A comparative introduction to statistical inference and hypothesis testing

Kimball Martin*

October 14, 2016

These are some notes on a very simple comparative introduction to four basic approaches of statistical inference—Fisher, Neyman–Pearson, Fisher/Neyman–Pearson hybrid, and Bayes—from a course on Quantitative & Statistical Reasoning at OU in Fall 2016. In particular, I hope to give a rough understanding of the differences between the frequentist and Bayesian paradigms, though they are not entirely disjoint.

This is not intended as a practical introduction to how to numerically perform various standard tests. For instance, I won't explain z-tests, t-tests, F-tests, χ^2-tests, or one-sided tests versus two-sided tests. I will only work with small samples of discrete distributions for transparency of computation. My goal is to focus on what I consider more basic conceptual issues in understanding the statistical framework of hypothesis testing.

I assume a little familiarity of the notion of a random variable, conditional probabilities, and a working understanding of the binomial distribution $\mathcal{B}(n, p)$, but no deeper study of statistics or probability is required.

1 Two scenarios

Statistical inference is the use of statistical methods to draw conclusions about the world from quantitative data. We will consider two simple scenarios.

1.1 Establishing a scientific fact

Scenario A: We want "statistically prove" some hypothesis H_1.

Example A: Suppose there is a large study which indicates that a person with the disease Cooties will recover in 1 week with probability 0.3. Dr Octopus thinks covering your body in octopi gets rid of Cooties faster. He finds 10 unwilling Cooties patients to cover in octopi and 6 recover in 1 week. (You're not sure about what happened to the other 4, but you think he ate at least 1 of them.) Based on this, should we conclude the octopus treatment speeds up recovery?

There are different possibilities for the statement of H_1, which we will come back to, but for simplicity let's just take the following:

*Department of Mathematics, University of Oklahoma, Norman, OK 73019

$$H_1 : \text{ the 1-week recovery rate of octopi-treated patients is not 0.3.}$$

In this scenario, the idea is that to statistically prove something we should be skeptical of the claim, and so just for the purposes of the argument, we will assume the claim has no truth, which is called the **null hypothesis** H_0. That is, H_0 is simply the statement that H_1 is not true. After a statistical evaluation, we can either *reject* H_0—i.e., accept (a statistical proof of) H_1—or *fail to reject* H_0. The idea is that we should reject the null, i.e., accept a statistical proof of H_1, only if there is overwhelming evidence that the data was very unlikely under the assumption of H_0.

Note: this is the exact same idea as for a mathematical proof by contraction—to prove something we start off by assuming its opposite and derive an impossibility. Of course, with statistics, nothing is impossible (at least in a reasonable model), but might just be a very small probability event.

In our example above, the null is then

$$H_0 : \text{ the 1-week recovery rate of octopi-treated patients is 0.3.}$$

That is, H_0 is the statement that the octopus treatment has no effect, which is the natural skeptical take on Dr Octopus's claim. Put another way, we can think of the control group as the population from previous study where the 1-week recovery rate is 0.3, and the test group as the population that undergoes the treatment (Dr Octopus's patients). Then H_0 is the statement that the 1-week recovery rate is the same for the test group as for the control group.

Now you might have thought to take H_1 to be H_1': the 1-week recovery rate of octopi-treated patients is > 0.3, or H_1'': the 1-week recovery rate of octopi treated patients is 0.6. The problem with these is that their negations, the corresponding null statements H_0' and H_0'', do not put a specific probability distribution on the test group, which make it more difficult to statistically test. (Actually, no collection of data would ever reject H_0'', because H_0'' is too broad—it allows for recovery rates like 0.5999999931—so even a huge amount of data suggesting H_1 can't rule out very similar recovery rates.)

1.2 Comparing two possibilities

Scenario B: We believe that one of 2 alternative hypotheses H_1 and H_2 is true, but we don't know which, and we want to determine which is more likely.

Example B: Let's say you're invited as a contestant on the Monty Hall show. Recall, at the end of the show, there are 3 doors—behind 1 is a car and the other 2 are goats, and the goal is to choose the one with the car. Monty lets you pick one door initially and stand in front of it, then opens one of the other two doors and reveals a goat, and lets you switch if you want. You've seen the show twice, both times contestants stay, and one won and one lost. You've heard two competing arguments—one says your probability q of winning if you stay is $\frac{1}{2}$, and one says it's $\frac{1}{3}$—but try as you might you can't figure out which is right. This gives 2 hypotheses:

$$H_1 : q = \frac{1}{2}$$

$$H_2 : q = \frac{1}{3}$$

So we're friends and I help you prep the show, and we play the game 8 times with the contestant staying each time, and find that the staying strategy won 3 out of these 8 times. Together with the data of when you saw the show, we saw a contestant who stays won 4 out of 10 times. To what extent does this help us determine if H_1 or H_2 is correct?[1]

Mathematically, there are a couple of important differences between Scenario A and B.

