Universiteit Leiden
The Netherlands

Optimal test statistics for anytime-valid hypothesis tests

Lardy, T.D.

1 | Introduction

> Statistical procedure and experimental design are only two different aspects of the same whole....
>
> *R. A. Fisher, The Design of Experiments*

At the core of science is a systematic process of inquiry, where observations lead to hypotheses, experiments provide evidence in favor of or against these hypotheses, and this in turn shapes our understanding of the world around us. To illustrate the complexities of this process, imagine that you frequently play table tennis with a friend. After some time, you get the feeling that you are not equally matched and decide to test this by playing a game against each other. A choice presents itself: one approach is to play until a fixed number of rallies has been played, another is to follow the official rules (first to 11, minimum of two-point difference), and a third option is to play until, for whatever reason, either of you has had enough. Whichever option is chosen, the outcome of the match will offer evidence in favor of or against the hypothesis that you are (not) equally matched. Once again, this requires deliberation. Would you consider winning by a single point to be evidence in favor of your hypothesis? Or should the difference in points be more extreme before you believe that you are not equally matched? In this analogy, the match represents an experiment and the questions about its outcome mirror the subsequent statistical analysis. The first match type listed above corresponds to experiments in which the number of data points to be collected is fixed in advance. The second option relates to experiments where the data collection plan is decided in advance, but the total number of data points collected is variable (you could win 11-0, or a closer game might end 12-10). Finally, the third variant corresponds to experiments without a predetermined sampling plan, that is, where experimentation may be stopped at any point in time. The analysis of data collected from this type of experiment lies at the heart of this thesis.

The remainder of this introduction serves to further outline this problem and contextualize the results that are given in the subsequent chapters. The structure is as follows: First, we introduce the general problem of statistical hypothesis testing, leading up to an explanation showing that experiments with a variable sample size require different statistical methodology than those with a fixed sample size. Next, we discuss one of the repercussions of not taking this into account properly. Following this (in Section 1.3), a review is provided of the first statistical methods that were designed to be used with a fixed data collection plan, but without a fixed sample size. Then, an introduction to the theory of hypothesis testing for experiments without any sampling plan is given. Finally, we give an overview of the content of this thesis.

1.1 Hypothesis Testing

A statistical hypothesis test is a data-driven method that is used to decide whether a hypothesis of interest, called the null hypothesis, can be falsified. The way in which hypothesis tests are used is comparable to a proof by contradiction. For example, suppose that you want to verify the claim that you and your friend are not equally good at table tennis. You would then take the null hypothesis to be the assumption that this claim is false, that is, that you are equally good. After collecting data (playing a match), a hypothesis test is used to determine whether the null hypothesis is still plausible in light of the results. If the outcome of the test is that the null hypothesis can be falsified, then the data are so incompatible with the null hypothesis that you should no longer believe it to be correct. This would happen if, for example, one of you wins by a landslide. You can then proceed as if your original claim were true, that is, as if you are not equally good at table tennis. This is commonly referred to as "rejecting" the null hypothesis. The other possibility is that the data do not give sufficient reason to doubt the null hypothesis (for example, if the match is very close), in which case you fail to reject the null hypothesis. This does not mean that the null hypothesis is true: absence of evidence is not evidence of absence. In this case, you can therefore not be sure about your original claim.

Of course, the conclusion drawn by a hypothesis test is useful only if it does not lead to too many mistakes. To quantify how often a hypothesis test is wrong, the null hypothesis is translated to a probabilistic model, which we denote by $\mathcal{H}_0$. This translation allows probabilistic statements about the data to be made. For example, in the table tennis example above, the null hypothesis postulates that you and your friend are equally good. One way of modeling this is by saying that both of you are

equally likely to win each point. If this model were actually true, then the probability that the difference in points is at least four in a match with a total of five points is equal to 0.375, or 37.5%. This follows because there are five ways in which the score could end up being 4-1 and only one way in which the score could be 5-0, and similarly for 1-4 and 0-5. Under $\mathcal{H}_0$, all of these scores have associated probability $\frac{1}{2}^5$, so the total probability of this happening is $12 \times \frac{1}{2}^5 = 0.375$.

