Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 483-499
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# Why It Is Problematic to Calculate Probabilities of Findings Given Range Null Hypotheses

**David Trafimow**

New Mexico State University, Las Cruces, NM, USA
Email: dtrafimo@nmsu.edu

**How to cite this paper:** Trafimow, D. (2017) Why It Is Problematic to Calculate Probabilities of Findings Given Range Null Hypotheses. *Open Journal of Statistics*, **7**, 483-499.
https://doi.org/10.4236/ojs.2017.73034

**Received:** March 17, 2017
**Accepted:** June 17, 2017
**Published:** June 20, 2017



## Abstract

An important problem with null hypothesis significance testing, as it is normally performed, is that it is uninformative to reject a point null hypothesis [1]. A way around this problem is to use range null hypotheses [2]. But the use of range null hypotheses also is problematic. Aside from the usual issues of whether null hypothesis significance tests can be justified at all, there is an issue that is specific to range null hypotheses. It is not straightforward how to calculate the probability of the data given a range null hypothesis. The traditional way is to use the single point that maximizes the obtained *p*-value. The Bayesian alternative is to propose a prior probability distribution and integrate across it. Because frequentists and Bayesians disagree about a variety of issues, especially those pertaining to whether it is permissible to assign probabilities to hypotheses, and what gets lost in the shuffle is that the two camps actually come to different answers for the probability of the data given a range null hypothesis. Because the probability of the data given the hypothesis is a precursor for both camps, for drawing conclusions about hypotheses, different values for this probability for the different camps is crucial but seldom acknowledged. The goal of the present article is to bring out the problem in a manner accessible to researchers without strong mathematical or statistical backgrounds.

## Keywords

Range Hypotheses, One-Tailed Test, Range Null Hypothesis, Uninformative Hypotheses

## 1. Introduction

Frequentists and Bayesians disagree about how to handle the inverse inference issue. How does a researcher traverse a pathway from the calculated probability of the finding given a hypothesis (such as the null hypothesis) to the probability

of the hypothesis given the finding? Bayesians argue that direct inverse inferences are invalid, thereby similarly invalidating the null hypothesis significance testing procedure. In contrast, frequentists criticize Bayesians for having to make unjustified assumptions about priori probabilities of null hypotheses to allow the Bayesian machinery to run. In marked contrast to this issue, there is little literature on what would seem to be an issue that precedes the inverse inference issue; namely, how does one calculate the probability of the finding given a hypothesis in the first place? It might seem that this is straightforward, and it is straightforward in the context of point hypotheses. But it is not straightforward in the context of range hypotheses which provide the present focus. Put simply, the question of interest is: Given a range hypothesis, how can one calculate the probability of the finding given it?

To understand why we should care about range null hypotheses at all, it is necessary to consider, in detail, that which is so well known that few consider it carefully. First, there is a preliminary issue about whether hypotheses can have probabilities at all. Second, there is an additional preliminary issue about precisely the logic by which frequentists decide between competing hypotheses. My immediate goal in these sections is not to take sides but rather to bring out the disagreements. My more general goal is to show that both sides can be faulted not just on the difficult issue of inverse inference, but even on the more basic issue of calculating the probability of the finding given a range hypothesis. If a calculation this basic already is problematic, the inverse inference issue may be even more intractable.

## Probabilities of Hypotheses

Bayesians and frequentists disagree with each other with respect to how they draw conclusions about hypotheses. Bayesians are willing to assume that hypotheses have probabilities anywhere between 0 and 1, whereas the furthest frequentists are willing to go is to allow that hypotheses can have probabilities of 0 (hypothesis is false) or 1 (hypothesis is true), but nothing between these extreme values. And of course, most frequentists will freely admit that they do not know whether to assign a probability of 0 or a 1 to a particular hypothesis. This admission causes most frequentists to focus on procedures for controlling the error rate, rather than assigning values to particular hypotheses [3] [4]. From a Bayesian perspective, frequentists might be considered to be too "conservative" but from a frequentist perspective, Bayesians can be considered to be too "liberal."

Graduate training in the sciences tends to stress scientific conservativism; scientists should demand reasonably impressive evidence before being willing to draw a conclusion. From this perspective, the fact that frequentists are more conservative than liberals easily can be taken as evidence that frequentists are more "scientific" than Bayesians. When contrasting the two perspectives against each other at the level of drawing conclusions about hypotheses, it is relatively easy to make this argument. It clearly is more conservative to admit to not knowing how to assign a probability to a hypothesis than to insist that one does

know how to do this. Saying "I do not know" is more conservative than making numerical assignments of numbers to hypotheses.

In the other direction, however, Bayesians could claim that frequentists are too liberal because they use 0.05 as the alpha level for deciding statistical significance. In general, Bayesians claim to be more conservative than frequentists because they insist that the probability of the favored hypothesis given the finding, when they get to that point, be at least 8 or even 10 times greater the probability of the hypothesis that is not favored, before believing the favored hypothesis [5]. It is possible to suggest that the two groups are talking about apples and oranges because of the difficulty of comparing *p*-values against posterior ratios.

Thus, the two sides disagree on whether the notion of a probability of a hypothesis (other than zero or one) makes sense at all, whether it is possible to calculate such an entity even if it did make sense, on the importance of the probabilities of findings given hypotheses, and on whether liberalism or conservativism is about ratios of hypothesis probabilities or about probabilities of findings given hypotheses. But the main issue of interest here is yet to come, and it falls out of the issue of the plausibility of point null hypotheses.

