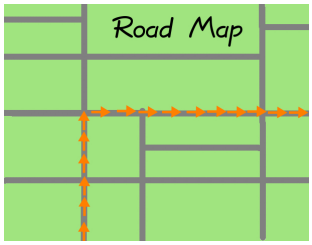
Module 3.2: Exploring Probability Through Problem Solving



This is the most important module in the entire chapter. I have painstakingly collected a set of problems, gathered over several years of teaching this course. These problems have been hand-chosen to address interesting questions. That's in stark contrast to the probability problems that you might have seen in high school about dice, cards, and pulling balls out of urns.

I have three strategies that I'd like you to learn and that can solve a surprising breadth of probability problems. The problems in this module have been arranged carefully, starting with easier ones and moving toward hard ones. Yet, before we get properly underway, I'd like to challenge you with a very fair and realistic but moderately difficult problem, to show you how subtle probability can be.

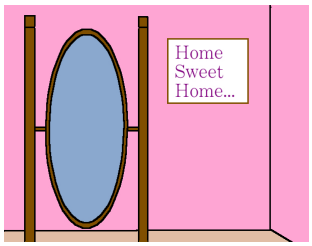
I'd like to address an important issue now. It is extremely likely that you've seen the topic of probability before, in other university classes, in high school, or even in middle school. Many of my students will say, "Probability is easy! What's the big deal? Why do we need these large readings for something that simple?!"



In the end, I am compelled to agree that probability is easy, as a topic, overall. However, there are some pitfalls that can trap you. These pitfalls can even trap very experienced professional scientists if they are not paying attention.

I have taught *Discrete Mathematics* many times. (The sixth time was in the Fall of 2017.) I have gone back in painstaking detail over the final exams of my students. I analyzed the errors to try to see if there were some common mistakes, usually during summer break or winter break.

For probability, I have found six core errors that are remarkably common. They will be listed, by order of frequency, at the end of this module (on Page 273).



A Pause for Reflection...

The remark which denigrates probability that I hear most often from students is "Isn't probability just all about problems where you divide two numbers? What's the big deal? It is just division!"

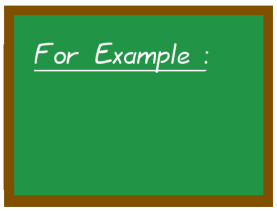
Technically, this is true. Most probability problems come down to a few very simple arithmetic operations. However, you have to pick the correct numbers for the division or other arithmetic step. As we are about to see, that isn't always easy.

We are about to consider an example with a very simple data set. From that simple data set, a whole host of questions can be asked. This will reveal that while probability might sometimes seem trivial, it is actually a very rich and deep topic.

Between January 31st and February 2nd, 2015, the magazine *The Economist* surveyed 998 Americans with the following question: "Do you think the government should or should not require parents to have their children vaccinated against infectious diseases (e.g. measles, mumps, whooping cough)?" I found the results on the magazine's webpage. The survey included three answers: "Should Require," "Not Sure," and "Should Not Require." The survey also asked the respondent their political party, with the answers "Democrat," "Independent," and "Republican."

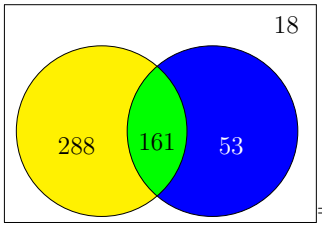
For our first attempt at this problem, to keep it simple and to keep the problem from becoming huge, we will exclude the "Not Sure" responses and any respondents who remarked "Independent." Otherwise, the problem would be much longer than it already is.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



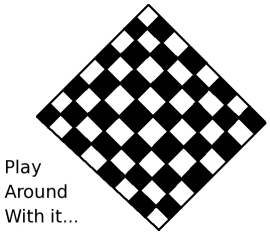
3-2-1

Here is a Venn Diagram representing the data from the survey in the previous box. There are 520 responses in the reduced data set. The left circle represents respondents who say that vaccines should be required, and republicans are the circle on the right.



=520

From this one very simple Venn Diagram, we can ask a whole host of questions. In fact, over the next few boxes, we will explore 16 different questions from this one data set. Some of my readers will be able to answer these questions without difficulty. Other readers will be unable to answer them. If you find yourself unable to answer these 16 questions correctly, then do not be discouraged, but complete the module, and return to retry these 16 questions after you have finished with the rest of the module.

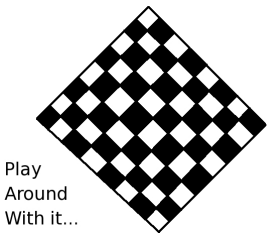


3-2-2

Referring back to the previous box, answer the following small questions.

- How many republicans are in the data set? [Answer: 214.]
- How many democrats are in the data set? [Answer: 306.]
- How many people in the data set wish to require vaccines? [Answer: 449.]
- How many people in the data set wish to not require vaccines? [Answer: 71.]

We will continue in the next box.



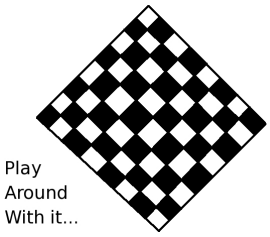
3-2-3

Continuing with the previous box,

- What is the probability that a random republican is in favor of requiring vaccines? [Answer: $161/214 = 75.2336\cdots\%$.]
- What is the probability that a random republican is not in favor of requiring vaccines? [Answer: $53/214 = 24.7663\cdots\%$.]
- What is the probability that a random democrat is in favor of requiring vaccines? [Answer: $288/306 = 94.1176\cdots\%$.]
- What is the probability that a random democrat is not in favor of requiring vaccines? [Answer: $18/306 = 5.88235\cdots\%$.]
- What is the probability that a random person in the data set is in favor of requiring vaccines? [Answer: $449/520 = 86.3461\cdots\%$.]
- What is the probability that a random person in the data set is not in favor of requiring vaccines? [Answer: $71/520 = 13.6538\cdots\%$.]

We will continue in the next box.

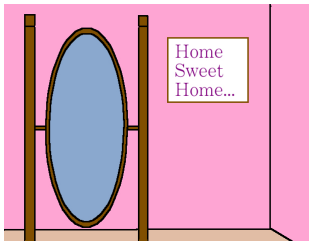
COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



3-2-4

Continuing with the previous box,

- What is the probability that a person in favor of requiring vaccines is a democrat? [Answer: $288/449 = 64.1425\cdots\%$.]
- What is the probability that a person in favor of requiring vaccines is a republican? [Answer: $161/449 = 35.8574\cdots\%$.]
- What is the probability that a person not in favor of requiring vaccines is a democrat? [Answer: $18/71 = 25.3521\cdots\%$.]
- What is the probability that a person not in favor of requiring vaccines is a republican? [Answer: $53/71 = 74.6478\cdots\%$.]
- What is the probability that a random person in the data set is a democrat? [Answer: $306/520 = 58.8461\cdots\%$.]
- What is the probability that a random person in the data set is a republican? [Answer: $214/520 = 41.1538\cdots\%$.]



A Pause for Reflection...
In conclusion, yes, most probability questions just are a matter of dividing two numbers. However, as you can see from the previous box, there are a lot of possible questions that can be asked about even an easily imaginable situation: a Venn Diagram with two circles.

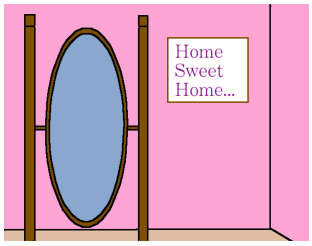


It is very common for someone to have a question in front of them, but to imagine that a slightly different question had been asked. Therefore, it is necessary to dwell on the most subtle distinctions in the wording. The tiniest change in phrasing can change the answer completely.

Observe, "What is the probability that a person in favor of requiring vaccines is a democrat?" had a probability of 64.1425%, whereas "What is the probability that a random democrat is in favor of requiring vaccines?" had a probability of 94.1176%.

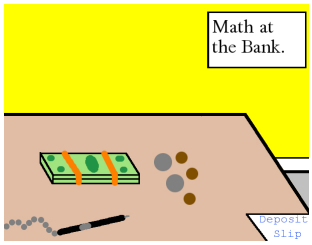
It is truly important to read a probability problem very carefully.

Remember, if you found yourself unable to correctly answer those 16 questions (about the vaccines), then do not be discouraged. Just complete the module, and return to retry those 16 questions after you have finished with the rest of the module.



A Pause for Reflection...
Here's another interesting perspective on those 16 questions that we just finished. Imagine that you happen to be a political writer, trying to write an article or a blog post about the survey that we just discussed. Then you have a large collection of twelve probabilities that you can choose from, to paint whatever picture you might want to paint.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



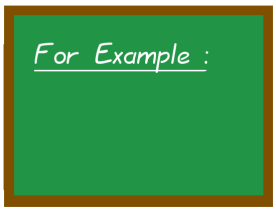
By the way, now would be a good time to mention that probabilities can be written as fractions, decimals, or percentages. For example, it is common to see 75%, 3/4, or 0.75. Each is entirely acceptable (unless an examination question were to specifically note otherwise), and all three formats can be found in higher-level textbooks on economics, finance, or business.



I would like to offer you two basic strategies, and a third strategy will follow later, on Page 259 of this module.

