## (BINOMIAL) SIGNIFICANCE TESTING

To avoid the base-rate fallacy, we should consider the base-rate when estimating how likely our evidence makes a hypothesis. But unfortunately, *very often* we don't know the base-rate. Consider the following example:

Example: *Gloria Stewart has a reputation of being clairvoyant. We decide to test her by taking an ordinary deck of cards, drawing a card, and having her identify the suit of the card without her looking. Suppose we give her 10 chances to do this. We discover that she is right 6 times. (Note that clairvoyance is still possible even if she is not infallible. Clairvoyance may be like eyesight where it is fallible but still able to provide knowledge.) So does the evidence indicate that she has clairvoyant power? Or is she just lucky in her guesses?*

It is impossible to say what the prior probability of clairvoyance is—and without that, we cannot readily calculate how likely the clairvoyance-hypothesis is, given our evidence. But we can answer a related question, namely: How likely is *our evidence* if Stewart is randomly guessing? If 6 out of 10 successes ends up being extremely unlikely under the guessing-hypothesis, then it may be reasonable to reject that hypothesis.

Statisticians would call the *null hypothesis* "$H_0$" the hypothesis that Stewart is just guessing. It is the hypothesis that our experimental evidence has a "null explanation," no explanation other than random chance. The *alternative hypothesis* "$H_1$" would be the hypothesis that Stewart has some clairvoyant power. Again, our plan is not to ask directly how probable $H_1$ is. Rather, the plan is to ask: How likely is 6 successes out of 10 trials, on the assumption that $H_0$ is correct? The conditional probability of our evidence in this context is called the "*p*-value." The basic idea here is that if the *p*-value is sufficiently low, then assuming our testing process was legitimate, we should reject $H_0$.[1]

[1] In fact, our test procedure here is not legitimate—ten trials is not enough to ensure that we know Stewart's real success-rate in identifying cards. However, I'm going to mostly ignore this issue. If we were to use an appropriately large number of trials, the example would become too complex. (E.g., in lieu of using the binomial theorem below, we would need to use the Stirling approximation to the binomial).

IMPORTANT: If our *p*-value justifies rejecting $H_0$, this does NOT mean we should conclude that $H_1$ is true! There may be other explanations for Stewart's success besides $H_1$, e.g., perhaps she was somehow cheating the test. (Still, if our *p*-value justifies rejecting $H_0$, that would be informative of something.[2]) Conversely: If $H_0$ survives our test, that does NOT mean we should conclude that $H_0$ is true! It would mean only that *our* evidence did not disprove $H_0$. But there still may be *other* readily available evidence against $H_0$. (This is why you often say that you "failed to reject the null" in the context of an experiment, rather than that you "proved the null.")

In our testing of Stewart, there are several steps in understanding how to determine the *p*-value. For convenience, assume for now that Stewart was correct in the first six trials specifically and incorrect in the last four trials. The probability of this happening, assuming $H_0$, would be the probability of her guessing correctly on the first trial AND the probability of her guessing correctly on the second trial AND… AND the probability of her guessing incorrectly on the seventh trial AND … AND the probability of her guessing incorrectly on the tenth trial. Presumably, moreover, the trials are independent of each other. So, as per the mathematics of probability, we can use the [Simple AND Rule] to multiply all these probabilities together, in order to know what the chances are of her being correct on the first six and incorrect on the last four (assuming that she is just randomly guessing).

Since there are 4 suits in an ordinary deck, Stewart would have a 1/4 chance of guessing correctly per trial, and a 3/4 chance of guessing incorrectly. And thus, the relevant calculation looks like this:

$$(1/4)^6 \times (3/4)^4 = 1/4096 \times 81/256 = 81/1{,}048{,}576$$

So the probability of guessing the first six correctly, and of guessing the last four incorrectly, is 81 chances out of 1,048,576.

