Common Fallacies in Applying Hypothesis Testing

X. Rong Li
Department of Electrical Engineering
University of New Orleans, New Orleans, LA 70148
504-280-7416 (phone), 504-280-3950 (fax), xli@uno.edu

Xiao-Bai Li
College of Management
University of Massachusetts Lowell
Lowell, MA 01854, xiaobai_li@uml.edu

Abstract—Although the theory of hypothesis testing is well developed and has a long history of application, practical application of hypothesis testing is plagued with fallacies, confusions, misconceptions, misuses, and abuses. This paper addresses four of the most widespread ones, particularly in statistical processing of signals, data, and information in uncertainty. They concern the decision on a single hypothesis, the assignment of the null hypothesis in the Neyman-Pearson framework, the confusion between two classes of significance tests, and the interpretation of a hypothesis not rejected. We articulate the principle underlying the tests, explain and analyze where the fallacies and confusions arise from, and present convincing arguments, clear conclusions, and succinct guidelines for practice, along with detailed, representative examples.

Keywords: Hypothesis testing, fallacy, confusion, application

I. INTRODUCTION

Statistical hypothesis testing, along with estimation, forms the core of statistical inference. It has a vast domain of application covering many fields of the applied world. Engineering is no exception. The theory of hypothesis testing underpins all processing of signals, data, images, and information in uncertainty. Its typical application in engineering includes signal detection, target classification, pattern recognition, fault identification, data association, and model selection.

Although the theory of hypothesis testing is well developed and has a long history of application, as with many other components of statistics, practical application of hypothesis testing is plagued with fallacies, confusions, misconceptions, misuses, abuses, and so on.

In this paper, we point out several fallacies, confusions, and misuses in applying hypothesis testing that are particularly widespread in statistical processing of signals, data, images, and information in uncertainty. They include those concerning the decision on a single hypothesis, the assignment of the null hypothesis in the Neyman-Pearson framework, the confusion between two classes of significance tests, and the interpretation of a hypothesis not rejected. To dispel various fallacies, not only do we articulate the principles and make detailed explanation and analysis, along with concrete examples, but we also present clear conclusions and guidelines for practice.

It is well known that part of the theory of statistical hypothesis testing is controversial. The first thing that comes to mind in this regard is the lasting debate between the Bayesians and the frequentists, which has been ongoing for more than half a century. It is not our intention to discuss this debate or any controversial theoretical issues in this paper. Rather, our emphasis is on the application side—we address inconsistency of some widespread practice with the established theory.

While the value of this paper to an expert on hypothesis testing is limited, the points it raises are important for practitioners.

II. FALLACIES IN APPLYING SINGLE-HYPOTHESIS TESTS

A. Single-Hypothesis Testing and Its Underlying Principle

Single-hypothesis testing. The problem of testing a single hypothesis is deceptively simple: Given a single hypothesis and a set of data, decide on the fate of the hypothesis. One source of confusion and fallacy among many practitioners is a narrow interpretation of the “fate” as either rejection or acceptance of the hypothesis or as a judgment on whether the hypothesis is true or false. This is addressed in Sec. V.

A single hypothesis is often formulated as

$$H : \theta = \theta_0$$

or more precisely

$$H : z \sim F(z|\theta = \theta_0)$$

where $z \sim F$ stands for “data $z$ has the distribution $F$,” and $F(z|\theta = \theta_0)$, or simply $F(z|\theta_0)$, is used to mean either $F(z|\theta_0)$ (for Bayesians) or $F(z; \theta)|\theta = \theta_0$ (for non-Bayesians). For a single-hypothesis testing problem, the parameter (or more precisely, distribution) when $H$ is not true is not specified at all.

Cournot’s principle. Developed mainly by R. A. Fisher, along with K. Pearson and others, in the early 20th century [8], single-hypothesis testing is based on the following principle of small-probability events. Following Glenn Shafer [16], we refer to it as Cournot’s principle: The occurrence of an event of an extremely small probability on a single trial has a profound implication—the assumptions based on which the probability is calculated is incorrect and should be abandoned. This is one of the most fundamental principles of science—it is based on this principle that a theory is confirmed and accepted (at least initially), no matter how bizarre it is. A good example is the Big-Bang theory in cosmology, which is much more bizarre than science fictions. Amazingly, it was accepted by most physicists in the same field right after only a few of its bold predictions were verified. Why was this the case? The answer lies in Cournot’s principle: Assuming the

---

1Research supported in part by Navy through Planning Systems Contract N68335-05-C-0382 and Project 863 through grant 2006AA01Z126.
theory is incorrect, a verification of any of its bold predictions would be extremely improbable; given such verifications, the underlying assumption that it is incorrect should be abandoned.