## 2. Are Null Hypotheses Plausible?

Over many decades, there has accumulated much criticism pertaining to the null hypothesis significance testing procedure (NHSTP) [1] [6]-[23]. The criticism that is of particular present relevance is that because the null hypothesis specifies an exact value when there is an infinitude of possible values, the null hypothesis almost certainly is not true [1]. Therefore, the NHSTP is a pointless exercise because it results in the rejection of a null hypothesis that is not plausible anyhow. Although there have been a couple of attempts to argue that the null hypothesis is plausible under particular circumstances [24] [25], this is not the main defense. The main defense is that one does not have to settle for using a point null hypothesis. It is possible to let the null hypothesis specify a range of values, as is the case when one performs a one-tailed test, and so the rejection of the null hypothesis is meaningful after all [26] [27] [28].

My goal is to examine this argument carefully to see where it leads. However, it is first necessary to review the syllogisms that come into play in discussions of this sort.

## 3. The Syllogisms

Let us commence with the usual logic that accompanies traditional two-tailed significance tests. In such cases, researchers define a point null hypothesis to be contrasted against a range alternative hypothesis. Because the arguments to be developed do not depend on the idiosyncrasies of any particular type of study, let us consider the simplest possible case of coin tosses and whether or not the coin is fair. We might define null and alternative hypotheses as follows where $P(H)$ refers to the probability of heads.

**Case 1**

$H_0$: $P(H) = 0.5$

$H_1$: $P(H) \neq 0.5$

In the foregoing case, the logic is simple and based on the ability to use a small $p$-value to reject the null hypothesis[1]. Specifically, we have the following syllogism.

**Syllogism 1**

| | |
|---|---|
| $H_0$ or $H_1$ | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

There can be no doubt that Syllogism 1 is valid. But as pointed out earlier, Syllogism 1 can be criticized as not being informative because it is extremely implausible, a priori, that a coin is *perfectly* fair.

It is easy to imagine another state of affairs, accompanied by another syllogism. Consider Case 2 and Syllogism 2 below.

**Case 2**

$H_0$: $P(H) = 0.5$

$H_1$: $P(H) > 0.5$

**Syllogism 2**

| | |
|---|---|
| $H_0$ or $H_1$ or something else [*i.e.*, $P(H) < 0.5$] | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

Syllogism 2 has a rather obvious flaw that stems from the fact that Premise 1 states three possibilities, which is necessitated by the fact that Case 2 leaves open the possibility that $P(H)$ can be greater than 0.5 (alternative hypothesis), equal to 0.5 (null hypothesis), or less than 0.5 (unstated hypothesis). Therefore, rejecting the null hypothesis that the probability of heads is equal to 0.5 does not allow an unambiguous conclusion about whether this probability is less than or greater than 0.5. Put simply, Syllogism 2 is blatantly invalid when based on Case 2. Perhaps it is a recognition of this invalidity that is responsible for some statistical authorities favoring range null hypotheses. As an example, consider Case 3 and Syllogism 3 below[2].

**Case 3**

$H_0$: $P(H) \leq 0.5$

$H_1$: $P(H) > 0.5$

**Syllogism 3**

| | |
|---|---|
| $H_0$ or $H_1$ | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

The combination of Case 3 and Syllogism 3 seems beautiful. It is logically valid and, at the same time, solves the problem that we had earlier of rejecting a non-plausible null hypothesis. Rejecting the null hypothesis in Case 3 is quite informative because doing so also causes half of the possibilities to be rejected,

[1]In the interest of full disclosure, I do not believe that a small $p$-value justifies rejecting the null hypothesis. However, let us accept this premise for the sake of argument.

[2]Note that Case 2 and Case 3 both would be tested using a one-tailed test according to the traditional null hypothesis significance testing procedure.

thereby allowing a directional hypothesis to be supported. Therefore, it is worth examining this combination in more detail.

## 4. Getting the *p*-Value

The standard way to handle the combination of Case 3 and Syllogism 3 is to use a one-tailed test. For coin tosses, one would use the binomial theorem. Suppose that one has obtained $k$ heads out of $N$ tosses. The one-tailed probability is simply the probability of having obtained $k$ heads out of $N$ tosses, plus the probability of having obtained $k + 1$ heads out of $N$ tosses, and so on, up to $N$ heads out of $N$ tosses. The binomial theorem is presented below as Equation (1):

$$P(k \text{ heads out of } N \text{ tosses}) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}. \quad (1)$$

Suppose that an investigator performed a study that involved $N = 20$ coin tosses and $k = 17$ heads. The normal procedure would be to use Equation (1) as follows. Set $p$ at the "fair coin" level of 0.5 (note that this $p$ is not the same $p$ as in $p$-value), and substitute 20 and 17 for $N$ and $k$ in Equation (1), respectively, but with three more iterations where 18, 19, and 20 are substituted for $k$. This is performed below:

$$\begin{aligned} &P(17 \text{ or more heads out of } 20 \text{ tosses}) \\ &= \frac{20!}{17!(20-17)!} 0.5^{17} (1-0.5)^{20-17} + \frac{20!}{18!(20-18)!} 0.5^{18} (1-0.5)^{20-18} \\ &\quad + \frac{20!}{19!(20-19)!} 0.5^{19} (1-0.5)^{20-19} + \frac{20!}{20!(20-20)!} 0.5^{20} (1-0.5)^{20-20} \\ &= 0.001 \end{aligned}$$

In other words, the result of our hypothetical experiment is highly significant and so Syllogism 3 can proceed without hindrance. Or so it seems.

