

Testing for HIV: Connecting Arithmetic, Functions and Probability to the Real World

Lew Romagnano
The Metropolitan State College of Denver
romagnal@mscd.edu <http://clem.mscd.edu/~romagnal>

The HIV Problem

- (a) What is the chance that Chris, a person selected at random from the population of the USA, is infected with HIV, the virus that causes AIDS?
- (b) What if Chris was one of 100 000 people selected at random from the population of the USA for an HIV test that is 99.5 percent accurate, and what if Chris's test result was "positive"? What is the chance that Chris is infected with HIV?

For question (a), in the absence of other information about Chris, the best guess of the chance that Chris is infected with HIV is $1/250 = 0.004$, because in the population of the USA, estimates suggest that one in every 250 people has been infected with HIV.

Question (b) is different, because we have new information about Chris. When selected as one of 100 000 people to be tested, the result was a positive test. Given that information, what is your estimate of the chance that Chris really is infected?

The Testing Procedure

A common procedure for testing for HIV is as follows: the ELISA (enzyme-linked immunosorbent assay) test is applied to a portion of the person's blood sample. If that very sensitive test is positive, then another ELISA test is applied, this time to a different portion of the sample. If that test is positive too, then the "Western Blot" test is applied to a third portion of the sample. If all three tests are positive, then a positive test result is reported.

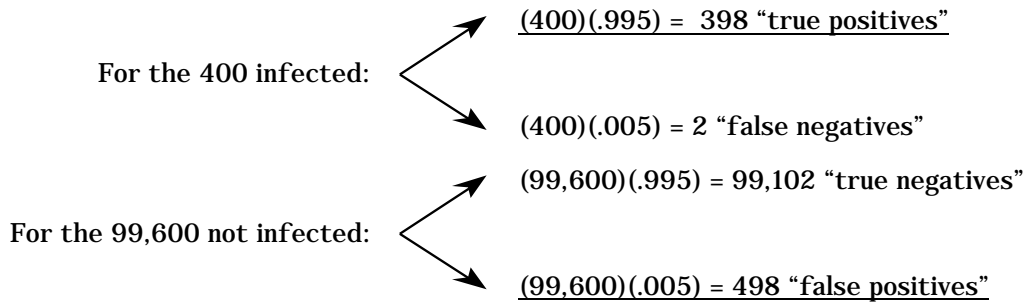
This testing procedure is very accurate; estimates range as high as 99.5 percent or higher. For this paper, we will assume that the accuracy of the test—both the test's *sensitivity* (its ability to detect the HIV antibodies if present) and its *specificity* (its ability to correctly determine that HIV antibodies are not present)—is 99.5 percent.¹

Arithmetic Solutions

One person out of each 250, or 0.4 percent of the population, is actually infected, so 400 of the 100,000 in the sample are infected and 99,600 are not infected. The arithmetic solution to this problem is to determine the number of correct and incorrect tests in each of these two subgroups of people being tested. Here are two ways to picture this:

¹ In general, the sensitivity and specificity of a test are not the same. For example, a CDC report published almost a decade ago stated, "In 1988 and 1989, the ELISA [test] correctly identified 98.5 percent of infected blood donors and 99.5 percent of HIV-positive samples. The Western blot detected 99.7 percent of positive samples and 91.6 percent of negative results in 1989..."

Testing for HIV



Sample Size = 100,000	Actually Infected = 400	Actually Not Infected = 99,600
Accurate Test	$(400)(.995) = 398$ True Positive	$(99,600)(.995) = 99,102$ True Negative
Inaccurate Test	$(400)(.005) = 2$ False Negative	$(99,600)(.005) = 498$ False Positive

In either case, the probability that a positive test result is accurate is:

$$P(\text{True Positive given a Positive Test}) = \frac{\text{True +}}{\text{Total +}} = \frac{398}{398 + 498} = \frac{398}{896} \approx .44$$

So, someone who tests positive will have learned less than if she or he had tossed a coin. The presence in the sample of such a large number of uninfected people produces a relatively large number of false positives, even with a test that is nearly perfect.

Functional Analysis

How does the infection rate of the sample affect these results? By repeating the calculations illustrated above, the following data may be collected:

For a test that is 99.5 percent accurate:

Infection Rate of Sample	Probability of True + if + Test
.004	.4442
.01	.6678
.02	.8024
.05	.9128
.1	.9567
.2	.9803
.4	.9925
.5	.995
.6	.9967

From this table, one could conclude that the probability does depend on the infection rate; as the number of infected people approaches half of the sample, the probability approaches the accuracy rating of the test. When more than half of the sample is infected, the probability exceeds the accuracy of the test, and the probability approaches 1, asymptotically.

Several other questions might come to mind at this point. How accurate must a test be in order to produce "acceptable" results for a given sample infection rate? (And how would you define "acceptable"?) Why does the probability equal the accuracy of the test when the infection rate is .5? Is this always the case, independent of the test's accuracy?

