13

How Good Is That Test?

In Chapter **??** we looked at probabilities of independent events — things that had nothing to do with one another. Here we think about probabilities in situations where we expect to see connections, such as in screening tests for diseases or DNA evidence for guilt in a criminal trial.

Chapter goals:

Goal 13.1. Interpret and build two way contingency tables.

Goal 13.2. Understand how to compute probabilities for dependent events.

Goal 13.3. Understand the implications of false positives.

13.1 UMass Boston enrollment

Table **??** summarizes student enrollment at UMass Boston in 2007 by category two ways: graduate/undergraduate and male/female. We can use the data to answer some probability questions about a random student.

- What is the probability that a student chosen at random is an undergraduate?

	Undergraduate	Graduate	Total
Female	5,680	2,388	8,068
Male	4,328	1,037	5,365
Total	10,008	3,425	13,433

Table 13.1. UMass Boston enrollment, 2007

The last row of the table has the numbers we need:

$$\begin{aligned}\frac{\text{number of undergraduates}}{\text{number of students}} &= \frac{10,008}{13,433}\\ &= 0.745\\ &\approx 75\%.\end{aligned}$$

Three quarters of the students are undergraduates.

- What is the probability that a student is female?
  For that computation we use the totals in the last column:

$$\begin{aligned}\frac{\text{number of females}}{\text{number of students}} &= \frac{8,068}{13,433}\\ &= 0.600610437\\ &\approx 60\%.\end{aligned}$$

- What is the probability that a student is a female undergraduate?
  Use the count in the first column of the first row:

$$\begin{aligned}\frac{\text{number of female undergraduates}}{\text{number of students}} &= \frac{5,680}{13,433}\\ &= 0.4228392764\\ &\approx 42\%.\end{aligned}$$

In each of these probability calculations we used the total number of students (13,433) in the denominator.

Continuing ...

- What is the probability that a female student is an undergraduate?
  Since this is a question about the female students, we need a different denominator:

$$\begin{aligned}\frac{\text{number of female undergraduates}}{\text{number of female students}} &= \frac{5,680}{8,068}\\ &= 0.70401586514\\ &\approx 70\%.\end{aligned}$$

- What is the probability that an undergraduate is female?

  That's a different question. This time we know the student is an undergraduate. That calls for a different denominator:

$$\begin{aligned}\frac{\text{number of female undergraduates}}{\text{number of undergraduates}} &= \frac{5,680}{10,008}\\ &= 0.56754596322\\ &\approx 57\%.\end{aligned}$$

The last two questions sound similar, but have different answers, because each begins with a different assumption. In the first we know the student is female and wonder whether she's an undergraduate. In the second, we know that the student is an undergraduate and wonder whether it's a she.

We're not finished thinking about these probabilities. We found that there's a 60% probability that a student is female. But if we know the student is an undergraduate then that probability drops to 57%, because the proportion of women is different for undergraduates than for the student body as a whole. This is not what happened when we thought about a coin and a die in Section **??**. The probability that the die shows a four is the same whether the coin comes up heads or tails. Those events are *independent*. The facts "is female" and "is an undergraduate" are *dependent*. When you know one of them you know something about the probability of the other.

We learned in Section **??** that when events are independent you multiply to compute the probability that both happen:

$$\begin{aligned}\text{probability(coin H and die 4)} &= \text{probability(coin H)} \times \text{probability(die 4)}\\ &= \frac{1}{2} \times \frac{1}{6}\\ &= \frac{1}{12}.\end{aligned}$$

For dependent events that won't work. We found that

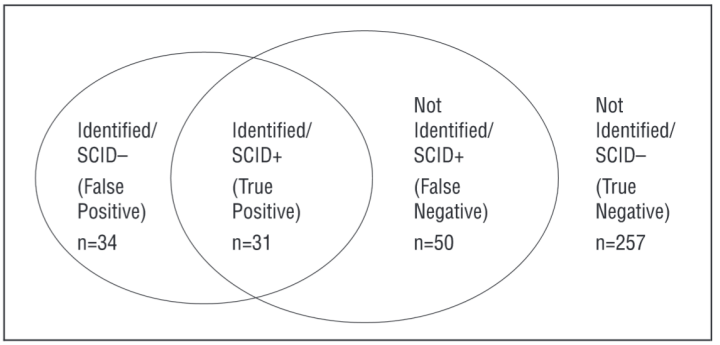
$$\text{probability(female and undergraduate)} = 42\%$$

but

$$\begin{aligned}\text{probability(female)} \times \text{probability(undergraduate)} &= 60\% \times 75\%\\ &= 45\%.\end{aligned}$$

Those answers are close, but not the same, because the proportion of females among the undergraduates is close to but not the same as the proportion among the graduate students.

In the rest of this chapter we will look at the probabilities for dependent events, working with displays like Table **??** in examples where the consequences matter much more than they do here.



Proportions of patients identified as depressed by physicians, Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition *(SCID), both methods, or neither method (N = 372).*

Figure 13.2. Diagnosing depression [R**??**]

13.2 False positives and false negatives

Many women have periodic mammograms to look for breast cancer. Many men have periodic PSA tests to look for prostate cancer. In each there are four possibilities. We'll spell them out for breast cancer.

