Clinical Statistics, Psychodiagnosis, and Bayes' Theorem

Stephen D. Benning
September 2008 (last updated December 2016)

I. Hit rates and the sensitivity and specificity of a test

In the course of assessment, one of our primary goals is to establish whether a person has a certain condition or doesn't; whether a person *is* a representative of something or not. These are fundamental, *binary* decisions that can be made based on cut scores on a distribution of test scores, assuming that one group has a lower mean score on the test and the other group has a higher mean score on the test (if this weren't the case, the test would be worthless for making these sorts of decisions). In all the subsequent examples, we'll assume that the group we're interested in finding through our test has the higher score on the test (or that they're *cases*), and the contrast group is the one that has the lower score (or *non-cases*).

Thus, one of the goals of assessment can be seen as correctly *classifying* individuals into certain groups that have utility in understanding the problems with which a particular patient presents and/or lead to different treatments, based on their group membership. For instance, even if there weren't anything you could do about it therapeutically due to limited resources or family engagement, it would be useful to understand whether an individual has conduct disorder, especially if it's in the context of other psychological disorders. You would also want to treat someone with a single major depressive episode very differently from a person with borderline personality disorder.

Assuming that we have some sort of quasi-definitive criterion against which the results of our test can be compared (like a genetic test to see whether a baby has the metabolic disorder phenylketonuria), the *hit rate* is the proportion of correct classifications that our test gives us:

$$hit\,rate = \frac{correct\,classifications}{total\,classifications}$$

However, this isn't usually sufficient to allow us to understand the diagnostic utility of a test, as there are two ways for a test to obtain classification “hits” and two ways to have test-based classifications “miss”. Take a look at the table below:

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	True positive (TP)	False positive (FP)
	- (absent)	False negative (FN)	True negative (TN)

Both true positives and true negatives contribute to the hit rate; conversely, both false positives and false negatives are test “misses”. From an assessment perspective, there are a couple of summary statistics that are important to calculate: How often the test hits when people are in fact members of the group it says they're in (or *sensitivity*) and how often the test hits when people are in fact not members of the group its says they're not (or *specificity*). Using the notation in the table above,

$$\boxed{sensitivity = \frac{TP}{TP+FN} \qquad specificity = \frac{TN}{TN+FP} \qquad hit\,rate = \frac{TP+TN}{TP+TN+FN+FP}}$$

Note how the hit rate conflates true positives and true negatives into a summary statistic.

Ideally, a test will have high sensitivity and high specificity; however, in practice, this is very difficult to achieve, especially without considerable expense. Thus, a two-stage testing methodology is often used. In the first stage, a test with high sensitivity and lower specificity is used to catch virtually all possible cases and is administered to everyone of interest (think of the tests used in National Depression Screening Day, for instance). If a positive result is obtained, a more expensive test with high sensitivity and specificity is run (a person may be called in for a clinical interview to follow up a positive result in National Depression Screening Day).

All things being equal, you can increase the sensitivity of the test by decreasing the cut score used to identify cases (at the expense of reducing specificity), whereas you can increase the test's specificity by increasing the cut score used to identify cases (at the expense of reducing sensitivity). For instance, if you assume that if an individual who endorses even one item on the National Depression Screening Day test is in fact depressed, you're guaranteed to catch nearly every person who's depressed – at the cost of having almost everyone who takes the test appear depressed and hence wasting lots of clinical time to confirm a ton of false positives. Likewise, if you insist that only those who endorse all items on the National Depression Screening Day test are the only ones you really want to call depressed, you'll likely weed out just about everyone who isn't truly depressed, but you'll also fail to assess a large number of people who are actually depressed (just not enough to endorse *every* item) and hence fail to provide clinical care to a large number of people who need it.

Let's look at how tests with different properties might appear in a four-way table.

High sensitivity, high specificity

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	2 (FP)
	- (absent)	5 (FN)	48 (TN)

sensitivity = 45 / (45+5) = 45/50 = .900
specificity = 48 / (48+2) = 48/50 = .960
hit rate = (45+48) / (45+48+5+2) = 93/100 = .930

High sensitivity, low specificity

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	48 (FP)
	- (absent)	5 (FN)	2 (TN)

sensitivity = 45 / (45+5) = 45/50 = .900
specificity = 2 / (2+48) = 2/50 = .040
hit rate = (45+2) / (45+2+5+48) = 47/100 = .470

Low sensitivity, high specificity

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	2 (TP)	5 (FP)
	- (absent)	48 (FN)	45 (TN)

sensitivity = 2 / (2+48) = 2/50 = .040
specificity = 45 / (45+5) = 45/50 = .900
hit rate = (2+45) / (2+45+5+48) = 47/100 = .470

Low sensitivity, low specificity

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	5 (TP)	48 (FP)
	- (absent)	45 (FN)	2 (TN)

sensitivity = 5 / (5+45) = 5/50 = .100
specificity = 2 / (2+48) = 2/50 = .040
hit rate = (5+2) / (5+2+48+45) = 7/100 = .070

Note how the high sensitivity/low specificity and low sensitivity/high specificity tests have exactly the same hit rate but for two very different reasons. Also note how if you just flip the values for the present and absent values in the low sensitivity/low specificity table, you get the high sensitivity/high specificity table. In other words, if you've got what looks like a really, really bad test, you might actually have a test coding error (perhaps someone reverse scored the whole test?).

So, where to set the cut score? Assuming that the distribution of test scores has an equal number of cases and non-cases, and that the variance of scores in both case and non-case groups is equal, the best place to locate the cut score is the point at which the two distributions overlap. This procedure will balance between false positives and false negatives, and it will also set equal numbers of true positives and true negatives.

II. Effects of base rates on measures of test accuracy: Positive and negative predictive value

All this is well and good, but the sensitivity and specificity of a test may be of relatively limited clinical utility. For many of the conditions for which we're interested in testing in the clinic, the number of cases and non-cases is far from equal, with there being many more non-cases than cases in the general population. Thus, we need statistics comparable to sensitivity and specificity that take into account the prevalence of the condition for which we're testing. These statistics are the *positive predictive value (PPV)* and *negative predictive value (NPV)*, respectively. They may be computed from a four-fold table:

$$\textit{positive predictive value} \;=\; \frac{TP}{TP+FP}$$

$$\textit{negative predictive value} \;=\; \frac{TN}{TN+FN}$$

Note how in the table, when the state of nature is presented in columns with labels on top (just like Omniscient Jones looking down upon the world with his shining beneficence) and the test outcomes are the rows with labels on the left, sensitivity and specificity are determined by the left and right columns, whereas positive and negative predictive value are determined by the top and bottom rows, respectively.