B. Analysis of a Common Fallacy

A common fallacy in application of single-hypothesis testing stems from ignorance or oversight of its underlying principle. There are two seriously flawed yet widely used versions of the single-hypothesis test (see, e.g., [2], [3]):

**Version 1:**

- Obtain an estimate \( \hat{\theta} \) of \( \theta \) from data.
- Find a (shortest) interval \( (\theta_l, \theta_u) \) such that \( P\{\theta \in (\theta_l, \theta_u) | H\} = 1 - \alpha \) for a small \( \alpha \) (e.g., 0.05).
- Reject \( H \) if \( \hat{\theta} \notin (\theta_l, \theta_u) \); accept \( H \) if \( \hat{\theta} \in (\theta_l, \theta_u) \).

It is applied most often concerning significance of a chi-square or Gaussian \( \hat{\theta} \).

**Version 2:**

- Find an interval estimate \( (\hat{\theta}_l, \hat{\theta}_u) \) from the data \( z \) such that \( P\{(\hat{\theta}_l, \hat{\theta}_u) \ni \theta | H\} = 1 - \alpha \) for a small \( \alpha \) (e.g., 0.05), where \( P\{(\hat{\theta}_l, \hat{\theta}_u) \ni \theta \} \) is the probability that the random interval \( (\hat{\theta}_l, \hat{\theta}_u) \) covers \( \theta \).
- Reject \( H \) if \( (\hat{\theta}_l, \hat{\theta}_u) \notin \theta_0 \); accept \( H \) if \( (\hat{\theta}_l, \hat{\theta}_u) \ni \theta_0 \).

Here, \( \alpha \) is the significance and \( (1 - \alpha) \) the confidence of the test; \( P\{\theta \in (\theta_l, \theta_u)\} \) and \( P\{(\theta_l, \theta_u) \ni \theta \} \) are equivalent numerically but have distinct interpretations:

- \( P\{\theta \in (\theta_l, \theta_u)\} \) is the probability of \( \theta \) being inside \( (\theta_l, \theta_u) \), which implies that \( \theta \) is random but \( (\theta_l, \theta_u) \) is not.
- \( P\{(\theta_l, \theta_u) \ni \theta \} \) is the probability of the interval \( (\theta_l, \theta_u) \) capturing \( \theta \), which implies that \( (\theta_l, \theta_u) \) is random but \( \theta \) is not.

We now present a quantitative analysis. Let

\[
B = \begin{cases} 
\{\hat{\theta} \in (\theta_l, \theta_u)\}, & \text{version 1} \\
\{(\theta_l, \theta_u) \ni \theta_0\}, & \text{version 2} 
\end{cases}
\]

\[
\tilde{B} = \begin{cases} 
\{\hat{\theta} \notin (\theta_l, \theta_u)\}, & \text{version 1} \\
\{(\theta_l, \theta_u) \ni \theta_0\}, & \text{version 2} 
\end{cases}
\]

\[
\alpha = P\{\tilde{B} | H\} \tag{1}
\]

**Deductive reasoning for rejecting \( H \) if \( \tilde{B} \) occurs.** Since \( P\{\tilde{B} | H\} = 1 - \alpha \), in the limit as \( \alpha \to 0 \), we have \( P\{\tilde{B} | H\} \to 1 \) and thus by deductive reasoning, if \( \tilde{B} \) then \( H \), where \( H \) is the hypothesis that \( H \) is false. This provides a justification for the principle of small-probability events. As such, if \( B \) occurs for an extremely small \( \alpha \) we should reject \( H \).

**Basic assumptions.** The selection of \( (\theta_l, \theta_u) \) or \( (\hat{\theta}_l, \hat{\theta}_u) \) is such that the distribution under \( H \) is more concentrated (has a smaller tail probability) than that under \( \tilde{H} \) in the following sense:

A1. For a very small \( \alpha \), \( \tilde{B} \) is less probable under \( H \) than under \( \tilde{H} \), that is, \( P\{\tilde{B} | H\} < P\{\tilde{B} | H\} \); and further,

A2. \( P\{\tilde{B} | H\} / P\{\tilde{B} | H\} \to 0 \) as \( \alpha \to 0 \).

Assuming \( P\{\tilde{H}\} \neq 0 \), Assumptions A1 and A2 are equivalent to the following assumptions, respectively:

A1’. For a very small \( \alpha \), \( \tilde{B} \) is less probable under \( H \) than under no hypothesis, that is, \( P\{\tilde{B} | H\} < P\{\tilde{B}\} \) (or equivalently, \( P\{B\} < P\{\tilde{B} | H\}\)); and further,

A2’. \( P\{B | H\} / P\{\tilde{B}\} \to 0 \) as \( \alpha \to 0 \); that is

\[
\lim_{\alpha \to 0} \frac{\alpha}{P\{\tilde{B}\}} = 0 \tag{2}
\]

The equivalence of A1 and A1’ follows directly from

\[
P\{\tilde{B}\} = P\{\tilde{B} | H\}P\{H\} + P\{\tilde{B} | \tilde{H}\}P\{\tilde{H}\}
\]

and the equivalence of A2 and A2’ follows directly from

\[
\frac{P\{B | H\}}{P\{B\}} = \frac{P\{\tilde{B} | H\}P\{H\} + P\{B | H\}P\{H\}}{P\{\tilde{B} | H\}P\{H\} + P\{\tilde{H}\}P\{\tilde{H}\}}
\]

