Cambridge University Press
978-1-108-73071-6 — Hypothesis Testing Reconsidered
Gregory Francis
Excerpt
More Information

1 Introduction

In many scientific fields, there is no better start to a results section than, "As predicted, we found a significant difference between . . . " Finding a significant difference (e.g., $p < 0.05$) allows authors to affirm their beliefs about, for example, color perception, attention, the workings of visual circuits, or how people search for targets in a cluttered display. Many scientists learned the basics of hypothesis testing as undergraduate students, and they learned to deal with more complicated tests (e.g., multi-way ANOVA, ANCOVA, mediation, moderation) as graduate students. Statistical analyses are central to modern investigations of psychology, including perception, and hypothesis testing is a common approach to statistical analysis.

Despite playing a central role, many properties of hypothesis tests are misunderstood. These misunderstandings can lead to scientific articles that make no sense and to experiments that are so poorly conceived that it was never appropriate to run them. Over the past seven years, psychology has experienced a "replication crisis," whereby some important findings do not hold up when independent scientists repeat the experiment; much of the crisis seems to be related to inappropriate uses of hypothesis testing.

With this issue in the background, it might be useful to characterize some confusions about hypothesis testing and to describe its assumptions and limitations. Throughout this Element, we provide examples of how the issues impact the design and interpretation of perception studies. This discussion is not meant to be a critique of hypothesis testing itself; although after considering all the challenges, you may decide that hypothesis testing is not worth the effort. Alternative approaches include a focus on estimation (Cumming, 2014), Bayesian methods (Kruschke, 2010; McElreath, 2016), and information criterion methods (Burnham & Anderson, 2002), but they are not discussed here.

The target audience for this Element is someone who has already taken one (or more) statistics courses and uses hypothesis testing. The discussion requires little explicit mathematics (and there are no theorems!), but a general understanding of sampling distributions, p-values, and power is probably going to be necessary for the reader to follow all the arguments. The selected topics represent issues that have been raised over the past few years in discussions with colleagues and students. Readers may be disappointed to discover that the text sometimes identifies problems without proposing solutions, but it may be useful to discover that there remain unsolved problems in the use of hypothesis testing. Indeed, an overall theme of the Element is that the proper use of hypothesis testing is rather more complicated than generally believed. While the basic idea is simple and appealing, the actual use is often quite complicated, and some common practices undermine the tenets of hypothesis testing.

2 The Basics of Hypothesis Testing

Hypothesis testing offers an appealing approach to data analysis. Follow the rules and you will make a Type I error (conclude there is an effect when there really is no effect) only 5% of the time (or whatever criterion you set). Such Type I error control sounds really good because it aligns with the natural skepticism of a scientist who doubts an effect exists unless there is sufficient reason to believe otherwise.

Hypothesis testing is also pretty easy to apply. We create a quantitative null hypothesis that indicates "no effect" (e.g., population means equal each other across two conditions) and then predict properties of our data set if that null hypothesis is true. A fundamental concept here is the sampling distribution, which describes how common it should be to find various values of a sample statistic if the null hypothesis is true. The test essentially checks whether the statistic computed from the observed data is among the "rare" statistics in the sampling distribution by computing the probability that the observed data or something even more extreme would occur. This probability is the p-value. See Figure 1.

The details get more complicated for other analyses, but the basic reasoning is the same as that given earlier. Assume the null is true and estimate the probability of the observed (or more extreme) statistic under that assumption. If the probability is low (e.g., less than 0.05), reject the null hypothesis: conclude statistical significance. By definition, if everything is done properly, you should only make a Type I error (reject the null hypothesis when it is actually true) at your criterion rate (e.g., 0.05).

A key part of that last sentence is "if everything is done properly." Lots of things can go wrong when doing hypothesis testing, even when scientists are operating with the best of intentions. As we will see in the following sections, even seemingly small deviations from the proper procedures for hypothesis testing can cause the Type I error rate to be much larger than intended.