	State of nature / Quasi-infallible criterion results			
		+ (has condition)	- (has not condition)	
Result of test	+ (present)	45 (TP)	2 (FP)	**PPV** **TP/(TP+FP)**
	- (absent)	5 (FN)	48 (TN)	**NPV** **TN/(TN+FN)**
		sensitivity **TP/(TP+FN)**	**specificity** **TN/(TN+FP)**	

But what if all you've got are sensitivity, specificity, and the prevalence of your rare condition for which you're screening, without access to a four-fold table? You're in luck; PPV and NPV can be calculated from those statistics, too.

$$\textit{positive predictive value} \;=\; \textit{prevalence} \;\times\; \frac{\textit{sensitivity}}{\left(\textit{prevalence}\times\textit{sensitivity}\right) \;+\; \left(1-\textit{prevalence}\right)\times\left(1-\textit{specificity}\right)}$$

$$\textit{negative predictive value} \;=\; \left(1-\textit{prevalence}\right) \;\times\; \frac{\textit{specificity}}{\textit{specificity}\times\left(1-\textit{prevalence}\right) \;+\; \textit{prevalence}\times\left(1-\textit{sensitivity}\right)}$$

where the *prevalence* can be computed as follows:

$$\textit{prevalence} \;=\; \frac{TP+FN}{TP+TN+FN+FP}$$

Let's look at how positive and negative predictive values work in the tables of data that were discussed above with respect to sensitivity and specificity, in which the prevalence was always .500.

High sensitivity, high specificity

	State of nature / Quasi-infallible criterion results		
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	2 (FP)
	- (absent)	5 (FN)	48 (TN)

PPV= 45 / (45+2) = 45/47 = .957

= .500 * (.900 / {[(.500*.900] + [(1-.500)*(1-.960)]}) = .50*.90 / {.45+.02} = .45/.47 = .957

NPV = 48 / (48+5) = 48/53 = .906

= (1-.500) * ([.960 / {[(1-.500)*.960] + [.500*(1-.900)]}) = .50*.96 / {.48+.05} = .48/.53 = .906

High sensitivity, low specificity

	State of nature / Quasi-infallible criterion results		
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	48 (FP)
	- (absent)	5 (FN)	2 (TN)

PPV= 45 / (45+48) = 45/93 = .484

= .500 * (.900 / {[(.500*.900] + [(1-.500)*(1-.040)]}) = .50*.90 / {.45+.48} = .45/.93 = .484

NPV= 2 / (2+5) = 2/7 = .286

= (1-.500) * (.040 / {(1-.500)*[.040] + [.500*(1-.900)]}) = .50*.04 / {.02+.05} = .02/.07 = .286

Low sensitivity, high specificity

	State of nature / Quasi-infallible criterion results		
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	2 (TP)	5 (FP)
	- (absent)	48 (FN)	45 (TN)

PPV= 2 / (2+5) = 2/7 = .286

= .500 * (.040 / {[(.500*.040] + [(1-.500)*(1-.900)]}) = .50*.04 / {.02+.05} = .02/.07 = .286

NPV= 45 / (45+48) = 45/93 = .484

= (1-.500) * (.900 / {[(1-.500)*.900] + [.500(1-.040)]}) = .50*.90 / {.45+.48} = .45/.93 = .484

Low sensitivity, low specificity

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	5 (TP)	48 (FP)
	- (absent)	45 (FN)	2 (TN)

PPV= 5 / (5+48) = 5/53 = .094
= .500 * (.100 / {[(.500*.100)] + [(1-.500)*(1-.040)]}) = .50*.10 / {.45+.02} = .05/.47 = .094
NPV= 2 / (2+45) = 2/47 = .043
= (1-.500) * (.040 / {[(1-.500)*.040] + [.500*(1-.100)]}) = .50*.04 / {.02+.45} = .02/.47 = .043

Note how neither PPV and NPV look good when either sensitivity or specificity are high, and both look abysmal when sensitivity and specificity are both low. However, let's see how these values change when we manipulate the prevalence of the condition for which we're screening, keeping exactly the same sensitivity and specificity statistics.

High sensitivity, high specificity, prevalence = .100 (multiply values in original right column by 9)

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	18 (FP)
	- (absent)	5 (FN)	432 (TN)

PPV= 45 / (45+18) = 45/63 = .714
= .100 * (.900 / {[(.100*.900)] + [(1-.100)*(1-.960)]}) = .10*.09 / {.09+.036} = .09/.126 = .714
NPV= 432 / (432+5) = 432/437 = .989
= (1-.100)*(.960 / {[(1-.100)*.960] + [.100*(1-.900)]}) = .10*.96/{.864+.010} = .864/.874 = .989

High sensitivity, high specificity, prevalence = .010 (multiply values in original right column by 99)

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 (TP)	198 (FP)
	- (absent)	5 (FN)	4752 (TN)

PPV= 45 / (45+198) = 45/243 = .185
= .010*(.900 / {[(.010*.900)] + [(1-.010)*(1-.960)]}) = .01*.90 / {.009+.040)} = .009/.049 = .185
NPV= 4752 / (4752+5) = 4752/4757 = .999
= (1-.010)*(.960 / {[(1-.010)*.960] + [.010*(1-.900)]} = .99*.96/{.950+.001} = .950/.951 = .999

What happened to our beautiful test? It now looks even better for screening people out as non-cases, but we devised our test to detect cases, and it seems almost worthless for this purpose! Conversely, if we shift the prevalence closer to 1, it'll identify lots of cases, but it'll also fail to screen out people who are non-cases (with a prevalence of .900, our test will have PPV = .995 and NPV = .484; with a prevalence of . 990, it'll have PPV = .9996 and NPV = .088). Clearly, the prevalence has something to do with the usefulness of our test...but what? The 1% prevalence table is highlighted, as it will be important later on.

III. Two ways to think about probability: The frequentist and subjectivist schools

To answer that question, we'll look at probability a couple of different ways. Since the dawn of statistical time, we've been trained to think of probability in what's usually called the *frequentist* way of understanding probability. That is, we think about how frequently a specific thing like a patient with depression will be observed, given a pool of possible observations, like all patients that walk into a clinic. In this case, all we care about is the relative frequency of cases in the grand scheme of all observations.

