Unit 4: Probability, Random Variables, and Probability Distributions

Unit 4 Introduction

Much of what we have done so far has been based on the assumption of correlation and causation between variables. But there are also things that are completely unrelated and determined by random chance. This is the domain of probability. Dice rolls, gambling, drawing cards, and many other things are governed by probability or random chance.

Many video games have a probabilistic element to them – if you are playing a game and you know that a certain enemy has a "rare drop" – where there is a low chance of you receiving a rare item upon defeating the enemy. Whoever programmed that had to have an understanding of probability and statistics to set this up in the game.

But beyond gambling and video games, randomness is incredibly important. A good knowledge of what randomness "looks like", mathematically, can ultimately help us figure out if a difference is statistically significant or not. If we understand probability well enough, we should be able to apply that knowledge (in a later unit) to determine if data that we collect is random or if it actually establishes a statistically significant relationship.

In other words, just because we encounter a pattern in our data, it does not mean the data was not random. Suppose I flipped a coin ten times and got the following results (H = heads, T = tails):

H H T T H H T T H H

It looks like there is a pattern to the data. I get two heads, then two tails, and so on. But that does not mean the coin is not random, nor does it mean that the next two flips will definitely be tails. Being able to establish what is actually a pattern and what is actually random is a huge part of statistics.

Imagine that you are a researcher, and instead of flipping coins you collected experimental data. If you see a pattern, you had better be able to make sure that it was an actual pattern and not just randomness. Reporting random data as a pattern can have disastrous results. (If you think I am overstating this, just a few years ago, a research team tried to recreate the experiments in a series of social science and psychology journals – 5 out of 13 studies were not replicable – the data that the original team was actually random and not a pattern. If you think that is not a big deal, psychologists will make treatment decisions for patients based on published studies; if some of those studies actually don't prove what we thought they did, people may be getting improper mental health care. Analyzing random data or patterns can be incredibly important.)

4.1 Estimating Probabilities Using Simulations and Probability in General

Objectives:

- Estimate probabilities using a simulation.
- Determine the sample space for a random experiment.
- Calculate probabilities for events and their complements.
- Interpret probabilities for events.

Before we talk about *simulations with probabilities*, we need to understand some basic vocabulary about **probability**:

- **Random Process:** A process that generates results determined by chance.
- **Random Experiment:** An experiment in which the outcome of each trial is determined by chance.
- **Outcome:** The result of a *trial* (individual occurence) of a random process.
- **Event:** A collection of possible outcomes.
- **Sample Space:** The collection of *all* possible outcomes.

Example 1: List the sample space of biological sex for a two child family.

Answer: M = Male, F = Female

S = {MM, MF, FM, FF}

Example 2: An experiment consists of flipping a coin and rolling a die.

- What is the sample space?
- What is the number of possible outcomes?

Answer: H = Heads, T = Tails

S = {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}

There are 12 possible outcomes.

Example 3: Given that a fair six-sided die is rolled, and a success is defined as rolling a 5 or a 6. List the sample space, the event that is a success, and the event that is a failure.

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

Success event: $E_1 \in \{5,6\}$

Failure event: $E_2 \in \{1,2,3,4\}$

Notice that I used S for the set that was the sample space, and E for the set that was the event. This is a fairly standard, but not essential, method of presenting these values.

Example 4: For each of the following determine the sample space. It may be helpful to use a **tree diagram** for parts (b) and (c).

(a) Toss 2 coins.

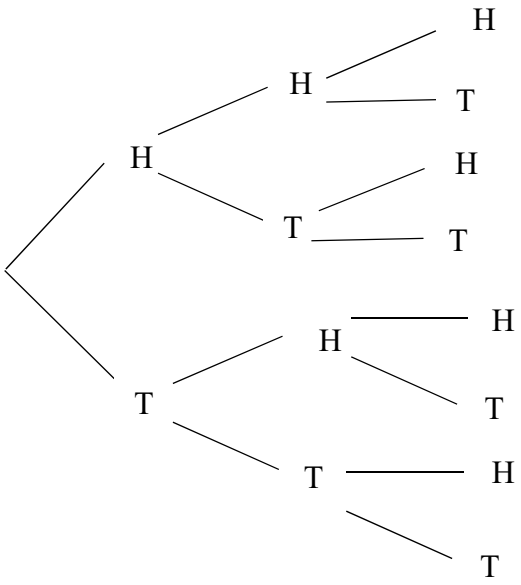
(b) Toss 3 coins.

(c) Toss 4 coins.

a. Toss 2 coins:

S = {HH, HT, TH, TT}

b. Toss 3 coins, using a tree diagram.



Then simply follow the "branches" of the tree to get the sample space:

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

c. Make a similar tree to the one above, but add the last set of branches for the fourth toss: HHTT,

S ={HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}

We already know that we can generate random numbers from a table, from selecting them from a container, or by using a calculator or computer. We will use this process to *simulate* real world events with probability. Why would we use a simulation? There are a number of reasons for this, but a simple one is that we can use simulations to project how something in the real world may actually happen. Suppose I know that a certain free throw shooter in the NBA hits 70% of his shots. If I want to get an idea of how many free throw that player might make in a game where he shoots 8 free throws, I could run a *simulation*. I could simulate this event many times, and I could not actually run a real experiment (I couldn't force this player to shoot 8 free throws in games a bunch of times, could I?)

Let's run this simulation 5 times, using our calculator.

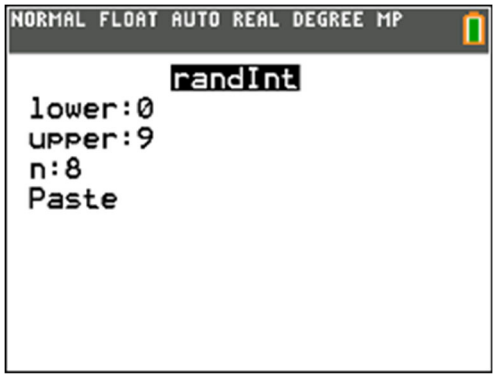
Since the chance of hitting a free throw is 70%, I can generate random numbers from 1 to 10.

I will define success and failure as follows:

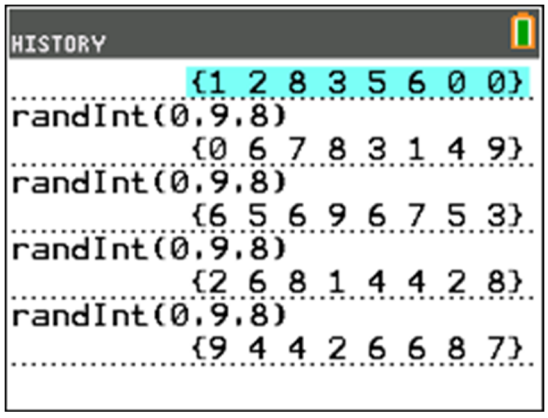
Making a shot: {1, 2, 3, 4, 5, 6, 7}

Missing a shot: {8, 9, 0}

Then I will use the calculator to run the simulation 5 times. Remember, hit the **MATH** button, go to the PROB Menu, and select option 5:randInt(.



After that, I will just hit enter 5 times to run the simulation five times.



Trial 1: 5 makes, 3 misses

Trial 2: 6 makes, 2 misses

Trial 3: 7 makes, 1 miss

Trial 4: 6 makes, 2 misses

Trial 5: 5 makes, 3 misses

If you totaled up the shots you would find for this simulation, the percentage of makes was 72.5%.

As you run more trials, your simulated probabilities will tend to approach the true probability.

1. **The Law of Large Numbers:** Simulated probabilities tend to approach empirical probabilities as the number of trials increases.

Sometimes, we have a more complicated question. Running a simulation can be used to estimate the probability of a given event happening.

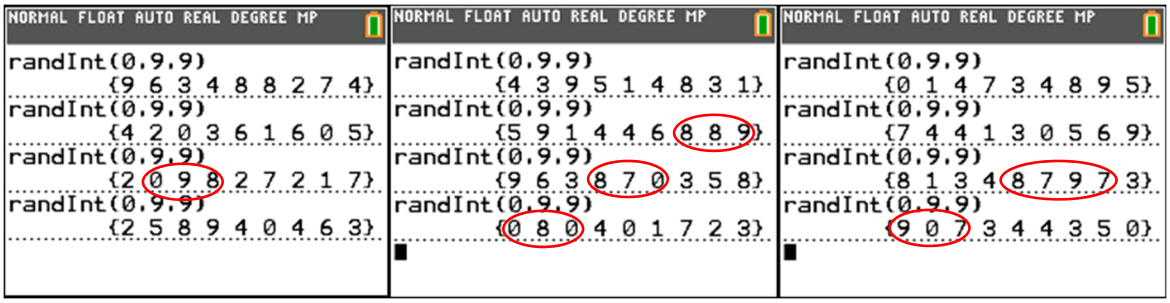
Example 5: In a certain season, Shaquille O'Neal had a free throw percentage of 60% and averaged 9 shots per game. Run a simulation of twelve games to determine how likely it is for him to miss 3 or more free throws in a row.

I will define success and failure as follows:

Making a shot: {1, 2, 3, 4, 5, 6}

Missing a shot: {7, 8, 9, 0}

Your answer will vary from mine as your random number generator will have different results from the one shown below.



According to my simulation, missing three or more free throws in these parameters is about 6/12 or 50%.

If I had asked for the probability of missing 4 or more in a row, this simulation would have shown a probability of 1/12 or about 8.3%

Steps to Running a **Simulation:**

1. Assign the digits for the random digit table or for your calculator.
2. Describe how the simulation will be run.
    - If using random digits, be sure to state whether duplicates are allowed.
3. Give a stopping rule (if applicable)
4. State what is to be measured.
5. Conduct the simulation with a reasonable number of replications.
6. State the conclusion reached in the context of the problem.

Example 6: A player's batting average is 0.400. He bats five times per game. How could you assign random digits to simulate his times at bat?

A trial/game would consist of how many numbers?

Use the following digits to simulate baseball games.

03672	04177	29104	61579	00123
49810	65214	43629	59012	43435

Given the digits above, how many games can be simulated?

Based on your simulation, what is the probability that he gets at least 2 hits in five times at bat in a game?

Answer: A trial would consist of 5 numbers (representing 5 at bats in a game).

Getting a hit: {1, 2, 3, 4}

Not getting a hit: {5, 6, 7, 8, 9, 0}

03672	04177	29104	61579	00123
49810	65214	43629	59012	43435

This player got at least 2 hits in 8/10 of the simulated games (or about 80%)

Example 7: In each of the following situations, describe a sample space *S* for the indicated random phenomenon. Do the sample spaces consist of **discrete** or **continuous** numerical data?

(a) Choose a student in your class at random. Ask how much time that student spent studying during the past 24 hours.

(b) The Physicians' Health Study asked 11,000 physicians to take an aspirin every other day and observed how many of them had a heart attack in a five-year period.

(c) In a test of new package design, you drop a carton of a dozen eggs from a height of 1 foot and count the number of broken eggs.

(d) A nutrition researcher feeds a new diet to a young male white rat. The response variable is the weight (in grams) that the rat gains in 8 weeks.

Answers:

(a) Continuous. $S \in [0, 24]$
(b) Discrete. $S = \{0, 1, 2, 3, \ldots 11000\}$
(c) Discrete. $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
(d) Continuous. $S \in (-\infty, \infty)$

Recall from the beginning of this section:

- **Event:** A collection of possible outcomes.
- **Sample Space:** The collection of *all* possible *non-overlapping* outcomes.

It is very important for the topic at hand that we emphasize the "non-overlapping" nature of the outcomes in this section (we will deal with overlapping outcomes in a later section). This allows us to have some basic definitions with which to work.

For example, if you roll a die with six sides, there are six non-overlapping possibilities that you could roll. Since all of the numbers are distinct values, the options do not overlap. The fact that the outcomes do not overlap lead us to the following:

- If all outcomes are equally distributed, then the probability of an event (E) will occur is given by

$$\frac{\text{number of outcomes in event } E}{\text{total number of outcomes in sample space } S}$$

- The probability of an event E is a number between 0 and 1 (inclusive).
  - The sum of all non-overlapping outcomes is 1 (100%)
- **Complement of an Event:** This is the outcomes for an event *not happening*.
  - If an event is noted by E, then the complement of the event is E^C or E'.

We should also get used to the following notation:

- $P(E)$ is the probability of event E. So $P(A)$ is the probability of event A.
  - The probability of the complement of an event is $P(E^C) = 1 - P(E)$

Example 8: Given that a fair six-sided die is rolled, and a success is defined as rolling a 5 or a 6. List the sample space, the event, and the complement of the event. Determine the probability of the event and its complement.

Answer: Note that this is the same basic question as in example 3

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

Event: $E \in \{5,6\}$ $\qquad P(E) = \frac{1}{3}$ or 0.333

Complement of the event: $E^C \in \{1,2,3,4\}$ $\qquad P(E^C) = \frac{2}{3}$ or 0.667

Another important concept is that probabilities of events in <u>*repeatable situations*</u> can be interpreted as the relative frequency with which the event will occur *in the long run*.

Example 9: Given the event described in example 8, how many successes would you expect if you rolled the die 60 times?

You would expect that $\frac{1}{3}$ of the rolls would give you a 5 or 6 (the defined success), so you would expect 20 successes. This does not guarantee you will get 20 successes; it just means that if you roll enough times about a third of the results will be successes. The more trials you run, the more likely you will get 1/3 of the rolls as successes.

Example 10: A store owner reports that the probability that a customer who purchases a television set will also purchase insurance for the television is 0.45.

Which of the following is the correct interpretation of the probability of 0.45?

(a) For all customers who purchase a television, 45% will also purchase insurance.
(b) For all customers of the store, 45% will purchase a television.
(c) For all customers who purchase insurance, 45% will use the insurance.
(d) From the next 20 customers, 9 will purchase insurance.
(e) From the next 20 customers, 9 will purchase a television.

Answer: The correct answer is (a) as it matches the "long haul" idea for probability. (b) and (c) are incorrect because they misapply the probability (they state the probability is for something other than what is in the initial problem). (d) and (e) make the mistake of trying to apply the probability to an exact number of transactions.