Such statements allow us to only consider hypothesis tests such that, if $\mathcal{H}_0$ were actually true, the probability of seeing data for which the test would reject $\mathcal{H}_0$ is not too high. What qualifies as too high should be determined by the experimenters beforehand and is referred to as the significance level α; common values include $0.1, 0.05$ and 0.01. False rejections of the null hypothesis are also called type-I errors, so the above can be summarized as follows: the probability of making a type-I error should be smaller than α. If this is true for a certain hypothesis test, we say that the test is valid at the significance level α. Returning to the example, if a match with five rallies is played, one might define a hypothesis test by rejecting the null hypothesis whenever the difference in points is four or more. This test has a type-I error probability of 0.375 according to the above reasoning. That is, even if the null hypothesis were true, this test would reject it with a probability of 37.5%. Therefore, it is not a valid hypothesis test at any significance level $\alpha < 0.375$. To obtain a smaller type-I error probability, one can consider a test that only rejects the null hypothesis if the difference in points is more extreme than four. In fact, it is a standard result in statistics that, for a fixed sample size of n, rejecting when the difference in points exceeds $1.96 \times \sqrt{n}$ gives a type-I error probability of approximately 0.05. For $n = 5$, this amounts to rejecting the null hypothesis whenever the difference in points is greater than $1.96 \times \sqrt{5} \approx 4.4$, which only happens when the score is 5-0 in favor of either player.

The type-I error probability of this test is calculated under the assumption that the sample size is fixed beforehand and that the data are only analyzed once. Such methods will be referred to as fixed-sample methods. Now, suppose that the null hypothesis is true but that you have a certain incentive to reject it nonetheless. If the sample size n has not been fixed beforehand, you could adopt a technique known as optional stopping: continue to play—increase the sample size—until the difference in points first exceeds $1.96 \times \sqrt{n}$. It follows from a standard result in probability theory that this will surely happen infinitely often. Hence, you are certain to come to a point at which the difference in points exceeds $1.96 \times \sqrt{n}$. By stopping at such a point, the hypothesis test will suggest to reject the null hypothesis, even though it is true. The intuition is that, even if you and your friend are equally matched, one of you will at

some point reach a long enough lucky streak to be able to reject the null hypothesis. This immediately disqualifies the use of fixed-sample methods whenever n is not fixed in advance. The problem is neither that the method for data collection is invalid nor that the hypothesis test is incorrect. They are simply not compatible. The test is designed for a fixed sample size, while the experiment follows a sequential sampling plan. The latter would be valid if the rejection rule were adjusted accordingly. That is, serial data require different statistical methods from data collected in a batch.

1.2 Optional Stopping in Academia

One of the first documented cases of optional stopping was when researchers saw it unfold in the study of extrasensory perception, studies to prove that the sixth sense exists (see e.g Greenwood, 1938; Greenwood and Greville, 1939; Feller, 1940). It has since been the subject of continuous debate among statisticians and has, more critically, been referred to as "sampling to a foregone conclusion" (Anscombe, 1954). The problem is that researchers, either aware or unaware, might engage in optional stopping and still report that they have conducted a fixed-sample hypothesis test at some significance level α, while the true type-I error probability can be much higher. This results in frequent controversies stemming from statistical investigations in various fields. For example, an article published in 2018 in a *Nature* journal showed that activation of certain neurons in animals suppresses REM sleep (Weber et al., 2018). This conclusion was based on hypothesis tests designed for a fixed sample size, yet the authors state that "we continuously increased the number of animals until statistical significance was reached to support our conclusions." This sparked some debate on social media among statisticians (Barnett, 2018), but it only came to light because the authors were transparent regarding their methodology.

An example in which there was no such transparency was when a team of researchers supposedly showed the benefits of power posing (Carney et al., 2010). They reported that assuming a powerful position helped boost confidence, increase testosterone, and decrease cortisol. Five years later, another research group tried to replicate their findings with a larger sample size, and while they indeed found that power posing increased subjective feelings of power, they found no significant effect on hormonal levels (Ranehill et al., 2015). In light of this and other failed replication attempts (see e.g. Simmons and Simonsohn, 2017), the first author of the original article eventually disassociated herself from the claims in the article and mentioned optional stopping as one of her concerns: "We ran subjects in chunks and checked the effect along the way.