1. The probability of a compound event is equal to the sum of the probabilities of the simple events (the outcomes) that comprise that compound event. While this technique is always correct, it does require you to know the probabilities of the simple events. (Sometimes you don't have those, or perhaps you don't have them right away.)

2. In a sample space where the "equally likely assumption" holds, you can often frame a simple event or a compound event as the probability that a random person or item from some interesting set happens to be from some very interesting subset. In this case, the probability is equal to the size of the very interesting subset divided by the size of the interesting set. This strategy must not be used when the "equally likely assumption" does not hold.



3-2-5

Imagine a manager who is frustrated with a particular cluster of employees (perhaps stock-room workers). He might estimate that there's a 5% chance one of them will be absent on a given day; a 10% chance that they'll be late; and an 85% chance they'll be on time. We could ask what is the probability that such an employee is "present," i.e. either late or on time, but not absent.

First, we ask ourselves if the set of outcomes "absent," "late," and "on time" represent a sample space. They appear to be mutually exclusive and collectively exhaustive, therefore, the set is a sample space. Second, we should verify that the probabilities do add to 100%.

$$5\% + 10\% + 85\% = 100\%$$

Since they add to 100% and none are negative, this is a valid probability distribution.

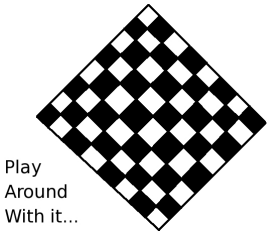
We will continue in the next box.

Continuing with the previous box, we must realize that the event we are asked about is a compound event, consisting of the simple events (outcomes) named "late" and "on time."

In this situation, we add the probabilities of the simple events that comprise the compound event. The probability of this event "present" is easily calculated by

$$10\% + 85\% = 95\%$$

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



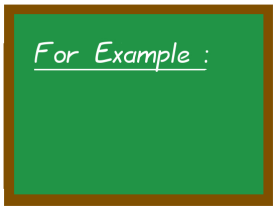
3-2-6

Returning to the stock-room employees of the previous box, suppose there is a form that the supervisor must fill out when a stock-room employee is either absent or late. What is the probability that the form must be used, for any particular employee on any particular day? [Answer: 15%.]



Did you notice how in the previous example (and the follow-up checkerboard box) the outcomes were not equally likely? That's okay, because our strategy—adding together the probabilities of several outcomes to get the probability of a compound event comprised of those several outcomes—works regardless if the outcomes are equally likely or not.

The other strategy, of dividing the size of a very interesting subset by the size of an interesting set, does require the outcomes to be equally likely. We called that the "equally likely assumption."



3-2-7

Suppose there is a survey with 487 respondents, to see if they read the local newspaper. The choices were "always read it," "sometimes read it," or "never read it." There were 68 people who said "always," 206 who said "sometimes," and 213 who said "never." What are the probabilities, as a percentage, that a random member of the set of respondents would give each of those answers, when asked whether they read the local newspaper?

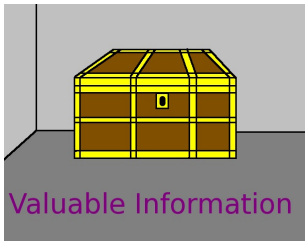
The sample space is

$$\mathcal{S} = \{Always, Sometimes, Never\}$$

First, we can calculate the probabilities of the simple events (the outcomes).

- $Pr\{Always\} = 68/487 = 0.139630\cdots = 13.9630\cdots\%$
- $Pr\{Sometimes\} = 206/487 = 0.422997\cdots = 42.2997\cdots\%$
- $Pr\{Never\} = 213/487 = 0.437371\cdots = 43.7371\cdots\%$

We will continue our analysis after a few important comments in the next box.



This is the first time we've used the notation $Pr\{E\}$, which indicates the probability of an event E. It can be used for simple or compound events. This notation is extremely common.

Moreover, our list of three outcomes in the previous box, "always," "sometimes," and "never" form a sample space, by being mutually exclusive and collectively exhaustive. When we assign probabilities to those events, we have a probability distribution.

Recall that a probability distribution is a list of outcomes in a sample space along with their probabilities.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

For Example :

3-2-8

Continuing with the previous example, we computed the probability that a random respondent from a survey reads the newspaper "always," "sometimes," or "never." Now, using that data, we can calculate the probability of some compound events. Here is the data we found earlier:

- $Pr\{Always\} = 68/487 = 0.139630\cdots = 13.9630\cdots\%$
- $Pr\{Sometimes\} = 206/487 = 0.422997\cdots = 42.2997\cdots\%$
- $Pr\{Never\} = 213/487 = 0.437371\cdots = 43.7371\cdots\%$

Based on that data, we can answer more questions:

- What is the probability, given the data in the previous problem, of someone reading the newspaper "at least once in awhile"?

$$\frac{68}{487} + \frac{206}{487} = \frac{274}{487} = 0.562628\cdots = 56.2628\cdots\%$$

- What is the probability, given the data in the previous problem, of someone reading the newspaper "less than every day"?

$$\frac{206}{487} + \frac{213}{487} = \frac{419}{487} = 0.860369\cdots = 86.0369\cdots\%$$

For Example :

3-2-9

When trains arrive at a busy train station, an important task for the computer running the station is to dispatch them to empty tracks. Suppose a train is pulling into a train station and the computer (or alternatively the signaling system connecting the train and the computer) has failed. The train's driver must choose a track at random, because it is very difficult for trains to stop quickly, or even moderately quickly. There are 20 tracks, and 3 of them are occupied. What is the probability that the train chooses an occupied track, and therefore crashes? What is the probability that a train chooses an unoccupied track, and arrives safely?

Since the driver is choosing randomly, we can assume that all the tracks are equally likely in the sense of the "equally likely assumption." There are 3 of them that are occupied out of 20, so the probability of a collision caused by two trains sharing the same track is given by the following:

$$3/20 = 0.15 = 15\%$$

which is far too high of a risk to bear.

We will continue in the next box.

In the previous box, we computed the probability that the incoming train will pick a track that is already occupied. Now we should compute the probability that the incoming train picks a track that is not already occupied, therefore avoiding a crash.

We know that $20 - 3 = 17$ tracks are unoccupied. Therefore, the chance of picking an unoccupied track and not crashing is given by the following

$$17/20 = 0.85 = 85\%$$

which is far too low.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



The way to analyze the problem in the previous box is to realize that we're talking about an interesting set—the set of train tracks that could be chosen, a set of size 20. That will be the denominator of our fraction. Then we are asking about a subset of the interesting set, namely the subset with some property.

In the first calculation, we are asking about the subset of tracks that happen to already have a train on them, a subset of size 3. That was our first numerator, resulting in 3/20.

In the second calculation, we are asking about the subset of tracks that happen to not already have a train on them, a subset of size $20 - 3 = 17$. That was our second numerator, resulting in 17/20.

You have to look to the heart of the problem and its context to be able to determine what the interesting set is. That insight does not come from some set of rules that you can memorize, but rather by trying to visualize what is happening inside of the problem.

Last, but not least, I should remind you that the "interesting set" must always meet the requirements of being a sample space as well as the requirements of the equally likely assumption. As we saw many times in the previous module ("A Formal Introduction to Probability Theory"), it is possible for a probability problem to have many different sample spaces—all of which are valid. It might be the case that for some of them, the equally likely assumption holds, and for some of them, it does not.

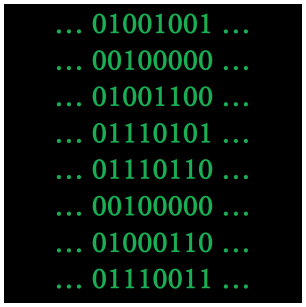
By the way, please understand that "interesting set" and "very interesting subset" are not technical terms or vocabulary terms. The interesting set is merely a set that is interesting in the context of a given problem.



In the previous example, we made use of the "equally likely assumption," and divided the size of a very interesting subset by the size of an interesting set. However, we could also see this as the other strategy, namely that of adding the probabilities of simple events to get the probability of a compound event.

Since the train driver chose one of the 20 tracks at random, each had probability 1/20. We essentially added 3 copies of 1/20 to get the probability of 3/20 of the train driver unluckily picking an occupied track. Similarly, we added 17 copies of 1/20 to get the probability of the train driver luckily picking an unoccupied track.

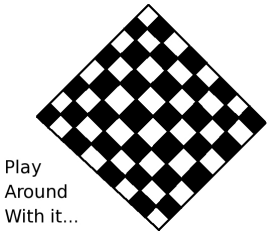
$$\underbrace{\frac{1}{20}+\frac{1}{20}+\frac{1}{20}+\cdots+\frac{1}{20}}_{17 \text{ copies}}=\frac{17}{20}$$



By the way, the previous example illustrates why the computers which run important transportation systems have protections to make sure they are always operating, including being attached to a UPS (*Uninterruptible Power Supply*). Many business people are unfamiliar with the existence of such power supplies, which usually guarantee anywhere from 30 minutes to 2 hours of power in the event of a power failure. This can help your business survive a minor natural disaster, like a hurricane or an earthquake, which is why I am bringing it to your attention.

Many of the techniques of this chapter, as well as two modules (Module 5.2: "Independence and Repetition" plus Module 5.3: "Bernoulli's Binomial Distribution Formula and Reliability") are useful in the design of high-reliability systems. In fact, the mathematics department in which I teach, the University of Wisconsin—Stout, lost a faculty member to a startup company that uses mathematical software to help analyze and design ultra-reliable digital equipment for critical systems like mass-transit controls.