- *True positive*: a woman has breast cancer and the mammogram says so.
- *True negative*: a woman does not have breast cancer and the mammogram says she doesn't.
- *False positive*: a woman doesn't have breast cancer but the mammogram mistakenly says she does.
- *False negative*: a woman does have breast cancer but the mammogram doesn't detect it.

If the test were perfect there would be no false positives and no false negatives — but there are very few perfect tests. In order to understand what the test results mean you can build a table like the one in the first section of this chapter.

We'll do that with a real example. Figure **??** appeared in the article "False positives, false negatives, and the validity of the diagnosis of major depression in primary care" in the September 1998 *Archives of Family Medicine*.

It summarizes the results of a study of 372 patients who were screened by family physicians for clinical depression.

The numbers in the four categories in the figure are easier to understand when we put them in Table **??**.

		depressed		
		yes	no	total
diagnosed	yes	31	34	65
	no	50	257	307
	total	81	291	372

Table 13.3. Diagnosing depression

		condition	
		present	absent
screened positive	yes	true positive	false positive
	no	false negative	true negative

Figure 13.4. A two way contingency table

Two by two tables like this are called *contingency tables*. Figure **??** shows the standard names for the four cells with raw data: true positive, false positive, false negative and true negative. In this example they have values 31, 34, 50 and 257.

The totals tell us that 65 people in the population of 372 (17%) were diagnosed as depressed, and that 81 (22%) were depressed. Those numbers are pretty close. But does that make it a good test? To answer that question we need to look at the columns separately.

- The first column tells us there were 31 true positives and 50 false negatives from the total of 81 subjects who were in fact depressed. So if a subject was depressed the probability that he or she was diagnosed correctly is only $31/81 \approx 38\%$. That is the true positive rate. There's a 62% chance the condition was missed. That 62% is the false negative rate.

- The second column says that even when a subject was not depressed the chance of a diagnosis of depression was $34/291 = 0.117 \approx 12\%$. That is the false positive rate.

Whether this is a "good" test is a difficult decision.

Although the chance of misdiagnosis of depression when it doesn't exist is fairly low — about 12% — the 62% false negative rate says that test will identify fewer than half the depressed people.

13.3 Screening for a rare disease

A test with a small false positive rate looks like a good candidate for screening large populations for a nasty disease. However, if the disease is rare, the test may not be as good as it looks. In this section we'll study two examples, one made up and one real.

Suppose a drug company has developed a test for the rare disease X. Clinical trials show that the test is 90% accurate at detection, so the false negative rate is 10%. Those trials also show that the false positive rate is only 1%.

These are the important questions:

(1) What is the probability that a person who suffers from X tests positive?

(2) What is the probability that a person who tests positive suffers from X?

If the test were perfect — no false positives, no false negatives — each question would have the same answer: 100%. But the two facts "suffers from X" and "tests positive for X" are not exactly the same. Knowing either one makes the other more likely, but not certain. We want to find out how much more likely in each case.

Question 1 is easy: the drug company's clinical trials found that there is a 90% probability that a person who suffers from X tests positive for X.

Whether that test is as good as it sounds depends in part on the answer to the second question. That answer depends on two things: the false positive rate and the number of people who actually have X. Suppose just one person in every 1,000 suffers from X (one tenth of one percent of the population). Then even though the false positive rate is only 1%, most of the positive results will come from healthy people. We can use a contingency table to find the actual value for "most of".

Since percentages (particularly small percentages) are often confusing, we'll build our table for an imaginary population of 100,000 people that just matches the statistical profile for this test. In a population of 100,000, one out of every 1,000 will have the disease. That's 100 people. Of those 100, 90% (so 90 people) will test positive. The other 10 will be the false negatives. Of the 99,900 healthy people, one percent (999) will test positive. The other 98,901 will be the true negatives. Table **??** shows the contingency table.

		suffers from X		
		yes	no	total
test + for X	yes	90	999	1,089
	no	10	98,901	98,911
	total	100	99,900	100,000

Table 13.5. Screening for disease X

Now we can answer the second question. The probability that someone who tests positive is actually ill with X is only $90/1089 = 8.26\%$.

Is this acceptable? Maybe, maybe not. If the test is inexpensive and there's a second test (perhaps more expensive) that can weed out the false positives, and the disease can be treated successfully if detected, perhaps the screening is a good idea. If all the people who test positive must undergo expensive painful unreliable treatment, which would be unnecessary for more than 90% of them, then the screening is probably a bad investment of scarce health care resources.

For a real application of this technique to the statistics of screening for breast cancer, work Exercise **??** .

13.4 Trisomy 18

In this section we'll work through the numbers when considering whether to call for routine prenatal screening for the rare birth defect trisomy 18.

> K Spencer and colleagues' claims for prenatal detection of trisomy 18 by measurement of maternal serum (alpha) fetoprotein and free β human chorionic gonadotrophin concentrations are impressive. Detection of 50% of cases for a false positive rate of only 1% seems to compare favourably with the detection rate for Down's syndrome when similar techniques are used, which is 70% for a false positive rate of 5%. Unfortunately, the authors fail to emphasise the importance of the relative incidence of the two conditions at birth before concluding that screening for trisomy 18 should be introduced. [R**??**]

Davies notes that there are about 12.6 instances of Down's syndrome per 10,000 births. The incidence for trisomy 18 is just 1.3 per 10,000 births.

