Chapter 1. An introduction to statistical significance 1단원. 통계적 유의성 서설

Much of experimental science comes down to measuring differences. Does one medicine work better than another? Do cells with one version of a gene synthesize more of an enzyme than cells with another version? Does one kind of signal processing algorithm detect pulsars better than another? Is one catalyst more effective at speeding a chemical reaction than another? We use statistics to make judgments about these kinds of differences. We will always observe some difference due to luck and random variation, so statisticians talk about statistically significant differences when the difference is larger than could easily be produced by luck. So first we must learn how to make that decision.

실험과학이 하는 대부분의 일은 차이를 측정하는 것으로 귀결된다. 하나의 약제가 다른 것보다 좋은가? 특정 버전의 유전자를 가진 세포가 다른 버전의 유전자를 가진 세포 보다 어떤 효소를 잘 합성하는가? 어떤 신호처리알고리즘이 다른 알고리즘보다 펄사를 더 잘 감지할 수 있는가? 어떤 촉매가 다른 촉매보다 어떤 화학반응을 가속하는 효과가 좋은가? 우리는 통계학을 써서 이와 같은 차이들에 관한 판정을 내린다. 우리 눈에 보이는 얼마간의 차이는 우연이나 임의변동에 의한 것일 수 있다. 그래서, 우연에 의해서 쉅사리 만들어지는 차이보다 훨씬 큰 차이일 때 비로소 통계학자들은 “통계학적으로 유의한 차이”라고 칭한다.

Suppose you’re testing cold medicines. Your new medicine promises to cut the duration of cold symptoms by a day. To prove this, you find 20 patients with colds, give half of them your new medicine, and give the other half a placebo. Then you track the length of their colds and find out what the average cold length was with and without the medicine. But not all colds are identical. Maybe the average cold lasts a week, but some last only a few days. Others might drag on for two weeks or more. It’s possible that the group of 10 patients who got the genuine medicine in your study all came down with really short colds. How can you prove that your medicine works, rather than just proving that some patients got lucky?

감기약을 시험한다고 치자. 주어진 새 감기약은 감기 증상의 지속을 만 하루만큼 줄인다고 전망한다. 이를 증명하고자 스무 명의 감기 환자를 동원하여 절반에게 신약을 처방하고 나머지 절만에게는 가짜약을 준다. 그리고 피험자들의 감기 증상의 길이를 추적하여 신약 처방 환자들의 평균 감기 유병기간과 가짜약 처방 환자들의 평균 감기 유병기간을 알아낸다. 그런데, 모든 감기가 똑같지는 않다는 게 문제다. 평균적으로는 일주일 지속할 것이나, 몇 명은 겨우 며칠 만에 낫고, 다른 몇 명은 두 주일 이상을 끌 수도 있다. 실험에서 진짜 약을 받은 열 명의 환자 모두 감기를 짧게 앓고 마친 경우를 상상할 수 있다. 몇 명이 그저 운이 좋아서였던 게 아니라 정말로 신약의 약효 덕분에 이렇게 된 것이라고 증명하려면 어떻게 해야 하나?

Statistical hypothesis testing provides the answer. If you know the distribution of typical cold cases—roughly how many patients get short colds, long colds, and average-length colds—you can tell how likely it is that a random sample of patients will all have longer or shorter colds than average. By performing a hypothesis test (also known as a significance test), you can answer this question: “Even if my medication were completely ineffective, what are the chances my experiment would have produced the observed outcome?”

통계학적 가설검정을 통해 답을 찾아낸다. 만약 전형적인 감기의, 경과가 어떻게 분포하는지 알고 있다면, (즉, 대략 얼마나 많은 환자가 잠깐 앓고, 오래 앓고, 얼마나 많은 환자가 평균적으로 앓는지의 분포를 안다는 뜻인데) 감기 환자들로부터 뽑은 임의의 표본이 평균보다 오래 감기를 앓을지, 짧게 감기를 앓을지에 대해서 알 수 있을 것이다. 가설검정을 행함으로써(이걸 유의성 검정이라고도 한다) 이 질문에 답할 수 있다. “만약 내가 가진 약이 아예 효과가 없더라도 실험을 통해 이와 같은 결과가 관찰될 가능성은 얼마나 되는가?”

If you test your medication on only one person, it’s not too surprising if her cold ends up being a little shorter than usual. Most colds aren’t perfectly average. But if you test the medication on 10 million patients, it’s pretty unlikely that all those patients will just happen to get shorter colds. More likely, your medication actually works.