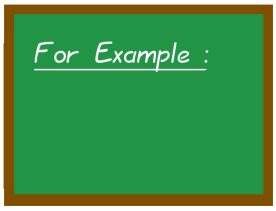
COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



3-2-10

Suppose that a large multinational company has 17 offices and 4 data centers in the USA. The Chief Information Officer (CIO) is going to inspect one location, choosing a facility at random. As it turns out, 11 out of 17 offices have installed a UPS. Moreover, each of the data centers has installed a UPS. What is the probability that the location which the CIO inspects has installed a UPS?

The answer will be given on Page 274 of this module.



3-2-11

Let's imagine three students, Marlann, Laura and Alan, who are taking an economics class. They have a small test today. The publisher of the economics textbook has provided a test bank, which the instructor uses. The instructor has told the class that the test is going to consist of one homework question from each of three chapters, chosen randomly from the set of assigned homework questions. The three students are really only worried about Chapter 5, as they've learned the other chapters very well.

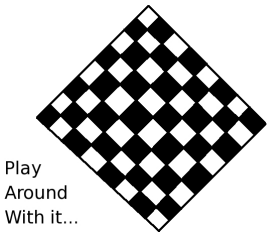
In Ch 5, section 5.1 had 20 questions, section 5.2 had 10 questions, section 5.3 had 15 questions, and section 5.4 had 5 questions. As it comes to pass, Alan studied 5.1 and 5.2, but not 5.3 and 5.4; Laura studied all the sections except 5.4; Marlann studied all the sections, without exception. We now wish to compute the probabilities that Marlann, Laura, and Alan get a question for Ch 5 drawn from a section that they studied. Of course, since they studied different amounts, the probabilities will not be the same.

We will solve the problem in the next box.

Looking at the data in the previous box, we can identify that our interesting set is the set of questions from Chapter 5, and there are $20+10+15+5=50$ of them. Because of the words in the problem statement, we know that the questions are chosen at random, so the "equally likely assumption" holds.

The numerator for Alan will be a subset of those 50 questions, namely those questions that come from sections which Alan studied. What is the size of that subset? The subset has $20+10=30$ questions in it. Therefore, the probability that Alan gets a question for Ch 5, drawn from one of the sections which he studied, is $30/50=0.6$.

In the next box, you will compute the probabilities for Laura and for Marlann.



3-2-12

Looking at the previous box, compute the probabilities that Marlann or Laura gets a question for Chapter 5, from a section that he or she has studied.

- What is the probability that Laura gets a question from a section that she studied? [Answer: $45/50=0.9$.]
- What is the probability that Marlann gets a question from a section that she studied? [Answer: $50/50=1.0$.]

As you can see, it is wise to make sure that you can answer all of the questions in a chapter, before showing up to a test.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



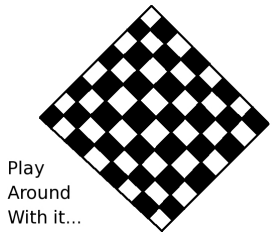
For the example about the automatically generated test (and the followup checkerboard box) it is nice to see how we can think of it as adding the probabilities of simple events (outcomes) to get the probability of a compound event. There were 50 questions in chapter 5, and since the questions are chosen at random, each had probability 1/50.

We essentially added 30 copies of 1/50 to get the probability of $30/50 = 0.6$ for Alan. In particular,

$$\underbrace{\frac{1}{50} + \frac{1}{50} + \frac{1}{50} + \cdots + \frac{1}{50}}_{\text{30 copies}} = \frac{30}{50} = 0.6$$

While the addition of all of that would be horribly time-consuming, it does show that both of our two fundamental strategies give us the same answer in this case.

Similarly, we added 45 copies of 1/50 to get $45/50 = 0.9$ for Laura. Finally, we added 50 copies of 1/50 to get $50/50 = 1.0$ for Marlann.



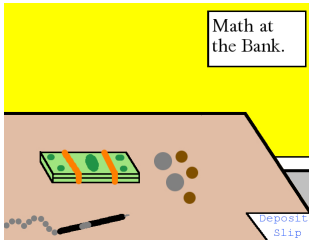
3-2-13

Depending on how you count, there are 238 countries in the world. The reason that the number isn't so clear has to do with places that have an intermediate-degree of self-government, like Hong Kong, Puerto Rico, or Svalbard. In any case, the CIA World Factbook estimated (for July of 2015) that the population of the world was 7,256,490,011. The top ten countries in population were as follows.

1:	China	1,367,485,388	6:	Pakistan	199,085,847
2:	India	1,251,695,584	7:	Nigeria	181,562,056
3:	USA	321,368,864	8:	Bangladesh	168,957,745
4:	Indonesia	255,993,674	9:	Russia	142,423,773
5:	Brazil	204,259,812	10:	Japan	126,919,659

If you picked a person at random from the world's population, then what is the probability that they come from one of these ten countries? [Answer: $58.1514\cdots\%$.]

For the sake of completeness, it might be good to mention that the sample space (the interesting set) was the set of people on the planet. The "very interesting subset" is the set of people who happen to be in the population of one of those ten listed countries.



For the previous box, please note that you definitely should not respond

$$\frac{4,219,752,402}{7,256,490,011}$$

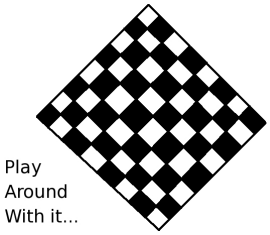
as this is not a human-readable number. It is hard to imagine that fraction in a concrete or tangible way, except perhaps we can see that it is slightly more than 1/2.

In any workplace, but especially in business, if you reply 58.1514%, 58.15%, or 58.1%, then people know what you mean. However, if you were to reply

$$\frac{4,219,752,402}{7,256,490,011}$$

then for sure, everyone is going to think that there is something wrong with you.

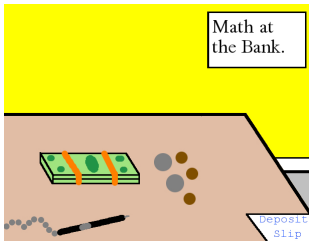
COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



3-2-14

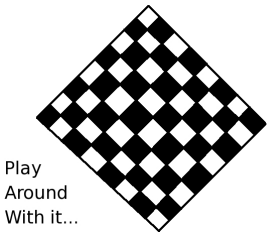
Looking at the previous checkerboard box and its data set, if you picked a random person in the world, what is the probability that they are...

- ...from China? [Answer: $18.8449\cdots\%$.]
- ...from the USA? [Answer: $4.42870\cdots\%$.]
- ...from Japan? [Answer: $1.74905\cdots\%$.]



This would be a good moment to mention that when you talk about probabilities in business or industry, usually you report them to the nearest basis point. A basis point is 1% of 1%. Therefore, we would normally write 18.84% or 18.85%, 4.42% or 4.43%, and 1.74% or 1.75%, for the three answers in the previous box.

I reported six digits there, because we are using six significant figures throughout this textbook.

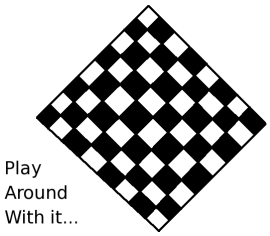


3-2-15

An ice-cream survey is being performed in a school cafeteria to determine what kind of flavors will sell the most. The choices are

Chocolate	213 votes	Strawberry	51 votes
Vanilla	88 votes	Mint Chocolate Chip	148 votes
Pistachio	23 votes	Salt-Water Taffy	2 votes
Banana Nut	18 votes		

- What is the probability that some random person from the survey likes strawberry? [Answer: $51/543 = 0.0939226\cdots$.]
- What is the probability a random person from the survey likes an ice-cream that contains chocolate in its name? [Answer: $361/543 = 0.664825\cdots$.]

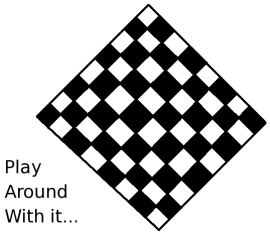


3-2-16

Let's continue with further questions from the previous box. Perhaps both the pistachio and the mint chocolate chip contain green food coloring, and a parent is concerned that a certain type of green food coloring was recalled by the manufacturer last year. No other flavors contain green food coloring, as it turns out.

What is the probability that a random student in the survey has chosen one of these two tainted ice creams as their choice?

[Answer: $171/543 = 0.314917\cdots$.]



3-2-17

Still continuing with the ice cream survey of the previous two boxes, let's consider the following questions.

- Both pistachio ice cream and banana-nut ice cream contain nuts, but none of the other flavors do. What is the probability that an ice-cream flavor containing nuts is chosen by a random student in the survey? [Answer: $41/543 = 0.0755064\cdots$.]
- Similar to the previous bullet, what is the probability that an ice-cream without nuts is chosen? [Answer: $502/543 = 0.924493\cdots$.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



There are two different ways to arrive at the answer $502/543 = 0.924493\cdots$, in the previous box. On the one hand, we could say that the non-nut flavors are Chocolate, Strawberry, Vanilla, Mint Chocolate Chip, and Salt-Water Taffy. We can then add these simple events to get a complex event.

$$\frac{213}{543} + \frac{51}{543} + \frac{88}{543} + \frac{148}{543} + \frac{2}{543} = \frac{502}{543} = 92.4493\%$$

On the other hand, we know that percentages add to 100% when considering the parts of a whole, and therefore if 7.55064% of surveyed students choose something with nuts, then surely

$$100\% - 7.55064\% \approx 92.4493\%$$

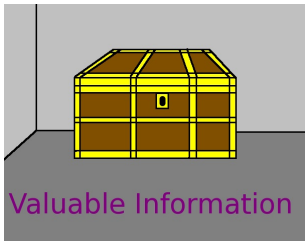
choose something without nuts.



