Scientists, engineers, and statisticians share similar concerns about evaluating the accuracy of their results, but they don't always talk about it in the same language. This can lead to misunderstandings when reading across disciplines, and the problem is exacerbated when technical work is communicated to and by the popular media.

The "Statistics to English Translation" series is a new set of articles that we will be posting from time to time, as an attempt to bridge the language gaps. Our goal is to increase statistical literacy: we hope that you will find it easier to read and understand the statistical results in research papers, even if you can't replicate the analyses. We also hope that you will be able to read popular media accounts of statistical and scientific results more critically, and to recognize common misunderstandings when they occur.

The first installment discusses some different accuracy measures that are commonly used in various research communities, and how they are related to each other. There is also a more legible PDF version of the articlehere.

The Basics

In informal language and in popular press articles, "accuracy" is often discussed as if it were a one-dimensional property of a diagnostic test or a classifier.

MyMoney Personal Finance Software
Automatically classifies your financial transactions with
99% accuracy
So you don't have to!
Budgets are a breeze with
MyMoney!

In general, though, a single number is not enough. A test or classifier should detect what's interesting, and ignore what's not. How well it accomplishes these two tasks is related to the two kinds of mistakes that a test or classifier can make: false negatives, and false positives.

For a classification task, *positive* means that an instance is labeled as belonging to the class of interest: we may want to automatically gather all news articles about Microsoft out of a news feed, or identify fraudulent credit card transactions. For a screening test, positive means that the test detects whatever it was designed to look for: an HIV test detects the presence of human immunodeficiency virus, for example, while an allergy test detects the presence of an allergic reaction. A *negative* is obviously the opposite of a positive.

A *false positive* is concluding that something is positive when it is not. False positives are sometimes called *Type I errors*. A *false negative* is concluding that something is negative when it is not. False negatives are sometimes called *Type II errors*. The terms "Type I error" and "Type II error" are not terribly mnemonic, but they are commonly used, and therefore worth knowing.

For binary classification or binary test procedures, the *False Positive Rate*, FPR , is the fraction of negative instances that are erroneously misclassified as positive.

$$FPR = \frac{\text{\#false positives}}{\text{all negative instances}} = \frac{\text{\#false positives}}{\text{\#false positives} + \text{\#true negatives}} \tag{1}$$

Likewise, the *False Negative Rate*, FNR , is the fraction of positive instances that are erroneously misclassified as negative.

$$FNR = \frac{\text{\#false negatives}}{\text{all positive instances}} = \frac{\text{\#false negatives}}{\text{\#false negatives} + \text{\#true positives}} \tag{2}$$

The *True Positive Rate*, TPR , is the fraction of positive instances that are correctly identified as such. It follows from the Definition 2 above that

$$TPR = 1 - FNR$$

.

The *True Negative Rate*, TNR , is the fraction of negative instances that are correctly identified as such. It follows from the Definition 1 above that

$$TNR = 1 - FPR$$

.

Sensitivity and Specificity

The terms sensitivity and specificity generally refer to diagnostic or screening procedures, such as an HIV or allergy tests. The *sensitivity* of a test is its true positive rate; the *specificity* is its true negative rate, although it can be more intuitive to think of specificity as the complement of the false positive rate: *Specificity* = $TNR = (1 - FPR)$.

The Wikipedia entry on Sensitivity and Specificity [Wiki] uses a nice example to illustrate the difference: think of a drug-sniffing dog as a screening test for illicit drugs. If the dog's nose is highly *sensitive* to the smell of drugs, then it will detect all the hidden packets of drugs; if it is less sensitive, then it will fail to detect some of the packets. At the same time, the dog should react *specifically* to drugs, and not, say, jambalaya or doggie biscuits. If the dog is highly specific in its reactions, it will only react to drugs; if it is less specific, then it will react to the occasional care package of yummy home cooking from Mom.

Screening tests may trade off specificity for sensitivity (and vice-versa). To go back to our drug-sniffing example, we might treat every suitcase and bag that comes through the airport as if it contained drugs; this procedure is perfectly sensitive (it will detect every packet of drugs, for sure), but not specific at all. Or, we might assume that no one is carrying drugs. This is perfectly specific (we will never make a false accusation), but not sensitive at all.