약을 딱 한 명에게만 투여하는 시험이라면 그 환자의 감기가 보통보다 조금 짧게 나왔다 하더라도 그닥 놀랄 일이 아니다. 모든 감기가 완벽하게 평균에 놓인 것이 아니기 때문이다. 반면에 천만 명의 환자에 대해서 약효를 시험했을 때 모든 환자들의 감기가 짧아졌다면 이것조차 우연에 의한다고 볼 수는 없다. 그보다는 그 약이 실제로 약효를 가졌을 것이다.

Scientists quantify this intuition with a concept called the p value. The p value is the probability, under the assumption that there is no true effect or no true difference, of collecting data that shows a difference equal to or more extreme than what you actually observed.

바로 이 직관을 정량화할 때 P값이라는 개념을 동원한다. 실제로는 효과가 없다고 전제하였을 때, 표본이 드러나 보이는 차이만큼, 또는 그 이상의 차이를 낼 수 있는 확률이 P값이다.

So if you give your medication to 100 patients and find that their colds were a day shorter on average, then the p value of this result is the chance that if your medication didn’t actually do anything, their average cold would be a day shorter than the control group’s by luck alone. As you might guess, the p value depends on the size of the effect—colds that are shorter by four days are less common than colds that are shorter by just one day—as well as on the number of patients you test the medication on.

자 그러면, 100명의 환자에게 약을 주고 환자들의 감기가 평균보다 하루 짧게 나왔다고 쳤을 때, 그 약이 실제로는 전혀 작용하지 않았어도 그 약을 먹은 환자들의 평균 감기 지속이 약을 먹지 않은 환자들의 평균보다 하루 짧을 확률 즉 순전히 운이 좋아서 그랬을 확률이 이 경우의 P값이 된다. 짐작하였듯이, P값은 일차적으로 효과의 크기에(효과의 크기는 여기서는 감기의 유병기간이다) 좌우되며 - 보자, 감기를 나흘 단축시키는 것은 단지 하루를 단축시키는 것보다 큰 효과이다. 당연히 약효를 시험하고자 동원한 환자의 수에도 좌우된다.

Remember, a p value is not a measure of how right you are or how important a difference is. Instead, think of it as a measure of surprise. If you assume your medication is ineffective and there is no reason other than luck for the two groups to differ, then the smaller the p value, the more surprising and lucky your results are—or your assumption is wrong, and the medication truly works.

꼭 기억하라. P값은 이런 것들이 아니다: 네가 얼마나 옳은지에 대한 척도가 아니다. 그 차이가 얼마나 중요한지에 대한 척도도 아니다. 대신, 이렇게 볼 수 있다. P값은, 놀라움의 척도이다. 약효가 전혀 없고 두 집단의 감기 유병기간이 다를 이유라고는 우연 말고는 없다고 가정했었는데도 아주 작은 P값을 얻었다면 이 결과는 더욱 놀랍고 운이 좋았던 것이다. 운이 좋았던 것이 아니라면, 그 약은 정말로 작용한 것이다.

How do you translate a p value into an answer to this question: “Is there really a difference between these groups?” A common rule of thumb is to say that any difference where p < 0.05 is statistically significant. The choice of 0.05 isn’t because of any special logical or statistical reasons, but it has become scientific convention through decades of common use.

자, 우리는 다음 질문에 대해서 P값을 가지고 설명해야 한다: “이 집단들 간의 차이가 정말로 존재하는 것인가?”에 대해서. 통용되는 일반칙은, P값이 0.05 미만의 차이일 때 통계적으로 유의한 차이라고 말한다. 특별한 논리학적, 통계학적 근거가 있어서 0.05가 선택된 것은 아니다. 과학계의 관례일 뿐으로서 그렇게 수십 년 동안 통용되었다.

Notice that the p value works by assuming there is no difference between your experimental groups. This is a counterintuitive feature of significance testing: if you want to prove that your drug works, you do so by showing the data is inconsistent with the drug not working. Because of this, p values can be extended to any situation where you can mathematically express a hypothesis you want to knock down.

P값을 계산할 때 “실험에 동원된 집단들 간의 차이가 없다”라고 전제했던 것에 주목하자. 이게 바로 유의성 검정이 직관에 반하는 현상이다. 어떤 약이 약효를 가진다는 것을 증명하려면 그 약이 약효가 없는 약과 비교해서 일치하지 않는 결과를 보임으로써 증명하는 셈이다. (시작이) 이렇기 때문에 (우리는) P값을 확대적용함으로써 폐지하고자 하는 가설이 있을 때 이 가설을 수학적으로 표현하는 방식으로 P값을 적용하는 것이다.