A positive test leads to a second procedure, an amniocentesis to check whether the positive is true or false. Davies calculates that screening 10,000 pregnant women for Down's syndrome "would result in 8.8 cases being detected at the cost of 500 amniocenteses (5% of 10,000). This means that one case of Down's syndrome is detected for every 57 amniocenteses performed." For trisomy 18 his figures are 0.65 cases detected, 100 amniocenteses, so 154 amniocenteses to detect one true case.

Let's check his arithmetic for trisomy 18. To build the contingency table we need three numbers:

- The false negative rate. Since the test detects 50% of cases the other 50% are the false negatives.
- The false positive rate. It's only 1%.
- The incidence rate. The second paragraph tells us it's 1.3 per 10,000 births.

Figure **??** shows a screenshot of the spreadsheet `ContingencyTable.xlsx` with entries for this problem. Cell `B12` is named `INCIDENCE`; it contains the formula `=1.3/10000`, formatted as a percent. Cell `B17` for the number of true positive results contains the formula

```
=POPULATION*INCIDENCE*(1-FALSENEG)
```

which in this example is

$$10,000 \times \frac{1.3}{10,000} \times (1-0.5) = 0.65,$$

Contingency Table Calculations
Ethan Bolker and Maura Mast, for *Common Sense Mathematics*

Set population to 1 to see percentages in the table.
Set population to a large round number like 1000 to see meaningful numbers of cases.

The incidence, false positive and false negative rates are formatted as percentages.
You can use actual values in Excel formulas to set them - for example, =1/1000 for 0.1%.

population:	10000
incidence:	0.0130%
false+:	1.00%
false-:	50.00%

	actual+	actual-	total	
test+	0.65	99.987	100.637	0.65% prob(actual+ when test+)
test-	0.65	9898.713	9899.363	99.99% prob(actual- when test-)
total	1.3	9998.7		

Figure 13.6. Screening for Trisomy 18

confirming Davies' "0.65 cases per 10,000 women tested." The spreadsheet shows 100.637 positive tests, which matches Davies' estimate of 100. So it would take 100 amniocenteses to find 0.65 cases of trisomy 18. That works out to $100/0.65 = 154$ amniocenteses to find each case. (Exercise **??** asks you to check Davies's arithmetic for Down's syndrome.)

You don't have to know what amniocentesis is, but you do have to know that it has some risks: there is a small chance that it will lead to a miscarriage. Davies claims that "a screening programme would cause the abortion of at least as many normal fetuses as it would detect cases of trisomy 18." ("Abortion" his synonym for "miscarriage," not the politically charged "abortion" so much in the news.)

That would be true if the risk of miscarriage from amniocentesis was about one in 150. It's probably smaller. Several web sources provide statistics like these:

> Miscarriage is the primary risk related to amniocentesis. The risk of miscarriage ranges from 1 in 400 to 1 in 200. In facilities where amniocentesis is performed regularly, the rates are closer to 1 in 400. [R**??**]

If we use one in 300 instead of one in 150 as the probability of miscarriage from an amniocentesis then it costs about one unnecessary miscarriage to detect two cases of trisomy 18. That's still a pretty high risk.

Davies compares his risk estimate to the much lower estimate for similar screening for Down's syndrome, noting that "In many places it is still undecided whether screening for Down's syndrome is worth the disbenefits for the prospective parents."

13.5 The prosecutor's fallacy

The Cornell University Legal Information Institute posted a discussion of *McDaniel v. Brown* when that case was on the docket of the Supreme Court. They wrote

> Following a state conviction for sexual assault, Troy Brown filed a petition for writ of habeas corpus in the United States District Court for the District of Nevada. The District Court allowed Brown to present new evidence: a report from Dr. Lawrence Mueller. This report detailed a statistical error ("prosecutor's fallacy") made by the prosecution during the presentation of DNA evidence. Based on Dr. Mueller's report, the District Court dismissed the DNA evidence from consideration, found insufficient evidence to convict Brown, and ordered a retrial.
>
> ...
>
> At trial, Renee Romero, a forensic scientist at the Washoe County Crime Lab, testified that the DNA found in the victim's underwear matched Brown's DNA; only one in three million people would match the DNA tested. The prosecutor asked Romero to express this statistic as "the likelihood that the DNA found ... is the same as the DNA found in [Brown's] blood." Romero concluded that the likelihood was 99.999967 percent. Based on this statistic, the prosecutor then asked Romero if it would be fair to conclude that there was a 0.000033 percent chance that the DNA did not belong to Brown. Romero agreed with the prosecutor, stating that that this was "not inaccurate." [R**??**]

Romero's arithmetic is right: one in three million is 0.000033 percent. But her thinking is wrong.

The prosecutor's fallacy is the claim that the one in three million probability of a random match is the same as the probability that the defendant is innocent. We can use a contingency table to show why those probabilities are different.

