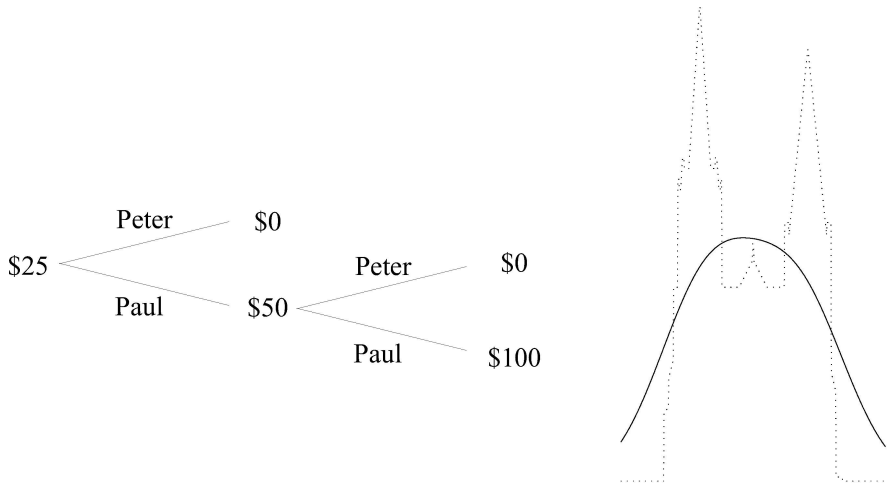
Game-Theoretic Significance Testing

Glenn Shafer
Rutgers Business School
gshafer@business.rutgers.edu



The Game-Theoretic Probability and Finance Project

Working Paper #49

First posted April 16, 2017. Last revised April 19, 2017.

Project web site:
http://www.probabilityandfinance.com

Abstract

How should we adjust p-values to account for multiple testing? This question, first discussed by Cournot in his *Exposition de la théorie des chances and des probabilités* (1843), still puzzles statistical theorists and practitioners. Modern game-theoretic probability, developed by Shafer and Vovk in *Probability and Finanace: It's Only a Game!* (2001), gives us a new way to think about the problem and concrete rules for adjusting and combining p-values.

1 Cournot's principle and Cournotian testing

The French mathematician and philosopher Antoine-Augustin Cournot (1801–1877), remembered by economists for his work on duopoly and on supply and demand, is also remembered for his dictum that probability acquires objective scientific content only by its predictions. This is *Cournot's principle* [46].

To predict using a probabilistic hypothesis, you identify an event to which it gives probability close to one: the prediction is that this event will happen. Or you identify an event to which it gives probability close to zero: the prediction is that this event will not happen. These predictions allow the hypothesis to be tested: if you single out an event to which it assigns very small probability and it happens, then the hypothesis is discredited and may need modification or merit rejection.

Some scholars (especially Bayesians) reject Cournot's principle on the grounds that the actual outcome of a complex process is always an event of small probability. This overlooks the role of the statistician, who selects a particular event in advance as a prediction or a test.

Cournot's writing is a good starting point for understanding aspects of statistical testing still being discussed today, if only because it reminds us that issues debated today were already on the table nearly 200 years ago. His comments on multiple testing are particularly timely.

1.1 When is a deviation attributable to chance?

As Cournot explained, the evidence provided by a predicted event failing is attenuated when the statistician makes many tests of similar hypotheses. Using the example of a statistician who looks for variation in the ratio between male and female births, first looking at whether the births are legitimate and then considering the age, profession, and religion of the parents, the season in which the child is born, and so on, he explained that

> ... for a statistician who undertakes a thorough investigation, the probability of a deviation of given size not being attributable to chance will have very different values depending on whether he has tried more or fewer groupings before coming upon the observed deviation.

He went on to say that because the statistician knows how many groupings he has tried before finding a notable deviation, the probability of its not being attributable to chance still has an objective value for him, though it is diminished by the number of groupings he has tried. But for a member of the public from whom the multiple testing is hidden, the probability loses all objective substance.[1]

[1] The passage quoted (in translation) is from Section 111 of Cournot's *Exposition de la théorie des chances et des probabilités* (1843 [6]). See [47] for additional translations from this and Cournot's other books.

Cournot's assertion that the probability of a deviation being attributable to chance diminishes as more tests are made can be elaborated as follows. Suppose you are looking for a deviation that would happen by chance with probability 0.005 or less. The first deviation you check is not that large, but by trying 9 more times you find one that is. The probability that you would find so large a deviation by chance in 10 tries does not exceed $10 \times 0.005 = 0.05$. So you are entitled to tell the public that the hypothesis was discredited by giving a 95% prediction that failed, but you are not entitled to make the stronger claim that it was discredited by giving a 99.5% prediction that failed.

1.2 The concept of a p-value

Statistical testing in Cournot's sense is now understood in terms of test statistics and p-values.

What Cournot called a deviation we now often call a *test statistic*. Under the hypothesis we are testing, this test statistic, say T, is a random variable. If T comes out equal to t, then the *p-value* is the probability under the hypothesis that T would come out that large or larger:

$$\{\text{p-value from observing } T = t\} := \mathbf{P}(T \geq t). \tag{1}$$

Being a function of what we observe, the p-value is itself a random variable before it is observed. Let us designate this random variable by P, and let us call a random variable of this type a *p-variable*. Thus a p-value p is the observed value of a p-variable P.

The probability that a p-variable P is less than or equal to 5% is less than or equal to 5%. This sounds a bit convoluted, but we get accustomed to thinking this way when we do statistical testing. More generally, a p-variable P will satisfy

$$\mathbf{P}(P \leq p) \leq p \tag{2}$$

for every $p \in [0, 1]$. We can take (2) as the definition of a p-variable and develop the classical theory of statistical testing from this starting point; see [48]. But most people find it more intuitive to start with the notion of a test statistic and define the notion of a p-value by (1).

1.2.1 Composite hypotheses

Often we test a hypothesis that only incompletely specifies probabilities for the test statistic we use. Such a hypothesis fixes only a class of probability distributions, asserting that good predictions can be made by one of the probability distributions in the class without saying which one. We call such a class of distributions a *composite hypothesis* or a *statistical model*. Indexing the probability distributions in the model by a parameter θ that ranges over a set $\mathcal{T}$,[2] we

[2]The use of *parameter* in this context goes back at least to R. A. Fisher's pathbreaking 1922 article, "On the mathematical foundations of theoretical statistics" [15]. The set $\mathcal{T}$ may consist of real numbers, vectors, or more complicated mathematical objects. In contemporary

generalize (1) to

$$\{\text{p-value from observing } T = t\} := \sup_{\theta \in \mathcal{T}} \mathbf{P}_\theta(T \geq t).$$

The condition (2) that characterizes a p-variable then becomes

$$\mathbf{P}_\theta(P \leq p) \leq p$$

for all $p \in [0, 1]$ and $\theta \in \mathcal{T}$ or, equivalently,

$$\sup_{\theta \in \mathcal{T}} \mathbf{P}_\theta(P \leq p) \leq p$$

for all $p \in [0, 1]$.

On the other hand, we can accommodate incompleteness or imprecision in the specification of the probabilistic hypothesis, along with other limitations on our ability to calculate probabilities for T and P from the hypothesis, without formally introducing the notion of a statistical model. Instead, we say that the hypothesis we are testing is a probability distribution $\mathbf{P}$ about whose probabilities we have only partial knowledge or limited computational facility. Instead of using (1) as our definition, we say that (a) the p-value p from observing $T = t$ is the least upper bound we can calculate on $\mathbf{P}(T \geq t)$, and (b) the p-variable P is the random variable whose realized value is this least upper bound. The inequality (2) follows.

If the probability distribution is fully specified, the test statistic T is continuous under $\mathbf{P}$, and we have no difficulties in computing T's probabilities, then we can replace (2) with

$$\mathbf{P}(P \leq p) = p. \tag{3}$$

This says that P is uniformly distributed on $[0, 1]$.

1.2.2 The venerability of p-values

As the quotations from Cournot demonstrate, hypothesis testing using what we now call p-values was already a familiar statistical tool in the middle of the 19th century. The idea is often traced back to a note published by John Arbuthnot (1667–1735) in the *Philosophical Transactions of the Royal Society of London* in 1710. Noting that male births had exceeded female births in London for 82 successive years, Arbuthnot argued that male and female births could not have equal chances. This and other 18th and 19th centuries examples of hypothesis testing, including some in the work of Laplace, are discussed by Stigler [51], Hald [25], and Gorroochurn [23].