But p values have their limitations. Remember, p is a measure of surprise, with a smaller value suggesting that you should be more surprised. It’s not a measure of the size of the effect. You can get a tiny p value by measuring a huge effect—“This medicine makes people live four times longer”—or by measuring a tiny effect with great certainty. And because any medication or intervention usually has some real effect, you can always get a statistically significant result by collecting so much data that you detect extremely tiny but relatively unimportant differences.

그런데 P값은 고유의 제약을 가지고 있다. 기억하겠지만 P값은 놀라움의 척도여서 작은 값이 산출되면 더 크게 놀라야 한다는 뜻이다. P값은 효과의 크기를 나타내는 척도가 아니다. 어떤 거대한 효과를 측정하였을 때도 아주 작은 P값을 얻을 수도 있지만… (가령, “이 약은 사람들을 네 배쯤 더 오래 살게 만든다”처럼… 하지만) 아주 미미하지만 대단히 확실한 효과를 측정해냈을 경우에도 아주 작은 P값을 얻을 수 있다. 모든 약물과 모든 중재술은 일반적으로 얼마간의 실제 효과를 가질 것이기 때문에, 우리가 정말 많은 자료를 모으기만 한다면 통계적으로 유의한 결과를 얻는 것이 항상 가능하기 때문에, 그저 중요치 않고 극도로 미미한 차이까지 검출하게 된다.

As Bruce Thompson wrote, Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. This tautology has created considerable damage as regards the cumulation of knowledge.

브루스 톰슨의 말이다: 통계학적 유의성검정에는 동어반복의 논리가 들어 있다. 수백 명 피험자의 자료를 모으고 이미 충분하다고 느끼고 있는 연구자가 피험자가 충분한지 아닌지 질문에 대한 답을 알아내기 위해서 통계학 검정을 한다. 연구자 스스로 자료를 모았으며 그들은 이미 더 모을 생각이 없으므로 질문의 답을 이미 알고 있기 때문이다. 이 동어반복은, 지식의 축적이라는 면에서 어마어마한 악영향을 끼치고 있다.

In short, statistical significance does not mean your result has any practical significance. As for statistical insignificance, it doesn’t tell you much. A statistically insignificant difference could be nothing but noise, or it could represent a real effect that can be pinned down only with more data. There’s no mathematical tool to tell you whether your hypothesis is true or false; you can see only whether it’s consistent with the data. If the data is sparse or unclear, your conclusions will be uncertain.

간단히 말해서, 통계학적으로 유의한 결과가 실용적으로 중요성을 가진다는 뜻이 아니다. (반대 상황) 통계학적으로 유의하지 않음에 대해서 말하자면 (이해하기 더 어려운데) 이게 뜻하는 바는 별로 없다. 즉, 통계학적으로 유의하지 않은 차이라는 것은 어쩌면 잡음 말고는 아무것도 아닐 수도 있고, 또 어쩌면 자료를 더 모았을 때 분명히 드러날 실제 효과인 경우도 있다.

Psychic Statistics 상식을 초월하는 통계학

Hidden beneath their limitations are some subtler issues with p values. Recall that a p value is calculated under the assumption that luck (not your medication or intervention) is the only factor in your experiment, and that p is defined as the probability of obtaining a result equal to or more extreme than the one observed. This means p values force you to reason about results that never actually occurred—that is, results more extreme than yours. The probability of obtaining such results depends on your experimental design, which makes p values “psychic”: two experiments with different designs can produce identical data but different p values because the unobserved data is different.

P값이 가진 한계의 저변에는 다소 미묘한 쟁점이 있다. P값이 계산될 때에는 실험에 개입한 유일한 요인이 (실험에 쓰인 약제나 중재술이 아니라) 행운뿐이라는 전제가 세워졌었다는 것을 기억하자. 그래서 눈에 보이는 차이와 같거나 더욱 큰 차이의 결과를 실제로도 얻을 수 있는 확률로 P값이 정의된다. 즉 (유의한) P값을 받았을 때 연구자는 억지로라도 실제로는 일어날 수 없는 결과(가 일어난 원인)에 대해서 추론해내야 한다. 이와 같은 결과를 얻을 확률은 실험의 디자인에 달려 있다. P값이 상식을 초월한다고 말하는 것이 이 때문이다. 두 실험연구의 디자인이 다르고 관찰 밖의 자료가 다르다면 동일한 자료를 관찰했어도 다른 P값을 가질 수 있다.