- In Scenario B both hypothesis H_1 and H_2 are what are called **simple hypotheses**—they exactly specify probability distributions. In Scenario A only the null H_0 should be simple, and the alternative H_1 may be a **composite hypothesis** (not simple).
- Scenario B is about deciding between two (specific) options and one is essentially forced to make a conclusion in favor of one of them. For Scenario A, we either statistically prove something or fail to—we are not trying to prove H_0 *or* H_1 is true, either we establish a belief in H_1 or we simply don't find enough evidence to conclude H_1.

The first point says that for Scenario B we must be in a situation where we know or assume our population can behave in only one of two specific ways. The second point says that Scenario B is a decision problem where we have to decide between two things, but Scenario A is more exploratory—we are just trying to see if we can provide convincing evidence for some hypothesis.

1.3 An index to approaches

We mentioned 4 methods we will compare in a very simplified setting. Here we briefly outline the major differences, in the context of our two simple scenarios:

- *Fisher's approach* is a soft, subjective approach for Scenario A. It uses statistical measures, notably the *p-value*, as evidence to help the researcher make claims in light of the context. (Fisher tests the statistical significance of the data, but does not use solely this to admit a statistical proof of a hypothesis.)
- The *Neyman–Pearson approach* is a deterministic test for Scenario B. It requires the choice of two parameters α, a significance level, and β, the power of the test (for one of the hypotheses, say H_1). The parameters α and β measure desired probability cutoffs for false positives and false negatives, respectively.
- The *hybrid Fisher/Neyman–Pearson approach* is a deterministic test for Scenario A. This uses Fischer's p-value with Neyman and Pearson's significance level α. This seems to be both the most widely used and most problematic approach of the 4, mainly because it is blindly used by scientists and other who don't really understand statistics.

[1] Of course, in practice, even if you can't figure out which is right, you should always switch because that is optimal no matter whether H_1 or H_2 holds. However, if you prefer, you can imagine a variant where Monty Hall charges you money to switch so that it is not worthwhile to switch if H_1 holds, and your rational behavior really will change according whether one believes H_1 or H_2.

- There are *Bayesian approaches* for both Scenarios A and B, though the treatment of Scenario B is much more elementary. The Bayesian paradigm relies on assigning prior probabilities (levels of belief) to alternative hypotheses, and using data with Bayes' theorem to revise the probabilities of these hypotheses.

The non-Bayesian approaches above are often called *frequentist approaches*. The term frequentist refers to specifying probability distributions just based on frequency of certain data. For instance, in Example A we say the distribution for H_0 is that the probability of recovery in 1-week is 0.3. The underlying assumption is that this came from a study, say of 10,000 patients where exactly 30% recovered in 1-week. So we just counted the frequency of 1-week recoveries to define this distribution. The Bayesian philosophy for Scenario A is that you should treat the probability of 1-week recovery itself as a random variable, because it may not be exactly 0.3. (For Scenario B this is not necessary as we assume in advance that only one of 2 possible distributions can happen.)

One of the biggest debates about statistics is what approach to use when, and I hope this note will make clear at least some of the main issues involved as well as limitations you should be aware of when reading statistical studies. In particular, no one approach is superior in all situations. Except for mine. Mine is the best.

2 Fisher's approach

Fisher considered Scenario A, so let's work with Example A.

Fisher's **p-value** is the probability p that a result as extreme as what was found occurs, assuming H_0. Fisher says p measures the *significance of the data*. In general, there are different ways to specify what we mean by "as extreme as." In our example, the most natural and obvious one is: (assuming H_0) what is the probability that at least 6 out of 10 patients recovered in 1 week. Here we can use the binomial distribution to calculate

$$p = P(X \geq 6) = \sum_{k=6}^{10} \binom{10}{k} (.3)^k (.7)^{n-k} \approx 0.047,$$

where X is a random variable with binomial distribution $\mathcal{B}(10, .3)$. In other words, there was less than a 5% chance that data as extreme as what we found occurs if the treatment has no effect.

This was a *one-sided p-value*: we only tested for extremes on one side of the expected value $E[X|H_0] = 3$. We could also do a *two-sided p-value* and look at the probability that X differs from $E[X] = 3$ by at least 3, i.e., $P(|X - 3| \geq 3)$. In this case, the only adds $P(X = 0) \approx 0.0001$ to the one-sided p-value, so the difference between these methods is negligible in this case (though not in general).