But there is a major problem with the foregoing mathematics in the context of Case 3. Specifically, the calculation performed gives the probability of 17 or more heads out of 20 tosses, based on a single population parameter ($p = 0.5$). But the calculation does not cover if $p$ equals 0.49, 0.48, and so on down to 0. Depending on one's philosophical perspective, it is far from clear that the calculation based on a single parameter applies to the whole range of values. Arguably, one would have to assign (prior) probabilities to all of the values within the range between $p = 0$ and $p = 0.5$, and integrate across that range. To render the arguments accessible to everyone, however, let us simplify the problem and see where the simplification takes us.

## 5. Simplifying the Problem

Let us commence with a null hypothesis that specifies only two values, rather than dealing with a range of values. Later, we will add more values.

### 5.1. The Example of a Null Hypothesis with Two Values

Suppose that we have a null hypothesis with two values instead of a null hypo-

thesis with a range of values. This is shown in Case 4.

**Case 4**

$H_0$: $P(H) = 0.5$ or $P(H) = 0$

$H_1$: $P(H) \neq 0.5$ and $P(H) \neq 0$

**Syllogism 4**

| | |
|---|---|
| $H_0$ or $H_1$ | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

Syllogism 4 is logically valid and it makes use of Case 4 where the null hypothesis specifies two values. Assuming, as usual, that a sufficiently low probability of the finding given the null hypothesis justifies rejecting the null hypothesis, how should the probability be calculated?

To answer this question, let us put aside the null hypothesis for a moment and consider the abstract case where we are concerned with the probability of $A$ given that $C$ or $D$ is true. For example, imagine we are invited to dinner and we are interested in the probability that our hostess will serve chocolate for dessert ($A$) given that she serves chicken ($C$) or fish ($D$) for dinner. In symbols, we are interested in $P\left(A|\left(C \text{ or } D\right)\right) = P\left(A|\left(C \cup D\right)\right)$.

Let us assume that $C$ and $D$ are mutually exclusive $\left(C \cap D = \varnothing\right)$. In the dinner example, our hypothetical hostess would never serve both chicken and fish for dinner, though she might serve either one or something else entirely. It is possible to rewrite the expression of interest so that we have only conditional and unconditional probabilities (see Equation (5) below):

$$P\left(A|\left(C \cup D\right)\right) = \frac{P\left(A \cap C\right) + P\left(A \cap D\right)}{P\left(C\right) + P\left(D\right)}, \tag{2}$$

$$P\left(A|\left(C \cup D\right)\right) = \frac{P\left(A \cap C\right)}{P\left(C\right) + P\left(D\right)} + \frac{P\left(A \cap D\right)}{P\left(C\right) + P\left(D\right)}, \tag{3}$$

$$P\left(A|\left(C \cup D\right)\right) = \frac{P\left(A \cap C\right)}{P\left(C\right)} \cdot \frac{P\left(C\right)}{P\left(C\right) + P\left(D\right)} + \frac{P\left(A \cap D\right)}{P\left(D\right)} \cdot \frac{P\left(D\right)}{P\left(C\right) + P\left(D\right)}, \tag{4}$$

$$P\left(A|\left(C \cup D\right)\right) = P\left(A|C\right) \cdot \frac{P\left(C\right)}{P\left(C\right) + P\left(D\right)} + P\left(A|D\right) \cdot \frac{P\left(D\right)}{P\left(C\right) + P\left(D\right)}. \tag{5}$$

Equation (5) makes clear that we need not only the conditional probability of $A$ given $C$ or $D$, but we also need the unconditional probability of $C$ and the unconditional probability of $D$, in order to calculate the conditional probability of $A$ given that $C$ or $D$ is true. Returning to our hostess, if we wish to calculate the probability that she will serve chocolate for dessert given that she serves chicken or fish for dinner, we need to know the unconditional probability that she will serve chicken for dinner and the unconditional probability that she will serve fish for dinner. If we do not know these two unconditional probabilities, there is no way for us to calculate the conditional probability that our hostess will serve chocolate for dessert given that she serves chicken or fish for dinner.

Let us now apply what we learned from our hostess to consider again the

combination of Case 4 and Syllogism 4, where the null hypothesis specifies that $p = 0$ or $p = 0.50$. Equation (5) tells us that in order to calculate the probability of getting, say, 17 heads out of 20 tosses, given that the population proportion of heads is 0 or 0.50, we would need to know the unconditional probability that the population proportion of heads is 0.50 and the unconditional probability that the population proportion of heads is 0. We saw earlier how to calculate the conditional probability of 17 heads out of 20 tosses given a single value ($p = 0.50$), but Equation (5) shows that this is insufficient when there are two population values to consider. Again, although we do need the conditional probability of 17 heads out of 20 tosses, given that $p = 0$ or $p = 0.50$, we also need the two unconditional probabilities concerning the 0 and 0.50 population values. Are we stuck?

It depends, to some extent, on one's philosophical position pertaining to whether hypothesized population values can take on probabilities. If one's answer is "yes," as would be the case with most Bayesians, then we are not stuck. The researcher would find an arbitrary way of assigning probabilities to $p = 0$ and $p = 0.50$, and then it would be easy to carry the calculation through. An example of an arbitrary approach would be to say that because we have no reason to favor $p = 0$ over $p = 0.50$, or to favor $p = 0.50$ over $p = 0$, we can assign a probability of 0.5 to each of these. Using this arbitrary system, we might perform the following calculation:

$$
\begin{aligned}
&P\left(\text{17 or more heads out of 20 tosses} \mid \left(p = 0 \text{ or } p = 0.5\right)\right) \\
&= P\left(\text{17 or more head out of 20 tosses} \mid p = 0\right) \cdot \frac{P\left(p = 0\right)}{P\left(p = 0\right) + P\left(p = 0.50\right)} \\
&\quad + P\left(\text{17 or more heads out of 20 tosses} \mid p = 0.50\right) \cdot \frac{P\left(p = 0.50\right)}{P\left(p = 0\right) + P\left(p = 0.50\right)} \\
&= 0 + 0.001\left(\frac{0.5}{0.5 + 0.5}\right) = 0.0005.
\end{aligned}
$$