It was something like 25 subjects run, then 10, then 7, then 5" (Carney, 2016, fact 5). Unfortunately, these are not isolated incidents. In 2012, 56% of participants in a survey of more than 2000 psychologists admitted to "deciding whether to collect more data after looking to see whether the results were significant" (John et al., 2012).

This compromises the reliability of results derived from statistical analyses and, in turn, might cause a distrust of statistics in general. Feller (1940) eloquently puts it as follows: "...statistics makes claim to mathematical rigor, and still its practical applications are often disputed or rejected as absurd. I fear it may sometimes produce a feeling that mathematics is, after all, 'one of those rational and scientific paths, which is nothing more than a narrow, short, and dirty dead end, at the end of which one hits their nose ingloriously'" (second part translated from French). Rather than taking such a bleak point of view, statisticians have recently emphasized the need to develop tools that can accommodate optional stopping. That is, to focus on the development of methods that can handle serial data collection without losing type-I error control (see e.g. Ramdas et al., 2023).

1.3 Sequential Methods

The first developments of methods that were designed for experiments with a variable sample size were based on a simple observation: There are situations where one can reach the same conclusions as when using fixed-sample methods but with a strictly smaller sample size. For example, for a fixed sample size of $n = 20$, the table tennis hypothesis test described above would reject the null hypothesis if the difference in points is greater than or equal to 9. Now suppose that after 15 points the score is 15-0. At this point, the game might as well be stopped, since no matter how it plays out, the null hypothesis will always be rejected. That is, the sample size can be strictly reduced while making the exact same decisions (a similar realization is already present in early work on lot inspection by Dodge and Romig, 1929, p. 626).

This insight led to the idea that it might be beneficial to directly design hypothesis tests that incorporate serial data collection. The culmination was the development of a widely applicable sequential method: the sequential probability ratio test (SPRT) (Wald, 1947). The SPRT works by updating a single number, the probability ratio, that represents the evidence in favor of or against the null hypothesis after each collected data point. Based on this number, the SPRT prescribes whether to collect another data point or to stop and draw conclusions about the null hypothesis. In this way, the total number of data points that will be collected in an experiment is variable

and depends on the data that are observed. However, if one were to repeatedly follow this procedure, the average sample size that is needed to reach a conclusion is smaller than the average sample size that would be needed to draw similar conclusions using any other method (fixed-sample and sequential alike). That is, the SPRT achieves the optimal expected sample size.

This was a highly desirable feat because, in the intended applications, data collection was extremely costly or even destructive. Indeed, the SPRT was developed in the context of World War II and, as such, its applications included testing whether ordnance (artillery, ammunition, explosives, etc.) was faulty or not (see also Wallis, 1980). The caveat to this cost efficiency is that the probability of falsely rejecting the null hypothesis is only bounded by the prescribed significance level α if the SPRT's sampling plan is followed to a tee. If one deviates from it, then the type-I error probability might still be higher, so the SPRT does not fix the problems with optional stopping that were outlined in the previous section.

1.4 Anytime-Valid Inference

The problem of optional stopping was specifically addressed in work by, among others, Robbins (1952, 1970), Darling and Robbins (1968), and Lai (1976). They developed hypothesis testing methods for which the type-I error probability remained below the prespecified level of significance regardless of the data sampling rule that was employed. Their methods are now referred to as anytime-valid. Similarly to the SPRT, anytime-valid methods are based on a test statistic that can be thought of as a measure of evidence against the null hypothesis at a certain time. This test statistic is updated after observing each data point. If the total evidence is large enough at some point, the null hypothesis is rejected. The key difference from the SPRT is that anytime-valid methods do not prescribe whether or not to collect more data. Instead, this choice is left to the experimenters. In particular, they might choose to keep collecting data until the test statistic gives enough evidence to reject the null hypothesis. The probability that this will ever happen if the null hypothesis is actually true, is smaller than α. As such, anytime-valid methods are unaffected by optional stopping.