Note we test for data as extreme as a given result rather than just looking at the probability of a given result because the probability of a given result may be small for any result. That is, it might be that all or most outcomes are low probability events, so it may not be insightful to look at just $P(X = 6)$. So a better measure of how significant an outcome is is to look at the probability of a range of outcomes similar to what was observed. You might think that $X = 7$ is similar to $X = 6$, but $X = 10$ is not—however we include all

extreme events as a conservative estimate, and since extreme events happen with very low probability, they may not effect the results very much. Now you might think that $X = 5$ is also similar to $X = 6$, but we have to make a cutoff somewhere, and $X \geq 6$ is the simplest one to make. We remark that there are alternative proposals to consider what outcomes should be considered as similar to the observed one for use in calculating a p-value.

Going back to our problem, we can say from the p-value that the data provides a fair amount of evidence for H_1. Now Fisher says we should use the p-value *in conjunction with other information* to decide whether to accept H_1 or not. In fact, this could be a successive process. Maybe with no other information, we can tell Dr Octopus: "okay, your method looks promising—maybe the octopi eat up the cooties—why don't you do try treating a few more patients, as long as you promise not to eat them." Then use the new data together with the old data to compute a more informed p-value, and be more confident about accepting or H_1 or not.

Key to this is that with more and more trials, if H_1 is true, the p-value will get closer and closer to 0, boosting our confidence in H_1. Conversely, if H_0 is true, the p-value will get closer to 1 with more data.

Actually, our above "conclusion", that it might be worthwhile for Dr Octopus to do more work, we did use a little bit of qualitative information: there is a possible explanation that the treatment is effective because octopi may physically affect cooties. If the treatment were something like "speak only in pig latin," and it had the same results on 10 patients, we would just chalk it up to chance as there is no plausible reason that the treatment could be effective. We'll mention some other things one might consider later when discussing the hybrid approach.

Pros:

- Particularly suitable for combining quantitative and qualitative information.
- Researcher can factor in possible biases/flaws in experiment and prior beliefs.
- Forces you to take other factors into account and think about what is important, and thus have some understanding both of statistics and of your study.

Cons:

- Is subjective and not entirely methodical.
- Requires you to take other factors into account and think about what is important, and thus have some understanding both of statistics and of your study.

3 Neyman–Pearson approach

Different sources say different things about what Neyman and Pearson actually proposed, and I'm not sure all of them are right. I'm no historian, or even statistician, but I've tried to only attribute to them what I get from Neyman and Pearson's paper *On the problem of the most efficient test of statistical hypotheses* (Phil. Trans. R. Soc. Lond. A 1933 231

289-337) applied to our Scenario B, though I will give more modern terminology as well. (That paper is concerned with much greater generality than our Example B, which is a trivial case in their setup.)

First of all, the Neyman–Pearson philosophy (for Scenario B) is that we are not trying to establish the truth (or even necessarily belief) in H_1 or H_2, rather we are simply trying to provide a decision rule for how to behave—should we behave as if H_1 holds, or as if H_2 holds? Their approach is such that over the long term of making many such decisions one's behavior follows the truth as much as possible (on average).

Perhaps the most obvious thing to do is compare how likely our data is given H_1 (switching doesn't matter) or given H_2 (we should switch), which are called **likelihood estimates**. In Example B, let X be a random variable representing the number of times a staying strategy wins in 10 trials. Recall we observed an instance of $X = 4$. We compute

$$P(X = 4|H_1) = \binom{10}{4} (\frac{1}{2})^4 (\frac{1}{2})^6 \approx .205$$

$$P(X = 4|H_2) = \binom{10}{4} (\frac{1}{3})^4 (\frac{2}{3})^6 \approx .228$$

The Neyman–Pearson philosophy says, in its simplest form, since the observed data is more likely under H_2, we should behave as if H_2 is true, i.e., switch.

However, the observed data is only slightly more likely under H_2 than H_1. In our Example B, if there is no cost of switching, the relative likelihoods of H_1 and H_2 don't matter, and we should always switch just in case H_2 holds. But if there is a cost for switching, then we might want strong evidence in favor of H_2 before we decide to switch, and the strength of the evidence we seek should depend upon the cost of switching.

In the general Neyman–Pearson framework, we are allowed to do one of 3 things: behave as if H_1 is true, behave as if H_2 is true, or remain on the fence. In Example B (with a cost for switching), behaving as if H_1 is true means you definitely stay, behaving as if H_2 is true means you definite switch, and remaining on the fence means you have no strong opinion and may just stay or switch according to a spur-of-the-moment feeling or audience input.