Indeed, absent any other information, the *base rate* is the frequentist probability associated with observing a case (e.g., patient with depression) in the population of interest – that is, the *base rate* is the *prevalence*. With a base rate of .500, something is as likely to be a case as not, so there are both lots of cases and non-cases. As the base rate approaches zero, there are many more non-cases than cases; conversely, as the base rate approaches 1, there are many more cases than non-cases.

From a frequentist perspective, greatest classification accuracy in the absence of any test information comes from just assuming that everyone is from the more frequent group. Indeed, in the case of extraordinarily rare events, having more information may cause you to be less accurate overall! For example, because the prevalence of schizophrenia is approximately 1%, in the long run, if you classify everyone you see as not having schizophrenia, you'll be right 99% of the time. Contrast that with the hit rate of our high sensitivity, high specificity test, which has a hit rate of 96% in the 1% prevalence case [(45+4732)/(45+4732+5+198)]. Thus, you'll have a higher hit rate by not administering your test and just *betting the base rates*!

The reason for this is that the number of false positives and false negatives (particularly the false positives for rare cases and particularly the false negatives for rare non-cases) exceeds the quantity (base rate) * (number of individuals tested). However, if you were to test all individuals who appeared to be cases upon the first test in the 1% prevalence example above with a second test with the same sensitivity and specificity, you would get the following result:

		State of nature / Quasi-infallible criterion results	
		+ (has condition)	- (doesn't have condition)
Result of test	+ (present)	45 * sensitivity = 40.5 (TP2)	198 * (1-specificity) = 7.92 (FP2)
	- (absent)	45 * (1-sensitivity) = 4.5 (FN2)	198 * specificity = 190.08 (TN2)

Now for the second test, PPV = [40.5 / (40.5+7.92)] = .836 and NPV = [190/(190+4.5)] = .977, which is much better. Why does the second schizophrenia test fare so much better, even though it's got the same diagnostic statistics as the first test? Because you've changed the base rate from .010 to [45/(45+198)], or .185 – the PPV of the first test. Thus, you've made cases with schizophrenia 18.5 times more common in the group of potential cases in your second test, so the second test has more favorable probabilities with which to work. The hit rate of this two-stage battery is [(TN+TP2+TN2)/(TN+TP2+TN2+FN+FP2+FN2)] = (4752+40.5+188)/(4752+40.5+188+4.5+10+5) = .996 – better than the base rate.

Note that TP and FP from the original test don't factor into this hit rate calculation because their values are split up completely among TP2, TN2, FN2, and FP2. Put another way, we retested everyone who tested as a case on the first test to reduce our errors of classification, so their data are transferred and split up in the cells of the second test; thus, they're "removed" from the cells of the first test. However, we have to take into account that the first test missed some real cases who were never identified as such and hence were lost forever – these are diagnostic misses that can't be recovered and hence must go into our calculation of the hit rate.

This is all well and good as a way of thinking about probability, but it assumes something very important: There's something physical we can count to assign frequentist probabilities. What if instead, we're interested in probabilities of things that we can't count readily, like a clinician's belief that a given person is a case? At this, frequentists throw up their hands unless this "belief" can somehow be operationalized into something that's countable and tallyable.

Not so with those who hold with the subjective school of probability. All that matters is that someone be able to articulate about how likely something is to be true on a scale of 0 to 1. Another way of looking at this is to have people give odds on how likely something is to be true, which can then easily be converted into a probability. Do you think that something's roughly three times as likely to be true compared to it being false? You've then just described an odds of belief of 3:1, which can be converted easily into a probability of 3/(3+1) = 3/4 = .750.

That's it.

No counting needed, no frequency tables required to compute – you just generated what's called a *prior probability*, which refers to the probability you believe something to be true before some other data impinges on your beliefs. Some criticize the subjective school of probability for this ability to use a probability pulled from any source – data, a black box, or an orifice of choice – but this is in fact one of the strengths of the subjective school of probability. It provides a way of modeling the garbage that some people have in their heads (remember "Why I Do Not Attend Case Conferences"?) to help understand why, when confronted with exactly the same data, people can come to opposite diagnostic conclusions.

IV. Modifying subjective probabilities: Bayes' Theorem, with a specific clinical application

However, in the clinic, we're often interested in allowing our assessments to modify those prior probabilities. The probability that results from some sort of updating of the prior probability is called the *posterior probability*. The Rev. Thomas Bayes initially discussed these kinds of *conditional probabilities* (probabilities that give how probable something is given that something else has occurred previously); later theorists expanded on the implications of his work and codified what is now known as Bayes' Theorem.

We'll start with a familiar example to see how Bayes' Theorem works. Let's return to our test to identify schizophrenia (a disorder with a 1% lifetime prevalence); the test has a sensitivity of .900 and a specificity of .960. We want to know the probability of believing that someone has schizophrenia given a positive test result. To do this, we also need to know the probability of believing that person has schizophrenia before any test data (which could be modeled as the prevalence), the probability that a positive test result will be obtained, and the probability that a positive test result will be obtained given that the person actually has schizophrenia (Scz).

We can write out Bayes' Theorem for this problem as follows:

$$P\left(\text{Scz}|\text{Test}^{+}\right) \;=\; P\left(\text{Scz}\right) \times \frac{P\left(\text{Test}^{+}|\text{Scz}\right)}{P\left(\text{Test}^{+}\right)}$$

where:

P(Scz|Test^{+}) is the *posterior probability of having schizophrenia*, given a positive test result
P(Scz) is the *prior probability of having schizophrenia*
P(Test^{+}) is the *prior probability of a positive test result*
P(Test^{+}|Scz) is the conditional probability of a positive test result, given having schizophrenia

P(Test^{+}|Scz) is also the quantity TP/(TP+FN), which is the *sensitivity* of the test; in this case, it's .900.
P(Scz) is .010, the *prevalence* of the disorder.
P(Test^{+}) is also the quantity (TP+FP)/(TP+FP+FN+TN), which is (45+198)/(45+198+5+4752) = .049.
Thus, P(Scz|Test^{+}) = .010*(.900*/.049) = <u>.185</u>.

The value P(Test^{+}|Scz) represents the probability of believing that a person has schizophrenia given a positive test result. Note that this posterior probability is identical to the value of PPV for this test for schizophrenia. If we take the obtained posterior probability and plug it into Bayes' Theorem again (which is just like running the second stage of a two-stage diagnostic testing sequence as before), we find that:

P(Test^{+}) = (TP2+FP2)/(TP2+FP2+FN2+TN2), which is (40.5+8)/(40.5+8+4.5+190) = .200
Thus, P(Scz|Test^{+}) = .185*(.900/.200) = <u>.835</u>.