Testing for HIV

To answer these and other questions about this situation, one could generalize the computational procedures illustrated above. If the size of the sample is n , the infection rate is i , and the accuracy of the test is a , then:

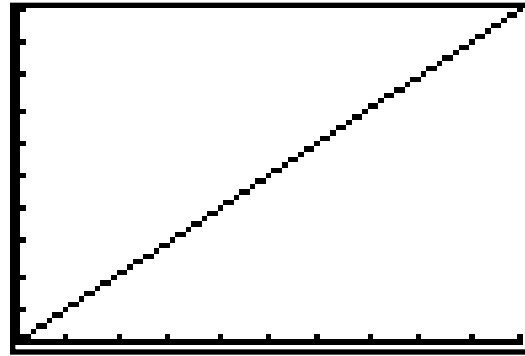
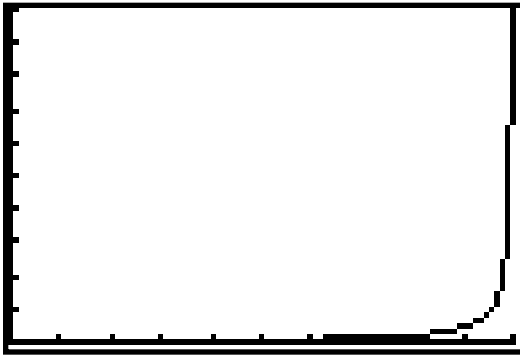
- The number of infected individuals is $(n)(i)$;
- The number of uninfected individuals is $(n)(1 - i)$;
- The number of true positives is $(n)(i)(a)$; and
- The number of false positives is $n(1 - i)(1 - a)$.

Therefore, the probability that a person is infected given a positive test is given by the following function:

$$P(\text{True+} | +\text{Test}) = \frac{nia}{nia + n(1-i)(1-a)}$$

$$P(\text{True+} | +\text{Test}) = \frac{ia}{ia + (1-i)(1-a)} = f(i, a)$$

The probability, regardless of the size of the population, is a function of both the infection rate (i) and the accuracy of the test (a). By treating one of these variables as a fixed parameter and allowing the other to vary, one could use graphical analysis to investigate the questions posed above. For example, if i is fixed at .004, the value from our original problem, then the graph of $P = f(a)$ for $0 < a < 1$ and $0 < P < 1$ (see left graph below) illustrates that the probability is extremely low for all but the most accurate tests.



Setting $i = .5$, the graph at right above shows that $P = a$ for all values of a between 0 and 1. The identity relationship between P and a when $i = .5$ can be confirmed symbolically as follows:

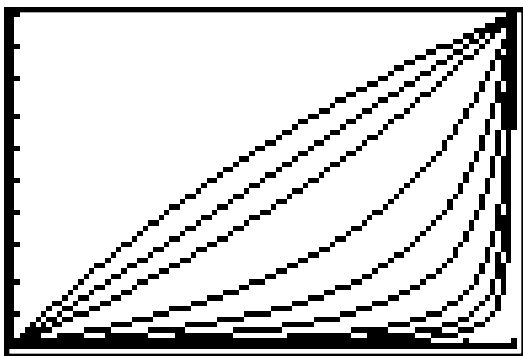
$$P = \frac{ia}{ia + (1-i)(1-a)} = \frac{ia}{ia + 1 - i - a + ia} = \frac{ia}{2ia - i - a}$$

If $i = .5$,

$$P = \frac{.5a}{2(.5)a - .5 - a} = \frac{.5a}{a - .5 - a} = \frac{.5a}{.5} = a$$

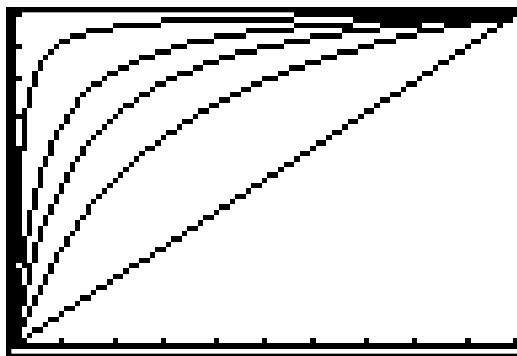
Finally, the families of graphs produced below contain the complete description of all of the relationships discussed in this problem.

Testing for HIV



$$y = \frac{ix}{ix + (1-i)(1-x)}$$

$x = \text{accuracy}$ $y = \text{probability}$
 $i = .004, .01, .02, .05, .1, .2, .4, .5, .6$



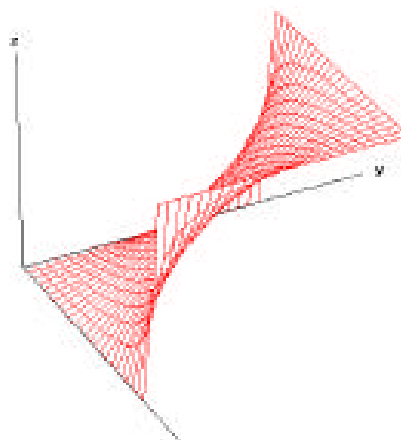
$$y = \frac{ax}{ax + (1-a)(1-x)}$$

$x = \text{infection rate}$ $y = \text{probability}$
 $a = .99, .95, .9, .8, .5$

The graph at right² shows the surface defined by the two-variable function

$$z = f(x, y) = \frac{xy}{xy + (1-x)(1-y)}$$

This graph is a 3-dimensional summary of this problem. Its cross-sections look like the two-dimensional graphs above.