2.1 An Example from Perception

The stimuli in Figure 2a show the Muller–Lyer illusion: the horizontal line with outward wings appears to be longer than the horizontal line with inward wings. To measure the size of the illusion, n=310 observers adjusted the length of a line with wings so that it appeared to be the same length as a comparison line of 100 pixels long with no wings. See the Appendix for details on how to get the data set. Each observer made eight matches for the inward wing and outward wing

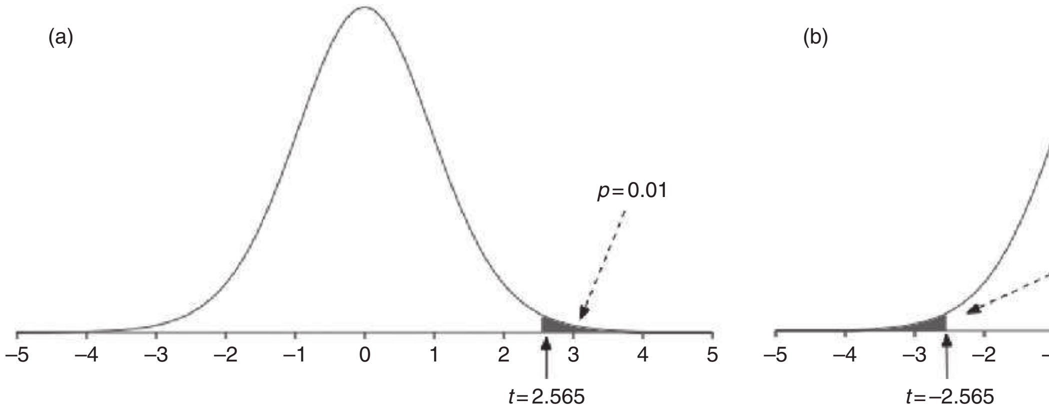


Figure 1 The *p*-value is the area under the curve of the sampling distribution beyond the obse distribution is for the *t*-value statistic that compares two sample means. (a) For a positive one-tai observed direction. (b) For a two-tailed test, the area is more extreme than the obs

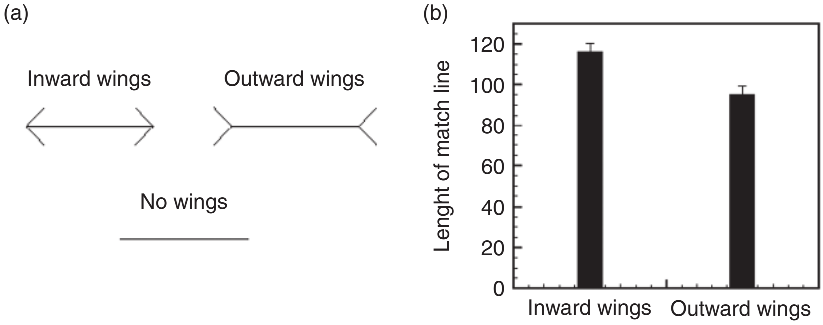


Figure 2 Stimuli and summary data for an experiment on the Muller–Lyer illusion. (a) A line with outward wings looks longer than a line with inward wings. (b) Mean line lengths for lines with the inward or outward wings so that they appeared to be the length of a 100-pixel line with no wings. The error bars indicate the standard deviation.

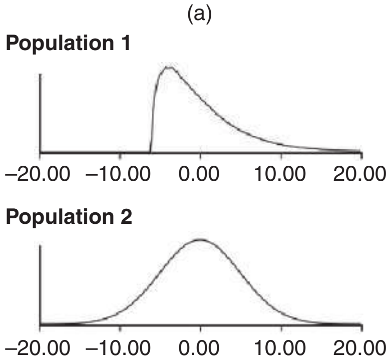
conditions, and the observer's score was the mean match length across the comparisons for each condition. Figure 2b plots the average match length across the 310 observers for each wing type. As expected, the match length is smaller than 100 pixels when the line has outward wings (a 92-pixel-long line with outward wings looks to be 100 pixels long). Likewise, the match length is longer than 100 pixels when the line has inward wings (a 112-pixel-long line with inward wings looks to be 100 pixels long).

A dependent two-sample hypothesis test comparing the means for the two wing conditions requires the sample size (in this case n=310) and computation of the sample means, standard deviations, and correlation of subject scores across the conditions ($\overline{X}_{\text{Inward}} = 112.3$, $s_{\text{Inward}} = 8.1$, $\overline{X}_{\text{Outward}} = 91.5$, $s_{\text{Outward}} = 8.0$, $r = 0.522$). With this information, the standard deviation of the difference of paired scores is computed to be $s_{\text{Difference}} = 7.87$ and the test statistic is t=46.6 with df=309, which corresponds to p<0.001. If there were truly no difference in the mean line lengths for the population of observers, then a random sample of 310 observers that produced a t-value test statistic at least as large as what we observed would be extremely rare. In practice, we say that the observed difference is "significant."