Again, note that P(Scz|Test^{+}) is identical to the PPV for the second stage of the test battery, showing how the PPV of a test is a frequentist representation of the subjective probability of believing a person to have schizophrenia, when the initial base rates are used as the prior probability. Thus, we can write:

$$P\left(\text{Scz}|\text{Test}^{+}\right) \;=\; \textit{prevalence} \;\times\; \frac{\textit{sensitivity}}{\left(\textit{prevalence}\times\textit{sensitivity}\right) \;+\; \left(1-\textit{prevalence}\right)\times\left(1-\textit{specificity}\right)}$$

This form of Bayes' Theorem is easier to use if you believe the population base rate isn't a good prior probability to use. Let's say that a clinician has a gut feeling that the probability of seeing someone with schizophrenia in her own work is .200. Why? Dunno – perhaps she just guessed it, perhaps she reviewed her records and saw that one out of every five patients had a diagnosis of schizophrenia, or perhaps a voice in her head commanded her to use .200 as the prior probability. *Subjectivist probability calculations are completely agnostic about the source of the priors fed into them.* So, dutifully cranking through Bayes' Theorem for the first test stage, using .200 as the "prevalence", or the prior P(Scz),

P(Scz|Test^{+}) = .200 * (.900/[.900*.200+(1-.960)(1-.200)]) = .18 / (.18 + .032) = <u>.849</u>

Note how this value is even higher than that when we used our two-stage test battery to diagnose schizophrenia in the population. Why is that? Because the prior probability we used in this case was higher than the posterior probability returned by the first stage of diagnostic testing. This equation also allows us to calculate these probabilities without having to have the four-fold table with actual counts in front of us. We could manufacture one to help us compute the value of P(Test^{+}|Scz), but that's a lot of computational work to find values to plug in, when the PPV formula lets us compute right away.

V. Bayes' Theorem and odds

Another way to think about this problem is to think about the odds of something being believable vs. not. *Odds* are frequently used in diagnostic statistics to tell us if a certain factor makes something more or less likely; odds >1 indicate the factor makes something more likely to happen, and odds < 1 indicate the that factor makes something less likely to happen.

For instance, having an early childhood loss (such as the death of a parental figure) may make the odds of experiencing depression more likely (hence odds of 2, 3, or maybe even 10), whereas having a strong network of social support may make the odds of experiencing depression less likely (hence odds of 1/2, 1/3, or possibly even 1/10). The further toward infinity the odds are, the more likely the event is to happen; the further toward zero they is, the less likely the event is to happen. Odds of 1 means that an event is just as likely to happen as not and corresponds with a probability of .5.

In our case, to generate odds that someone has schizophrenia from a prior probability (whatever its source may be), we can use the following formula:

$$O(\mathrm{Scz}) \;=\; \frac{P(\mathrm{Scz})}{P(\sim\!\mathrm{Scz})}$$

where:

O(Scz) is the odds of having schizophrenia vs. not having schizophrenia
P(Scz) is the probability of having schizophrenia
P(~Scz) is the probability of not having schizophrenia, or 1 – P(Scz)

Thus, with our clinician's assumed P(Scz) of .200, we get the following:

O(Scz) = .200 / (1-.200) = .200 / .800 = <u>.250</u>

In this scenario, the clinician is four times as likely to believe that any individual in her clinic does not have schizophrenia as to believe that any individual has it.

Formally speaking, to generate the odds associated with believing a person has schizophrenia, given that person's testing positive for it on our initial assessment, we use the following formula:

$$O(\mathrm{Scz}|\mathrm{Test}^{+}) \;=\; O(\mathrm{Scz}) \;\times\; \Lambda(\mathrm{Scz}|\mathrm{Test}^{+}), \;\; \text{and}\,\Lambda(\mathrm{Scz}|\mathrm{Test}^{+}) \;=\; \frac{P(\mathrm{Test}^{+}|\mathrm{Scz})}{P(\mathrm{Test}^{+}|\sim\!\mathrm{Scz})} \;\left(=\;\frac{L(\mathrm{A}|\mathrm{B})}{L(\sim\!\mathrm{A}|\mathrm{B})}\right)$$

where:

$O(\mathrm{Scz}|\mathrm{Test}^{+})$ is the odds of having schizophrenia given a positive test vs. not having Scz
$\Lambda(\mathrm{Scz}|\mathrm{Test}^{+})$ & L(A|B)/L(~A|B) are the *likelihood ratio* that a positive test has Scz vs. doesn't
$P(\mathrm{Test}^{+}|\mathrm{Scz})$ is the probability that someone who tests positive for schizophrenia actually has it
$P(\mathrm{Test}^{+}|\sim\!\mathrm{Scz})$ is the probability that someone who tests positive for schizophrenia doesn't have it

Note that $P(\mathrm{Test}^{+}|\mathrm{Scz})$ is equal to the test's sensitivity; $P(\mathrm{Test}^{+}|\sim\!\mathrm{Scz})$ is also (1- specificity). So to put this form of Bayes' Theorem in terms of the various probabilities we know,

$$O(\mathrm{Scz}|\mathrm{Test}^{+}) \;=\; \frac{P(\mathrm{Scz})}{P(\sim\!\mathrm{Scz})} \;\times\; \frac{P(\mathrm{Test}^{+}|\mathrm{Scz})}{P(\mathrm{Test}^{+}|\sim\!\mathrm{Scz})}$$

or in terms of diagnostic statistics,

$$O(\mathrm{Scz}|\mathrm{Test}^{+}) \;=\; \frac{P(\mathrm{Scz})}{P(\sim\!\mathrm{Scz})} \;\times\; \frac{\text{sensitivity}}{1-\text{specificity}}$$

Using this formulation of Bayes' Theorem – with our prior probability of .200, sensitivity of .900, and specificity of .960 – we can calculate the odds as follows: O(Scz|Test$^+$) = [.200/(1-.200)]*[.900/(1-.960)] = .25*22.5 = 5.625. Thus, it is now 5.625 times as likely that the clinician believes this patient to have schizophrenia as that she believes the patient does not have schizophrenia. Ooooo-K, how do we get this back into a probability that we can use in future computations? It turns out that the formula for converting odds into probabilities is an easy one. In this case,

$$P(\text{Scz}|\text{Test}^+) = \frac{O(\text{Scz}|\text{Test}^+)}{O(\text{Scz}|\text{Test}^+)+1}$$

Thus, P(Scz|Test$^+$) = 5.625/(5.625+1) = .849, which is exactly the same answer that we got for the PPV way of calculating this probability. In our two-stage diagnostic testing with a prevalence of 1%, you can just multiply the sensitivity and (1-specificity) of each stage together to obtain the odds. Therefore, O(Scz|Test$^+$) = [(.010/(1-.010)]*[.900/(1-.040)]*[.900/(1-.040)] = 5.114; P(Scz|Test$^+$) = 5.114/(5.114+1) = .836, precisely the same answer as we got from the frequentist approach and PPV.