**Assumption justification.** These assumptions are needed in order to reject \( H \) when \( \tilde{B} \) occurs for a very small \( \alpha \). Suppose a very small \( \alpha \) is used. If \( P\{\tilde{B} | H\} \neq P\{\tilde{B}\} \), or equivalently, \( P\{H | \tilde{B}\} \neq P\{H\} \), since

\[
P\{H | \tilde{B}\} = \frac{P\{\tilde{B} | H\}P\{H\}}{P\{B\}}
\]

then by probabilistic reasoning \( \tilde{B} \) does not refute \( H \) at all, let alone to reject \( H \) when \( \tilde{B} \) occurs. So, Assumption A1’ (or equivalently, A1) is absolutely necessary in order to reject \( H \) when \( \tilde{B} \) occurs for a very small \( \alpha \). If for a very small \( \alpha \), \( P\{B | H\} \ll P\{B\} \) (or equivalently, \( P\{H | \tilde{B}\} \ll P\{H\} \)) is not true, then there is no strong reason to reject \( H \) when \( \tilde{B} \) occurs. So, Assumption A2 is needed to reject \( H \) when \( \tilde{B} \) occurs for a very small \( \alpha \).

In other words, these assumptions underlie the above application of the principle of small-probability events for the rejection of \( H \) when \( \tilde{B} \) occurs for a very small \( \alpha \). They are also reasonable for the two common cases with, respectively, a Gaussian distribution and a chi-square distribution under \( H \).

**What happens if \( B \) occurs?** Now suppose that \( B \) occurs. Assumption A1’ implies that \( P\{B | H\} > P\{B\} \), or equivalently, \( P\{H | B\} > P\{H\} \) since

\[
P\{H | B\} = \frac{P\{B | H\}P\{H\}}{P\{B\}}
\]

However, the increase from \( P\{B\} \) to \( P\{B | H\} \) (or equivalently, from \( P\{H\} \) to \( P\{H | B\} \)) is small for a small \( \alpha \) and insignificant for a very small \( \alpha \). This can be seen as follows. Assumption A2’ (2) implies that

\[
\lim_{\alpha \to 0} \frac{P\{B | H\}}{P\{B\}} = 1, \text{ or equivalently, } \lim_{\alpha \to 0} P\{H | B\} = P\{H\}
\]

whenever \( P\{B\} \to 0 \) as \( \alpha \to 0 \), which usually holds. Further, this apparent support, however weak, is on shaky ground because the confidence to declare the occurrence of \( B \) is low for a small \( \alpha \), as discussed in more detail later.

**Rejection vs. acceptance: relative strengths.** Note the difference in the case of \( B \) and \( \tilde{B} \): For the same small \( \alpha \),
what $B$ refutes $H$ is much stronger than what $B$ supports $H$. This can be seen as follows with $x = P(B)$ and A1:

$$\eta = \frac{P(B)}{P(B|H)} - \frac{P(B|H)}{P(B)} = \frac{P(H|B)}{P(H)} - \frac{P(H|B)}{P(H)} = \frac{(x-\alpha)(1-(x+\alpha))}{\alpha(1-x)} > 0,$$

where $0 < \alpha + x < 1$, $\alpha < x$ provided $P(B) \neq 1$. Further, for an extremely small $\alpha$, it follows from A2* (i.e., $\alpha/x \to 0$ as $\alpha \to 0$) that the strength of rejecting $H$ when $B$ occurs is always by far stronger than that of accepting $H$ when $B$ occurs, because

$$\eta = \frac{(1-\alpha/x)(1-(x+\alpha))}{\alpha(x)(1-x)} \to +\infty$$

Note that usually $P(B|H) \to 0$ as $\alpha \to 0$. Fig. 1 shows $P(H|B)$ and $P(H|B)$ vs. $P(H)$ for $P(B|H) = \alpha^{1/10}$, $\alpha^{1/5}$, $\alpha^{1/3}$ and $\alpha = P(B|H)$ = 0.01, 0.05, 0.1, and 0.2. For a small $\alpha$, while the curves of $P(H|B)$ are much lower than the diagonal line (the smaller the $\alpha$, the lower), those of $P(H|B)$ are only slightly above the diagonal line (the smaller the $\alpha$, the closer to the diagonal line). These plots clearly demonstrate that for a very small $\alpha$, the relative strength of rejecting $H$ should $B$ occur is much stronger than accepting $H$ if $B$ occurs. From the Bayesian perspective, $P(H|B)$ is usually low (unless $P(H)$ is very high) and $P(H|B)$ is rarely high since $P(H)$ is unknown, not to mention that the meager support of $B$ for $H$ is on shaky grounds. As such, it is certainly not prudent at all to accept $H$ if $B$ occurs since the evidence is too inconclusive. For non-Bayesians, the same results are obtained by plotting $P(H|B)$ vs. $P(B)$ and $P(H|B)$ vs. $P(B)$, respectively, since $P(B|H)/P(B)$ = $P(H|B)/P(H)$ and $P(B|H)/P(B) = P(H|B)/P(H)$.

