$$\begin{array}{ccccccccccc}
 87 & + & 95 & + & 76 & + & 88 & + & x & = & 425 \\
 & & 346 & & & + & x & & & = & 425 \\
 & & & & x = 79 & & & & & &
 \end{array}$$

He needs to get at least a 79 on the last test.

Have you understood?

Q1.The number of cars sold by each of the 10 salespeople in an automobile dealership during a particular month, arranged in ascending order is :2,4,7,10,10,10,12,12,14,15.Determine the a)range b)interquartile range and c)middle 80 percent range for these data.

Q2.The weights of a sample of outgoing packages in a mailroom, weighted to the nearest ounce, are found to be:21,18,30,12,14,17,28,10,16,25 oz.Determine the a)range and b)interquartile range for these weights.

Q3.The number of accidents which occurred during a given month in the 13 manufacturing departments of an industrial plant was :2,0,0,3,3,12,1,0,8,1,0,5,1.Determine the a)range and b)interquartile range for the number of accidents.



STATISTICS - PROPORTION

Estimation

Introduction

Quality assurance manager may be interested in estimating the proportion defective of the finished product before shipment to the customer. Manager of the credit department needs to estimate the average collection period for collecting dues from the customers. How confident are they in their estimates? An estimator is a sample statistic used to estimate a population parameter. For example, sample mean is the estimator of population mean. Sample proportion is the estimator of population proportion. An estimate is a specific observed value of a statistic.

Learning objectives

After reading this unit, you will be able to:

Define and compute point estimation

Define and compute interval estimation

Determine sample size based on confidence interval

Criteria for selecting an estimator

Unbiasedness: An estimator is unbiased if its expected value equals the parameter for all sample sizes.

Relative efficiency: An unbiased estimator is relatively efficient when its s.e., is smaller than that of another unbiased estimator of the same parameter.

Consistency: An estimator is consistent if the probability that its value is very near the parameter's value which increasingly approaches 1 as the sample size increases.

Sufficiency: an estimator is sufficient if it makes so much use of the information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

Types of estimation:

There are two types of estimation. They are

- a) Point estimation
- b) Interval estimation

A point estimate is a single valued estimate. For example, estimation of the population mean to be 410 is equal to the sample mean.

An interval estimate is an estimate that is a range of values. For example, estimation of the population means to be 400 to 420.

Sample Proportions and Point Estimation

Sample Proportions

Let p be the proportion of successes of a sample from a population whose total proportion of successes is π and let \bar{p} be the mean of p and σ_p be its standard deviation.

Then

The Central Limit Theorem For Proportions

1. $\bar{p} = \pi$
2. $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
3. For n large, p is approximately normal.

Example

Consider the next census. Suppose we are interested in the proportion of Americans that are below the poverty level. Instead of attempting to find all Americans, Congress has proposed to perform statistical sampling. We can concentrate on 10,000 randomly

selected people from 1000 locations. We can determine the proportion of people below the poverty level in each of these regions. Suppose this proportion is .08. Then the mean for the sampling distribution is

$$\mu_p = 0.08$$

and the standard deviation is

$$\sigma_p = \sqrt{\frac{0.08 \cdot 0.92}{10,000}} = 0.0027$$

Point Estimations

A *Point Estimate* is a statistic that gives a plausible estimate for the value in question.

Example

\bar{x} is a point estimate for μ

s is a point estimate for σ

A point estimate is *unbiased* if its mean represents the value that it is estimating.

Confidence Intervals for a Mean

Point Estimations

Usually, we do not know the population mean and standard deviation. Our goal is to estimate these numbers. The standard way to accomplish this is to use the sample mean and standard deviation as a best guess for the true population mean and standard deviation. We call this "best guess" a *point estimate*.

A *Point Estimate* is a statistic that gives a plausible estimate for the value in question.

Example:

\bar{x} is a point estimate for μ

s is a point estimate for σ

A point estimate is *unbiased* if its mean represents the value that it is estimating.

Confidence Intervals

We are not only interested in finding the point estimate for the mean, but also determining how accurate the point estimate is. The Central Limit Theorem plays a key role here. We assume that the sample standard deviation is close to the population standard deviation (which will almost always be true for large samples). Then the Central Limit Theorem tells us that the standard deviation of the sampling distribution is

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

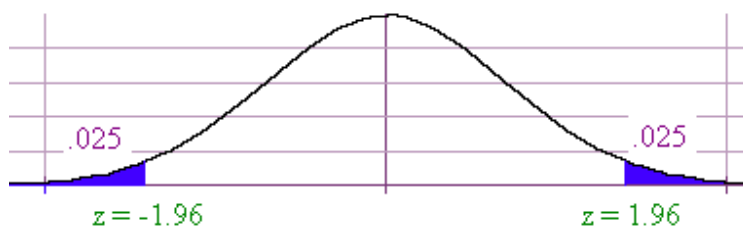
We will be interested in finding an interval around \bar{x} such that there is a large probability that the actual mean falls inside of this interval. This interval is called a *confidence interval* and the large probability is called the *confidence level*.

Example

Suppose that we check for clarity in 50 locations in Lake Tahoe and discover that the average depth of clarity of the lake is 14 feet with a standard deviation of 2 feet. What can we conclude about the average clarity of the lake with a 95% confidence level?

Solution

We can use \bar{x} to provide a point estimate for μ and s to provide a point estimate for σ . How accurate is \bar{x} as a point estimate? We construct a *95% confidence interval for μ* as follows. We draw the picture and realize that we need to use the table to find the z-score associated to the probability of .025 (there is .025 to the left and .025 to the right).



We arrive at $z = -1.96$. Now we solve for \bar{x} :

$$-1.96 = \frac{\bar{x} - 14}{2/\sqrt{50}} = \frac{\bar{x} - 14}{0.28}$$

Hence

$$\bar{x} - 14 = -0.55$$

We say that ± 0.55 is the *margin of error*.

We have that a 95% confidence interval for the mean clarity is

$$(13.45, 14.55)$$

In other words there is a 95% chance that the mean clarity is between 13.45 and 14.55.

In general if z_c is the z value associated with $c\%$ then a $c\%$ confidence interval for the mean is

$$\bar{x} \pm \frac{z_c \sigma}{\sqrt{n}}$$

Confidence Interval for a Small Sample

When the population is normal the sampling distribution will also be normal, but the use of σ to replace s is not that accurate. The smaller the sample size the worse the approximation will be. Hence we can expect that some adjustment will be made based on the sample size. The adjustment we make is that we do not use the normal curve for this approximation. Instead, we use the *Student t* distribution that is based on the sample size. We proceed as before, but we change the table that we use. This distribution looks like the normal distribution, but as the sample size decreases it spreads out. For large n it nearly matches the normal curve. We say that the distribution has $n - 1$ *degrees of freedom*.

Example

Suppose that we conduct a survey of 19 millionaires to find out what percent of their income the average millionaire donates to charity. We discover that the mean percent is 15 with a standard deviation of 5 percent. Find a 95% confidence interval for the mean percent.

Solution

We use the formula:

$$\bar{x} \pm \frac{t_c s}{\sqrt{n}} \quad (\text{Notice the } t \text{ instead of the } z)$$

We get

$$15 \pm t_c 5 / \sqrt{19}$$

Since $n = 19$, there are 18 degrees of freedom. Using the table in the back of the book, we have that

$$t_c = 2.10$$

Hence the margin of error is

$$\pm 2.10 (5) / \sqrt{19} = \pm 2.4$$

We can conclude with 95% confidence that the millionaires donate between

12.6% and 17.4% of their income to charity.

Confidence Intervals For Proportions and

Choosing the Sample Size

A Large Sample Confidence Interval for a Population Proportion

Recall that a confidence interval for a population mean is given by

Confidence Interval for a Population Mean

$$\bar{x} \pm z_c \frac{s}{\sqrt{n}}$$

We can make a similar construction for a confidence interval for a population proportion. Instead of x , we can use p and instead of s , we use $\sqrt{p(1-p)}$, hence, we can write the confidence interval for a large sample proportion as

Confidence Interval Margin of Error for a Population Proportion

$$E = z_c \sqrt{\frac{p(1-p)}{n}}$$

Example

1000 randomly selected Americans were asked if they believed the minimum wage should be raised. 600 said yes. Construct a 95% confidence interval for the proportion of Americans who believe that the minimum wage should be raised.

Solution:

We have

$$p = 600/1000 = .6 \quad z_c = 1.96 \quad \text{and} \quad n = 1000$$

We calculate:

$$0.6 \pm 1.96 \sqrt{\frac{(0.6)(0.4)}{1000}} = 0.6 \pm 0.03$$

Hence we can conclude that between 57 and 63 percent of all Americans agree with the proposal. In other words, with a margin of error of .03, 60% agree.

Calculating n for Estimating a Mean**Example**

Suppose that you were interested in the average number of units that students take at a two year college to get an AA degree. Suppose you wanted to find a 95% confidence interval with a margin of error of .5 for σ knowing $\sigma = 10$. How many people should we ask?

Solution

Solving for n in

$$\text{Margin of Error} = E = \pm z_c \sigma / \sqrt{n}$$

we have

$$E \sqrt{n} = z_c \sigma$$

$$\sqrt{n} = \frac{z_c \sigma}{E}$$

Squaring both sides, we get

$$n = \left(\frac{z_c \sigma}{E} \right)^2$$

We use the formula:

$$n = \left(\frac{1.96(10)}{0.5} \right)^2 = 1,536$$

Example

A Subaru dealer wants to find out the age of their customers (for advertising purposes). They want the margin of error to be 3 years old. If they want a 90% confidence interval, how many people do they need to know about?

Solution:

We have

$$E = 3, \quad z_c = 1.65$$

but there is no way of finding sigma exactly. They use the following reasoning: most car customers are between 16 and 68 years old hence the range is

$$\text{Range} = 68 - 16 = 52$$

The range covers about four standard deviations hence one standard deviation is about

$$\square \square 52/4 = 13$$

We can now calculate n:

$$n = \left(\frac{1.65(13)}{3} \right)^2 = 51.1$$

Hence the dealer should survey at least 52 people.

Finding n to Estimate a Proportion

Example

Suppose that you are in charge to see if dropping a computer will damage it. You want to find the proportion of computers that break. If you want a 90% confidence interval for this proportion, with a margin of error of $\pm 4\%$, How many computers should you drop?

Solution

The formula states that

$$E = z_c \sqrt{\frac{p(1-p)}{n}}$$

Squaring both sides, we get that

$$E^2 = \frac{z_c^2 p(1-p)}{n}$$

Multiplying by n, we get

$$nE^2 = z_c^2 [p(1-p)]$$

$$n = p(1-p) \left(\frac{z_c}{E} \right)^2$$

This is the formula for finding n.

Since we do not know p, we use .5 (A conservative estimate)

$$n = .5(1-.5) \left(\frac{1.65}{0.04} \right)^2 = 425.4$$

We round 425.4 up for greater accuracy

We will need to drop at least 426 computers. This could get expensive.

Estimating Differences

Difference Between Means

I surveyed 50 people from a poor area of town and 70 people from an affluent area of town about their feelings towards minorities. I counted the number of negative comments made. I was interested in comparing their attitudes. The average number of negative comments in the poor area was 14 and in the affluent area was 12. The standard deviations were 5 and 4 respectively. Let's determine a 95% confidence for the difference in mean negative comments. First, we need some formulas.

Theorem

The distribution of the difference of means $\bar{x}_1 - \bar{x}_2$ has mean

$$\mu_1 - \mu_2$$

and standard deviation

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For our investigation, we use s_1 and s_2 as point estimates for μ_1 and μ_2 . We have

$$\begin{array}{llllll} x_1 = 14 & x_2 = 12 & s_1 = 5 & s_2 = 4 & n_1 = & \\ 50 & n_2 = 70 & & & & \end{array}$$

Now calculate

$$\bar{x}_1 - \bar{x}_2 = 14 - 12 = 2$$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{5^2}{50} + \frac{4^2}{70}} = 0.85$$

The margin of error is

$$E = z_c s = (1.96)(0.85) = 1.7$$

The confidence interval is

$$2 \pm 1.7$$

or

$$[0.3, 3.7]$$

We can conclude that the mean difference between the number of racial slurs that poor and wealthy people make is between 0.3 and 3.7.

Note: To calculate the degrees of freedom, we can take the smaller of the two numbers $n_1 - 1$ and $n_2 - 1$. So in the prior example, a better estimate would use 49 degrees of freedom. The t-table gives a value of 2.014 for the $t_{.95}$ value and the margin of error is

$$E = z_c s = (2.014)(0.85) = 1.7119$$

which still rounds to 1.7. This is an example that demonstrates that using the t-table and z-table for large samples results in practically the same results.

Small Samples With Pooled Standard Deviations (Optional)

When either sample size is small, we can still run the statistics provided the distributions are approximately normal. If in addition we know that the two standard deviations are approximately equal, then we can pool the data together to produce a *pooled standard deviation*. We have the following theorem.

Pooled Estimate of σ^2

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

with $n_1 + n_2 - 2$ degrees of freedom

You've gotta love the beautiful formula!

Note

After finding the pooled estimate we have that a confidence interval is given by

$$\bar{x}_1 - \bar{x}_2 \pm t_c s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example

What is the difference between commuting patterns for students and professors. 11 students and 14 professors took part in a study to find mean commuting distances. The mean number of miles traveled by students was 5.6 and the standard deviation was 2.8. The mean number of miles traveled by professors was 14.3 and the standard deviation was 9.1. Construct a 95% confidence interval for the difference between the means. What assumption have we made?

Solution

We have

$$x_1 = 5.6 \quad x_2 = 14.3 \quad s_1 = 2.8 \quad s_2 = 9.1$$

$$n_1 = 11 \quad n_2 = 14$$

The pooled standard deviation is

$$s = \sqrt{\frac{(11-1)2.8^2 + (14-1)9.1^2}{11+14-2}} = 7.09$$

The point estimate for the mean is

$$14.3 - 5.6 = 8.7$$

and

$$\sqrt{\frac{1}{11} + \frac{1}{14}} = .403$$

Use the t-table to find t_c for a 95% confidence interval with 23 degrees of freedom and find

$$t_c = 2.07$$

$$8.7 \pm (2.07)(7.09)(.403) = 8.7 \pm 5.9$$

The range of values is [2.8, 14.6]

The difference in average miles driven by students and professors is between 2.8 and 14.6. We have assumed that the standard deviations are approximately equal and the two distributions are approximately normal.

Difference Between Proportions

So far, we have discussed the difference between two means (both large and small samples). Our next task is to estimate the difference between two proportions. We have the following theorem

And a confidence interval for the difference of proportions is

Confidence Interval for the difference of Proportions

$$p_1 - p_2 \pm z_c \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Note: in order for this to be valid, we need all four of the quantities

$$p_1 n_1 \quad p_2 n_2 \quad q_1 n_1 \quad q_2 n_2$$

to be greater than 5.

Example

300 men and 400 women we asked how they felt about taxing Internet sales. 75 of the men and 90 of the women agreed with having a tax. Find a confidence interval for the difference in proportions.

Solution

We have

$$p_1 = 75/300 = .25 \quad q_1 = .75 \quad n_1 = 300$$

$$p_2 = 90/400 = .225 \quad q_2 = .775 \quad n_2 = 400$$

We can calculate

$$.25 - .225 \pm 1.96 \sqrt{\frac{(.75)(.25)}{300} + \frac{(.775)(.225)}{400}} = -.025 \pm .06$$

We can conclude that the difference in opinions is between -8.5% and 3.5%.

Confidence interval for a mean when the population S.D known

Solve:

As the owner of custom travel, you want to estimate the mean time that it takes a travel agent to make the initial arrangements for vacation package. You have asked your office manager to take a random sample of 40 vacation requests and to observe how long it takes to complete the initial engagements. The office manager reported a mean time of 23.4 min. You want to estimate the true mean time using 95% confidence level. Previous time studies indicate that the S.d of times is a relatively constant 9.8min

Solve:

A machine produces components which have a standard deviation of 1.6cm in length. A random sample of 64 parts is selected from the output and this sample has a mean length of 90cm. The customer will reject the part if it is either less than 88cm or more than 92 cm. Does the 95% confidence interval for the true mean length of all components produced ensure acceptance by the customer?

Confidence interval for the population proportion for large samples

Solve:

In a health survey involving a random sample of 75 patients who developed a particular illness, 70% of them are cured of this illness by a new drug. Establish the 95% confidence interval for the population proportion of all the patients who will be cured by the new drug. This would help to assess the market potential for this new drug by a pharmaceutical company.

Confidence interval for population means for small samples Using T-distribution

Solve:

The average travel time taken based on a random sample of 10 people working in a company to reach the office is 40 minutes with a standard deviation of 10 minutes. Establish the 95% confidence interval for the mean travel time of everyone in the company and re-design the working hours.

Determining the sample size using confidence interval

Solve

A marketing manager of a fast food restaurant in a city wishes to estimate the average yearly amount that families spend on fast food restaurants. He wants the estimate to be within \pm Rs100 with a confidence level of 99%. It is known from an earlier pilot study that the standard deviation of the family expenditure on fast food restaurant is Rs500. How many families must be chosen for this problem?