VI. Bayes' Theorem: Generally Speaking

We can represent the general form of Bayes' Theorem as:

$$P(X|Y) = P(X) \times \frac{P(Y|X)}{P(Y)}$$

where:

P(X|Y) is the *posterior probability of X*, given Y
P(X) is the *prior probability of X*
P(Y) is the *prior probability of Y*
and P(Y|X) is the conditional probability of Y, *given X*

In a diagnostic framework,

$$\boxed{P(Condition|Test^+) = P(Condition) \times \frac{P(Test^+|Condition)}{P(Test^+)}}$$

which is identical to the PPV for the test, so the PPV formula can be used for computing the value P(Condition|Test$^+$), especially when the computation of P(Test$^+$) is tricky. For ease of computation, we can also represent the odds form of Bayes' Theorem as follows:

$$\boxed{O(\text{Condition}|\text{Test}^+) = \frac{P(\text{Condition})}{P(\sim\text{Condition})} \times \frac{\text{sensitivity}}{1-\text{specificity}}}$$

and to return the probability from the odds,

$$\boxed{P(\text{Condition}|\text{Test}^+) = \frac{O(\text{Condition}|\text{Test}^+)}{O(\text{Condition}|\text{Test}^+)+1}}$$

VII. The reconciliation of subjectivist and frequentist probabilities?

One important thing to note from a historical perspective is that frequentists and subjectivists (also known as Bayesians, even though it's not clear that Bayes himself would have taken a truly subjective stance on the nature of probability) like to argue with each other about how the other side doesn't really get it. They whine, they moan, and they bitch. Almost incessantly, really. The frequentists complain that the lack of well-defined priors makes Bayesians hopelessly muddle-headed, and the Bayesians retort that they can explain things that leave the frequentists stumped because the Bayesians allow for subjective probability determinations.

However, from a practical standpoint, it may not matter if one is a frequentist or a subjectivist, as long as one always uses an empirically derived base rate as a prior to obtain posterior probabilities. Notice how when the frequentist base rate was used as the prior probability, the calculations returned exactly the same value for PPV and a Bayesian posterior probability (and could in fact reduce to the same equation). It's only when people insist on using wacky priors that don't derive from data that discrepancies in clinical judgments arise, from the standpoint of the material in this course. If you anchor your priors in data, you'll at least be able to replicate your train of thought if challenged about it in subsequent case conferences or court cases, unlike those who set their priors based on hunches or faulty data.

Nevertheless, it's worth noting that all of the clinical statistics described here have error terms associated with them that aren't discussed here. The errors for sensitivity and specificity can be derived using errors associated with proportions, as could the errors for PPV and NPV. Because the PPV has an error associated with it, so too does the Bayesian estimate of the probability of belief. Thus, a person will have a certain "spread" of belief around the point estimate that will tend to shrink as more and consistent experience is gained.

Therefore, it's possible that people refuse to alter their priors in the face of a mountain of contrary data because the spread of their priors is so narrow that it would take nearly a miracle of diagnostic revelation to budge the point estimate appreciably. This is what happens when overdiagnosis runs rampant, when clinicians see cases of ADHD everywhere or mistake borderline personality symptoms for "ultra-rapid cycling bipolar disorder". Beware having priors that are too narrowly spread in the absence of good data to firm them up; that way lies delusion, dilettante diagnosis, and despair.

VIII. The reference class problem

There still remains the vexing consideration of the relative infrequency of cases in the population as a whole. It would be much more convenient to be able to conduct some sort of prescreening that would allow you to make the probability of observing a case approach .5 and thus increase the diagnostic efficiency of your tests. One way of doing this is to find a way to preclassify a patient into a *reference class* that includes a narrower group of individuals that share characteristics and that also has a higher base rate of cases.

Let's take the case of generalized anxiety disorder (GAD). DSM-IV-TR claims a one-year community prevalence of 3% for GAD. However, it also notes that "[i]n anxiety disorder clinics, up to a quarter of the individuals have [GAD] as a presenting or comorbid diagnosis." Let's also say that you're working in an anxiety disorders clinic and are assessing a new patient for GAD. Congratulations! By dint of your patient's being a member of the "patient of an anxiety disorder clinic" reference class, you've increased the probability of your patient having GAD eightfold from assuming the patient's membership in the "general population" reference class. Assuming a positive result on a diagnostic test with sensitivity of . 85 and specificity .75, your odds of identifying a true case of GAD are [.250/(1-.250)]*[.850/(1-.750)] = . 333 * 3.4 = 1.13, and your probability of believing this person to have GAD is .531. Without the clinical base rate, the case odds are [.030/(1-.030)]*[.850/(1-.750)] = .031*3.4 = .105, for a case probability of .

095.

Thus, if you don't narrow your reference class sufficiently, you might miss true cases by the sheer limitations that low base rate phenomena place on the diagnostic efficiency of your tests. Therefore, in your clinical work, one of the most important things to ask yourself is: What's the right base rate to use as my prior probability? Some hospitals or clinics might note the prevalence of a variety of different disorders in their own setting; if you work in one such hospital, that kind of local reference class is a great one to use. The mere fact that someone is showing up in your clinic suggests that the probability of finding just about any psychopathology is greater than it would be in the general population; after all, what's one of the major reasons why people seek psychological services?

However, there are other ways to narrow reference classes besides looking at the specific setting in which your patients present. For example, according to DSM-IV-TR, depression is twice to three times as common in women as men. If you're assessing a woman, use that information to narrow down the reference class to “woman”, who tend to have a 5%-9% point prevalence of depression in the community, compared to a 2%-3% point prevalence for men. Family history can provide another way of altering the reference class – DSM-IV-TR states that depression is 1.5-3 times as frequent in individuals with first-degree relatives who have had a major depressive episode.