**Use only a high confidence level.** According to Cournot’s principle, the hypothesis $H$ should be rejected only if a rare event occurred on a single trial, that is, when $\hat{\theta}$ is outside a very high (say, 99%) interval, and this decision has a confidence of at least 99%. As demonstrated in Fig. 1, this is solid only if the decision is to reject $H$ and the confidence is very high (say, 95% or above). When $\hat{\theta}$ is outside an interval of a not high probability (say, 85%), rejection of $H$ and thinking it has a 85% or higher confidence is questionable because the underlying Cournot’s principle is not applicable here. When $\hat{\theta}$ is inside a very high (say, 95%) interval, $B$ should be deemed not against $H$ at the 95% confidence rather than for $H$ (loosely speaking, $B$ may be deemed evidence for $H$ only with a confidence that could be as low as 5%).

**Shaky ground for $B$ to support $H$.** Let $A$ be the event that $\hat{\theta}$ is located at an end point of the 98% confidence interval. Then $A$ is deemed $B$ or $B$, respectively, if a 95% or 99% confidence is used. So, the same $A$ would refute $H$ strongly or
support $H$ (weakly) depending on if a 95% or 99% confidence is used. This is not paradoxical, though: $A$ is strong evidence against $H$ at the 95%, but not the 99%, confidence; on the other hand, there is a chance of not higher than 5% and possibly as low as 1% that the same $A$ is weak evidence for $H$. This should make it clear that the weak support of $B$ for $H$ is indeed on shaky ground.

C. Conclusions

The test widely used in practice is: For a (very) small $\alpha$,
- Reject $H$ if $\hat{\theta} \notin (\theta_1, \theta_u)$;
- Accept $H$ if $\hat{\theta} \in (\theta_1, \theta_u)$. This conclusion is not well supported and thus not reasonable—it should be avoided.

The above analysis clearly shows the following:
- For a small (but not very small) $\alpha$, 
  - $\hat{\theta} \notin (\theta_1, \theta_u)$ refutes $H$ on solid grounds;
  - $\hat{\theta} \in (\theta_1, \theta_u)$ weakly supports $H$ on shaky ground; but the evidence is inconclusive.
- For an extremely small $\alpha$, we should:
  - Reject $H$ if $\hat{\theta} \notin (\theta_1, \theta_u)$;
  - Make no decision if $\hat{\theta} \in (\theta_1, \theta_u)$ since it only supports $H$ extremely weakly on very shaky ground.

These conclusions are quite intuitive. Version 2 is parallel.

In some applications, a conclusion must be drawn. If $B$ occurs in this case, our recommendation is: Resist the pressure against $H$. This should make it clear that the weak support of $B$ for $H$ is indeed on shaky ground.

B. Discussion

Given a data set, type I and type II error probabilities cannot be reduced simultaneously to an arbitrary level. In the NP framework, they are treated differently: while type I error probability may be maintained or controlled to an arbitrarily low level specified by the user, type II error probability can only be minimized, which could still be large. Simply put, it is helpful to deem the null and alternative hypotheses primary and secondary hypotheses, respectively. As such, assigning of the null and alternative hypotheses is usually not arbitrary in practice. Here are some guidelines:
- The type I error should be assigned to the one with a more serious consequence [4].
- The null hypothesis should be the one whose rejection is of most interest [13], [17].
- The generally accepted (or currently prevailing) hypothesis should be designated as the null one.
- The null hypothesis should be the one under which the distribution of the test statistic is better known [5].

The alternative hypothesis is also known as the research hypothesis (see, e.g., [13]): The main purpose of the testing is to reject the null hypothesis so as to establish the alternative (research) hypothesis. The last guideline above is an outgrowth of the need to be able to determine the decision threshold. A corollary is that the null hypothesis should include one with a parameter equality if one or more hypotheses involve such equalities [13].

Criminal-court example. A good example that demonstrates the importance and complexity of the choice of the null and alternative hypotheses is the criminal court cases: The defendant is assumed not guilty unless proven otherwise beyond a reasonable doubt. This is the well-known “presumed innocence.” In the language of hypothesis testing, the “not guilty” hypothesis should be the null and the “guilty” hypothesis should be the alternative. That is mainly because the mistake of convicting an innocent person (type I error) is usually considered more serious than setting a criminal free (type II error). In some countries, it is the defendant’s burden to prove his or her innocence. Such a case is equivalent to having the null hypothesis as “the defendant is guilty,” a real tragedy of mankind!
formulated as “type I error probability cannot exceed a small threshold,” although this threshold may have been set so small that type II error (setting criminals free) probability is too large.

