Five Minute Science Lesson: Test Accuracy Metrics

Raymond Nelson

The practical goal of any scientific test is to *classify, predict,* and *quantify* a phenomenon of interest, often referred to as an *unknown parameter*, for which neither deterministic observation (which is theoretically perfect), nor physical measurement (subject only to random measurement error) are possible. Quite often, such as when dealing with human behavior and social phenomena, the unknown parameter of interest is amorphous. That is, there is no physical substance, and therefore no physical unit of measure. For this reason, all use of scientific testing methods for the processes of classifying, predicting, and quantifying will rely on probability theory and statistics as a foundation. For this reason, it is useful to understand the conceptual vocabulary of test accuracy metrics.

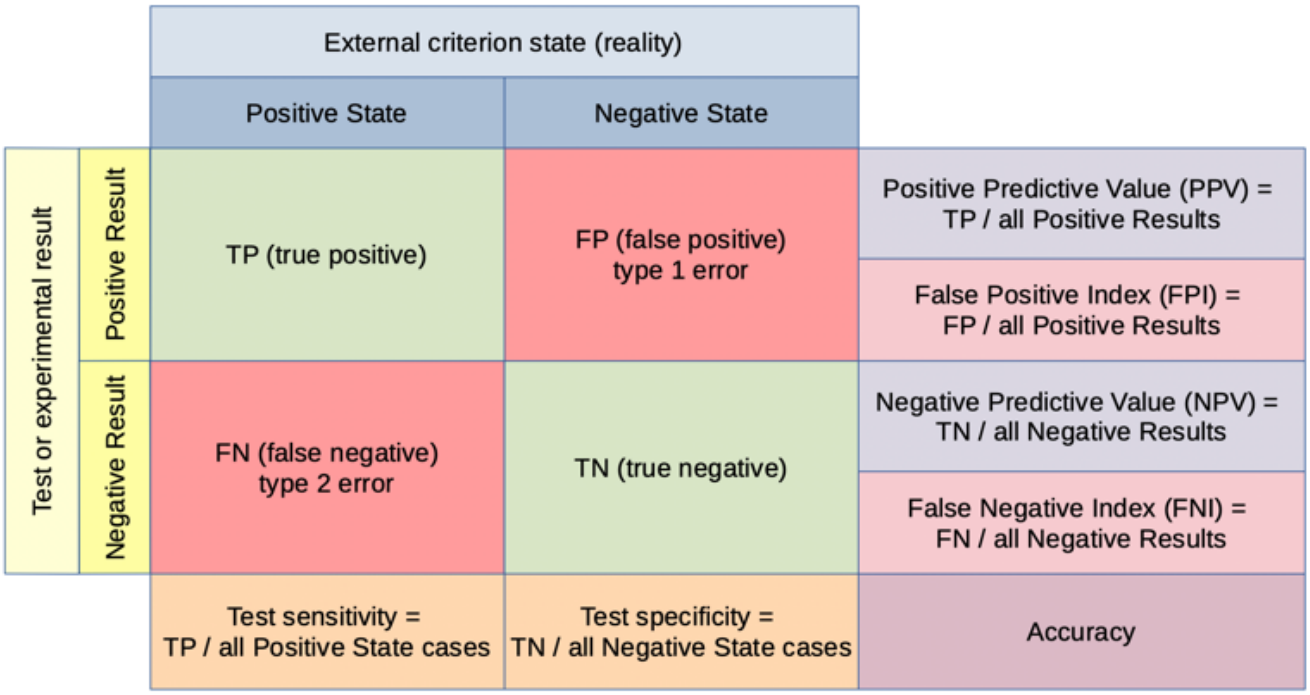
The purpose of any scientific test is to quantify a phenomenon of interest that cannot be easily subject to direct physical measurement. Actual measurement requires both a physical phenomena and a standardized unit of measure for that phenomena. Scientists and engineers can define a standardized unit of measure for any physical phenomena. One interesting aspect of scientific validity is described by the *Buckingham pi theorem*, for which a consequence of the theorem is that conclusions about quantitative questions, if answered through procedures based in mathematical and logical proof, will be invariant around the unit of measure. In other words valid conclusions will be exactly the same regardless of the unit of measurement. For example, using the Imperial measurement system that is familiar to adults in the U.S., a child who is 39.37 inches tall is exactly one-half the height of an adult of an adult who is 78.74 inches tall. Using metric measurements, common to most countries, that same child is 100 cm tall and the 200 cm adult is still exactly twice the height of the child. Another example involves common

units of temperature for which a daily temperature of 15 degrees Celsius does not indicate exactly one half the thermodynamic activity (heat) of 30 degrees Celsius. Using the Fahrenheit temperature scale those same temperatures are 59 degrees and 86 degrees. Mathematical inferences using the Celsius and Fahrenheit scales are not valid. (This is the reason that physicists use the Kelvin scale to mathematically model thermodynamic energy transfer.) It is important, for all experts who make use of scientific tests, to have some understanding of both basic principles of measurement and the basic concepts of probabilities and test outcomes.

Scientific tests, because they are fundamentally probabilistic, are not expected to be infallible or deterministic. Anyone who is interested in asserting any level of expertise in the social sciences – especially when attempting to explain the actual meaning of a scientific test or experiment – will be obligated to learn to understand something about statistics and probability. (Interestingly, scientist have also found, at the atomic and subatomic level, that attempting to understand some physical phenomena also depends on an understanding of probability theory.) More practically, one easily observable clue that can be observed when dealing with pseudoscience is a tendency towards overconfident claims that one's conclusions are virtually infallible – which serves to attempt to make a powerful social/psychological impression, and also relieves the burdensome task of actually quantifying the level of confidence and margin of uncertainty associated with a conclusion. Actual science is very often fraught with uncertainty. Theoretical scientists, applied scientists, and expert practitioners alike are expected to know that their task is to try to find ways to reduce that uncertainty – and to know that this begins by actually quantifying the uncertainty. From a practical perspective, having some understanding of probability and probabilistic test outcomes will permit more useful and realistic discussion of the actual value of a test result.