Sample size determination:-population proportion

Solve:

A company manufacturing sports goods wants to estimate the proportion of cricket players among high school students in India. The company wants the estimate to be within \pm 0.03 with a confidence level of 99%. A pilot study done earlier reveals that out of 80 high school students, 36 students play cricket. What should be the sample size for this study?

Summary

This unit has given a conceptual framework of statistical estimation. In particular, this unit has focused on the following:

- The definition and meaning of point estimation for the population mean and population proportion.
- The role of sample mean and sample proportion in estimating the population mean and population proportion with their property of unbiasedness.
- The conceptual framework of interval estimation with its key elements.
- The methodology for establishing the confidence interval for the population mean and the population proportion based on the sample mean and the sample proportion.
- Examples giving the 95% and 99% confidence interval for the population mean and the population proportion for large samples.
- Establishing confidence interval for small samples using the t distribution after explaining the role of degree of freedom in computing the value of t .
- Determining the optimal sample size based on precision, confidence level, and a knowledge about the population standard deviation.



STATISTICS - HYPOTHESIS

Hypothesis Testing

whenever we have a decision to make about a population characteristic, we make a hypothesis. Some examples are:

$$\mu > 3$$

or

$$\mu \neq 5.$$

Suppose that we want to test the hypothesis that $\mu \neq 5$. Then we can think of our opponent suggesting that $\mu = 5$. We call the opponent's hypothesis the *null hypothesis* and write:

$$H_0: \mu = 5$$

and our hypothesis the alternative hypothesis and write

$$H_1: \mu \neq 5$$

For the null hypothesis we always use equality, since we are comparing μ with a previously determined mean.

For the alternative hypothesis, we have the choices: $<$, $>$, or \neq .

Procedures in Hypothesis Testing

When we test a hypothesis we proceed as follows:

Formulate the null and alternative hypothesis.

1. Choose a level of significance.
2. Determine the sample size. (Same as confidence intervals)
3. Collect data.

4. Calculate z (or t) score.
5. Utilize the table to determine if the z score falls within the acceptance region.
6. Decide to
 - a. Reject the null hypothesis and therefore accept the alternative hypothesis or
 - b. Fail to reject the null hypothesis and therefore state that there is not enough evidence to suggest the truth of the alternative hypothesis.

Errors in Hypothesis Tests

We define a *type I error* as the event of rejecting the null hypothesis when the null hypothesis was true. The probability of a type I error (α) is called the significance level.

We define a type II error (with probability β) as the event of failing to reject the null hypothesis when the null hypothesis was false.

Example

Suppose that you are a lawyer that is trying to establish that a company has been unfair to minorities with regard to salary increases. Suppose the mean salary increase per year is 8%.

You set the null hypothesis to be

$$H_0: \alpha = .08$$

$$H_1: \alpha < .08$$

Q. What is a type I error?

A. We put sanctions on the company, when they were not being discriminatory.

Q. What is a type II error?

A. We allow the company to go about its discriminatory ways.

Note: Larger σ results in a smaller σ , and smaller σ results in a larger σ .

Hypothesis Testing For a Population Mean

The Idea of Hypothesis Testing

Suppose we want to show that only children have an average higher cholesterol level than the national average. It is known that the mean cholesterol level for all Americans is 190. Construct the relevant hypothesis test:

$$H_0: \mu = 190$$

$$H_1: \mu > 190$$

We test 100 only children and find that

$$\bar{x} = 198$$

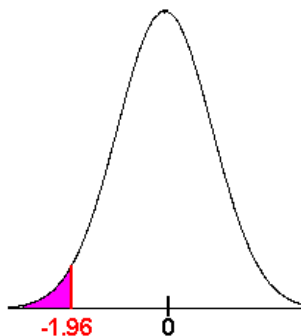
and

$$s = 15.$$

Do we have evidence to suggest that only children have an average higher cholesterol level than the national average? We have

$$z = \frac{198 - 190}{15/\sqrt{100}} = 5.33$$

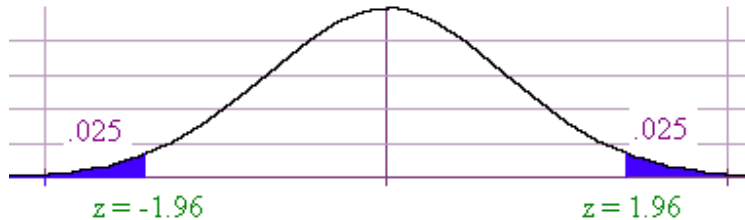
z is called the *test statistic*.



Since z is so high, the probability that H_0 is true is so small that we decide to reject H_0 and accept H_1 . Therefore, we can conclude that only children have a higher cholesterol level on the average than the national average.

Rejection Regions

Suppose that $\alpha = .05$. We can draw the appropriate picture and find the z score for -.025 and .025. We call the outside regions the rejection regions.



We call the blue areas the *rejection region* since if the value of z falls in these regions, we can say that the null hypothesis is very unlikely so we can reject the null hypothesis

Example

50 smokers were questioned about the number of hours they sleep each day. We want to test the hypothesis that the smokers need less sleep than the general public which needs an average of 7.7 hours of sleep. We follow the steps below.

Compute a rejection region for a significance level of .05.

- A. If the sample mean is 7.5 and the standard deviation is .5, what can you conclude?

Solution

First, we write down the null and alternative hypotheses

$$H_0: \mu = 7.7 \quad H_1: \mu < 7.7$$

This is a left tailed test. The z-score that corresponds to .05 is -1.96. The critical region is the area that lies to the left of -1.96. If the z-value is less than -1.96 there we will reject the null hypothesis and accept the alternative hypothesis. If it is greater than -1.96, we will fail to reject the null hypothesis and say that the test was not statistically significant.

We have

$$z = \frac{7.5 - 7.7}{.5 / \sqrt{50}} = -2.83$$

Since -2.83 is to the left of -1.96, it is in the critical region. Hence we reject the null hypothesis and accept the alternative hypothesis. We can conclude that smokers need less sleep.

p-values

There is another way to interpret the test statistic. In hypothesis testing, we make a yes or no decision without discussing borderline cases. For example with $\alpha = .06$, a two tailed test will indicate rejection of H_0 for a test statistic of $z = 2$ or for $z = 6$, but $z = 6$ is much stronger evidence than $z = 2$. To show this difference we write the *p-value* which is the lowest significance level such that we will still reject H_0 . For a two tailed test, we use twice the table value to find p , and for a one tailed test, we use the table value.

Example:

Suppose that we want to test the hypothesis with a significance level of .05 that the climate has changed since industrialization. Suppose that the mean temperature throughout history is 50 degrees. During the last 40 years, the mean temperature has been 51 degrees with a standard deviation of 2 degrees. What can we conclude?

We have

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

We compute the z score:

$$z = \frac{51-50}{2/\sqrt{40}} = 3.16$$

The table gives us .9992

so that

$$p = (1 - .9992)(2) = .002$$

since

$$.002 < .05$$

we can conclude that there has been a change in temperature.

Note that small p-values will result in a rejection of H_0 and large p-values will result in failing to reject H_0 .

Hypothesis Testing for a Proportion and for Small Samples

Small Sample Hypothesis Tests For a Normal population

When we have a small sample from a normal population, we use the same method as a large sample except we use the t statistic instead of the z-statistic. Hence, we need to find the degrees of freedom ($n - 1$) and use the t-table in the back of the book.

Example

Is the temperature required to damage a computer on the average less than 110 degrees? Because of the price of testing, twenty computers were tested to see what minimum temperature will damage the computer. The damaging temperature averaged 109 degrees with a standard deviation of 3 degrees. (use $\alpha = 0.05$)

We test the hypothesis

$$H_0: \mu = 110$$

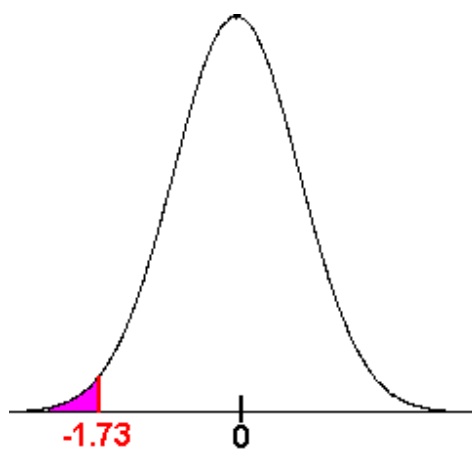
$$H_1: \mu < 110$$

We compute the t statistic:

$$t = \frac{109 - 110}{3/\sqrt{20}} = -1.49$$

This is a one tailed test, so we can go to our t-table with 19 degrees of freedom to find that

$$t_c = 1.73$$



Since

$$-1.49 > -1.73$$

We see that the test statistic does not fall in the critical region. We fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that the temperature required to damage a computer on the average less than 110 degrees.

Hypothesis Testing for a Population Proportion

We have seen how to conduct hypothesis tests for a mean. We now turn to proportions. The process is completely analogous, although we will need to use the standard deviation formula for a proportion.

Example

Suppose that you interview 1000 exiting voters about who they voted for governor. Of the 1000 voters, 550 reported that they voted for the democratic candidate. Is there sufficient evidence to suggest that the democratic candidate will win the election at the .01 level?

$$H_0: p = .5$$

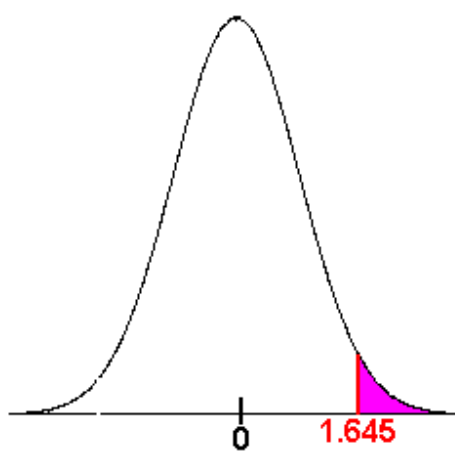
$$H_1: p > .5$$

Since it a large sample we can use the central limit theorem to say that the distribution of proportions is approximately normal. We compute the test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.6 - 0.5}{\sqrt{0.5(1 - 0.5)/1000}} = 3.16$$

Notice that in this formula, we have used the hypothesized proportion rather than the sample proportion. This is because if the null hypothesis is correct, then .5 is the true proportion and we are not making any approximations. We compute the rejection region using the z-table. We find that $z_c = 2.33$.

The picture shows us that 3.16 is in the rejection region. Therefore we reject H_0 so can conclude that the democratic candidate will win with a p-value of .0008.



Example

1500 randomly selected pine trees were tested for traces of the Bark Beetle infestation. It was found that 153 of the trees showed such traces. Test the hypothesis that more than 10% of the Tahoe trees have been infested. (Use a 5% level of significance)

Solution

The hypothesis is

$$H_0: p = .1$$

$$H_1: p > .1$$

We have that

$$\hat{p} = \frac{153}{1500} = .102$$

Next we compute the z-score

$$z = \frac{0.102 - 0.1}{\sqrt{0.1(1-0.1)/1500}} = 0.26$$

Since we are using a 95% level of significance with a one tailed test, we have $z_c = 1.645$. The rejection region is shown in the picture. We see that 0.26 does not lie in the rejection region, hence we fail to reject the null hypothesis. We say that there is insufficient evidence to make a conclusion about the percentage of infested pines being greater than 10%.

Exercises

- A. If 40% of the nation is registered republican. Does the Tahoe environment reflect the national proportion? Test the hypothesis that Tahoe residents differ from the rest of the nation in their affiliation, if of 200 locals surveyed, 75 are registered republican.
- B. If 10% of California residents are vegetarians, test the hypothesis that people who gamble are less likely to be vegetarians. If the 120 people polled, 10 claimed to be a vegetarian.

Difference Between Means

Hypothesis Testing of the Difference Between Two Means

Do employees perform better at work with music playing. The music was turned on during the working hours of a business with 45 employees. Their productivity level averaged 5.2 with a standard deviation of 2.4. On a different day the music was turned off and there were 40 workers. The workers' productivity level averaged 4.8 with a standard deviation of 1.2. What can we conclude at the .05 level?

Solution

We first develop the hypotheses

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

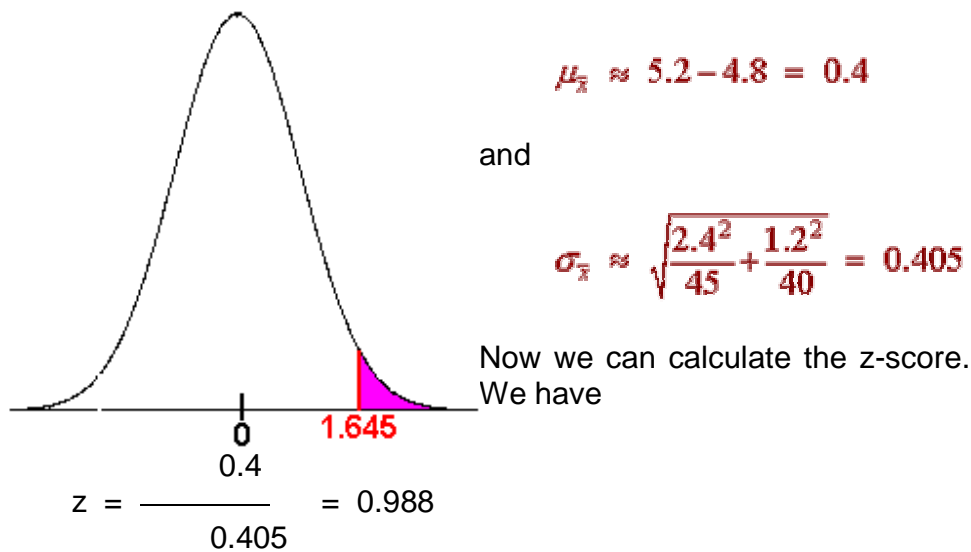
Next we need to find the standard deviation. Recall from before, we had that the mean of the difference is

$$\bar{x} = \bar{x}_1 - \bar{x}_2$$

and the standard deviation is

$$s_x = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We can substitute the sample means and sample standard deviations for a point estimate of the population means and standard deviations. We have



Since this is a one tailed test, the critical value is 1.645 and 0.988 does not lie in the critical region. We fail to reject the null hypothesis and conclude that there is insufficient evidence to conclude that workers perform better at work when the music is on. Using the P-Value technique, we see that the P-value associated with 0.988 is

$$P = 1 - 0.8389 = 0.1611$$

which is larger than 0.05. Yet another way of seeing that we fail to reject the null hypothesis.

Note: It would have been slightly more accurate had we used the t-table instead of the z-table. To calculate the degrees of freedom, we can take the smaller of the two numbers $n_1 - 1$ and $n_2 - 1$. So in this example, a better estimate would use 39 degrees of freedom. The t-table gives a value of 1.690 for the $t_{.95}$ value. Notice that 0.988 is still smaller than 1.690 and the result is the same. This is an example that demonstrates that using the t-table and z-table for large samples results in practically the same results.

Hypothesis Testing For a Difference Between Means for Small Samples Using Pooled Standard Deviations (Optional)

Recall that for small samples we need to make the following assumptions:

1. Random unbiased sample.
2. Both population distributions are normal.

3. The two standard deviations are equal.

If we know σ , then the sampling standard deviation is:

$$s = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$= \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If we do not know σ then we use the pooled standard deviation.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Putting this together with hypothesis testing we can find the t-statistic.

$$t = \frac{x_1 - x_2 - \text{hypothesized difference}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and use $n_1 + n_2 - 2$ degrees of freedom.

Example

Nine dogs and ten cats were tested to determine if there is a difference in the average number of days that the animal can survive without food. The dogs averaged 11 days with a standard deviation of 2 days while the cats averaged 12 days with a standard deviation of 3 days. What can be concluded? (Use $\alpha = .05$)

Solution

We write:

$$H_0: \square_{\text{dog}} - \square_{\text{cat}} = 0$$

$$H_1: \square_{\text{dog}} - \square_{\text{cat}} \neq 0$$

We have:

$$n_1 = 9, \quad n_2 = 10$$

$$x_1 = 11, \quad x_2 = 12$$

$$s_1 = 2, \quad s_2 = 3$$

so that

$$s_p = \sqrt{\frac{(9-1)(4) + (10-1)(9)}{9+10-2}} = 2.58$$

and

$$t = \frac{12-11-0}{2.58\sqrt{\frac{1}{9} + \frac{1}{10}}} = 0.84$$

The t-critical value corresponding to $\alpha = .05$ with $10 + 9 - 2 = 17$ degrees of freedom is 2.11 which is greater than .84. Hence we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that there is a difference between the mean starvation time for cats and dogs.

Hypothesis Testing for a Difference Between Proportions**Inferences on the Difference Between Population Proportions**

If two samples are counted independently of each other we use the test statistic:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

where

$$p = \frac{r_1 + r_2}{n_1 + n_2}$$

and

$$q = 1 - p$$

Example

Is the severity of the drug problem in high school the same for boys and girls? 85 boys and 70 girls were questioned and 34 of the boys and 14 of the girls admitted to having tried some sort of drug. What can be concluded at the .05 level?