Let's look again at the work of the previous box.

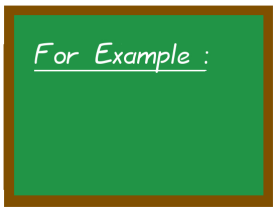
As you can see, if we had already known the value 7.55064%, then the second option is a much faster way of arriving at 92.4493%. We had to do fer mathematical operations to reach that point.

This technique is extremely common in probability. If we work with percentages, we will subtract a probability from 100%. If we work with decimals or fractions, we will subtract a probability from 1. In fact, this technique is so common, that it has a name. We'll present that in the next box.



If p is the probability of an event happening, then $1 - p$ is the probability of an event not happening.

This is called *the complement principle* of probability. It is our third main strategy for common probability problems. (The other two strategies were given on Page 252 of this module.)



3-2-18

On Page 253 we had a problem about people who read the newspaper "always," "sometimes," and "never." We computed the following probabilities:

- $Pr\{Always\} = 68/487 = 0.139630\cdots$
- $Pr\{Sometimes\} = 206/487 = 0.422997\cdots$
- $Pr\{Never\} = 213/487 = 0.437371\cdots$

Several boxes after that, we computed the probability of someone reading the newspaper "sometimes or less often." This is the same as "not always." Therefore, we could have computed

$$1 - \frac{68}{487} = \frac{487}{487} - \frac{68}{487} = \frac{419}{487}$$

Similarly, we were asked about the probability of someone reading the newspaper "sometimes or more often." That's the same as "not never." Therefore, we could have computed

$$1 - \frac{213}{487} = \frac{487}{487} - \frac{213}{487} = \frac{274}{487}$$

though I do confess that I find the phrase "not never" to be extremely awkward phrasing.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

For Example :

3-2-19

Returning to the ice cream survey (from Page 258), what is the probability of a random student in the survey *not* choosing strawberry?

The data says that 51 students out of 543 students have chosen strawberry. Therefore, the probability of a student not choosing strawberry is computed by

$$1 - \frac{51}{543} = \frac{492}{543} = 0.906077\cdots$$

Isn't that much easier than calculating the following?

$$\frac{213}{543} + \frac{88}{543} + \frac{23}{543} + \frac{18}{543} + \frac{148}{543} + \frac{2}{543} = \frac{492}{543}$$

That is why the use of the complement principle is so popular. It's just a huge timesaver.

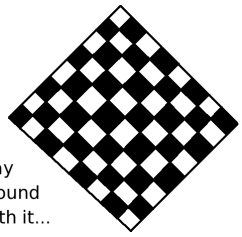
For Example :

3-2-20

I'd like to pose one last follow up question about the ice-cream survey of the last few boxes. If someone asked you to take the data, and form a probability distribution, you have to know what that means. It means you're going to list all the outcomes in the sample space along with their assigned probabilities. In this case, it comes out as

Chocolate	$0.392265\cdots$	Strawberry	$0.0939226\cdots$
Vanilla	$0.162062\cdots$	Mint Chocolate Chip	$0.272559\cdots$
Pistachio	$0.0423572\cdots$	Salt-Water Taffy	$0.00368324\cdots$
Banana Nut	$0.0331491\cdots$		

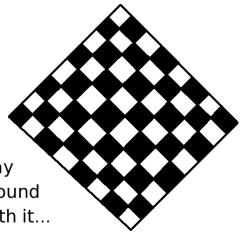
We don't really need six-significant figures here, but that's the standard used in this textbook. By the way, these numbers were obtained by taking the original entries from Page 258 and dividing them by the total number of survey respondents, 543.



3-2-21

Let's look back at the checkerboard problem about the world's population and the ten most populous countries (see Page 257 of this module). Let's say that you are helping a younger sibling with their homework, and that they are writing a report about the world's population. Your younger sibling wants to know what percentage of people *do not* live in one of the top ten most populous countries. In other words, what percentage of the world's population live in the other $238 - 10 = 228$ countries? Must you now add up 228 numbers? That would be very tedious, because most of the populations would be in the millions, so you'd have to make a lot of button presses on your calculator.

The answer will be given on Page 274.



3-2-22

In Australian rules football, draws or ties are much more common than they are in some other sports. Suppose Ian's favorite team has had 7 wins and 2 draws in the last twenty games. Ian is thinking of betting on the next game, and therefore he asks his friend Cooper to estimate the probability that this team will lose their next game. What does the data predict for that probability? (Note: I am technically obligated to add that the skill level of the next scheduled opposing team does not appear to be any higher nor any lower than the skill levels of the teams already played.)

[Answer: 55%.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



The previous box has some very interesting points in it. First, we made the tacit assumption that the reader knows that three outcomes are a win, a draw, and a loss. This is probably common knowledge. Most students would either compute $20-7-2=11$ losses and compute $11/20=0.55=55\%$, or compute

$$Pr\{\text{Loss}\}=1-Pr\{\text{Win}\}-Pr\{\text{Draw}\}=1-\frac{7}{20}-\frac{2}{20}=\frac{11}{20}=0.55=55\%$$

and clearly the same answer is obtained either way. That's kind of cute, because computing $20-7-2=11$ is using the complement operation from set theory, and the other calculation uses the complement principle of probability.

Second, I'd like to share with you a common mistake that I've seen during this problem, particularly when I used to teach business students. We'll discuss that in the next box.



A very frequent answer to the previous checkerboard box is to claim

$$\frac{2}{7}=0.285714\cdots=28.5714\%$$

This is startling because it is wrong. It is also startling for another reason. The team clearly isn't doing well. The team has lost the majority of the games. Surely the answer should be more than 50%, right?

Unfortunately, most people don't think that way. Many students, especially business students in my experience, will simply perform some sequence of operations on the given numbers and hope for success, or at least "partial credit." That's not the path to success.

An exacerbating factor is that "twenty" was written using letters, and not as "20." For some students, this makes the quantity invisible.



It is really very important to read exactly what has been written in a problem about probability. The slightest change of a few words can completely change a problem, as I also mentioned earlier on Page 251.

This is not unlike the story of "The Man who was Hanged by a Comma," summarized on Page 99 of the module "Intermediate Venn Diagram Problems."



Before we move on, I'd like to address yet another point about the Australian rules football question. Why was it essential that I add the following note: "I am technically obligated to add that the skill level of the next scheduled opposing team does not appear to be any higher nor any lower than the skill levels of the teams already played."

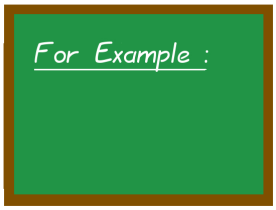
When we use past data to predict future behavior, we are making the assumption that the past and the future are related. This is often the case, but not always. There can also be hidden variables. For example, if the past 20 opposing teams were very good, and the next scheduled team is very unskilled, then that would change the probability of a loss. Likewise, if the past 20 opposing teams were significantly below average, and a top team is scheduled next, the probability of a loss would be underestimated by this calculation.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



Continuing with the question about Australian rules football, other hidden variables could include something like a sports injury. If a member of the team is unable to play because of an injury, then the team composition has changed. In many ways, it isn't even the same team anymore.

Another way to have resolved that issue is to say, "Ian has recorded all of the previous 20 games, is going to pick one at random, to watch with his friend Cooper. What is the probability that Ian picks a game that was a loss, by coincidence?" Then we are choosing uniformly from the set of the previous 20 games, and the equally likely assumption applies. Here, the probability would be rock-solid reliable.

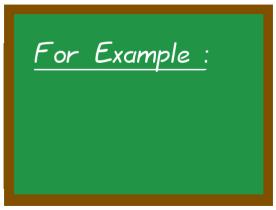


3-2-23

Cameron has the misfortune of having five tests this week, one on each of Monday, Tuesday, Wednesday, Thursday, and Friday. On Thursday evening, he tells you that out of the four tests he's had this week, unfortunately he failed three of them. What is the probability that Cameron will fail Friday's test?

Many students would simply answer 75%. This is a delightful example of a broken problem. There is no reason to believe that Cameron's performance on the first four tests will predict his success on the fifth test. For the academic systems in most countries, tests so close together would almost surely be in different courses. If he's an engineering major, and passed *Calculus* on Monday, failed *Art History* on Tuesday, failed *Caribbean Literature* on Wednesday, and failed *Music Appreciation* on Thursday, then we simply can't assume that he's likely to fail *Physics* on Friday. Since the problem didn't tell us what subjects Cameron failed, and what subject will be on Friday's test, we simply have no way of knowing the probability that Cameron will fail on Friday.

Nonetheless, I would also not recommend betting cash on his success.