There are two possible ways in which this approach forces a wrong decision: we might treat H_1 as false when it is true, or we might treat H_1 as true when it is not. (We don't count staying on the fence as a wrong decision.) An important point is that we can estimate how often our decision procedure will result in these errors. If we think of H_1 as a null hypothesis, and we are testing for the effectiveness of switching (H_2), the first type of error (switching when we shouldn't) is called a **Type I error** or **false positive**, and the second is a **Type II error** (not switching when we should) or **false negative**. (Neyman and Pearson did not use this terminology, at least in this paper.)

Let's say our observed value for X is k (so $k = 4$ in the original statement of Example B). Neyman and Pearson introduce the likelihood statistic for our observation k

$$\lambda = \frac{P(X = k|H_1)}{\max\{P(X = k|H_1),\, P(X = k|H_2)\}},$$

which measures how likely H_1 is compared to H_2 with observation k. Note H_1 suggests k should be larger, and $\lambda = 1$ if $P(X = k|H_1) \geq P(X = k|H_2)$, which happens when $k \geq 5$. The complete values for λ in this example are given in Table 1.

Table 1: Likelihood statistic (rounded) for H_1

k	0	1	2	3	4	5	6	7	8	9	10
λ	.06	.11	.23	.45	.9	1	1	1	1	1	1

Table 2: Type I error estimates (rounded) for incorrectly treating H_2 as true

C_1	0	.06	.11	.23	.45	.9	1
k range	—	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 10
Type I error	0	.00	.01	.05	.17	.38	.62

They propose a decision rule of the form: behave according to H_2 (switch) if $\lambda \leq C_1$ and behave according to H_1 (stay) if $\lambda \geq C_2$, and stay in doubt if $C_1 < \lambda < C_2$, for some numbers C_1, C_2. The choice of C_1 and C_2 should be made based on the situation and allow us to control the chance of Type I and Type II errors. (In our example, our choices should depend upon the cost of switching and how much we want the car.) In our simple situation, the choice of C_1 can be translated as treat H_2 as true whenever $k \leq K_1$, for some appropriate K_1 (see Table 2). Then the probability of a Type I error, that we "accept H_2" assuming H_1 is true, is simply $P(X \leq K_1|H_1)$. Similarly, rule to stay if $\lambda \geq C_2$ is translated to stay if $k \geq K_2$ for some K_2, and the probability of a Type II error—that is we "accept H_1" assuming H_2 is true—is just $P(X \geq K_2|H_2)$. We tabulated the values for Type I and Type II errors for "critical values" of C_1 and C_2 in Table 2 and Table 3.

In particular, if we just use the naive decision rule to treat H_2 as true (switch) if $X \leq 4$ and treat H_1 as true (stay) if $X \geq 5$, then the chance of a Type I error (wrongly switching) is about 38% and the chance of a Type II error (wrongly staying) is about 44%). Of course, if it turns out our observation is extreme rather than being a borderline case for the decision rule, we can be more confident of not making an error. (However, one should choose the decision rule before considering the data.)

In the modern terminology, we use the following notation:

— the **significance level** α of our test is the probability of a Type I error

— the **power** $1 - \beta$ of our test is the probability of a Type II error

Usually the test is presented with just a single cutoff C (or K in our example), where you make one decision with a statistic on one side of C and the other on the other, so you never have the remain in doubt option, though Neyman–Pearson explicitly give these 3 options in their paper. Then the recommended way to use the Neyman–Pearson test is to specify α and β in advance and then calculate how large of a sample size n you need to get a test with significance level α and power $1 - \beta$ before doing your experiment. Using larger samples will allow you to lower α (good) and increase β (good). For instance, in our naive decision rule above we get $\alpha \approx 0.38$ and $\beta \approx .79$, but if we used a sample size of $n = 20$,

Table 3: Type II error estimates (rounded) for incorrectly treating H_1 as true

C_2	0	.06	.11	.23	.45	.9	1
k range	≥ 0	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	—
Type II error	1	.98	.90	.70	.44	.21	0

then the analogous decision rule is to switch if $X \leq 8$ and stay if $X \geq 9$. In this case we get, $\alpha \approx 0.25$ and $\beta \approx 0.81$, i.e., there's about a 25% chance of a Type I error and a 19% chance of a Type II error.

Pros:

- Provides systematic and practical way of making decisions.
- Can control likelihood of errors.

Cons:

- Requires specific (simple) alternative hypotheses to choose from (though more than 2 is okay).
- Doesn't take into account other evidence you may have for H_1 over H_2 or vice versa, which may be very helpful when you only have small or potentially biased samples.
- Typically there is no "optimal choice" of parameters α and β, so their choice is often arbitrary, which can make a big difference in the targeted sample size n.