Suppose I ask you a series of 12 true-or-false questions about statistical inference, and you correctly answer 9 of them. I want to test the hypothesis that you answered the questions by guessing randomly. To do this, I need to compute the chances of you getting at least 9 answers right by simply picking true or false randomly for each question. Assuming you pick true and false with equal probability, I compute p = 0.073.[3] And since p > 0.05, it’s plausible that you guessed randomly. If you did, you’d get 9 or more questions correct 7.3% of the time.2

선생이 12개의 오엑스 문제를 연속으로 냈고, 학생이 9개를 올바르게 답했다고 하자. 선생은 학생이 무작위로 답했을 것이라는 가설을 시험하고자 한다. 한 문제마다 임의의 답을 했을 때 적어도 9개의 정답을 만날 확률을 계산하면 된다. 계산해 보면 P = 0.073이다. 이 숫자는 0.05보다 크므로 학생이 무작위로 답했다는 가설 쪽이 그럴싸하게 들린다. 즉, 무작위로 답해도 전체의 7.3퍼센트에서는 9개 또는 그 이상의 정답을 맞힐 수 있다는 뜻이다.

But perhaps it was not my original plan to ask you only 12 questions. Maybe I had a computer that generated a limitless supply of questions and simply asked questions until you got 3 wrong. Now I have to compute the probability of you getting 3 questions wrong after being asked 15 or 20 or 47 of them. I even have to include the remote possibility that you made it to 175,231 questions before getting 3 questions wrong. Doing the math, I find that p = 0.033. Since p < 0.05, I conclude that random guessing would be unlikely to produce this result.

그런데, 선생의 처음 계획은 단지 12개의 문제만 내는 것이 아니었을 수도 있다. 학생의 오답이 3개가 나올 때까지 컴퓨터를 이용해서 문제가 무한히 나오게 했었을 수도 있다. 그렇다면 15개 중에서 3개를 틀린 것, 20개 중에서 3개를 틀린 것, 또는 47개 중에서 3개를 틀린 경우의 확률도 계산할 수 있어야 한다. 무척 길게 잡아서 17만5천 개의 질문 가운데 3개를 틀린 경우도 포함할 수 있어야 한다. (12개의 경우를 새로) 계산하면 P = 0.033이다. 0.05보다 작으므로 무작위 답변으로는 이런 결과를 낳을 수 없다는 결론을 내린다.

This is troubling: two experiments can collect identical data but result in different conclusions. Somehow, the p value can read your intentions.

골치아픈 일이다. 두 실험은 동일한 결과를 나타내면서도 결론은 상이하다. P값에는 어쩐지 연구자의 의도가 반영되어 있는 것 같다.

Neyman-Pearson Testing 나이만-피어슨 검정

To better understand the problems of the p value, you need to learn a bit about the history of statistics. There are two major schools of thought in statistical significance testing. The first was popularized by R.A. Fisher in the 1920s. Fisher viewed p as a handy, informal method to see how surprising a set of data might be, rather than part of some strict formal procedure for testing hypotheses. The p value, when combined with an experimenter’s prior experience and domain knowledge, could be useful in deciding how to interpret new data.

P값이 가진 문제를 이해하려면 통계학의 역사에 대해서 조금은 알아 둘 필요가 있다. 통계학적인 유의성 검정에 관해 상이한 의견을 가진 두 개의 학파가 있다. 첫 번째는 1920년대에 로널드 피셔가 대중화한 내용이다. 피셔는, P값이, 가설을 검정하는 엄격하고 공식적인 절차라기 보다는 일군의 자료가 보일 수 있는 놀라움의 척도이므로, (P값을) 가볍고 비격식적인 수단이라고 간주했다. 그리하여 P값에다가 연구자의 사전 경험과 해당 분야의 지식을 결합함으로써 새로운 자료를 어떻게 해석해야 할지 결정할 때(만) 유용할 수 있다고 보았다.

After Fisher’s work was introduced, Jerzy Neyman and Egon Pearson tackled some unanswered questions. For example, in the cold medicine test, you can choose to compare the two groups by their means, medians, or whatever other formula you might concoct, so long as you can derive a p value for the comparison. But how do you know which is best? What does “best” even mean for hypothesis testing?

피셔의 제안 이후로, 저지 나이만과 에곤 피어슨이 해답이 나오지 않는 문제들과 씨름했다. 가령, 감기약 시험으로 치면, 연구자는 두 군을 비교할 때 P값만 계산해낼 수 있다면, 평균을 써서 비교할 수도 있고, 중간값을 사용할 수도 있고, 또는 지어낼 수 있는 모든 수식으로부터 나온 어떠한 값이라도 사용할 수도 있다. 그렇다면 어느 것이 최선인가? 그리고 가설검정에서 “최선”이라는 게 있기는 한 건가? 하는 문제들.