First we need an estimate of the population in which a possible DNA match might be found. To make the arithmetic easier, we'll take that to be 9 million people (Los Angeles is near enough to Nevada). Then the "one in three million" statistic says we should expect three DNA matches from the innocent people in that population. This contingency table summarizes the data:

		truth		
		guilty	innocent	total
DNA	match	1	3	4
	nonmatch	0	8,999,996	8,999,996
	total	1	8,999,999	9,000,000

Make sure you understand the first row of the table. One person is guilty and is a DNA match. The three matches from innocent people are false positives. In other words, the first row of that table tells us that if the only evidence in the case is the DNA match the odds

are 3 : 1 that the suspect is innocent! The probability that he's guilty is only 25%. That's a far cry from the "99.999967% guilty" that the prosecutor asked the jury to believe.

The defense didn't make this argument using a hypothetical 9,000,000 population of potential suspects. Instead they questioned the "one in three million" chance of a match. The defendant had near relatives in the area which increased the chances of a match to about one in 6,500, according to a defense specialist. That would reduce the chance of an accidental match to $6499/6500 = 0.999846154 \approx 99.98\%$. We're not surprised that the change from 99.999967 percent to 99.98% did not convince the jury to acquit. 99.98% still sounds very much like a sure thing.

But it's not, because of the prosecutor's fallacy. That was the basis for the appeal. Suppose we reduce the population from which the match might come to just 100,000 — the nearby area where there may be close relatives. Then the 1 in 6,500 chance of a match means there will be about 15 matches in that population in addition to the one match for the guilty party. The numbers in the revised contingency table below show there is now a 15 : 1 chance that the DNA match fingers an innocent person rather than the true criminal.

		truth		
		guilty	innocent	total
DNA	match	1	15	16
	nonmatch	0	99,984	99,984
	total	1	99,999	100,000

Nevertheless, the story did not end well for Brown.

> The Supreme Court [overturning the appeals court order for a retrial] said in a per curiam opinion that overstated estimates of a DNA match at trial did not warrant reversal of a conviction when there is still "convincing evidence of guilt." [R**??**]

13.6 The boy who cried "Wolf"

After unusual disasters like terrorist attacks, earthquakes, severe storms or airplane crashes you often hear finger-pointing discussions about the incompetence of the agencies charged with predicting (perhaps even preventing) what happened. Those discussions may start with a search that discovers warning signs that were ignored.

Sometimes there were real lapses, and policies and practices must be designed to prevent a recurrence. But often blame is unjustified. Table **??** explains why, even without numbers. You might call this *qualitative reasoning*.

With numbers in the first column you can compute the probability that a disaster occurs with no warning at all. That's not good. To guard against it, there should be more warnings. Then with numbers in the first row you can compute the probability that a particular

		what happens		
		disaster	nothing	total
warning?	yes	rare	usually	infrequent
	no	rare	almost always	almost always
	total	rare	almost always	always

Table 13.7. Should it have been predicted?

warning actually corresponds to a disaster about to happen. But more warning don't lead to more disasters, just to more false positives.

That means there are often good reasons for ignoring a warning. State and governmental agencies have to balance the severity of the warning with the cost and inconvenience of asking the public to respond. For example, an earthquake warning may lead to an order to evacuate an entire city. The expense and disruption from repeated evacuations that are not followed by an earthquake may be worse than the consequences in the rare instance when the earthquake happens. Just because after the fact you look back and find clues in the seismic record that suggested an earthquake might be imminent doesn't mean evacuation was the right call.

13.7 Exercises

Exercise 13.7.1. [S][Section **??**] [Goal **??**][Goal **??**] Chronic fatigue syndrome.

On August 24, 2010 a headline in *The Boston Globe* read "Researchers link chronic fatigue syndrome to class of virus". The story reported on a study of 37 patients with the disease. 32 tested positive for a particular suspicious virus. Only 3 of 44 healthy people tested positive. [R**??**]

A 2003 study in the *Archives of Internal Medicine* reported that "The overall . . . prevalence of CFS . . . was 235 per 100,000 persons." [R**??**]

(a) Construct the contingency table for this diagnostic tool. You may do this by hand, or with the spreadsheet `ContingencyTable.xlsx`.

(b) Explain why this test is potentially important for research on chronic fatigue syndrome but might not be a good screening test.

[See the back of the book for a hint.] The first quote tells you the false positive and false negative rates. The second tells you the incidence.

(a) Construct the contingency table for this diagnostic tool.

I did the computations in Excel, entering the false positive rate as `=3/44` (6.82%), the false negative rate as `=5/37` (13.51%) and the incidence as `=235/100000`, for a population of 100,000.

Excel computed the contingency table (rounded):

		chronic fatigue syndrome		
		yes	no	total
tested positive	yes	203	6,802	7,005
	no	32	92,963	92,885
	total	235	99,765	100,000

(b) Explain why this test is potentially important for research on chronic fatigue syndrome but might not be a good screening test.

The test suggests pretty clearly that a virus may be involved in chronic fatigue syndrome. That is a lead worth pursuing with further research. However, the false positive rate is almost 7% and the incidence is low so most of the positives will be false positives, causing lots of anxiety and expense. In fact the probability that a positive test means the patient has CFS is just 2.9%.

The test might be good for people who already show symptoms suggesting they have the disease.