Take away message: When done properly, hypothesis testing controls the Type I error rate and the calculations are fairly easy to perform.

3 Robustness of the Two-sample *t*-test

A canonical hypothesis test is the two-sample *t*-test that compares two independent means. Our undergraduate classes told us that the *t*-test requires two



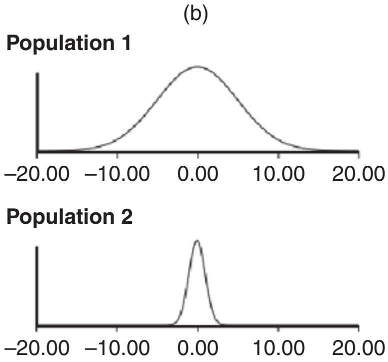


Figure 3 Exploring robustness of the *t*-test for two independent sample means. Here, every population distribution has a mean of zero. (a) Although the *t*-test assumes normal population distributions, even very skewed population distributions do not cause severe problems. For these populations the Type I error rate is 0.051. (b) Normal population distributions with unequal standard deviations. Here, the Type I error rate can be very different from the intended 0.05.

assumptions: the population distributions are normally distributed and the population standard deviations are the same. The mathematical theorems about Type I error rates no longer hold if the population distributions are non-normal, but in practice it matters only a little bit. For example, the distribution for population 1 in Figure 3a is strongly skewed, while the distribution for population 2 is a normal distribution; but both distributions have the same mean value (0), so a test of population means is for a true null hypothesis. Out of 10,000 simulated *t*-tests based on samples drawn from these distributions, the Type I error rate for the standard *t*-test is 0.051, which is just a bit above the intended 0.05. (See the Appendix for access to the simulation code.) In general, as long as the population distributions are unimodal and close to a normal distribution, the Type I error rate will be close to the intended value.

As long as the samples drawn from each population are of equal sizes, the *t*-test is also quite robust when the population standard deviations are different. In Figure 3b, population 1 has a standard deviation of 5, while population 2 has a standard deviation of 1. From 10,000 simulated *t*-tests with equal sample sizes (n_1=n_2=25), the Type I error rate is 0.059, which is only somewhat bigger than the intended 0.05.

In contrast to these situations, unequal standard deviations coupled with unequal sample sizes can be a disaster. If a large sample size (n_1=25 scores) is combined with the large standard deviation for population 1 and a smaller sample size (n_2=5 scores) is combined with the small standard deviation for population 2, then the Type I error rate is around zero (none of the 10,000 simulated *t*-tests rejects the null hypothesis). On the other hand, if the larger sample size is paired with the smaller standard deviation, then the Type I error rate is around 0.38, even when the 0.05 criterion is used to decide statistical significance.

The good news is that there is an easy solution to this problem. Welch's test is an alternative to the *t*-test that maintains the desired Type I error rate even when unequal standard deviations are paired with unequal sample sizes. In the cases presented earlier, Welch's test produces Type I error rates of 0.046 and 0.05, respectively. Welch's test is not perfect; for example, if the population standard deviations are equal but the sample sizes are different, a Type I error rate of around 0.06 is produced. Nevertheless, it avoids the really egregious cases that can occur for the standard *t*-test.

Take away message: The *t*-test is quite robust to deviations from some of its assumptions; but if you have unequal sample sizes, you should use Welch's test rather than a standard *t*-test.

4 Adding Data Increases the Type I Error Rate: Optional Stopping

A not uncommon situation is that after gathering some initial data, your analysis produces a promising but not significant result (e.g., p=0.08). Some people describe such a result as a "marginal effect" and move on, but that feels unsatisfying since the whole point of your experiment was to test for the effect (and it is not clear what "marginal" means anyhow). What some scientists do is add more subjects to the data set and rerun the analysis. That approach is problematic because when you make a final decision, you have given yourself two chances to reject the null hypothesis. Since the first decision (assuming everything else is appropriate) had a 5% chance of making a Type I error, the second decision inflates the error rate. The amount of increase in Type I error depends on a variety of factors (notably the sizes of the first and added samples). Moreover, suppose after adding some subjects to the original data set, your analysis produces p=0.07. You face the same issue and may decide to add still more subjects to the data set. If you are willing to keep adding subjects, the probability of making a Type I error approaches 1.0!