Summary:

- **Random Process:** A process that generates results determined by chance.
- **Random Experiment:** An experiment in which the outcome of each trial is uncertain (determined by chance).
- **Outcome:** The result of a *trial* (individual occurence) of a random process.
- **Event:** A collection of possible outcomes.
- **Sample Space:** The collection of *all* possible *non-overlapping* outcomes.
- **Steps to Creating a Simulation:**
  1. Assign the digits for the random digit table or for your calculator.
  2. Describe how the simulation will be run.
     - If using random digits, be sure to state whether duplicates are allowed.
  3. Give a stopping rule (if applicable)
  4. State what is to be measured.
  5. Conduct the simulation with a reasonable number of replications.
  6. State the conclusion reached in the context of the problem.
- If all outcomes are equally distributed, then the probability of an event (E) will occur is given by

$$\frac{\text{number of outcomes in event } E}{\text{total number of outcomes in sample space } S}$$

- The probability of an event E is a number between 0 and 1 (inclusive).
  - The sum of all non-overlapping outcomes is 1 (100%)
- **Complement of an Event:** This is the outcomes for an event *not happening*.
  - If an event is noted by E, then the complement of the event is E^C or E'.
- $P(E)$ is the probability of event E. So $P(A)$ is the probability of event A.
  - The probability of the complement of an event is $P(E^C) = 1 - P(E)$

Checkpoint 4.1

Multiple Choice

1. The theoretical probability of rolling a 2 or a 6 on one roll of a die is $\frac{1}{3}$. A student who tries to reproduce this probability by counting successes in repeated trials is most likely to come closest to $\frac{1}{3}$ with

   (a) one roll of the die
   (b) two rolls of the die
   (c) three rolls of the die
   (d) ten rolls of the die
   (e) one hundred rolls of the die

2. Suppose there are five outcomes to an experiment and a computer calculates the respective probabilities of the outcomes to be 0.4, 0.5, 0.3, 0, and -0.2. The proper conclusion is that

   (a) the sum of the individual probabilities is 1.
   (b) one of the outcomes will never occur.
   (c) one of the outcomes will occur 50% of the time.
   (d) all of the above are true.
   (e) there is a something wrong with the computer program as probabilities cannot be negative.

3. Lexus automobiles come in the following colors and are manufactured in these percentages:

White	Black	Red	Silver	Gold
0.46	0.22	0.09	0.11	0.12

   If I choose a car at random, what is the probability that I do not choose a white or black car?

   (a) 0.68
   (b) 0.46
   (c) 0.20
   (d) 0.32
   (e) None of these

Free Response

1. Explain how you could use a 10–sided number cube to simulate the answer to a true or false question.

2. The probability that a soccer player will score a goal is $\frac{2}{5}$. How could random digits be assigned to simulate whether or not the player makes a goal?

3. A student's daily homework average is 70% (7 out of 10 correct). She has a new assignment to turn in.

   (a) How could you use random digits to simulate the grade on this assignment?

   (b) A trial would consist of _____ numbers.

   (c) Based on the simulation below, what is the probability that she gets at least half of her homework problems correct?

   16658 50687 42935 12647 15989 81604 74563 61773 48433 60048

   (d) How many assignments are represented above?

4.1 Homework

1) A teacher determines that average rate at which students in her class of twenty-four students turn in homework is 80%.
    a. How could you use random digits to simulate the likelihood of a student turning in homework in her class? What number of trials would you use to simulate a class turning in homework?
    b. Use the random digit table below to run a simulation for her class turning in an assignment. Determine the percentage of assignments students would have turned in in your simulation.

19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095	32863	29485	82226	90056
52711	38889	93074	60227	40011	85848	48767	52573
95592	94007	69971	91481	60779	53791	17297	59335

    c. Suppose the teacher has 5 assignments in a given quarter. Use your calculator to run 15 simulations of this. What is the percentage of assignments turned in for each of your simulations? What is the overall percentage of assignments turned in?

2) An NBA player has a success rate of 90% on free throws, and he averages 5 free throw attempts per game.
    a. How could you use random digits to simulate the number of free throws he makes in a given game? What number of trials would you use to simulate this?
    b. Use the random digit table below to run a simulation of 8 games for this player. Determine the free throw percentage over the duration of this simulation.

95592	94007	69971	91481	60779	53791	17297	59335

    c. Suppose you want to know how likely it is for this player to get streaks of many free throws in a row in a given season. Suppose the player will play 40 games averaging 3.9 free throws a night for a total of at most 160 free throw attempts. Use the random digit table below to simulate this and determine the longest streak of made free throws as well as his average streak length.

93074	60227	40011	85848	48767	52573	73676	47150
99400	01927	27754	19223	95034	05756	28713	96409
12531	42544	82853	42648	82425	36290	77558	00095
32863	29485	82226	90056	45467	71709	52711	38889

3) Lexus automobiles come in the following colors and are manufactured in these percentages:

White	Black	Red	Silver	Gold
0.46	0.22	0.09	0.11	0.12

Use this information to determine each of the following:

a. P(Black)
b. P(Black^C)
c. P(Silver or Gold)
d. P(Red^C)
e. P(Blue)
f. P(Blue^C)

4) Which of the following do you think is more likely if you are flipping a fair coin?

I. Getting 7 or more heads flipping the coin 10 times.
II. Getting 70 or more heads flipping the coin 100 times.

Explain why you think the option you chose was correct.

5) Consider the chance experiment in which the type of transmission of the next two cars purchased from a particular dealership is being recorded (transmissions are either automatic (A) or manual (M)).

a. What is the set of all possible outcomes for this experiment (the sample space)?
b. Draw a tree diagram displaying all possible outcomes.

6) A college library has four copies of a book that are numbered 1, 2, 3, and 4. Two of these books are selected at random. One of the two can be put on a two-hour hold, and the other can be checked out overnight.

a. Construct a tree diagram to display all of the outcomes in the sample space. How many possible outcomes are there?
b. Suppose *A* is the event that <u>at least</u> one of the books selected is odd-numbered. What outcomes are in *A*?
c. Suppose that books 1 and 2 are soft-cover, and books 3 and 4 are hard-cover. Let *B* denote the event that <u>exactly one</u> of the books selected is a soft-cover book. What outcomes are contained in *B*?

4.2 Mutually Exclusive Events, Conditional Probability, and Independent Events

Objectives:

- Explain why two events are or are not mutually exclusive.
- Calculate conditional probabilities.
- Calculate the probabilities for independent events.
- Calculate the probabilities for unions of two events.

Recall our definition of events and complements from the previous section. This leads us to the ideas of *unions* and *intersections.*

- **Event:** A subset of outcomes of the sample space.
- Let A and B denote two events.

1. The event ***not A*** consists of all experimental outcomes that are not in event A. Not A is sometimes called the **complement of *A*** and is usually denoted by A^C, A', or rarely $\overline{A}$.
2. **Union of two events:** The event ***A or B*** consists of *all experimental outcomes* that are in *at least* one of the two events, that is, in A or in B or in both of these. It is also denoted by $A \cup B$.
3. **Intersection of two events:** The event ***A and B*** consists of *all experimental outcomes* that are *in both* the event A and the event B. It is also denoted by $A \cap B$.

Example 1: Let A be the event of rolling a number less than four from a fair six-sided die. Let B be the event of an even number between 0 and 9, inclusive.

(a) What are the elements of $A \cup B$?

(b) What are the elements of $A \cap B$?

(c) What are the elements of A^C?

(d) What are the elements of $A^C \cup B$?

(e) What are the elements of $A^C \cap B$

Answers:

(a) $A \cup B = \{1, 2, 3, 4, 6, 8\}$

Notice that 2 and 4 were members of both sets, while the rest of the numbers were members of either *A* or *B*.

(b) $A \cap B = \{2, 4\}$

Since this is an intersection, we only want the numbers that occur in both sets – that is, 2 and 4.

(c) $A^C = \{5, 6\}$

(d) $A^C \cup B = \{5, 6, 8\}$

(e) $A^C \cap B = \{6\}$

Example 2: Let *A* be the event of rolling a number a three or a five from a fair six-sided die. Let *B* be the event of an even number between 0 and 9, inclusive. What are the elements of $A \cap B$?

There are no shared outcomes of *A* and *B*, so we could write $A \cap B = \{\ \}$ or $A \cap B = \varnothing$

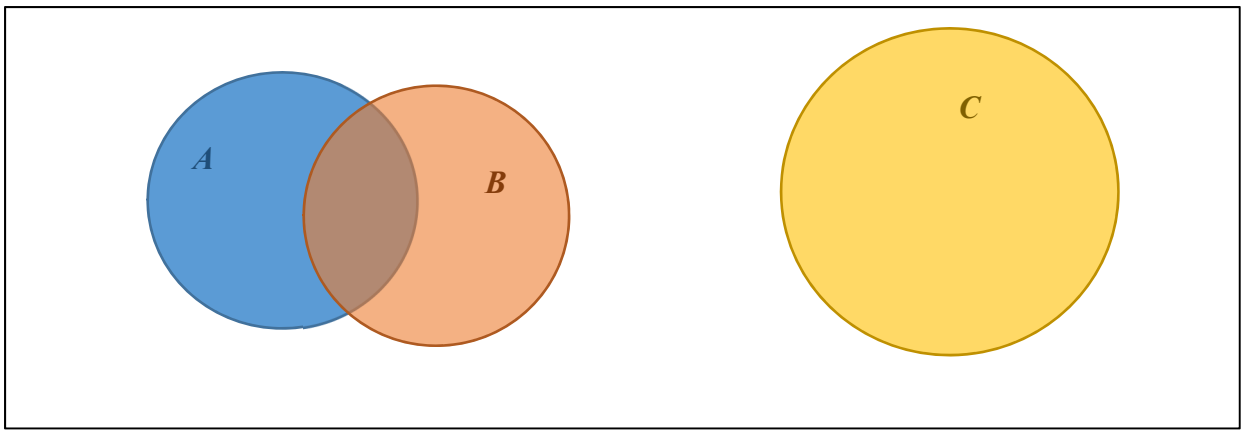
This is also called the "null set" (a set with no elements).

The two sets in Example 2 are called ***mutually exclusive***:

- **Mutually exclusive events:** Events that have no common outcomes.
  - These are often called ***disjoint*** as well.
  - They are denoted by $A \cap B = \varnothing$.

Some sets are more complicated and require that we have a visual display to show the variety of the sets, unions, and intersections. One tool for this is the **Venn Diagram**. Venn diagrams represent sets with overlapping or non-overlapping circles (or other shapes, circles are just the most common). They are often enclosed in a large box which encompasses all of the sets.

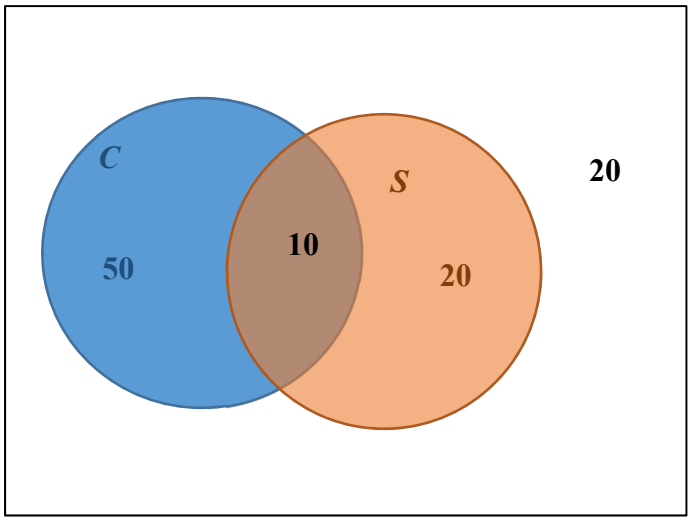
See the image on the following page for an example.



In the Venn diagram above, you can see that sets *A* and *B* overlap. The portion that overlaps is the *intersection of sets A and B*. You can see that set *C* does not overlap *A* or *B*, indicating it is mutually exclusive with those sets. The entire blue and orange region would be the *union of sets A and B*.

Example 3: In a sample of 100 college students, 60 said that they own a car, 30 said that they own a stereo, and 10 said they own both a car and a stereo. Display this information in a **Venn diagram**. Use the diagram to determine how many students own neither a car nor a stereo.

Answer:



C = owns a car

S = owns a stereo

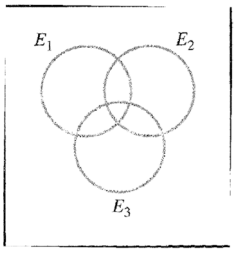
Notice that the initial information in the problem did not specify that the values given were *non-overlapping*. Notice that there are a total of 60 members in the car diagram: 50 exclusively in the car, and 10 in the overlap. S has been broken up similarly. This means that there are 20 students who have neither a car nor a stereo.

Example 4: An engineering firm is currently working on power plants at three different sites. Define the events *E1, E2, E3* as follows.

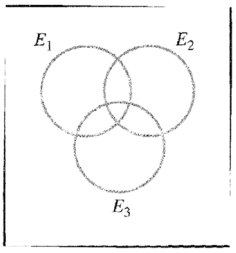
E1 = the plant at Site 1 is completed by the contract date
E2 = the plant at Site 2 is completed by the contract date
E3 = the plant at Site 3 is completed by the contract date

Write up the region using proper set notation and shade the correct region in the **Venn diagram**.

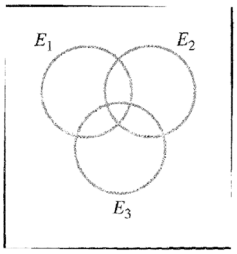
(a) All plants are complete date.



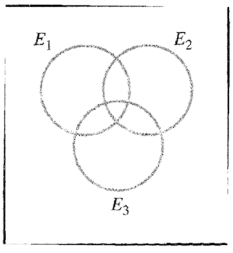
(c) At least one plant is completed by the contract by the contract date.



(b) None of the plants are completed by the contract date.



(d) Only the plant at Site 1 is completed by the contract date.



Below are the fundamentals of probability (some of which were mentioned in the previous sections, but I have included them here to have them all in one place):

Basic Properties of Probability

- For any event A, $0 \le P(A) \le 1$.
- If S is the sample space for an experiment, $P(S) = 1$.
- For any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- If two events A and B are disjoint (mutually exclusive), i.e. $P(A \cap B) = 0$, then $P(A \cup B) = P(A) + P(B)$.
- For any event A, $P(A) + P(A^C) = 1$.
  Therefore, $P(A^C) = 1 - P(A)$ and $P(A) = 1 - P(A^C)$.

Example 5: The probability of your favorite college basketball team's winning a game is 0.6. What is the probability of the team not winning the next game?

Example 6: If A and B are mutually exclusive events and $P(A) = 0.5$ and $P(B) = 0.4$, then $P(A \cup B) =$

(a) 0
(b) 0.2
(c) 0.9
(d) 0.8

Answer: Since A and B are mutually exclusive, the probabilities for the union of the sets simply add, so the answer is (c).