Note that because it is impossible to get 17 (or any) heads out of 20 tosses if the population proportion of heads is 0, the whole first term is 0. Regarding the second term, we saw earlier that the probability of 17 heads out of 20 tosses, if the population proportion is 0.5, is 0.001[3]. Thus, the final value is 0.0005, which is half the value we had obtained earlier. We could have assigned a probability of 1 to the 0.5 value and a probability of 0 to the 0 value, in which case we would have obtained the same result as the calculation performed earlier. Or we could have assigned a probability of 0.75 to the 0.5 value and a probability of 0.25 to the 0 value, or anything else we wished. The advantage of arbitrary unconditional probabilities is that they allow us to run the mathematical machinery. A disadvantage is that admitting the arbitrariness of assignments of unconditional probabilities to population values renders difficult a convincing argument that

[3]Because the probability of the finding given the null hypothesis is computed directly, there is no test statistic. If the dependent measure were continuous, rather than discrete (binary), it would be necessary to calculate a *t*-value.

any particular answer is the correct answer. After all, a different assignment of probabilities to unconditional probabilities would result in different "correct" answers. Is there another way to think about this?

In fact, there is, though it also depends on arbitrariness [29] [30]. To move in this direction, consider that in the foregoing reasoning, we commenced with the assumption that it is reasonable to assign probabilities to hypotheses. But it is possible not to assume this and there are two possibilities for not making the assumption. One possibility is to assert that hypotheses do not have any probabilities whatsoever. If we make this assertion, then there obviously is no way to use Equation (5) to calculate the probability of obtaining 17 heads out of 20 coin tosses, and so there is no way to enable Syllogism 4. A second possibility is to assert that hypotheses are true or false—that is, they have probabilities of 1 or 0—but we do not know which. From this point of view, it is possible to make progress as follows. Suppose that we assign a probability of 1 to the 0.5 population value and a probability of 0 to the 0 population value. In that case, the probability of heads comes out to the same 0.001 that we saw earlier by using the binomial theorem. Why might we be justified in performing this operation?

An argument that can be used is as follows. Based on the binomial theorem, we would obtain a larger probability of 17 heads out of 20 tosses using a population value of 0.5 than using a value of 0. Therefore, by using 0.5 as the population value, we can be sure that we are obtaining the largest possible probability of the finding, given the null hypothesis. Because the goal eventually will be to use a low probability of the finding, given the null hypothesis, to reject the null hypothesis, using 0.5 as the population value renders a conservative judgment. If the probability of 17 heads out of 20 tosses is 0.001 assuming $p = 0.50$, the probability of 17 heads out of 20 tosses is lower than 0.001 using $p = 0.49$, $p = 0.48$, and so on, down to $p = 0$. In fact, as we saw earlier, the probability of 17 heads out of 20 tosses given a population value of 0, is 0.

So what is correct? If one assumes that hypotheses about population values can have probabilities other than 0 or 1, this results in a dilemma for those who wish to perform one-tailed tests to reject null hypotheses. That is, to make the logic work out so that rejecting the null hypothesis is both meaningful (because it specifies a range rather than a point) and also really does force acceptance of the alternative hypothesis, it is necessary to have a range null hypothesis rather than a point null hypothesis. However, again from the perspective that it is reasonable to assign probabilities to hypotheses, Equation (5) shows that the mathematics typically used gives wrong answers! That is, the mathematics of using only the binomial theorem, without taking unconditional probabilities into account, is blatantly wrong by Equation (5). On the other hand, if one uses the argument that it only is permissible to assign a value of 0 or 1 to each population value, perhaps it is possible to justify the binomial calculation that results in a value of 0.001 for the probability of 17 heads out of 20 tosses. We will discuss this further, but let us come closer to considering the whole range of population values first.

### 5.2. Using 0, 0.01, 0.02, ···, 0.50 as Population Values for the Null Hypothesis

The combination of Case 4 and Syllogism 4 used a null hypothesis that specified two values: these were 0 and 0.50. Let us now consider 51 values: 0, 0.01, 0.02, ···, 0.50. There is no expectation that anyone performing substantive research would use these value. Rather, the purpose is to move us closer to the continuous case, where a traditionalist would maximize and a Bayesian would integrate.

**Case 5**

$H_0$: $P(H) = 0$ or $P(H) = 0.01$ or $P(H) = 0.02$, ···, or $P(H) = 0.50$

$H_1$: $P(H) \neq 0$ and $P(H) \neq 0.01$ and $P(H) \neq 0.02$, ···, and $P(H) \neq 0.50$

**Syllogism 5**

| | |
|---|---|
| $H_0$ or $H_1$ | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

Case 5 is similar to Case 4 except that 51 values are specified rather than two values. Using such a spread allows us to dramatize the difference between the two philosophical perspectives pertaining to whether hypotheses about population values can have probabilities.