3-2-24

This problem was suggested by Prof. Mark Fenton. Suppose a father and young son attend a sporting event together, and there is a raffle for a football helmet worn by a famous athlete. The young son really wants the helmet, and (as you might guess) the father will be buying raffle tickets. It turns out that the drawing is coming up soon. The raffle-ticket salesman tells the dad that he can purchase as many tickets as he likes, but then those will be the last tickets sold—he's going to the stage right away, so that the winning ticket can be drawn. The raffle-ticket salesman also mentions that only 35 tickets have been sold. How many raffle tickets would the father have to buy in order to have a 60% chance of winning? 70%? 80%? 90%?

The key to this problem is to realize that if the father buys x tickets, then $35 + x$ tickets have been bought. The set of tickets bought comprise the interesting set, and we're concerned with the subset of tickets that were bought by the dad—there are x of those. Therefore, for any x, the probability that the father wins will be given by

$$f(x) = \frac{x}{35 + x}$$

which we can use to find the sample probabilities. We'll continue in the next box.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

Let's consider how many tickets the father would have to buy, in order to cause a 60% probability of success.

$$\begin{aligned} f(x) &\geq 0.6 \\ \frac{x}{35+x} &\geq 0.6 \\ x &\geq (0.6)(35+x) \\ x &\geq 21 + 0.6x \\ 0.4x &\geq 21 \\ x &\geq 21/0.4 \\ x &\geq 52.5 \end{aligned}$$

Of course, since he cannot buy half a ticket, he would have to buy 53 tickets to obtain a 60% chance of success.



Before we continue, I should mention a technicality. We only have the right to multiply both sides by $35+x$ in an inequality, if we can be certain that $35+x$ is always positive.

Note, it is impossible for the father to buy a negative number of raffle tickets. Since $x \geq 0$, adding 35 to both sides tells us that $35+x \geq 35$. In plain English, if the father buys zero tickets, then $35+x$ is positive; if he buys some non-zero number of tickets, $35+x$ gets bigger, so it certainly stays positive. Therefore, $35+x$ is always positive for this problem, and we were justified in what we did.

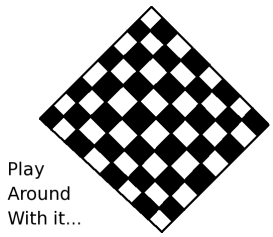
Admittedly, most students would not notice this, but would solve the problem correctly anyhow.



Let's check our work from the previous example. If we've solved the problem correctly, then 52 tickets purchased should result in a probability just under 60%, and 53 tickets purchased should result in a probability just over 60%. Let's see if that's true.

- $f(52) = \frac{52}{35+52} = \frac{52}{87} = 0.597701\cdots$
- $f(53) = \frac{53}{35+53} = \frac{53}{88} = 0.6022\overline{727}$

Alright, now we are certain that we've successfully solved for 60%. I will let you solve for 70%, 80%, and 90% yourself, in the next box.



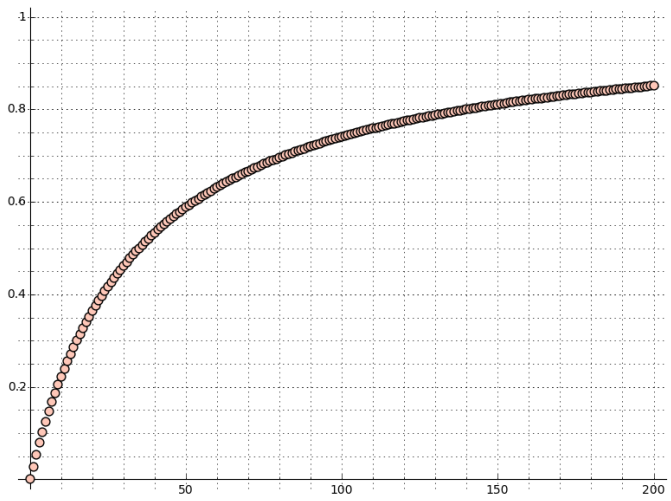
3-2-25

Continuing with the previous example, how many tickets would the father have to buy in order to win...

- ...with probability 70%? [Answer: 82 tickets.]
- ...with probability 80%? [Answer: 140 tickets.]
- ...with probability 90%? [Answer: 315 tickets.]
- ...with probability 99%? [Answer: 3465 tickets.]

It is shocking how many tickets must be purchased in order to guarantee 99% success. Luckily, it is far easier to guarantee 90% or 80% success. Let's explore this in more detail, in the next box.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



The graph at the left shows our function from the previous example:

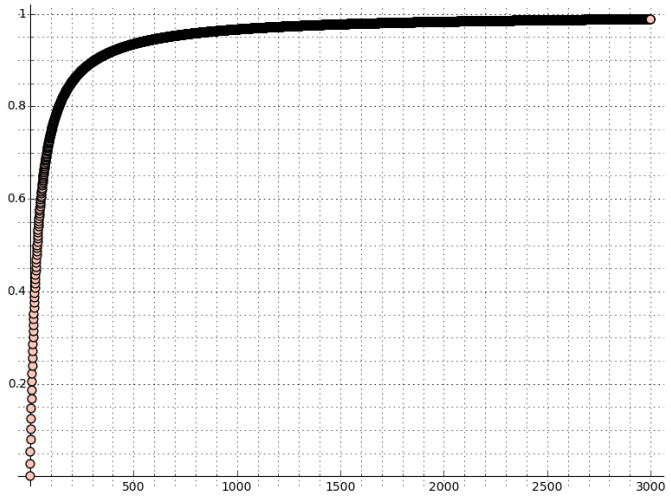
$$f(x) = \frac{x}{35 + x}$$

As you can see, the function grows rapidly for the first bunch of tickets, but it eventually levels out. Since only integer values of x make sense, the function is plotted as a collection of dots representing

$$x \in \{0, 1, 2, 3, \ldots\}$$

instead of a smooth curve, representing all real values of x for some interval. In this case, our plot represents all integers between 0 and 200, inclusive.

It can be informative to see where the function crosses the $y = 0.7$ and $y = 0.8$ lines, and see that they appear to be at approximately $x = 82$ or $x = 140$, as you calculated in the previous box. Moreover, you can see that the function becomes very flat after about 100 tickets. Let's explore that more in the next box.



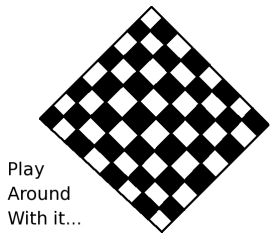
As you saw in the previous box, the graph of the function levels out after about 100 or 200 tickets. That's why we could reach 80% probability with only 140 tickets, but required 315 tickets to reach 90% probability, and a whopping 3465 tickets to reach 99% probability.

On the left, I've redrawn the plot to include all integers between 0 and 3000, inclusive. As you can see, the function is clearly leveling out. The probability will approach 1, getting closer and closer, but it will never reach 1. This behavior is an example of an *asymptote*.

One neat way to see that the probability will never reach one is to compute what happens if one million tickets are bought. In that case, we have

$$f(1,000,000) = \frac{1,000,000}{1,000,000 + 35} = \frac{1,000,000}{1,000,035} = 0.999965001\cdots$$

which is really close to 1, but not equal to 1.

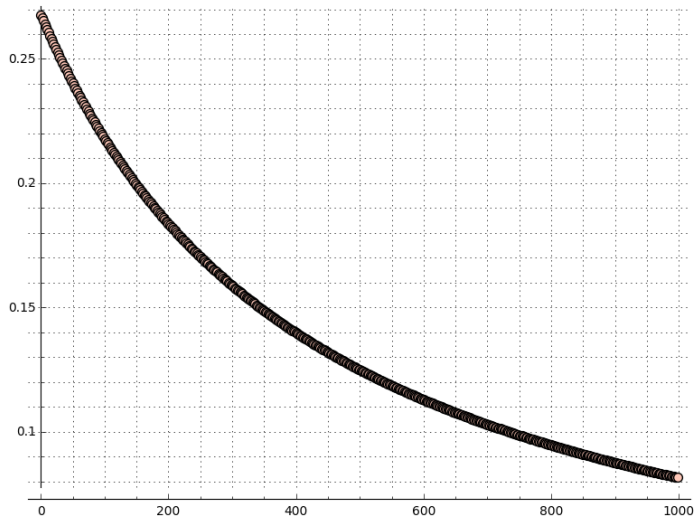


3-2-26

Let's imagine that a recent production run of 437 tires had 117 turn out to be defective due to something going horribly wrong in the manufacturing process. An inspector is coming, and the shift manager is going to take tires, known to be good, from the warehouse and add them to the 437 tires from the recent production run. The inspector will take 1 tire at random, and inspect it. What is the probability $p(x)$ that that the inspector gets a defective tire, if x tires get moved?

[Answer: $p(x) = 117/(437 + x)$.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



Here you can see a plot of all possible values of x in the range $0 \leq x \leq 1000$, though I only plotted only the even values of x. As you can see, the probability decays. As the manager adds more tires that are known to be good, the probability of the inspector picking a defective dire decreases.

However, the manager will have to add an extraordinarily large number of tires to get the probability to become very small. The manager will have to add $x = 343$ tires just to bring the probability down to 15%.

$$p(343) = \frac{117}{437 + 343} = \frac{117}{780} = 15\%$$

For Example :

3-2-27

According to "Table 163: Osteopathic Physicians" of *The Statistical Abstract of the United States*, 131st Edition (2012–2013), in the year 2001 there were 46,962 osteopaths in the United States. Furthermore, they have the following age structure:

- 4838 were age 65 and over.
- 9544 were age 55 and over.
- 22,298 were age 45 and over.
- 37,096 were age 35 and over.