Drug-development example. Another good example is the following. A new drug is developed. The developer would like to test if it is effective. Then the null hypothesis should be “the new drug is not effective.” He/she can thus set up a maximum allowable type I error probability (say, 0.05) and try to reject this hypothesis. If the test does not reject it, then nothing would be declared; otherwise the new drug would be declared at least 95% effective. Note that this designation follows the guidelines: (a) Type I error (declare the new drug to be effective while it is in fact not) is usually considered more serious than type II error (dismiss the new drug that is effective). (b) The rejection of the null hypothesis is of more interest than the rejection of the alternative. (c) It is more generally believed that a new drug is not effective (unless shown otherwise) because, for instance, much more new drugs are developed that are not effective than effective.

Scientific-theory example. Suppose that a new theory is developed that competes with the generally accepted (currently prevailing) theory. To test these two competing theories, the null hypothesis $H_0$ should be assigned to the generally accepted one for several reasons. First, most people would agree that incorrect rejection of the generally accepted theory (type I error) is more detrimental than incorrect rejection of the new theory (type II error). Second, rejection of the generally accepted theory is certainly more sensational than rejection of the new theory. Third, the distribution of the test statistic (whatever it is) under the prevailing theory is usually better determined (than under the new theory) since it is better studied. In short, all the other three guidelines stated above justify that a generally accepted or currently prevailing hypothesis should be designated as the null. Of course, were this not the case, new “accepted theory” would come out too frequently and there would be an instability of the prevailing theory.

Type I error $\neq$ false alarm. In detection theory, however, it is unfortunate that type I error is always assigned automatically to false alarm (declare signal present while signal is absent) and type II error to miss (declare signal absent while signal is present). Since a false alarm is normally not as bad as a miss (e.g., miss detection of an incoming missile or a fatal disease), assigning “signal absent” to $H_0$ and “signal present” to $H_1$ does not follow the general guidelines. It seems more appropriate in many cases to assign the null hypothesis $H_0$ to “signal present” and the alternative $H_1$ to “signal absent” such that the miss detection will not exceed a certain level. This unfortunate situation has a historical reason. The theory of detection is based largely on radar signal detection. Since radar is continuously monitoring the space and has many cells that can make a detection, if the null hypothesis were assigned to “signal present,” there would be way too many false alarms, which would render the system useless. Moreover, here miss detection at any particular moment in a given cell is not serious because the target would be detected anyway immediately afterwards almost certainly. Also, the literal interpretation of the null hypothesis as one that represents “no effect” or “no difference” to be tested against an alternative hypothesis that postulates the presence of an effect or difference (see, e.g., [6], [17]) is another source of this widespread fallacy. It should be clear now that for most other applications (e.g., detecting an indicator or factor of a disease) that do not involve a continuous detection effort and/or over a large number of detection cells, blindly assigning the null hypothesis to “signal absent” is questionable. Instead, the assignment should be determined carefully case by case.

C. Example: Detection of HIV by T Cell Count

Suppose that the normalized T cell count $z$ of a person has a Gaussian $\mathcal{N}(\theta, \sigma^2)$ distribution, where $\theta = \theta_1$ ($\theta_1 > 0$) if the person has HIV and $\theta = 0$ otherwise. We want to decide if the person has HIV using the NP test with $z$. If we blindly follow what is done normally by the following assignment

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta = \theta_1$$

the NP test is

$$f(z|H_1) = \exp \left[ -\frac{(z - \theta_1)^2}{2\sigma^2} \right] \mathbb{I}_{H_1}$$

$$f(z|H_0) = \mathbb{I}_{H_0}$$

where the threshold $\lambda_0$ is determined by the type I error probability $\alpha$ as

$$P_I = P\{z > \lambda_0|H_0\} = \int_{\lambda_0}^{\infty} \mathcal{N}(z; 0, \sigma^2) dz = \alpha$$

For example, if $\alpha = 5\%$, then $\lambda_0 = 1.65\sigma$ and thus the NP test is

$$H_1 \overset{H_0}{\Rightarrow} 1.65\sigma$$

This test has the type II error (miss detection) probability $P_{md}$ given in Table I, obtained by

$$P_{II} = P\{z < \lambda_0|H_1\} = \int_{-\infty}^{-1.65\sigma} \mathcal{N}(z; 0, \sigma^2) dz$$

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{md}$ with (3)</td>
<td>0.0004</td>
<td>0.046</td>
<td>0.196</td>
<td>0.359</td>
<td>0.74</td>
</tr>
<tr>
<td>$P_{II}$ with (4)</td>
<td>0.0004</td>
<td>0.046</td>
<td>0.196</td>
<td>0.359</td>
<td>0.74</td>
</tr>
</tbody>
</table>

The probability of miss detection $P_{md}$ is quite large if $\sigma$ is not substantially smaller than unity—it is as large as 0.74 for $\sigma = 1$! This is very bad and often fatal: There would be no HIV treatment if a miss detection occurs. Using the above questionable assignment of the null hypothesis, one has to live with such terrible results since the test is optimal. Instead, if we use the following assignment

$$H_0 : \theta = \theta_1 \text{ vs. } H_1 : \theta = 0$$

TABLE I

MISS-DETECTION PROBABILITY WITH (3) AND FALSE-ALARM PROBABILITY WITH (4).
with the NP test, we can guarantee \( P_{\text{md}} \leq \alpha \) for an arbitrarily small \( \alpha \). In this case with \( \alpha = 5\% \), the false alarm probability \( P_{fa} \) is given in Table I. Yes, \( P_{fa} \) can be large, but this would not be as bad as \( P_{\text{md}} \) is large, especially if the person is made aware of the fact that \( P_{fa} \) can large.