---

[2] In some significance tests, $H_1$ is identified simply with the denial of $H_0$. (Thus, instead of the hypothesis that Stewart "is clairvoyant," $H_1$ could instead be the hypothesis that Stewart is "not guessing") Moreover, if $H_1$ just is $\sim H_0$, then rejecting $H_0$ indeed means accepting $H_1$. But usually, significance tests are not formulated this way. And our example above is a case where rejecting $H_0$ is not the same as accepting $H_1$.

But at this point, we should recognize that we are not just interested in the chances of Stewart guessing *the first six* correctly. We are rather interested in the chances of Stewart guessing *any* of the six correctly, in any order. After all, any success-rate of 60% should affect how we regard the null hypothesis. So in more precise terms, we want to know what the probability is of 6 successes out of 10 in *some* order, under the null hypothesis.

Now it is crucial here that, no matter what the order of correct and incorrect guesses, the chances of getting any particular combination of 6-and-4 will actually be *the same as any other such combination* under $H_0$. Hence, since we already found what the chances are for one of the combinations, we just need to multiply that by the number of total combinations.

Let $\binom{10}{6}$ be the total number of combinations where Stewart gets exactly 6 right out of 10 trials. The general formula to determine the number of combinations is as follows (where *n* is the total number of trials and *k* is the number of successes):

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

So in our example, $\binom{10}{6} = \frac{10!}{6! \times 4!} = \frac{3{,}628{,}800}{720 \times 24} = 210$. Hence, there are 210 ways to get exactly 6 successes out of 10 trials. Thus, if we multiply 210 by our earlier number, 81/1,048,576, we get the probability of Stewart guessing 6 successes out of 10 trials in any order. That number is 17,010/1,048,576.

But here things get a bit tricky. We now know the chances that she gets *exactly* 6 successes in 10 trials. Yet there are other ways of getting 6 successes—namely, she could get 6 successes by getting 7 successes! And she could get 6 successes by getting 8, 9, or 10 successes. We have to count these possibilities as well, if we want to know the true chances of getting 6 successes under the null hypothesis.

Now the chances of just guessing the first 7 correctly, and the last 3 incorrectly, is:

$(1/4)^7 \times (3/4)^3 = 1/16{,}384 \times 27/64 = 27/1{,}048{,}576$

Moreover, there are $\binom{10}{7} = \frac{10!}{7! \times 3!} = \frac{3{,}628{,}800}{5040 \times 6} = 120$ combinations possible of guessing 7 correctly and 3 incorrectly. So we multiply 120 by 27/1,048,576 to get the probability of Stewart getting any combo of 7-right-3-wrong under the null hypothesis. That number is 3240/1,048,576.

We repeat the analogous procedure to get the chances of guessing any combo of 8-right-2-wrong, the chances of guessing any combo of 9-right-1-wrong, and the chances of guessing 10 right out of 10. In shortened form, our thinking here works out as follows:

$p(\text{Exactly8}/H_0) = \binom{10}{8} \times (1/4)^8 \times (3/4)^2 = 45 \times 1/65{,}536 \times 9/16 = 405/1{,}048{,}576$
$p(\text{Exactly9}/H_0) = \binom{10}{9} \times (1/4)^9 \times (3/4)^1 = 10 \times 1/262{,}144 \times 3/4 = 3/1{,}048{,}576$
$p(\text{Exactly10}/H_0) = \binom{10}{10} \times (1/4)^{10} \times (3/4)^0 = 1 \times 1/1{,}048{,}576 \times 1 = 1/1{,}048{,}576$

We then add together the probabilities of guessing any 6-and-4 combo, any 7-and-3 combo, any 8-and-2 combo, any 9-and-1, and the probability of guessing 10 out of 10. Summing those numbers then yields the probability of her getting *at least* 6 out of 10 successes under the null hypothesis. (This is a correct application of the [Simple OR Rule], given each combination mutually excludes all other combinations.) The summation is as follows:

17,010/1,048,576 + 3240/1,048,576 + 405/1,048,576 + 3/1,048,576 + 1/1,048,576
= **20,659/1,048,576**

Thus, there is just under a 2% chance of Stewart getting 6 successes in 10 trials, assuming $H_0$.