On the other hand, the framework in which statistical testing is now most often conducted, correlation and multiple regression, was developed only in the late 19th and early 20th centuries, primarily by British statisticians. The

usage, *parameter* may refer the entire object θ or to a component of θ or to an arbitrary real-valued function of θ.

statistical concept of correlation first emerged in the work of the gentleman scholar and eugenicist Francis Galton (1822–1911) in the 1880s. It was taken up by Karl Pearson (1857–1936), who organized a statistical laboratory at the University of London and founded two journals, *Biometrika* in 1901 and the *Annals of Eugenics* in 1925. The first exercise in multiple regression, on the relationship between poverty and the generosity of government welfare, was carried out in 1895 by George Udny Yule (1871–1951), who began his career in statistics as an assistant to Pearson. R. A. Fisher (1890–1962), who shared Galton's and Pearson's zeal for eugenics, extended all this work, establishing paradigms still dominant in many areas of applied statistics. This group of statisticians is often referred to collectively as the English or British (Yule was a Scot) school of statistics or biometry. (See [2, 23, 37, 51].)

The word *significant* was used as a technical term in statistical testing well before the beginning of the twentieth century. We find it in boldface in the first edition of Yule's *An Introduction to the Theory of Statistics* ([59], 1911, page 262): "...if we observe a different proportion in one sample from that which we have observed in another, the question again arises whether this difference may be due to fluctuations of simple sampling alone, or whether it indicates a difference between the conditions subsisting in the universes from which the two samples were drawn: in the latter case the difference is often said to be **significant**."

The name *p-value* came into use only beginning in the 1970s. The British school used p-values as an indication of significance but did not have a formal name for them. They sometimes used the phrase "value of P". Fisher referred informally to the "value of P" in his influential *Statistical Methods for Research Workers*, from its first edition in 1925 to its thirteenth in 1958 (see also [17]).

Fisher's originality with respect to p-values is sometimes exaggerated. In 1993 ([21], page 486), Steven N. Goodman wrote that, "Fisher was not the first to use the p value, but he was the first to outline formally the logic behind its use, as well as the means to calculate it in a wide variety of situations." In 2001 ([22], page 295), Goodman wrote that, "The references on this topic encompass innumerable disciplines, going back almost to the moment that P-values were introduced (by R.A. Fisher in the 1920s)." Citing Goodman, Campbell Harvey [26] states that, "The idea of using a p-value for hypothesis testing was introduced by Fisher (1925)." As of this writing, similar statements appear in Wikipedia articles on significance testing.

1.3 The concept of a significance level

According to the Neyman-Pearson theory of testing, formulated by Jerzy Neyman (1894–1981) and E. S. Pearson (1895–1980) in the late 1920s, you should decide before looking at the data on the probability of false rejection you will tolerate.[3] This probability α is the *significance level*. In many areas of research,

[3]E. S. Pearson, Karl Pearson's son, worked in the department of statistics in London founded by his father. Neyman collaborated with Pearson by correspondence from Poland during the late 1920s and early 1930s, joined Pearson in London in 1934, and then immigrated

a significance level of 5% is conventional.

The Neyman-Pearson theory does not require the use of a test statistic. It merely asks us to specify, in advance of seeing the data, an event E (the *rejection region*) that has probability α or less. We then reject the hypothesis if and only if E happens. But as Neyman and Pearson expected and intended, their theory is most often implemented using a test statistic T: we perform a level α test by choosing a value c such that $\mathbf{P}(T \geq c) = \alpha$ and rejecting if $T \geq c$—i.e., rejecting if the p-value given by (1) is less than or equal to α.

Neyman and Pearson considered not only the probability of rejecting the hypothesis when it is true (*Type I error*) but also the probability of failing to reject it when it is false (*Type II error*). They advocated balancing the probabilities of the two types of errors in light of the costs and benefits of the actions associated with rejecting or failing to reject. This balancing might lead to a value of α much larger or much smaller than 5%.

The Neyman-Pearson foundation for statistical testing is clear, mathematically interesting, and conceptually deeper than the mere notion of calculating the probability that a deviation can be attributed to chance. For these reasons it became dominant in mathematical statistics in the mid-twentieth century. The idea of balancing the probabilities of Type I and Type II error was implemented in a number of engineering and business contexts.

1.3.1 Neyman-Pearson vs. p-values

In spite of popularity of the Neyman-Pearson theory among mathematical statisticians, many researchers in the natural and social sciences have continued to focus on p-values, ignoring or paying only lip service to the notion of a fixed significance level α. We can distinguish three arguments for this continued emphasis on p-values:

1. The goal is a scientific conclusion or the assessment of the evidence for or against a scientific conclusion, not an action with definable costs and benefits.

2. When assessment of the evidence is the goal, no one wants to stop after merely stating that the hypothesis was discredited by the occurrence of an event E of low probability α, where E and α were fixed before looking at the data. The scientist wants his or her public to know just how strong the evidence against the hypothesis turned out to be. A p-value of one in a million is surely much stronger evidence than a p-value of one in twenty, and everyone should hear about it.

3. By providing a p-value, a scientific article enables each reader to choose their own level of significance α.

The contrast between "forward looking" Neyman-Pearson significance levels and "backward-looking" p-values was one element of the vigorous debate about sig-

to the United States in 1938.

nificance testing that was already underway in the 1950s and 1960s. See, for example, the reader published by Morrison and Henkel in 1970 [38].

Another aspect of the long-running debate around hypothesis testing, probably more important, concerns the difference between statistical significance and substantive significance. As many authors have noted, statistically significant differences are often reported, published, and acted upon even though they are substantively insignificant. See [60] for one persuasive jeremiad against this widespread and continuing phenomenon. See also Section 2.3 below.

1.3.2 Efficient test statistics

Our choice of a test statistic T or rejection region E depends on what sort of deviation we want to detect. Usually we have in mind some alternative to the hypothesis being tested, perhaps a different probabilistic hypothesis, or perhaps some less precise notion of what might happen. Neyman and Pearson's concept of Type II error provides one way to formalize this insight.

Suppose the hypothesis being tested is a single probability distribution $\mathbf{P}$ (rather than a composite hypothesis), and suppose the alternative is another probability distribution $\mathbf{Q}$ (rather than a class of probability distributions or something less precise). Then as Neyman and Pearson showed in 1933 [40], the best trade-off between Type I and Type II errors is achieved by the *likelihood ratio*, the test statistic T given by

$$T(y) := \frac{\mathbf{q}(y)}{\mathbf{p}(y)}, \tag{4}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the probability densities for $\mathbf{P}$ and $\mathbf{Q}$, respectively, and y is the complete outcome; this is the *Neyman-Pearson lemma*. The name *likelihood* had been coined by R. A. Fisher in 1921 [14, 10], and statisticians had been using the notion without the name for over a century [53].

Because the hypothesis being tested is usually composite, and the alternative usually composite or ill defined, Neyman and Pearson's insight concerning the likelihood ratio is not always directly usable. But as we will see in Sections 2.1 and 3.1, this ratio also emerges when we look at hypothesis testing from other perspectives.

1.4 Are p-values and significance levels frequentist?

It is conventional to distinguish two principal schools of thought in mathematical statistics: frequentist and Bayesian. According to frequentists, probabilities are frequencies; according to Bayesians, they are degrees of belief. Both p-values and Neyman-Pearson statistical testing are considered tools of the frequency school. But as we have just seen, p-values and testing with a fixed significance level can be understood without appealing to the notion of frequency. It is enough that a probability close to zero be understood as a prediction, and that we test the theory (or statistician or forecaster) producing the prediction by checking whether the prediction is successful [24, 39].

As an experiment, to emphasize that we can understand p-values and fixed-level significance testing without bringing in the notion of frequency, I will call these tools *Cournotian* rather than *frequentist*.