Exercise 13.7.2. [S][W][Section **??**][Goal **??**] [Goal **??**] Pregnancy tests.

An online website on pregnancy testing says that

> Usually, if all care has been taken, [home] pregnancy tests are 97% accurate. [R**??**]

Assume that "97% accurate" means a false positive rate and a false negative rate of 3%. Since a woman is unlikely to use a home pregnancy test unless she thinks she's probably pregnant, assume that 80% of the women who try one are in fact pregnant.

Explain why a positive test indicates a pregnancy more than 99% of the time even though the false positive rate is 3%.

Explain why a positive test indicates a pregnancy more than 99% of the time even though the false positive rate is 3%.

I did the calculations in the spreadsheet, imagining a population of 1000 possibly pregnant women:

population:	1000
incidence:	80.0000%
false+:	3.00%
false-:	3.00%

	actual+	actual-	total
test+	776	6	782
test-	24	194	218
total	800	200	

99.23% prob(actual+ when test+)
88.99% prob(actual- when test-)

The probability is 99.23%. It's *higher* than the incidence in the tested population, because in that population the probability that a woman is pregnant is alread high.

Exercise 13.7.3. [U][Section **??**] [Goal **??**][Goal **??**] Prenatal screening.

Check the calculations for Down's syndrome testing using the data in the quotation in Section **??**.

Exercise 13.7.4. [U][C][Section **??**][Goal **??**] [Goal **??**] Spam.

Spam is junk email. Most mail systems have a spam filter that tries to decide whether each piece of email you get is spam. When the spam filter finds something it thinks is spam, it may throw it away, or put it in a junk mail folder so that you can decide whether to throw it away without reading it.

Before my university department set up a spam filter I ran my own. (The "I" here is Ethan Bolker, one of the authors, not the generic authorial "we" we use in most of the book.)

I got about 250 emails each day. My spam filter trapped about 175 of them. Of those about five were legitimate, and should have been delivered directly to me. My inbox, which should have contained just the legitimate messages was usually about half spam. So (in words) my spam filter is pretty good (but not perfect) at recognizing legitimate email but not very good at calling spam spam.

(a) Build a two way contingency table with row categories "marked spam" and "not marked spam", column categories "spam" and "legitimate".

(b) Compute and interpret the false positive and false negative rates.

(c) Explain why both the false positives and the false negatives make dealing with my email harder.

(d) I can adjust the settings in my spam filter to reduce the false positive rate. Explain why that would increase the false negative rate.

(e) Is the number of spam emails I received consistent with the claim in the August 6, 2008 issue of *The New Yorker* that there are more than a hundred billion spam emails every day? [R**??**]

(f) What is the original meaning of the word "spam"? Does the company that sells (the real) spam object to the new meaning?

(g) How do you deal with spam? (If your email provider does all the filtering for you, you may not even know it's throwing things away before you see them, so you may need to do some research on your email provider's web site to find the answers to these questions.)

- Who provides your email service (your university, your internet service provider, Google, Yahoo, . . .) ?
- Do you have any say in how your email provider filters spam for you? If so, what do you tell it?
- Estimate the data you need to build the two way table for your spam statistics and compute the false negative and false positive rates.

Here are some web sites to look at if you want to find out more about spam.

- `www.imediaconnection.com/content/3649.asp` . There are some useful tips here about how to keep other people's spam filters from thinking mail from you is spam.
- Tools your system administrator might use: `www.spamcop.net/` , `www.spamhaus.org/`

Exercise 13.7.5. [S][Section **??**][Goal **??**] [Goal **??**] Plagiarism.

In 2006 UMass Boston experimented with the plagiarism detection software described at `www.turnitin.com` that claims it can identify plagiarism in essays students write. UMass did not purchase the software after the experiment. Perhaps the possibility of false positives contributed to that decision.

Suppose that the software can actually detect every cheater and that it's 99% accurate in declaring honest students honest. (We made up these numbers since the company does not advertise them.) Sounds like a pretty good test.

(a) Estimate how many papers are submitted by students at your school each semester.

(b) Suppose that most students are honest. Estimate how many students will be falsely accused of cheating.

(c) What are the advantages and disadvantages of using the software? (There are several arguments on both sides of the question. Think of as many as you can.)

(d) Read and write about this article from *The New York Times*: `www.nytimes.com/2010/07/06/education/06cheat.html`

(a) Estimate how many papers are submitted by students at your school each semester.

In the spring of 2011 there were about 13,000 students at UMass Boston. If each one wrote six papers a semester that would come to about 80,000 papers — a nice round number in the right ballpark.

(b) Suppose that most students are honest. Estimate how many students will be falsely accused of cheating.

Since most of the 80,000 papers are honest, the false positive rate applies — one percent of them, or 800 papers, will be falsely tagged as plagiarized. That might not be quite 800 students, since some students might be unjustly accused twice, but the order of magnitude is right.

(c) What are the advantages and disadvantages of using the software? (There are several arguments on both sides of the question. Think of as many as you can.)

An advantage is that some plagiarists will be caught who might otherwise get away with it. Another is that students might be less likely to cheat knowing that this software was being used.

I can think of several disadvantages. One is the anxiety caused by the false accusations. Another is the cost.