In summary, the assignment of the null hypothesis (i.e., the correspondence between \( \text{H} \) and \( \text{H} \) and false alarm and miss probabilities) in the NP framework should be determined at least case by case, rather than blindly following the current common practice.

Note that the assignment can be done arbitrarily in other hypothesis testing framework, such as the Bayesian and the sequential probability ratio test frameworks, where the null and alternative (and other) hypotheses are treated symmetrically.

IV. SIGNIFICANCE TESTS: SINGLE VS. BINARY

There is much confusion about significance tests, especially between those for a single hypothesis and those for binary hypotheses. The two classes of significance tests are:

Class A: Only a single hypothesis \( H \) is specified; that is, the distribution when \( H \) is not true is unknown.

Class B: Two hypotheses are specified: one as the null and the other as the alternative.

We use the following example to illustrate the fallacies and drive home the points we want to make.

A. Example: Statistical Significance of Estimate

Binary-hypothesis case. This example is modified from one in [3]. A parameter \( \theta \) is estimated based on \( n \) measurements, each corrupted by \( N(0,10^2) \) noise. For

\[
H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0
\]

since the NP test for \( z \sim N(\theta, \sigma^2) \) is, with \( \alpha = 95\% \),

\[
\hat{\theta} \mid H_1 \geq \lambda = 1.96 \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \frac{\hat{\theta}}{\sigma/\sqrt{n}} \geq 1.96 \quad H_1 \quad H_0
\]

the test statistic takes the following values, assuming \( \hat{\theta} = 2 \),

\[
\frac{\hat{\theta}}{\sigma/\sqrt{n}} = \begin{cases} 
2.0 & \text{for } n = 1000 \\
100 & \text{for } n = 100 \\
20 & \text{for } n = 10
\end{cases}
\]

Compared with the threshold 1.96, we conclude that with \( \alpha = 95\% \) this estimate \( \hat{\theta} \) should be deemed

- statistically significant, if \( \hat{\theta} = 2 \) is obtained from 1000 measurements;
- marginally statistically significant, if \( \hat{\theta} = 2 \) is obtained from 100 measurements;
- statistically insignificant if \( \hat{\theta} = 2 \) is obtained from 10 measurements.

We emphasize that the distribution under \( H_1 \) is well specified in this example—\( H_1 : z \sim N(\theta, \sigma^2) \) with \( \theta \neq 0 \).

Single-hypothesis case. Assume now everything is the same as before except that the distribution under \( H_1 \) is not known at all. The same test statistic and threshold can be used since

\[
P\{\sqrt{n\hat{\theta}/\sigma} \in (-1.96,1.96)|H_0\} = 0.95
\]

We conclude that with \( \alpha = 95\% \) the estimate \( \hat{\theta} \) is deemed

- statistically significant, if \( \hat{\theta} = 2 \) is obtained from 1000 measurements;
- marginally statistically significant, if \( \hat{\theta} = 2 \) is obtained from 100 measurements.

However, if \( \hat{\theta} = 2 \) is obtained from 10 measurements, we no longer can conclude that the estimate \( \hat{\theta} \) should be deemed statistically insignificant. It is even more inappropriate to think that the estimate is statistically insignificant with a 95% confidence; otherwise all estimates are insignificant at the 100% level because every \( \hat{\theta} \) falls inside the 100% interval, which is \((−\infty,\infty)\). In fact, as discussed in Sec. II, in this case no conclusion should be drawn in principle, although a widespread misconception in practice is to conclude that the estimate is statistically insignificant. Were this correct, we could always draw this conclusion by using a large enough probability, say, 0.999999, so that the interval becomes large enough to include \( \hat{\theta} \). In this case, data may even suggest that the null hypothesis is false but the evidence is not strong enough to make a convincing case (beyond a reasonable doubt). Put it differently, a significance test for the single-hypothesis case can only be used to reject the null hypothesis since it is the only case where Cournot’s principle applies. If \( \hat{\theta} \) is inside the interval, we do not know whether \( \theta = 0 \) or not, let alone to conclude \( \theta = 0 \), as presented in Sec. II-B. This subtle drawback is inherent in general in all significance tests for a single hypothesis. However, it is not well known in the engineering community, probably due to a lack of knowledge and warning as well as the deceptively simple structure of the test.

In summary, a significance test for the single-hypothesis case can only be used to reject the null hypothesis (e.g., to conclude that the estimate is statistically significant) and this can be done only with an extremely high confidence. Any other conclusion lacks solid grounds.

For a detailed discussion of a concrete erroneous use of the significance test in the context of testing the credibility of a filter’s self-assessment (error covariance and/or bias) using the so-called normalized estimation error squared [1], see [11], [12].