Let us generalize Equation (5) to the case where there are 51 values rather than two, whose union qualifies under the null hypothesis. Generalization renders Equation (6):

$$\begin{aligned} & P\left(A \mid \left(C \cup D \cup \cdots \cup \text{fifty-first letter}\right)\right) \\ & = P\left(A \mid C\right) \cdot \frac{P\left(C\right)}{P\left(C\right)+P\left(D\right)+\cdots+P\left(\text{fifty-first letter}\right)} \\ & \quad + P\left(A \mid D\right) \cdot \frac{P\left(D\right)}{P\left(C\right)+P\left(D\right)+\cdots+P\left(\text{fifty-first letter}\right)} + \cdots \\ & \quad + P\left(A \mid \text{fifty-first letter}\right) \cdot \frac{P\left(\text{fifty-first letter}\right)}{P\left(C\right)+P\left(D\right)+\cdots+P\left(\text{fifty-first letter}\right)}. \end{aligned} \tag{6}$$

We might ask what the probability is of getting 17 or more heads out of 20 trials given Equation (6). From the point of view that hypotheses about population values can have probabilities, we would need to compute the probability of 17 or more heads out of 20 tosses given each of the population values (0, 0.01, 0.02, ···, 0.50), which is not much of a challenge. But we also would need to assign probabilities to the population values (that is, we need to assign unconditional probabilities). This is much more of a challenge. For example, should we assign an equal probability to each of the 51 population parameters included in the null hypothesis? Should we assign larger probabilities to values near 0.5? Alternatively, as 0.25 is in the middle of the range from 0 to 0.5, should we assign larger probabilities to values near 0.25? In fact, we could assume a normal-like distribution of values so that values near 0.25 are more likely than values at either extreme (0 or 0.5)[4]. Although there is no easy answer to this question, two points

[4]The distribution is "normal-like" rather than "normal" because it is not continuous, nor can the tails extend infinitely.

should be apparent. First, different decisions about assigning probabilities to population values will render different probabilities of getting 17 or more heads out of 20 tosses. Second, although it is possible to make decisions that would result in findings similar to the binomial calculation with which we commenced, it also is possible to make decisions that would result in findings that differ markedly from that obtained by the binomial calculation.

Or, we could resort again to the strategy of maximizing the calculated probability of obtaining 17 or more heads out of 20 tosses. In this case, we would assign a probability of 1 to the population value of 0.5, and a probability of 0 to all of the other population values. In this case, Equation (6) would reduce down to a single term that is equivalent to the binomial calculation with which we commenced—namely, the probability of 17 or more heads out of 20 tosses is 0.001.

## 6. The Continuous Null Hypothesis

Let us now return to Case 3 and Syllogism 3, copied below for the reader's convenience.

**Case 3**

$H_0$: $P(H) \leq 0.5$

$H_1$: $P(H) > 0.5$

**Syllogism 3**

| | |
|---|---|
| $H_0$ or $H_1$ | {Premise 1} |
| Not $H_0$ | {Premise 2} |
| Therefore, $H_1$ | {Conclusion} |

In the combination of Case 3 and Syllogism 3, we have that which truly represents the thinking in one-tailed tests, as opposed to merely approximating as in Combination 4 and Combination 5.

In the combination of Case 3 and Syllogism 3, we have a continuous range, with an infinite number of points contained within the range from 0 to 0.5. Nevertheless, the issues that were raised still apply. From the philosophical point of view that hypothesized population values can have probabilities, the question in this continuous case is: What is the density distribution that the researcher should assign to the range going from 0 to 0.5? Here, if one desired, one could apply a uniform distribution, an approximation of the normal distribution, a triangular distribution, or many others. Further, the researcher might wish to define the peak, if there is one, at 0.5 but might instead choose 0.25, or even 0, with the choice being influenced by the shape of the distribution one wishes to assume. To find the probability of the finding given the range null hypothesis, the researcher would have to integrate across the range of values for the assumed distribution. The philosophical weak point, perhaps, is that it is difficult to know what distribution to use and also, if the chosen distribution has a peak, it may be difficult to know where that peak should be.

Or, we can be traditional, and again assign a prior probability of 1 to the population value of 0.5 and a probability of 0 to everything else. As usual, this results in a binomial calculation for the probability of 17 or more heads out of 20 tosses

as being 0.001[5]. If a researcher decided to assign probabilities other than 0 or 1 to hypotheses, there are many ways of integrating across the range from 0 to 0.5, depending on the assumed distribution, that would result in values for the probability of 17 heads out of 20 tosses that differ markedly from each other, and also from the strict binomial calculation based on a probability of 1 for the 0.5 population value.

## 7. Discussion

Having taken considerable trouble to mark out the issues, does it make sense to compute the probabilities of findings given range null hypotheses? As we will see below, there is more than one way to think about it.

### 7.1. Three Perspectives on Probabilities of Findings Given Hypotheses

Calculations of probabilities of findings given hypotheses play a role in three schemes: Bayesian [31] [32] [33] [34], Fisher [10] [35] [36], and Neyman-Pearson [3] [4]. These are considered below, in turn.

For many researchers, the goal of research is to come to probabilities of hypotheses. From the standard frequentist point of view, as explained earlier, hypotheses are true or false but do not have probabilities (other than zero or one). But if one is a Bayesian, then hypotheses have probabilities. In turn, if the ultimate goal is to assign probabilities to hypotheses (but informed by the data), there is no need to go through the formalism of accepting or rejecting them, and so there is no point in using any of the foregoing syllogisms. Rather, the exercise of determining probabilities of findings is a preliminary step to the eventual use of the famous theorem by Bayes to determine the probabilities of hypotheses. Thus, there is no reason to perform either one-tailed or two-tailed significance tests because statistical significance is irrelevant. Importantly for the issue of conditional probabilities of findings, a Bayesian would assume a prior density distribution of unconditional probabilities to aid in calculating the conditional probability of a finding, given a hypothesis, rather than assigning a probability of 1 to the fair coin value of 0.5. From a Bayesian point of view, then, the usual calculation of the probability of the finding given a range null hypothesis is blatantly wrong because of the failure to assign unconditional probabilities to all of the values in the range specified by the null hypothesis. Let me reiterate for emphasis. From a Bayesian point of view, the calculation of the probability of the finding given a range null hypothesis is wrong. This is not just a matter of the frequentist being too conservative, but rather a matter of plain mathematical wrongness.