Exercise 13.7.6. [U][Section **??**][Goal **??**] Airport screening.

In response to the article "Screening programme evaluation applied to airport security" in the December 10, 2007 issue of the *British Medical Journal*, Ganesan Karthikeyan wrote

> It is probably true that airport security in its present form is not an efficient screening measure. However, one important difference exists between screening for disease in individual patients and screening for, say, explosives in airports. While one missed cancer on screening can cause the loss of at the most, one life, the number of potential lives lost per missed screening at airports can be substantially larger. This has to be factored into any attempts at evaluation of the process. [R**??**]

It's clear that a false negative is a disaster. Discuss the consequences of a high false positive rate.

Exercise 13.7.7. [S][W][Section **??**] [Goal **??**][Goal **??**] Breast cancer screening.

In his "Chances Are" column in *The New York Times* on April 25, 2010 Steven Strogatz wrote about a diagnostic puzzle presented to several doctors:

> The probability that [a woman in this cohort] has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?
>
> . . .
>
> [When 24 doctors were asked this question], their estimates whipsawed from 1 percent to 90 percent. Eight of them thought the chances were 10 percent or less, 8 more said 90 percent, and the remaining 8 guessed somewhere between 50 and 80 percent. Imagine how upsetting it would be as a patient to hear such divergent opinions. [R**??**]

(a) What is the correct answer?

[See the back of the book for a hint.] Build the contingency table, based on a population of 1,000 women tested. You may do this by hand or with the spreadsheet `ContingencyTable.xlsx`.

(b) What percentage of the 24 doctors got the correct answer?

(a) What is the correct answer?

Here is the contingency table, based on 1,000 women screened.

		has breast cancer		
		yes	no	total
screened positive	yes	7	70	77
	no	1	922	923
	total	8	992	1,000

So the probability that a woman with a positive mammogram actually has cancer is just 7/77 = 1/11, or about 9%.

(b) What percentage of the 24 doctors got the correct answer?

Eight doctors thought the correct answer was less than 10%, which it is. One doctor thought it was just 1%, so I won't count that as a correct answer. That means 7/24 or about 30% got the answer right.

Exercise 13.7.8. [U][S][Section **??**][Goal **??**] [Goal **??**] Identity fraud.

On July 17, 2011 in an article in *The Boston Globe* headlined "Identity fraud dragnet hardly seems worth the expense or trouble" you could read that in 2010 the Massachusetts Registry of Motor Vehicles used software that cost $1.5 million to send 1,500 suspension letters a day, leading to 100 arrests for fraudulent identity and 1,860 revoked licenses. [R**??**]

On July 24 Jane Allen wrote a letter to the editor in response, to say that the time and money hardly seemed worth it since only

> about 390,000 people were questioned for the sake of finding fewer than 2,000 transgressors. [R**??**]

(a) Check Allen's arithmetic in the second paragraph.

(b) Identify the false positives and calculate the false positive rate. Explain the costs and benefits.

(a) Check Allen's arithmetic in the second paragraph.

Since $1,500 \times 360 = 540,000$ the $390,000$ people in the second paragraph is an underestimate. Maybe it's 1,500 letters each business day, for $390,000/1,500 = 260$ business days.

(b) Identify the false positives and calculate the false positive rate. Explain the costs and benefits.

The population "tested" in this case is all the people who got letters. 2000 of those are true positives: they are the ones properly targeted. The false negatives would be people who should have gotten a letter but didn't — there's no way to know how many of those there are. Probably very few.

The false positive rate is $388,000/390,000 \approx 99.5\%$.

Allen's letter explains the costs: the $1.5 million grant, the personnel time — not to mention the headaches for the 388,000 falsely accused. The benefit: finding 2000 people who broke the law, 100 seriously.

Exercise 13.7.9. [U][C][Section **??**][Goal **??**] Candy leads to crime.

An article headlined "Happy Halloween! Kids who eat candy every day grow up to be violent criminals" in the October 2, 2009 *Daily Finance*, begins

> Quick, hide the candy jar! Feeding your child candy every day could help turn Junior into a violent criminal, according to a large study in Britain, which found that 69 percent of the participants who had committed violence by 34 had eaten sweets or chocolate nearly every day during childhood. [R**??**]

You can find the full text at `www.aol.com/2009/10/02/happy-halloween-kids-who-eat-candy-every-day-grow-up-to-be-viol/`

(a) Read the rest of the article. Build the contingency table with columns for whether or not someone ate candy as a child, rows for whether or not they committed violence as an adult.

(b) Explain why this is an example of the prosecutor's fallacy.

(c) Some of the online comments on that article recognize the fallacy — for example

> 10-03-2009 @ 10:21PM
> Bski said...
> I bet you, 99% of criminals ate bread daily by the time they were 10 years old!!!!

Write your own blog entry, using your understanding of two way contingency tables to enlighten any readers. If you like what you've written you may still be able to post your comment on the article's blog.

Exercise 13.7.10. [U][Section **??**][Goal **??**] Domestic violence.