A more realistic example, inspired by a discussion of mandatory AIDS testing by Joshua Rosenau [Ros06], is the use of the ELISA screening test to detect HIV-infected blood donations. The ELISA test is designed to be very sensitive: it detects 99.7% of the cases of HIV-infection, which gives a false negative rate of 3×10^{-3}. On the other hand, it is not very specific: it has a 1.9% false positive rate[1]. If you assume that the incidence of HIV-positive individuals in the general population is about 448 out of every 100,000 people [Hig08], then a positive test result is correct only about 19% of the time: one case of true infection out of every five positives. This error rate may be appropriate for screening blood donations, since it is better to discard four perfectly good pints of blood, "just in case", than to allow a pint of HIV-infected blood into the blood bank. But it is *not* appropriate to assume that all five of those poor blood donors are HIV-positive, without follow-up tests to increase the specificity of the screening procedure.

Sensitivity, Specificity, and Prevalence

The example above brings up an important point. Sensitivity and specificity are properties *of the test itself*, not *how the test performs in a given population*. **The absolute accuracy** (as the term is commonly understood) **of a screening test will change, depending on the prevalence of the condition that the test is screening for.**

Let's imagine the ELISA test described above as an HIV-screening daemon, who uses two coins to generate uncertainty. When the daemon is shown a pint of infected blood, she flips an unfair quarter. If the quarter comes up heads (which it does 3 times out of every 1000 flips), then she lies and says the blood is uninfected, otherwise she tells the truth. When the daemon is shown a pint of uninfected blood, she flips a silver dollar that comes up heads about 2 times out of every 100 flips. If the silver dollar comes up heads, she lies and says the blood is infected, or else she tells the truth. The quarter and the silver dollar encode ELISA's sensitivity and specificity, respectively.