Note that 5% (95%) and 1% (99%) are not magic numbers—they are not really special, except that they are widely used due to the availability of statistical tables for them [10]. In general, they must be very small. 5% is about the largest that can be used here since it is based on the small-probability-event principle. The use of 5% as the significance point, which has some merits, was first used repeatedly by Fisher, but even Fisher himself did not stick to it; he used sometimes a larger value and often a smaller value [7].

B. SINGLE-HYPOTHESIS TESTING VS. BINARY HYPOTHESIS TESTING

To help the reader understand why such a distinction exists between the two classes of significance tests, we highlight the

\footnote{Except that we cannot conclude with 95% or higher confidence that the estimate is statistically significant.}
differences between single-hypothesis testing and binary/M-ary hypothesis testing as follows.

- Binary (and M-ary) hypothesis testing *presumes* that one and only one of the hypotheses should not be rejected. This is not the case for single hypothesis testing.
- Based on the above, binary (and M-ary) hypothesis testing relies on optimization. For single-hypothesis testing, however, the concept of optimality is elusive. Rather, it is underpinned by the principle of small-probability events.
- Also based on the above, in single-hypothesis testing, the question is *whether* the hypothesis should be rejected; in binary (and M-ary) hypothesis testing, the question is *which* hypothesis is better (or best) not rejected.
- While a single-hypothesis testing problem with $H_0 : \theta = \theta_0$ can be formally formulated as a binary problem of testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, the distribution under $H_1$ is not specified. As such, the standard tests (e.g., the NP test, the maximum likelihood test, and the Bayes test) are not applicable, which all rely on knowledge of the distribution under $H_1$. Therefore, the formulation of single hypothesis testing is more appropriate than its “equivalent” binary formulation if the distribution when $H_0$ is not true cannot be specified reasonably well.
- The binary tests are more suitable when we are concerned with whether to reject (rather than to confirm) the null hypothesis $H_0$. While this sounds similar to the rationale for the single-hypothesis test, they are actually quite different: For the single-hypothesis test, no solid conclusion can be drawn if it cannot reject $H_0$ because the decision is absolute (i.e., whether $H_0$ is implausible, rather than whether $H_0$ is more plausible than $H_1$ or not); for binary tests, failure to reject $H_0$ justifies to “accept” $H_0$ since the error probability of this decision is minimized. In this sense, the NP decision is *relative*. The “accepted” $H_0$ is usually more likely to be better than $H_1$. This is because the single-hypothesis tests rely on the principle of small-probability events while the binary tests are based on certain optimization principle. Note that as the price of being more solid than the optimization principle, the principle of small-probability events is more conservative—in some cases it cannot lead to a reasonable conclusion.

In the statistical community, the single-hypothesis tests and tests for two or more hypotheses are often referred to as significance tests and hypothesis tests, respectively [9].

To gain better insight into the difference between the single-hypothesis testing $T_1 = \text{"testing $H_0$: } \theta \in \Theta_0\text{"}$ and the NP binary hypothesis testing $T_2 = \text{"testing ($H_0$: } \theta \in \Theta_0\text{ vs. } H_1$: } \theta \in \Theta_1\text{"}$, we provide the following further comments.

- Type I error probability $P_1$ in both tests can be controlled as low as wish—their thresholds are determined by distributions of the test statistics under $H_0$. But there are infinitely many test statistics and thresholds that yield identical $P_1$.
- The single-hypothesis testing $T_1$ tries to use the shortest “acceptance” interval in the hope of having a small type II error probability $P_{11}$; whereas the NP test $T_2$ uses one in the optimal class where $P_{11}$ is minimized.
- For a simple $H_1$, $P_{11}$ can always be minimized and so we regard $T_2$ to be superior to $T_1$.
- For a composite $H_1$, $P_{11}$ in general depends on values of $\theta \in \Theta_1$. Note the following:
  - It is minimized by the uniformly most powerful test (or some variant such as the uniformly most powerful invariant (or unbiased) test), if it exists, and so we regard $T_2$ to be better than $T_1$.
  - In most other cases, $T_2$ is still preferred to $T_1$ since $P_{11}$ is bounded (i.e., $P_{11} \leq \beta$) or is minimized for at least some $\theta \in \Theta_1$.
- There is no preference between $T_1$ and $T_2$ if there is no knowledge about $P_{11}$.
- $T_2$ cannot be used if $P_{11}$ is unknown, such as when the distribution under $H_1$ cannot be specified.

**Conclusion:** $T_2$ beats $T_1$ if the assumed $P_{11}$ is known and correct.

V. NOT REJECTED $\neq$ ACCEPTED TO BE TRUE

Even in the binary or M-ary case, a hypothesis can only be rejected, not confirmed (to be true). A hypothesis that is not rejected should not be interpreted as “accepted to be true” [18]; such a hypothesis can only be accepted for the moment on a provisional basis [9]. In other words, a hypothesis that is not rejected should not be deemed true. This is so for several reasons. For example, the truth need not be one of the hypotheses. In the binary or M-ary case, we can reject a hypothesis if it is worse than another one, but the best hypothesis in a given set is still not necessarily true. We offer an extreme example next.