In science, it is important to limit two kinds of errors: false positives, where you conclude there is an effect when there isn’t, and false negatives, where you fail to notice a real effect. In some sense, false positives and false negatives are flip sides of the same coin. If we’re too ready to jump to conclusions about effects, we’re prone to get false positives; if we’re too conservative, we’ll err on the side of false negatives.

과학에서는 두 가지 오류를 줄이는 것이 중요하다. 하나는 거짓양성으로서 실제로는 없는 효과를 있다고 판단하는 것이고, 다른 하나는 거짓음성으로서, 정말로 효과가 있는데 밝혀내지 못한 경우이다. 어떤 의미에서는, 거짓양성과 거짓음성은 동전의 양면이다. 효과가 있는 결론을 향해 우리가 서둘러 덤비면 거짓양성을 얻기 십상이고 우리가 만약 너무 보수적일 땐 거짓음성 쪽의 실수를 범할 것이다.

Neyman and Pearson reasoned that although it’s impossible to eliminate false positives and negatives entirely, it is possible to develop a formal decision-making process that will ensure false positives occur only at some predefined rate. They called this rate α, and their idea was for experimenters to set an α based upon their experience and expectations. So, for instance, if we’re willing to put up with a 10% rate of false positives, we’ll set α = 0.1. But if we need to be more conservative in our judgments, we might set α at 0.01 or lower. To determine which testing procedure is best, we see which has the lowest false negative rate for a given choice of α.

거짓양성과 거짓음성을 완전히 없앨 수는 없지만 규격화된 의사결정 절차를 개발함으로써 미리 정한 일정한 빈도 아래로 거짓양성의 발생을 줄일 수 있다고 추론했던 것이 나이만과 피어슨이다. 이 빈도를 알파라고 불렀다. 이런 착상을 통해 실험자들은 그들의 경험과 기대에 기초하여 알파를 설정하게 되었다. 그래서, 이를 테면, 거짓양성을 10퍼센트까지만 용인하고자 할 때 알파를 0.1로 설정하면 된다. 만약 보수적인 판단이 필요하다면 알파를 0.01이나 그 미만으로 잡으면 될 것이다. 그리고 어떤 검정 절차가 최선인지 알기 위해서 주어진 알파에서 가장 낮은 거짓음성률을 나타내는 절차가 어떤 것이지 확인한다.

How does this work in practice? Under the Neyman–Pearson system, we define a null hypothesis—a hypothesis that there is no effect—as well as an alternative hypothesis, such as “The effect is greater than zero.” Then we construct a test that compares the two hypotheses, and determine what results we’d expect to see were the null hypothesis true. We use the p value to implement the Neyman-Pearson testing procedure by rejecting the null hypothesis whenever p < α. Unlike Fisher’s procedure, this method deliberately does not address the strength of evidence in any one particular experiment; now we are interested in only the decision to reject or not. The size of the p value isn’t used to compare experiments or draw any conclusions besides “The null hypothesis can be rejected.”

이게 실무에서는 어떻게 작동하는가? 나이만-피어슨 체계에서는, 귀무가설이라는 것을 정의한다. 귀무가설은 실제로는 효과는 없다는 가설이다. 그리고 대립가설을 세운다. 대립가설은 실제 효과는 0보다 크다는 가설이다. 그리고나서 연구자는 두 가설을 비교하는 검정을 구성하며 연구자가 알고자 하는 것은 귀무가설이 참이라는 결론인가 하는 것이다. 이때 P값이 알파보다 작으면 귀무가설을 기각한다는 나이만-피어슨 검정 절차를 수용할 때 P값을 사용한다. 피셔의 절차와는 달리, 이 방법은 의도적으로 어느 한쪽 가설이 더 높은 증거력을 가졌다고 간주하지 않는다. 연구자들이 결정을 내릴 때 단지 기각하거나 기각하지 않는 것이다. P값의 크기는 실험을 비교하는 데 사용되지 않으며 “귀무가설을 기각할 수 있다” 대신에 어떠한 결론도 내리지 않는다.

As Neyman and Pearson wrote, We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

나이만과 피어슨이 썼던 것처럼, 특정한 가설을 염두에 두었을 때, 확률론에 기초한 어떠한 검정도 그 자체로는 그 가설이 참인지 거짓인지에 대한 쓸만한 가치를 제공하지 않는다.