2 The Bayesian challenge to Cournot's principle

Cournot's effort to distinguish between the subjective and the objective aspects of probability came in the wake of the monumental work of the French mathematician Pierre Simon Laplace (1749–1827), who developed both Bayesian estimation and non-Bayesian methods of estimation and testing without troubling himself about possible conflicts. In the second half of the nineteenth century, the conflicts attracted increasing attention, and debate began in earnest between those who insisted on probability's subjectivity those who insisted on its objectivity, and between those who gave priority to Bayesian estimation (then called *inverse probability* in English) and those who favored older non-Bayesian methods.[4] In Germany, Carl Stumpf (1848–1936) argued that probability can only be subjective, while Johannes von Kries (1853–1928) argued for an objective conception. In Britain, inverse probability was sharply criticized by John Venn (1834–1923) and George Chrystal (1851–1911) and defended with equal vigor by W. Allen Whitworth (1840–1905).[5]

The most influential mathematical statisticians of the mid twentieth century, R. A. Fisher and Jerzy Neyman, were outspoken critics of Bayesian methods, but Bayesian statistics also had its proponents during this period, notably the British physical scientist Harold Jeffreys (1891–1989),[6] the Italian mathematician Bruno de Finetti (1906–1985), and the American statistician Leonard J. Savage (1917–1971). In the spirit of Laplace, Jeffreys sought to make inverse probability objective as well as subjective; he advanced suggestions for choosing prior probability distributions that would reflect lack of information about parameter values. De Finetti and Savage, on the other hand, favored a thoroughly subjective interpretation of probability.

Jeffreys, de Finetti, and Savage had no use for Cournotian statistical testing.

[4] Campbell Harvey is mistaken when he guesses ([26], page 19) that "The long-standing debate between the Bayesian and frequentist statisticians likely originated with Berkson's (1938) observation that you can reject any null hypothesis with enough data." The names "Bayesian" and "frequentist" came into the widespread use, however, only starting in the 1970s [13].

[5] See [8] for a detailed account of inverse probability from Thomas Bayes to Karl Pearson. As Stephen M. Stigler has noted [51, 52, 54], applications of inverse probability during this period tended to rely on Laplace's principle that the probabilities of causes after an effect is observed should be proportional to the probabilities the causes would have given to the effect. This corresponds to the assumption of a uniform prior distribution. In [35], Laplace had mentioned that an additional factor would enter if the causes were unequally probable before the effect was observed, but this was seldom or never implemented in applications. Bayes's picture, in which a joint probability for cause and effect (parameter and observation, in twentieth-century terminology) is constructed and then conditioned on the observed effect, was not usually used, perhaps because his argument for what we now call conditional probabilities for the parameter given the observation was unpersuasive [44].

[6] Jeffreys and Fisher debated the meaning of probability at length in the 1930s; see [1, 29].

When we ask whether a hypothesis is true, they argued, we should ask about its probability, not about whether some apparent deviation "can be attributed to chance". Jeffreys ridiculed the notion of a p-value as defined by Equation (1). If we observe $T = t$, then why, he asked, should we add to $\mathbf{P}(T = t)$ the probability $\mathbf{P}(T > t)$, the probability of an event that did not happen? Should we reject a hypothesis for assigning a small probability to something that did not happen?

By the beginning of the twenty-first century, the methods for selecting estimates and test statistics championed by Fisher, Neyman and their successors were proving inadequate for the increasingly complex models needed to handle more extensive and complex data, whereas Bayesian methods were looking more adaptable and more amenable to large-scale computation. Some philosophically-inclined statisticians see in this development a triumph for de Finetti and Savage's thoroughly subjective point of view, but others who use Bayesian methods still aspire to objectivity in Cournot's sense.

To gain a deeper understanding of this state of play, let us review the Bayesian alternative Jeffreys proposed to Cournotian testing in the 1930s.

2.1 Jeffreys's Bayesian significance test

Jeffreys presented his Bayesian method of significance testing in a 1935 article [31] and in his book *Theory of Probability* [32].[7] He began by considering the familiar situation where we ask whether a particular parameter β in a statistical model (perhaps the coefficient of a particular independent variable in a multiple regression) should be set to zero, and he took advantage of R. A. Fisher's non-Bayesian work on the likelihood function and the method of estimation by maximum likelihood.

Suppose, following Fisher, that the model is sufficiently regular and the number of observations is sufficiently large that

1. the maximum-likelihood estimator $\hat{\beta}$ is approximately normal with mean β and a standard deviation s well enough estimated that we can take it as known, and

2. the likelihood function is approximately proportional to a normal density with mean equal to the observed value of $\hat{\beta}$ and standard deviation s.

A Bayesian analysis requires prior probabilities for β. Jeffreys assigned half his prior probability to the null hypothesis $\beta = 0$ and distributed the other half over β's possible values according to a continuous probability density $f(\beta)$. His "significance test" consisted of calculating the posterior odds in favor of $\beta = 0$. By Bayes's theorem,

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}. \tag{5}$$

[7] Antecedents in Jeffreys's work with Dorothy Wrinch and in the work of others are discussed by Etz and Wagenmakers [12].

Because $\mathbf{P}(\beta = 0) = \mathbf{P}(\beta \neq 0) = 1/2$, the prior odds reduces to unity and so (5) reduces to

$$\begin{aligned} &\text{posterior odds} = \text{likelihood ratio} \\ &\approx \frac{\frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(0-\hat{\beta})^2}{2s^2}\right)}{\int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(b-\hat{\beta})^2}{2s^2}\right) f(b)db}. \end{aligned} \tag{6}$$

Suppose the range of values of β that we consider reasonably possible is large relative to the expected accuracy s of the estimate $\hat{\beta}$. Then we may assume that the prior density f is fairly constant over the range $\hat{\beta} \pm 3s$, say, and in this case the integral in (6) will approximate $f(\hat{\beta})$, reducing (6) to

$$\text{posterior odds} \approx \frac{1}{f(\hat{\beta})s\sqrt{2\pi}} \exp\left(-\frac{\hat{\beta}^2}{2s^2}\right). \tag{7}$$

How do these posterior odds for $\beta = 0$ compare to conclusions we might draw from a p-value? It is natural to ask the question when the usual test statistic $|\hat{\beta}/s|$ is just barely statistically significant, say approximately equal to 2. In this case, Jeffreys's posterior odds can diverge sharply from the p-value:

- If $\beta = 0$, $\hat{\beta}/s$ is approximately normal with mean 0 and variance 1. The p-value is $\mathbf{P}(|\hat{\beta}/s| \geq 2) \approx 0.05$, suggesting we should reject $\beta = 0$.
- For clarity, choose the units so that s=1 and thus $\hat{\beta} = 2$. We assumed that the range of reasonably possible values for β is great relative to s; pushing this to something of an extreme, suppose the prior density f is uniform on the range from -100 to 100. Then (7) comes out to

$$\frac{1}{(1/200)\sqrt{2\pi}} \exp(-2) \approx 10.8,$$

  corresponding to a probability of over 90% for $\beta = 0$.[8]

This divergence suggests that a p-value may drastically overstate the evidence against a hypothesis. On the other hand, we can question the prior probabilities for β. We are giving probability 1/2 to the precise hypothesis $\beta = 0$ while choosing a density f that expresses great uncertainty about β. This dissonance can be seen as a symptom of the fundamental impossibility of using probabilities to express ignorance.[9]

Jeffreys's effort to find probabilities that express ignorance was situated in a long tradition. Laplace had used a uniform distribution of probabilities to

[8] The odds o in favor of an event are related to its probability p by $o = p/(1-p)$, so that $p = o/(o+1)$. So if the posterior odds come out approximately 10.8, the posterior probability is approximately $10.8/11.8 \approx 0.92$.

[9] A very spread-out density f may be based on experience rather than on ignorance. But in this case the relevance of the experience to the particular case is always at issue, especially when extreme probabilities are deduced from it. See [45] and the following discussion.

express ignorance (he called this the *principle of insufficient reason*), and this idea was still popular in the late 19th century and early 20th centuries, especially among logicians and mathematicians working on geometric probability. It is still has support among some physicists and philosophers, who see it as a foundation for statistical mechanics. But its shortcomings were also already well known in the late 19th century. John Maynard Keynes, who renamed it the *principle of indifference* and tried unsuccessfully to defend it in a limited form ([34], Chapter IV), listed some of the shortcomings: There is no uniform probability distribution when there are infinitely many discrete possibilities, the meaning of uniformity depends on the parametrization when the possibilities lie in a continuous range, and different ways of counting arise even when there are only a few possibilities.[10] Jeffreys attempted to overcome these problems by basing prior probabilities for parameters in statistical models on the properties of the models, but his recommendations had their own inconsistencies.