According to Fisher [10] [35] [36], a *p*-value is useful as a preliminary indication of the strength of the evidence, but it is not the most important or decisive factor in interpreting findings. This point of view can be interpreted as being

[5]Because the dependent variable concerns coin tosses, which are binary, the *p*-value is computed directly, without the need for a test statistic. If the dependent variable were continuous, a test statistic, such as a *t*-value, would be necessary.

consistent with the recent American Statistical Association (ASA) statement that *p*-values should be used in concert with other information, rather than being the sole piece of information used to decide whether or not to reject hypotheses [37]. But a researcher who wishes to use a *p*-value as a preliminary indication of the strength of the empirical evidence would not undergo the exercise of constructing a syllogism to reject or accept hypotheses. If the *p*-value is not used to accept or reject hypotheses, the worry explored earlier, about implausible null hypotheses, is no longer relevant. That is, because the researcher is not accepting or rejecting a null hypothesis, there need not be concern pertaining to whether or not one is rejecting an implausible null hypothesis. Three conclusions are consistent with this point of view. First, a point null hypothesis is fine because it does not have to be plausible. It only has to be useful in coming up with a *p*-value that, in turn, is merely a preliminary way to assess the strength of the empirical evidence. Second, because the purpose of the *p*-value is merely to help the researcher form a preliminary assessment of the strength of the empirical evidence, its importance should be much less than it typically is taken as having, by researchers, journal reviewers, and journal editors. Third, if the computed *p*-value is to be used for the purpose of a preliminary assessment of the strength of the empirical evidence, it seems desirable to have as precise a value as possible. Therefore, the use of a range null hypothesis, along with the strategy of overestimating the *p*-value by an amount that is impossible to determine, is far from being the best possible practice. With all of this having been said, however, it is worth noting that Vieland and Hodge have shown that no existing statistical procedure validly indexes the state of the empirical evidence, and this includes *p*-values, whether they are one-tailed or two-tailed[6] [38].

The probability of the null hypothesis, given the finding, is not the same as the probability of the finding given the null hypothesis (Trafimow & Marks, 2015). From a frequentist point of view, the null hypothesis does not have a probability (other than zero or one) and even if one is not a frequentist, it should be obvious that the probability of the hypothesis given the finding is not the same as the probability of the finding given the hypothesis. Nevertheless, based on Neyman and Pearson [3] [4], it is possible to use the concept of Type I error to draw conclusions. A Type I error is when the researcher wrongly concludes that the null hypothesis is false. By using $p < 0.05$ as a cutoff for rejecting or failing to reject null hypotheses, the researcher can be assured of wrongly rejecting the null hypothesis only 0.05 of the time. From the point of view of this cutoff strategy, the exact value of $p$ is irrelevant; what matters is simply whether $p < 0.05$ or not. Here the combination of Case 3 and Syllogism 3, along with the assignment of 1 to the population probability of the coin being 0.50, can be argued to be sensible. By using the "fair value" of 0.50 to substitute for the whole range of values between 0 and 0.5, the researcher can be assured that if the obtained conditional probability of the finding given the 0.5 value is less than 0.05, the conditional

[6]Bayes' factors also do not satisfy the Vieland and Hodge criteria for indexing the state of the evidence. [38]

probability also would be less than 0.05 no matter how unconditional probabilities might be assigned to values between 0 and 0.50.

From the point of view of controlling Type I error, we have seen that assigning an unconditional probability of 1 to the 0.5 value serves to provide the most conservative way to arrive at the probability of the finding given the null hypothesis. That is, it provides the largest possible probability and thus the smallest chance of meeting the required 0.05 criterion. But from another point of view, using a one-tailed test is not at all conservative. Consider again our original binomial calculation that the probability of 17 or more heads out of 20 tosses is 0.001. But if we were using a traditional point hypothesis with a two-tailed test, we would be interested in the probability of obtaining 17 or more heads out of 20 tosses, given that the coin is fair; and we also would be interested in the probability of obtaining 3 or fewer heads out of 20 tosses, given that the coin is fair. The probability of 17 or more heads out of 20 tosses, or 3 or fewer heads out of 20 tosses, is double the value that we calculated earlier. That is, the value is $0.001 \times 2 = 0.002$. Obviously, then, although using a one-tailed test is conservative from the point of view of assigning a prior probability of 1 to the maximum value contained in the range specified by the null hypothesis (*i.e.*, 0.5), it is extremely liberal relative to the usual two-tailed test with a point null hypothesis that $P(H) = 0.5$.

Well then, from the first and second points of view above, it is silly to engage in null hypothesis significance testing in the first place. That is, the goal either should be to come to a conclusion about the probability of the hypothesis given the finding, in which case one assumes that it is reasonable to assign probabilities to hypotheses; or to form a preliminary assessment of the strength of the evidence. In the former case, to calculate the conditional probability of the finding, one needs to assign unconditional probabilities of hypotheses to carry the calculation through. From this perspective, the usual one-tailed calculation gives blatantly wrong answers (again, not just conservative answers but wrong answers). And from the perspective of using the calculation as a preliminary assessment of the state of the evidence, a precise value is desirable, and the usual one-tailed calculation clearly is not precise. Consequently, it is only if one wishes to control Type I error that (a) significance testing makes sense and (b) one-tailed tests make sense to avoid the problem of rejecting an implausible null hypothesis. However, Trafimow and Earp have suggested that this point of view also is problematic. [39] Specifically, these researchers questioned the domain over which one wishes to exercise the control of Type I error. Is it all of science, all of psychology, a researcher's lifetime, a substantive domain, a single experiment, or a single hypothesis. No matter what domain one chooses, Trafimow and Earp show that important problems result that far outweigh the gain researchers enjoy by controlling Type I error. And even aside from the domain issue, there are two other issues. First, no less an authoritative body than the American Statistical Association has come out with a statement indicating that scientists should consider a variety of factors, rather than just using *p*-values