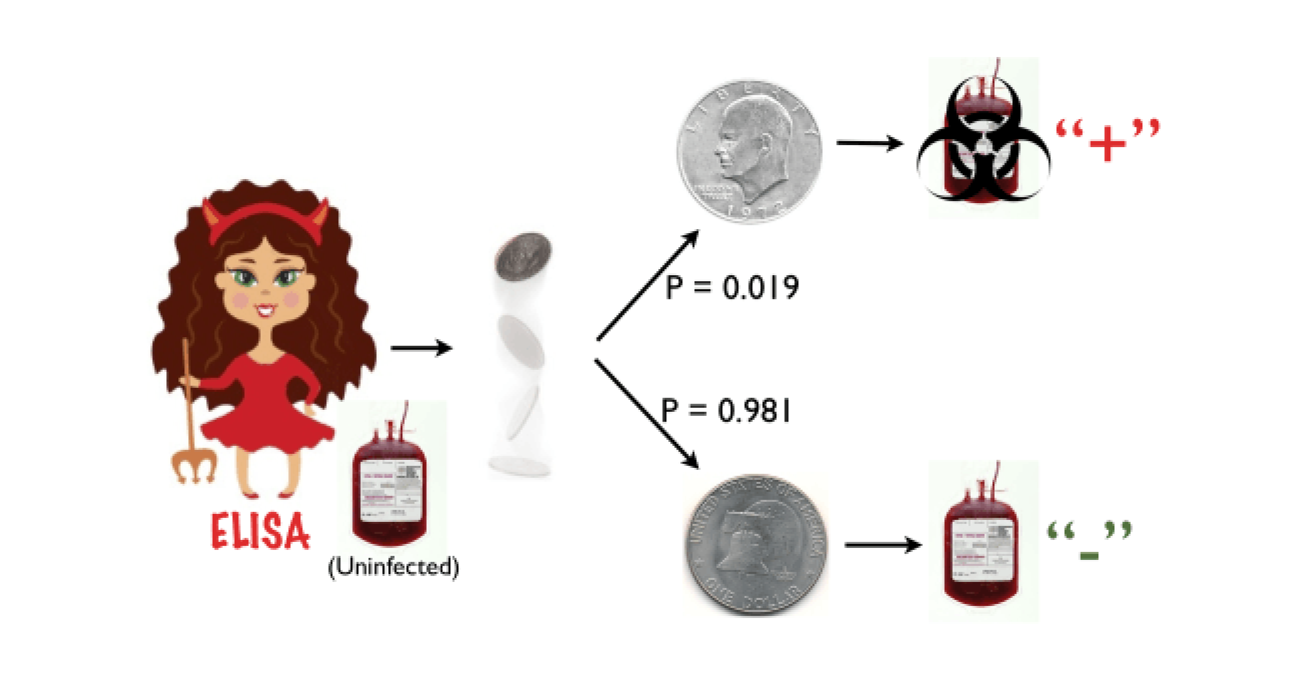

Figure 1: The ELISA daemon screening an uninfected pint of blood

Suppose ELISA looks at the blood of 1000 people a day, drawn from the general population. We can expect that about 5 of them are truly infected. That means that ELISA flips her silver dollar 995 times; it will come up heads about 20 times. That's about twenty false positives a day. She'll flip the quarter about 5 times, and with high probability, won't ever see a head. That's near zero false negatives a day. In total, ELISA will read positive for about 25 pints of blood every day, and she will be wrong for 80% of those cases.

But suppose ELISA looks at the blood of 1000 people from a high-risk population, where one out of four people are infected. Then ELISA flips her silver dollar about 750 times, and it will come up heads about 15 times: 15 false positives. She'll also flip the quarter 250 times; the coin just might come up heads one time. Let's say it does. Then ELISA will read positive for 249+15 = 264 pints of blood, and she'll be wrong for only about 6% of those cases – plus that case of infection that she missed.

Same test, same sensitivity and specificity, but different overall accuracy. The percentage of positives that are actually true positives in a given population is called the *positive predictive value* (PPV) of the test within that population; it is the probability *for that population*that a positive test result correctly predicts a positive instance.

$$PPV = \frac{TPR \times P(+)}{TPR \times P(+) + FPR \times P(-)} \qquad (3)$$

where $P(+)$ is the probability of a positive instance, or in other words the prevalence of the condition in the population. $P(-)$ is the probability of a negative instance, and of course $P(+) + P(-) = 1$.[2]

Likelihood Ratios

Likelihood ratios are another measure of diagnostic test accuracy. The *positive likelihood ratio* is the true positive rate over the false positive rate: $LR_P = TPR/FPR$. For our example ELISA test, the positive likelihood ratio is 0.997/0.019 = 52.47. The *negative likelihood ratio* is the false negative rate over the true negative rate, $LR_N = FNR/TNR$. For our ELISA example, the negative likelihood ratio is 0.003/.981 = 0.003058.

Likelihood ratios are a property of the screening test, independent of the prevalence of the condition in the population. If you know the odds of infection for the population of interest, $odds_{pop} = P(+)/P(-)$, then you can calculate the posterior odds of infection for someone who has tested positive:

$$odds_{post} = LR_P \times odds_{pop}$$

and the posterior odds of infection for someone who has tested negative:

$$odds_{post} = LR_N \times odds_{pop}$$

It's been argued that likelihood ratios make it easier for non-statistically-minded practitioners to interpret the results of a test than sensitivity and specificity do [JGS94]. It's also been argued the other way [PSBtR05]. Which framework makes more sense depends on if you prefer thinking in odds or probabilities. In either case you should be leery of "guidelines" of the sort: " $LR_P > 10$ indicates large and often conclusive increase in the likelihood of the disease." There is certainly a large increase in the posterior likelihood of infection if $LR_P > 10$, but as the ELISA coin-flipping example should have made clear, this posterior likelihood can still be quite small, if the disease is sufficiently rare.

I occasionally see something called the *diagnostic odds ratio*. It was developed as "a single indicator of test performance," and I've seen it described as "the odds of the true positive rate divided by the odds of the false positive rate" [HC07]. I could give you the actual formula here, but frankly – it makes no sense. The whole point of having two measures for accuracy is that one is not enough, and if you absolutely must have one number, you are better off using something like the F_1 measure that we describe in the next section.

Precision and Recall

Precision and recall are similar (but not identical) to sensitivity and specificity. The measures are popular in the information retrieval and machine learning communities.