The problem is actually worse than it seems because Type I error control in hypothesis testing is not a property of any individual test. Rather, it is a property of the *procedure* you use to make a final decision about whether an effect exists (e.g., your result is statistically significant). If your procedure has many decision points (e.g., you will add subjects before making your final decision if p=0.08, but not if p=0.3), then you have to consider all those decision points, whether or not you actually follow them in a given situation. Thus, if your first data set produces p=0.02 and you report a significant result as your final decision, then your Type I error rate may be much higher than your intended 0.05. The Type I error rate has to consider what you *would have done* with results different from what you observed. Thus, if you would have added subjects had the p-value been larger, then that fact has to be included when considering the Type I error rate of your procedure.

The more principled way of describing the problem is to flip it around and describe it as *optional stopping*. It is not the adding of subjects that is truly problematic; rather the problem comes from stopping data collection when you are satisfied with the outcome. What is the absolute upper limit of resources (e.g., sample size) you would commit to a study? In many situations, scientists pick a sample size to "start," but they know that they will run more subjects if necessary. Having possible stopping points along the way up to that absolute upper limit sample size must inflate the Type I error rate. Oftentimes, scientists do not know their absolute upper limit sample size, nor (until faced with the choice) do they know what they would do if they found p=0.07 on their third analysis check. Such scientists cannot know the Type I error rate for their hypothesis-testing procedure.

What to do? There are sequential sampling methods that let you specify stopping points in advance and still maintain a desired Type I error rate. A simple approach is called the composite open adaptive sequential test (COAST; Frick, 1998). Here you gather an initial data set and run a t-test. If the p-value is below 0.01, you stop and conclude that you found a significant result. If the p-value is above 0.36, you stop and conclude that you did not find a significant result. Otherwise, you add another score and repeat. This procedure has a Type I error rate of 0.05, and it tends to use fewer subjects than a traditional t-test where sample size has been identified by a power analysis. There are costs, of course; you cannot decide whether or not to use COAST *after* looking at your data set. For example, if your first sample produces 0.02, you cannot claim significance; instead, the COAST procedure requires you to keep adding subjects. Moreover, for a given sample size, sequential sampling approaches have (somewhat) lower power than the traditional t-test. Finally, COAST does not have an upper limit on the sample size. As a result, if data collection stops with a p-value between 0.01 and 0.36, the scientist would not conclude evidence for an effect, and so COAST has a Type I error rate a bit below the intended 0.05. Other sequential sampling approaches allow for upper limits on the sample size, but you must have the resources to generate such sample sizes, even though you are unlikely to use them. (You cannot say

Cambridge University Press
978-1-108-73071-6 — Hypothesis Testing Reconsidered
Gregory Francis
Excerpt
More Information

that you will run up to 250 subjects if you only have enough money to pay for 75 subjects.)

You might say that the solution to optional stopping is obvious: pick a sample size in advance and stick to it. That can work in some situations, but then what do you do when you get p=0.08? If you run another experiment with entirely new data, then you inflate the Type I error rate by having multiple chances to reject the null. Meta-analysis (pooling data across experiments) does not help either because it is just a variation of optional stopping; you would not have run the follow-up studies if the original study were sufficiently convincing (Ueno, Fastrich, & Murayama, 2016). Even worse, although you might have a fixed sample size in mind for your study, someone else might have a different maximum sample size in mind and use your study as a starting point for further investigation. These different analyses would have different procedures and therefore different Type I error rates, even if they reported the same results for the same samples.

Taken to an extreme, the fixed sample size requirement for hypothesis testing seems to suggest that each experiment can only be run once, that you have to specify the sample size in advance, and then you (and everyone else) have to accept the decision of that experiment. That extreme view seems rather ludicrous, but if you relax the fixed sample size requirement of hypothesis testing, then you lose control of the Type I error rate, which is the whole point of hypothesis testing. In some sense, this view emphasizes that science cannot be too closely tied to statistical analyses. Statistical analysis is a means of double-checking scientific reasoning, but it cannot do the reasoning itself.