2.2 Bayes factors

The Bayesian revival of the late twentieth and early twenty-first centuries has largely abandoned the tradition of Laplace and Jeffreys in favor of de Finetti and Savage's thoroughly subjective Bayesianism. In the thoroughly subjective view, (1) applied statisticians always have prior information and (2) rationality demands that they somehow find probabilities that represent it.[11]

Suppose we drop the assumption that $\mathbf{P}(\beta = 0) = 1/2$, on the grounds that each person should provide their own subjective prior probability for $\beta = 0$. In this case, (5) still tells you to multiply your prior odds by the likelihood ratio to get your posterior odds, and that the information in the data affects your prior odds only through the likelihood ratio. To emphasize this role for the likelihood ratio (and perhaps to acknowledge that it too has a subjective component, the prior density f), many statisticians now call it the *Bayes factor*.[12]

The divergence we just noted, between a p-value of 0.05 that seems to refute $\beta = 0$ and a Bayes factor that favors it by more than 10 to 1, is still of interest when we drop the assumption that $\mathbf{P}(\beta = 0) = 1/2$, but we should also consider the Bayes factors that might result from other choices for the prior f. What is the least Bayes factor, the one least favorable to $\beta = 0$, that we could obtain?[13]

The integral in the denominator of (6) is a weighted average of the likelihood

$$\frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(b-\hat{\beta})^2}{2s^2}\right) \tag{8}$$

[10] See for example [51], page 127, and [43], pages 22–25.

[11] This view has now entered financial economics; Harvey declares in his presidential address that, "If you are rational, you are a Bayesian." ([26], page 18)

[12] Etz and Wagenmakers [12] discuss the history of the name. It has been popular at least since 1995, when Robert Kass and Adrian Raftery used "Bayes factors" as the title of an article in the *Journal of the American Statistical Association* [33].

[13] As Harvey notes ([26], page 21), attention was already being directed to this minimum Bayes factor by Edwards, Lindman and Savage in 1963 [11].

Table 1: Minimum Bayes factors (MBF) corresponding to various p-values. The second column gives to two significant figures the value of $|\hat{\beta}/s|$ that would produce the given p-value. The fourth column gives (p-value)/(p-value+MBF), which is the prior probability for $\beta = 0$ that would produce a posterior probability equal to the p-value if the MBF were actually the Bayes factor.

p-value	$\lvert\hat{\beta}/s\rvert$	MBF	implied prior
0.05	2.0	0.15	0.25
0.02	2.3	0.067	0.23
0.01	2.6	0.036	0.22
0.005	2.8	0.019	0.20
0.001	3.3	0.0045	0.18
0.0001	3.9	0.00052	0.16
0.00001	4.4	0.000058	0.15
0.000001	4.9	0.0000064	0.14

over different values of b. This average is maximized and hence the Bayes factor is minimized when f puts all its probability on the value of b that maximizes (8), namely the maximum-likelihood estimate $\hat{\beta}$. So the *minimum Bayes factor* is

$$\exp\left(-\frac{\hat{\beta}^2}{2s^2}\right).$$

Table 1 shows the minimum Bayes factors corresponding to conventional p-values ranging from 5% to one in a million. As we see from this table, the relationship between the two numbers is fairly stable in this range; when the p-value is 5%, the MBF is about 3 times as large; when the p-value is one in a million, the MBF is about 6 times as large. So we obtain a posterior probability for $\beta = 0$ equal to the p-value if we set its prior probability to about 20% (14% to 25% for the range of p-values in the table) and put the remaining prior probability on the observed value of $\hat{\beta}$.[14]

The Bayesian is supposed to specify $f(\beta)$ before seeing the observations, when he or she does not yet know the value of the maximum-likelihood estimate $\hat{\beta}$. So there is no reason to suppose that $f(\beta)$ will concentrate its probability on this value. Very possibly it will be more spread out or concentrated elsewhere, and thus the discrepancy between the p-value and the posterior probability will be greater than the already noticeable discrepancy we see for the MBF. So Jeffreys's method of significance testing can be expected in general to be substantially more favorable to the hypothesis being tested than the p-value.

[14] Harvey [26] also considers the Vovk-Sellke Bayes factor (he calls it the SD-MBF), first suggested by Vovk [55] and rediscovered by Sellke, Bayarri and Berger in 2001 [42].

2.3 Why omit a statistically significant variable?

Suppose, to fix ideas, that β is the coefficient of an independent variable x in a multiple regression, where y is the dependent variable. Jeffreys presented his Bayesian significance test as a way of deciding whether x should be included or omitted from the model ([32], Section 5.1). This is not the same question as whether that x has absolutely no effect on y—i.e., that β is exactly zero. We may also be asking whether β is large enough to matter. This is the question of substantive (or scientific or economic or material) significance, as opposed to the statistical significance marked by a small p-value.

The cases where a statistically significant effect (p-value too small) is not substantively significant (β too small) tend to be the same as the cases where Jeffreys's posterior probability for $\beta = 0$ is large in spite of the small p-value. When the number of observations is very large, the standard deviation s of the least-squares estimate $\hat{\beta}$ is very small, and so even a small $\hat{\beta}$ can be many times as large as s and hence achieve a small p-value. When we say that $\hat{\beta}$ is small, we probably mean that x would make no meaningful difference in the determination or prediction of y if its coefficient β were that small—the difference would be substantively insignificant. But if we had initially thought x might make a difference, then this value $\hat{\beta}$ is also small relative to the values of β that we initially considered reasonably possible. So Jeffreys would tell us to choose a prior density f very spread out relative to s, and as in our numerical example, the Bayes factor will be much more favorable to $\beta = 0$ than the p-value.

In the situation just described, where a variable x in a multiple regression appears to be statistically but not substantively significant, Cournotians[15] usually agree that it should be left out of the regression, because the confidence interval $\hat{\beta} \pm 2s$ (this is the set of values of β that would not be rejected by a 5% test) indicates that the coefficient β is too small to make any substantive difference, because introducing variables that can make so little difference adds noise, and because apparent evidence for a small effect can result from small imperfections in the model. So Jeffreys and the Cournotians arrive in the same place by different reasoning.

2.4 Reconciling Bayes with Cournot's principle

Although the concept of the Bayes factor as an alternative to Cournotian significance testing is now popular among philosophically-inclined Bayesian statisticians, it is less popular among applied statisticians who use Bayesian models.

It is easy to question the importance of the Bayes factor from a Bayesian point of view. If we are uncertain about the effect of a variable x, why should we put nonzero probability on its effect being exactly zero, as opposed to being relatively small? Why not instead spread all our probability continuously over the different possible values of its coefficient β? When we do this the Bayes factor is not so interesting. Our prior odds in favor the exact hypothesis $\beta = 0$

[15] I hasten to repeat that I have just now invented this use of "Cournotian". The non-Bayesian statisticians of whom I am writing may well refuse the name.

are zero, and so our posteriori odds will still be zero, no matter how large the likelihood (a.k.a. Bayes factor) in (5). The Bayesian statistician who takes this approach will remove variables from a model when their possible effect is too small to be substantively significant on grounds consistent with the grounds invoked by Cournotian statisticians. Rather than appeal to a confidence interval that indicates that β has a substantively insignificant value, he or she will note a posterior probability for the hypothesis that β has a substantively insignificant value.

Many applied statisticians who use Bayesian methods also find a place for Cournotian testing. The usual account of the difference between Bayesian and Cournotian statisticians begins with the assumption that they agree on a statistical model with unknown parameters; they agree that if the value θ of the parameters were known, the observation y would have a given probability distribution $\mathbf{P}_\theta$. (The objects θ and y may be single numbers, vectors of numbers, or more complicated mathematical objects.) The Bayesian adds probabilities for θ, obtaining a joint probability distribution for (θ, y) and then conditions that joint probability distribution on the observation y to obtain a probability distribution for θ. The Cournotian estimates θ in other ways and may perform Cournotian testing to see whether particular values of θ are plausible and even whether the entire model $\mathbf{P}_\theta$ is plausible. But at this point we can distinguish between thoroughly subjective Bayesians such as de Finetti and Bayesians who take a more Cournotian view of their enterprise:

- The thoroughly subjective Bayesians see the strategy of forming the prior distribution for θ and then the joint probability distribution for (θ, y) as one application of a general principle that all uncertainties should be dealt with by adopting numerical probabilities and that all evidence should be taken into account by conditioning on such probabilities. They may change their probabilities, but they do not step outside the probability model to check them. De Finetti made this point in the 1950s in a discussion with the French mathematician Maurice Fréchet; see [18].