What is desired is a probability distribution of the ages of osteopaths. The outcomes are "65 and over," "age 55–64," "age 45–54," "age 35–44," and "below age 35." We certainly can't just divide the above numbers by 46,962, because that won't add up to 100%. Moreover, an osteopath who is age 66 will be counted in all four of those categories, violating mutual exclusivity.

Instead, we need only perform some subtractions. We'll do those in the next box.

Continuing with the previous box, first we are going to find out how many osteopaths are in each age bracket.

- There are $46,962 - 37,096 = 9866$ osteopaths who are under age 35.
- There are $37,096 - 22,298 = 14,798$ osteopaths who are 35–44.
- There are $22,298 - 9544 = 12,754$ osteopaths who are 45–54.
- There are $9544 - 4838 = 4706$ osteopaths who are 55–64.
- We already know that there are 4838 osteopaths who are 65 or older.

Then it is an easy matter to covert these to percentages, using division. We'll do that in the next box.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

Continuing with the previous two boxes, we obtain the following:

- Under age 35 $= 9866/46,962 = 21.0084\cdots\%$.
- Age 35–44 $= 14,798/46,962 = 31.5105\cdots\%$.
- Age 45–54 $= 12,754/46,962 = 27.1581\cdots\%$.
- Age 55–64 $= 4706/46,962 = 10.0208\cdots\%$.
- Age 65 and over $= 4838/46,962 = 10.3019\cdots\%$.

Those five bullets above assign probabilities to five events. Those five outcomes are mutually exclusive and collectively exhaustive, and thus form a sample space. The set of outcomes, taken with their assigned probabilities, form a probability distribution.



Now we can check the work of the previous two boxes with some sums.

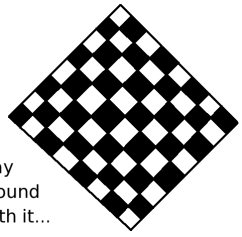
- $4838 + 4706 + 12,754 + 14,798 + 9866 = 46,962$ osteopaths are in the data set. ✓
- $4838 + 4706 + 12,754 + 14,798 = 37,096$ osteopaths are age 35 and over. ✓
- $4838 + 4706 + 12,754 = 22,298$ osteopaths are age 45 and over. ✓
- $4838 + 4706 = 9544$ osteopaths are age 55 and over. ✓

It is probably a good idea to check that the percentages get very close to 100%.

$$21.0084\% + 31.5105\% + 27.1581\% + 10.0208\% + 10.3019\% = 99.9997\% \checkmark$$

and surely it is clear that the last bit there is due to rounding error.

It looks like we got this one correct!



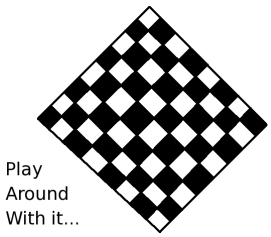
3-2-28

Also according to "Table 163: Osteopathic Physicians" of *The Statistical Abstract of the United States*, 131st Edition (2012–2013), in the year 2010 there were 70,068 osteopaths in the United States. That's an impressive increase in only nine years, wouldn't you say? Furthermore, they have the following age structure:

- 6528 were age 65 and over.
- 17,723 were age 55 and over.
- 33,673 were age 45 and over.
- 53,791 were age 35 and over.

Figure out how many osteopaths are in each age band, and then construct a probability distribution for the age band of a randomly selected osteopath. The answers will be given on Page 274.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



3-2-29

The Pew Research Center did a study, published in the summer of 2015, on how Americans get their news. A portion of the research was studying how Americans use Twitter for news. A sample of 176 Twitter users were voluntarily monitored. For each user, four randomly selected weeks between August 2014 and February 2015 were analyzed. The tweets were read, categorized, and tweets that were not about one's friends and family, but about other events and issues, were tabulated. Here is a table that summarizes their data.

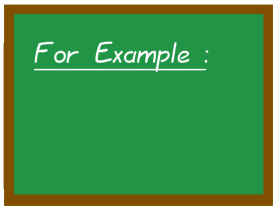
Never	28%	At least 50 times	18%
At least once	72%	At least 100 times	12%
At least 10 times	39%	At least 200 times	5%
At least 25 times	26%		

The questions follow in the next box.

Here are some questions about the data in the previous box.

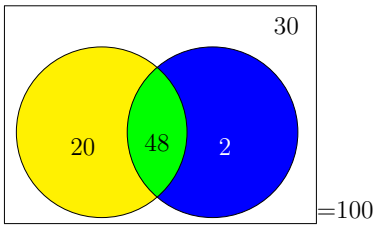
- Hint: It is easier to do this problem by first constructing a probability distribution.
- What percentage of people sent at least one tabulated tweet?
- What percentage of people sent between 10 and 24 tabulated tweets?
- What percentage of people sent between 50 and 199 tabulated tweets?

The answers will be given on Page 275. You can read an article about this data, by Michael Barthel and Elisa Shearer, called "How do Americans use Twitter for News?", posted to the blog "`KnightBlog`" on August 19th, 2015.



3-2-30

On Page 60 of the module "Basic Venn Diagram Problems," we studied a problem about a survey of graduating seniors and their employment. It was summarized by the following Venn Diagram. The left circle represent students who've had internships, and the right circle represent students with a job lined up for after graduation.

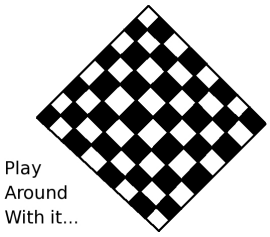


Based on the above, we can ask questions like "What is the probability that a student who has done an internship has a job lined up for after graduation?" and "What is the probability that a student who has not done an internship has a job lined up for after graduation?"

There are 68 students with internships, and 48 of them have a job lined up after graduation. Therefore, the probability that a student who had an internship has a job lined up is $48/68 = 70.5882\cdots\%$.

We will leave the others for you to find, in the next box.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

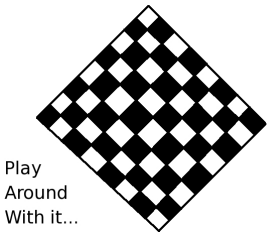


3-2-31

Continuing with the previous box, ...

Hint: For the next two questions, the "interesting set" is the set of students who did not have an internship.

- What is the probability that a student who did not have an internship has a job lined up? [Answer: $2/32 = 6.25\%$.]
- What is the probability that a student who did not have an internship has no job lined up? [Answer: $30/32 = 93.75\%$.]

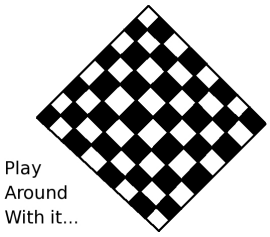


3-2-32

Continuing with the previous two boxes, ...

Hint: For the next two questions, the "interesting set" is the set of students who have a job lined up.

- What is the probability that a student who has a job lined up had an internship? [Answer: $48/50 = 96\%$.]
- What is the probability that a student who has a job lined up did not have an internship? [Answer: $2/50 = 4\%$.]

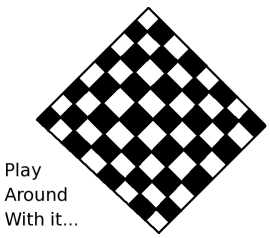


3-2-33

Continuing with the previous three boxes, ...

Hint: For the next two questions, the "interesting set" is the set of students who did have an internship.

- What is the probability that a student who did have an internship has no job lined up? [Answer: $20/68 = 29.4117\cdots\%$.]
- What is the probability that a student who did have an internship has a job lined up? [Answer: $48/68 = 70.5882\cdots\%$.]



3-2-34

Continuing with the previous four boxes, ...

Hint: For the next two questions, the "interesting set" is the set of students who have no job lined up.

- What is the probability that a student who has no job lined up had an internship? [Answer: $20/50 = 40\%$.]
- What is the probability that a student who has no job lined up did not have an internship? [Answer: $30/50 = 60\%$.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



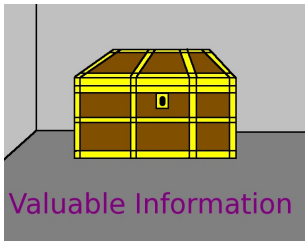
At first, it might look as though the answers in the previous two boxes are violations of the complement principle. After all,

- What is the probability that a student who did have an internship has no job lined up? $29.4117\cdots\%$
- What is the probability that a student who did not have an internship has no job lined up? $30/32 = 93.75\%$
- Yet, $29.4117\cdots\% + 93.75\% = 123.161\cdots\% \neq 100\%$.

Similarly, here is another example,

- What is the probability that a student who has a job lined up had an internship? 96%
- What is the probability that a student who has no job lined up had an internship? 40%
- Yet, $96\% + 40\% = 136\% \neq 100\%$.

On the other hand, if you look at the pairs of answers (in the way that I had paired them up inside the checkerboard boxes), each of the four pairs individually adds to 100%. So the complement principle did work on those four occasions.



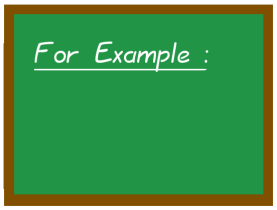
The reason that the complement principle didn't work in the two situations of the previous box is that it does not apply in the first place.