4 Fisher/Neyman–Pearson hybrid

The Fisher/Neyman–Pearson hybrid is a way to apply the Neyman–Pearson approach to Scenario A. Recall Scenario A means we are trying to establish some hypothesis H_1, and we begin by assuming its opposite, the null hypothesis H_0.

The first step in this hybrid method is to choose a significance level α for which we will reject H_0. Common values are $\alpha = 0.1$, 0.05 and 0.01. The choice of α should depend on the situation, and smaller values of α correspond to setting higher standards for believing H_1. For instance, if we are doing a DNA test for a murder trial (H_0: not guilty; H_1: guilty) we want a very small α to be very sure that DNA evidence is incriminating before claiming someone is guilty. Since the choice of α is very important, people don't usually think about it and just choose $\alpha = 0.05$ because everyone else does.

Now we do our experiment and collect our data, and the p-value p as in Fisher's approach, and follow the simple rule

$$\begin{cases} p < \alpha & \implies \text{reject } H_0 \\ p > \alpha & \implies \text{fail to reject } H_0 \end{cases}$$

For instance, in Example A the p-value was .047 so if $\alpha = 0.05$ we would reject H_0, i.e., believe (or at least choose to act) as if Dr Octopus' treatment is effective. But if we chosen $\alpha = 0.01$, or even $\alpha = 0.046$, we would have failed to reject H_0, i.e., not be convinced that the treatment is effective. This should make it clear that the choice of α plays a crucial role and is one of the major concerns with this method—with the same data, choosing α slightly differently gives us different results.

On the other hand, let's consider what the possible (1-sided) p-values are if Dr Octopus' results had turned out a bit different. If k out of 10 treatments were effective ($k \geq 4$),

Table 4: approximate 1-sided p-values for Example A

k	4	5	6	7	8	9	10
p	.35	.15	.047	.011	.0016	.00014	.0000059

then we get the p-values listed in Table 4. Note that if we had chosen $\alpha = 0.01$, the only results that would "test positive" (H_1 seems effective) are $k = 8, 9, 10$, which have p-values at most 0.0016 whereas the $k = 7$ case which has the closest p-value $\alpha = 0.011$ would not test positive. Put another way, $\alpha = 0.01$ isn't any weaker than the very strong $\alpha = 0.0017$. Thus, in some sense, the $\alpha = 0.01$ choice doesn't really give us the level of significance that we want. This is mainly an issue because we are working with a discrete random variable with a small sample size, but the earlier issue of our test results being very sensitive to the choice of α is a concern even for large samples.

Now let's think about what this test tells us and point out some common misconceptions.

As in the Neyman–Pearson approach, a **Type I error** is a false positive, i.e., rejecting H_0 when we shouldn't (thinking the treatment may be effective when it's not). Similarly, a **Type II error** is a false negative, i.e., failing to reject H_0 when it's not true (e.g., not realizing that a treatment works when it actually does).

The significance level α is the probability of a Type I error on a single run of this test (when H_0 holds). **WARNING:** a positive test result, say with $\alpha = 0.05$, does not mean we are 95% certain this drug is effective. A positive test result only means the observations are relatively unlikely (past some threshold), but certainly not impossible, if there is no effect. This is the *only thing* that is meant by the term **statistical significance**. I think this point is *the* most misunderstood point about hypothesis tests like this.

For instance, let's say we get a fingerprint sample at a crime scene and test it against a fingerprint database of 10,000 people. If we choose, $\alpha = 0.001$, we expect (on average) to get 10 positive matches.[2] Say we get exactly 10 matches. Clearly we are not 99.9% certain that each of these 10 people were at the crime scene. There's not even a good reason (a priori) to believe even one of them was at the crime scene. And unless we know the probability of Type II errors is very very small, we can't even rule out the other 9,990 people in the database (let alone all the people not in the database).

If one just reports positive results, without giving p-values or saying how many things were tested, knowing that something got a positive result is not so informative. If one is trying to find some result and publish a paper, trying only 1 test with $\alpha = 0.05$ and getting a positive result is a lot more convincing than trying 12 tests and getting 1 positive result. Maybe this is Dr Octopus' 9th go. Or maybe other scientists have tried the octopus treatment and didn't get statistically significant results so they didn't publish, but the one time it does get a positive test result, it does get published. (Because the other studies failed, other people don't know about this approach and may try it again and get lucky.)