- More Cournotian Bayesians view their joint probability distribution for (θ, y) more in the way Cournotians view the statistical model $\mathbf{P}_\theta$; it is a tool for making predictions and its ability to do so should be checked. The British-American statistician George Box (1919–2103) was known for this attitude [5], and it has been defended by Andrew Gelman, Donald B. Rubin, and others [20, 19].

The attitude of the Cournotian Bayesians is consistent with Cournot's own views, as he thought that probability is initially subjective and acquires an objective status, marked by interpersonal consensus, as it is validated by experience. Similar views were expressed by Cournot's French successors, including Emile Borel (1871–1956) and Paul Lévy (1886–1971), well into the twentieth century; see [50, 46, 3, 36].

3 Game-theoretic Cournotian testing

Mathematical probability began as a theory of betting. The notion of betting still underlies the theory's structure and, in the last analysis, its power. An event is probable if you can bet on it, and the most persuasive way to refute a probabilistic prediction or a probabilistic hypothesis is to bet and win. The modern game-theoretic framework for probability[16] builds on these insights in a way that accommodates both the subjective and the objective aspects of probability.

Cournot's principle becomes game-theoretic as soon as we interpret probabilistic predictions as betting offers. This allows a statistician to test a probabilistic hypothesis by selecting and betting on an event E to which the hypothesis gives a small probability α. If he bets $\$\alpha$ and E happens, he wins \$1: he has discredited the hypothesis by multiplying the money he risked by the large factor $1/\alpha$.[17] We can think of α as the Neyman-Pearson significance level. The bet is merely another way of interpreting Neyman-Pearson "rejection". Rejecting the hypothesis means discrediting it by multiplying one's money by a large factor.

The bet may be merely notional—i.e., imagined. The statistician can easily enough make the bet, as he needs only risk pennies to make his point, but usually the hypothesis will not really be backing its predictions up with money. The essential requirement is that the statistician announce his bet, real or imagined, in advance of seeing the observations.

3.1 General game-theoretic testing

The game-theoretic picture also allows the statistician to test a probabilistic hypothesis in a more general way. He can try to multiply his money by selecting any nonnegative variable T to which the hypothesis assigns positive expected value $\mathbf{E}(T)$, paying $\mathbf{E}(T)$ and receiving in return the realized value t of T. Let us call T the statisticians game-theoretic *test statistic*, and let us call the factor by which he multiplies the money he risks,

$$s(t) := \frac{t}{\mathbf{E}(T)}$$

the *test score* achieved by T. The test score $s(t)$ is a measure of how much the statistician has discredited the hypothesis. When $s(t) > 1$, we can think of its inverse, $1/s(t)$, in the same way as we think of a significance level. If $s(t) = 20$, the statistician has discredited the hypothesis at the significance level 0.05.

The test score is unaffected when T is multiplied by a positive constant, and so we may assume without loss of generality that its expected value $\mathbf{E}(T)$ under the hypothesis $\mathbf{P}$ is unity and thus that the test score is simply T's realized

[16] See my 2001 book with Vladimir Vovk [49] and papers by Vovk, myself, and others posted at `www.probabilityandfinance.com`.

[17] Or he can bet any positive amount $\$C$, losing it if E fails and receiving $\$C/\alpha$ if E happens. No matter what the value of C, he multiplies the money he risks by $1/\alpha$ if E happens.

value t. Supposing for simplicity that $\mathbf{P}$ is discrete with probability density $\mathbf{p}$ and setting $\mathbf{q} := T\mathbf{p}$, we see that

$$1 = \mathbf{E}(T) = \sum_y T(y)\mathbf{p}(y) = \sum_y \mathbf{q}(y),$$

so that $\mathbf{q}$ is a probability density and thus T is a likelihood ratio:

$$T(y) = \frac{\mathbf{q}(y)}{\mathbf{p}(y)}. \tag{9}$$

This is consistent with (4) and hence with Neyman and Pearson's insights into how a test statistic should be selected. But we obtain (9), just as Neyman and Pearson obtained (4), only in the case where the hypothesis being tested fully specifies the probability distribution $\mathbf{P}$ for the outcome y.

On the other hand, the premise that a hypothesis is discredited when we multiply substantially the money we risk betting against it is a basic principle, not a conclusion drawn from other principles. The principle is convincing when we are betting at odds given by the hypothesis and equally convincing when we are betting at less favorable odds. So a test score carries just as much weight when $\mathbf{E}(T)$ is only an upper bound on the expected value of T under a probability distribution that is not fully specified.

Game-theoretic Cournotian testing also extends to settings where a theory (or an individual or some sort of forecasting system) makes successive forecasts that can be interpreted as betting offers. In this case, the statistician can test the theory by making a sequence of gambles; see again [49].

3.2 Offering bets vs. selecting from offered bets

In order to put betting in the framework of modern game theory, we must distinguish between a player who offers bets and a player who decides which of the offered bets to take. So it bears repeating that the game-theoretic interpretation of Cournotian testing identifies the hypothesis being tested as the player who offers bets. The statistician who tests the hypothesis is the player who decides which of the offered bets to take.

Contemporary subjective Bayesianism also emphasizes betting and decision, but it puts the statistician in the role of the player who offers bets. The statistician is supposed to create his own system of probabilities (not merely test someone else's), and once he has created it, to follow it when dealing with the choices nature presents—i.e., to allow nature to bet against him at the odds he has established. As we saw in Section 2.4, this leaves no space or need for Cournot's principle.

3.3 Game-theoretic adjustment of p-values

A p-value does not enjoy the same clear betting interpretation as a Neyman-Pearson significance level. The statistician identifies a test statistic T (or equiv-

Table 2: Test scores and adjusted p-values produced by betting equal amounts of money at the significance levels 0.05, 0.01, and 0.001. All numbers are to two significant figures.

unadjusted p-value	test score	adjusted p-value
$p > 0.05$	0	no evidence
$0.01 < p \leq 0.05$	6.7	0.15
$0.001 < p \leq 0.01$	40	0.025
$p \leq 0.001$	370	0.0027

alently a p-variable P) in advance, but does not fix in advance a level of T (or P) on which to bet.

Why do applied statisticians sometimes report a p-value instead of fixing a significance level α and a rejection region E in advance of seeing the data and reporting only whether E happened? As we noted in Section 1.3, they often do so because they want to be able to report even stronger evidence if it comes their way. Even if they would be content to find evidence at the 5% level, they do not want to forgo, if it appears, the stronger evidence represented by a much smaller p-value.

The game-theoretic point of view can accommodate the desire to recognize evidence at different levels of strength if we specify in advance all the levels we want to recognize. Suppose, for example, that we want to recognize evidence at significance levels 0.05, 0.01 and 0.001. So we bet a dollar on each: a dollar on $P \leq 0.05$, a dollar on $P \leq 0.01$, and a dollar on $P \leq 0.001$. What will happen?

- If the p-value p comes out greater than 0.05, we lose the \$3 we risked. The test score is zero. We have not discredited the hypothesis.
- If $0.01 < p \leq 0.05$, we win \$20 from the first bet but lose the other two bets. We have turned \$3 into \$20, achieving a test score of $20/3 \approx 6.7$, corresponding to rejection at the significance level $3/20 = 0.15$.
- If $0.001 < p \leq 0.01$, then we win \$20 from the first bet and \$100 from the second. This produces a test score of $120/3 = 40$, corresponding to the significance level $3/120 = 0.025$.
- If $p \leq 0.001$, then we win \$20 from the first bet, \$100 from the second, and \$1000 from the third. This produces the test score $1120/3 \approx 373$, corresponding to the significance level $3/1120 \approx 0.0027$.