- If you are using my technique (an interesting set with a very interesting subset), and two probabilities have the same interesting set, but complementary very interesting subsets, then the complement principle applies. The probabilities will add to 100%. You will get the right answer if you use the complement principle for this situation.
- If you are using my technique (an interesting set with a very interesting subset), and two probabilities have the same very interesting subsets, but complementary interesting sets, then the complement principle does not apply. The probabilities will (in most cases) not add to 100%. You will (in most cases) get the wrong answer if you use the complement principle for this situation. In particular, we got $123.161\cdots\%$ and 136% in the previous box.

The discussion in the previous two boxes is extremely important. Please study the previous two boxes very carefully. Perhaps when this module is completed, you can investigate pairs of probabilities from the vaccine problem, near the start of this module, and see if you understand why some pairs add to 100%, but some pairs do not add to 100%.

A professor in Louisiana gave me the following problem that had appeared on one of his tests. He wishes to keep his name and his university's name out of my textbook, which I shall respect. I have updated the numbers to reflect current populations.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



3-2-35

Consider the following:

- 1 out of every 3500 babies born in the United States contracts neonatal herpes.
- In Louisiana, 1 out of every 2500 babies contracts neonatal herpes.
- The population of Louisiana is 4,682,000.
- The population of the USA is 323,127,000.
- What is the probability that a baby *not* born in Louisiana (but born in the USA) contracts neonatal herpes?

Several students saw the Louisiana rate "1 out of every 2500 babies," and the phrase "a baby *not* born in Louisiana" and used the complement principle. They wrote

$$\text{WRONG!} \rightarrow \qquad 1 - \frac{1}{2500} = \frac{2499}{2500} \qquad \leftarrow \text{WRONG!}$$

which is really horrible.

This handful of incorrect students were claiming that 2499/2500 babies (or 99.96%) born in USA but not Louisiana, were born with neonatal herpes. Really?! 99.96%?!



Here's another way to look at this distinction of when the complement principle does, or does not, apply. We have no reason to believe that the following two probabilities will add to 100%.

- The probability that a random person born in Louisiana is born with neonatal herpes.
- The probability that a random person *not* born in Louisiana is born with neonatal herpes.
- Let's be honest, both probabilities are going to be very small.

However, we have every reason to believe that the following two probabilities will add to 100%, because of the complement principle.

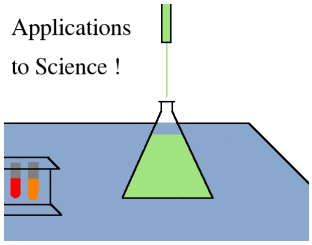
- The probability that a random person born in Louisiana is born with neonatal herpes.
- The probability that a random person born in Louisiana is *not born* with neonatal herpes.

This is because our interesting set in the second pair is the set of people born in Louisiana, for both questions. The very interesting subsets (in the second pair) are those who are born with neonatal herpes, and those who are not born with neonatal herpes.

In stark contrast to this, the interesting sets of the first pair are not the same. Instead, they are the set of people born in Louisiana, and the set of people not born in Louisiana.

By the way, the correct solution to the neonatal herpes problem requires a technique that I didn't teach you. I include the solution at the end of this module (on Page 275), only for completeness. The purpose of this problem was to alert you to a common misconception surrounding the complement principle.

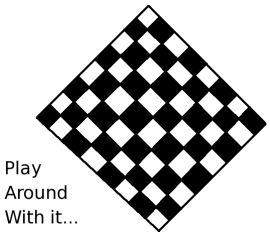
COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



In case you are curious, according to the website `childrenshospital.org`, about 1 out of every 3,500 babies born in the United States contracts neonatal herpes, usually from the mother's birth canal. However, treatments are available with high success rates. You can read more about this, below.

`https://en.wikipedia.org/wiki/Neonatal_herpes_simplex`

`http://www.childrenshospital.org/conditions-and-treatments/conditions/neonatal-herpes-simplex`

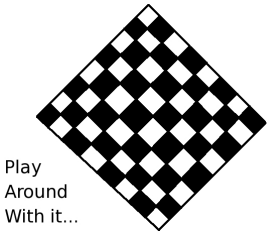


3-2-36

Let's try another survey-analysis problem. You might want to look back to the previous example about college seniors, on Page 267 of this module. I hope that you can do this problem. However, if you have trouble, then take heart: I've made the solutions a bit more detailed than typical, because this problem is challenging.

As it comes to pass, Gallup Poll surveyed 1021 Americans during July 5-9, 2017, about whether they use marijuana. The exact question was "Keeping in mind that all of your answers in this survey are confidential, do you, yourself, smoke marijuana?" For males, there were 740 who said no, and 111 who said yes. For females, there were 12 who said yes, and 158 who said no. I have some probability questions for you in the next two boxes, but first it might be good to summarize this data with a Venn Diagram. Let the left circle be survey respondents who smoke marijuana, and the right circle be males.

The solution is given on Page 276.

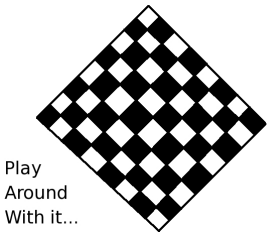


3-2-37

Now that you have a Venn Diagram for the data of the previous box, answer the following:

- I'd like the answer to be a percentage with six significant figures.
- What is the probability that a random male survey respondent smokes marijuana?
- What is the probability that a random female survey respondent smokes marijuana?
- What is the probability that a random survey respondent smokes marijuana?
- What is the probability that a random male survey respondent doesn't smoke marijuana?

The answers are given on Page 276 of this module.



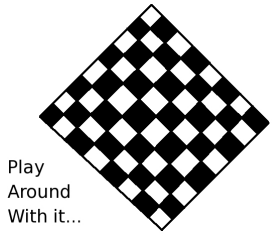
3-2-38

Continuing with the previous two boxes, answer the following:

- I'd like the answer to be a percentage with six significant figures.
- What is the probability that a random female survey respondent doesn't smoke marijuana?
- What is the probability that a random survey respondent doesn't smoke marijuana?
- What is the probability that a random marijuana-smoking survey respondent is male?
- What is the probability that a random marijuana-smoking survey respondent is female?

The answers are given on Page 276 of this module.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

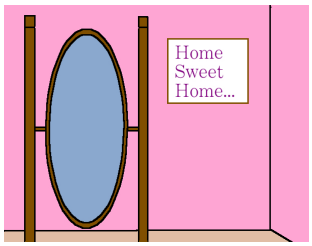


3-2-39

Continuing with the previous three boxes, answer the following:

- I'd like the answer to be a percentage with six significant figures.
- What is the probability that a random non-marijuana-smoking survey respondent is male?
- What is the probability that a random non-marijuana-smoking survey respondent is female?
- What is the probability that a random survey respondent is male?
- What is the probability that a random survey respondent is female?

The answers are given on Page 277 of this module.

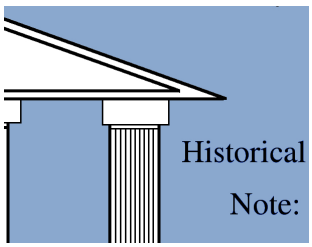


A Pause for Reflection...

An interesting point about the previous three boxes is that marijuana smoking, in most but not all US states, is illegal. With that in mind, some survey respondents might have lied. How can we know how many lied?

Also, some people who were being surveyed might have hung up the phone rather than complete the survey. What if the person being contacted was a college student, with conservative parents within ear-shot? Last but not least, it completely overlooks the use of brownies (and other pastries) baked with marijuana.

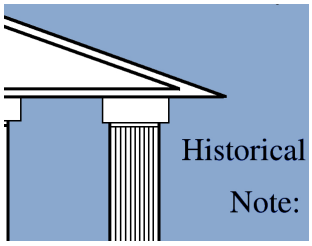
Research data compiled from survey responses should always be taken "with a grain of salt." Too many things can go wrong.



As we've seen here, probability has many practical applications. One of the most important can be the analysis of survey data, to explain what is going on in the population around us, whether it might be pollution, crime, disease, income, education, lifespan, or opinions. The etymology of the word *statistics*, discussed over the next two boxes, shows that heritage.

During the Renaissance, the arbitrary rule of a noble overlord, with or without the consultation of his immediate vassals (his "tenants in chief" or *tenens in capite*), was replaced by the advice of a panel of educated advisors. The name for such a council of state (which we might call "The Cabinet" today) was *statisticum collegium*. Members of the *statisticum collegium* were called *statista*.

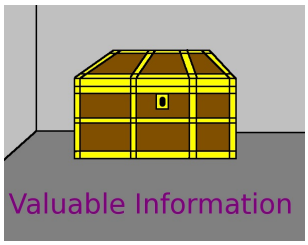
In 1749, Gottfried Achenwall (1719–1772) published a book about the analysis of data and its role in governing, and he introduced the word Statistik into German. Relatively soon after, in 1791, John Sinclair (1754–1835) published the 21-volume *Statistical Account of Scotland* which is not unlike our census. That massive work introduced the word "statistics" into English.



While the word was not found in English until 1791, the concept was already familiar to the English government about six generations earlier, but it was called "political arithmetic." William Petty (1623–1687) was an advisor to Oliver Cromwell (1599–1658), the dictator who took power in England after Charles I (1600–1649) was beheaded.