NOTE WELL: This does NOT mean that $H_0$ *itself* has about 2% chance of being true, given the evidence. It goes the other way around: There is around 2% chance of *getting this evidence*, under $H_0$. Even so, since that success-rate is quite unlikely under $H_0$, the usual view is that we would be reasonable in rejecting $H_0$ (assuming our testing procedure is otherwise legit). $H_0$ could still be true, but it would be true only if something really unlikely happened during our testing.

Our process can be summarized by the following formula:

**[Binomial Theorem] If our evidence E is *k* successes in *n* binomial trials, and if $p(S/H_0)$ is the probability of a successful trial under the null hypothesis, then:**

$$p(E/H_0) = \sum_{k}^{n} \binom{n}{k} p(S/H_0)^k \times (1 - p(S/H_0))^{n-k}$$

The formula instructs us to take the probability (assuming $H_0$) of *k* successes, and multiply it by the probability (given $H_0$) of $n - k$ failures. We then multiply *that* by the number of combinations where you get *k* successes in *n* trials. Call the result $r_k$. We then do this again for $k+1$ successes: We take the probability (assuming $H_0$) of $k+1$ successes, multiply by the probability (given $H_0$) of $n - k+1$ failures. And we then multiply *that* by the number of combinations where you get $k+1$ successes in *n* trials. Call the result $r_{k+1}$. Subsequently, we calculate $r_{k+2}$ for $k+2$ successes, and so on, up to result $r_n$ for *n* successes. After getting all those results, the formula tells us to sum together $r_k$, $r_{k+1}$, $r_{k+2}$, …, $r_n$. (That summing operation is what the Greek letter Σ indicates.) The resulting sum is the probability of getting *at least k* successes in *n* trials under the null hypothesis, i.e., it is $p(E/H_0)$, our *p*-value.

Notice that this thinking is apt only if we are dealing with *binomial* trials—that is, trials where the results are *binary*—either 'yes' or 'no' (or in our example, 'right' or 'wrong'). The *p*-value must be calculated in a different way when the results are not binary (e.g., with measurements along a continuum). But these other methods are more complicated and we'll skip them here.

Getting back to Stewart: We found that there is only about a 2% chance that she would get 6 out of 10 right by guessing. To repeat, that does NOT mean we should conclude that she is clairvoyant. What we can say that *if* she was just guessing during our trials, then her success-rate was really unlikely. And so, if our testing procedure was otherwise legitimate, that is a reason to conclude that she was not just guessing during the trials.

Again, none of this suggests that $H_0$ itself has about 2% chances of being true. Many statisticians (so-called "frequentists") would say it does *not even make sense* to say that $H_0$ has a probability

of being true. They would insist that either Stewart was randomly guessing during our trials or not (just like she is either pregnant or not). Other statisticians will disagree however. Regardless, everyone would agree that our evidence concerning Stewart does not show that *the null hypothesis* has about a 2% chance. Rather, what has approximately 2% chances is our *evidence* under the null hypothesis. That is what the *p*-value represents.

Even so, if the testing procedure is otherwise legitimate, a *p*-value at or below 5% is standardly regarded as enough to reject the null hypothesis. Such a *p*-value suggests that the evidence is "sufficiently incompatible" with $H_0$ to warrant rejecting $H_0$. But in some scientific contexts, a lower *p*-value is required, e.g., at or below 1%. So whether scientists would reject our null hypothesis about Stewart would depend partly on a *decision* concerning how much risk shall be tolerated.

That may seem odd. Whether our evidence "disproves" the null hypothesis will depend on someone's preferences about risk-tolerance. And there's a question about why 5% or 1% represents the appropriate level of tolerance. Granted, there seems to be good reason not to set it higher than 5%—if we did, our science would seem less trustworthy. But why not a require a *p*-value at or below .005% or even .0000001%? After all, if we want to speak of "scientific proof," a stricter criterion might be in order.