Example 7: A survey of local car dealers revealed that 64% of all cars sold last month had iPhone ports, 28% had alarm systems, and 22% had both iPhone ports and alarm systems. (Drawing a Venn diagram in the space provided below may help your work.)

(a) What is the probability one of these cars selected at random has neither an iPhone port nor an alarm system?

(b) What is the probability a car had an iPhone port unprotected by an alarm system?

(c) Are having an iPhone port and alarm system disjoint events?

Answers:

(a) We could draw a Venn diagram, or analyze the numeric quantities. Since 22% had both iPhone ports and alarm systems, this is the overlap. So subtract that value from each individual quantity to find out how much in each group you have without overlap.

iPhone, no alarm: 42% alarm, no iPhone: 6% Both: 22%

So the union of iPhone and alarm is 70%. Therefore, the probability that a randomly selected car has neither is 0.30 or 30%.

(b) See the work in part (a) and there is a 0.42 (42%) probability of an iPhone port with no alarm.

(c) No they are not disjoint – we were told that there was an overlap of 22% in the initial problem.

Example 7: In a survey of 120 college students living in the dorms, 60 said that they had only a TV set in their rooms, 40 said they had only a laptop in their rooms, and 15 said they had both a TV and a laptop in their rooms. The remaining 5 students had neither. If a student is randomly chosen from this group, what is

I. the probability that the student has both a TV and a laptop?
(a) 0.1250 (b) 0.2174 (c) 0.2143 (d) 0.8333

II. the probability that the student has a TV or a laptop?
(a) 0.9583 (b) 0.8333 (c) 0.8000 (d) 0.7273

III. the probability that the student does not have a TV?
(a) 0.4783 (b) 0.4583 (c) 0.5000 (d) 0.3750

IV. the probability that the student does not have a laptop?
(a) 0.6522 (b) 0.2727 (c) 0.5417 (d) 0.6667

V. the probability that the student does not have either a TV or a laptop?
(a) 0.0435 (b) 0.0417 (c) 0.8696 (d) 0.8750

Answers: Note that they gave all options as non-overlapping events.

I. (a) II. (a) III. (d) IV. (c) V. (a)

Example 8: A survey of an introductory statistics class in Fall 2011 asked students whether or not they ate breakfast the morning of the survey. Results are as follows:

		Breakfast		
		Yes	**No**	**Total**
Gender	**Male**	66	66	132
	Female	125	74	199
	Total	191	140	331

(a) What is the probability that a randomly selected student is female?

(b) What is the probability that a randomly selected student ate breakfast?

(c) What is the probability that a student is female and ate breakfast?

Answers:

(a) There are 199 female students, and there are 331 total students, so the probability is $\frac{199}{331}$, or 0.601

(b) 191 students ate breakfast, again out of 331 total students, so the probability is $\frac{191}{331}$, or 0.577.

(c) We see that 125 students were female and ate breakfast, so the probability is $\frac{125}{331}$, or 0.378.

		Breakfast		
		Yes	**No**	**Total**
Gender	**Male**	66	66	132
	Female	125	74	199
	Total	191	140	331

As you can see from the two-way table above, there are a lot of different probabilities that can be expressed. Sometimes we want to find the probability of an event given that another event has occurred. This is referred to as a **conditional probability**.

The name, hopefully, is obvious – it is the likelihood of something occurring given the *condition* that something else has also occurred.

For example, using the table above, if I randomly select someone who is male, what is the probability that he has eaten breakfast?

Since I am not looking for the proportion of people who have eaten breakfast in total – I have a condition. In this case the condition is that I am randomly selecting a *male*. So my probability is $\frac{66}{132}$ or 0.50.

- **Conditional Probability:** Let A and B be two events with $P(B) > 0$. The conditional probability of A given that B has occurred, $P(A|B)$ is given by the following formula:

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

It is important to notice the *denominator* is $P(B)$. This should imply to you that conditional probability **is not "commutative"**. That is, the probability of B given that A occurs is different than the probability of A given that B occurs.

You may have also noticed that in the initial example, I did not calculate the conditional probability using the formula, rather, I simply referenced the values on the chart – both approaches are totally valid, and in some cases one way may be easier than the other.

In addition, you could rearrange the conditional probability formula to give you the following:

For dependent events A and B, $\quad P(A \cap B) = P(A|B) \cdot P(B)$

		Breakfast		
		Yes	**No**	**Total**
Gender	**Male**	66	66	132
	Female	125	74	199
	Total	191	140	331

Example 9: Given the same table from example 8, calculate each of the probabilities:

a) The probability that someone who is female has eaten breakfast.

b) The probability that someone who has eaten breakfast is female.

Answer: It is important to note that these two statements **are not** the same thing!

a) Here, the condition is being female – we want to know how likely is it, if you are female, that you have eaten breakfast:

$$P(\text{breakfast}|\text{female}) = \frac{P(\text{breakfast} \cap \text{female})}{P(\text{female})} = \frac{125/331}{199/331} = \frac{125}{199} = 0.628$$

b) Here, the condition is eating breakfast – we want to know how likely is it, if you have eaten breakfast, that you are female:

$$P(\text{female}|\text{breakfast}) = \frac{P(\text{female} \cap \text{breakfast})}{P(\text{breakfast})} = \frac{125/331}{191/331} = \frac{125}{191} = 0.654$$

Example 10: If $P(A)=0.3$, $P(B)=0.5$, and $P(A\cap B)=0.2$, then $P(A|B)$ is

(a) 0.500 (b) 0.833 (c) 0.400 (d) 0.450 (e) 0.600

Answer: $P(A|B)=\dfrac{P(A\cap B)}{P(B)}=\dfrac{(0.2)}{(0.5)}=0.400$, so the answer is (c).

Example 11: If $P(A)=0.5$, $P(B)=0.4$, and $P(B|A)=0.3$, then is $P(A\cap B)$ is

(a) 0.750 (b) 0.600 (c) 0.120 (d) 0.150 (e) 0.375

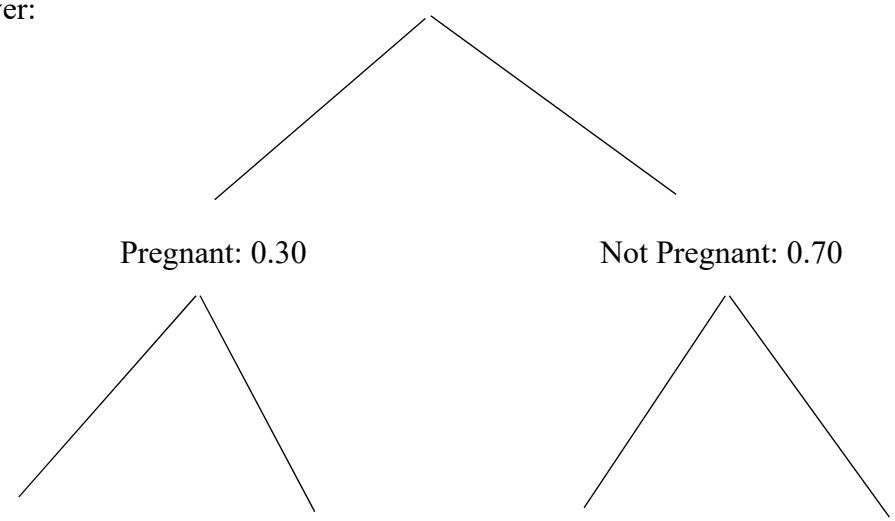
Answer: We could use the rearrange the conditional probability formula to $P(B\cap A)=P(B|A)\cdot P(A)$. We also know that $P(B\cap A)=P(A\cap B)$, because for the intersection of two sets (the overlapping portion) – the order doesn't matter.

$P(B\cap A)=P(B|A)\cdot P(A)=0.3\cdot 0.5=0.150$, so the answer is (d).

Example 12: Home pregnancy test kits have grown in popularity. Research shows that only 30% of those using a particular kit are actually pregnant. When a pregnant woman uses this kit, it correctly indicates pregnancy 96% of the time. A woman who is not pregnant gets a correct indication 90% of the time. What is the probability that a woman is pregnant given that this test gives a positive result? (Hint: draw a tree diagram)

(a) About 96%
(b) About 86%
(c) About 80%
(d) About 36%
(e) About 21%

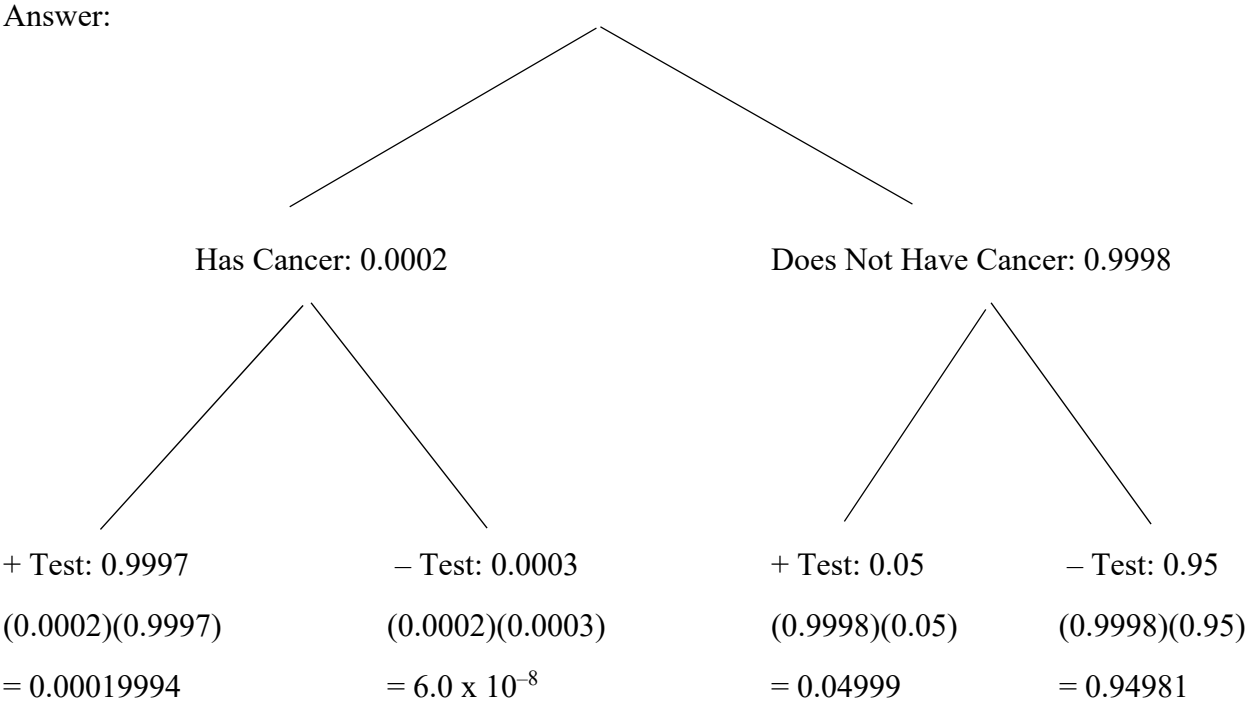
Answer:



There are two categories of positive test: 0.288 + 0.07 = 0.358 or about 36% (d).

Example 13: Ovarian cancer afflicts 1 of every 5000 women. In April 2011, Science magazine reported on a new computer based test for ovarian cancer detection that examines blood samples. The test is highly sensitive and able to correctly detect the presence of ovarian cancer in 99.97% of women who have the disease. However, it is unlikely to be used as a screening test because the test gave false positives 5% of the time. Draw a tree diagram and determine the probability that a woman who tests positive using this method actually has ovarian cancer.

Answer:



The total of all positive tests is 0.00019994 + 0.04999 = 0.05018994

Simply take the probability of having cancer and testing positive, and divide by the probability of testing positive:

$= \frac{0.00019994}{0.05018994} = 0.0039836668$ or approximately 0.398% of the time a woman tests positive, she has cancer.

Now you should be able to see why we would not want to use this as a diagnostic tool. Most (99.602% of the time) a woman who tested positive did not have cancer. This is why widespread testing regimens are very seldom implemented. Even very accurate tests can have a large amount of false positives, especially when the disease is not very widespread.

These two examples apply probability to the real world with medical applications. Tests like these have two new terms associated with them:

- **Sensitivity:** The probability of testing positive on a test given that one is actually positive for what is being tested.
- **Specificity:** The probability of testing negative given that one is actually negative for what is being tested.

Two events are said to be **independent** if the occurrence of one does not alter the probability of the other. We have two formulas that can be used to prove independence or to calculate the probabilities of independent events.

An example of independent events is flipping two separate coins. The result of the first coin toss has no effect on the outcome of the second toss.

- **Conditional Independence Formula:**
  Two events, *A* and *B*, are independent if $P(A|B) = P(A)$
- **Multiplication Rule for Two Independent Events:**
  Two events, *A* and *B*, are independent if and only if ("iff") $P(A \cap B) = P(A) \cdot P(B)$

Example 14: If *A* and *B* are independent, and $P(A) = 0.3$ and $P(B) = 0.6$, then $P(A \cap B) =$

(a) 0.72 (b) 0.90 (c) 0.18 (d) 0.50 (e) 1.10

Answer: Since the events are independent, just multiply the probabilities and we get (c) 0.18

Example 15: The probability that any one of two engines on an aircraft will fail is 0.001. Assuming that the engines operate independently of each other, then the probability that both engines will not fail is

(a) $(0.001)^2$ (b) 0.002 (c) (0.001)(0.999) (d) $(0.999)^2$ (e) $(0.001)^2(0.999)$

Answer: The chance of an engine *not* failing is 0.999, and they operate independently of one another, the probability of both not failing is (d) $(0.999)^2$.

Example 16: A medical treatment has a success rate of 0.8. Two patients will be treated with this treatment. Assuming the results are independent for the two patients, what is the probability that neither one of them will be successfully cured?

(a) 0.5 (b) 0.36 (c) 0.2 (d) 0.25 (e) 0.04

Answer: The chance of one not being cured is 0.2, so the chance of both not being cured is (0.2)(0.2) = 0.04. Therefore, the answer is (e)

Summary:

- Let A and B denote two events.

  1. The event ***not A*** consists of all experimental outcomes that are not in event A. Not A is sometimes called the **complement of *A*** and is usually denoted by A^C, A', or rarely $\bar{A}$.
  2. **Union of two events:** The event ***A or B*** consists of *all experimental outcomes* that are in *at least* one of the two events, that is, in A or in B or in both of these. It is also denoted by $A \cup B$.
  3. **Intersection of two events:** The event ***A and B*** consists of *all experimental outcomes* that are *in both* the event A and the event B. It is also denoted by $A \cap B$.