There are a variety of ways to alter the reference class you're considering in completely justifiable ways to get your prior probabilities as close to .500 as possible before you administer a single test. Use data-driven methods for selecting the correct reference class, and you'll be giving your diagnostic tests much more power than you would otherwise. Use too generic a reference class, and the psychopathology you seek to assess may be so rare that not even the most powerful test available could hope to detect it. Though using the right base rate in your assessments requires some judgment, it's one of the most important things you can do when performing clinical assessments.

Glossary

Base rate – The rate at which cases are observed in the population as a whole. Ranging from 0 to 1, clinical diagnostic tests are most effective if the base rate is near .5.

Four-fold table – A table in which the two rows represent whether a test reveals a person meets a cutpoint for a condition or not, and the two columns represent whether the person (in the unseen state of nature or according to a quasi-infallible criterion measure) has a condition or does not. A true positive (TP) occurs when the test and the state of nature agree a person has a condition. A true negative (TN) occurs when the test and the state of nature agree a person does not have a condition. A false positive (FP) occurs when the test indicated the person has a condition but the state of nature is that the person does not. A false negative (FN) occurs when the test indicates the person does not have a condition but the state of nature is that the person does have the condition.

Frequentist probability – Probability is defined by how often a case is observed in the total number of observations. Requires observable, countable, measurable entities to be calculated.

Negative predictive power – The probability that a negative result on a test actually means that the person being assessed does not have the condition that would make him a case. Also, the frequentist representation of the probability of believing that an individual is a non-case. Operationalized in the four-fold table as TN/(TN+FN).

Odds – The relative probability of one thing happening vs. its not happening, or of a case being present vs. a non-case being present. Odds can be converted into probabilities using the formula P = O/(O+1).

Positive predictive power – The probability that a positive result on a test actually means that the person being assessed has the condition that would make him a case. Also, the frequentist representation of the probability of believing that an individual is a case. Operationalized in the four-fold table as TP/(TP+FP).

Posterior probability – The probability of believing something after a set of data is collected. This probability accounts for the effects that assessment data have on your belief that a patient is a case.

Prevalence – See *base rate*.

Prior probability – The probability of believing something before a set of data is collected. Before any assessment data are collected, this is the *base rate* in the frequentist interpretation of probability.

Reference class – The class of individuals for whom a base rate is defined. For a base rate defined in the population as a whole, the population as a whole is the reference class. A base rate for a condition found in a particular clinic's assessment practice is based on that clinic's patient composition as the reference class. Proper selection of the reference class can make relatively rare cases more common, making diagnostic testing more effective.

Sensitivity – The probability that a test will correctly detect cases as such. Operationalized as TP/(TP+FN).

Specificity – The probability that a test will correctly detect non-cases as such. Operationalized as TN/(TN+FP).

Subjectivist probability – Probability is defined as that of holding a certain belief. These probabilities are identical to the frequentist probabilities when the frequentist base rate is used as the prior, but values of the prior may be pulled from any source and are not constrained by frequentist interpretations.

Appendix: Derivations of PPV, NPV, and Bayesian formulas from the four-fold table

January 2013 (last updated December 2016)

Recall that from the four-fold table:
sensitivity = TP/(TP+FN)
specificity = TN/(TN+FP)
prevalence = (TP+FN)/(TP+FN+TN+FP)
PPV = TP/(TP+FP)
NPV = TN/(TN+FN)

The goal is to derive formulas for PPV and NPV that don't require us to have access to the four-fold table. These formulas would allow us to account for a "prevalence" that's not rooted in empirical, frequentist base rates. That way, we can just use prevalences and test statistics to calculate our probabilities of belief based on test data rather than having to know the precise number of observations in each cell of the table.

To aid in this derivation, I'm going to introduce one last quantity:
whole sample (W) = TP+FN+TN+FP
making prevalence = (TP+FN)/W

Now, recalling our initial definition of PPV, and using an algebraic trick to introduce W by representing 1 as (1/W)/(1/W) and multiplying through (because multiplying any quantity by 1 yields that quantity):

$$1)\quad PPV = \frac{TP}{TP+FP} = \frac{TP}{TP+FP} \times \frac{1/W}{1/W} = \frac{\frac{TP}{W}}{\frac{TP+FP}{W}}$$

Huh. Well, that doesn't seem to have gotten us anywhere. But I recognize TP as the numerator of sensitivity, and I can transform TP into sensitivity by dividing by (TP+FN). Thus, I can use a similar algebraic trick to include (TP+FN) into the equation and employ the commutative property of multiplication to move TP and TP+FN terms around in the numerator of the big ol' fraction:

$$2)\quad PPV = \frac{\frac{TP}{W}}{\frac{TP+FP}{W}} \times \frac{\frac{TP+FN}{TP+FN}}{\frac{1}{1}} = \frac{\frac{TP}{W} \times \frac{TP+FN}{TP+FN}}{\frac{TP+FP}{W} \times \frac{1}{1}} = \frac{\frac{TP+FN}{W} \times \frac{TP}{TP+FN}}{\frac{TP+FP}{W}}$$

Wait. Now we're getting somewhere! I recognize both prevalence and sensitivity in the numerator, so I can make some replacements and move more terms around! I can also expand out the denominator a bit:

$$3)\quad PPV = \frac{prevalence \times sensitivity}{\left(\frac{TP+FP}{W}\right)} = prevalence \times \frac{sensitivity}{\left(\frac{TP+FP}{W}\right)} = prevalence \times \frac{sensitivity}{\left(\frac{TP}{W}+\frac{FP}{W}\right)}$$

A few things strike me here. First, it's now explicit how PPV depends on the initial prevalence estimate – and is an update of the prevalence. Second, the denominator looks like the *total probability of getting a positive test result*, or P(Test^{+}) from the Bayesian formula! That is, the quantity (TP+FP)/W gives me the number of positive test results divided by the total number of people assessed with the test. That's a nifty result to keep in mind, but I didn't include P(Test^{+}) in the quantities I was concerned with above.

The third thing will really help move the derivation move forward. Notice that at the end of equation 1 and the beginning of equation 3, I've already shown that TP/W = prevalence x sensitivity, so I can substitute that into the denominator:

$$4)\quad PPV = prevalence \times \frac{sensitivity}{(prevalence \times sensitivity) + \frac{FP}{W}}$$

Now, I just need to represent FP/W in terms of sensitivity, specificity, and prevalence. The problem is that I don't have any terms that have FP as the numerator. Drat.