This issue is called **publication bias**. It means there are a lot of studies with statistically significant results out there, but we shouldn't expect most of such one-off studies to be true, statistically speaking. Rather, we should—as Fisher said—look at other factors:

[2] I know nothing about the science of fingerprint testing. For the purposes of this argument, let's say we have a computer program that can ID fingerprints to an accuracy level where 1/1000 different people will match a given fingerprint. This corresponds to the mathematical choice of $\alpha = 0.001$.

bigger sample sizes and smaller p-values are generally better, and a convincing physical explanation would be great. Moreover, the hypotheses, significance level and experimental design (including sample size) should all be determined in advance of the study so they are not biased by the data to give a positive result. (Or worse, one does not just vary these things until they give the desired result—some people may confuse getting $p < \alpha$ as the goal, rather than finding out the truth.) Remember, if there is an actual effect to the treatment, we should see this by p-values getting smaller using data from sufficiently many independent trials.

My general suggestion for how to think about a positive test result is that it means there is some (preliminary, if it's a first test) evidence that, say, a treatment may be effective. And now we should evaluate it in context, and see if it is worth doing further experiments to provide more evidence for something. If repeated, independent studies are getting similar results, then we can be fairly confident that there's something "real" behind the studies, rather than chance.

Remarks on perceptions of science: There seem to be a number of people who think "most science is wrong." I believe this is in large part due the way science is portrayed in the media. The media might report "Scientist prove green jelly beans cause acne!" (cf. xkcd comic on course page) and convey this as a "scientific fact" when really there was just one experiment that came up positive, perhaps by chance. Then this may get "disproved" in another study, and these sorts of things happen all the time, so you don't know when you should trust any reports. The media also typically simplifies the conclusions of a scientific paper which exacerbates this (out of honest misunderstanding, or a desire to be sensational, or quite possibly both)—e.g., maybe you get "Red wine leads to longer, healthier lives!" for some study indicates red wine is correlated with lower incidence of heart attacks for people with certain predispositions, but then there is another study that says it increases the risks of something else and you get the headline "Red wine may kill you!"

In fact (and to repeat a bit, but this is important) whether or not there is any effect of a treatment, statistically speaking if you do enough tests you will get some positive results and some negative results, so you need to look at a consensus view (in addition to qualitative information) when multiple studies have been performed. Citing a single study that supports your point of view is not necessarily very meaningful when there sufficiently many studies to suggest all possible points of view.

I worry that many people think most scientists aren't doing research right or because (i) later refutations of all the (statistically unavoidable) false positives which are made to sound more shocking than they are, and (ii) occasional scandals of some scientists fabricating data/results. Some people might claim Ionniadis' paper[3] as proof that scientists either don't know what they're doing or are dishonest most of the time. Sure, scandals do happen, and some scientists don't follow the scientific method correctly (consciously or not) and may engage in "p-hacking" (manipulating data/tests to get a desired p-value), but because of the large number of false positives that are inevitable together with publication bias, one shouldn't conclude we're going about science in a bad way. (And scientists are working on addressing these other issues too.) If you wanted to really conclude we're doing science badly, you would need to do a statistical analysis of studies which were later refuted,

[3]Why most published research findings are false, *PLOS*, 2005.

taking into account the above factors as well as other important ones I'm not discussing. Also, it shouldn't even be that surprising that a lot of current science is coming to wrong conclusions—historically, science has always been a process of experimenting and theorizing and revising to find the truth, slowly getting more and more accurate.

Another caution: if the statistical test fails, that does not necessarily mean we should believe H_0. This process is designed with the idea that to statistically "prove" a statement (H_1), you should show overwhelming evidence that its opposite (H_0) is not correct. It may be that, without any a priori assumptions, the data favors H_1 over H_0, but if we don't meet the desired significance level, we will be cautious and not assert H_1 without more information. The choice of H_0 and H_1 at the start is important here, as alternate choices of H_0 and H_1 will often give different results.

Some people argue against this hybrid approach, and here is my perspective. The central problem with the hybrid approach is people not understanding Fisher's perspective or Neyman–Pearson's. Fisher's approach was to take things in context and try to see if something is true, which may be a gradual process by gathering more and more evidence. The Neyman–Pearson philosophy is that: okay, we're going to get a lot of decisions wrong, but we want a systematic rule to make decisions to get as many right as possible. The trouble comes when people want a systematic approach like Neyman–Pearson but interpret a statistically significant result as "proof" of an effect, and don't think about contextual evidence as Fisher wanted. (Of course, the same problem exists using the Neyman–Pearson approach for Scenario B, but this hybrid test seems to be more common so the issue arises more for the hybrid.) And the usefulness of preliminary results are even worse when people don't follow the systematic approach correctly.