Solution

The hypotheses are

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

We have

$$p_1 = 34/85 = 0.4 \quad p_2 = 14/70 = 0.2$$

$$p = 48/155 = 0.31 \quad q = 0.69$$

Now compute the z-score

$$z = \frac{0.4 - 0.2}{\sqrt{\frac{(0.31)(0.69)}{85} + \frac{(0.31)(0.69)}{70}}} = 2.68$$

Since we are using a significance level of .05 and it is a two tailed test, the critical value is 1.96. Clearly 2.68 is in the critical region, hence we can reject the null hypothesis and accept the alternative hypothesis and conclude that gender does make a difference for drug use. Notice that the P-Value is

$$P = 1 - .9963 = 0.0037$$

is less than .05. Yet another way to see that we reject the null hypothesis.

Paired Differences

Paired	Data:	Hypothesis	Tests
--------	-------	------------	-------

Example

Is success determined by genetics?

The best such survey is one that investigates identical twins who have been reared in two different environments, one that is nurturing and one that is non-nurturing. We could measure the difference in high school GPAs between each pair. This is better than just pooling each group individually. Our hypotheses are

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

where μ_d is the mean of the differences between the matched pairs.

We use the test statistic

$$z = \frac{\mu_d - 0}{s_d / \sqrt{n}}$$

where s_d is the standard deviation of the differences.

For a small sample we use $n - 1$ degrees of freedom, where n is the number of pairs.

Paired Differences: Confidence Intervals

To construct a confidence interval for the difference of the means we use:

$$\bar{x}_d \pm t s_d / \sqrt{n}$$

Example

Suppose that ten identical twins were reared apart and the mean difference between the high school GPA of the twin brought up in wealth and the twin brought up in poverty was 0.07. If the standard deviation of the differences was 0.5, find a 95% confidence interval for the difference.

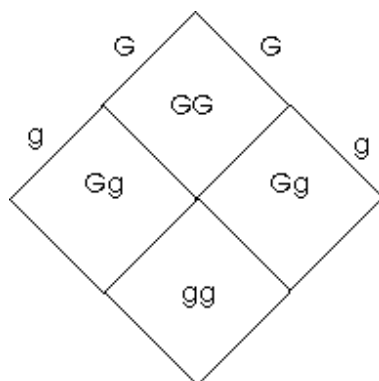
Solution

We compute

$$0.07 \pm 2.26 \frac{0.5}{\sqrt{10}} = 0.07 \pm 0.36$$

or

$$[-0.29, 0.43]$$



We are 95% confident that the mean difference in GPA is between -0.29 and 0.43. Notice that 0 falls in this interval, hence we would fail to reject the null hypothesis at the 0.05 level

Chi-Square Test

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if, according to Mendel's laws, you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the "goodness to fit" between the observed and expected. Were the

deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before you, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what scientists call the **null hypothesis**, which states that there is no significant difference between the expected and observed result.

The formula for calculating chi-square (χ^2) is:

$$\chi^2 = \sum (o-e)^2/e$$

That is, chi-square is the sum of the squared difference between observed (o) and the expected (e) data (or the deviation, d), divided by the expected data in all possible categories.

For example, suppose that a cross between two pea plants yields a population of 880 plants, 639 with green seeds and 241 with yellow seeds. You are asked to propose the genotypes of the parents. Your *hypothesis* is that the allele for green is dominant to the allele for yellow and that the parent plants were both heterozygous for this trait. If your hypothesis is true, then the predicted ratio of offspring from this cross would be 3:1 (based on Mendel's laws) as predicted from the results of the Punnett square (Figure B. 1).

- **Punnett Square.** Predicted offspring from cross between green and yellow-seeded plants. Green (G) is dominant (3/4 green; 1/4 yellow).

To calculate χ^2 , first determine the number *expected* in each category. If the ratio is 3:1 and the total number of observed individuals is 880, then the *expected numerical values* should be 660 green and 220 yellow.

Chi-square requires that you use numerical values, not percentages or ratios.

Then calculate χ^2 using this formula, as shown in Table B.1. Note that we get a value of 2.668 for χ^2 . But what does this number mean? Here's how to interpret the χ^2 value:

1. Determine degrees of freedom (df). Degrees of freedom can be calculated as the number of categories in the problem minus 1. In

our example, there are two categories (green and yellow); therefore, there is 1 degree of freedom.

2. Determine a relative standard to serve as the basis for accepting or rejecting the hypothesis. The relative standard commonly used in biological research is $p > 0.05$. The p value is the *probability* that the deviation of the observed from that expected is due to chance alone (no other forces acting). In this case, using $p > 0.05$, you would expect any deviation to be due to chance alone 5% of the time or less.

3. Refer to a chi-square distribution table (Table B.2). Using the appropriate degrees of freedom, locate the value closest to your calculated chi-square in the table. Determine the closest p (probability) value associated with your chi-square and degrees of freedom. In this case ($\chi^2 = 2.668$), the p value is about 0.10, which means that there is a 10% probability that any deviation from expected results is due to chance only. Based on our standard $p > 0.05$, this is within the range of acceptable deviation. In terms of your hypothesis for this example, the observed chi-square is not significantly different from expected. The observed numbers are consistent with those expected under Mendel's law.

Step-by-Step Procedure for Testing Your Hypothesis and Calculating Chi-Square

1. State the hypothesis being tested and the predicted results. Gather the data by conducting the proper experiment (or, if working genetics problems, use the data provided in the problem).

2. Determine the expected numbers for each observational class. Remember to use numbers, not percentages.

Chi-square should not be calculated if the expected value in any category is less than 5.

3. Calculate χ^2 using the formula. Complete all calculations to three significant digits. Round off your answer to two significant digits.

4. Use the chi-square distribution table to determine significance of the value.

- a. Determine degrees of freedom and locate the value in the appropriate column.

- b. Locate the value closest to your calculated χ^2 on that degrees of freedom df row.
 - c. Move up the column to determine the p value.
5. State your conclusion in terms of your hypothesis.
 - a. If the p value for the calculated χ^2 is $p > 0.05$, accept your hypothesis. The deviation is small enough that chance alone accounts for it. A p value of 0.6, for example, means that there is a 60% probability that any deviation from expected is due to chance only. This is within the range of acceptable deviation.
 - b. If the p value for the calculated χ^2 is $p < 0.05$, reject your hypothesis, and conclude that some factor other than chance is operating for the deviation to be so great. For example, a p value of 0.01 means that there is only a 1% chance that this deviation is due to chance alone. Therefore, other factors must be involved.

The chi-square test will be used to test for the "goodness to fit" between observed and expected data from several laboratory investigations in this lab manual.

Table
Calculating Chi-Square

B.1

	Green	Yellow
Observed (o)	639	241
Expected (e)	660	220
Deviation (o - e)	-21	21
Deviation ² (d2)	441	441
d^2/e	0.668	2
$\chi^2 = d^2/e = 2.668$.	.

Table
Chi-Square Distribution

B.2

Degrees of Freedom (df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83

2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Source: R.A. Fisher and F. Yates, Statistical Tables for Biological Agricultural and Medical Research, 6th ed., Table IV, Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Chi-square test

Purpose:

The chi-square test (Snedecor and Cochran, 1989) is used to test if a sample of data came from a population with a specific distribution.

An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the chi-square test. However, the value of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

The chi-square test is defined for the hypothesis:

H_0 : The data follow a specified distribution.

H_a : The data do not follow the specified distribution.

Test statistic:

For the chi-square goodness-of-fit computation, the data are divided into k bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^k (O_i - E_i)^2 / E_i$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i . The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l))$$

where F is the cumulative Distribution function for the distribution being tested, Y_u is the upper limit for class i , Y_l is the lower limit for class i , and N is the sample size.

This test is sensitive to the choice of bins. There is no optimal choice for the bin width (since the optimal bin width depends on the distribution). Most reasonable choices should produce similar, but not identical, results. Dataplot uses $0.3*s$, where s is the sample standard deviation, for the class width. The lower and upper bins are at the sample mean plus and minus $6.0*s$, respectively. For the chi-square approximation to be valid, the expected frequency should be at least 5. This test is not valid for small samples, and if some of the counts are less than five, you may need to combine some bins in the tails.

Significance interval: alpha

Critical region:

The test statistic follows, approximately, a chi-square distribution with $(k - c)$ degrees of freedom where k is the number of non-empty cells and $c =$ the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution + 1. For example, for a 3-parameter Weibull distribution, $c = 4$.

Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha, k-c)}$$

Where $\chi^2_{(\alpha, k-c)}$ is the chi-square percent point function with $k - c$ degrees of freedom and a significance level of α .

In the above formulas for the critical regions, the Handbook follows the convention that χ^2_{α} is the upper critical value from the chi-square distribution and $\chi^2_{1-\alpha}$ is the lower critical value from the chi-square distribution. Note that this is the opposite of what is used in some texts and software programs. In particular, Data plot uses the opposite convention.

The chi-square test can be used to answer the following types of questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?

- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?

Are the data from a binomial distribution?

Importance:

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the distribution of the data follows a Weibull distribution.

There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique, such as the sign test, that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumption can be confirmed, the parametric techniques are generally preferred.

If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, a non-parametric or robust technique may be required.

Example

The chi-square statistic for the above example is computed as follows:

$$\begin{aligned} X^2 &= (49 - 46.1)^2/46.1 + (50 - 54.2)^2/54.2 + (69 - 67.7)^2/67.7 + \dots + \\ &\quad (28 - 27.8)^2/27.8 \\ &= 0.18 + 0.33 + 0.03 + \dots + 0.01 \\ &= 1.51 \end{aligned}$$

The degrees of freedom are equal to $(3-1)(3-1) = 2 \times 2 = 4$, so we are interested in the probability $P(\chi^2 > 1.51) = 0.8244$ on 4 degrees of freedom. This indicates that there is no association between the choice of most important factor and the grade of the student -- the difference between observed and expected values under the null hypothesis is negligible.

Examples of statistical tests used to analyze some basic experiments

Click on the name of the test to see an example, or scroll down to the examples given below the table. This Table is modified from Motulsky, H., **Intuitive Biostatistics**. Oxford University Press, New York, 1995, p. 298.

Data comparisons you are making	When your data are normally distributed	When your data are not normally-distributed, or are ranks or scores	When your data are Binomial (possess 2 possible values)
You are studying one set of data	Find the mean, standard deviation	Find the median, interquartile range (Q_3-Q_1)	Calculate a proportion
Compare one set of data to a hypothetical value	Run a one-sample t-test	Run a Wilcoxon Test	Run a χ^2 (chi-square) test
Compare 2 sets of independently-collected data	Run a 2-sample t-test	Run a Mann-Whitney Test	Run a Fisher test, or a χ^2 (chi-square) test
Compare 2 sets of data from the same subjects under different circumstances	Run a t-test on the differences between the data values (a matched-pairs t-test)	Run a Wilcoxon Test	Run a McNemar's test
Compare 3 or more sets of data	Run a one-way ANOVA test	Run a Kruskal-Wallis test	Run a chi-square test
Look for a relationship between 2 variables	Calculate the Pearson Correlation coefficient	Calculate the Spearman Correlation coefficient	Calculate Contingency Correlation coefficients
Look for a linear relationship between 2 variables	Run a linear regression	Run a nonparametric linear regression	Run a simple logistic regression

Data comparisons you are making	When your data are normally distributed	When your data are not normally-distributed, or are ranks or scores	When your data are Binomial (possess 2 possible values)
Look for a non-linear relationship between 2 variables	Run a power, exponential, or quadratic regression	Run a nonparametric power, exponential, or quadratic regression	
Look for linear relationships between 1 dependent variable and 2 or more independent variables	Run a multiple linear regression		Run a multiple logistic regression
See your teacher for specific details for analyses required in your particular class. pcbryan@prodigy.net 7-			

1. You read of a survey that claims that the average teenager watches 25 hours of TV a week and you want to check whether or not this is true in your school (too simple a project!).

Predicted value of the variable variable: the predicted 25 hours of TV

Variable under study: actual hours of TV watched

Statistical test you would use: **t-test**

Use this test to compare the mean values (averages) of **one set of data** to a predicted mean value.

2. You grow 20 radish plants in pH=10.0 water and 20 plants in pH=3.0 water and measure the final mass of the leaves of the plants (too simple an experiment!) to see if they grew better in one fluid than in the other fluid.

Independent variable: pH of the fluid in which the plants were grown

Dependent variable: plant biomass

Statistical test you would use: **2-sample t-test**

Use this test to compare the mean values (averages) of **two sets of data**.

A **Mann-Whitney** test is a 2-sample t-test that is run on data that are given rank numbers, rather than quantitative values. For example, You want to compare the overall letter-grade GPA of students in one class with the overall letter-grade GPA of students in another class. You rank the data from low to high according to the letter grade (here, A = 1, B = 2, C = 3, D = 4, E = 5 might be your rankings; you could also have set A = 5, B = 4, ...).

3. You give a math test to a group of students. Afterwards you tell ? of the students a method of analyzing the problems, then re-test all the students to see if use of the method led to improved test scores.

Independent variable: test-taking method (your method vs. no imparted method)

Dependent variable: (test scores after method - test scores before method)

Statistical test you would use: **matched-pairs t-test**

Use this test to compare data **from the same subjects** under two different conditions.

4. You grow radish plants given pesticide-free water every other day, radish plants given a 5% pesticide solution every other day, and radish plants given a 10% pesticide solution every other day, then measure the biomass of the plants after 30 days to find whether there was any difference in plant growth among the three groups of plants.

Independent variable: pesticide dilution

Dependent variable: plant biomass

Statistical test you would use: **ANOVA**

Use this test to compare the mean values (averages) of **more than two sets of data where there is more than one independent variable but only one dependent variable**. If you find that your data differ significantly, this says only that at least two of the data

sets differ from one another, not that all of your tested data sets differ from one another.

If your ANOVA test indicates that there is a statistical difference in your data, you should also run Bonferroni paired t-tests to see which independent variables produce significantly different results. This test essentially penalizes you more and more as you add more and more independent variables, making it more difficult to reject the null hypothesis than if you had tested fewer independent variables.

One assumption in the ANOVA test is that your data are normally-distributed (plot as a bell curve, approximately). If this is not true, you must use the **Kruskall-Wallis** test below.

5. You ask children, teens, and adults to rate their response to a set of statements, where 1 = strongly agree with the statement, 2 = agree with the statement, 3 = no opinion, 4 = disagree with the statement, 5 = strongly disagree with the statement, and you want to see if the answers are dependent on the age group of the tested subjects.

Independent variables: age groups of subject

Dependent variable: responses of members of those age groups to your statements

Statistical test you would use: **Kruskall-Wallis Test**. Use this test to compare the mean values (averages) of **more than two sets of data** where the data are chosen from some limited set of values or if your data otherwise don't form a normal (bell-curve) distribution. This example could also be done using a **two-way chi-square test**.

An example of the Kruskal-Wallis Test for non-normal data is: You compare scores of students on Math and English tests under different circumstances: no music playing, Mozart playing, rock music playing. When you score the tests, you find in at least one case that the average score is a 95 and the data do not form a bell-shaped curve because there are no scores above 100, many scores in the 90s, a few in the 80s, and fewer still in the 70s, for example.

Independent variables: type of background music

Dependent variable: score on the tests , with at least one set of scores not normally-distributed

6. You think that student grades are dependent on the number of hours a week students study. You collect letter grades from students and the number of hours each student studies in a week.

Independent variables: hours studied

Dependent variable: letter grade in a specific class

Statistical test you would use: **Wilcoxon Signed Rank Test**. Use this test to compare the mean values (averages) of **two sets of data**, or the mean value of one data set to a hypothetical mean, where the data are ranked from low to high (here, A = 1, B = 2, C = 3, D = 4, E = 5 might be your rankings; you could also have set A = 5, B = 4, ...).

7. You ask subjects to rate their response to a set of statements that are provided with a set of possible responses such as: strongly agree with the statement, agree with the statement, no opinion, disagree with the statement, strongly disagree with the statement.

Independent variable: each statement asked

Dependent variable: response to each statement

Statistical test you would use: **χ^2 (chi-square) test** (the 'chi' is pronounced like the 'chi' in 'chiropractor') for within-age-group variations.

For this test, typically, you assume that all choices are equally likely and test to find whether this assumption was true. You would assume that, for 50 subjects tested, 10 chose each of the five options listed in the example above. In this case, your observed values (O) would be the number of subjects who chose each response, and your expected values (E) would be 10.

The chi-square statistic is the sum of: $(\text{Observed value} - \text{Expected value})^2 / \text{Expected value}$

Use this test when your data consist of a limited number of possible values that your data can have. **Example 2:** you ask subjects which toy they like best from a group of toys that are identical except that they come in several different colors. Independent variable: toy color; dependent variable: toy choice.

McNemar's test is used when you are comparing some aspect of the subject with that subject's response (i.e., answer to the survey compared to whether or not the student went to a particular middle school). McNemar's test is basically the same as a chi-square test in calculation and interpretation.

8. You look for a relationship between the size of a letter that a subject can read at a distance of 5 meters and the score that the subject achieves in a game of darts (having had them write down their experience previously at playing darts).