In Andrew Gelman's blog on "Statistical Modeling, Causal Inference, and Social Science" commenter Mike Spagat writes that

> Even within exceptionally violent environments most households will still not have a violent death. So a very small false positive rate in a household survey will cause substantial upward bias in violence estimates. [R**??**]

Write a paragraph or two explaining this to someone who is interested and smart enough to understand this but has not studied the material in this chapter. Consider making up some numbers to illustrate your argument.

Exercise 13.7.11. [U][Section **??**][Goal **??**] Surgery for prostate cancer?

An article in *The Boston Globe* headlined "Surgery offers no advantage for early prostate cancer, study finds" reported on a clinical trial involving 731 men diagnosed with prostate cancer. About half had surgery; the rest were monitored.

> After 12 years, nearly 6 percent of men who had immediate surgery died of the cancer, compared with slightly more than 8 percent of those patients who were observed, which was not a great enough difference to reach statistical significance. [R**??**]

(a) About how many men were in each category?

(b) About how many deaths were there in each category?

(c) Construct the contingency table for this study.

Exercise 13.7.12. [S][A][W][Section **??**][Goal **??**][Goal **??**] Teenage drug use.

Here's a made up story.

The dean at a fancy private high school is very worried. She suspects that about 20% of the 1000 students on campus are using drugs. She has asked all the parents to administer a home drug test to their kids (since it's a private school she can actually require them to do it). She has read on the web that

> With home drug testing methods believed to produce reliable and accurate results, many of us overlook the cases of false positives and draw conclusions on the suspect before reconfirming the result. But, researchers from the Boston University have found out that drug tests may produce false positives in 5-10% of cases and false negatives in 10-15% of cases. [R**??**]

We found several blogs that seem to report on this same study. None gives a link or a precise reference. We haven't been able to locate the original.

Answer the following questions, assuming the worst cases (10% false positive rate, 15% false negative rate).

(a) Build the contingency table for this drug screening scenario. To do that you will have to figure out

> How many students are drug users?
> How many of the drug users test positive? How many test negative?
> How many students are drug free?
> How many of the drug free students test positive? How many test negative?

You may do the arithmetic with by hand or with the spreadsheet at `ContingencyTable.xlsx`.

(b) What is the true positive rate?

(c) Student John Smith tested positive. What is the probability that he is really on drugs?

(d) Student Jane Doe tested negative. What is the probability that she is really drug free?

(e) Answer the previous two questions if you assume the best cases for reported false values in the Boston University study.

		bought a ticket		
		yes	no	total
won the lottery	yes	1	0	1
	no	many	very many	very many
	total	many	very many	very many

Table 13.8. Playing the lottery

(a) Build the contingency table for this drug screening scenario. To do that you will have to figure out

How many students are drug users?
20% of 1000, so 200.
How many of the drug users test positive? How many test negative?
170 of the 200 users test positive. The other 30 test negative (these are the false negatives).
How many students are drug free?
The other 800.
How many of the drug free students test positive? How many test negative?
720 of the 800 clean students test negative. 80 are false positives.

(b) What is the true positive rate? $100\% - 15\% = 85\%$.

(c) Student John Smith tested positive. What is the probability that he is really on drugs?
That's $170/250 = 0.68 = 68\%$.

(d) Student Jane Doe tested negative. What is the probability that she is really drug free?
$760/780 = 0.96 = 96\%$.

(e) Answer the previous two questions if you assume the best cases for reported false values in the Boston University study.
82% and 97% — Excel did the work for me.

Exercise 13.7.13. [U][Section **??**][Goal **??**][Goal **??**] The boy who cried "wolf".

Use Table **??** to analyze the children's story with that title.

Exercise 13.7.14. [S][Section **??**][Goal **??**] Playing the lottery.

Table **??** illustrates the ultimate example of the error you can make reading a column instead of a row.

(a) Suppose you won the lottery. What is the probability that you bought a ticket?

(b) Suppose you bought a ticket. What is the probability that you won the lottery?

(a) Suppose you won the lottery. What is the probability that you bought a ticket?
You can't win if you don't play. The probability is 1.

(b) Suppose you bought a ticket. What is the probability that you won the lottery?
Essentially zero.

Exercise 13.7.15. [N][Section **??**][Goal **??**] Mad cow disease.

Bovine Spongiform Encephalopathy(BSE) is a disease fatal to people who eat infected beef products.

Here is a paragraph from the United States Department of Agriculture website on screening for BSE:

> After the first confirmation of BSE in an animal in Washington State in December 2003, USDA evaluated its BSE surveillance efforts in light of that finding. We determined that we needed to immediately conduct a major surveillance effort to help determine the prevalence of BSE in the United States. Our goal over a 12-18 month period was to obtain as many samples as possible from the segments of the cattle population where we were most likely to find BSE if it was present. This population was cattle exhibiting some signs of disease. We conducted this enhanced surveillance effort from June 2004 - August 2006. In that time, we collected a total of 787,711samples and estimated the prevalence of BSE in the United States to be between 4-7 infected animals in a population of 42 million adult cattle. We consequently modified our surveillance efforts based on this prevalence estimate to a level we can monitor for any potential changes, should they occur. Our statistical analysis indicated that collecting approximately 40,000 samples per year from the targeted cattle population would enable us to conduct this monitoring. [R**??**]

Exercise 13.7.16. [N] Correlation and causation.