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.3

그러나 연구자들은 검정을 하면서 그 취지를 또다른 관점에서 조망하기도 한다. 각각의 가설이 참인지 거짓인지 당장 알려고 희망하는 게 아니라 그것을 둘러싼 연구자들의 행동들을 지배하는 규칙을 탐색하는 것이다. 경험이 길게 축적될 때(자료가 축적될 수록) 연구자가 오류를 범하는 일이 너무 자주 일어나서는 안 된다는 점에 대해서 확실히 해야 한다는 것이다.

Although Neyman and Pearson’s approach is conceptually distinct from Fisher’s, practicing scientists often conflate the two.4,5,6 The Neyman-Pearson approach is where we get “statistical significance,” with a prechosen p value threshold that guarantees the long-run false positive rate. But suppose you run an experiment and obtain p = 0.032. If your threshold was the conventional p < 0.05, this is statistically significant. But it’d also have been statistically significant if your threshold was p < 0.033. So it’s tempting—and a common misinterpretation—to say “My false positive rate is 3.2%.”

나이만과 피어슨의 접근법이 개념적으로 피셔 쪽과 아주 다르기는 하지만, 현장에서 일하는 과학자들은 이 둘을 융합하는 일이 흔하다. 즉, (아주 많은 자료로부터 위양성을 범하는 비율을 통제할 수 있는) 미리 정해진 P값 역치를 가지고 “통계적 유의성”을 얻으면서 나이만-피어슨을 써먹고, (가령) 실험을 해서 P = 0.032를 얻었다고 쳤을 때, 기왕의 역치가 P < 0.05였다면 통계적으로 유의한 것이고, 역치가 P < 0.033이었더라도 통계적으로 유의했을 것이므로 흔히 범하는 실수를 하는 것이다. 이렇게, “내 연구의 거짓양성률은 3.2퍼센트이다”라고.

But that doesn’t make sense. A single experiment does not have a false positive rate. The false positive rate is determined by your procedure, not the result of any single experiment. You can’t claim each experiment had a false positive rate of exactly p, whatever that turned out to be, when you were using a procedure to get a long-run false positive rate of α.

이건 이치에 맞는 해석이 아니다. (주목하라) 단 한 번의 실험은 거짓양성률을 가질 수 없다. 거짓양성률은 절차를 통해서 결정될 뿐, 단발 실험결과로부터 나올 수 없다. 연구자들은 개별 실험의 개별 P값으로부터 거짓양성률이 얼마라고 주장할 수 없다. 일련의 같은 실험을 여러 차례 반복하여 알파값의 거짓양성률을 얻고 났을 때 그때 그 값이 맞았다고 하더라도 말이다.

Significance tests tend to receive lots of attention, with the phrase “statistically significant” now part of the popular lexicon. Research results, especially in the biological and social sciences, are commonly presented with p values. But p isn’t the only way to evaluate the weight of evidence. Confidence intervals can answer the same questions as p values, with the advantage that they provide more information and are more straightforward to interpret.

“통계적으로 유의함”이라는 말이 이제 흔해빠진 문구가 되어 버리면서 유의성검정은 관심을 끌고 있다 . 생물학, 사회학을 비롯한 연구들은 결과에서 P값을 제시하기 마련이다. 하지만 P값 말고도 증거의 무게를 묘사할 수 있다. 즉, 신뢰구간도 같은 질문에 대해서 같은 답변을 할 수 있으며, 오히려 (P값보다) 더 많은 정보를 줄 수 있고 더욱 쉽게 해석할 수 있도록 돕는다.

A confidence interval combines a point estimate with the uncertainty in that estimate. For instance, you might say your new experimental drug reduces the average length of a cold by 36 hours and give a 95% confidence interval between 24 and 48 hours. (The confidence interval is for the average length; individual patients may have wildly varying cold lengths.) If you run 100 identical experiments, about 95 of the confidence intervals will include the true value you’re trying to measure.

신뢰구간은 점추정을 묘사하면서 동시에 그 추정값의 불확실성을 결합한 값이다. 연구자들은 실험 약물이 감기의 유병기간을 36시간만큼 줄였다고 말할 때, 95퍼센트 신뢰구간이 24시간에서 48시간이라고 말하게 된다. (평균 유병기간에 대한 신뢰구간으로서 거칠게 표현해서 각각의 환자들이 겪을 수 있는 유병기간의 폭을 뜻한다) 만약 연구자가 100번의 동일한 실험을 수행하여 100번의 신뢰구간을 측정하면 그 중 95번은 연구자가 알고자 했던 그 값(모집단의 추정값)을 포함할 것이라는 뜻이다.