These results can be thought of as a rule for adjusting the observed p-value, as laid out in Table 2.

There are, of course, many other strategies for betting on a p-variable—many different ways of spreading our money over bets against $P \leq p$ for different p, and each leads to a different rule for adjusting the p-value. It might be useful to have a standard rule of thumb that could be used by consumers of research

when that research announces a p-value rather than rejection or failure to reject at a fixed significance level.

3.3.1 The truncated square-root rule

One simple and easily remembered rule of thumb arises when we spread our money, say \$1, continuously over all the possible p-values from 0 to 1 according to the probability density

$$f(q) = \frac{1}{2\sqrt{q}}. \tag{10}$$

This means dividing $[0, 1]$ into increments of length dq and dividing our dollar into corresponding increments following this density, the amount $\$f(q)dq$ being assigned to the increment at q. We bet this $\$f(q)dq$ on the p-value being q or less, winning $\$\frac{1}{q}f(q)dq$ if we win the bet. Given the actual p-value p, we win altogether

$$\int_p^1 \frac{1}{q} f(q)dq = \frac{1}{2}\int_p^1 q^{-\frac{3}{2}} dq = -q^{-\frac{1}{2}}\Big|_p^1 = \frac{1}{\sqrt{p}} - 1$$

dollars. Turning \$1 into this amount corresponds to winning at significance level

$$\frac{1}{\frac{1}{\sqrt{p}} - 1} = \frac{\sqrt{p}}{1 - \sqrt{p}}, \tag{11}$$

which is very close to $\sqrt{p}$ when p is 5% or less. This is a very simple rule of thumb: adjust a p-value by taking its square root. The adjustment is severe; to get significance at the conventional 0.05 level, you need an unadjusted p-value less than 0.0025; to get significance at the one in a thousand level you need an unadjusted p-value of one in a million or less.

The density (10) puts more than 2/3 of our money on p-values greater than 0.10, and this may be unreasonable, as we would not test the hypothesis by betting on such large p-values. We might also want to consider putting more of our money on p-values closer to 0.05 and less on p-values that are much smaller. This suggests that we compare the following strategies:

Rule 1 (square-root rule). Spread our money over the interval $[0, 1]$ of p-values using the density $f(q)$ given by (10). As we have just seen, produces the payoff g_1 given by

$$g_1(p) := \frac{1}{\sqrt{p}} - 1 \tag{12}$$

and hence leads us to replace a p-value p by $\sqrt{p}/(1 - \sqrt{p})$.

Rule 2 (truncated square-root rule). Spread our money over the interval $[0, 0.10]$ of p-values using the density $f(q)$ truncated to this interval. By a similar calculation, this produces the payoff g_2 given by

$$g_2(p) := \begin{cases} \sqrt{\frac{10}{p}} - 10 & \text{if } 0 \le p \le 0.10 \\ 0 & \text{if } 0.10 < p \le 1 \end{cases} \tag{13}$$

Table 3: Three rules for adjusting p-values. Numbers are given to the nearest percentage or to one significant figure. The last column is a recommended rule of thumb. As the numbers in boldface show, it approximates Rule 2 well when the p-value is less than 1%.

p-value	Rule 1	Rule 2	Rule 3	$\sqrt{\text{p-value}}/3$
0.07	0.40	0.51	0.28	0.09
0.05	0.25	0.24	0.14	0.07
0.02	0.16	0.08	0.06	0.05
0.01	0.11	**0.05**	0.04	**0.03**
0.005	0.08	**0.03**	0.03	**0.02**
0.001	0.03	**0.01**	0.02	**0.01**
0.0001	0.01	**0.003**	0.01	**0.003**
0.00001	0.003	**0.001**	0.01	**0.001**
0.000001	0.001	**0.0003**	0.009	**0.0003**

and hence leads us to replace a p-value p in the interval $[0, 0.10]$ by $\sqrt{p/10}/(1-\sqrt{10p})$ and to ignore a p-value exceeding 0.10.

Rule 3 (truncated logarithmic rule). Spread our money uniformly over the interval $[0, 0.10]$. This produces the payoff g_3 given by

$$g_3(p) := \begin{cases} 10 \ln\left(\frac{0.10}{p}\right) & \text{if } 0 \le p \le 0.10 \\ 0 & \text{if } 0.10 < p \le 1 \end{cases} \tag{14}$$

and hence leads us to replace a p-value in the interval $[0, 0.10]$ by $1/(10\ln(0.10/p))$ and to ignore a p-value exceeding 0.10.

Numerical values for these three adjustments are shown in Table 3.

Which of the rules in Table 3 provides the most reasonable adjustments? I recommend Rule 2, the truncated square-root rule. It is a little less severe than Rule 1 (the square-root rule) for p-values less than 5%. As the last column of the table shows, it is very well approximated for p-values less than 1% by a simple rule of thumb: take the square root and divide by 3. Rule 3, which results from putting more of our money on p-values near 5%, is slightly less severe on those p-values but perhaps unreasonably severe on more extreme p-values.

Comparing Tables 1 and 3, we see that game-theoretic adjustment of small p-values is generally much more severe than the adjustment suggested by the minimum Bayes factor. The game-theoretic adjustment comes into play, however, only when the statistician fails to fix a significance level in advance. Were the statistician to state in advance, when choosing his test statistic, that he is looking for a one in a million deviation and will reject the hypothesis only if his p-value is this extreme, then an actual one in a million deviation can be taken at face value.

Statisticians often think of statistical significance in terms of a test statistic $|Z|$, where Z has a standard normal distribution. The p-value is 0.05 when $|Z| = 1.96$. How much do we need to raise the cutoff 1.96 in order to obtain a p-value with adjusted value 0.05 according to the truncated square-root rule? Solving

$$\frac{\sqrt{p/10}}{1 - \sqrt{10p}} = 0.05$$

for p, we obtain $p = 1/90 \approx 0.011$. This corresponds to the cutoff 2.54.

In [27], Harvey, Liu, and Zhu recommend that financial economists raise their cutoff for statistical significance from 2 to 3 in order to account for multiple testing. The analysis here suggests that, on this scale, about half this increase (the part from $2 \approx 1.96$ to 2.54) is needed to account for the way p-values exaggerate the significance level, before we even think about multiple testing. We can also ask how the truncated square-root rule adjusts the p-value corresponding to the cutoff 3, namely 0.0027. Its adjusted p-value is 0.02, which is less than 0.05 but perhaps not enough less to account for much multiple testing.

3.3.2 Gambling on p-variables more abstractly

As we have just seen, the probabilities associated with a p-variable by the inequality (2) can be interpreted as betting offers made by the hypothesis the p-variable P is testing, and we can exploit these betting offers to effectively buy certain payoffs, such as the payoffs g_1, g_2, and g_3 given by (12), (13), and (14), respectively. As we saw, the price for each of these payoffs is 1.

In fact, (2) authorizes us to buy any function g of p that is nonnegative and nonincreasing at the price $\int_0^1 g(q)dq$. It follows that $1/g(p)$ is a legitimate rule for adjusting p-values for any $g : [0, 1] \to \mathbb{R}$ that is nonnegative and nonincreasing and integrates to 1. See [48, 9].

4 Game-theoretic multiple testing

As Cournot lucidly explained, it is difficult to impossible to judge whether an extreme deviation is attributable to chance when a statistician discovers it by searching across different ways of analyzing a body of data. A statistician who remembers everything he or she has tried may be able to judge how much the significance of the striking result finally found should be discounted. Others are in the dark. Nowadays we call this "publication bias": tests that do not produce statistically significant results are not published [30, 57].

In today's competitive research communities, including the community of empirical financial economics as Harvey and Ohlson describe it, the evaluation of statistical results is even more difficult than in Cournot's time. The quest for striking results is pursued by numerous individuals or teams who do not observe each other's searches, and so no one really knows the extent of the search. As consumers of such research, we must nevertheless evaluate its significance as

best we can. This may involve modeling the extent of the search, even if the modeling only formalizes rather than mitigating our uncertainty.

Accounting for multiple testing is one aspect of the larger enterprise of combining multiple studies or sources of evidence. In recent decades, this enterprise has often been called *meta-analysis* in science and *data fusion* in various fields of technology.[18] The analyses by Harvey, Liu, and Zhu in [27] are impressive and informative examples of meta-analysis. I will not attempt a general discussion of meta-analysis or data fusion here, but I will look at how Cournotian testing, understood game-theoretically as explained in the preceding section, can be used to combine or otherwise evaluate multiple p-values.