Recall is the same as sensitivity, or the true positive rate: the number of true examples correctly classified as such. *Precision* is defined as the fraction of instances classified as positive that really are positive:

$$\text{precision} = \frac{\#\ \text{true positives}}{\#\ \text{true positives} + \#\ \text{false positives}}$$

This is *not* the same as specificity; it is the same as the positive predictive value, and is a joint property of the classifier and the population that it was evaluated on.

Information retrieval research concerns itself with efficient discovery of relevant documents from document collections, and that domain motivates the definitions of precision and recall. A library patron queries the library catalog for books on a given topic; the catalog's search engine should return all of the books relevant to her query, and only those books. Recall is a measure of how well the search delivers "all of the relevant books", and precision is a measure of how well it delivers "only the relevant books". If recall is poor, then the patron will miss finding many relevant books; if precision is poor, then she will be inundated with a bunch of book suggestions that have nothing to do with her search.

As with diagnostic procedures, classifiers may trade precision for recall, and vice-versa. Suppose our library patron is looking for novels about vampires. She could request all novels with the word “vampire” in the title. This search would have almost perfect precision, since presumably a novel with the word “vampire” in the title is going to be about vampires. It would not have perfect recall, since many novels about vampires – like *Dracula*, or the books from the *Twilight* series – don’t announce themselves quite that blatantly. Now suppose she is only interested in novels from Anne Rice’s *Vampire Chronicles* series. She could request all novels authored by Anne Rice. This search would have perfect recall, but not perfect precision, since Ms. Rice did in fact write several novels that are not about vampires.

These examples show that neither high precision nor high recall guarantee a useful classifier. It is the tension between achieving high precision and high recall that leads to good classifiers.

As we discussed above, the primary difference between precision and specificity is that precision is a property of*the algorithm and the population*. One could argue that precision is a more appropriate measure than specificity for many classification and machine learning tasks, especially those related to text or natural language. The fundamental assumption, after all, is that such algorithms are trained on data that is representative of the population that the classifier will be deployed in.

If you insist: Single Score Measures

There is another measure called F_1, the harmonic mean of precision and recall:

$$F_1 = \frac{2}{(1/\text{precision} + 1/\text{recall})} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

F_1 is near one when both precision and recall are high, and near zero when they are both low. It is a convenient single score to characterize overall accuracy, especially for comparing the performance of different classifiers.

Using F_1 to compare classifiers assumes that precision and recall are equally important for the application. If one criterion is more important than the other, then one can also use the weighted geometric mean:

$$F_\alpha = (1+\alpha)(\text{precision} \times \text{recall})/(\alpha\ \text{precision + recall}).$$

α describes how much more important recall is than precision: use F_2 if recall is twice as important as precision, $F_{0.5}$ if precision is twice as important as recall.

It is still better to have separate target goals for precision and recall that a candidate classifier must meet. Still, F_1 and F_α are found in the literature, so they are presented here.

ROC Curves

Not all diagnostic tests or classifiers return a simple “yes-or-no” answer. In fact, most probably don’t. Generally, a classification or diagnostic procedure will return a score along a continuum; ideally, the positive instances score towards one end of the scale, and the negative examples towards the other end. It is up to the scientist or the analyst to set a threshold on that score that separates what is considered a positive result from what is considered a negative result. The Receiver Operating Characteristic Curve, or *ROC Curve*, is a tool that helps set the best threshold.