This question and and answers at the statistics stackexchange site has nice examples. The answers are written using conditional probabilities but can be rewritten as contingency tables.

`stats.stackexchange.com/questions/283133/relationships-between-correlation-and-causation`

R?? M. S. Klinkman, J. C. Coyne, S. Gallo and T. L. Schwenk,, False Positives, False Negatives, and the Validity of the Diagnosis of Major Depression in Primary Care, *Arch Fam Med.* 1998;7(5):451 – 461, www.ncbi.nlm.nih.gov/pubmed/9755738 (last visited October 4, 2015). Licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License. (creativecommons.org/licenses/by-nc-nd/3.0/

R?? T, Davies, Prenatal screening for trisomy 18: Should not be contemplated, *British Medical Journal*, February 12, 1994, doi: doi.org/10.1136/bmj.308.6926.471, www.bmj.com/content/308/6926/471.1 (last visited July 28, 2019).

R?? Amniocentesis, American Pregnancy Association, americanpregnancy.org/prenataltesting/amniocentesis.html (last visited July 15, 2015).

R?? M. Lynn and C. Maier, McDaniel v. Brown (08-559) Legal Information Institute, LII Supreme Court Bulletin, Cornell University Law School, www.law.cornell.edu/supct/cert/08-559 (last visited September 18, 2015). Michelle Jessica Lynn and Chris Maier are the authors of that Supreme Court Preview. Legal Information Institute at Cornell Law School.

R?? D. Badertscher, U.S. Supreme Court Update: McDaniel v. Brown, Criminal Law Library Blog (January 26, 2010), www.criminallawlibraryblog.com/2010/01/us_supreme_court_update_mcdani.html (last visited July 25, 2015).

R?? R. Stein, Researchers link chronic fatigue syndrome to class of virus, Washington Post report in *The Boston Globe* (August 24, 2010), www.boston.com/news/nation/articles/2010/08/24/researchers_link_chronic_fatigue_syndrome_to_class_of_virus (last visited March 30, 2020).

R?? M. Reyes *et. al.*, Prevalence and incidence of chronic fatigue syndrome in Wichita, Kansas. *Arch Intern Med.* 2003 Jul 14;163(13):1530–6, www.ncbi.nlm.nih.gov/pubmed/12860574 (last visited December 15, 2015).

R?? When should I test with a pregnancy test?, Yourdays (free information for women), www.yourdays.com/when-pregnancy-test.htm (last visited March 12, 2020).

R?? M. Specter, Damn Spam, Annals of Technology, *The New Yorker* (August 6, 2007), www.newyorker.com/reporting/2007/08/06/070806fa_fact_specter (last visited July 31, 2019).

R?? G. Karthikeyan, The cost of a "negative test", response to Screening programme evaluation applied to airport security, *British Medical Journal* (December 27 2007), www.bmj.com/rapid-response/2011/11/01/cost-negative-test (last visited September 4, 2015). Quoted with permission.

R?? S. Strogatz, Chances Are, *The New York Times* (April 25, 2010), opinionator.blogs.nytimes.com/2010/04/25/chances-are/ (last visited March 2, 2016).

R?? M. E. Irons, Caught in a dragnet, *The Boston Globe*, July 17, 2011, archive.boston.com/news/local/massachusetts/articles/2011/07/17/man_sues_registry_after_license_mistakenly_revoked/ , (last visited March 12, 2020).

R?? J. Allen, Identity fraud dragnet hardly seems worth the expense or trouble, *The Boston Globe* (July 24, 2011), www.boston.com/bostonglobe/editorial_opinion/letters/articles/2011/07/24/identity_fraud_dragnet_hardly_seems_worth_the_expense_or_trouble/ (last visited March 30, 2020).

R?? E. Wahlgren, Happy Halloween! Kids who eat candy every day grow up to be violent criminals, www.aol.com/2009/10/02/happy-halloween-kids-who-eat-candy-every-day-grow-up-to-be-viol/ originally published on DailyFinance.com (October 2, 2009), (last visited March 12, 2020). Quoted with permission.

R?? A. Gelman, The Reliability of Cluster Surveys of Conflict Mortality: Violent Deaths and Non-Violent Deaths, Statistical Modeling, Causal Inference, and Social Science (August 11 2011), andrewgelman.com/2011/08/the_reliability/ (last visited July 25, 2015). Quoted with permission.

R?? D. Kotz, Surgery offers no advantage for early prostate cancer, study finds, *The Boston Globe* (July 18, 2012), bostonglobe.com/lifestyle/health-wellness/2012/07/18/surgery-offers-survival-advantage-for-older-men-with-early-stage-prostate-cancer-study-finds/T5XM7APIuoZuav6PbJzYuI/story.html (last visited July 25, 2015).

R?? How to Avoid False Positives While Conducting a Home Drug Test, lapoliticaesotracosa.blogspot.com/2012/05/how-to-avoid-false-positives-while.html (last visited July 25, 2015).

R?? BSE (Mad Cow Disease) Ongoing Surveillance Information Center, U.S. Department of Agriculture, www.usda.gov/wps/portal/usda/usdahome?contentid=BSE_Ongoing_Surveillance_Information_Center.html (last visited November 15, 2015).