The only other place where FP even occurs in our definitions above outside of PPV and W is in the definition of specificity. Let's use the “one isn't necessarily the loneliest number” algebraic trick again to introduce the denominator for specificity and see what happens:

$$\begin{aligned} 5)\quad PPV &= prevalence \times \frac{sensitivity}{(prevalence \times sensitivity) + \frac{FP}{W}} \times \frac{\frac{1}{1}}{\frac{TN+FP}{TN+FP}} \\ &= prevalence \times \frac{sensitivity \times \frac{1}{1}}{(prevalence \times sensitivity) \times \left(\frac{TN+FP}{TN+FP}\right) + \frac{FP}{W} \times \left(\frac{TN+FP}{TN+FP}\right)} \\ &= prevalence \times \frac{sensitivity}{(prevalence \times sensitivity) \times \frac{1}{1} + \left(\frac{FP}{W} \times \frac{TN+FP}{TN+FP}\right)} \\ &= prevalence \times \frac{sensitivity}{(prevalence \times sensitivity) + \left(\frac{TN+FP}{W} \times \frac{FP}{TN+FP}\right)} \end{aligned}$$

This looks...sort of promising. But we still don't have things expressed as quantities we recognize fully. Hmm. Well, I know that TP+FN+TN+FP = W, so (TP+FN+TN+FP)/W = W/W = 1.

Expanding out, TP/W + FN/W + TN/W + FP/W = 1.
Rearranging terms, TN/W + FP/W = 1 – TP/W – FN/W.

Whoa.

Let's put this together another way:

(TN+FP)/W = 1 – (TP+FN)/W

Recognize that last term? (TP+FN)/W = *prevalence*! That means (TN+FP)/W = 1 – prevalence!

$$6)\quad PPV = prevalence \times \frac{sensitivity}{(prevalence \times sensitivity) + (1 - prevalence) \times \frac{FP}{TN+FP}}$$

One...last...term...to replace. Any other tricks up the algebraical sleeve like before? Actually, yes.

TN/(TN+FP) + FP/(TN+FP) = (TN+FP)/(TN+FP) = 1. See where I'm going with this?
FP/(TN+FP) = 1 – TN/(TN+FP), and TN/(TN+FP) = *specificity*, so FP/(TN+FP) = 1 – specificity!

$$7)\quad PPV \;=\; prevalence \;\times\; \frac{sensitivity}{(prevalence \times sensitivity) \;+\; (1 - prevalence) \times (1 - specificity)}$$

QEMFD. All right, so MF isn't proper Latin.

Similar algebraic legerdemain can provide a formula for negative predictive value:

$$8)\quad NPV \;=\; \frac{TN}{TN+FN} \;=\; \frac{1/W}{1/W} \times \frac{TN}{TN+FN} \;=\; \frac{\dfrac{TN}{W}}{\dfrac{TN+FN}{W}}$$

Multiplying in another 1 (this time, the denominator of specificity over itself) into the equations, we get:

$$9)\quad NPV \;=\; \frac{\dfrac{TN}{W}}{\dfrac{TN+FN}{W}} \times \frac{\dfrac{TN+FP}{TN+FP}}{\dfrac{1}{1}} \;=\; \frac{\dfrac{TN}{W} \times \dfrac{TN+FP}{TN+FP}}{\dfrac{TN+FN}{W} \times \dfrac{1}{1}} \;=\; \frac{\dfrac{TN+FP}{W} \times \dfrac{TN}{TN+FP}}{\dfrac{TN+FN}{W}}$$

Substituting in terms we already know and skipping ahead a few steps,

$$10)\quad NPV \;=\; (1 - prevalence) \;\times\; \frac{specificity}{(1 - prevalence) \times specificity \;+\; \dfrac{FN}{W}}$$

Multiplying in another 1 (this time, the denominator of sensitivity over itself) and switching around terms,

$$11)\quad NPV \;=\; (1 - prevalence) \;\times\; \frac{specificity}{(1 - prevalence) \times specificity \;+\; \dfrac{TP+FN}{W} \times \dfrac{FN}{FN+TP}}$$

We can replace (TP+FN)/W with prevalence, because that's how we defined it earlier:

$$12)\quad NPV \;=\; (1 - prevalence) \;\times\; \frac{specificity}{(1 - prevalence) \times specificity \;+\; prevalence \times \dfrac{FN}{FN+TP}}$$

then using more "X = 1 – Y" subtractive goodness from the definition of W to replace FN/(FN+TP) with 1 – sensitivity,

$$13)\quad NPV \;=\; (1 - prevalence) \;\times\; \frac{specificity}{(1 - prevalence) \times specificity \;+\; prevalence \times (1 - sensitivity)}$$

Again, QED.

But remember how I mentioned something about the Bayesian way of approaching probability earlier? Specifically, with the end product of equation 3?

Well, in the handout, we talked earlier about how:

PPV = P(Condition|Test$^+$), or the posterior probability of having a condition given a positive test,
sensitivity = P(Test$^+$|Condition), or the probability of a positive test given having the condition,
prevalence = P(Condition), or the frequentist base rate
TP/W + FP/W = P(Test$^+$), or the overall probability of having a positive test result,

So if you substitute the terms on the right side of each equality into equation 3, you get:

$$14) \quad P\left(Condition|Test^{+}\right) = P\left(Condition\right) \times \frac{P\left(Test^{+}|Condition\right)}{P\left(Test^{+}\right)}$$

The Bayesian equation for updating your probability of belief that a person has a condition given a test's sensitivity, base rate, and probability of a positive test result is a simple “find and replace” job on the PPV equation. Can you do a similar find and replace job on the NPV equation to give P(~Condition|Test$^-$)?

Appendix: From diagnostic statistics to Bayesian statistical analysis, in brief

December 2016

You may have heard about Bayesian methods for analyzing data statistically – how these analyses are better than the ones you learned in (under)grad courses that are all about the frequentist model. So, how would they work?