As consumers of research, we can evaluate p-values $p_1, \ldots, p_n$ game-theoretically without actually betting. As when we adjust a single p-value, the bet can be imaginary. But the evaluation will be convincing (to ourselves and others) only if we have decided how to bet (or imagine betting) in advance of seeing the actual p-values. In practice, because we are likely to see p-values before we think about combining them, this means we need general policies for how to bet on multiple p-values. Ideally, these policies should lead to rules as simple as the rule just proposed for adjusting a single p-value, the truncated square-root rule.

We can confront p-values resulting from multiple tests with two different questions:

1. To what extent does the evidence as a whole refute the hypothesis being tested? Or, more generally, in the case where different hypotheses were being tested, to what extent does the evidence indicate that at least one of the hypotheses is false?

2. If different hypotheses were being tested, to what extent does the evidence as a whole refute the particular hypothesis tested by the p-value that came out the smallest? Or some group of hypotheses for which tests produced relatively low p-values?

I will consider these two questions in turn.

4.1 Testing the same hypothesis multiple times

Suppose a hypothesis is tested n times, by one or multiple researchers. The results are reported as p-values: $p_1, \ldots, p_n$. How might a consumer of this research evaluate the overall evidence against the hypothesis?

4.1.1 One p-value from many

A crude but frequently used way of obtaining a single p-value from p-values $p_1, \ldots, p_n$ is to take their minimum and multiply by n. To see that this produces

[18] A number of authors, including Hedges and Olkin [28], O'Rourke [41], and Borenstein et al. [4], have discussed the history of meta-analysis before the invention of the name. The term *data fusion* encompasses Bayesian and Dempster-Shafer methods [43].

a p-value, note that because the probability of the union of a finite number of events is always less than or equal to the sum of their probabilities,

$$\mathbf{P}\left(\min_{1\le i\le n} P_i \le p\right) = \mathbf{P}\left(\bigcup_{i=1}^{n}\{P_i \le p\}\right) \le \sum_{i=1}^{n}\mathbf{P}\left(P_i \le p\right) \le np,$$

where P_i is the p-variable that produced the p-value p_i. It follows that

$$\mathbf{P}\left(n\min_{1\le i\le n} P_i \le p\right) \le p.$$

In other words, $n\min_{1\le i\le n} P_i$ is a p-variable and hence its value, say

$$p := n\min_{1\le i\le n} p_i, \tag{15}$$

is a p-value. Let us call $n\min_{1\le i\le n} P_i$ *Bonferroni's p-variable* and (15) *Bonferroni's p-value*.[19]

Another equally general way of obtaining a single p-value from p-values $p_1,\dots,p_n$ is to take twice their average:

$$p := 2\frac{\sum_{i=1}^{n} p_i}{n}. \tag{16}$$

See [56] for a proof that this is a p-value.

From a game-theoretic point of view, we can treat the p-value (15) like any other. If we come to the data with the policy of testing at a particular significance level α, then we say that the hypothesis is discredited at level α if $p \le \alpha$. If instead we come to the data with the policy of betting on the p-value using the truncated square-root rule, then we can say that the hypothesis is discredited at level $\sqrt{p/10}/(1-\sqrt{10p})$ if $p \le 0.10$.

If the n p-variables $P_1,\dots,P_n$ are based on different data sets, then it may be reasonable to assume that they are independent, and in this case the product $P_1\cdots P_n$ may be a more efficient test statistic. If the test statistics have continuous probability distributions under the hypotheses, so that the P_i are uniformly distributed on $[0,1]$ (see (3) in Section 1.2.1), then as R. A. Fisher pointed out in [16], their independence implies that $-\ln(P_1\cdots P_n)$ has a chi-square distribution with n degrees of freedom. We can take this as our test-statistic and treat its p-value game-theoretically. This method is usually more efficient than betting on the Bonferroni p-variable, and it will be valid even if the test statistics do not have continuous distributions, provided they are independent. But the assumption of independence is not generally applicable to financial economics, where different tests are often based on overlapping data.

[19] The English logician and philosopher George Boole (1815–1865) is often cited for having noticed that the probability of a disjunction does not exceed the sum of the probabilities of the disjuncts. But the Italian mathematician Carlo Bonferroni (1892–1960), who adduced more general inclusion/exclusion inequalities, is usually cited when this inequality is used in significance testing.

When we cannot assume that the P_i are independent, is there some other function of the p-values that is more powerful than Bonferroni's p-variable—i.e., that consistently discredits the hypothesis more strongly when the p_i are small? Generally not. In principle, we could replace the minimum by any other function of the p_i, say $T(p_1, \ldots, p_n)$, that is small when the p_i are small, and define a p-value by

$$\{\text{p-value from observing } T = t\} := \sup\{\mathbf{P}(T(P_1, \ldots, P_n) \leq t)\},$$

where the supremum is over all joint probability distributions $\mathbf{P}$ for $P_1, \ldots, P_n$ that satisfy $\mathbf{P}(P_i \leq p) \leq p$ for $p \in [0, 1]$ and $i = 1, \ldots, n$. But in general this supremum will be difficult to calculate and often be too large to be interesting. Experts on meta-analysis usually prefer, when possible, to leave aside the idea of combining p-values and try instead to combine the test statistics that produced them [28].

4.1.2 Multiple bets

The game-theoretic picture gives us another option. Instead of considering a test statistic $T(P_1, \ldots, P_n)$ and betting on its p-variable, we can bet on each of the p-variables $P_1, \ldots, P_n$ and combine the bets.

Here is one policy of this type:

1. Choose a significance level α that we will use whenever we want to combine multiple p-values for the same hypothesis.
2. When the number of p-values we are combining is n, bet $\$\frac{1}{n}$ on $P_i \leq \alpha$ for $i = 1, \ldots, n$.

This multiplies the dollar we risk by

$$\frac{N}{n\alpha}, \text{ where } N \text{ is the number of } P_i \text{ satisfying } P_i \leq \alpha. \tag{17}$$

The expected value of N under the hypothesis being $n\alpha$ or less, we expect some rejections but do not expect the test score (17) to be much greater than 1.

An alternative policy would fix a nominal level α that we will use whenever we combine p-values but vary the level for testing each p-variable with n, perhaps taking it to be α/n or $\alpha/\sqrt{n}$. If the policy is to spread one dollar equally over the n p-variables and to test each at level α/n, then we will multiply the dollar we risk by

$$\frac{N}{\alpha}, \text{ where } N \text{ is the number of } P_i \text{ satisfying } nP_i \leq \alpha. \tag{18}$$

The expected value of N under the hypothesis now being α or less, we are less likely to see any individual p-values rejecting the hypothesis. As with the test score (17), we do not expect the text score (18) to be much greater than 1.

The test score (18) can be compared with the test score obtained by testing the Bonferroni p-variable $n \min_{1 \leq i \leq n} P_i$ at level α. Risking one dollar on

$n \min_{1 \le i \le n} P_i \le \alpha$ returns $\$N/\alpha$ when $N = 0$ or 1 but only $\$1/\alpha$ when $N \ge 2$. As this is less than or equal to (18), we can say that the Bonferroni p-variable is unnecessarily conservative from the game-theoretic point of view.

Another of the infinitely many possible policies for combining p-values is to bet on each of the n p-variables using the truncated square-root rule. Again distributing one dollar uniformly over the n p-variables, this produces the test score

$$\frac{1}{n} \sum_{\substack{1 \le i \le n \\ p_i \le 0.10}} \left(\sqrt{\frac{10}{p_i}} - 10 \right). \tag{19}$$

4.2 Testing multiple hypotheses

Now we come to our second question. In the case where p-values $p_1, \ldots p_n$ result from tests of different hypotheses (or perhaps different aspects of a hypothesis) to what extent does the success of our betting discredit particular hypotheses (or particular aspects)? Can we lay the whole of the discredit at the door of the particular hypothesis that produced the smallest p-value? Surely not. But can we not say something about how much this particular hypothesis is discredited?