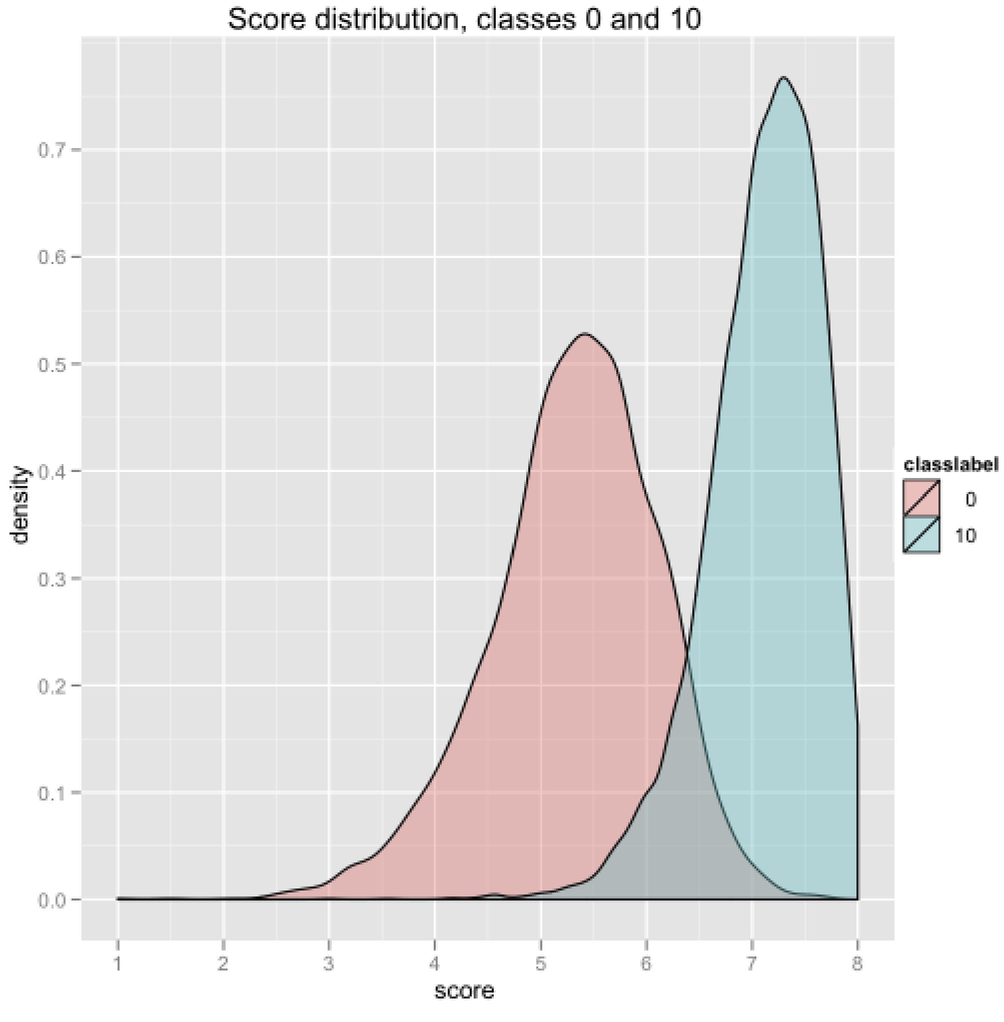

Figure 2: Plot of score distributions for positive and negative instances (Class 10 is positive)

Suppose we are trying to classify a set of instances into one of two classes, positive and negative[3]. We've gathered a test set of representative samples, and we've developed a scoring procedure to try to separate them. Positives tend to score on the high end of the scale, negatives toward the low end. We want to pick a threshold value.

Figure 2 shows what happens when score the test set. We can see that the scores of the positive instances (class 10) are in a cluster centered just above 7, and the scores of the negatives (class 0) are in a cluster centered near 5. Still, there is an interval where the two clusters overlap substantially. If we pick a threshold to the right of that interval (say, $T = 7$), almost everything that scores greater than T will be truly positive (high precision/specificity), but we miss a lot of positives, too (low recall/sensitivity). If we pick a threshold to the left of that interval (say $T = 5$), we will catch almost all the positives (high recall/sensitivity), but we will also pick up a lot of negatives (low specificity/precision). So we want the threshold to be somewhere in the overlap interval, but where?