For a long time, anytime-valid methods made up only a minor segment of the area of sequential analysis. A shift occurred when interest in them took off in recent years, as demonstrated by the mere existence of this thesis. The key year for this shift was 2019, when at least four breakthrough articles on the subject were made available online (Wasserman et al. (2020); Shafer (2021); Vovk and Wang (2021); Grünwald

et al. (2024); all of these articles first appeared on ArXiv in 2019). In addition to these breakthroughs, the surge of interest may partly be caused by what is known as the replication crisis, which refers to the widespread phenomenon that many scientific studies cannot be reliably replicated or reproduced. One of the many facets of this problem is that the number of scientific conclusions based on false-positive results is much higher than one might theoretically expect. As argued in Section 1.1, this could be caused by practices such as optional stopping. Therefore, anytime-valid methods are sometimes presented as a partial solution to the replication crisis. Furthermore, anytime-valid methods might be more appealing now than they used to be because technological advances have transformed the way data is collected and stored (Cukier and Mayer-Schoenberger, 2013; Sagiroglu and Sinanc, 2013). Datasets used to be collected manually and stored on paper, film, or other analogue media with careful thought as to what was stored and what was not. Today, the standard in many fields—finance, commerce, and many more—has become storing data digitally, and as much of it as possible. Think of stock prices, the amount of time customers spend at a certain webshop, or atmospheric conditions at various points in time. These are all examples of data that are inherently serial, and digital records allow them to be accessed easily and at all times. Anytime-valid methods enable researchers to analyze these data in real time and draw conclusions accordingly without breaking the type-I error control, that is, without making too many mistakes.

1.5 Contributions of This Thesis

In the discussion of hypothesis testing so far, only one type of error has been considered: false rejection of the null hypothesis. If this were the only basis on which hypothesis tests are evaluated, then one should be content with never rejecting the null hypothesis, and the experiment might as well not have been performed. Therefore, to determine whether testing methods are actually useful, it is custom to consider a second type of mistake: failure to reject the null hypothesis when it is actually false. This is commonly referred to as a type-II error. For a fixed significance level α, the typical approach for fixed-sample methods is to find the hypothesis test that minimizes the probability of making a type-II error, among all methods with a type-I error probability below α. Equivalently, one can consider maximizing one minus the probability of making a type-II error, which is referred to as the power of the test. Which test has the most power will depend on the specific hypotheses under consideration and should be determined on a case-by-case basis.

For anytime-valid tests, the concept of power is ambiguous. Indeed, the advantage of anytime-valid tests lies in the fact that they can be used with an unknown data sampling rule. One might keep collecting data forever, in which case the null hypothesis will almost certainly be rejected if it is false, or one might stop after a single data point, making it highly unlikely to be rejected. Due to this ambiguity, it is customary to consider optimality criteria different from power for anytime-valid methods. To this end, Koolen and Grünwald (2022); Grünwald et al. (2024) propose to consider anytime-valid tests that are based on log-optimal test statistics. A formal definition of this concept, as well as a rigorous introduction to the theory of anytime-valid testing, is given in Chapter 2. As is the case for power, the criterion of log optimality gives an abstract notion of how optimal tests can be found. In practice, actually finding them is not a straightforward exercise. This thesis is concerned with the criterion of log-optimality and finding log-optimal test statistics for a variety of hypotheses. We now give a brief overview of all the chapters.

Log Optimality

Grünwald et al. (2024) show that, under certain conditions, there exists a log-optimal test statistic that takes the form of a likelihood ratio—an object that is well studied in the classical literature on testing. However, when these conditions are not met, the log-optimality criterion cannot fully differentiate between different test statistics. That is, there are settings for which a wide range of test statistics all seem equally good when judged by this criterion. One can even construct examples where a certain test statistic always provides more evidence against the null hypothesis than another, yet both statistics seem equally good according to the log-optimality criterion. Chapter 3 discusses a way to redefine the log-optimality criterion in such situations to avoid this. Optimal test statistics are shown to still take on the form of likelihood ratios.