A confidence interval quantifies the uncertainty in your conclusions, providing vastly more information than a p value, which says nothing about effect sizes. If you want to test whether an effect is significantly different from zero, you can construct a 95% confidence interval and check whether the interval includes zero. In the process, you get the added bonus of learning how precise your estimate is. If the confidence interval is too wide, you may need to collect more data. For example, if you run a clinical trial, you might produce a confidence interval indicating that your drug reduces symptoms by somewhere between 15 and 25 percent. This effect is statistically significant because the interval doesn’t include zero, and now you can assess the importance of this difference using your clinical knowledge of the disease in question. As when you were using p values, this step is important—you shouldn’t trumpet this result as a major discovery without evaluating it in context. If the symptom is already pretty innocuous, maybe a 15–25% improvement isn’t too important. Then again, for a symptom like spontaneous human combustion, you might get excited about any improvement.

(그렇게) 신뢰구간은 연구자들이 내린 결론의 불확실성을 정량화함으로써 P값보다 훨씬 많은 정보를 제공한다. P값은 효과의 크기에 대해서 아무 정보도 주지 못하는데 말이다. 효과가 영보다 크다는 것(영과는 다르다는 것) 을 검정하고자 할 때, 연구자는 95퍼센트 신뢰구간을 계산함으로써 신뢰구간 안에 영이 포함되는지 파악할 수 있다. 그 과정에서 덤으로, 연구자는 계산된 (평균)추정값이 얼마나 정밀한지도 알 수 있다. 만일 신뢰구간이 지나치게 넓다면 연구자는 자료를 더 모으는 게 좋을 것이다. 예를 들어서, 임상시험을 하면서, 어떤 약이 증상을 줄여 주는 효과의 신뢰구간을 계산하여 “15내지 25퍼센트 줄인다”라고 나왔을 때 그 신뢰구간이 영을 포함하지 않으므로 그 약의 효과는 통계적으로 유의한 차이를 가지게 된다. 그리고나서 이 정도 효과의 차이가 얼마나 중요한지 판단하려면 연구자가 가진 질병에 대한 임상적 지식에 기초하여야 한다. P값을 써서 결정할 때도 그랬던 것처럼, (이 과정이 정말 중요한데) (통계적으로 유의한 차이라고 해서) 맥락에 대한 이해 없이 이것이 정말 대단한 발견이라고 팡파르를 울려서는 곤란하다. 그 질병의 증상들이 애초부터 그리 무해한 것이 아니었다면 15내지 25퍼센트 개선은 (설령 통계적으로 유의하다 하더라도) 썩 대단한 것이 아니다. 다시 말하건대, 무척 심각한 증상들, 가령, 사람의 몸이 돌연 불타버린다든가 하는 증상이라면 아무리 작은 개선이 보이더라도 신을 내도 좋다.

If you can write a result as a confidence interval instead of as a p value, you should.7 Confidence intervals sidestep most of the interpretational subtleties associated with p values, making the resulting research that much clearer. So why are confidence intervals so unpopular? In experimental psychology research journals, 97% of research papers involve significance testing, but only about 10% ever report confidence intervals—and most of those don’t use the intervals as supporting evidence for their conclusions, relying instead on significance tests.8 Even the prestigious journal Nature falls short: 89% of its articles report p values without any confidence intervals or effect sizes, making their results impossible to interpret in context.9

P값 대신에 신뢰구간으로 결과를 쓸 수 있다면 신뢰구간을 쓰는 것이 좋다. 신뢰구간을 통해서 P값을 해석하면서 발생하는 해석상의 온갖 미묘한 함정들을 피할 수 있다. 결과적으로 연구는 훨씬 자명해진다. 그렇다면 신뢰구간은 왜 이리도 주목 받지 못해 왔을까? 실험심리학연구 저널들에서 연구논문의 97퍼센트가 유의성검정을 행하고 있는 반면에 겨우 10퍼센트에서만 신뢰구간을 표기하고 있다. 게다가 신뢰구간을 표기한 논문들마저도, 신뢰구간에 기초하여 결론을 보증한 게 아니라 유의성검정의 결과에만 의존해서 결론을 이끌고 있다. 저명한 저널인 네이처조차도 이 점에 대해서는 모자라다. 89퍼센트의 논문이 신뢰구간이나 효과크기 없이 P값만 보고하고 있어서 문맥만으로 결과를 해석하는 것이 힘들다.