- **Mutually exclusive events:** Events that have no common outcomes.
  - These are often called ***disjoint*** as well.
  - They are denoted by $A \cap B = \varnothing$.
- **Conditional Probability:** Let A and B be two events with $P(B) > 0$. The conditional probability of A given that B has occurred, $P(A|B)$ is given by the following formula:

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

- **Sensitivity:** The probability of testing positive on a test given that one is actually positive for what is being tested.
- **Specificity:** The probability of testing negative given that one is actually negative for what is being tested.
- **Independent Events:** The occurrence of one event does not alter the probability of the occurrence of a second event.
  - **Conditional Independence Formula:**
    Two events, A and B, are independent if $P(A|B) = P(A)$
  - **Multiplication Rule for Two Independent Events:**
    Two events, A and B, are independent if and only if ("iff")
    $P(A \cap B) = P(A) \cdot P(B)$

Basic Properties of Probability

- For any event A, $0 \le P(A) \le 1$.
- If S is the sample space for an experiment, $P(S) = 1$.
- For any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- If two events A and B are disjoint (mutually exclusive), i.e. $P(A \cap B) = 0$, then

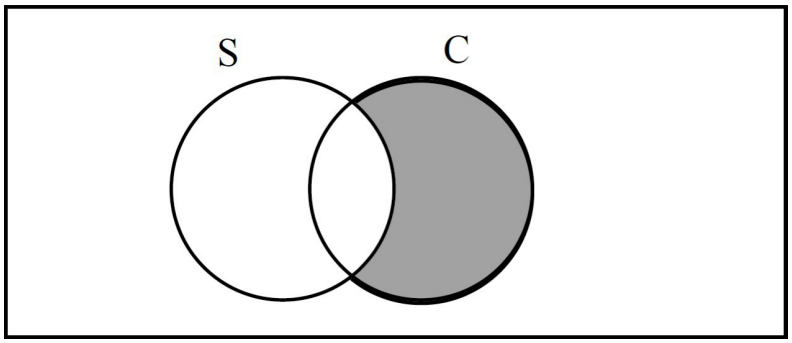
  $P(A \cup B) = P(A) + P(B)$.
- For any event A, $P(A) + P(A^C) = 1$.

  Therefore, $P(A^C) = 1 - P(A)$ and $P(A) = 1 - P(A^C)$.

Checkpoint 4.2

Multiple Choice

1. The following Venn diagram uses the following definitions: C = students taking calculus, S = students taking statistics. What does the shaded region represent?

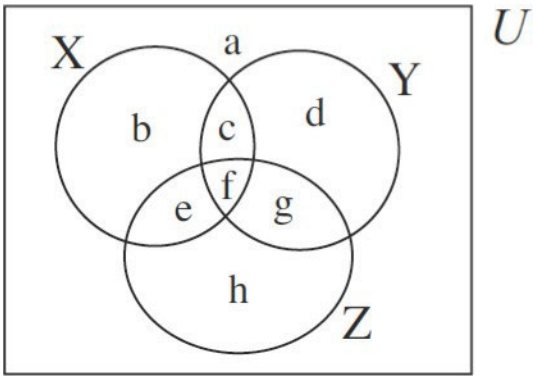


(a) Students taking Calculus
(b) Students not taking Calculus or Statistics
(c) Students taking Calculus but not Statistics
(d) Students taking both Calculus and Statistics
(e) Students taking only one of Calculus or Statistics

2. If $U = \{1,2,3,...,10\}$ and $S = \{4,5,6,7,8\}$, then $S' =$
Note: $U =$ 'universe' and $S =$ 'subset of universe'

(a) {9,10} (b) {1,2,3} (c) {1,2,3,9} (d) {1,2,3,9,10} (e) {1,2,3, … ,10}

3. In the figure below, $U = \{a, b, c, d, e, f, g, h\}$, find $(X \cup Y) \cap Z$



(a) {c, e, f, g, h} (b) {e, f, g} (c) {f} (d) {c, f, h} (e) {c, e, g}

4. At Kennett High School, 5% of athletes play both football and some other contact sport, 30% play football, and 40% play other contact sports. If there are 200 athletes, how many play neither football nor any other contact sport?
   (a) 20
   (b) 70
   (c) 80
   (d) 100
   (e) 130

5. If $P(A) = 0.5$, $P(B) = 0.6$, and $P(A \cap B) = 0.3$, then is $P(A \cup B)$ is
   (a) 0.800 (b) 0.500 (c) 0.600 (d) 0 (e) 0.375

6. If $P(A) = 0.6$, $P(B) = 0.3$, and $P(B|A) = 0.4$, then is $P(A \cup B)$ is
   (a) 0.680 (b) 0.120 (c) 0.667 (d) 0.220 (e) 0.780

7. If $P(A) = 0.6$, $P(B) = 0.3$, and $P(B|A) = 0.4$, then is $P(A^C)$ is
   (a) 0.400 (b) 0.100 (c) 0.600 (d) 0.760 (e) 0.412

Questions 8 and 9 refer to the following table and information:

Five hundred people used a home test for HIV, and then all underwent more conclusive hospital testing. The accuracy of the home test was evidenced in the following table:

	HIV	Healthy
Positive Test	35	25
Negative Test	5	435

8. What is the *sensitivity* of the test? That is, what is the probability of testing positive given that the person has HIV?
   (a) 0.070 (b) 0.130 (c) 0.538 (d) 0.583 (e) 0.875

9. What is the specificity of the test? That is, what is the probability of testing negative given that the person does not have HIV?
   (a) 0.125 (b) 0.583 (c) 0.879 (d) 0.946 (e) 0.950

10. You flip four coins. What is the probability of getting exactly four heads?

(a) 1/16
(b) 4/16
(c) 6/16
(d) 2/16
(e) 5/16

11. If $P(B) = 0.4$, and $P(A \cap B) = 0.21$, if A and B are independent, then $P(A)$ is

(a) 0.084
(b) 0.475
(c) 0.525
(d) 0.600
(e) Not enough information is given to find the value of $P(A)$

12. You have two specially created dice. Each die has six sides. The first die has the numbers $\{1,3,5,7,9,11\}$ on the sides. The second die has the numbers $\{2,4,8,10,11,15\}$ on the sides. Which of the following would **not** represent an instance of independent events?

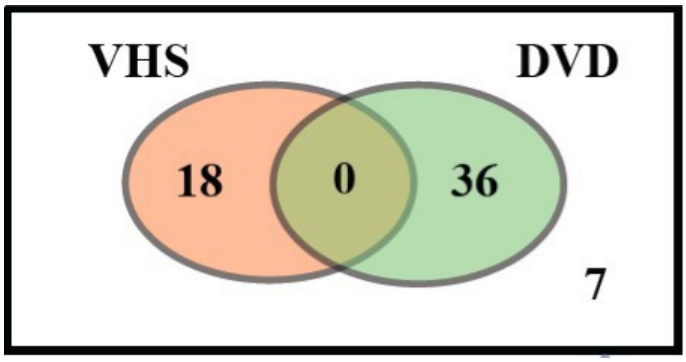
(a) Getting an odd number on the first die AND getting an odd number on the second die.
(b) Getting a 9 on the first die AND getting a sum of 19 on the two dice.
(c) Getting a 2 on the first die AND getting a value greater than 10 on the second die.
(d) Getting an even number on the first die AND getting a 15 on the second die.
(e) All of the above are examples of independent events.

13. Two events, A and B, have the following probabilities: $P(A) = 0.54$, $P(B) = 0.2$, and $P(A \cap B) = 0.108$. Which of the following can be concluded from these probabilities?

(a) A and B are mutually exclusive events.
(b) A and B are independent events.
(c) A and B are dependent events.
(d) A and B are complementary events.
(e) A and B are independent and mutually exclusive events.

Free Response:

1. Consider the following diagram of an experiment concerning people who own a VHS player, a DVD player, both, or neither.



   a. Are these events mutually exclusive? Explain.

   b. How many people are in the sample space?

   c. How many people own both a VHS and a DVD player?

   d. How many people own neither?

2. A survey was taken among a group of people. The probability that a person chosen likes Italian food (event A) is 0.79, the probability that a person likes Chinese food (event B) is 0.48, and the probability of that a person likes both foods is 0.31.

(a) Draw a Venn diagram that shows the relation between the events A and B.

Use your Venn diagram to find the probability

(b) a person likes Italian but not Chinese

(c) a person likes at least one of these foods

(d) a person likes at most one of these foods

(e) If a person likes Italian, he also likes Chinese

(f) If a person likes Chinese, he also likes Italian

3. On average, suppose a baseball player hits a home run once in every 10 times at bat, and suppose he gets exactly two "at bats" in every game. Using simulation with the following randomly generated numbers, estimate the likelihood that the player will hit 2 home runs in a single game.

42 99 02 65 04 14 30 09 70 88 89 85 95 40 53 67 25 50 48 79 86 92 76
24 53 39 08 73 78 17 72 81 08 01 68 94 43 43 95 12 36 90 28 88 34 69
18 69 91 79 14 82 26 94 15 26 19 41 74 02 17 20 38 84 74 30 34 96 09
46 61 41 02 93 94 90 00 71 84 98 30 82 80 11 92 97 81 29 85 44 40 05
83 22 04 86 13 33 00 99 74 75 27 43 68 22 59 20 66 00 24 01 96 84 19
14 57 26 47 58 51 73 06 08 49 52 70 15 79 35 65 28 40 77 93 73 33 24
25 22 32 03 89 03 62 13 85 16 23 28 12 61 16 75 45 37 15 54 36 18 45
64 31 31 06 80 32 75 99 27 91 25 98 05 55 32 27 16 51 45 89 31 78 90
82 05 11 39 80 83 01 20 10 67 97 33 72 09 98 78 39 56 57 54 63 35 21
35 93 18 17 48 55 60 44 92 21 07 77 42 46 86 41 49 76 96 36 62 38 11
64 07 04 58 23 56 29 37 87 37 59 47 83 77 21 63 10 95 87 10 42 71 12
88 06 52 42 99 02 65 04 14 30 09 70 88 89 85 95 40 53 67 25 50 48 79
86 92 76 24 53 39 08 73 78 17 72 81 08 01 68 94 43 43 95 12 36 90 28
88 34 69 18 69 91 79 14 82 26 94 15 26 19 41 74 02 17 20 38 84 74 30
34 96 09 46 61 41 02 93 94 90 00 71 84 98 30 82 80 11 92 97 81 29 85
44 40 05 83 22 04 86 13 33 00 99 74 75 27 43 68 22 59 20 66 00 24 01
96 84 19 14 57 26 47 58 51 73 06 08 49 52 70 15 79 35 65 28 40 77 93
73 33 24 25 22 32 03 89 03 62 13 85 16 23 28 12 61 16 75 45 37 15 54
36 18 45 64 31 31 06 80 32 75 99 27 91 25 98 05 55 32 27 16 51 45 89
31 78 90 82 05 11 39 80 83 01 20 10 67 97 33 72 09 98 78 39 56 57 54
63 35 21 35 93 18 17 48 55 60 44 92 21 07 77 42 46 86 41 49 76 96 36
62 38 11 64 07 04 58 23 56 29 37 87 37 59 47 83 77

4.2 Homework

1) Ed Wine is a student who has the usual student complaint: "Usually I do my homework and the teacher never checks it and when I don't do my homework, the teacher usually checks it."

   Here are the statistics: The teacher checks homework 38% of the time. When she doesn't check homework, Ed has done his homework 84% of the time. When she does check homework, Ed has done it 52% of the time.

   Make a tree diagram that illustrates these statistics and find the following probabilities:
   a) If Ed does his homework, the teacher doesn't check it.
   b) If Ed doesn't do his homework, the teacher does check it.
   c) Is Ed justified in his complaint based on what you just found? Why?

2) The following two-way table shows the breakdown of SI students in AP Calculus (AB or BC) and AP Statistics in 2022 – 2023 school year*. Use that information to calculate each of the following:

	Male	Female	Total
AP Calculus	95	130	225
AP Statistics	72	69	141
Total	167	199	366

   a) What is the probability that a student is taking AP Calculus given that she is female?
   b) What is the probability that a student who is taking AP Calculus is female?
   c) What is the probability that a student is taking AP Calculus given that he is male?
   d) What is the probability that a student who is taking AP Calculus is male?
   e) What is the probability that a student who is male is taking AP Statistics?
   f) What is the probability that a student who is female is taking AP Statistics?

*Note that this data is not strictly accurate because some students are "double-counted" – that is, they are in both AP Calculus and AP Statistics. However, for illustrative purposes of the idea of a two-way table, it is a useful example without overly complicating.

3) An engineering firm is constructing three power plants at three distinct locations. Define events A, B, and C as follows:

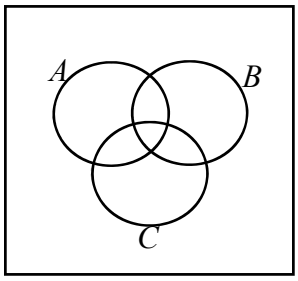
A = the plant at the first site is finished on time

B = the plant at the second site is finished on time

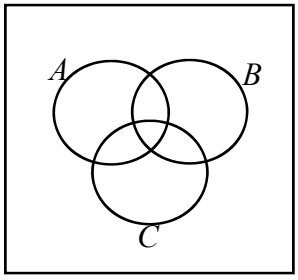
C = the plant at the third site is finished on time

Shade the region in the Venn diagram that corresponds to each of the following events.

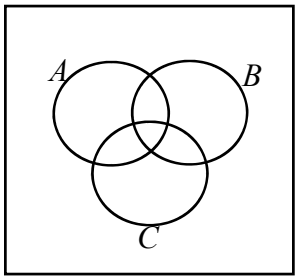
a) At least one plant is finished on time.



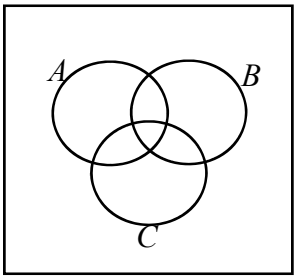
b) All three plants are finished on time.



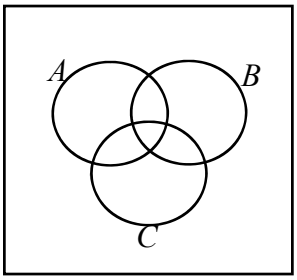
c) Both A and B are finished on time, but C is not.