Independent variable #1: vision-test result (letter size)

Independent variable #2: darts score

Statistical test you would use: **Correlation (statistics: r^2 and r)**

Use this statistic to identify whether changes in one independent variable are matched by changes in a second independent variable. Notice that you didn't change any conditions of the test, you only made two separate sets of measurements

9. You load weights on four different lengths of the same type and cross-sectional area of wood to see if the maximum weight a piece of the wood can hold is directly dependent on the length of the wood.

Independent variable: length of wood

Dependent variable: weight that causes the wood to break

Statistical test you would use: **Linear regression (statistics: r^2 and r)**

Fit a line to data having only one independent variable and one dependent variable.

10. You load weights on four different lengths and four different thicknesses of the same type of wood to see if the maximum weight a piece of the wood can hold is directly dependent on the length and thickness of the wood, and to find which is more important, length or weight.

Independent variables: length of wood, weight of wood

Dependent variable: weight that causes the wood to break

Statistical test you would use: **Multiple Linear regression (statistics: r^2 and r)**

Fit a line to data having two or more independent variables and one dependent variable.

11. You load weights on strips of plastic trash bags to find how much the plastic stretches from each weight. Research that you do indicates that plastics stretch more and more as the weight placed on them increases; therefore the data do *not* plot along a straight line.

Independent variables: weight loaded on the plastic strip

Dependent variable: length of the plastic strip

Statistical test you would use: **Power regression of the form $y = ax^b$, or Exponential regression of the form $y = ab^x$, or Quadratic regression of the form $y = a + bx + cx^2$ (statistics: r^2 and r)**

Fit a curve to data having only one independent variable and one dependent variable.

There are numerous polynomial regressions of this form, found on the STAT:CALC menu of your graphing calculator.

Paired Sample t-test

A paired sample t-test is used to determine whether there is a significant difference between the average values of the same measurement made under two different conditions. Both measurements are made on each unit in a sample, and the test is based on the paired differences between these two values. The usual null hypothesis is that the difference in the mean values is zero. For example, the yield of two strains of barley is measured in successive years in twenty different plots of agricultural land (the units) to investigate whether one crop gives a significantly greater yield than the other, on average.

The null hypothesis for the paired sample t-test is

$$H_0: d = \mu_1 - \mu_2 = 0$$

where d is the mean value of the difference.

This null hypothesis is tested against one of the following alternative hypotheses, depending on the question posed:

H1:	d	$=$	0
H1:	d	$>$	0
H1:	$d < 0$		

The paired sample t-test is a more powerful alternative to a two sample procedure, such as the two sample t-test, but can only be used when we have matched samples.

Student's t Distribution

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic** (also known as the **t score**), whose values are given by:

$$t = [x - \mu] / [s / \sqrt{n}]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the t statistic is called the **t distribution** or the **Student t distribution**.

Degrees of Freedom

There are actually many different t distributions. The particular form of the t distribution is determined by its **degrees of freedom**. The degrees of freedom refers to the number of independent observations in a set of data.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the t statistic from samples of size 8 would be described by a t distribution having $8 - 1$ or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.

Properties of the t Distribution

The t distribution has the following properties:

- The mean of the distribution is equal to 0 .
- The variance is equal to $v / (v - 2)$, where v is the degrees of freedom (see last section) and $v > 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

When to Use the t Distribution

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The central limit theorem states that the sampling distribution of a statistic will be normal or nearly normal, if any of the following conditions apply.

- The population distribution is normal.
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

The t distribution should *not* be used with small samples from populations that are not approximately normal.

Probability and the Student t Distribution

When a sample of size n is drawn from a population having a normal (or nearly normal) distribution, the sample mean can be transformed into a t score, using the equation presented at the beginning of this lesson. We repeat that equation below:

$$t = [x - \mu] / [s / \sqrt{n}]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, n is the sample size, and degrees of freedom are equal to $n - 1$.

The t score produced by this transformation can be associated with a unique cumulative probability. This cumulative probability represents the likelihood of finding a sample mean less than or equal to x , given a random sample of size n .

The easiest way to find the probability associated with a particular t score is to use the T Distribution Calculator, a free tool provided by Stat Trek.

Notation and t Scores

Statisticians use t_α to represent the t-score that has a cumulative probability of $(1 - \alpha)$. For example, suppose we were interested in the t-score having a cumulative probability of 0.95. In this example, α would be equal to $(1 - 0.95)$ or 0.05. We would refer to the t-score as $t_{0.05}$.

Of course, the value of $t_{0.05}$ depends on the number of degrees of freedom. For example, with 2 degrees of freedom, that $t_{0.05}$ is equal to 2.92; but with 20 degrees of freedom, that $t_{0.05}$ is equal to 1.725.

Note: Because the t distribution is symmetric about a mean of zero, the following is true.

$$t_{\alpha} = -t_{1 - \alpha} \quad \text{And} \quad t_{1 - \alpha} = -t_{\alpha}$$

Thus, if $t_{0.05} = 2.92$, then $t_{0.95} = -2.92$.

T Distribution Calculator

The T Distribution Calculator solves common statistics problems, based on the t distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables menu item, which appears in the header of every Stat Trek web page.

T Distribution
Calculator

Test Your Understanding of This Lesson

Problem 1

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

Note: There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented below. Solution A is the traditional approach. It requires you to compute the t score, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability. Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t score "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

Solution A

The first thing we need to do is compute the t score, based on the following equation:

$$t = \frac{[x - \mu]}{[s / \sqrt{n}]} \\ t = (290 - 300) / [50 / \sqrt{15}] = -10 / 12.909945 = -0.7745966$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

Now, we are ready to use the T Distribution Calculator. Since we know the t score, we select "T score" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to $15 - 1 = 14$.
- The t score is equal to -0.7745966 .

The calculator displays the cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

Solution B:

This time, we will work directly with the raw data from the problem. We will not compute the t score; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to $15 - 1 = 14$.
- Assuming the CEO's claim is true, the population mean equals 300.
- The sample mean equals 290.
- The standard deviation of the sample is 50.

The calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

Problem 2

Suppose scores on an IQ test are normally distributed, with a mean of 100. Suppose 20 people are randomly selected and tested. The standard deviation in the sample group is 15. What is the probability that the average test score in the sample group will be at most 110?

Solution:

To solve this problem, we will work directly with the raw data from the problem. We will not compute the t score; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to $20 - 1 = 19$.
- The population mean equals 100.
- The sample mean equals 110.
- The standard deviation of the sample is 15.

We enter these values into the T Distribution Calculator. The calculator displays the cumulative probability: 0.996. Hence, there is a 99.6% chance that the sample average will be no greater than 110.

t-Test for the Significance of the Difference between the Means of Two Independent Samples

This is probably the most widely used statistical test of all time, and certainly the most widely known. It is simple, straightforward, easy to use, and adaptable to a broad range of situations. No statistical toolbox should ever be without it.

Its utility is occasioned by the fact that scientific research very often examines the phenomena of nature two variables at a time, with an eye toward answering the basic question: Are these two variables related? If we alter the level of one, will we thereby alter the level of the other? Or alternatively: If we examine two different levels of one variable, will we find them to be associated with different levels of the other?

Here are three examples to give you an idea of how these abstractions might find expression in concrete reality. On the left of each row of cells is a specific research question, and on the right is a brief account of a strategy that might be used to answer it. The first two examples illustrate a very frequently employed form of experimental design that involves randomly sorting the members of a subject pool into two separate groups, treating the two groups differently with respect to a certain independent variable, and then measuring both groups on a certain dependent variable with the aim of determining whether the differential treatment produces differential effects. (Variables: Independent and Dependent.) A quasi-experimental variation on this theme, illustrated by the third example, involves randomly selecting two groups of subjects that already differ with respect to one variable, and then measuring both

groups on another variable to determine whether the different levels of the first are associated with different levels of the second.

Question	Strategy
Does the presence of a certain kind of mycorrhizal fungus enhance the growth of a certain kind of plant?	Begin with a "subject pool" of seeds of the type of plant in question. Randomly sort them into two groups, A and B. Plant and grow them under conditions that are identical in every respect except one: namely, that the seeds of group A (the experimental group) are grown in a soil that contains the fungus, while those of group B (the control group) are grown in a soil that does not contain the fungus. After some specified period of time, harvest the plants of both groups and take the relevant measure of their respective degrees of growth. If the presence of the fungus does enhance growth, the average measure should prove greater for group A than for group B.
Do two types of music, type-I and type-II, have different effects upon the ability of college students to perform a series of mental tasks requiring concentration?	Begin with a subject pool of college students, relatively homogeneous with respect to age, record of academic achievement, and other variables potentially relevant to the performance of such a task. Randomly sort the subjects into two groups, A and B. Have the members of each group perform the series of mental tasks under conditions that are identical in every respect except one: namely, that group A has music of type-I playing in the background, while group B has music of type-II. (Note that the distinction between experimental and control group does not apply in this example.) Conclude by measuring how well the subjects perform on the series of tasks under their respective conditions. Any difference between the effects of the two types of music should show up as a difference between the mean levels of performance for group A and group B.
Do two strains of mice, A and B, differ with respect to their ability to learn to avoid an aversive stimulus?	With this type of situation you are in effect starting out with two subject pools, one for strain A and one for strain B. Draw a random sample of size N_a from pool A and another of size N_b from pool B. Run the members of each group through a standard aversive-conditioning procedure, measuring for each one how well and quickly the avoidance behavior is acquired. Any difference between the avoidance-learning abilities of the

	two strains should manifest itself as a difference between their respective group means.
--	--

In each of these cases, the two samples are **independent** of each other in the obvious sense that they are separate samples containing different sets of individual subjects. The individual measures in group A are in no way linked with or related to any of the individual measures in group B, and vice versa. The version of a t-test examined in this chapter will assess the significance of the difference between the means of two such samples, providing: (i) that the two samples are randomly drawn from normally distributed populations; and (ii) that the measures of which the two samples are composed are equal-interval.

To illustrate the procedures for this version of a t-test, imagine we were actually to conduct the experiment described in the second of the above examples. We begin with a fairly homogeneous subject pool of 30 college students, randomly sorting them into two groups, A and B, of sizes $N_a=15$ and $N_b=15$. (It is not essential for this procedure that the two samples be of the same size.) We then have the members of each group, one at a time, perform a series of 40 mental tasks while one or the other of the music types is playing in the background. For the members of group A it is music of type-I, while for those of group B it is music of type-II. The following table shows how many of the 40 components of the series each subject was able to complete. Also shown are the means and sums of squared deviates for the two groups.

Group A music of type-I	Group B music of type-II
26 21 22	18 23 21
26 19 22	20 20 29
26 25 24	20 16 20
21 23 23	26 21 25
18 29 22	17 18 19
$N_a=15$ $M_a=23.13$ $SS_a=119.73$	$N_b=15$ $M_b=20.87$ $SS_b=175.73$
$M_a - M_b = 2.26$	

Null

Hypothesis

Recall from Chapter 7 that whenever you perform a statistical test, what you are testing, fundamentally, is the null hypothesis. In

general, the null hypothesis is the logical antithesis of whatever hypothesis it is that the investigator is seeking to examine. For the present example, the research hypothesis is that the two types of music have different effects, so the null hypothesis is that they do not have different effects. Its immediate implication is that any difference we find between the means of the two samples should significantly differ from zero.

If the investigator specifies the direction of the difference in advance as either

task performance will be better with type-I music than with type-II	which would be supported by finding the mean of sample A to be significantly greater than the mean of sample B ($M_a > M_b$)
---	--

or

task performance will be better with type-II music than with type-I	which would be supported by finding the mean of sample B to be significantly greater than the mean of sample A ($M_b > M_a$)
---	--

then the research hypothesis is directional and permits a one-tail test of significance. A non-directional research hypothesis would require a two-tail test, as it is the equivalent of saying "I'm expecting a difference in one direction or the other, but I can't guess which." For the sake of discussion, let us suppose we had started out with the directional hypothesis that task performance will be better with type-I music than with type-II music. Clearly our observed result, $M_a - M_b = 2.26$, is in the hypothesized direction. All that remains is to determine how confident we can be that it comes from anything more than mere chance coincidence.

¶ Logic

and

Procedure

- (1) The mean of a sample randomly drawn from a normally distributed source population belongs to a sampling distribution of sample means that is also normal in form. The overall mean of this sampling distribution will be identical with the mean of the source population:

$$\mu_M = \mu_{\text{source}}$$

- (2) For two samples, each randomly drawn from a normally distributed source population, the difference between the means of the two samples,

$$M_a - M_b$$

belongs to a sampling distribution that is normal in form, with an overall mean equal to the difference between the means of the two source populations

$$\mu_{M-M} = \mu_{\text{source A}} - \mu_{\text{source B}}$$

- (2) On the null hypothesis, $\mu_{\text{source A}}$ and $\mu_{\text{source B}}$ are identical, hence

$$\mu_{M-M} = 0$$

- (3)

For the present example, the null hypothesis holds that the two types of music do not have differential effects on task performance. This is tantamount to saying that the measures of task performance in groups A and B are all drawn indifferently from the same source population of such measures. In items 3 and 4 below, the phrase "source population" is a shorthand way of saying "the population of measures that the null hypothesis assumes to have been the common source of the measures in both groups."
-

- (3) If we knew the variance of the source population, we would then be able to calculate the standard deviation (aka "standard error") of the sampling distribution of sample-mean differences as

$$\sigma_{M-M} = \sqrt{\frac{\sigma_{\text{source}}^2}{N_a} + \frac{\sigma_{\text{source}}^2}{N_b}}$$

- (3) This, in turn, would allow us to test the null hypothesis for any particular $M_a - M_b$ difference by calculating the appropriate z-ratio

$$z = \frac{M_{Xa} - M_{Xb}}{\sigma_{M-M}}$$

(3) and referring the result to the unit normal distribution.

In most practical research situations, however, the variance of the source population, hence also the value of σ_{M-M} , can be arrived at only through estimation. In these cases the test of the null hypothesis is performed not with **z** but with **t**:

$$t = \frac{M_{Xa} - M_{Xb}}{\text{est. } \sigma_{M-M}}$$

(3) The resulting value belongs to the particular sampling distribution of **t** that is defined by **df**=(**N_a**—1)+(**N_b**—1).

(4) To help you keep track of where the particular numerical values are coming from beyond this point, here again are the summary statistics for our hypothetical experiment on the effects of two types of music:

Group A music of type-I	Group B music of type-II
N_a =15 M_a =23.13 SS_a =119.73	N_b =15 M_b =20.86 SS_b =175.73
M_a—M_b =2.26	

(3) As indicated in Chapter 9, the variance of the source population can be estimated as

$$\{s_p^2\} = \frac{SS_a + SS_b}{(N_a - 1) + (N_b - 1)}$$

(3) which for the present example comes out as

$$\{s_p^2\} = \frac{119.73 + 175.73}{14 + 14} = 10.55$$

(3) This, in turn, allows us to estimate the standard deviation of the sampling distribution of sample-mean differences as

$$\begin{aligned}
 \text{est. } i \sigma_{M-M} &= \sqrt{\frac{\{s_p^2\}}{N_a} + \frac{\{s_p^2\}}{N_b}} \\
 &= \sqrt{\frac{10.55}{15} + \frac{10.55}{15}} = \pm 1.19
 \end{aligned}$$

- (4) And with this estimated value of $i \sigma_{M-M}$ in hand, we are then able to calculate the appropriate t -ratio as

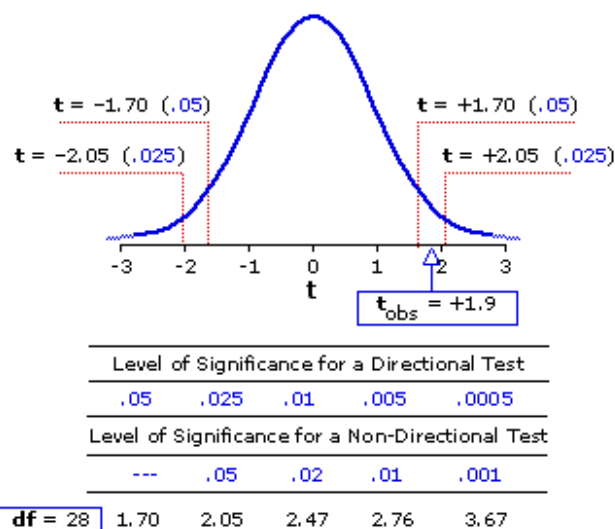
$$\begin{aligned}
 t &= \frac{M_{Xa} - M_{Xb}}{\text{est. } i \sigma_{M-M}} \\
 &= \frac{23.13 - 20.87}{1.19} = +1.9
 \end{aligned}$$

- (4) with $df = (15 - 1) + (15 - 1) = 28$

In the calculation of a two-sample t -ratio, note that the sign of t depends on the direction of the difference between M_{Xa} and M_{Xb} . $M_{Xa} > M_{Xb}$ will produce a positive value of t , while $M_{Xa} < M_{Xb}$ will produce a negative value of t .