Pros:

- Systematic and simple-to-use criterion.
- Can control Type I errors (and Type II with appropriate design).

Cons:

- Choice of α is rather arbitrary.
- Meaning of a positive test is widely misinterpreted, and the simplicity of the test makes people think they know what they're doing when they don't (the Dunning–Kruger effect). Teaching this test without solid theory is like giving a loaded gun to a toddler.
- Knowing the value of a single positive test result, without the p-value and the total number of tests tried (not even just the ones by the same researcher), provides little information on how "surprising" the results are.
- If H_0 is slightly incorrect to start with or there is a treatment effect but it is very minor (e.g., the probability of 1-week recovery is actually .3000017), then the probability of a positive test result is very high with large samples (simply because H_0 is technically not true, even though the efficacy of the treatment may be miniscule). (Someone who knows what they're doing can correct for this.)

5 Bayesian approach

Just like the original Neyman–Pearson approach doesn't use a null hypothesis, neither does the Bayesian approach. Here one considers a set of possible underlying distributions for the sample space, and assigns probabilities to them and updates these probabilities when new data is available. This will be easiest to illustrate in the case of just two alternate (simple) hypotheses H_1 and H_2, i.e., Scenario B.

Let's return to Example B. The possible distributions for how are data should behave are specified by the hypotheses $H_1 : q = \frac{1}{2}$ and $H_2 : q = \frac{1}{3}$. The idea with the Bayesian approach is that we want to *assign probabilities to the hypotheses* and they use Bayes' theorem to update these probabilities in light of the data.

To do this, we need to assign **prior probabilities:** to our hypotheses. If we're completely on the fence, then we can reflect this by saying H_1 and H_2 seem equally likely to start, i.e., we set our priors as:

$$P(H_1) = 0.5, \quad P(H_2) = 0.5.$$

On the other hand, if we were slightly more inclined to H_1 say (maybe because it was the first argument we heard/though of, or the people making that argument seemed a little more trustworthy), we might take instead take something like $P(H_1) = 0.6$ and $P(H_2) = 0.4$. But let's work with the notion that H_1 and H_2 are equally likely a priori.

As in the Neyman–Pearson section, let X be a random variable representing the number of types a staying strategy wins in 10 trials, and say the outcome of our experiment is k out 10 wins. The idea is once we've observed $X = k$, this affects the probabilities of H_1 versus H_2. That is, our new belief for how likely H_1 is should be the conditional probability $P(H_1|X = k)$, the probability of H_1 given that $X = k$, and similarly for H_2.

Recall Bayes' theorem says $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, where $P(A|B) = \frac{P(A\cap B)}{P(B)}$. In our case, this means

$$P(H_1|X = k) = \frac{P(X = k|H_1)P(H_1)}{P(X = k)}.$$

Let's work this out when $k = 4$. Recall we computed in the Neyman–Pearson section that

$$P(X = 4|H_1) \approx .117, \quad P(X = 4|H_2) \approx .130.$$

We now know every term on the right except the "absolute probability" $P(X = k)$. We don't know it directly, but we can compute it in terms of conditional probabilities:

$$P(X = k) = P(X = k|H_1)P(H_1) + P(X = k|H_2)P(H_2),$$

which when $k = 4$ gives

$$P(X = 4) \approx .117 \cdot .5 + .130 \cdot .5 \approx .124.$$

Thus we can revise our probabilities in light of $X = 4$ to get

$$P(H_1|X = 4) \approx \frac{.117 \cdot .5}{.124} \approx .47$$

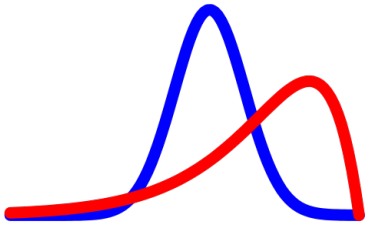
and we can similarly compute

$$P(H_2|X = 4) \approx .53.$$

Since these probabilities must sum to 1, we just computed the latter probability as 1 minus the first. In other words, they say the data means H_2 is ever so slightly likely than H_1, which corresponds with the observation that $\frac{4}{10}$ is slightly closer to $\frac{1}{3}$ than $\frac{1}{2}$. Now you can forget about the experiment and update your beliefs as $P(H_1) = .47$, $P(H_2) = .53$, and if you do another experiment, you can use Bayes approach again to update your beliefs with these new probabilities as your priors. While it is true that at each step your beliefs depend on your choice of priors, it is a theorem that with more and more experiments this process will approach the limiting case $P(H_1) = 0$, $P(H_2) = 1$ since H_2 is true. This is the Bayesian approach in the simplest setting.