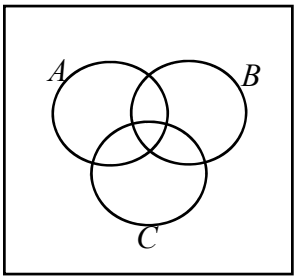
d) No plants are finished on time.



e) Only plant C is finished on time.



f) Either A or C is finished on time, but B is not.

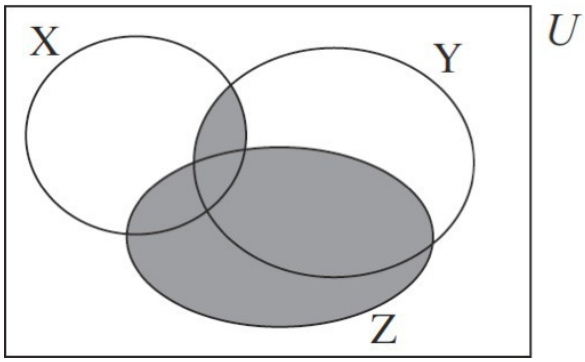


4) A certain test for opioids has a likelihood of correctly identifying a person with opioids in their system of 98.1%. However, there is a false positive rate of 7% (that is, 7 percent of people with no opioids in their system will still test positive for opioids). If a business that is concerned with opioid use on the job has 10,000 employees, and if we can assume that 2.5% of them are using opioids, find each of the following. If all employees are tested using the test, find each of the following:

   a) How many of the employees test positive who are actually using opioids?
   b) How many of the employees test positive who are not using opioids?
   c) For this group, what is the chance of testing positive given that you are not using opioids?
   d) For this group, what is the chance of testing negative given that you are using opioids?
   e) Given that opioid use on the job is grounds for firing an employee, would you advise this business to do widespread testing with this test or not? Explain.

5) In the figure to the right, the shaded portion represents which of the following:

   (a) $(X \cap Z) \cup Y$
   (b) $(X \cap Y) \cup Z$
   (c) $(X \cup Y) \cap Z$
   (d) $(Y \cap Z) \cup X$
   (e) $(Y \cup Z) \cap X$



Problems 6 to 8 deal with the following table and information:

A study has been done to determine whether or not a certain drug leads to an improvement in symptoms for patients for a particular medical condition. The results are shown in the following table:

	Improvement	No Improvement	Total
Drug	270	530	800
No Drug	120	480	600
Total	390	1010	1400

6) Based on this table, what is the probability that a patient shows improvement if it is known that the patient was given the drug?

(a) 0.3250
(b) 0.3375
(c) 0.2250
(d) 0.4355
(e) None of the above

7) Based on this table, what is the probability that a patient shows improvement if it is known that they were not given the drug?

(a) 0.200
(b) 0.250
(c) 0.279
(d) 0.501
(e) None of the above

8) Based on this table, what is the probability that a patient shows improvement if they participated in this study?

(a) 0.200
(b) 0.250
(c) 0.279
(d) 0.501
(e) None of the above

4.3 Random Variables and Probability Distributions

Objectives:

- Represent the probability distribution for a discrete random variable.
- Calculate discrete probabilities from a table.
- Calculate expected values.
- Interpret a probability distribution.

Remember that a *discrete variable* was a variable that took on specific (often, but not always, integer) values. That is, the values were not infinitely subdivisible – for example, the number of students in a class can only be a whole number. You cannot have 27.325 students in a class, you could have 27 students or 2G8 students. Another example would be the price of a pair of shoes, which could have decimal values ($159.99 for a pair of running shoes, for example) but are still not infinitely subdivisible – you won't have a pair of shoes cost $145.34721.

We will be looking at *probabilities* and *probability distributions* for discrete random variables in this section.

- **Probability Distribution of a Discrete Random Variable (*x*):** Gives the probability associated with each *x*-value.
    - Common ways to display probability distributions for discrete random variables are tables, a probability histogram, or a formula.

- **Properties of Discrete Probability Distributions:**
    - For every possible *x*-value, the probability of *x* occurring is between 0 and 1, inclusive: $0 \le P(x) \le 1$
    - The sum of the probabilities of all *x*-values is 1: $\sum P(x) = 1$

- **Cumulative Probability Distribution:** A probability distribution that shows the probability of being *less than or equal to* each value of the random variable (*x*).

It is important to remember, probabilities can never be negative and can never be greater than 1. It is fairly common for the AP test to include questions in the multiple choice section that focus on this seemingly basic concept, so do not throw away easy points on the test because you don't recall the basics!

Example 1: Which of the following is a valid discrete probability distribution?

(A)

x	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1
$P(x)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

(B)

x	−2	1	2	4
$P(x)$	0.2	0.6	0.2	0.1

(C)

x	1	2	3
$P(x)$	0.3	0.2	0.1

(D)

x	1	2	3	4
$P(x)$	0.1	0.2	0.3	−0.1

(E)

x	−2	−1	1	2
$P(x)$	−0.3	−0.2	0.2	0.3

Answer: (A) is the only valid response, all other responses have either negative probabilities or have probabilities that sum to something other than 1.

Example 2: Let's say a family decides to have two children. Let X represent the number of girls in the family. Assuming that there is an equal probability of having a boy or a girl, what is a probability distribution for X?

Answer: You could do a tree diagram or other method to find the probabilities for each option (0, 1, or 2 girls). Keep in mind that having a boy and then a girl represents a distinct and different result than having a girl and then a boy:

X	0	1	2
$P(X)$	0.25	0.50	0.25

Example 3: A consumer organization that evaluates new automobiles customarily reports the number of major defects on each car examined. Let x denote the number of major defects on a randomly selected car of a certain type. A large number of automobiles were evaluated, and a probability distribution consistent with these observations is given below:

x	0	1	2	3	4	5	6	7	8	9	10
$P(x)$	0.041	0.130	0.209	0.223	0.178	0.114	0.061	0.028	0.011	0.004	0.001

i. Calculate $P(2 \le x \le 5)$.

ii. Calculate $P(2 \le x < 5)$.

iii. Calculate $P(2 < x \le 5)$.

iv. Calculate $P(2 < x < 5)$.

Answers: All you have to do is find the values on the distribution and add them up to get the total probability. Pay attention to whether or not each inequality is "or equal to" or not.

a) $P(2 \le x \le 5) = 0.624$

b) $P(2 \le x < 5) = 0.510$

c) $P(2 < x \le 5) = 0.415$

d) $P(2 < x < 5) = 0.301$

Example 4: A statistics professor has established a probability distribution for grades a student can earn in that course, based on his past 20 years of teaching this course. The following table gives the distribution, where x is the grade and P(x) is the corresponding probability for the grade:

Grade	**F**	**D**	**C**	**B**	**A**
x	0	1	2	3	4
P(*x*)	0.12	0.15	0.35	0.30	0.08

I. What is the probability that the student will earn a grade of A for this course?

(a) 0.20 (b) 0.08 (c) 0.92 (d) 0.12

II. What is the probability that the student will earn a grade of C or better for this course?

(a) 0.38 (b) 0.62 (c) 0.73 (d) 0.35

III. What is the probability that the student will earn a grade better than a C for this course?

(a) 0.38 (b) 0.62 (c) 0.73 (d) 0.35

Answers: I. (b) 0.08 II. (c) 0.73 III. (a) 0.38

Example 5: Create a cumulative probability distribution for the table from Example 4:

Grade	**F**	**D**	**C**	**B**	**A**
x	0	1	2	3	4
***P*(*x*)**	0.12	0.15	0.35	0.30	0.08

Answer: Remember, the cumulative distribution should show the probability of getting less than the associated *x*-value.

Grade	F	D	C	B	A
x	0	1	2	3	4
$\sum P(x)$	0.12	0.27	0.62	0.92	1.00

There are some specific things we need to know about the random variable *x*:

- **Mean Value of a Random Variable, *X*:** This is denoted by μ_X, and it describes where the probability distribution of *x* is centered.
- **Expected Value:** This is denoted by $E(X)$, and it is sometimes used in place of the mean value, μ_X.
- **Standard Deviation:** This is denoted by σ_X, and is a measure of the spread of the data about the mean.
    - **Variance:** Denoted by σ_X^2 is simply the standard deviation, squared.

The formulas are provided on the AP Formula Sheet, and are listed here as well:

$$\mu_X = E(X) = \sum x_i P(x_i) \qquad \sigma_X = \sqrt{\sum (x_i - \mu_X)^2 P(x_i)}$$

Where x_i represents and individual value of the random variable, *X*, and $P(x_i)$ is the probability associated with that individual *x*-value, x_i.

Example 6: Individuals applying for a certain license are allowed up to four attempts to pass the exam. Let X denote the number of attempts made by a randomly selected applicant. The probability distribution of X is given below:

X	1	2	3	4
$P(X)$	0.10	0.20	0.30	0.40

Find μ_X and σ_X.

Answers:

$$\mu_X = 1(0.10) + 2(0.20) + 3(0.30) + 4(0.40) = 3.0$$

$$\sigma_X = \sqrt{(1-3.0)^2(0.10) + (2-3.0)^2(0.2) + (3-3.0)^2(0.30) + (4-3.0)^2(0.40)} = 1.0$$

Example 7: The number of T-shirts a store sells daily has the following probability distribution: If each T-shirt sells for \$10 but costs the store \$4 to purchase, what is the expected daily T-shirt profit?

# of T-shirts, X	0	1	2	3	4	5	6	7	8	9	10
$P(X)$	0.02	0.15	0.18	0.21	0.14	0.08	0.08	0.04	0.03	0.02	0.05

(a) \$3.78 (b) \$15.12 (c) \$22.68 (d) \$30.00 (e) \$37.80

Since the store profits $6 per shirt, so you have two options:

Calculate the expected value for the probability distribution (same as the mean value) and then multiply by 6.

OR

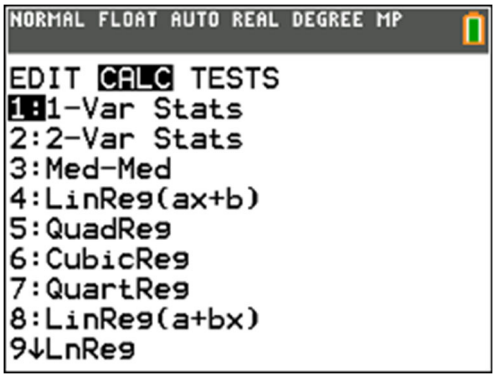
Multiply each number of T-shirts by 6 and then calculate the expected value using these new values for your x_i. (Note this is actually just the same as the first process, but with the 6 distributed.)

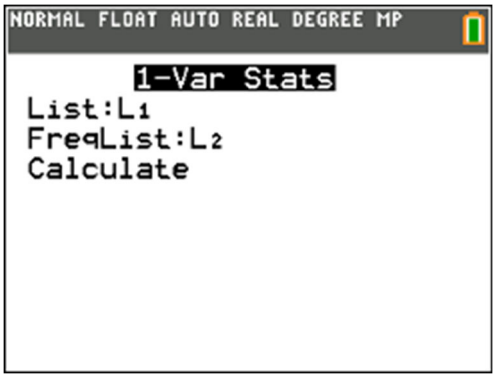
Either way you end up with 6 x 3.78 = $22.68 or (c).

Example 8: Calculate the standard deviation for the probability distribution in Example 7.

Answer: Note that the mean value for the distribution (not the profit) was 3.78, so calculate the standard deviation by hand to get 2.47

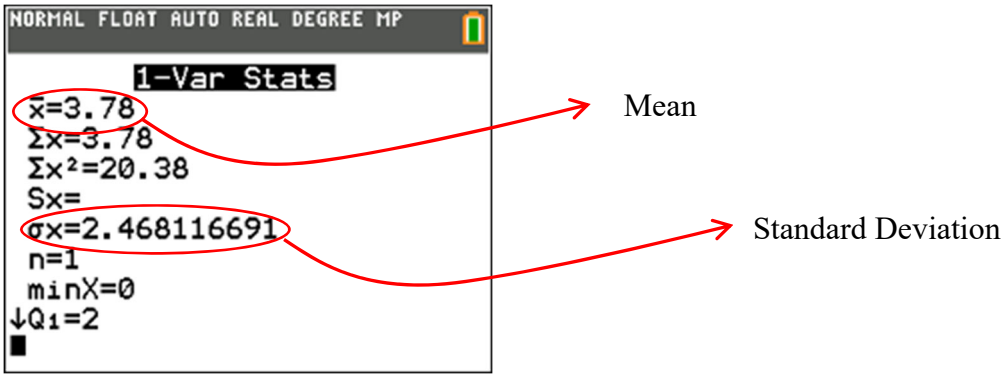
You could also accomplish this on your calculator in a couple of ways. Enter the values of X as L_1 and the probabilities as L_2. Then hit your **stat** button and go to 1–Var Stat. We will now use the FreqList that we left blank before.





Essentially, you put your list with of your X variable in the space for List, and then your list for your probabilities in FreqList.

Below is the screen you get after running these statistics:



Example 9: Suppose you are one of 7.5 million people who send in their name for a drawing with one top prize of $1 million, five second place prizes of $10,000, and 20 third place prizes of $100. Is it worth the $0.32 postage it cost you to send in your name?

(a) Yes, because $\dfrac{1{,}000{,}000}{0.32} = 3{,}125{,}000$ which is less than 7,500,000

(b) No, because your expected winnings are only $0.14

(c) Yes, because $\dfrac{7{,}500{,}000}{(1+5+20)} = 288{,}462$

(d) No, because 1,052,000 < 7,500,000

(e) Yes, because $\dfrac{1{,}052{,}000}{(1+5+20)} = 40{,}462$

Answer: If you calculate the expected value, you get the following:

$$\frac{1}{7{,}500{,}000}(\$1{,}000{,}000) + \frac{5}{7{,}500{,}000}(\$10{,}000) + \frac{20}{7{,}500{,}000}(\$100) = \$0.14$$

Since the expected value of the prize (14 cents) is less than the cost of the stamp (32 cents), it is not worth it.