[37]. Clearly, if one is to use a cutoff, as is necessitated by the insistence on controlling Type I error, there is no room left to consider other factors. Either the computed conditional probability of the finding, given the null hypothesis, is under or over the cutoff and that is it! Second, few philosophers of science would agree that it is the job of scientists to reject or not reject hypotheses after obtaining single findings. Rather, scientists are supposed to propose and test larger theories, and each individual hypothesis is part of a larger network of theoretical assumptions, auxiliary assumptions, substantive hypotheses, and statistical hypotheses [12] [13] [19] [40]. From this perspective, a strong focus on controlling Type I error, and the cutoff strategy that goes with it, seems philosophically naïve.

### 7.2. Conclusions

We have seen that, depending on one's perspective, the traditional calculation for conditional probabilities of findings given range null hypotheses via maximization at the largest value in the range, is either blatantly wrong, quite imprecise, a conservative overestimate of the actual probability of the finding (in which case the risk of Type II error is quite large), or a quite liberal underestimate of the actual probability of the finding (relative to two-tailed calculations with point hypotheses). I underscore that these contradictory assessments pertain to evaluating probabilities of findings given hypotheses. This contrasts with the usual demonstration that different points of view give contradictory assessments about how researchers should evaluate hypotheses given findings. To my knowledge, this is the first demonstration to emphasize that these contradictions occur at the level of data evaluation rather than just at the level of hypothesis evaluation.

It is interesting that if researchers only used point null hypotheses, although different philosophical perspectives would still demand differences in hypothesis evaluation, at least the calculations would not differ pertaining to data evaluation. That is, for example, the probability of 17 or more heads out of 20 tosses, given a fair coin, would be calculated the same way by everybody and in accordance with the binomial theorem as we saw earlier (probability = 0.001). Thus, data evaluation, at least, would not be controversial, though hypothesis evaluation would remain controversial. But matters change when range null hypotheses are used, which places researchers in the Neyman-Pearson tradition in a dilemma. On the one hand, they can use point null hypotheses, where the calculations pertaining to data evaluation are not controversial, but at the cost of rejecting null hypotheses that are not plausible anyway. Or, they can use range null hypotheses that have better plausibility; but where the calculation of the probability of the obtained findings, given the null hypothesis, is potentially quite problematic. More generally, whatever the philosophical perspective, range null hypotheses are problematic even from a data evaluation point of view prior to hypothesis evaluation. Unfortunately, it is not obvious how to defend the computational technique chosen to calculate the probability of the obtained finding

given a range null hypothesis. A balanced assessment might be that all of them stand on rickety foundations.

## References


[1] Meehl, P.E. (1967) Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, **34**, 103-115. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.8918&rep=rep1&type=pdf https://doi.org/10.1086/288135

[2] Leventhal, L. (1999) Answering Two Criticisms of Hypothesis Testing. *Psychological Reports*, **85**, 3-18. https://doi.org/10.2466/pr0.1999.85.1.3

[3] Neyman, J. and Pearson, E.S. (1928) On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika*, **20A**, 175-240.

[4] Neyman, J. and Pearson, E.S. (1933) The Testing of Statistical Hypotheses in Relation to Probabilities a Priori. *Proceedings of the Cambridge Philosophical Society*, **29**, 492-510. https://doi.org/10.1017/S030500410001152X

[5] Etz, A. and Vandekerckhove, J. (2016) A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE*, **11**, e0149794.

[6] Bakan, D. (1966) The Test of Significance in Psychological Research. *Psychological Bulletin*, **66**, 423-437. http://www.tc.umn.edu/~nydic001/docs/teaching/Fall2011_PSY3801H/readings/Readings%20-%2003Bakan%201966.pdf https://doi.org/10.1037/h0020412

[7] Carver, R.P. (1978) The Case against Statistical Significance Testing. *Harvard Educational Review*, **48**, 378-399. http://healthyinfluence.com/wordpress/wp-content/uploads/2015/04/Carver-SSD-1978.pdf https://doi.org/10.17763/haer.48.3.t490261645281841

[8] Carver, R.P. (1993) The Case against Statistical Significance Testing, Revisited. *The Journal of Experimental Education*, **61**, 287-292. http://www.jstor.org/stable/20152382 https://doi.org/10.1080/00220973.1993.10806591

[9] Cohen, J. (1994) The Earth Is Round ($p < 0.05$). *American Psychologist*, **49**, 997-1003. http://qpsy.snu.ac.kr/teaching/gradstat/Cohen.pdf https://doi.org/10.1037/0003-066X.49.12.997

[10] Fisher, R.A. (1973) Statistical Methods and Scientific Inference. 3rd Edition, Hafner Press., New York.

[11] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf https://doi.org/10.1080/01621459.1995.10476572

[12] Meehl, P.E. (1978) Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, **46**, 806-834. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.200.7648&rep=rep1&type=pdf https://doi.org/10.1037/0022-006X.46.4.806

[13] Meehl, P.E. (1990) Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant Using It. *Psychological Inquiry*, **1**, 108-

141.
https://pdfs.semanticscholar.org/2a38/1d2b9ae7e7905a907ad42ab3b7e2d3480423.pdf
https://doi.org/10.1207/s15327965pli0102_1

[14] Meehl, P.E. (1997) The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In: Harlow, L., Mulaik, S.A. and Steiger, J.H., Eds., *What If There Were No Significance Tests*? Erlbaum, Mahwah, NJ, 393-425.