A positive response seems to require an extension of the basic principle of game-theoretic testing, the principle that success in betting using odds given by a hypothesis is evidence against the hypothesis. The extended principle says that when we spread our money over bets against different hypotheses, we can count the dollars returned by bets against particular hypotheses as evidence against them, but that we must take account of how much money we risked altogether in evaluating the weight of this evidence. If we accept this extended principle, we can use the following definition:

> **Extended game-theoretic testing principle.** Suppose we distribute one dollar over a set H of probabilistic hypotheses, betting the amount assigned to hypothesis $h \in H$ at odds given by h. For each subset A of H, let T_A be the total winnings in dollars of the bets for hypotheses in A. We call T_A the *test score* for A.

When using this concept of a test score, we think of the observed value t of T_A as a measure of evidence against the assertion that *all the hypotheses in A are true*, and we think of this assertion as being discredited at significance level $1/t$.

Suppose for example, that H contains n hypotheses, and each hypothesis $h \in H$ has been tested, producing a p-value p_h. Then if we spread one dollar over the n hypotheses and apply the truncated square-root rule to each p-variable, then for each we obtain a test score against each hypothesis $h \in H$ of

$$T_{\{h\}} = \begin{cases} \frac{1}{n} \left(\sqrt{\frac{10}{p_h}} - 10 \right) & \text{if } p_h \le 0.10 \\ 0 & \text{if } p_h > 0.10 \end{cases}$$

Notice the discounting due to the fact that we had bet against $n - 1$ other hypotheses in addition to h before focusing on h; had we decided before seeing

the p-values to test only h using its p-value p_h, then our test score against it would have been n times as large. Notice also that T_H, the test score against the assertion that all n hypotheses are true, is (19).

The extended game-theoretic testing principle can also be used when the hypothesis space H is continuous, and it can be used to produce systems of game-theoretic scored intervals, analogous to the confidence intervals and confidence distributions derived from the Neyman-Pearson theory [58].

5 Acknowledgements

This working paper began as a memorandum for the students in my doctoral course on evidence in Rutgers University's Department of Accounting and Information Systems, who were reading Campbell Harvey's recent meta-analyses of the study of factors affecting expected stock returns [27, 26]. I would like to thank the students in course for their comments and suggestions. I have also benefited from conversations with Ren-Raw Chen, Suresh Govindaraj, Dan Palmon, Bharat Sarath, Steve Stigler, and Vladimir Vovk.

References

For GTP Working Papers, see the web site `http://probabilityandfinance.com`.

[1] John Aldrich. The statistical education of Harold Jeffreys. *International Statistical Review*, 73(3):289–308, 2005.

[2] John Aldrich. The origins of modern statistics: The English statistical school. In Alan Hájek and Christopher Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*, pages 112–129. Oxford, New York, 2016.

[3] Émile Borel. *Valeur pratique et philosophie des probabilités*. Gauthier-Villars, Paris, 1939.

[4] Micahel Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. *Introduction to Meta-Analysis*. Wiley, New York, 2009.

[5] George E. P. Box. Comment on "The unity and diversity of probability" by Glenn Shafer. *Statistical Science*, 5:448–449, 1990.

[6] Antoine Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [7].

[7] Antoine Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The books in this collection are numbered I through XI, but VI and XI are double volumes, so that there are 13 volumes altogther.

[8] Andrew W. Dale. *A History of Inverse Probability from Thomas Bayes to Karl Pearson*. Springer, New York, second edition, 1999.

[9] A. P. Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability, 2010. GTP Working Paper 34. A version appeared in *Statistics and Probability Letters* 81:157–162, 2011.

[10] A. W. F. Edwards. The history of likelihood. *International Statistical Review*, 42(1):9–15, 1974.

[11] Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychologists. *Psychological Review*, 70:193–242, 1963.

[12] Alexander Etz and Eric-Jan Wagenmakers. J.B.S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, page ??, 2017.

[13] Stephen E. Fienberg. When did Bayesian inference become Bayesian? *Bayesian Analysis*, 1(1):1–40, 2006.

[14] Ronald A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921.

[15] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A)*, 222:309–368, 1922.

[16] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925. The thirteenth edition appeared in 1958.

[17] Ronald A. Fisher. Combining independent tests of signficance. *The American Statistician*, 2(5):30, 1948.

[18] Maurice Fréchet. *Les mathématiques et le concret*. Presses Universitaires de France, Paris, 1955.

[19] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, third edition, 2013.

[20] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013.

[21] Steven N. Goodman. p values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137(5):485–495, 1993.

[22] Steven N. Goodman. Of P-values and Bayes: A modest proposal. *Epidemiology*, 12:295–297, 2001.

[23] Prakash Gorroochurn. *Classic Topics on the History of Modern Mathematical Statistics from Laplace to More Recent Times*. Wiley, New York, 2016.

[24] Trygve Haavelmo. The probability approach to econometrics. *Econometrica*, 12(Supplement):1–115, 1944.

[25] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York, 1998.

[26] Campbell R. Harvey. The scientific outlook in financial economics. Technical report, Duke University, 2017. http://dx.doi.org/10.2139/ssrn.2893930.

[27] Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

[28] Larry V. Hedges and Ingram Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, 1985.

[29] David Howie. *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge University Press, Cambridge, 2002.

[30] John P. A. Ioannidis. Why most research findings are false. *PLOS Medicine*, 2(8):696–701, 2005.

[31] Harold Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophy Society*, 31:203–222, 1935.

[32] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1939. Second edition 1948, third 1961.

[33] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[34] John Maynard Keynes. *A Treatise on Probability*. Macmillan, London, 1921.

[35] Pierre Simon de Laplace. *Théorie analytique des probabilités*. Courcier, Paris, first edition, 1812. This monumental work had later editions in 1814 and 1820. The third edition was reprinted in Volume 7 of Laplace's *Œuvres complètes*.

[36] Paul Lévy. *Calcul de probabilités*. Gauthier-Villars, Paris, 1925.

[37] Donald A. MacKenzie. *Statistics in Britain 1865–1930*. Edinburgh University Press, Edinburgh, 1981.

[38] Denton E. Morrison and Ramon E. Henkel, editors. *The Significance Test Controversy–A Reader*. Aldine, Chicago, 1970.

[39] Jerzy Neyman. Indeterminism in science and new demands on statisticians. *Journal of the American Statistical Association*, 55:625–639, 1960.

[40] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (A)*, 36:289–337, 1933.

[41] Keith O'Rourke. A historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12):579–582, 2007.

[42] Thomas Sellke, J. J. Bayarri, and James O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[43] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[44] Glenn Shafer. Bayes's two arguments for the rule of conditioning. *Annals of Statistics*, 10:1075–1089, 1982.

[45] Glenn Shafer. Lindley's paradox. *Journal of the American Statistical Association*, 77:325–334, 1982.

[46] Glenn Shafer. From Cournot's principle to market efficiency, March 2006. GTP Working Paper 15. Published as Chapter 4 of: Jean-Philippe Touffut, editor, *Augustin Cournot: Modelling Economics*. Edward Elgar, Cheltenham, UK, 2007.

[47] Glenn Shafer. Cournot in English, April 2017. GTP Working Paper 48.

[48] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors, and p-values, 2010. GTP Working Paper 33. A version with the color missing from the figures appeared in *Statistical Science* 26:84–101, 2011.

[49] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[50] Glenn Shafer and Vladimir Vovk. The origins and legacy of Kolmogorov's *Grundbegriffe*, April 2013. GTP Working Paper 4. Abridged version published as "The sources of Kolmogorov's *Grundbegriffe*" in *Statistical Science* 21:70–98, 2006.

[51] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA, 1986.

[52] Stephen M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986.

[53] Stephen M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999.

[54] Stephen M. Stigler. *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge, MA, 2016.

[55] Vladimir Vovk. A logic of probability, with applications to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society. Series B*, 55(2):317–351, 1993.

[56] Vladimir Vovk. Combining p-values via averaging. `https://arxiv.org/abs/1212.4966v2`, 2012.

[57] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[58] Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review*, 81(1):3–77, 2013.

[59] George Udny Yule. *An Introduction to the Theory of Statistics*. Griffin, London, first edition, 1911.

[60] Stephen T. Ziliak and Deirdre N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Cost Us Jobs, Justice, and Lives*. University of Michigan Press, Ann Arbor, 2008.