Summary:

- **Probability Distribution of a Discrete Random Variable (*x*):** Gives the probability associated with each *x*-value.
  - Common ways to display probability distributions for discrete random variables are tables, a probability histogram, or a formula.
- **Properties of Discrete Probability Distributions:**
  - For every possible *x*-value, the probability of *x* occurring is between 0 and 1, inclusive: $0 \le P(x) \le 1$
  - The sum of the probabilities of all *x*-values is 1: $\sum P(x) = 1$
- **Mean Value of a Random Variable, *X*:** This is denoted by μ_X, and it describes where the probability distribution of *x* is centered.
- **Expected Value:** This is denoted by $E(X)$, and it is sometimes used in place of the mean value, μ_X.
- **Standard Deviation:** This is denoted by σ_X, and is a measure of the spread of the data about the mean.
  - **Variance:** Denoted by $\sigma_X{}^2$ is simply the standard deviation, squared.

$$\mu_X = E(X) = \sum x_i P(x_i) \qquad \sigma_X = \sqrt{\sum (x_i - \mu_X)^2 P(x_i)}$$

Checkpoint 4.3

Free Response

1. Of all airline flight requests received by a certain discount ticket broker, 70% are for domestic travel (D) and 30% are for international flights (I). Let X be the number of requests among the next three requests received that are for domestic flights. Assuming independence of successive requests, determine the probability distribution of X. (*Hint*: One possible outcome is D-I-D, with the probability of (0.7)(0.3)(0.7) = 0.147). What is the probability that there are fewer than 2 requests for domestic flights?

Multiple Choice

1. Companies proved to have violated pollution laws are being fined various amounts with the following probabilities:

Fine ($):	1000	10,000	50,000	100,000
Probability:	0.4	0.3	0.2	0.1

What are the mean and standard deviation for the fine variable?

(a) $\mu_x = 40,250, \sigma_x = 39,118$
(b) $\mu_x = 40,250, \sigma_x = 45,169$
(c) $\mu_x = 23,400, \sigma_x = 31,350$
(d) $\mu_x = 23,400, \sigma_x = 45,169$
(e) $\mu_x = 23,400, \sigma_x = 85,185$

2. At a warehouse sale 100 customers are invited to choose one of 100 identical boxes. Five boxes contain $700 color television sets, 25 boxes contain $540 camcorders, and the remaining boxes contain $260 cameras. What should a customer be willing to pay to participate in the sale?

(a) $260
(b) $352
(c) $500
(d) $540
(e) $699

4.3 Homework

1. Define a random variable x to be the number of courses for which a randomly selected university student at a certain university is registered in a given semester. Suppose that the probability distribution of x is given in the following table:

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

   a. What is $P(x=4)$?
   b. What is $P(x\le 4)$?
   c. What is $P(x\ge 4)$?
   d. What is the probability that the selected student is taking at most 5 courses?
   e. What is the probability that the selected student is taking at least 5 courses?
   f. What is the probability that the selected student is taking more than 5 courses?

2. Suppose a manufacturer of video game systems receives processors in lots of five. Two boards are selected at random from each lot for inspection. We can represent the possible outcomes of the selection process by numbered pairs; for example, (1,2) represents the outcome where Processors 1 and 2 are selected for inspection.
   a. List the ten different possible outcomes for this situation.
   b. Suppose that processors 3 and 5 are defective in a given lot of five. Two boards are chosen at random from the lot. Define x as the number of defective processors observed among the those that were inspected. Find the probability distribution of x.

3. In a given carton of a dozen eggs at a certain supermarket, there is a probability that some eggs are broken. Let y denote the number of eggs broken in a randomly selected carton. The probability distribution of y is given below:

y	0	1	2	3	4
$p(y)$	0.65	0.20	0.10	0.04	0.01

   a. Calculate μ_y and σ_y.
   b. In the long run, for what percentage of cartons is the number of broken eggs less than μ_y? Does this seem unusual to you?
   c. Why doesn't $\mu_y = (1+2+3+4+5)/5 = 2.0$? Explain.
   d. Let z = the number of unbroken eggs in a carton. Write a linear function of z in terms of y, and use that to find the mean value of z.

4. The probability distribution for the number of courses a university student is taking in a semester is given below (note that this is the same probability distribution as in problem 1):

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

   a. Use your calculator to find the mean and standard deviation for this distribution.
   b. Given that you should have found $\mu_x = 4.66$ and $\sigma_x = 1.20$, what is the probability that x is more than one standard deviation **below** the mean? more than one standard deviation **above** the mean?
   c. What x values are more than 2 standard deviations **away from** the mean? What is the probability that x is more than 2 standard deviations away from the mean?

4.4 Combining Random Variables

Objectives:

- Calculate the mean and standard deviation of a transformed discrete random variable.
- Calculate the mean and standard deviation of a linear combination of a discrete random variable.

Oftentimes, there will be situations where we change a discrete random variable via some kind of transformation, either multiplying by a factor, or by multiplying by a factor and adding a number. You may notice that multiplying then adding is the same as creating a linear function (multiplying for the slope, adding the intercept). Another thing we could do is add two random variables to create a new probability distribution of the combined variables.

When we transform a variable, we do not necessarily want to go through the process in the last section of writing out a complete probability distribution and calculating the mean, variance, and standard deviation, so we have several formulas to accomplish this.

Important Note: These formulas are not provided on the AP Test!

You must memorize them – I have highlighted in yellow the formulas that you must memorize.

If you have two random variables, X and Y, and real numbers a and b, the mean of $aX + bY$ is given by the following formula:

- $\mu_{aX+bY} = a\mu_x + b\mu_y$

  Note that this simply means that the means will add up. If you multiply by a linear factor, then multiply the means and add them.

Example 1: An AP statistics teacher wants to change the weighting of sections of his final exam – he wants to make the multiple choice worth 3 times as much and the free response worth 2/3 as much, and then he will add together the two sections for the exam score. If the mean of the multiple choice is 14 and the mean of the free response is 48, what is the mean of the total test score.

Answer:

Let T = the new test score, X = multiple choice score, and Y = free response score.

We know that $T = aX + bY$, with $a = 3$, and $b = 2/3$; $\mu_x = 14$ and $\mu_y = 48$

$$\mu_T = a\mu_x + b\mu_y$$

$$\mu_T = 3(14) + \frac{2}{3}(48) \qquad \mu_T = 78$$

Standard deviation and variance rely on the fact that two variables are *independent* of each other. This is very similar to the concept of probabilities being independent:

- **Independent Random Variables:** When knowing information about one variable does not change the probability distribution of the other variable.

For independent random variables, X and Y, and real numbers a and b, the variance and standard deviation are given as follows:

- **Variance:** $\sigma^2_{aX+bY} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}$
- **Standard Deviation:** $\sigma_{aX+bY} = \sqrt{a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}}$

Note that this does repeat iteratively if you have more variables you are adding, just continue adding terms. For example, if you needed the variance for X, Y, and Z, you would just add a term to the formula: $\sigma^2_{aX+bY+cZ} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2} + c^2\sigma_z^{\,2}$

Example 2: For the same final exam, the standard deviation of the multiple choice is 1.2, and the standard deviation of the free response is 6.8. Find the variance and standard deviation of the transformed exam score that uses the same transformation as Example 1 ($T = 3X + \frac{2}{3}Y$).

Let T = the new test score, X = multiple choice score, and Y = free response score.

We know that $T = aX + bY$, with $a = 3$, and $b = 2/3$; $\mu_x = 14$ and $\mu_y = 48$

Variance:

$$\sigma^2_{aX+bY} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}$$

$$\sigma_T^2 = (3)^2(1.2)^2 + \left(\frac{2}{3}\right)^2(6.8)^2$$

$$\sigma_T^2 = 33.511$$

Standard Deviation:

$$\sigma_{aX+bY} = \sqrt{a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}}$$

$$\sigma_T = \sqrt{(3)^2(1.2)^2 + \left(\frac{2}{3}\right)^2(6.8)^2}$$

$$\sigma_T = 5.789$$

Example 3: A young woman works two jobs and receives tips for both jobs. As a hair dresser, her distribution of weekly tips has mean \$65 and standard deviation \$5.75. As a waitress, her distribution of weekly tips has mean \$154 and standard deviation \$8.02. What are the mean and standard deviation of her combined weekly tips? (Assume independence for the two jobs.)

(a) Mean \$167.16; Standard deviation \$9.87
(b) Mean \$167.16; Standard deviation \$13.77
(c) Mean \$219.00; Standard deviation \$2.27
(d) Mean \$219.00; Standard deviation \$9.87
(e) Mean \$219.00; Standard deviation \$13.77

The answer would be (d), since there are no factors to multiply by the means simply add together, and the standard deviation is the square root of the sum of the squared standard deviations.

Example 4: A nationwide standardized exam consists of a multiple-choice section and a free-response section. The means and standard deviations are reported in the following table:

	Mean	**Standard Deviation**
Multiple Choice	38	6
Free Response	30	7

Let's define x_1 and x_2 as the multiple choice score and the free response score, respectively, of a student selected at random from the students taking this exam. We are also interested in the variable y = total score, where $y = x_1 + 2x_2$. Find the mean and standard deviation of y. Assume that the scores are independent of one another.

Answer: $\mu_y = 1(38) + 2(30)$ $\sigma_y = \sqrt{1^2(6)^2 + 2^2(7)^2}$

$\mu_y = 98$ $\sigma_y = 15.232$

As we mentioned at the beginning, we can also do a *linear transformation of a random variable.* This happens, as we said previously, when we multiply our variable, x by some coefficient, and add a constant value to the result. The mean, variance, and standard deviation are all affected by this transformation.

- **Linear Function of the Random Variable, *x*:** If x is a random variable, and a and b are constants, the random variable y is a linear function of x if $y = a + bx$.
    - **Note that the formulas are given with *a* = *y*-intercept and *b* = slope.** If you use the formula $y = ax + b$ instead, the values of a and b are switched in the formulas.

Mean of y (for $y = a + bx$) is

$$\mu_y = \mu_{a+bx} = a + b\mu_x$$

Variance of y (for $y = a + bx$) is

$$\sigma_y^2 = \sigma_{a+bx}^2 = b^2\sigma_x^2$$

It is worth noting a couple of things about these two formulas:

- To find the mean, simply plug the mean of your x variable into the linear transformation equation for the x.
- The constant value added in a linear transformation <u>*has no effect on variance*</u> (or standard deviation for that matter.
- Standard deviation is just the square root of the variance, so standard deviation is

  $$\sigma_y = \sigma_{a+bx} = |b|\sigma_x$$

  The absolute value is because standard deviation must be positive.

Example 5: Consider the experiment in which a customer of a propane gas company is randomly selected. The company serves homes heated by propane gas, and customers call when they wish to have their propane tank filled. Suppose that the mean and standard deviation of the random variable x = number of gallons required to fill the tank are known to be 318 gal and 42 gal, respectively. The company is considering two different pricing models:

Model 1: \$2 per gal
Model 2: service charge of \$50 + \$1.80 per gal

The company is interested in the variable y = amount billed. For each of the two models, y can be expressed as a function of the random variable x:

Model 1: $y_{\text{model1}} = 2x$

Model 2: $y_{\text{model2}} = 50 + 1.8x$

Find the mean and standard deviation for both models.

Answer: $\mu_{y_{\text{model 1}}} = \$636,\ \sigma_{y_{\text{model 1}}} = \$84$

$\mu_{y_{\text{model 1}}} = \$622.40,\ \sigma_{y_{\text{model 1}}} = \$75.60$

Example 6: For a given school year, a reporter has been told that the average teacher's salary was $59,500 with a standard deviation of $17,200. The reporter also knows that teachers will be receiving raises of 3.25% for the next school year. What would the reporter write for the new average teacher's salary and standard deviation?

(a) Mean $1934; Standard deviation $559
(b) Mean $59,500; Standard deviation $17,200
(c) Mean $59,500; Standard deviation $17,759
(d) Mean $61,434; Standard deviation $17,200
(e) Mean $61,434; Standard deviation $17,759

Answer: Since there is no a value added to the salaries, and we know that $b = 1.0325$ (3.25% increase above the base pay rate). Therefore, the answer would be (e).

Example 7: A random variable has a standard deviation of 1.3. A new variable is created by transforming the values of the random variable using the following rule: Multiply each value by 2 and then add 5. What is the value of the standard deviation for this transformed variable?

(a) 1.3
(b) 2.6
(c) 6.3
(d) 7.6
(e) 8.5

Answer: Since the standard deviation is unaffected by $a = 5$, the new standard deviation is simply the old one multiplied by 2. Thus, the answer is (b).

Example 8: If a certain company has an average salary this year of $158,500 with a standard deviation of $12,500, and based on hitting certain sales goals, they decide to give all of their employees a bonus of $8,000 as well as 1.5% of their salary. What will the new average and standard deviation of the salary be for this company after the bonus is applied?

Answer: $168877.50 is the new average salary, and $12,687.50 is the new standard deviation.

Just so that you are aware, the reason the standard deviation and variance is unchanged by adding a constant value is that they are both measures (roughly) of how far a series of points is from the mean value. If all of the values in a set of values increase by the same amount, they remain the same distance from the mean value – therefore, there is no change in the value of the variance or standard deviation.

Example 9: AP Statistics test scores on Discrete Random Variables are described by the following probability distribution:

Score	40	50	60	70	80
P(Score)	0.1	0.2	0.3	0.3	0.1

2. Determine the mean and variance of the scores.

3. Mr. Murphy, in his great wisdom and benevolence, decides to scale the scores to better help his students in their college application process. He decides the actual grades will be transformed by the following linear function: **Grade = 1.5*Score – 20**. Determine the mean and variance of the transformed grades.

4. Which scores, if any, did not increase?

Answer:

(a) $\text{Mean}_{\text{initial}} = 61$, $\text{Var}_{\text{initial}} = 129$
(b) $\text{Mean}_{\text{transformed}} = 71.5$, $\text{Var}_{\text{transformed}} = 290.25$
(c) Scores of 40 did not increase, multiplying by 1.5 and subtracting 20 gets you back to 40 if you do the transformation.

Summary:

- If you have two random variables, X and Y, and real numbers a and b, the mean of $aX + bY$ is given by the following formula:
  - $\mu_{aX+bY} = a\mu_x + b\mu_y$
- **Independent Random Variables:** When knowing information about one variable does not change the probability distribution of the other variable.
- For independent random variables, X and Y, and real numbers a and b, the variance and standard deviation are given as follows:
  - **Variance:** $\sigma^2_{aX+bY} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}$
  - **Standard Deviation:** $\sigma_{aX+bY} = \sqrt{a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2}}$
- **Linear Function of the Random Variable, *x*:** If x is a random variable, and a and b are constants, the random variable y is a linear function of x if $y = a + bx$.
  - **Note that the formulas are given with *a* = *y*-intercept and *b* = slope.** If you use the formula $y = ax + b$ instead, the values of a and b are switched in the formulas.