¶ Inference

Figure 11.1 shows the sampling distribution of t for $df=28$. Also shown is the portion of the table of critical values of t (Appendix C) that pertains to $df=28$. The designation " t_{obs} " refers to our observed value of $t=+1.9$. We started out with the directional research hypothesis that task performance would be better for group A than for group B, and as our observed result, $M_{Xa} - M_{Xb} = 2.26$, proved consistent with that hypothesis, the relevant critical values of t are those that pertain to a directional (one-tail) test of significance: 1.70 for the .05 level of significance, 2.05 for the .025 level, 2.47 for the .01 level, and so on.

Figure 11.1. Sampling Distribution of t for $df=28$ 

If our observed value of t had ended up smaller than 1.70, the result of the experiment would be non-significant vis-à-vis the conventional criterion that the mere-chance probability of a result must be equal to or less than .05. If it had come out at precisely 1.70, we would conclude that the result is significant **at** the .05 level. As it happens, the observed t meets and somewhat exceeds the 1.70 critical value, so we conclude that our result is significant somewhat **beyond** the .05 level. If the observed t had been equal to or greater than 2.05, we would have been able to regard the result as significant at or beyond the .025 level; and so on.

The same logic would have applied to the left tail of the distribution if our initial research hypothesis had been in the

opposite direction, stipulating that task performance would be better with music of type-II than with music of type-I. In this case we would have expected M_{Xa} to be smaller than M_{Xb} , which would have entailed a negative sign for the resulting value of t .

If, on the other hand, we had begun with no directional hypothesis at all, we would in effect have been expecting

either $M_{Xa} > M_{Xb}$ or $M_{Xa} < M_{Xb}$

and that disjunctive expectation ("either the one or the other") would have required a non-directional, two-tailed test. Note that for a non-directional test our observed value of $t = +1.9$ (actually, for a two-tailed test it would have to be regarded as $t = \pm 1.9$) would **not** be significant at the minimal .05 level. (The distinction between directional and non-directional tests of significance is introduced in Chapter 7.)

In this particular case, however, we did begin with a directional hypothesis, and the obtained result as assessed by a directional test is significant beyond the .05 level. The practical, bottom-line meaning of this conclusion is that the likelihood of our experimental result having come about through mere random variability—mere chance coincidence, "sampling error," the luck of the scientific draw—is a somewhat less than 5%; hence, we can have about 95% confidence that the observed result reflects something **more** than mere random variability. For the present example, this "something more" would presumably be a genuine difference between the effects of the two types of music on the performance of this particular type of task.

Step-by-Step Computational Procedure: t-Test for the Significance of the Difference between the Means of Two independent Samples

Note that this test makes the following assumptions and can be meaningfully applied only insofar as these assumptions are met:
 That the two samples are independently and randomly drawn from the source population(s).
 That the scale of measurement for both samples has the properties of an equal interval scale.
 That the source population(s) can be reasonably supposed to have a normal distribution.

Step 1. For the two samples, A and B, of sizes of N_a and N_b respectively, calculate

M_{Xa} and SS_a the mean and sum of squared deviates of sample A

M_{xb} and SS_a	the mean and sum of squared deviates of sample B
---------------------	--

Step 2. Estimate the variance of the source population as

$$\{s_p^2\} = \frac{SS_a + SS_b}{(N_a - 1) + (N_b - 1)}$$

Recall that "source population" in this context means "the population of measures that the null hypothesis assumes to have been the common source of the measures in both groups."

Step 3. Estimate the standard deviation of the sampling distribution of sample-mean differences (the "standard error" of $M_{xa} - M_{xb}$) as

$$\text{est. } \sigma_{M-M} = \text{sqrt} \left[\frac{\{s_p^2\}}{N_a} + \frac{\{s_p^2\}}{N_b} \right]$$

Step 4. Calculate t as

$$t = \frac{M_{xa} - M_{xb}}{\text{est. } \sigma_{M-M}}$$

Step 5. Refer the calculated value of t to the table of critical values of t (Appendix C), with $df = (N_a - 1) + (N_b - 1)$. Keep in mind that a one-tailed directional test can be applied only if a specific directional hypothesis has been stipulated in advance; otherwise it must be a non-directional two-tailed test.

Note that this chapter includes a subchapter on the Mann-Whitney Test, which is a non-parametric alternative to the independent-samples t -test.

Have you understood?

1. The average travel time taken based on a random sample of 10 people working in a company to reach the office is 40 minutes with a standard deviation of 10 minutes. Establish the 95% confidence interval for the mean travel time of everyone in the company and redesign the working hours.
2. A marketing manager of a fast-food restaurant in a city wishes to estimate the average yearly amount that families spend on fast food restaurants. He wants the estimate to be

within \pm Rs100 with a confidence level of 99%. It is known from an earlier pilot study that the standard deviation of the family expenditure on fast food restaurant is Rs500. How many families must be chosen for this problem?

3. a company manufacturing sports goods wants to estimate the proportion of cricket players among high school students in India. The company wants the estimate to be within ± 0.03 with a confidence level of 99%. A pilot study done earlier reveals that out of 80 high school students, 36 students play cricket. What should be the sample size for this study?



5

INTRODUCTION TO STEADY-STATE QUEUEING THEORY

This handout introduces basic concepts in steady-state queueing theory for two very basic systems later labeled M/M/1 and M/M/k ($k > 2$). Both assume Poisson arrivals and exponential service. While the more general Little equations are shown, more in-depth texts such as chapter 6 of the Banks, Carson, Nelson, Nicole 4th edition *Discrete-Event System Simulation* (Prentice-Hall, 2005, ISBN 0-13-144679-7). There are other more comprehensive books that I can recommend if you desire further study in this area. You must keep in mind that the formulas presented here are strictly for the steady-state or long-term performance of queueing systems. Additionally only the simplest statistical distribution assumptions (Poisson arrivals, exponential service) are covered in this brief handout. Nonetheless, if a closed form queueing solution exists to a particular situation, then use it instead of putting together a simulation. You are engineers and will be paid to use your brain to find cost effective timely solutions. Try doing so in this class as well.

The vast majority of this handout comes from a wonderful book *Quantitative Methods for Business*, 9th Ed., by Anderson, Sweeney, & Williams, ISBN#-324-18413-1 (newer editions are now out). This is a very understandable non-calculus operations research book that I highly recommend to everyone – get a copy and you will finally understand things that have been confusing to you during

your tough ISE classes. As you bust your posterior (assuming that you desire to obtain that hard to get passing mark in our class), you should check your skills on solving homework problems in one of at least the two following ways:

- Try the homework problems stored on our web page. The problems are in a handy Word file. The solutions are also in another Word file on the web page.
- Use the Excel template on the web page to check your solutions especially on finding minimal total cost solutions prior to our killer exams.

Hint: Do many problems of all types before our 2 hour exams. This pertains to queueing, and anything covered by the syllabus. You control your destiny in this course. No whining is anticipated after the exams – suck it up. Imagine being a civil engineer that does not do a full job on a bridge design and people die – partial solutions do not make my day. So learn this stuff and learn it well. Then consider taking ISE 704 later as you will have a major head start on the grad students just learning simulation for the 1st time in that class. Perhaps if you are really good, you can work on your MBA at Otterbein later in your career.

Now into the wide world of queueing...

Recall the last time that you had to wait at a supermarket checkout counter, for a teller at your local bank, or to be served at a fast-food restaurant. In these and many other waiting line situations, the time spent waiting is undesirable. Adding more checkout clerks, bank tellers, or servers is not always the most economical strategy for improving service, so businesses need to determine ways to keep waiting times within tolerable limits.

Models have been developed to help managers understand and make better decisions concerning the operation of waiting lines. In quantitative methods terminology, a waiting line is also known as a **queue**, and the body of knowledge dealing with waiting lines is **known as queueing theory**. In the early 1900s, A. K. Erlang, a Danish telephone engineer, began a study of the congestion and waiting times occurring in the completion of telephone calls. Since then, queueing theory has grown far more sophisticated with applications in a wide variety of waiting line situations.

Waiting line models consist of mathematical formulas and relationships that can be used to determine the **operating characteristics** (performance measures) for a waiting line. Operating characteristics of interest include the following:

1. The probability that no units are in the system

2. The average number of units in the waiting line
3. The average number of units in the system (the number of units in the waiting line plus the number of units being served)
4. The average time a unit spends in the waiting line
5. The average time a unit spends in the system (the waiting time plus the service time)
6. The probability that an arriving unit has to wait for service

Managers who have such information are better able to make decisions that balance desirable service levels against the cost of providing the service.

STRUCTURE OF A WAITING LINE SYSTEM

To illustrate the basic features of a waiting line model, we consider the waiting line at the Burger Dome fast-food restaurant. Burger Dome sells hamburgers, cheeseburgers, french fries, soft drinks, and milk shakes, as well as a limited number of specialty items and dessert selections. Although Burger Dome would like to serve each customer immediately, at times more customers arrive than can be handled by the Burger Dome food service staff. Thus, customers wait in line to place and receive their orders.

Burger Dome is concerned that the methods currently used to serve customers are resulting in excessive waiting times. Management wants to conduct a waiting line study to help determine the best approach to reduce waiting times and improve service.

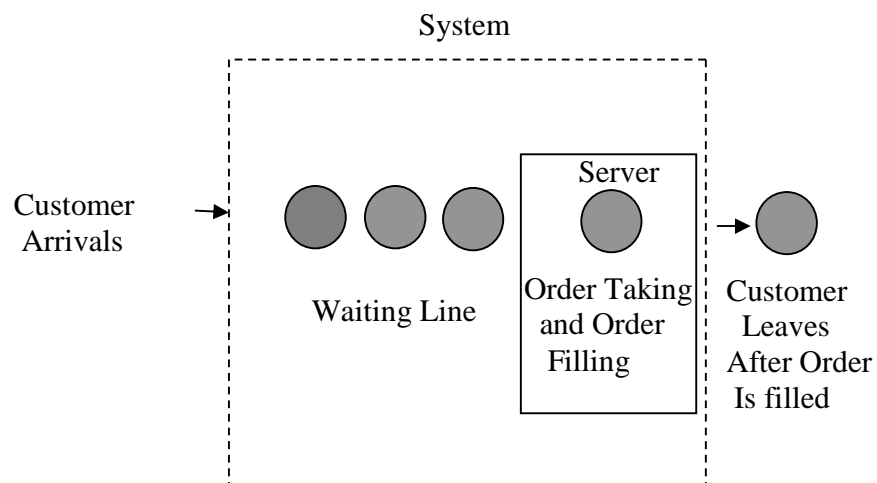
Single-Channel Waiting Line

In the current Burger Dome operation, a server takes a customer's order, determines the total cost of the order, takes the money from the customer, and then fills the order. Once the first customer's order is filled, the server takes the order of the next customer waiting for service. This operation is an example of a **single-channel waiting line**. Each customer entering the Burger Dome restaurant must pass through the *one* channel—one order-taking and order-filling station—to place an order, pay the bill, and receive the food. When more customers arrive than can be served immediately, they form a waiting line and wait for the order-taking and order-filling station to become available. A diagram of the Burger Dome single-channel waiting line is shown in Figure .1.

Distribution of Arrivals

Defining the arrival process for a waiting line involves determining the probability distribution for the number of arrivals in a given period of time. For many waiting line situations, the arrivals occur *randomly and independently* of other arrivals, and we cannot predict when an arrival will occur. In such cases, quantitative analysts have found that the **Poisson probability distribution** provides a good description of the arrival pattern.

FIGURE 14.1 THE BURGER DOME SINGLE-CHANNEL WAITING LINE



The Poisson probability function provides the probability of x arrivals in a specific time period. The probability function is as follows.

$$(14.1)$$

where
 e

x = the number of arrivals in the time period

λ = the mean number of arrivals per time period

Suppose that Burger Dome analyzed data on customer arrivals and concluded that the mean arrival rate is 45 customers per hour. For a one-minute period, the mean arrival rate would be $\lambda = 45$ customers/60 minutes = 0.75 customers per minute. Thus, we can use the following Poisson probability function to compute the probability of x customer arrivals during a one-minute period:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{0.75^x e^{-0.75}}{x!} \quad (14.2)$$

Thus, the probabilities of 0, 1, and 2 customer arrivals during a one-minute period are

$$\begin{aligned} P(0) &= \frac{(0.75)^0 e^{-0.75}}{0!} = e^{-0.75} = 0.4724 \\ P(1) &= \frac{(0.75)^1 e^{-0.75}}{1!} = 0.75e^{-0.75} = 0.75(0.4724) = 0.3543 \\ P(2) &= \frac{(0.75)^2 e^{-0.75}}{2!} = \frac{(0.75)^2 e^{-0.75}}{2!} = \frac{(0.5625)(0.4724)}{2} = 0.1329 \end{aligned}$$

The probability of no customers in a one-minute period is 0.4724, the probability of one customer in a one-minute period is 0.3543, and the probability of two customers in a one-minute period is 0.1329. Table 14.1 shows the Poisson probabilities for customer arrivals during a one-minute period.

The waiting line models that will be presented in Sections 14.2 and 14.3 use the Poisson probability distribution to describe the customer arrivals at Burger Dome. In practice; you should record the actual number of arrivals per time period for several days or weeks: and compare the frequency distribution of the observed number of arrivals to the Poisson probability distribution to determine whether the Poisson probability distribution provides a reasonable approximation of the arrival distribution.

TABLE 1 POISSON PROBABILITIES FOR THE NUMBER OF CUSTOMER ARRIVALS AT A BURGER DOME RESTAURANT DURING A ONE-MINUTE PERIOD ($\lambda = 0.75$)

Number of Arrivals	Probability
0	0.4724
1	0.3543
2	0.1329

3	0.0332
4	0.0062
5 or more	0.0010

Distribution of Service Times

The service time is the time a customer spends at the service facility once the service has started. At Burger Dome, the service time starts when a customer begins to place the order with the food server and continues until the customer receives the order. Service times are rarely constant. At Burger Dome, the number of items ordered and the mix of items ordered vary considerably from one customer to the next. Small orders can be handled in a matter of seconds, but large orders may require more than two minutes.

Quantitative analysts have found that if the probability distribution for the service time can be assumed to follow an **exponential probability distribution**, formulas are available for providing useful information about the operation of the waiting line. Using an exponential probability distribution, the probability that the service time will be less than or equal to a time of length t is fixed

$$P(\text{service time} \leq t) = 1 - e^{-\mu t} \quad (14.3)$$

where

μ = the mean number of units that can be served per time period

A property of the exponential probability distribution is that there is a 0.6321 probability that the random variable takes on a value less than its mean.

Suppose that Burger Dome studied the order-taking and order-filling process and found that the single food server can process an average of 60 customer orders per hour. On a one minute basis, the mean service rate would be $\mu = 60$ customers/60 minutes = 1 customer per minute. For example, with $\mu = 1$, we can use equation (14.3) to compute probabilities such as the probability an order can be processed in 1/2 minute or less, 1 minute or less, and waiting line 2 minutes or less. These computations are

$$P(\text{service time} \leq 0.5 \text{ min.}) = 1 - e^{-1(0.5)} = 1 - 0.6065 = 0.3935$$

$$P(\text{service time} \leq 1.0 \text{ min.}) = 1 - e^{-1(1.0)} = 1 - 0.3679 = 0.6321$$

$$P(\text{service time} \leq 2.0 \text{ min.}) = 1 - e^{-1(2.0)} = 1 - 0.1353 = 0.8647$$

Thus, we would conclude that there is a 0.3935 probability that an order can be processed in 1/2 minute or less, a 0.6321 probability that it can be processed in 1 minute or less, and a 0.8647 probability that it can be processed in 2 minutes or less.

In several waiting line models presented in this chapter, we assume that the probability distribution for the service time follows

an exponential probability distribution. In practice, you should collect data on actual service times to determine whether the exponential probability distribution is a reasonable approximation of the service times for your application.

Queue Discipline

In describing a waiting line system, we must define the manner in which the waiting units are arranged for service. For the Burger Dome waiting line, and in general for most customer oriented waiting lines, the units waiting for service are arranged on a **first-come, first. served** basis; this approach is referred to as an **FCFS** queue discipline. However, some situations call for different queue disciplines. For example, when people wait for an elevator, the last one on the elevator is often the first one to complete service (i.e., the first to leave the elevator). Other types of queue disciplines assign priorities to the waiting units and then serve the unit with the highest priority first. In this chapter we consider only waiting lines based on a first-come, first-served queue discipline.

Steady-State Operation

When the Burger Dome restaurant opens in the morning, no customers are in the restaurant. Gradually, activity builds up to a normal or steady state. The beginning or start-up period is referred to as the **transient period**. The transient period ends when the system reaches the normal or **steady-state operation**. The models described in this handout cover the steady-state operating characteristics of a waiting line.

STEADY-STATE SINGLE-CHANNEL WAITING LINE MODEL WITH POISSON ARRIVALS AND EXPONENTIAL SERVICE TIMES

In this section we present formulas that can be used to determine the *steady-state* operating characteristics for a single-channel waiting line. The formulas are applicable if the arrivals follow a Poisson probability distribution and the service times follow an exponential probability distribution. As these assumptions apply to the Burger Dome waiting line problem introduced in Section 14.1, we show how formulas can be used to determine Burger Dome's operating characteristics and thus provide management with helpful decision-making information.