**Example: Comparison of Credibility of Two Estimators**

As introduced in [11], [12], an estimator is said to be *credible* if its self-calculated bias and mean-square error (MSE) matrix are deemed statistically equal to the actual bias and MSE matrix. Given two estimators $\hat{x}_1$ and $\hat{x}_2$ having errors with unknown ($\mu_1^*, P_1^*$) and self-assessed biases and MSE matrices ($\mu_i, P_i$), $i = 1, 2$, we want to know which estimator is *less* credible. This can be formulated as a hypothesis testing problem of $H_1$ vs. $H_2$, where

$$H_1: \hat{x}_1 \text{ is less credible than } \hat{x}_2$$
$$H_2: \hat{x}_2 \text{ is less credible than } \hat{x}_1$$

While this does not imply that $\hat{x}_1$ or $\hat{x}_2$ is credible—both may be non-credible—we may nevertheless specify the hypotheses as [12]

$$H_1: (\mu_1^*, P_1^*) \neq (\mu_1, P_1) \text{ and } (\mu_2^*, P_2^*) = (\mu_2, P_2)$$
$$H_2: (\mu_1^*, P_1^*) = (\mu_1, P_1) \text{ and } (\mu_2^*, P_2^*) \neq (\mu_2, P_2)$$

(Note that $(\mu_1^*, P_1^*) = (\mu_1, P_1)$ implies that $\hat{x}_i$ is (perfectly) credible and that neither $H_1$ nor $H_2$ is likely to be true.) This formulation is justified by that statistical hypothesis testing is
not concerned with which hypothesis is true, but with which hypothesis is unlikely (or less likely) to be true and should be rejected; in other words, the hypothesis not rejected is not necessarily deemed true.

In this example, no matter what rigorous definition of credibility is, “(μ₁, P₁) ≠ (μ₂, P₂)” is a good representative of “μ₁ is less credible than μ₂.” This formulation is similar in spirit to that of the (generalized) likelihood ratio test, where a composite hypothesis with unknown parameters is represented by its most likely member. It is also similar to that of the Bayes test, where a composite hypothesis with random parameters is surrogated by the simple one obtained by averaging the composite hypothesis over the random parameters. Note that this surrogate simple hypothesis is not even necessarily a member of the original composite hypothesis.

As we made clear in the previous sections, a hypothesis that is not rejected should not even be accepted in the case of single-hypothesis testing. Secs. II and IV provides concrete examples of the single-hypothesis case. For binary hypothesis testing, rejection of H₁ is better interpreted as “the alternative H₁ is rejected” or “the null hypothesis H₀ is not rejected,” rather than “the null hypothesis H₀ is (accepted to be) true.” This is in analogy to the jury verdict that “the defendant is not guilty,” meaning that not sufficient evidence has been shown to convict the defendant beyond a reasonable doubt, rather than “the defendant is innocent,” which means that sufficient evidence has been shown to prove the defendant’s innocence (beyond a reasonable doubt).

The above interpretation agrees with probabilistic inductive reasoning, which maintains that while evidence may disprove a hypothesis, it does not confirm a hypothesis, but only increases (or decreases) its credibility.

This interpretation is also justified by the falsificationist viewpoint in the philosophy of science, spearheaded by Karl Popper, that evidence may or may not refute a theory but does not prove it. According to this philosophy, a theory cannot be a scientific one if it is not falsifiable (i.e., there is no possibility to be proven wrong by any evidence) [15].

Simply put, a hypothesis should not be deemed true simply because it is not rejected. It should be so deemed only if (a) no evidence disproves it and (b) its bold predictions are verified or overwhelming evidence of various kinds supports it. However, no statistical method is available that achieves this. This unavailability is a serious deficiency of the existing theory of statistical hypothesis testing. In view of the fact that many scientific theories are accepted to be true based (at least initially) on the principle of small-probability events, it is worthwhile to try to develop statistical hypothesis tests that may actually confirm a hypothesis based on the same principle. Until something like that is done, however, we cannot really deem a hypothesis to be “true” no matter how many statistical hypothesis tests it has succeeded to pass (i.e., not rejected).

VI. Conclusions

We have pointed out several widespread fallacies and confusions in application of hypothesis testing. They are: (a) inappropriate acceptance of the single hypothesis when it is not rejected based on the data; (b) blind assignment of the null hypothesis to the “signal absence” hypothesis in the Neyman-Pearson framework; (c) confusion between two classes of significance tests; and (d) improper acceptance of a hypothesis not rejected as being true. Simply put, they are inappropriate because of the following facts: Single-hypothesis testing is reliant on the principle of small-probability events; the null hypothesis is treated more seriously in the Neyman-Pearson setting than the alternative; the two classes of significance tests are based on two distinctive principles; and the truth is not necessarily one of the hypotheses.

Acknowledgments

Fig. 1 was plotted by Mr. Yongqi Liang, a visiting student. The authors would like to acknowledge this help.

References