Mean of y (for $y = a + bx$) is $\mu_y = \mu_{a+bx} = a + b\mu_x$

Variance of y (for $y = a + bx$) is $\sigma_y^2 = \sigma_{a+bx}^2 = b^2\sigma_x^2$

- Standard deviation is just the square root of the variance, so standard deviation is $\sigma_y = \sigma_{a+bx} = |b|\sigma_x$

Checkpoint 4.4

Multiple Choice

5. A bike relay race is conducted with two person teams. One person rides on a level roadway and the other rides over hilly terrain and the times are combined. The mean time on the level roadway is 15 minutes with a standard deviation of 3 minutes. The mean time on the hilly course is 21 minutes with a standard deviation of 4 minutes. Assuming that the times from both portions of the race are normally distributed and the riders' times are independent of each other, what is the standard deviation of the combined time?
   (a) 1 minute
   (b) 5 minutes
   (c) 6 minutes
   (d) 7 minutes
   (e) Cannot be determined since the number of contestants is unknown.

2. A company with 16 employees gives everyone a \$2000 bonus. What will be the change in the standard deviation of the employees' incomes after the bonus is awarded?
   (a) It will increase by \$2000.
   (b) It will be the same.
   (c) It will increase by \$500.
   (d) It will increase by $\$\sqrt{2000}$
   (e) It will be multiplied by \$2000.

3. Suppose X and Y are random variables with $E(X) = 25$, $\text{Var}(X) = 3$, $E(Y) = 30$, and $\text{Var}(Y) = 4$. What are the expected value and variance of the random variable $X + Y$?
   (a) $E(X+Y) = 55, \text{Var}(X+Y) = 3.5$
   (b) $E(X+Y) = 55, \text{Var}(X+Y) = 5$
   (c) $E(X+Y) = 55, \text{Var}(X+Y) = 7$
   (d) $E(X+Y) = 27.5, \text{Var}(X+Y) = 7$
   (e) There is insufficient information to answer this question.

Use the following information for questions 4 through 6. The independent random variables X and Y are defined by the following probability distribution tables.

X	1	3	6
$P(X)$	0.6	0.3	0.1

Y	2	3	5	7
$P(Y)$	0.1	0.2	0.3	0.4

4. Determine the mean of $X + Y$
   (a) 7.2
   (b) 8.4
   (c) 5.1
   (d) 9
   (e) 4.3

5. Determine the standard deviation of $3Y + 5$
   (a) 0.44
   (b) 3.62
   (c) 0
   (d) 5.1
   (e) 5.44

6. Determine the mean of $2X + 1$
   (a) 1.6
   (b) 2.1
   (c) 3.2
   (d) 4.2
   (e) 5.2

4.4 Homework

1. **(Multiple Choice)** Suppose the average height of policemen is 71 inches with a standard deviation of 4 inches, while the average for policewomen is 66 inches with a standard deviation of 3 inches. If a committee looks at all ways of pairing up one male with one female officer, what will be the mean and standard deviation for the difference in heights for the set of possible partners.
   (a) Mean of 5 inches with a standard deviation of 1 inch.
   (b) Mean of 5 inches with a standard deviation of 3.5 inches.
   (c) Mean of 5 inches with a standard deviation of 5 inches.
   (d) Mean of 68.5 inches with a standard deviation of 1 inch.
   (e) Mean of 68.5 inches with a standard deviation of 3.5 inches.

2. **(Multiple Choice)** Suppose that X and Y are independent random variables with $\mu_x = 720,\ \sigma_x = 6$ and $\mu_y = 240,\ \sigma_y = 8$. Given that X and Y are independent, what are the mean and standard deviation of $X + Y$?
   (a) $\mu_{X+Y} = 960, \sigma_{X+Y} = 7$
   (b) $\mu_{X+Y} = 960, \sigma_{X+Y} = 10$
   (c) $\mu_{X+Y} = 960, \sigma_{X+Y} = 14$
   (d) $\mu_{X+Y} = 480, \sigma_{X+Y} = 14$
   (e) $\mu_{X+Y} = 480, \sigma_{X+Y} = 7$

3. A particular phone app sells 15-second, 30-second and 45-second advertising spots. Let x represent the length of a randomly selected advertisement appearing on the app. The probability distribution for x is given below:

x	15	30	45
$p(x)$	0.1	0.6	0.3

   a. Find the mean length of advertisements appearing on this app.
   b. If a 15-second spot sells for \$40, a 30-second spot sells for \$70, and a 45-second spot costs \$100, find the average amount paid for advertisements on this app.

4. An appliance dealer sells three different models of refrigerators, with interior space of 13.5, 15.9, and 19.1 cubic feet of storage space, and x = the amount of storage space in a refrigerator purchased by a customer. Suppose x has a probability distribution given below:

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

   a. Calculate the mean and standard deviation of x.
   b. Suppose the price of the refrigerator is dependent on the size of the storage space, and the relationship is linear. If $Price = 25x - 8.5$, what are the mean and standard deviation of the price paid by someone who buys a refrigerator?

5. To assemble a piece of furniture, a wooden peg must be inserted into a pre-drilled hole. Suppose the diameter of a randomly selected peg is a random variable with a mean of 0.25 inches and a standard deviation of 0.006 inches, and the diameter of a randomly selected hole has a mean of 0.253 inches and a standard deviation of 0.002 inches. Let x_1 = peg diameter and x_2 = hole diameter.
   a. Why would the quantity $y = x_2 - x_1$ be important to the furniture manufacturer?
   b. What is the mean of the random variable y defined in part a?
   c. Assuming that x_1 and x_2 are independent of each other, what is the standard deviation of y?
   d. Based on you answers to parts b and c, do you think finding a peg too big to fit in the pre-drilled hole would be a common or an uncommon occurrence? Explain.

6. A standardized multiple choice exam has 100 questions. Each question has 5 choices, only one of which is correct, and the total score on the exam is calculated as follows:

   $$y = x_1 - 0.25x_2$$

   $x_1 =$ the number of correct answers

   $x_2 =$ the number of incorrect answers

   Calculating a score on a standardized exam by penalizing incorrect answers is a way of discouraging random guessing on the exam. It is known that a student who guesses completely at random will get an average of 10 answers correct and 40 answers incorrect. What is the mean value of the total score y? Why do you think this method is used on some standardized tests?

4.5 The Binomial Distribution

Objectives:

- Calculate the mean and standard deviation of a binomial random variable.
- Calculate probabilities for a binomial distribution.

When we have a number of trials in an experiment where a successful outcome always has the same probability, and when we are looking for a number of successes in those trials, we refer to that as a *binomial probability distribution*. This is a frequency distribution of the possible number of successes out of the number of trials.

A **Binomial Experiment** has the following properties:

1. The experiment consists of a fixed number of observations, called trials.
2. Each trial has only two, mutually exclusive, outcomes – we label these as success (S) or failure (F).
3. Outcomes of each trial are independent of one another.
4. The probability that a trial is a success is the same for each trial.

The **Binomial Random Variable, *x*,** is defined as follows:

- x = the number of successes observed when the experiment is performed.

A probability distribution of this variable, x, is referred to as **the Binomial Probability Distribution.**

An example of a binomial distribution would be if I decided to flip a coin 10 times, and count the number of heads that come up.

1. There are a fixed number of observations: $n = 10$ in this case.
2. A success (S) = "heads" and a failure (F) = "tails"
3. Each trial is independent, as one coin flip does not affect the next.
4. The probability of success is $p = 0.50$

If, instead, I had decided to keep flipping the coin until I got 8 heads, that ***would not be*** a binomial distribution, as I do not have a fixed number of trials – I keep trying until I reach a set number of successes.

Example 1: Determine whether each of the following experiments meets the criteria for a binomial experiment. If they do, identify n = the number of trials, and p = the probability of a success.

(a) A couple is planning on having 5 kids, and they will keep track of the number of girls that they have.

(b) A couple plans on continuing to have kids until they have a girl.

(c) You will draw 10 cards from a deck without replacement and count the number of successes, with success defined as drawing a heart.

(d) You will draw 10 cards from a deck with replacement and count the number of successes, with success defined as drawing a heart.

(e) An airline has a probability that flights are on time 91.3% of the time, and the probability that each flight is on time is independent of other flights. You want to calculate that no more than 3 of the next 10 flights will be delayed.

(f) You want to determine the probability that in the next 6 rolls of a fair, six-sided die, you will roll 3 or more ones.

(g) An airline has an on time rate of 96.5%, but as one flight is delayed the likelihood of the next one being on time decreases. You want to know the likelihood that 4 of the next 5 flights are on time.

Answers:

(a) Binomial Experiment: $n = 5$, $p = 0.50$
(b) Since they plan on continuing to have kids until they have a girl, so there are not a fixed number of trials, so it is not a binomial experiment.
(c) Since you are drawing without replacement, the probabilities change each time, so this is not a binomial experiment.
(d) Binomial Experiment: $n = 10$, $p = 0.25$
(e) Binomial Experiment: $n = 10$, $p = 0.913$
(f) Binomial Experiment: $n = 6$, $p = 1/6$
(g) Since the probabilities change when a flight is delayed, this is not a binomial experiment.

Binomial Probability

Let n be the number of independent trials in a binomial experiment, and let p be the constant probability that any particular trial results in a success. Then

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad \text{or} \qquad P(X=k)={}_nC_k\cdot p^k(1-p)^{n-k}$$

n = number of trials
k = number of successes
p = probability of a success

$$\binom{n}{k}={}_nC_k=\frac{n!}{(n-k)!k!}$$

- Note that some books and calculators will use r instead of k
- ${}_nC_r$ is located on your calculator by using the **math** button
  - Move over to the PROB menu, and it is option 3:nCr
- The "factorial" notation ($n!$) = n (n – 1) (n – 2) … (2) (1)

Example 2: Sixty percent of all computer monitors sold by a large computer retailer are laptops and 40% are desktop models. The type of computer purchased by each of the next 12 customers will be noted. Define a random variable x by

x = number of computers among these 12 that are laptops

(a) Is this a binomial distribution?

(b) Calculate $P(X=4)$

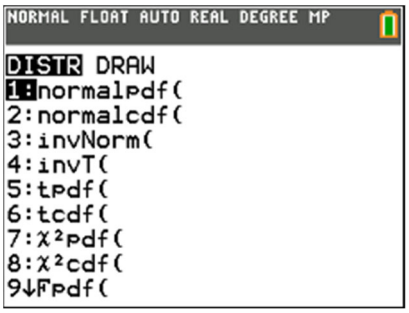
(c) Calculate $P(X\le 4)$

(d) Calculate $P(4\le X\le 7)$

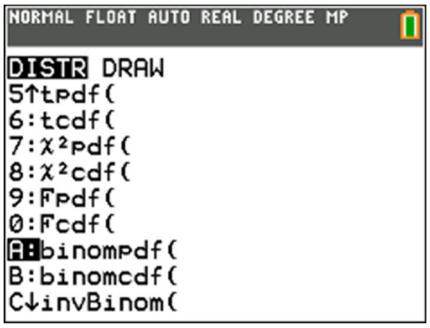
Answer:

(a) Yes, this is a binomial distribution. Success is a laptop, $n = 12$, $p = 0.40$, and the trials are independent:

(b) $P(X = 4) = 0.212$. You can arrive at this longhand (by plugging in to the formula above), but you can also find a binomial distribution on your calculator. Above the **vars** button you might notice **distr**. When you hit 2nd and **distr** then you get the following menu:

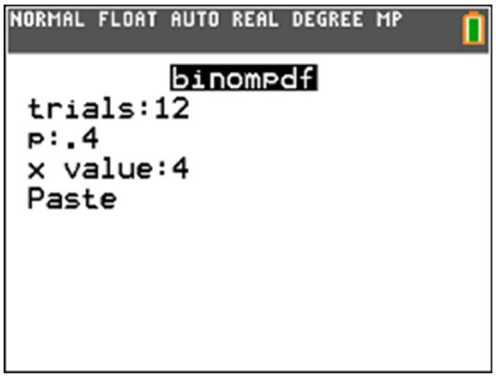


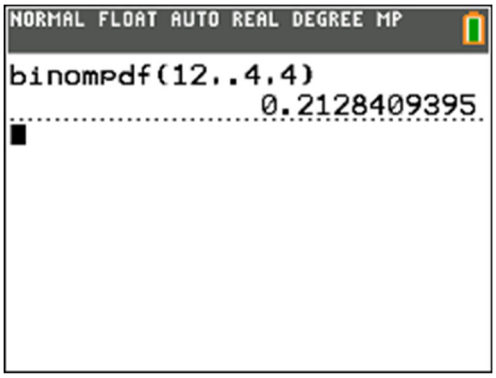
Scroll down, and you will ultimately see binompdf and binomcdf:



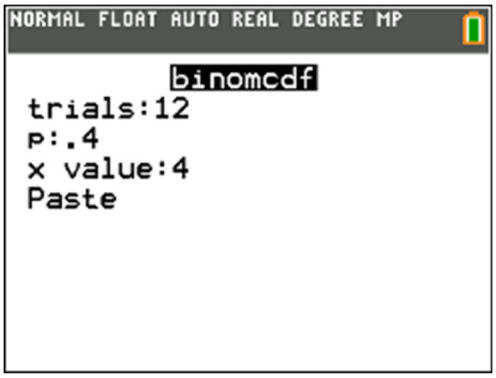
- **binompdf(** - this function is the *binomial probability density function*, and it will get us the probability of an *exact number of successes* (in this case $n = 4$)
- **binomcdf(** - this function is the *binomial cumulative density function*, and it will get us the probability of *less than or equal to* a set number of success.

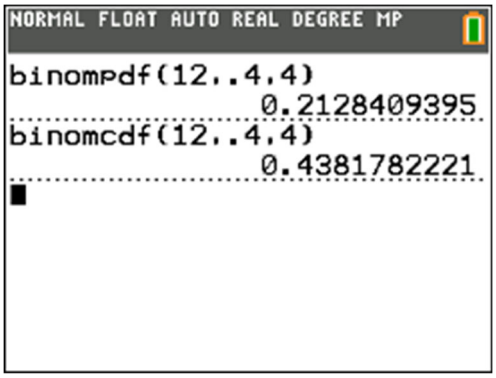
In this case, we will use **binompdf(** because we want an exact number of successes:





(c) Since we want less than or equal to 4 successes, we will use **binomcdf(**
we could also add up the 4 individual probabilities, but this takes a lot of work:
$P(X \le 4) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$





$P(X \le 4) = 0.438$

(d) If I want $P(4 \le X \le 7)$, that is the same as $P(X \le 7) - P(X \le 3)$, so I simply do the process twice on my calculator and subtract the values. I end up with the result
$P(4 \le X \le 7) = 0.717$

What if I had asked for $P(4 < X < 7)$? In this case, I could not include the 4 or 7, so I am really looking for $P(X = 5) + P(X = 6)$, so I could either do that using binompdf(, or I could do $P(X \le 6) - P(X \le 4)$ using binomcdf(. Remember, cdf finds values <u>*less than or equal to*</u> a value.