The mathematical methodology used to derive the formulas for the operating characteristics of waiting lines is rather complex. However, our purpose in this chapter is not to provide the theoretical development of waiting line models, but rather to show how the formulas that have been developed can provide information about operating characteristics of the waiting line.

Operating Characteristics

The following formulas can be used to compute the steady-state operating characteristics for a single-channel waiting line with Poisson arrivals and exponential service times, where

λ = the mean number of arrivals per time period (the mean arrival rate)

μ = the mean number of services per time period (the mean service rate)

Equations (14.4) through (14.10) do not provide formulas for optimal conditions. Rather, these equations provide information about the steady-state operating characteristics of a waiting line.

1. The probability that no units are in the system:

$$P_0 = 1 - \frac{\lambda}{\mu} \quad (14.4)$$

2. The average number of units in the waiting line:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (14.5)$$

3. The average number of units in the system:

$$L = L_q + \frac{\lambda}{\mu} \quad (14.6)$$

4. The average time a unit spends in the waiting line:

$$W_q = \frac{L_q}{\lambda} \quad (14.7)$$

5. The average time a unit spends in the system:

$$W = W_q + \frac{1}{\mu} \quad (14.8)$$

6. The probability that an arriving units has to wait for service:

7. The probability of n units in the system:

$$P_n = \left(\frac{\lambda}{\mu} \right)^n P_0 \quad (14.10)$$

$$P_w = \frac{\lambda}{\mu} \quad (14.9)$$

The values of the **mean arrival rate** λ and the **mean service rate** μ are clearly important components in determining the operating characteristics. Equation (14.9) shows that the ratio of the mean arrival rate to the mean service rate, λ/μ , provides the probability that an arriving unit has to wait because the service facility is in use. Hence, λ/μ often is referred to as the *utilization factor* for the service facility.

The operating characteristics presented in equations (14.4) through (14.10) are applicable only when the mean service rate μ is *greater than* the mean arrival rate λ --in other words, when $\lambda/\mu < 1$. If this condition does not exist, the waiting line will continue to grow without limit because the service facility does not have sufficient capacity to handle the arriving units. Thus, in using equations (14.4) through (14.10), we must have $\mu > \lambda$.

Operating Characteristics for the Burger Dome Problem

Recall that for the Burger Dome problem we had a mean arrival rate of $\lambda = 0.75$ customers per minute and a mean service rate of $\mu = 1$ customer per minute. Thus, with $\mu > \lambda$, equations (14.4) through (14.10) can be used to provide operating characteristics for the Burger Dome single-channel waiting line:

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{0.75}{1} = 0.25$$

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{0.75^2}{1(1 - 0.75)} = 2.25 \text{ customers}$$

$$L = L_q + \frac{\lambda}{\mu} = 2.25 + \frac{0.75}{1} = 3 \text{ customers}$$

$$W_q = \frac{L_q}{\lambda} = \frac{2.25}{0.75} = 3 \text{ minutes}$$

$$W = W_q + \frac{1}{\mu} = 3 + \frac{1}{1} = 4 \text{ minutes}$$

$$P_w = \frac{\lambda}{\mu} = \frac{0.75}{1} = 0.75$$

Equation (14.10) can be used to determine the probability of any number of customers in the system. Applying it provides the probability information in Table 14.2.

Use of Waiting Line Models

The results of the single-channel waiting line for Burger Dome show several important things about the operation of the waiting line. In particular, customers wait an average of three minutes before

beginning to place an order, which appears somewhat long for a business based on fast service.

TABLE .2 STEADY-STATE PROBABILITY OF n CUSTOMERS IN THE SYSTEM FOR THE BURGER DOME WAITING LINE PROBLEM

Number of Customers	Probability
0	0.2500
1	0.1875
2	0.1406
3	0.1055
4	0.0791
5	0.0593
6	0.0445
7 or more	0.1335

In addition, the facts that the average number of customers waiting in line is 2.25 and that 75% of the arriving customers have to wait for service are indicators that something should be done to improve the waiting line operation. Table 14.2 shows a 0.1335 probability that seven or more customers are in the Burger Dome system at one time. This condition indicates a fairly high probability that Burger Dome will experience some long waiting lines if it continues to use the single-channel operation.

If the operating characteristics are unsatisfactory in terms of meeting company standards for service, Burger Dome's management should consider alternative designs or plans for improving the waiting line operation.

Improving the Waiting Line Operation

Waiting line models often indicate where improvements in operating characteristics are desirable. However, the decision of how to modify the waiting line configuration to improve the operating characteristics must be based on the insights and creativity of the analyst.

After reviewing the operating characteristics provided by the waiting line model, Burger Dome's management concluded that improvements designed to reduce waiting times are desirable. To make improvements in the waiting line operation, analysts often focus on ways to improve the service rate. Generally, service rate improvements are obtained by making either or both the following changes:

1. Increase the mean service rate μ by making a creative design change or by using new technology.
2. Add service channels so that more customers can be served simultaneously.

Assume that in considering alternative 1, Burger Dome's management decides to employ an order filler who will assist the order taker at the cash register. The customer begins the service process by placing the order with the order taker. As the order is placed, the order taker announces the order over an intercom system, and the order filler begins filling the order. When the order is completed, the order taker handles the money, while the order filler continues to fill the order. With this design, Burger Dome's management estimates the mean service rate can be increased from the current service rate of 60 customers per hour to 75 customers per hour. Thus, the mean service rate for the revised system is, $\mu = 75 \text{ customers}/60 \text{ minutes} = 1.25 \text{ customers per minute}$. For $\lambda = 0.75 \text{ customers per minute}$ and $\mu = 1.25 \text{ customers per minute}$, equations (14.4) through (14.10) can be used to provide the new operating characteristics for the Burger Dome waiting line. These operating characteristics are summarized in Table 14.3.

TABLE .3 OPERATING CHARACTERISTICS FOR THE BURGER DOME SYSTEM WITH THE MEAN SERVICE RATE INCREASED TO $\mu = 1.25 \text{ CUSTOMERS PER MINUTE}$

Probability of no customers in the system	0.400
Average number of customers in the waiting line	0.900
Average number of customers in the system	1.500
Average time in the waiting line	1.200 minutes
Average time in the system	2.000 minutes
Probability that an arriving customer has to wait	0.600
Probability that seven or more customers are in the system	0.028

The information in Table 14.3 indicates that all operating characteristics have improved because of the increased service rate. In particular, the average time a customer spends in the waiting line has been reduced from 3 to 1.2 minutes and the average time a customer spends in the system has been reduced from 4 to 2 minutes. Are any other alternatives available that Burger

Dome can use to increase the service rate? If so, and if the mean service rate μ can be identified for each alternative, equations (14.4) through (14.10) can be used to determine the revised operating characteristics and any improvements in the waiting line system. The added cost of any proposed change can be compared to the corresponding service improvements to help the manager determine whether the proposed service improvements are worthwhile.

As mentioned previously, another option often available is to provide one or more additional service channels so that more than one customer may be served at the same time. The extension of the single-channel waiting line model to the multiple-channel waiting line model is the topic of the next section.

Notes:

1. The assumption that arrivals follow a Poisson probability distribution is equivalent to the assumption that the time between arrivals has an exponential probability distribution. For example, if the arrivals for a waiting line follow a Poisson probability distribution with a mean of 20 arrivals per hour, the time between arrivals will follow an exponential probability distribution, with a mean time between arrivals of $1/20$ or 0.05 hour.
2. Many individuals believe that whenever the mean service rate μ is greater than the mean arrival rate λ , the system should be able to handle or serve all arrivals. However, as the Burger Dome example shows, the variability of arrival times and service times may result in long waiting times even when the mean service rate exceeds the mean arrival rate. A contribution of waiting line models is that they can point out undesirable waiting line operating characteristics even when they $\mu > \lambda$ condition appears satisfactory.

STEADY-STATE MULTIPLE-CHANNEL WAITING LINE MODEL WITH POISSON ARRIVALS AND EXPONENTIAL SERVICE TIMES

You may be familiar with multiple-channel systems that also have multiple waiting lines. The waiting line model in this section has multiple channels, but only a single waiting line. Operating characteristics for a multiple-channel system are better

A **multiple-channel waiting line** consists of two or more service channels that are assumed to be identical in terms of service capability. In the multiple-channel system, arriving units wait in a single waiting line and then move to the first available channel to be

served. The single-channel Burger Dome operation can be expanded to a two-channel system by opening a second service channel. Figure 14.3 shows a diagram of the Burger Dome two-channel waiting line.

In this section we present formulas that can be used to determine the steady-state operating characteristics for a multiple-channel waiting line. These formulas are applicable if the following conditions exist.

1. The arrivals follow a Poisson probability distribution.
2. The service time for each channel follows an exponential probability distribution.
3. The mean service rate μ is the same for each channel.
4. The arrivals wait in a single waiting line and then move to the first open channel for service.

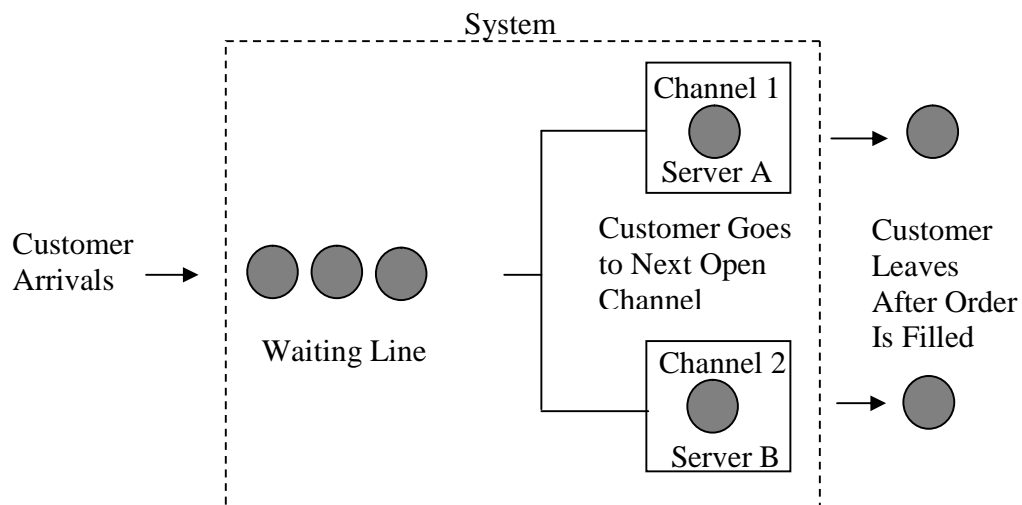


Figure 14.3: Two-channel ($k=2$) Queue (assuming M/M/2 conditions)

M/M/k Steady-State Operating Characteristics

The following formulas can be used to compute the steady-state operating characteristics for multiple-channel waiting lines, where

- λ = the mean arrival rate for the system
- μ = the mean service rate for *each* channel
- k = the number of channels

1. The probability that no units are in the system:

$$P_0 = \frac{1}{\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{k!} \left(\frac{k\mu}{k\mu - \lambda} \right)} \quad (14.11)$$

2. The average number of units in the waiting line:

$$L_q = \frac{(\lambda/\mu)^k \lambda \mu}{(k-1)!(k\mu - \lambda)^2} P_0 \quad (14.12)$$

3. The average number of units in the system:

$$L = L_q + \frac{\lambda}{\mu} \quad (14.13)$$

4. The average time a unit spends in the waiting line:

$$W_q = \frac{L_q}{\lambda} \quad (14.14)$$

5. The average time a unit spends in the system:

$$W = W_q + \frac{1}{\mu} \quad (14.15)$$

6. The probability that an arriving unit has to wait for service:

$$P_w = \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k \left(\frac{k\mu}{k\mu - \lambda} \right) P_0 \quad (14.16)$$

7. The probability of n units in the system:

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \quad \text{for } n \leq k \quad (14.17)$$

$$P_n = \frac{(\lambda/\mu)^n}{k! k^{(n-k)}} P_0 \quad \text{for } n > k \quad (14.18)$$

Because μ is the mean service rate for each channel, $k\mu$ is the mean service rate for the multiple-channel system. As was true for the single-channel waiting line model, the formulas for the operating characteristics of multiple-channel waiting lines can be applied only in situations where the mean service rate for the system is greater than the mean arrival rate for the system; in other words, the formulas are applicable only if $k\mu$ is greater than λ .

Some expressions for the operating characteristics of multiple-channel waiting lines are more complex than their single-channel counterparts. However, equations (14.11) through (14.18) provide the same information as provided by the single-channel model. To help simplify the use of the multiple-channel equations, Table 14.4 contains values of P_0 for selected values of λ/μ and k . The values provided in the table correspond to cases where $k\mu > \lambda$, and hence the service rate is sufficient to process all arrivals.

Operating Characteristics for the Burger Dome Problem

To illustrate the multiple-channel waiting line model, we return to the Burger Dome fastfood restaurant waiting line problem. Suppose that management wants to evaluate the desirability of opening a second order-processing station so that two customers can be served simultaneously. Assume a single waiting line with the first customer in line moving to the first available server. Let us evaluate the operating characteristics for this two channel system.

We use equations (14.12) through (14.18) for the $k = 2$ channel system. For a mean arrival rate of $\lambda = 0.75$ customers per minute and mean service rate of $\mu = 1$ customer per minute for each channel, we obtain the operating characteristics:

$$P_0 = 0.4545 \text{ (from Table 14.4 with } \lambda/\mu = 0.75\text{)}$$

$$L_q = \frac{(0.75/1)^2(0.75)(1)}{(2-1)![2(1)-0.75]^2}(0.4545) = 0.1227 \text{ customer}$$

$$L = L_q + \frac{\lambda}{\mu} = 0.1227 + \frac{0.75}{1} = 0.8727 \text{ customer}$$

$$W_q = \frac{L_q}{\lambda} = \frac{0.1227}{0.75} = 0.1636 \text{ minute}$$

$$W = W_q + \frac{1}{\mu} = 0.1636 + \frac{1}{1} = 1.1636 \text{ minutes}$$

$$P_w = \frac{1}{2!} \left(\frac{0.75}{1} \right)^2 \left[\frac{2(1)}{2(1)-0.75} \right] (0.4545) = 0.2045$$

Using equations (14.17) and (14.18), we can compute the probabilities of n customers in the system. The results from these computations are summarized in Table 14.5.

TABLE .4 VALUES OF P_0 FOR MULTIPLE-CHANNEL WAITING LINES WITH POISSON ARRIVALS AND EXPONENTIAL SERVICE TIMES

Ratio λ / μ	Number of Channels (k)			
	2	3	4	5
0.15	0.8605	0.8607	0.8607	0.8607
0.20	0.8182	0.8187	0.8187	0.8187
0.25	0.7778	0.7788	0.7788	0.7788
0.30	0.7391	0.7407	0.7408	0.7408
0.35	0.7021	0.7046	0.7047	0.7047
0.40	0.6667	0.6701	0.6703	0.6703
0.45	0.6327	0.6373	0.6376	0.6376
0.50	0.6000	0.6061	0.6065	0.6065
0.55	0.5686	0.5763	0.5769	0.5769
0.60	0.5385	0.5479	0.5487	0.5488
0.65	0.5094	0.5209	0.5219	0.5220
0.70	0.4815	0.4952	0.4965	0.4966
0.75	0.4545	0.4706	0.4722	0.4724
0.80	0.4286	0.4472	0.4491	0.4493
0.85	0.4035	0.4248	0.4271	0.4274
0.90	0.3793	0.4035	0.4062	0.4065
0.95	0.3559	0.3831	0.3863	0.3867
1.00	0.3333	0.3636	0.3673	0.3678
1.20	0.2500	0.2941	0.3002	0.3011
1.40	0.1765	0.2360	0.2449	0.2463
1.60	0.1111	0.1872	0.1993	0.2014
1.80	0.0526	0.1460	0.1616	0.1646
2.00	0.1111	0.1304	0.1343	
2.20	0.0815	0.1046	0.1094	
2.40	0.0562	0.0831	0.0889	
2.60	0.0345	0.0651	0.0721	
2.80	0.0160	0.0521	0.0581	
3.00	0.0377	0.0466		
3.20	0.0273	0.0372		
3.40	0.0186	0.0293		
3.60	0.0113	0.0228		
3.80	0.0051	0.0174		
4.00	0.0130			
4.20	0.0093			
4.40	0.0063			
4.60	0.0038			
4.80	0.0017			

We can now compare the steady-state operating characteristics of the two-channel system to the operating characteristics of the original single-channel system discussed in Section 14.2.

1. The average time a customer spends in the system (waiting time plus service time); is reduced from $W = 4$ minutes to $W = 1.1636$ minutes.
2. The average number of customers in the waiting line is reduced from $L_q = 2.25$ customers to $L_q = 0.1227$ customers.

TABLE .5 STEADY-STATE PROBABILITY OF n CUSTOMERS IN THE SYSTEM FOR THE BURGER DOME TWO-CHANNEL WAITING LINE

Number of Customers	Probability
0	0.4545
1	0.3409
2	0.1278
3	0.0479
4	0.0180
5 or more	0.0109

3. The average time a customer spends in the waiting line is reduced from $W_q = 3$ minutes to $W_q = 0.1636$ minutes.
4. The probability that a customer has to wait for service is reduced from $P_W = 0.75$ to $P_W = 0.2045$.