One journal editor noted that “p values are like mosquitoes” in that they “have an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them.”10 One possible explanation is that confidence intervals go unreported because they are often embarrassingly wide.11 Another is that the peer pressure of peer-reviewed science is too strong—it’s best to do statistics the same way everyone else does, or else the reviewers might reject your paper. Or maybe the widespread confusion about p values obscures the benefits of confidence intervals. Or the overemphasis on hypothesis testing in statistics courses means most scientists don’t know how to calculate and use confidence intervals.

한 저널편집인이 지적했듯, “P값은 모기와 같다.” 모기들은 “진화를 통해 그들만의 적소를 찾았으며, 아무리 긁어도, 아무리 찰싹 때려도, 또는 아무리 약을 뿌려댄다 해도 불행히도 모기를 뿌리뽑을 수는 없다.” (신뢰구간이 잘 쓰이지 않는 또다른 이유로서) 어쩌면 신뢰구간이 너무 넓게 계산되었을 때 저자들이 신뢰구간을 보고하기 꺼리는 것도 한 이유일 것이다. 또 어쩌면, 과학계는 전문가심사에 의존하므로 전문가들의 압력이 너무 셌기 때문일 수도 있다. – 즉, 과학을 하려면 남들이 다 하는 방식으로 하는 게 상책이며 그렇게 따라하지 않으면 심사자들이 논문을 거절할 테니까. 또 어쩌면, P값을 둘러싼 만연한 착각이 신뢰구간의 장점을 가리고 있을지도 모른다. 또 어쩌면, 가설검정에 대한 지나친 맹신을 통계학 수업에서 배움으로써 여러 학자들이 신뢰구간을 계산하거나 해석할 능력을 아예 갖추지도 못했을 수 있다.

Journal editors have sometimes attempted to enforce the reporting of confidence intervals. Kenneth Rothman, an associate editor at the American Journal of Public Health in the mid-1980s, began returning submissions with strongly worded letters:

저널의 편집인들은 신뢰구간을 표기하도록 한때 애써 왔다. 케네스 로스먼이 대표적으로, 1980년대 중반에 American Journal of Public Health의 부편집장으로 있을 때 무척 강경한 어조로 투고된 논문들을 되돌려 보내기 시작했다. 다음과 같다:

All references to statistical hypothesis testing and statistical significance should be removed from the paper. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests), you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere.12

통계학적인 가설검정과 통계학적 유의성에 대해서 언급하지 마십시오. 요청하건대, 저자들은 P값과 통계적 유의성에 대한 기술은 모두 지워야 합니다. 유의성검정이 부적절하다고 보는 내 기준에 동의하지 않으신다면 언제라도 논박하셔도 좋습니다. 논박하지 않고, 그저 내 의견이 잘못되었다고 평가하신 뒤, 논문은 다른 저널에 보내십시오.

During Rothman’s three-year tenure as associate editor, the fraction of papers reporting solely p values dropped precipitously. Significance tests returned after his departure, although subsequent editors successfully encouraged researchers to report confidence intervals as well. But despite reporting confidence intervals, few researchers discussed them in their articles or used them to draw conclusions, preferring instead to treat them merely as significance tests.12

로스먼이 부편집장으로 재직한 3년간 논문들에서 P값만 표기하는 비율은 급격히 감소하였다. (하지만) 그가 떠난 뒤에 뒤이은 편집인들이 비록 로스먼을 이어서 신뢰구간 표기를 격려한 것은 다행이었지만 유의성검정을 표기하는 관행이 되살아났다. 신뢰구간을 표기하기는 하지만, 본문 안에서 그걸 언급하거나 결론을 유도하면서 신뢰구간을 이용하는 저자들은 거의 없어졌다. 모두 단순히 유의성검정에만 매달렸다.

Rothman went on to found the journal Epidemiology, which had a strong statistical reporting policy. Early on, authors familiar with significance testing preferred to report p values alongside confidence intervals, but after 10 years, attitudes had changed, and reporting only confidence intervals became common practice.12 Perhaps brave (and patient) journal editors can follow Rothman’s example and change statistical practices in their fields.

로스먼은 저널 Epidemiology를 창간했고 당연히 거기선 엄격한 통계학 기술 정책을 유지했다. 그 저널의 저자들은 진작부터 신뢰구간도 쓰긴 하지만 P값을 이용하는 유의성검정에 익숙하였지만 10년이 지나고 나니까 저자들의 행태가 바뀌었다. 이제 (그 저널에서는) 신뢰구간만 표기하는 것이 보편화되었다. 그처럼 용감하고 인내심을 가진 저널 편집인이라면 로스먼의 전례에 배워서 자기 분야의 통계 관행을 개혁할 수 있다.