[15] Rozeboom, W.W. (1960) The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin*, **57**, 416-428.
http://www.ufrgs.br/psico-laboratorio/textos_classicos_9.pdf
https://doi.org/10.1037/h0042040

[16] Rozeboom, W.W. (1997) Good Science Is Abductive, Not Hypothetico-Deductive. In: Harlow, L., Mulaik, S.A. and Steiger, J.H., Eds., *What If There Were No Significance Tests*? Erlbaum, Mahwah, NJ, 335-391.

[17] Schmidt, F.L. (1996) Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers. *Psychological Methods*, **1**, 115-129. http://qpsy.snu.ac.kr/teaching/stat_dir/Schmidt(1996).pdf
https://doi.org/10.1037/1082-989X.1.2.115

[18] Schmidt, F.L. and Hunter, J.E. (1997) Eight Objections to the Discontinuation of Significance Testing in the Analysis of Research Data. In: Harlow, L., Mulaik, S.A. and Steiger, J.H., Eds., *What If There Were No Significance Tests*? Erlbaum, Mahwah, NJ, 37-64.

[19] Trafimow, D. (2003) Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights from Bayes's Theorem. *Psychological Review*, **110**, 526-535.
https://doi.org/10.1037/0033-295X.110.3.526

[20] Trafimow, D. (2006) Using Epistemic Ratios to Evaluate Hypotheses: An Imprecision Penalty for Imprecise Hypotheses. *Genetic, Social, and General Psychology Monographs*, **132**, 431-462. https://doi.org/10.3200/MONO.132.4.431-462

[21] Trafimow, D. and Marks, M. (2015) Editorial. *Basic and Applied Social Psychology*, **37**, 1-2. https://doi.org/10.1080/01973533.2015.1012991

[22] Trafimow, D. and Marks, M. (2016) Editorial. *Basic and Applied Social Psychology*, **38**, 1-2. https://doi.org/10.1080/01973533.2016.1141030

[23] Valentine, J.C., Aloe, A.M. and Lau, T.S. (2015) Life after NHST: How to Describe Your Data without "*p*-ing" Everywhere. *Basic and Applied Social Psychology*, **37**, 260-273. https://doi.org/10.1080/01973533.2015.1060240

[24] Frick, R.W. (1995) Accepting the Null Hypothesis. *Memory & Cognition*, **23**, 132-138. https://doi.org/10.3758/BF03210562

[25] Hagen, R.L. (1997) In Praise of the Null Hypothesis Significance Test. *American Psychologist*, **52**, 15-24. https://doi.org/10.1037/0003-066X.52.1.15

[26] Greenwald, A.G. (1975) Consequences of Prejudice against the Null Hypothesis. *Psychological Bulletin*, **82**, 1-20. https://doi.org/10.1037/h0076157

[27] Serlin, R.C. and Lapsley, D.K. (1985) Rationality in Psychological Research. *American Psychologist*, **40**, 73-83.
https://pdfs.semanticscholar.org/0ecb/48d1ad3747b4dd78ddcf2c8fd1f546481aa4.pdf
https://doi.org/10.1037/0003-066X.40.1.73

[28] Serlin, R.C. and Lapsley, D.K. (1993) Rational Appraisal of Psychological Research and the Good-Enough Principle. In: Keren, G. and Lewis, C., Eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Erlbaum, Hillsdale,

NJ, 199-228.

[29] Hays, W.L. (1994) Statistics. 5th Edition, Harcourt Brace College Publishers, Fort Worth, TX.

[30] Kirk, R.E. (1984) Elementary Statistics. 2nd Edition, Brooks/Cole Publishing Company, Belmont, CA,

[31] Howson, C. (1990) Fitting Your Theory to the Facts: Probably Not Such a Bad Thing after All. In: Savage, C.W., Ed., *Minnesota Studies in the Philosophy of Science*, Vol. 14, University of Minnesota Press, Minneapolis, 224-244.

[32] Howson, C. and Urbach, P. (1989) Scientific Reasoning: The Bayesian Approach. Open Court, La Salle, IL.

[33] Howson, C. and Urbach, P. (1994) Probability, Uncertainty and the Practice of Statistics. In: Wright, G. and Ayton, P., Eds., *Subjective Probability*, Wiley, Chichester, England, 39-51.

[34] Kruschke, J.K. (2011) Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press, USA.

[35] Fisher, R.A. (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, Scotland.

[36] Fisher, R.A. (1930) Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, **26**, 528-538. https://doi.org/10.1017/S0305004100016297

[37] Wasserstein, R.L. and Lazar, N.A. (2016) The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, **70**, 129-133. https://doi.org/10.1080/00031305.2016.1154108

[38] Vieland, V.J. and Hodge, S.E. (2011) Measurement of Evidence and Evidence of Measurement. *Statistical Applications in Genetics and Molecular Biology*, **10**, Article 35. https://doi.org/10.2202/1544-6115.1682

[39] Trafimow, D. and Earp, B.D. (2017) Null Hypothesis Significance Testing and the Use of P Values to Control the Type I Error Rate: The Domain Problem. *New Ideas in Psychology*, **45**, 19-27. https://doi.org/10.1016/j.newideapsych.2017.01.002

[40] Lakatos, I. (1978) The Methodology of Scientific Research Programmes. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511621123