Clearly the two-channel system will significantly improve the operating characteristics of the waiting line. However, adding an order filler at each service station would further increase the mean service rate and improve the operating characteristics. The final decision regarding the staffing policy at Burger Dome rests with the Burger Dome management. The waiting line study simply provides the operating characteristics that can be anticipated under three configurations: a single-channel system with one employee, a single-channel system with two employees, and a two-channel system with an employee for each channel. After considering these results, what action would you recommend? In this case, Burger Dome adopted the following policy statement: For periods when customer arrivals are expected to average 45 customers per hour, Burger Dome will open two order-processing channels with one employee assigned to each.

By changing the mean arrival rate λ , to reflect arrival rates at different times of the day, and then computing the operating characteristics, Burger Dome's management can establish guidelines and policies that tell the store managers when they should schedule service operations with a single channel, two channels, or perhaps even three or more channels.

NOTE: The multiple-channel waiting line model is based on a single waiting line. You may have also encountered situations where each of the k channels has its own waiting line. Quantitative analysts have shown that the operating characteristics of multiple-channel systems are better if a single waiting line is used. People like them better also; no one who comes in after you can be served ahead of you. Thus, when possible, banks, airline reservation counters, food-service establishments, and other businesses frequently use a single waiting line for a multiple-channel system.

.4 LITTLE'S GENERAL RELATIONSHIPS FOR STEADY-STATE WAITING LINE MODELS

In Sections 14.2 and 14.3 we presented formulas for computing the operating characteristics for single-channel and multiple-channel waiting lines with Poisson arrivals and exponential service times. The operating characteristics of interest included

line	L_q = the average number of units in the waiting
	L = the average number of units in the system
waiting line	W_q = the average time a unit spends in the
system	W = the average time a unit spends in the

John Little showed that several relationships exist among these four characteristics and that these relationships apply to a variety of different waiting line systems. Two of the relationships, referred to as *Little's flow equations*, are

$$L = \lambda W \quad (14.19)$$

$$L_q = \lambda W_q \quad (14.20)$$

Equation (14.19) shows that the average number of units in the system, L , can be found by multiplying the mean arrival rate, λ , by the average time a unit spends in the system, W .

Equation (14.20) shows that the same relationship holds between the average number of units in the waiting line, L_q , and the average time a unit spends in the waiting line, W_q .

Using equation (14.20) and solving for W_q , we obtain

$$W_q = \frac{L_q}{\lambda} \quad (14.21)$$

Equation (14.21) follows directly from Little's second flow equation. We used it for the single-channel waiting line model in Section 14.2 and the multiple-channel waiting line model in Section 14.3 [see equations (14.7) and (14.14)]. Once L_q is computed for either of these models, equation (14.21) can then be used to compute W_q .

Another general expression that applies to waiting time models is that the average time in the system, W , is equal to the average time in the waiting line, W_q , plus the average service time. For a system with a mean service rate μ , the mean service time is $1/\mu$. Thus, we have the general relationship

Recall that we used equation (14.22) to provide the average time in the system for both the single- and multiple-channel waiting line models [see equations (14.8) and (14.15)].

The importance of Little's flow equations is that they apply to *any waiting line model* regardless of whether arrivals follow the Poisson probability distribution and regardless of whether service times follow the exponential probability distribution. For example, in study of the grocery checkout counters at Murphy's Foodliner, an analyst concluded that arrivals follow the Poisson probability distribution with the mean arrival rate of 24 customers per hour or $\lambda = 24/60 = 0.40$ customers per minute. However, the analyst found that service times follow a normal probability distribution rather than an exponential probability distribution. The mean service rate was found to be 30 customers per hour or $\mu = 30/60 = 0.50$ customers per minute. A time study of actual customer waiting times showed that, on average, a customer spends 4.5 minutes in the system (waiting time plus checkout time); that is, $W = 4.5$. Using the waiting line relationships discussed in this section, we can now compute other operating characteristics for this waiting line. First, using equation (14.22) and solving for W_q , we have

$$W_q = W - \frac{1}{\mu} = 4.5 - \frac{1}{0.50} = 2.5 \text{ minutes}$$

With both W and W_q known, we can use Little's flow equations, (14.19) and (14.20), to

$$L = \lambda W = 0.40(4.5) = 1.8 \text{ customers}$$

$$L_q = \lambda W_q = 0.40(2.5) = 1 \text{ customer}$$

Murphy's Foodliner can now review these operating characteristics to see whether action should be taken to improve the service and to reduce the waiting time and the length of the waiting line.

Note: In waiting line systems where the length of the waiting line is limited (e.g., a small waiting area), some arriving units will be blocked from joining the waiting line and will be lost. In this case,

the blocked or lost arrivals will make the mean number of units entering the system something less than the mean arrival rate. By defining λ as the mean number of units *joining the system*, rather than the mean arrival rate, the relationships discussed in this section can be used to determine W , L , W_q , and W_q .

.5 ECONOMIC ANALYSIS OF WAITING LINES

Frequently, decisions involving the design of waiting lines will be based on a subjective evaluation of the operating characteristics of the waiting line. For example, a manager may decide that an average waiting time of one minute or less and an average of two customers or fewer in the system are reasonable goals. The waiting line models presented in the preceding sections can be used to determine the number of channels that will meet the manager's waiting line performance goals.

On the other hand, a manager may want to identify the cost of operating the waiting line system and then base the decision regarding system design on a minimum hourly or daily operating cost. Before an economic analysis of a waiting line can be conducted, a total cost model, which includes the cost of waiting and the cost of service, must be developed.

To develop a total cost model for a waiting line, we begin by defining the notation to be used:

Waiting cost is based on average number of units in the system. It includes the time spent waiting in line plus the time spent being served. Adding more channels always improves the operating characteristics of the waiting line and reduces the waiting cost. However, additional channels increase the service cost. An economic analysis of waiting lines attempts to find the number of

c_w = the waiting cost per time period for each unit

L = the average number of units in the system

c_s = the service cost per time period for each channel

k = the number of channels

TC = the total cost per time period

The total cost is the sum of the waiting cost and the service cost; that is,

$$TC = c_w L + c_s k \quad (14.23)$$

To conduct an economic analysis of a waiting line, we must obtain reasonable estimates of the waiting cost and the service cost. Of these two costs, the waiting cost is usually the more difficult to evaluate. In the Burger Dome restaurant problem, the waiting cost would be the cost per minute for a customer waiting for service. This cost is not a direct cost to Burger Dome. However, if Burger Dome ignores this cost and allows long waiting lines, customers ultimately will take their

business elsewhere. Thus, Burger Dome will experience lost sales and, in effect, incur a cost.

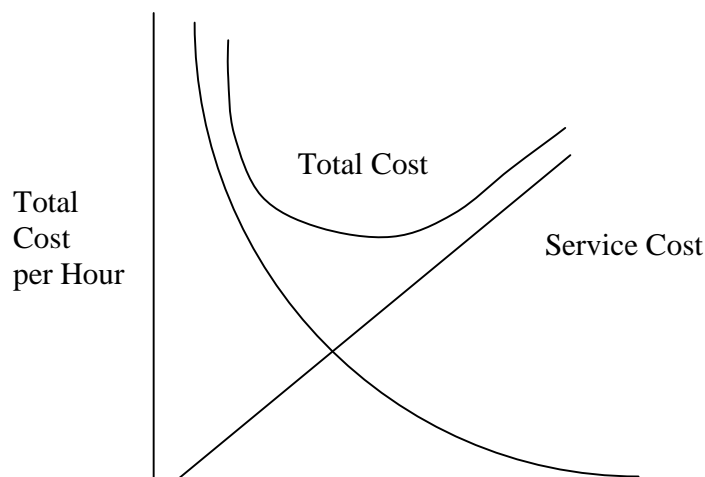
The service cost is generally easier to determine. This cost is the relevant cost associated with operating each service channel. In the Burger Dome problem, this cost would include the server's wages, benefits, and any other direct costs associated with operating the service channel. At Burger Dome, this cost is estimated to be \$7 per hour.

To demonstrate the use of equation (14.23), we assume that Burger Dome is willing to: assign a cost of \$10 per hour for customer waiting time. We use the average number of units in the system, L , as computed in Sections 14.2 and 14.3 to obtain the total hourly cost for the single-channel and two-channel systems:

Thus, based on the cost data provided by Burger Dome, the two-channel system provides the most economical operation. *Harper note: The value assigned to the waiting time of customers was very high in this problem. In most cases, the value assigned is less than the cost of a server or employee. Thus I would recommend something like 10-50% of the employee cost be assigned to the customer waiting time. Keep in mind that this is an intangible cost and is used solely to help balance good customer service with real operational staffing costs such as the cost of employees.*

Figure 14.4 shows the general shape of the cost curves in the economic analysis of waiting lines. The service cost increases as the number of channels is increased. However, with more channels, the service is better. As a result, waiting time and cost decrease as the number of channels is increased. The number of channels that will provide a good approximation of the minimum total cost design can be found by evaluating the total cost for several design alternatives.

FIGURE .4 THE GENERAL SHAPE OF WAITING COST, SERVICE COST, AND TOTAL COST CURVES IN WAITING LINE MODELS



Single-channel system ($L = 3$ customers):
 $TC = c_w L + c_s k = 10(3) + 7(1) = \37.00 per hour

Two-channel system ($L = 0.8727$ customer):
 $TC = c_w L + c_s k = 10(0.8727) + 7(2) = \22.73 per hour

Waiting Cost

Number of Channels (k)

.6 OTHER WAITING LINE MODELS

D. G. Kendall suggested a notation that is helpful in classifying the wide variety of different waiting line models that have been developed. The three-symbol Kendall notation is as follows:

$A/B/k$

where

A denotes the probability distribution for the arrivals
 B denotes the probability distribution for the service time
 k denotes the number of channels

Depending on the letter appearing in the A or B position, a variety of waiting line systems can be described. The letters that are commonly used are as follows:

- M designates a Poisson probability distribution for the arrivals or an exponential probability distribution for service time. The M stands for Markovian (a memory-less distribution).
- D designates that the arrivals or the service time is deterministic or constant
- G designates that the arrivals or the service time has a general probability distribution with a known mean and variance

Using the Kendall notation, the single-channel waiting line model with Poisson arrivals and exponential service times is classified as an $M/M/1$ model. The two-channel waiting line model with Poisson arrivals and exponential service times presented in Section 14.3 would be classified as an $M/M/2$ model.

NOTES AND COMMENTS

In some cases, the Kendall notation is extended to five symbols. The fourth symbol indicates the largest number of units that can be in the system, and the fifth symbol indicates the size of the population. The fourth symbol is used in situations where the waiting line can hold a finite or maximum number of units, and the fifth symbol is necessary when the population of arriving units or customers is finite. When the fourth and fifth symbols, of the Kendall notation are omitted, the waiting line; system is assumed to have infinite capacity, and the population is assumed to be infinite.

Single channel queuing theory

Example

On an average .6 customers reach a telephone booth every hour to make calls. determine the probability that exactly 4 customers will reach in 30 minute period, assuming that arrivals follow Poisson distribution.

Queueing Theory

To solve the queueing system we rely on the underlying Markov chain theory. However we can abstract away from this and use the traffic intensity to derive measures for the queue. The traffic intensity, t , is the service time divided by the inter-arrival time. From this we calculate the measures used in the applet as follows:

Let p_i represent the probability that there are i customers in the system. The probability that the system is idle, p_0 , (ie the Markov chain is in state 0) is given by

$$p_0 = 1 - t.$$

The Utilisation, U , of the system is $1 - p_0$. ie. the proportion of the time that it is not idle.

$$U = 1 - p_0 = t.$$

The probability that the queue is non-empty, B , is the probability of not being in state 0 or state 1 of the Markov chain ie.

$$1 - p_0 - p_1 = 1 - (1-t) - ((1-t)t) = 1 - 1 + t - t + t^2 = t^2.$$

The expectation of the number of customers in the service centre, N , is the sum over all states of the number of customers multiplied by the probability of being in that state.

This works out to be $t/(1-t)$.

The expectation of the number of customers in the queue is calculated similarly but one multiplies by one less than the number of customers.

This works out to be $t^2/(1-t)$.

Markov chains

A Markov chain is a system which has a set S of states and changes randomly between these states in a sequence of discrete steps. The length of time spent in each state is the 'sojourn time' in

that state, T . This is an exponentially distributed random variable and in state i has parameter q_i . If the system is in state i and makes a transition then it has a fixed probability, p_{ij} , of being in state j .

We can construct a Markov chain from a queueing system as follows; assign each possible configuration of the queue a state, define the probability of moving from one state to another by the probability of a customer arriving or departing. Thus state 0 corresponds to there being no customers in the system, state 1 to there being one customer and so on. If we are in state i , then the probability of moving to state $i-1$ is the probability of a customer departing the system, and the probability of moving to state $i+1$ is the probability of a customer arriving in the system (apart from the special case of state 0 when we cannot have a departure).

The fact that we can construct Markov chains from queueing systems means we can use standard techniques from Markov chain theory to find, for example, the probability of the queue having a particular number of customers (by finding the probability of the corresponding Markov chain being in the corresponding state).

Queueing Theory Basics

We have seen that as a system gets congested, the service delay in the system increases. A good understanding of the relationship between congestion and delay is essential for designing effective congestion control algorithms. Queueing Theory provides all the tools needed for this analysis. This article will focus on understanding the basics of this topic.

Communication Delays

Before we proceed further, let's understand the different components of delay in a messaging system. The total delay experienced by messages can be classified into the following categories:

Processing Delay

- This is the delay between the time of receipt of a packet for transmission to the point of putting it into the transmission queue.
- On the receive end, it is the delay between the time of reception of a packet in the receive queue to the

point of actual processing of the message.

- This delay depends on the CPU speed and CPU load in the system.

Queuing Delay

- This is the delay between the point of entry of a packet in the transmit queue to the actual point of transmission of the message.
- This delay depends on the load on the communication link.

Transmission Delay

- This is the delay between the transmission of first bit of the packet to the transmission of the last bit.
- This delay depends on the speed of the communication link.

Propagation Delay

- This is the delay between the point of transmission of the last bit of the packet to the point of reception of last bit of the packet at the other end.
- This delay depends on the physical characteristics of the communication link.

Retransmission Delay

- This is the delay that results when a packet is lost and has to be retransmitted.
- This delay depends on the error rate on the link and the protocol used for retransmissions.

In this article we will be dealing primarily with queueing delay.

Little's Theorem

We begin our analysis of queueing systems by understanding Little's Theorem. Little's theorem states that:

The average number of customers (N) can be determined from the following equation:

$$N = \lambda T$$

Here λ is the average customer arrival rate and T is the average service time for a customer.

Proof of this theorem can be obtained from any standard textbook on queueing theory. Here we will focus on an intuitive understanding of the result. Consider the example of a restaurant where the customer arrival rate (λ) doubles but the customers still spend the same amount of time in the restaurant (T). This will double the number of customers in the restaurant (N). By the same logic if the customer arrival rate remains the same but the customers service time doubles, this will also double the total number of customers in the restaurant.

Queueing System Classification

With Little's Theorem, we have developed some basic understanding of a queueing system. To further our understanding we will have to dig deeper into characteristics of a queueing system that impact its performance. For example, queueing requirements of a restaurant will depend upon factors like:

- How do customers arrive in the restaurant? Are customer arrivals more during lunch and dinner time (a regular restaurant)? Or is the customer traffic more uniformly distributed (a cafe)?
- How much time do customers spend in the restaurant? Do customers typically leave the restaurant in a fixed amount of time? Does the customer service time vary with the type of customer?
- How many tables does the restaurant have for servicing customers?

The above three points correspond to the most important characteristics of a queueing system. They are explained below:

Arrival Process

- The probability density distribution that determines the customer arrivals in the system.
- In a messaging system, this refers to the message arrival probability distribution.

Service Process

- The probability density distribution that determines the customer service times in the system.
- In a messaging system, this refers to the message transmission time

distribution. Since message transmission is directly proportional to the length of the message, this parameter indirectly refers to the message length distribution.

Number of Servers

- Number of servers available to service the customers.
- In a messaging system, this refers to the number of links between the source and destination nodes.

Based on the above characteristics, queueing systems can be classified by the following convention:

A/S/n

Where A is the arrival process, S is the service process and n is the number of servers. A and S can be any of the following:

M (Markov)	Exponential probability density
D (Deterministic)	All customers have the same value
G (General)	Any arbitrary probability distribution

Examples of queueing systems that can be defined with this convention are:

- **M/M/1:** This is the simplest queueing system to analyze. Here the arrival and service time are negative exponentially distributed (poisson process). The system consists of only one server. This queueing system can be applied to a wide variety of problems as any system with a very large number of independent customers can be approximated as a Poisson process. Using a Poisson process for service time however is not applicable in many applications and is only a crude approximation. Refer to M/M/1 Queueing System for details.
- **M/D/n:** Here the arrival process is poisson and the service time distribution is deterministic. The system has n servers. (e.g. a ticket booking counter with n cashiers.) Here the service time can be assumed to be same for all customers)
- **G/G/n:** This is the most general queueing system where the arrival and service time processes are both arbitrary. The system has n servers. No analytical solution is known for this queueing system.