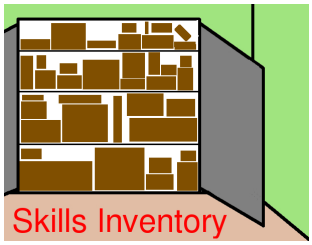
William Petty introduced techniques that we would now call statistics and economics into government. He remained a prominent advisor when the monarchy was restored in 1660 under Charles II (1630–1685). William Petty even advised Charles II's successor, James II (1633–1701), but he died before James II was overthrown in 1688.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



As promised, I now present you with the top six probability mistakes that I have found, looking at the exams and projects of students. (These are in order of most frequent to least frequent.)

1. The most common one is adding probabilities of events when that is not possible. The probabilities of outcomes (simple events) can be added. The probabilities of compound events cannot be added, in general. Recall, a compound event is an event that consists of more than one outcome, whereas a simple event has only one outcome in it. This pitfall is so dangerous that an entire module is dedicated to discussing this matter. That module is titled "You Can't Just Add Probabilities!"
2. Making the "equally likely assumption" when it does not apply.
3. The third error is hard to explain, but it deals with assuming events will be independent when they are not. This will be explained in Module 5.2: "Independence and Repetition."
4. Using the complement principle, when it does not apply. (Think about our example involving neonatal herpes.)
5. Using a sample population that is not representative of the population that matters. This is not so much from exams, but from particularly badly designed student projects.
6. Attempting to use a set that is not mutually exclusive, or that is not collectively exhaustive, as if it were a sample space. This is the least common one, but it is definitely not rare.



Here is a summary of what we have learned in this module.

- We tested ourselves by looking at a survey about political parties and vaccines. Then we had 16 questions to answer about that survey.
- We learned three fundamental strategies for basic probability problems, and we practiced those strategies with several problems. We can compute the probability of an event by . . .
  - . . . adding together the probabilities of the simple events (outcomes) in a sample space that comprise a given compound event.
  - . . . dividing, with the size of some interesting set in the denominator, and some very interesting subset in the numerator, (but only in the cases when the "equally likely assumption" applies).
  - . . . the complement principle, but only under certain conditions.
- We saw some discussion of common mistakes in probability.
- We learned how to convert survey data into a probability distribution.
- We saw how probability can sometimes be a function, and not a number.
- We analyzed several data sets, and computed probabilities from those data sets.
- We learned when the complement principle applies, and when illegally using it will produce a wrong answer.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.

I hope at this point that you've understood most of the module. If you did not get the 16 questions about the vaccine survey correct, then please retry those at this time. You'll find them on Page 250. As you do them, try to keep in mind the questions:

- What is the interesting set? (This will be your denominator.)
- What subset am I looking for? (This will be your numerator.)

The module is now complete. Here are a few answers to questions posed earlier in the reading.



Back on Page 256, we had a problem about CIO of a large multinational company inspecting 17 offices and 4 data centers in the USA.

The Chief Information Officer (CIO) is going to inspect one location, choosing a facility at random. As it turns out, 11 out of 17 offices have installed a UPS. Moreover, each of the data centers has installed a UPS. What is the probability that the location which the CIO inspects has installed a UPS?

The answer is given by

$$\frac{11+4}{17+4} = \frac{15}{21} = \frac{5}{7} = 0.714285\cdots = 71.4285\cdots\%$$



On Page 260, you were asked to find the probability that a random person in the world was not from one of the top ten most populous countries.

Of course, we do not have to add up all those 228 population estimates. We can use the complement principle of probability, and simply calculate

$$100\% - 58.1514\% = 41.8486\%$$

if we already have the number 58.1514%.

If we don't have the 58.1514% already computed, then we can compute

$$1 - \frac{\text{sum of the top ten}}{\text{population of the world}} = 1 - \frac{4,219,752,402}{7,256,490,011} = 1 - 0.581514\cdots = 0.418486\cdots = 41.8486\%$$



Here are the answers for converting the osteopath data from Page 266 into a probability distribution. First, the actual counts of the number of osteopaths by age band.

- There are 16,277 osteopaths that are under age 35.
- There are 20,118 osteopaths that are 35–44.
- There are 15,950 osteopaths that are 45–54.
- There are 11,195 osteopaths that are 55–64.
- There are 6528 osteopaths that are 65 or older.

We will continue in the next box.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



Continuing with the previous box (about the osteopaths), as percentages we have the following:

- Under age 35: $23.2302\cdots\%$.
- Age 35–44: $28.7121\cdots\%$.
- Age 45–54: $22.7636\cdots\%$.
- Age 55–64: $15.9773\cdots\%$.
- Age 65 or over: $9.31666\cdots\%$.

I'd like to note something. It is particularly odd to report the last few decimal places. The sixth significant figure represents much less than 1 person in this case. However, I have kept six significant figures in this problem because that is our standard throughout the textbook.



These are the answers about Twitter usage, taken from the question on Page 267.

- What percentage of people sent at least one tabulated tweet? [Answer: 72%.]
- What percentage of people sent between 10 and 24 tabulated tweets? [Answer: 13%.]
- What percentage of people sent between 50 and 199 tabulated tweets? [Answer: 13%.]

Here is the probability distribution. I think it is a lot easier "to see" how to get the answers to the above questions after constructing the probability distribution. However, this is not (strictly speaking) necessary.

0 tabulated tweets	28%	25–49 tabulated tweets	8%
1–9 tabulated tweets	33%	50–99 tabulated tweets	6%
10–24 tabulated tweets	13%	100–199 tabulated tweets	7%
		200 or more tabulated tweets	5%



This is the correct way to solve the neonatal herpes problem from Page 269. However, I am including it only for completeness. I did not teach you this technique.

First, we compute the number of people in the USA, but outside Louisiana by

$$323,127,000 - 4,682,000 = 318,445,000$$

We have to write down the following equation, which is a weighted average

$$\frac{4,682,000}{323,127,000} \cdot \frac{1}{2500} + \frac{318,445,000}{323,127,000} \cdot x = \frac{1}{3500}$$

and then we solve for x. We obtain

$$x = 0.000279918\cdots \approx \frac{1}{3572.46\cdots}$$

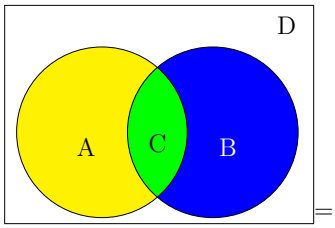
which differs only slightly from the 1/3500 probability nationwide.

This makes sense. Louisiana should not throw off the national average by too much, since it has only 4,682,000 people in it. Due to the errors involved in predicting probabilities, it surely is the case that all the probabilities surrounding neonatal herpes have very large uncertainties, because it is a rare event. We will learn more about that on Page 387 of the module "The Square Root of NPQ Rule" when we learn about $\hat{p}$.

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



Here is the Venn Diagram from the Gallup poll about marijuana, from Page 271.



=1021

Remember, the left circle represents survey respondents who smoke marijuana, and the right circle represents male respondents.



Here are the solutions to the first set of four questions about marijuana from Page 271.

- What is the probability that a random male survey respondent smokes marijuana?
  [Answer: $111/(111+740) = 111/851 = 13.0434\cdots\%$.]
- What is the probability that a random female survey respondent smokes marijuana?
  [Answer: $12/(12+158) = 12/170 = 7.05882\cdots\%$.]
- What is the probability that a random survey respondent smokes marijuana?
  [Answer: $(111+12)/1021 = 123/1021 = 12.0470\cdots\%$.]
- What is the probability that a random male survey respondent doesn't smoke marijuana?
  [Answer: $740/(111+740) = 740/851 = 86.9565\cdots\%$.]



Here are the solutions to the second set of four questions about marijuana from Page 271.

- What is the probability that a random female survey respondent doesn't smoke marijuana?
  [Answer: $158/(12+158) = 158/170 = 92.9411\cdots\%$.]
- What is the probability that a random survey respondent doesn't smoke marijuana?
  [Answer: $(740+158)/1021 = 898/1021 = 87.9529\cdots\%$.]
- What is the probability that a random marijuana-smoking survey respondent is male?
  [Answer: $111/(111+12) = 111/123 = 90.2439\cdots\%$.]
- What is the probability that a random marijuana-smoking survey respondent is female?
  [Answer: $12/(111+12) = 12/123 = 9.75609\cdots\%$.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.



Here are the solutions to the last set of four questions about marijuana from Page 272.

- What is the probability that a random non-marijuana-smoking survey respondent is male?
  [Answer: $740/(740+158) = 740/898 = 82.4053\cdots\%$.]
- What is the probability that a random non-marijuana-smoking survey respondent is female?
  [Answer: $158/(740+158) = 158/898 = 17.5946\cdots\%$.]
- What is the probability that a random survey respondent is male?
  [Answer: $(740+111)/1021 = 851/1021 = 83.3496\cdots\%$.]
- What is the probability that a random survey respondent is female?
  [Answer: $(158+12)/1021 = 170/1021 = 16.6503\cdots\%$.]

COPYRIGHT NOTICE: This is a work in-progress by Prof. Gregory V. Bard, which is intended to be eventually released under the Creative Commons License (specifically agreement # 3 "attribution and non-commercial.") Until such time as the document is completed, however, the author reserves all rights, to ensure that imperfect copies are not widely circulated.