A number of useful concepts have emerged around our scientific conclusions and scientific knowledge, including discussions about the margin of error or uncertainty, degree of strength, confidence interval, credible intervals, likelihoods, likelihood ratios, conditional probabilities, and more. The most common way of framing our discussions about the degree of certainty associated with scientific tests, and scientific experiments, is to talk about accuracy – which, in reality, can mean a number of different things depending on the practical problem of interest. When discussion the accuracy of a scientific test it is useful to have a coherent and well-organized conceptual vocabulary that will be easily recognizable to other professionals. Figure 1 shows is a graphic that illustrates the relationships between a number of conceptual terms that are useful towards understanding test accuracy.

Figure 1. Conceptual terms for understanding accuracy.



Criterion state (external criterion): is any phenomena of interest that we wish to quantify, classify or predict. This is sometime referred to as an *unknown parameter* or *unknown phenomena* of interest.

Test result: often refers to the categorical test result. Scientific test results are not discussed in terms of pass or *fail* – though in some contexts there may be a tendency towards practical interpretations at this level – but instead use the terms *positive* and *negative* to signify whether the likelihood is sufficient to support a categorical conclusion about the presence or absence of the unknown parameter.

Positive result: is a term that signifies when a probabilistic test result supports a categorical conclusion that the unknown parameter or unknown phenomena of interest is *present* in a case.

Negative result: signifies when a probabilistic test result supports a categorical conclusion that the unknown parameter or unknown phenomena of interest is *absent* in a case.

Positive state: refers to whether the unknown parameter or unknown phenomena of interest is actually *present* (in reality) for a case.

Negative state: refers to whether the unknown parameter or unknown phenomena of interest is actually *absent* (in reality) for a case.

True positive (TP): describes a *positive result* that concurs with a *positive state.* A test has *correctly* identified the *presence* of the unknown phenomena of interest for a case.

True negative (TN): describes a *negative result* that concurs with a *negative state.* A test has *correctly* identified the *absence* of the unknown phenomena of interest for a case.

False positive (FP): describes a p*ositive result* that concurs with a *negative* state. A test has *incorrectly* identified the *presence* of the unknown phenomena in a case for which the external criterion state (reality) is actually negative (the unknown phenomena of interest is actually *absent*). FP is sometimes calculated as 1-specificity. However, the 1-specificity calculation will be incorrect for tests that include the use of inconclusive classifications (when a probabilistic result is not statically significant for either positive or negative classification).

False negative (FN): describes a negative result that concurs with a *negative state*. A test has *incorrectly* identified the *presence* of the unknown phenomena in a case for which the external criterion state (reality) is actually positive (the unknown phenomena of interest is actually *present*). FP is sometimes calculated as 1-sensitivity. However, the 1-sensitivity calculation will be incorrect for tests that include the use of inconclusive classifications (when a test result is neither positive nor negative).

Positive predictive value (PPV): refers to the proportion of TP outcomes to (TP + FP) outcomes for a group of cases for which the actual positive state or negative state is known. (Data scientists sometimes refer to these as *labelled cases*.) Can be useful to estimate the likelihood that a positive result is *correct* for an unknown case. However, PPV is non-resistant to differences in the prior proportion of positive state and negative state cases (base rate or incidence rate). That is, PPV – the likelihood that a positive result is correct – will be a function of the proportion of positive state cases in the group of cases, in addition to being influenced by the test sensitivity and FP rates.

Negative predictive value (NPV): refers to the proportion of TN outcomes outcomes to (TN + FN) outcomes for a group of labelled cases (i.e., cases for which the actual positive state or negative state is known). Can be useful to estimate the likelihood that a negative result is *correct* for an unknown case. NPV is also non-resistant to differences in the prior proportion of positive state and negative state cases (base rate or incidence rate). NPV – the likelihood that a negative result is correct – will vary with both the test specificity rate and the proportion of positive state cases (i.e., prior probability, base rate or incidence rate) in a group of cases.

False positive index (FPI): refers to the proportion of FP outcomes to all positive (FP + TP) outcomes for a group of labelled cases. Can be useful to estimate the likelihood that a positive result is *incorrect* for an unknown case. FPI is non-resistant to group imbalance (i.e., differences in the proportion of positive state and negative state cases), and will vary with both the

prior base rate or incidence rate and the test sensitivity rate.

False negative index (FNI): refers to the proportion of FN outcomes to all negative (FN + TN) outcomes for a group of labelled cases. Can be useful to estimate the likelihood that a negative result is *incorrect* for an unknown case. FNI is non-resistant to group imbalance. FNI will vary with both the test specificity rate and also will differences in the proportion of negative state and positive state cases (prior base rate or incidence rate).

Sensitivity: refers to the proportion of a group of positive state cases for which a test can *correctly* identify the *presence* of the unknown phenomena of interest. Test sensitivity, because it is calculated only within the subgroup of positive state cases, is resistant to differences in prior incidence rate. That is, test sensitivity will be invariant to group imbalance.

Specificity: refers to the proportion of a group of negative state cases for which a test can *correctly* identify the *absence* of the unknown phenomena of interest. Test specificity is calculated within the subgroup of negative state cases and is resistant to differences group imbalance (prior probability, base rate or incidence rate).