Poisson Arrivals

M/M/1 queueing systems assume a Poisson arrival process. This assumption is a very good approximation for arrival process in real systems that meet the following rules:

1. The number of customers in the system is very large.
2. Impact of a single customer on the performance of the system is very small, i.e. a single customer consumes a very small percentage of the system resources.
3. All customers are independent, i.e. their decision to use the system are independent of other users.

Cars on a Highway

As you can see these assumptions are fairly general, so they apply to a large variety of systems. Lets consider the example of cars entering a highway. Lets see if the above rules are met.

1. Total number of cars driving on the highway is very large.
2. A single car uses a very small percentage of the highway resources.
3. Decision to enter the highway is independently made by each car driver.

The above observations mean that assuming a Poisson arrival process will be a good approximation of the car arrivals on the highway. If any one of the three conditions is not met, we cannot assume Poisson arrivals. For example, if a car rally is being conducted on a highway, we cannot assume that each car driver is independent of each other. In this case all cars had a common reason to enter the highway (start of the race).

Telephony Arrivals

Lets take another example. Consider arrival of telephone calls to a telephone exchange. Putting our rules to test we find:

1. Total number of customers that are served by a telephone exchange is very large.
2. A single telephone call takes a very small fraction of the systems resources.
3. Decision to make a telephone call is independently made by each customer.

Again, if all the rules are not met, we cannot assume telephone arrivals are Poisson. If the telephone exchange is a PABX catering to a few subscribers, the total number of customers is small, thus we cannot assume that rule 1 and 2 apply. If rule 1 and 2 do apply but telephone calls are being initiated due to some disaster, calls cannot be considered independent of each other. This violates rule 3.

Poisson Arrival Process

Now that we have established scenarios where we can assume an arrival process to be Poisson. Lets look at the probability density distribution for a Poisson process. This equation describes the probability of seeing n arrivals in a period from 0 to t .

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Where:

- t is used to define the interval 0 to t
- n is the total number of arrivals in the interval 0 to t .
- λ is the total average arrival rate in arrivals/sec.

Negative Exponential Arrivals

We have seen the Poisson probability distribution. This equation gives information about how the probability is distributed over a time interval. Unfortunately it does not give an intuitive feel of this distribution. To get a good grasp of the equation we will analyze a special case of the distribution, the probability of no arrivals taking place over a given interval.

Its easy to see that by substituting n with 0, we get the following equation:

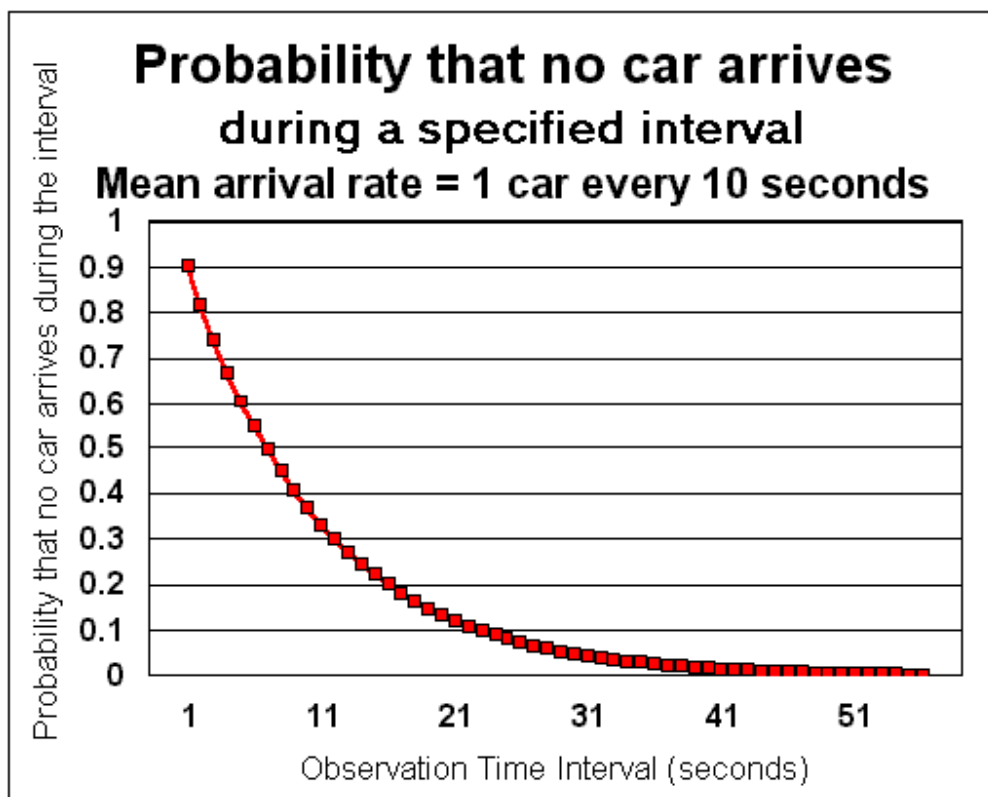
$$P_0(t) = e^{-\lambda t}$$

This equation shows that probability that no arrival takes place during an interval from 0 to t is negative exponentially related to the length of the interval. This is better illustrated with an example.

Consider a highway with an average of 1 car arriving every 10 seconds (0.1 cars/second arrival rate). The probability distribution with t is given below. You can see here that the probability of not seeing a single car on the highway decreases dramatically with the

observation period. If you observe the highway for a period of 1 second, there is 90% chance that no car will be seen during that period. If you monitor the highway for 20 seconds, there is only a 10% chance that you will not see a car on the highway. Put another way, there is only a 10% chance two cars arrive less than one second apart. There is a 90% chance that two cars arrive less than 20 seconds apart.

In the figure below, we have just plotted the impact of one arrival rate. If another graph was plotted after doubling the arrival rate (1 car every 5 seconds), the probability of not seeing a car in an interval would fall much more steeply.



Poisson Service Times

In an M/M/1 queueing system we assume that service times for customers are also negative exponentially distributed (i.e. generated by a Poisson process). Unfortunately, this assumption is not as general as the arrival time distribution. But it could still be a reasonable assumption when no other data is available about service times. Lets see a few examples:

Telephone Call Durations

Telephone call durations define the service time for utilization of various resources in a telephone exchange. Lets see if telephone call durations can be assumed to be negative exponentially distributed.

1. Total number of customers that are served by a telephone exchange is very large.
2. A single telephone call takes a very small fraction of the systems resources.
3. Decision on how long to talk is independently made by each customer.

From these rules it appears that negative exponential call hold times are a good fit. Intuitively, the probability of a customers making a very long call is very small. There is a high probability that a telephone call will be short. This matches with the observation that most telephony traffic consists of short duration calls. (The only problem with using the negative exponential distribution is that, it predicts a high probability of extremely short calls).

This result can be generalized in all cases where user sessions are involved.

Transmission Delays

Lets see if we can assume negative exponential service times for messages being transmitted on a link. Since the service time on a link is directly proportional to the length of the message, the real question is that can we assume that message lengths in a protocol are negative exponentially distributed?

As a first order approximation you can assume so. But message lengths aren't really independent of each other. Most communication protocols exchange messages in a certain sequence, the length distribution is determined by the length of the messages in the sequence. Thus we cannot assume that message lengths are independent. For example, internet traffic message lengths are not distributed in a negative exponential pattern. In fact, length distribution on the internet is bi-modal (i.e. has two distinct peaks). The first peak is around the length of a TCP ack message. The second peak is around the average length of a data packet.

Single Server

With M/M/1 we have a single server for the queue. Suitability of M/M/1 queueing is easy to identify from the server standpoint. For example, a single transmit queue feeding a single link qualifies as a single server and can be modeled as an M/M/1 queueing system. If

a single transmit queue is feeding two load-sharing links to the same destination, M/M/1 is not applicable. M/M/2 should be used to model such a queue.

M/M/1 Results

As we have seen earlier, M/M/1 can be applied to systems that meet certain criteria. But if the system you are designing can be modeled as an M/M/1 queueing system, you are in luck. The equations describing a M/M/1 queueing system are fairly straight forward and easy to use.

First we define ρ , the traffic intensity (sometimes called occupancy). It is defined as the average arrival rate (λ) divided by the average service rate (μ). For a stable system the average service rate should always be higher than the average arrival rate. (Otherwise the queues would rapidly race towards infinity). Thus ρ should always be less than one. Also note that we are talking about average rates here, instantaneous arrival rate may exceed the service rate. Over a longer time period, the service rate should always exceed arrival rate.

$$\rho = \frac{\lambda}{\mu}$$

Mean number of customers in the system (N) can be found using the following equation:

$$N = \frac{\rho}{1 - \rho}$$

You can see from the above equation that as ρ approaches 1 number of customers would become very large. This can be easily justified intuitively. ρ will approach 1 when the average arrival rate starts approaching the average service rate. In this situation, the server would always be busy hence leading to a queue build up (large N).

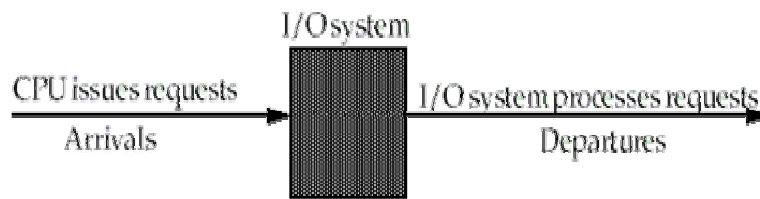
Lastly we obtain the total waiting time (including the service time):

$$T = \frac{1}{\mu - \lambda}$$

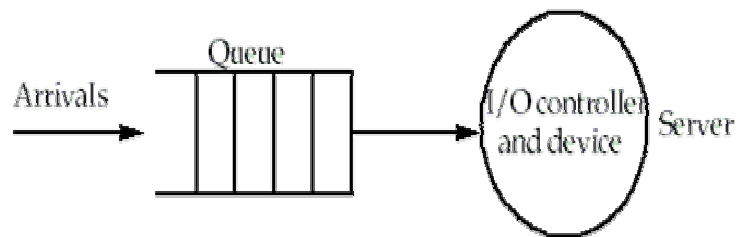
Again we see that as mean arrival rate (λ) approaches mean service rate (μ), the waiting time becomes very large.

Queuing theory

- Given the importance of response time (and throughput), we need a means of computing values for these metrics.
- Our black box model:



- Let's assume our system is in steady-state (input rate = output rate).
- The contents of our black box.



- I/O requests "depart" by being completed by the server.

Queuing theory

- Elements of a queuing system:
 - Request & arrival rate
 - This is a single "request for service".
 - The rate at which requests are generated is the arrival rate.
 - Server & service rate
 - This is the part of the system that services requests.
 - The rate at which requests are serviced is called the service rate.
 - Queue

- This is where requests wait between the time they arrive and the time their processing starts in the server.

Queuing theory

- Useful statistics
- $Length_{queue}$, $Time_{queue}$
 - These are the average length of the queue and the average time a request spends waiting in the **queue** .
- $Length_{server}$, $Time_{server}$
 - These are the average number of tasks being serviced and the average time each task spends in the **server** .
 - Note that a server may be able to serve more than one request at a time.
- $Time_{system}$, $Length_{system}$
 - This is the average time a request (also called a task) spends in the **system** .
 - It is the sum of the time spent in the queue and the time spent in the server.
 - The length is just the average number of tasks anywhere in the system.

Queuing theory

- Useful statistics
- Little's Law
 - The mean number of tasks in the **system** = arrival rate * mean response time .

$$Length_{System} = Arrival\ Rate \times Time_{System}$$

- This is true only for systems in equilibrium.
 - We must assume any system we study (for this class) is

in such a state.

- **Server utilization**

- This is just

$$\text{Server utilization} = \frac{\text{Arrival Rate}}{\text{Server Rate}} \quad \text{where Rate} = 1/\text{Time}$$

- This must be between 0 and 1.

- If it is larger than 1, the queue will grow infinitely long.

- This is also called traffic intensity .

Queuing theory

- **Queue discipline**

- This is the order in which requests are delivered to the server.

- Common orders are FIFO, LIFO, and random.

- For FIFO, we can figure out how long a request waits in the queue by:

$$\text{Time}_{\text{System}} = \text{Length}_{\text{Queue}} \times \text{Time}_{\text{Server}} + ?$$

Mean time for server to finish current tasks when request arrives

- The last parameter is the hardest to figure out.

- We can just use the formula:

$$\text{Average residual service time} = \frac{1}{2} \times \text{Weighted mean time} \times (1 + C)$$

- C is the coefficient of variance, whose derivation is in the book.

- (don't worry about how to derive it -

*this isn't a
class on
queuing
theory.)*

Queuing theory

- *Example: Given:*
 - *Processor sends 10 disk I/O per second (which are exponentially distributed).*
 - *Average disk service time is 20 ms.*

- *On average, how utilized is the disk?*

$$\text{Server utilization} = \frac{\text{Arrival Rate}}{\text{Server Rate}} = \frac{10}{\frac{1}{0.02}} = 0.2$$

- *What is the average time spent in the queue?*
 - *When the service distribution is exponential, we can use a simplified formula for the average time spent waiting in line:*

$$\text{Time}_{\text{queue}} = \text{Time}_{\text{server}} \times \frac{\text{Server utilization}}{(1 - \text{Server utilization})} = 20\text{ms} \times \frac{0.2}{(1 - 0.2)} = 5\text{ms}$$

- *What is the average response time for a disk request (including queuing time and disk service time)?*

$$\text{Time}_{\text{queue}} + \text{Time}_{\text{server}} = 5 + 20\text{ms} = 25\text{ms}$$

Queuing theory

- *Basic assumptions made about problems:*
 - *System is in equilibrium.*
 - *Interarrival time (time between two successive requests arriving) is exponentially distributed.*
 - *Infinite number of requests.*
 - *Server does not need to delay between servicing requests.*
 - *No limit to the length of the queue and queue is FIFO.*
 - *All requests must be completed at some point.*

- *This is called an M/G/1 queue*
 - *M = exponential arrival*

- G = general service distribution (i.e. not exponential)
 - 1 = server can serve 1 request at a time
- It turns out this is a good model for computer science because many arrival processes turn out to be **exponential** .
- Service times , however, may follow any of a number of distributions.

Disk Performance Benchmarks

- We use these formulas to predict the performance of storage subsystems.
- We also need to measure the performance of real systems to:
 - Collect the values of parameters needed for prediction.
 - To determine if the queuing theory assumptions hold (e.g., to determine if the queueing distribution model used is valid).
- Benchmarks:
 - Transaction processing
 - The purpose of these benchmarks is to determine how many small (and usually random) requests a system can satisfy in a given period of time.
 - This means the benchmark stresses **I/O rate** (number of disk accesses per second) rather than **data rate** (bytes of data per second).
 - Banks, airlines, and other large customer service organizations are most interested in these systems, as they allow simultaneous updates to little pieces of data from many terminals.

Disk Performance Benchmarks

- TPC-A and TPC-B

- *These are benchmarks designed by the people who do transaction processing.*
- *They measure a system's ability to do random updates to small pieces of data on disk.*
- *As the number of transactions is increased, so must the number of requesters and the size of the account file .*
 - *These restrictions are imposed to ensure that the benchmark really measures disk I/O.*
 - *They prevent vendors from adding more main memory as a database cache, artificially inflating TPS rates.*
- *SPEC system-level file server (SFS)*
 - *This benchmark was designed to evaluate systems running Sun Microsystems network file service, NFS.*

Disk Performance Benchmarks

- *SPEC system-level file server (SFS)*
 - *It was synthesized based on measurements of NFS systems to provide a reasonable mix of reads, writes and file operations.*
 - *Similar to TPC-B, SFS **scales** the size of the file system according to the reported throughput , i.e.,*
 - *It requires that for every 100 NFS operations per second, the size of the disk must be increased by 1 GB.*
 - *It also limits average response time to 50ms.*
- *Self-scaling I/O*

*This method of I/O benchmarking uses a program that **automatically scales***

several parameters that govern performance.

- *Number of unique bytes touched.*
 - *This parameter governs the total size of the data set.*
 - *By making the value large, the effects of a cache can be counteracted.*

Disk Performance Benchmarks

- *Self-scaling I/O*
- *Percentage of reads.*
- *Average I/O request size.*
 - *This is scalable since some systems may work better with large requests, and some with small.*
- *Percentage of sequential requests.*
 - *The percentage of requests that sequentially follow (address-wise) the prior request.*
 - *As with request size, some systems are better at sequential and some are better at random requests.*
- *Number of processes.*
 - *This is varied to control concurrent requests, e.g., the number of tasks simultaneously issuing I/O requests.*

Disk Performance Benchmarks

- *Self-scaling I/O*

- *The benchmark first chooses a nominal value for each of the five parameters (based on the system's performance).*
- *It then varies each parameter in turn while holding the others at their nominal value.*

Performance can thus be graphed using any of five axes to show the effects of changing parameters on a system's performance.