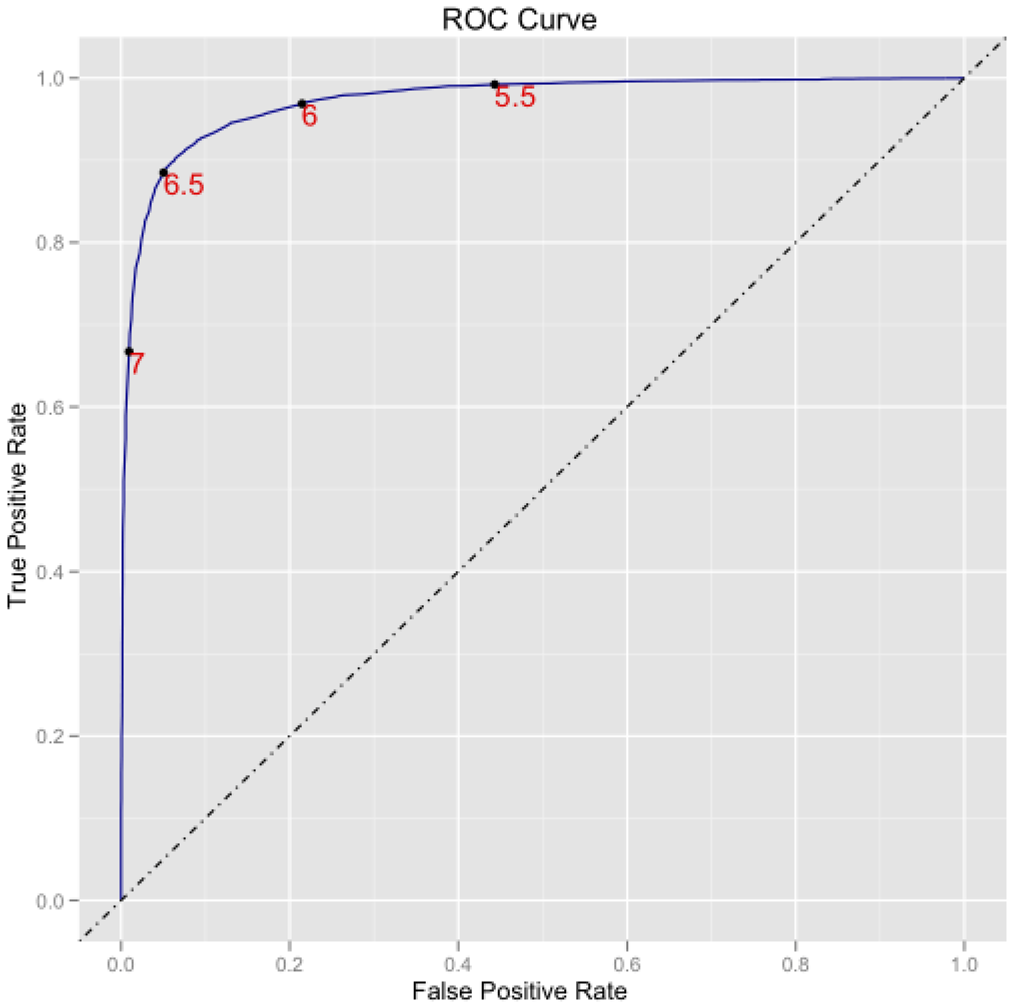

Figure 3: ROC Curve corresponding to Figure 2. Selected thresholds are marked on the curve.

ROC curves plot the false positive rate on the x-axis and the true positive rate on the y-axis, as we vary the threshold. The point $(0,0)$ corresponds to rejecting everything; the point $(1,1)$ corresponds to accepting everything. The ideal point is $(0,1)$: accept all positive instances and reject all negative instances. The line $x=y$ corresponds to random guessing: that is, a procedure that assigns each instance a score uniformly drawn from (in this example) the interval $[1,8]$ without even checking if the instance is positive or negative.

The ROC curve represents the tradeoff between true positives and false positives that we make as we increase the threshold from accepting everything to rejecting everything. Figure 3 gives the ROC curve for our example, with a few example thresholds marked on the curve.

The area between the ROC curve and the $x=y$ line can be considered a measure of accuracy; the smaller that area, the more the scoring procedure is like random guessing. The larger the area, the better separated the two classes are. We can use the curve to help us decide how to set a threshold that will give us the most acceptable tradeoff between true positives and false positives. For this example, we would probably want to select a threshold somewhere between 6 and 6.5.

In Conclusion

Some points to remember:

- Classifier and diagnostic test performance are not one-dimensional.
- Different fields use different (but related) measures of accuracy.
- Classifier and diagnostic test performance depend on the relative cost of Type I and Type II errors, as well as on the proportion of positive and negative instances in the population of interest.