What if We Re-Test?

It makes sense, given the results of the previous analysis, to re-test everyone who tested positive the first time around. What would be the results of this second test?

The infection rate of this new “population” of 896 people is .44 (398/896), and the accuracy of the test is still .995, so our function gives the following:

$$P(\text{True} + \text{2nd} + \text{Test}) = \frac{(444)(.995)}{(444)(.995) + (1 - .444)(1 - .995)} = 0.994.$$

Because the actually infected portion of this new population is now close to half, the chance that a second positive test is correct is close to the accuracy of the test. Re-testing is a good idea.

Note that the infection rate of the original test population (in this example, $1/250 = 0.004$) can be thought of as the probability that someone is really infected given no test. Similarly, the infection rate of the re-test population (in our case, 0.444) is probability that someone is really infected

² The graph was produced using PEANut Software, a free package of programs for Windows computers written by Rick Parris, Phillips Exeter Academy. The programs are available for download at <http://math.exeter.edu/rparris>.

Testing for HIV

given one positive test. In other words, the probability function we derived above can be written recursively:

$$P_0 = i_0 = 0.004$$

$$P_1 = \frac{P_0 a}{P_0 a + (1 - P_0)(1 - a)} = 0.444 = i_1$$

$$P_2 = \frac{P_1 a}{P_1 a + (1 - P_1)(1 - a)}$$

$$\vdots$$

$$P_n = \frac{P_{n-1} a}{P_{n-1} a + (1 - P_{n-1})(1 - a)}$$

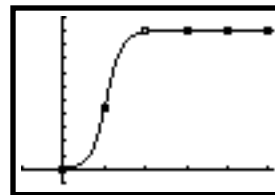
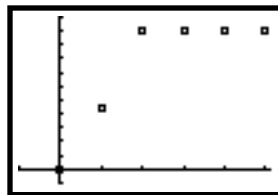
Using the TI-83's sequence mode, we can define this recursive function and use it to generate the results of repeated tests.

```

P1ot1 P1ot2 P1ot3
nMin=0
u(n)=(.995u(n-1)
)/(.995u(n-1)+.
005(1-u(n-1))
u(nMin)=.004
v(n)=
v(nMin)=
    
```

n	u(n)
0	.004
1	.4442
2	.99375
3	.99997
4	1
5	1
6	1
u(n)=.9999998412	

The graph of the table's values is shown at left below. The graph at right shows that the data fit the classic "S-shaped" curve of the logistic function.



The HIV Problem and Conditional Probability

Part (b) of the HIV problem may be re-stated as follows: What is the chance that Chris is actually infected, given a positive test result? In other words, how can you improve your estimate of the chance that Chris is HIV positive, if you know the chance without any other information (because you know the infection rate in the population), you have some evidence that Chris is HIV positive (a positive test result) and you know the reliability of that evidence (the accuracy of the test)? The information given in this problem can be restated as probabilities, in the following way.

1. The infection rate, i , may also be interpreted as a probability: the probability that someone chosen at random from the sample is actually infected. That is, $i = P(\text{True Positive})$. Therefore, $(1 - i) = P(\text{Not Positive})$.
2. The accuracy of the HIV test, a , may be thought of as the probability that someone who is positive will test positive. That is, $a = P(\text{Positive Test} | \text{True Positive})$. Therefore, $(1 - a) = P(\text{Positive Test} | \text{Not Positive})$.

Using these restatements of the information in this problem, the function we derived earlier:

Testing for HIV

$$P(\text{True } + | \text{Test}) = \frac{ia}{ia + (1-i)(1-a)}$$

may be rewritten:

$$P(\text{True } + | \text{Test}) = \frac{P(+\text{Test} | \text{True } +)P(\text{True } +)}{P(+\text{Test} | \text{True } +)P(\text{True } +) + P(+\text{Test} | \text{Not } +)P(\text{Not } +)}$$

If we call being a true positive A and getting a positive test B, then “not positive” is the complement of A, written A^c , and

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | A^c)P(A^c)}$$

This is Bayes' Theorem, a formula for computing the probability of A given the non-independent condition B.

References

- Devlin, K. (2000, February). The Legacy of Reverend Bayes. *MAA Online* [Online]. Available: http://www.maa.org/devlin/devlin_2_00.html [2000, October 27].
- Garfunkel, S., Godbold, L. & Pollak, H. (1998). *Mathematics: Modeling our World, Course 1*. Cincinnati, OH: South-Western Educational Publishing.
- Moore, D. S. (1990). Uncertainty. In L. A. Steen (ed.) *On the Shoulders of Giants: New Approaches to Numeracy* (pp. 95 – 137). Washington, D.C.: National Academy Press.