Bayes-ically, it all comes down to manipulating prior probabilities into posterior probabilities using the odds form of Bayes' Theorem:

$$\boxed{\begin{array}{lrcl} & P(\text{Condition}|\text{Test}^{+}) & = & \dfrac{O(\text{Condition}|\text{Test}^{+})}{O(\text{Condition}|\text{Test}^{+})+1}, \\ \textit{where} & O(\text{Condition}|\text{Test}^{+}) & = & \dfrac{P(\text{Condition})}{P(\sim\text{Condition})} \times \dfrac{P(\text{Test}^{+}|\text{Condition})}{P(\text{Test}^{+}|\sim\text{Condition})} \end{array}}$$

What's critical here is this term:

$$\frac{P(\text{Test}^{+}|\text{Condition})}{P(\text{Test}^{+}|\sim\text{Condition})}$$

That computes the *Bayes factor*, a measure of the strength of evidence for a particular diagnostic test. Note that when *P*(Condition) = *P*(~Condition) = .500, a Bayes factor of 3 represents a probability of .750 of believing that someone has a condition, and a Bayes factor of 10 represents a probability of .909 of believing that someone has a condition. Conversely, with a prevalence of .500, a Bayes factor of 0.333 represents a probability of .250 of believing that someone has a condition, and a Bayes factor of 0.100 represents a probability of .091 of believing that someone has a condition.

But the Bayes factor isn't just for interpreting diagnostic tests. In fact, the Bayes factor is the Bayes-is for an entire framework of statistical analysis.

Let's rewrite that term with some different symbols. Instead of a positive test result particularly, let's think broadly about obtaining some set of data. Rather than hypothesizing someone has a specific psychological condition, imagine we're testing for what we'd expect to find under an alternative hypothesis H_1. Thus, the absence of a condition corresponds to what's expected under a null hypothesis H_0. That gives:

$$\frac{P(\text{Data}|H_1)}{P(\text{Data}|H_0)}$$

If we follow this notation through, we can then rewrite the whole formula as:

$$\boxed{\begin{array}{lrcl} & P(H_1|\text{Data}) & = & \dfrac{O(\text{Data}|H_1)}{O(\text{Data}|H_1)+1}, \\ \textit{where} & O(\text{Data}|H_1) & = & \dfrac{P(H_1)}{P(H_0)} \times \dfrac{P(\text{Data}|H_1)}{P(\text{Data}|H_0)} \end{array}}$$

This...this would be a magical thing in statistical hypothesis testing. The big problem with frequentist *p* values is that they only tell us $P(\text{Data}|H_0)$, or the probability that we would get data as extreme or more extreme (in the case of two-tailed tests) given that the null hypothesis is true.

But that's the exact opposite conditional probability about which we really care in our analyses! We don't really care about the null hypothesis; we want information about the alternative hypothesis. And we don't want to know how likely it is to see data (e.g., test results) given certain hypotheses; we want to know how likely our hypotheses are given the data (e.g., test results) we collected. Alas, our typical hybrid Fisher-Neyman-Pearson null hypothesis significance tests don't give us much of what we want, so it's hard for us to get what we need. Ultimately, we really care about how likely our alternative hypothesis is given the data we observe...or $P(H_1|\text{Data})$.

And $P(H_1|\text{Data})$, dear friends, is *exactly* what Bayesian reasoning promises to give us.

Some simulation studies have shown that for Bayesian analyses analogous to *t* tests, a Bayes factor of 3 has the same evidential value as $p = .05$ for reasonably large samples, and that a Bayes factor of 10 has the same evidential value as $p = .01$ for the same such samples.

In fact, Bayesian methods can even give us evidence *for* the null hypothesis! For example, a Bayes factor of 1/3 gives somewhat weak evidence for the null compared to the particular alternative hypothesis being tested, and a Bayes factor of 1/10 gives reasonably strong evidence for the null compared to the alternative. Note that these are just the inverses of the Bayes factors *for* the alternative hypothesis.

So if Bayesian methods promise to give us exactly what we want, and we've got rules of thumb to interpret Bayes factors like we do *p* values to ease our transition, why don't we use them all the time, every time? Well, there are a couple of problems that crop up in Bayesville compared to the Land of Frequentism.

1) Computing the Bayes factor for most analyses is darn near impossible to do with strict formulas, unlike *p* values in frequentist analyses – and some people think they're not all that meaningful anyway.

In frequentist statistical analyses, we only have to compute $P(\text{Data}|H_0)$, which we can do by calculating the area under the curve of the probability distribution for whatever test statistic we're using (often using integral calculus). That is, we just need to get the *p* value for the scores as extreme or more extreme than the one we obtained with our data – under the assumption that there's no effect. Because this *p* value can be analyzed by plugging numbers into formulas that describe the distributions of our test statistics, we call this an *analytic* strategy.

Bayesian analyses require the computation of three additional probabilities, none of which are as easy to calculate from formulas as $P(\text{Data}|H_0)$. In particular, the intractability of calculating $P(\text{Data}|H_1)$ analytically is a major reason we engage in null hypothesis significance testing, not alternative or substantive hypothesis testing. Modern computational techniques can quickly create *numerical* approximations of what would be expected given certain assumptions, but they're not as analytically clean as the frequentist approach. Thus, the less-precisely defined math of these analyses turns off some practitioners.

Furthermore, some arch-Bayesians believe that Bayes factors are just a poor-man's link to null hypothesis significance testing and that reporting them instead of posterior probabilities is not valid Bayesian reasoning. In my opinion, that's a step too far. For me, Bayes factors represent a nice way to model the strength of evidence, allowing whatever priors exist to be informed by that evidence. Rather than assuming some kind of uniformly relevant prior (which is an oddly frequentist notion in my view), Bayes factors allow plugging in whatever priors need to be plugged in to understand a person's belief structure.

2) There's a lot of controversy about what the $P(H_0)$ and $P(H_1)$ priors should be – and their values strongly influence the conclusions these analyses support.

In the psychodiagnostic case, we have prevalence data to plug into $P(H_1)$, which makes $P(H_0)$ 1 – prevalence. But how should we approach assigning a prior probability to H_1 and H_0 in most analyses, especially in the absence of prior data?

Some argue for a uniform distribution to represent absolute ignorance, making all possible values equally probable. However, many others argue that this isn't a reasonable default prior: It shouldn't be equally as likely to expect a huge effect as a small one. For that reason, they recommend a Cauchy distribution, which represents the ratio of two normally distributed standardized variables. Even then, there's controversy about exactly what scaling factor should be used in the Cauchy distribution (1? $\sqrt{2}/2$? something else?), so this remains an area of active debate.

As noted above, the prior that's used in any Bayesian analysis strongly influences the results of that analysis. Bayesians view that as a feature, not a bug. However, I think there's a place for common ground. To the extent that prevalence rates for various conditions can inform the shapes of these priors, Bayesian and frequentist psychodiagnosticians can live together in harmony. Whether Bayesian and frequentist researchers would draw the same conclusion from the same data – well, that's a different story.