Exponential Families

Many common statistical models, such as the normal, Bernoulli, and Poisson model, are instances of exponential families. The latter are collections of probabilistic models with a specific form that offers great mathematical convenience for a variety of objectives. As such, exponential families are used to model data by researchers in a wide range of scientific fields for many different applications.

An example is for k-sample tests, where k groups of data (or samples) are observed and the goal is to test whether they have the same distribution, under the assumption

that this distribution comes from a certain exponential family. In Chapter 4, three different test statistics for this problem are studied. The first is the log-optimal test statistic for a specific alternative, which, as it turns out, cannot always be computed efficiently. The second test statistic is generally suboptimal but can always be easily evaluated. Finally, the third test statistic is designed to be robust against misspecification of the exponential family. That is, it is still valid if the data have the same distribution in all groups but this distribution is not a member of the exponential family. For small effects, that is, if the distributions of the different streams under the alternative are not too different, it is shown that all three statistics give a surprisingly similar amount of evidence.

Although the third statistic mentioned is robust against misspecification of the exponential family, the first two are not. This means that if those statistics are used but the model is wrong, then the type-I error probability is not guaranteed to be below the desired significance level. Similar reliability issues can occur with many statistical tools when the underlying models are misspecified. It is therefore important to test whether the exponential family is well specified, that is, whether the data could indeed have been generated by a specific exponential family. This problem is studied in Chapter 5. Conditions are given under which log-optimal test statistics are easy to compute. We furthermore discuss a variety of exponential families for which these conditions hold, so that anytime-valid tests can be constructed to test whether they are well specified.

Model-X

An important task in many branches of science is to detect whether there is an association, or dependence, between a response and an explanatory variable. Consider, for example, testing whether a certain medication (explanatory variable) has an impact on a patient's health (response). Often, there are also other variables that could potentially impact the response (e.g. age) and therefore need to be controlled for. These variables are called covariates. To capture this in a hypothesis testing setting, one needs to construct a probabilistic model for the data. The difficulty with that is that there is often little prior information about what the explanatory variable and response will behave like. However, in specific cases, it is known how the explanatory variable should be modeled conditional on knowledge of the covariates. The assumption that this conditional distribution is known, is referred to as the model-X assumption. For example, in many clinical trials, the medication is administered to

patients in a randomized manner, irrespective of the covariates. Therefore, the distribution of the explanatory variable (whether the medication has been administered to a certain patient) conditional on the covariates (that patient's characteristics) is known: It is fully characterized by the randomization. Chapter 6 shows that, under the model-X assumption, it is possible to construct anytime-valid tests of independence without further assumptions on the way in which the data are generated. In particular, log-optimal test statistics for specific alternatives are derived.

Group Invariance

Transformations of the data often do not have any meaningful effect from a statistical point of view. For example, changing the units of some measurement from kilometers to miles should generally not impact the information that can be extracted from the data. It is therefore an accepted principle that statistical inferences should exhibit certain invariance properties. In particular, if the data are not impacted by a certain transformation, then the conclusions drawn from a hypothesis test should also not be. Tests that have this property are referred to as invariant, and similarly for the corresponding test statistics. Anytime-valid tests that are invariant are the subject of Chapter 8. In particular, conditions are given under which the log-optimal test statistic is invariant. That is, the conclusion that will be reached by the optimal test will not depend on irrelevant aspects of the data.

However, it is not always entirely clear which facets of the data are actually irrelevant, that is, which transformations have a meaningful impact on the data. For example, it is a common assumption that changing the order of the data points does not have a meaningful effect on any important aspects of the data. However, there are also situations where, for example, seasonal effects are lost by reordering the data. Carrying out analyses under the assumption that certain transformations do not impact the data, while in reality they do, might cause researchers to draw wrong conclusions. It is therefore also important to be able to test whether certain transformations have an effect on the data. General methodology to construct anytime-valid tests for this purpose is discussed in Chapter 7. Furthermore, conditions are shown under which the corresponding test statistics are log optimal.