4.1 An Example from Perception

Optional stopping causes problems in addition to an inflation of the Type I error rate. Consider the Muller–Lyer experiment that produced the results in Figure 2, but suppose that your research interest was the correlation across subjects of matching lengths for inward and outward wings. Using the entire data set (n=310), this correlation is r=0.52, which is significant (t_{308}=10.7, p<0.001). If instead of gathering all the data and then analyzing, you analyzed data from just the first 30 participants and then added data one participant at a time until finding a significant result, then you would stop after getting data from n=53 subjects, when (for this data set) the correlation is r=0.3, which just satisfies the significance criterion (t_{51}=2.25, p=0.03). Analyses with earlier data sets do not produce significant results. For example, with the first n=52 subjects, the correlation is r=0.24, which corresponds to t_{50}=1.78, p=0.08. Generally speaking, optional stopping tends to produce results that just satisfy the significance

Cambridge University Press
978-1-108-73071-6 — Hypothesis Testing Reconsidered
Gregory Francis
Excerpt
More Information

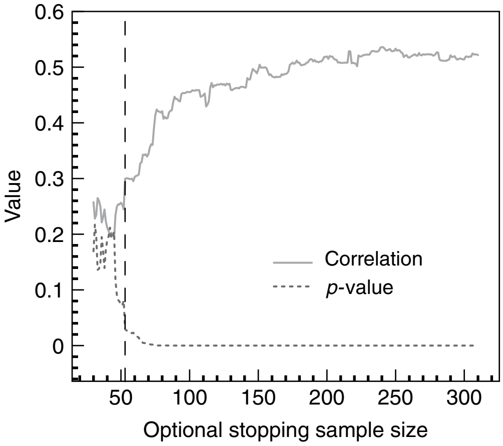


Figure 4 Using optional stopping for the Muller–Lyer data set from Figure 2 dramatically underestimates the correlation. The vertical line indicates the first time the updated sample produced a p-value less than the 0.05 criterion. Here, the sample correlation is small compared to what it would be with the full data set.

criterion. This means that the estimated effect size can overestimate or underestimate the true effect size, depending on its magnitude in the initial sample. For example, if the estimated effect happens to be small for the initial sample, then you rarely find strong effects because data collection stops before a strong result appears. Figure 4 demonstrates this property by plotting the sample r and p values generated by an optional stopping approach for the Muller–Lyer data set in Figure 2. The early samples happen to underestimate the correlation, and significance is found before the correlation is pulled toward the value calculated from the entire data set.

Take away message: Unless you are in a situation where you can fix the sample size, hypothesis testing does not necessarily do a good job controlling the Type I error rate. Unfortunately, it is difficult to avoid optional stopping.

5 ANOVA Can Be Extremely Conservative

Undergraduate statistics classes often introduce analysis of variance (ANOVA) as a way to resolve the multiple testing problem. If you have multiple tests (for example, to compare means against one another), then each test has a risk of making a Type I error and that risk accumulates, so that the probability of making at least one Type I error from the multiple comparisons is much larger than the intended 0.05 (or whatever rate you choose). ANOVA cleverly solves this problem by testing an omnibus null hypothesis (all means equal one another).

The cost of using an omnibus null hypothesis is that it does not indicate which means differ from other means. Thus, a significant ANOVA is usually followed up with additional tests to compare means (or groups of means) against one another. These additional tests have the feel of being the “dark arts” of hypothesis testing because they all seem a bit ad hoc. In many cases, these methods err on the side of being extra conservative.

For example, consider a situation where a scientist is testing search times for four visual maps. The scientist wants to compare her preferred map design to the other three designs. To convince other scientists that her design is better, she needs to show the following outcomes:

- A significant one-way ANOVA, which indicates that there is some difference among the map designs.
- A significant contrast of design 1 compared to design 2.
- A significant contrast of design 1 compared to design 3.
- A significant contrast of design 1 compared to design 4.

The three contrasts are necessary to conclude that her preferred design is better than each of the other designs. What is the Type I error rate for concluding that her preferred design is better than the other designs? Each hypothesis test has a Type I error rate of 0.05. But if all of the nulls are true and there really is no difference between any of the map designs, the Type I error rate of all four tests is around 0.003. It should intuitively make sense that requiring three significant contrasts *in addition to* a significant ANOVA has to reduce the Type I error rate. The reader can verify these calculations and create variations using the online ANOVA power calculator in Francis (2018) by setting up four levels, entering zero for each mean, and creating three appropriate contrasts.[1] Since all the population means are equal, the computed power for all tests will correspond to the Type I error rate.

Thus, if a scientist has specific comparisons in mind for drawing her conclusions, following standard analysis approaches may be setting up enormous statistical hurdles. Simulation studies using the power calculator in Francis (2018) show that a Type I error rate of just under 0.05 is generated across the full set of four tests if you set the significance criterion to be α=0.3 for each test. With such a criterion, each test has a fairly high risk of making a Type I error, but it is rather unlikely that *all* the tests will make a Type I error.

[1] https://introstatsonline.com/chapters/calculators/OneWayANOVAPower.shtml