$P(4 < X < 7) = 0.404$

Example 3: What is the probability that on five rolls of a fair, six-sided die that you will roll three or more ones?

(a) 99.7% (b) 96.5% (c) 40.2% (d) 3.5% (e) 0.3%

Answer: (d) is the correct answer – you should have done $1 - P(X \leq 2)$. $P(X \leq 2)$ would have calculated the probability of less than or equal to two ones, and we want the likelihood of greater than 3 ones, so we are looking for the complement.

We can also calculate the ***mean value*** (or expected value) and the ***standard deviation*** for binomial distributions using the following formulas (these are on the AP formula sheet):

- **Mean:** $\mu_X = np$
- **Standard Deviation:** $\sigma_X = \sqrt{np(1-p)}$

Example 4: In a binomial distribution with sample size of $n = 65$ and a probability of success of $p = 0.8$, which of the following would be the approximate mean of the distribution?

(a) 52
(b) 65
(c) 10.4
(d) 3.22
(e) 8

Answer: Mean is given by $\mu_X = np$, substituting in the values, we get $\mu_X = (65)(0.8) = 52$, so the answer is (a).

Example 5: You are to take a multiple-choice exam consisting of 100 questions with 5 possible responses to each question. Suppose that you have not studied and so must guess (select one of the five answers in a completely random fashion) on each question. Let x represent the number of correct responses on the test.

(a) What kind of distribution does x have?

(b) What is your expected score on the exam?

(c) Compute the variance and standard deviation of x.

(d) Based on your answers to parts (b) and (c), is it likely that you would score over a 50 on the exam? Show the work that supports your answer.

Answers:

(a) This is a binomial distribution with $n = 100$, and $p = 0.20$.

(b) The expected value, or mean value is $\mu_X = np$, so $\mu_X = (100)(0.2) = 20$

(c) Standard deviation is $\sigma_X = \sqrt{np(1-p)}$, so $\sigma_X = \sqrt{(100)(.2)(1-0.20)} = 4$. Recall that the variance is simply the standard deviation squared, so it is 16.

(d) Since the standard deviation is 4, from the Empirical Rule, 99.7% of the values fall between 8 and 32 (assuming that this is distributed normally), so it is very unlikely that you will score above a 50.

Important Note: Context is vital in Binomial Distributions. Make sure you use appropriate units and explanations when answering free response questions.

Summary:

- A **Binomial Experiment** has the following properties:
  1. The experiment consists of a fixed number of observations, called trials.
  2. Each trial has only two, mutually exclusive, outcomes – we label these as success (S) or failure (F).
  3. Outcomes of each trial are independent of one another.
  4. The probability that a trial is a success is the same for each trial.
- **Binomial Probability:** Let n be the number of independent trials in a binomial experiment, and let p be the constant probability that any particular trial results in a success. Then

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{or} \qquad P(X=k) = {}_nC_k \cdot p^k (1-p)^{n-k}$$

  - n = number of trials, k = number of successes, p = probability of a success
- You can use functions on your calculator to generate various probabilities:
  - **binompdf(** - this function is the *binomial probability density function*, and it will get us the probability of an *exact number of successes*.
  - **binomcdf(** - this function is the *binomial cumulative density function*, and it will get us the probability of *less than or equal to* a set number of success.
- **Statistics for the binomial distribution:**
  - **Mean:** $\mu_X = np$
  - **Standard Deviation:** $\sigma_X = \sqrt{np(1-p)}$

Checkpoint 4.5

Multiple Choice

1. Suppose we have a random variable, X, where the probability associated with the value k is $P(X=k)=\binom{10}{k}(0.37)^{k}(0.63)^{n-k}$ for $k = 0, 1, 2, \ldots, 10$. What is the mean of X.

   (a) 0.37
   (b) 0.63
   (c) 3.7
   (d) 6.3
   (e) None of the above.

2. Which of the following is *not* a condition for a binomial experiment?

   (a) There are only two possible outcomes for each trial.
   (b) The probability of success is the same for each trial.
   (c) The trials are independent.
   (d) There are a variable number of trials.
   (e) The variable of interest is the number of successes.

3. Statistics show that 7.3% of workers between the ages of 16 and 24 earn the minimum wage or less. What is the probability that if three young adults between the ages of 16 and 24 are polled, two or more will earn the minimum wage or less?

   (a) 0.0004
   (b) 0.0148
   (c) 0.0152
   (d) 0.0627
   (e) 0.677

Free Response

1. Suppose that $T = 2X + 3Y$, where X is your score on the multiple-choice part of a test, Y is your score on the written part of a test, and T is your total score. In this case, your total score is calculated by doubling your multiple-choice score and tripling your written score. Suppose $X \sim N(30, 7)$ and $Y \sim N(20,13)$.

   (a) What is the probability that T will be greater than 130?
   (b) Suppose that we find three students' T scores independently of one another. What is the probability that <u>at least one</u> of these measurements will be greater than 130?

4.5 Homework

1. Suppose we know that in a certain population, 8% of the adults are left-handed. Consider 10 randomly selected individuals from that population, and define the random variable x as follows:

   x = the number of people in the sample who are left-handed.

   Find the following probabilities:

   a. $p(x < 3)$
   b. $p(x \le 3)$
   c. $p(x \ge 4)$
   d. $p(1 \le x \le 3)$
   e. $p(1 < x < 3)$
   f. $p(x = 2)$

2. Suppose we know 90% of people carry a smartphone when they fly on an airplane. Consider that we randomly select 8 people and define the random variable x is the number of people in the sample carrying a smartphone. The probability distribution of x is the binomial distribution with $n = 8$ and $p = 0.9$.
   a. Calculate $p(5)$ and interpret the result.
   b. Calculate $p(8)$, the probability that all of the selected people are carrying a smartphone.
   c. Calculate the value of $p(x \ge 5)$

3. Suppose that we know a ride-sharing app has an on-time pick-up rate of 88.2%. Assuming that we have a binomial distribution, calculate each of the following:
   a. What is the probability that no more than 3 of your next 10 rides will be late?
   b. What is the probability that more than 3 of your next 10 rides will be late?
   c. What is the probability that none of your next 10 rides will be late?
   d. What is the probability that only 1 of the next 10 rides will be late?

4.6 The Geometric Distribution

Objectives:

- Calculate the mean and standard deviation of a geometric distribution
- Calculate probabilities for a geometric random variable.

The **Geometric Probability Distribution** is very similar to the *binomial distribution* from the last section – so much so that they are often confused with each other. A lot of times, we are not concerned with a fixed number of trials as much as we would like to know the probability of getting a certain number of successes. For example, if a rare prize is in some cereal boxes contain a rare prize, how many boxes would I need to open to get the prize. Another similar application could be in a video game a certain enemy has a rare drop – how many of that enemy would I have to defeat to get that rare drop?

These are cases where a *geometric distribution* is a useful tool.

- **Properties of a Geometric Experiment**

    1. There is a sequence of trials (that is uncertain in length).
    2. The trials are independent.
    3. Each trial can result in one of two possible outcomes, success or failure.
    4. The probability of success is the same for all trials.

- **Geometric Random Variable, *x*:** x = the number of trials until the first success is observed (including the success trial).

The probability distribution of this random variable, x, is called the **geometric probability distribution**.

Example 1: Which of the following represents a geometric distribution?

(a) The number of red M&M's in a handful of 25 M&M's.
(b) The number of cards dealt from a deck before you get a 10.
(c) The amount of time you wait in line at a bank before you get to the counter.
(d) The number of random telephone numbers you dial until you get an answer.
(e) The number of people entering the intensive care unit at a particular hospital on any day.

Answer: The correct answer is (d). It cannot be (a) because there are a fixed number of trials. It is not (b) because the trials are not independent (as you draw cards the probabilities change). (c) represents a continuous variable, and (e) has a timespan parameter, not a number of trials.

Geometric Probability

If x is a geometric random variable with probability of success p for each trial, then

$$P(X = k) = (1 - p)^{k-1} p$$

k = number of trials needed to get the first success p = probability of success

Example 2: Suppose that 40% of the students attending a certain college who drive to campus carry jumper cables. Define a random variable x = number of students who must be stopped before finding a student with jumper cables.

(a) What kind of distribution does x have?

(b) What is the probability that it takes two students before we find one with jumper cables?

(c) What is the probability that it takes three students before we find one with jumper cables?

(d) What is the probability that 3 or fewer students must be stopped before we find jumper cables?

Answer:

(a) It is a *geometric distribution*. We have a variable number of trials that have a defined success or failure with fixed, independent probabilities.

(b) $P(X=k)=(1-p)^{k-1}\,p$, with k =2 and p = 0.40

$P(X=2)=(0.60)^{2-1}(.40)$

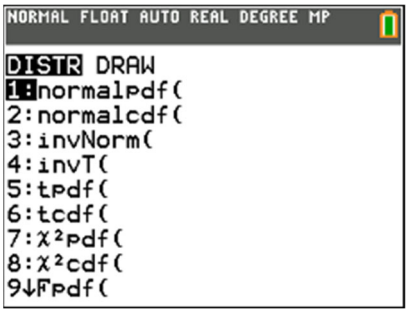
$P(X=2)=0.24$

(c) $P(X=3)=(0.60)^{3-1}(.40)$

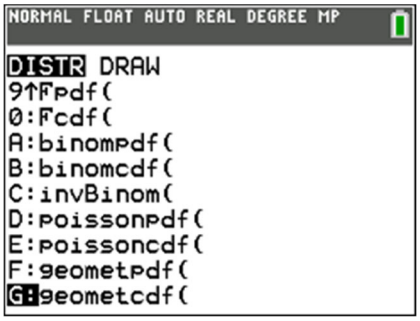
$P(X=2)=0.144$

(d) $P(X\le 3)=P(X=1)+P(X=2)+P(X=3)=0.784$

Similar to the processes in the previous section, we can use our calculators to generate these probabilities. Above the **vars** button you might notice **distr**. When you hit 2nd and **distr** then you get the following menu:



Scroll down, and you will ultimately see geompdf and geomcdf:



- **geompdf(** - This function is the *geometric probability density function*, and it will get us the probability of an *exact number of trials* (like k = 3, for example).
- **geomcdf(** - This function is the *geometric cumulative density function*, and it will get us the probability of *less than or equal to* a set number of trials.

Again, similar to the *binomial distribution*, we can also calculate the ***mean value*** (or expected value) and the ***standard deviation*** for ***geometric distributions*** using the following formulas (these are on the AP formula sheet):

- **Mean:** $\mu_X = \frac{1}{p}$
- **Standard Deviation:** $\sigma_X = \frac{\sqrt{1-p}}{p}$

Example 3: A certain batter has a batting average of 0.337. If each at bat is an independent event, find each of the following:

(a) Find the mean number of at bats before this batter would get a hit.

(b) Find the standard deviation for this geometric distribution.

(c) What is the probability that the batter gets a hit by his third at bat?

(d) What is the probability that the batter gets his first hit in his fifth at bat?

Answers: (a) 2.97 at bats (b) 2.416 (c) 0.709 (d) 0.065

Checkpoint 4.6

Multiple Choice:

1. Which of the following is *not* a condition for a geometric setting?

   (a) There are only two possible outcomes for each trial.
   (b) The probability of success is the same for each trial.
   (c) The trials are independent.
   (d) There are a fixed number of observations.
   (e) The variable of interest is the number of trials required to reach the first success.

2. A baseball recruiter visits a high school where a player has a batting average of 0.450 (This means that he gets a hit in 45% of his at-bats.) What is the probability that the recruiter won't see the player get a hit until his third at-bat?

   (a) $(0.450)^2 (0.550)$
   (b) $(0.550)^2 (0.450)$
   (c) $(0.450)^3 (0.550)^2$
   (d) $\binom{3}{1}(.450)(.550)^2$
   (e) $\binom{3}{2}(.450)(.550)^2$

Free Response

1. A certain enemy in Elden ring has been programmed to drop an item at a rate of 2.5% when defeated. Use that fact to find each of the following:
    a. What is the mean number of that enemy you must defeat to get that item?

    b. What is the standard deviation for this geometric distribution?

    c. What is the probability that you will get that item by defeating the enemy five times or fewer?

    d. What is the probability that you will get that item by defeating the enemy 20 times or fewer?

    e. How many times would you have to defeat this enemy to have a 0.90 probability of getting the item?

4.6 Homework

1. Suppose the inside of a bottle cap on a particular juice drink has a message written on it. There is a 5% chance that the message is "Congratulations, you have won a free bottle of juice" and there is a 95% chance that it says, "Sorry, you are not a winner, please try again". Consider the random variable x = the number of bottles purchased until a prize is found.
   a. Would this be a binomial or a geometric distribution? Explain.
   b. What is the probability that at most three bottles must be purchased?
   c. What is the probability that exactly 5 bottles must be purchased?
   d. What is the probability that more than 5 bottles must be purchased?

2. Suppose that a certain rare item in a video game will be dropped by a defeated enemy 0.25% of the time (that is, the probability is 0.0025). The rest of the time, nothing is dropped by this particular enemy. Consider the random variable x = the number of times you must defeat the enemy before the item drops.
   a. What is the probability of getting the item by defeating the enemy ten times or fewer?
   b. What is the probability of getting the item by defeating the enemy exactly 50 times? In 50 times or fewer?
   c. What is the probability of getting the item by defeating the enemy more than 50 times?
   d. What is the probability of getting the item by defeating the enemy fewer than 100 times?

3. Suppose that 20% of all vehicles undergoing emissions inspections at a particular inspection station fail the inspection. Determine whether each of the following is a binomial or geometric distribution, and calculate the probability.
   a. What is the probability that among 15 randomly selected cars, at most 5 cars will fail?
   b. What is the probability that a car will fail before 6 cars are tested?
   c. What is the probability that a car will not fail until after 10 cars have been tested?
   d. What is the probability that between 5 and 10 cars will fail if 30 cars are tested?

4. Suppose that when you restart a game on a particular app, there is an 80% chance that an advertisement will play, the remainder of the time, no ad plays. Let x = the number of restarts before an advertisement plays.
   a. What is the probability that an advertisement plays before the third restart?
   b. What is the probability that an advertisement will play only after three or more restarts?
   c. What is the probability that an advertisement plays before the fourth restart?

Unit 4 Practice Test