Note one key difference between the Bayesian approach and the Neyman–Pearson one, is that the Bayesian approach quantifies how much more you believe H_2 than H_1 (or vice versa), whereas Neyman–Pearson simply says you choose one or the other with controlled errors. (Stricter controls on errors in Neyman–Pearson, or the hybrid, is similar to saying how much more you believe one hypothesis over the other, but you may need to have a really large sample size to get the desired control over errors.)

Finally, just roughly indicate the idea of the Bayesian approach for Scenario A. Recall Example A. We had the initial hypothesis H_0: the chance of 1-week recovery is $q = .3$. However, this was based on a previous study which also has some uncertainty. To build in this uncertainty, we can think of q as a random variable and assign it a probability distribution, represented by the blue line in the graph below. Then if there is a lot of evidence for a positive effect in the treatment, one can a version of Bayes' theorem for continuous random variables to update the distribution for q which may look like the red line in the graph below.



Pros:

- Allows for uncertainty about the null hypothesis (particularly for Scenario A).
- Good for situations where you have well-informed prior beliefs about what is true.
- Quantifies beliefs in various hypotheses.
- Provides natural framework for gradually refining beliefs based on new evidence.

Cons:

- Requires a choice of prior distributions, which potentially are quite arbitrary (particularly for Scenario A).
- If the amount of sample data is small, the results of this process are highly dependent on the choice of priors.
- For Scenario A, the approach becomes considerably more complicated than the hybrid test.

6 Exercises

For some of these exercises, you may want to use a mathematical software system. One possibility is to use Sage (which uses the programming language Python, if you're familiar with that). You can use this online at:

https://sagecell.sagemath.org/

Here is how to use it to do some probability calculations. The following code

```
def p(n,k,q):
    return binomial(n,k)*q^k*(1-q)^(n-k)
sum([p(10,k,.3) for k in range(6,11)])
```

computes $\sum_{k=6}^{10} \binom{n}{k} (.3)^k (.7)^{n-k}$, which was the p-value for Example A. The first two lines of code defines a function $p(n, k, q)$ which returns $P(X = k)$ if X is a random variable with binomial distribution $\mathcal{B}(n, q)$, i.e., the probability of k success in n independent trials if each trial has success probability q. You should have no need to change the first 2 lines of code for these exercises. The last line of code says to sum up $p(10, k, .3)$ for $6 \leq k < 11$.

1. Let's modify Example A so there are n trials. What's the smallest value of n for which it is possible to get a p-value less than 0.05?
2. Make a table of p-values analogous to Table 4 for Example A but with a sample size of $n = 6$ instead of $n = 10$.
3. Let's suppose you do an experiment to test a hypothesis H_1 with a sample size of 10, and you get a p-value of 0.047. Then you do a follow up experiment with sample size of 10 and you again get a p-value of 0.047. Does the second experiment convince you of H_1 more than the first?
4. Let's suppose you do a hybrid test with $\alpha = 0.05$. Suppose there are two experiments (not necessarily about the same thing) which have sample sizes 10 and 1000, and they both test positive. Is one more convincing than the other because of the difference in sample sizes?
5. Say you have 2 independent hybrid tests, both with $\alpha = 0.05$. Can you determine the probability that both of them give false positive results?

6. Make a table analogous to Table 3 of Type II errors for possible Neyman–Pearson decision rules for Example B but with a sample size of $n = 8$ instead of $n = 10$.

7. Let's modify Example B so there are n trials and use the following naive Neyman–Pearson test. Think of H_1 as a null. If we observe k successes, then we pretend H_2 is true if $\frac{k}{n}$ is closer to $\frac{1}{3}$ than $\frac{1}{2}$, and pretend H_1 is true otherwise.

   (a) What is the smallest value of n for which this rule has significance level $\alpha < 0.2$?

   (b) What is the smallest value of n for which this rule has power $1 - \beta < 0.2$?

8. Going through Example B with Bayes approach, what are your updated probabilities if you start with priors $P(H_1) = 0.6, P(H_2) = 0.4$? What about with priors $P(H_1) = 0.4, P(H_2) = 0.6$?

9. Let's modify Example B, and say we win k times out of n trials with the staying strategy. Assuming H_2 is true (which it is), then the observation being within one standard deviation of the mean is saying that $n - 2\sqrt{n} \leq 3k \leq n + 2\sqrt{n}$. If we use Bayes approach with priors $P(H_1) = P(H_2) = 0.5$, what is the smallest value of n which gives an updated probability $P(H_2) > 0.9$ for any observation k with $n - 2\sqrt{n} \leq 3k \leq n + 2\sqrt{n}$.