Regardless, a *p*-value of 5% is standardly regarded as the threshold of "statistical significance." And this is where "significance testing" gets its name—it is a test for whether the evidence under the null hypothesis falls below the *p*-value, the chosen level of statistical significance.

To summarize, then, we have surrendered the ambition of giving an exact assessment of the following abductive argument:

(P1) The probability of Stewart identifying a suit correctly is 1/4, under the null hypothesis.
(P2) Stewart identified a suit correctly 6 out of 10 times.
(C1) So, Stewart is clairvoyant.

We are rather giving a *deductive* argument where, instead of (C1), we use (P1) and (P2) to conclude (C2):

(C2) So, (P2) has about 2% chances of being true under the null hypothesis.

We are then adding that, according to the standards one typically sees in science—and assuming our testing procedure was legitimate—the argument for (C2) also justifies the abductive inference to (C3):

(C3) So, the null hypothesis is false: Stewart is not just randomly guessing.

But this in no way implies that (C1) is true.

**Quick Reference Sheet**

Key Definitions

1. The *absolute* or *prior* probability of **P** is $p(\mathbf{P})$
2. The *posterior* probability of **P** given **Q** is $p(\mathbf{P}/\mathbf{Q})$
3. **P** is *independent* of **Q** iff $p(\mathbf{P}) = p(\mathbf{P}/\mathbf{Q})$
4. **P** and **Q** are *mutually exclusive* iff (**P** & **Q**) cannot be true.
5. The *p-value* in a significance test is $p(\mathrm{E}/\mathrm{H}_0)$, where E is the evidence from the test and $\mathrm{H}_0$ is the null hypothesis. This is almost always NOT the same as $p(\mathrm{H}_0/\mathrm{E})$!

Axioms

1a. $0 \leq p(\mathbf{P}) \leq 1$
1b. $0 \leq p(\mathbf{P}/\mathbf{Q}) \leq 1$
2. If **P** cannot be false, then $p(\mathbf{P}) = 1$
3. [Simple OR Rule] If **P** and **Q** are mutually exclusive, then $p(\mathbf{P} \vee \mathbf{Q}) = p(\mathbf{P}) + p(\mathbf{Q})$

Basic Theorems

1. [NOT Rule] $p(\sim\mathbf{P}) = 1 - p(\mathbf{P})$
2. [Simple AND Rule] If **Q** is independent of **P**, then $p(\mathbf{P}\ \&\ \mathbf{Q}) = p(\mathbf{P}) \times p(\mathbf{Q})$
3. [AND Rule] $p(\mathbf{P}\ \&\ \mathbf{Q}) = p(\mathbf{P}) \times p(\mathbf{Q}/\mathbf{P})$
4. [OR Rule] $p(\mathbf{P} \vee \mathbf{Q}) = p(\mathbf{P}) + p(\mathbf{Q}) - p(\mathbf{P}\ \&\ \mathbf{Q})$

Theorems about Posterior Probabilities

5. [Bayes' Theorem] If $p(\mathbf{Q}) \neq 0$, then:

$$p(\mathbf{P}/\mathbf{Q}) = \frac{p(\mathbf{P}) \times p(\mathbf{Q}/\mathbf{P})}{p(\mathbf{Q})}$$

6. [Binomial Theorem] If our evidence E is $k$ successes in $n$ binomial trials, and if $p(\mathrm{S}/\mathrm{H}_0)$ is the probability of a successful trial assuming $\mathrm{H}_0$, then:

$$p(E/\mathrm{H}_0) = \sum_{k}^{n} \binom{n}{k} p(\mathrm{E}/\mathrm{H}_0)^k \times (1 - p(\mathrm{S}/\mathrm{H}_0))^{n-k}$$

…where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$