LIKELIHOOD AND ITS EVIDENTIAL FRAMEWORK

Jeffrey D. Blume

1 INTRODUCTION

Statistics is the discipline responsible for the interpretation of data as scientific evidence. Not surprisingly, there is a broad statistical literature dealing with the interpretation of data as statistical evidence, the foundations of statistical inference, and the various statistical paradigms for measuring statistical inference. However, this literature is surprisingly diverse; the range of viewpoints, opinions and recommendations regarding methods for measuring statistical evidence is as varied as it is vast. This diversity is due, in part, to the complex philosophical nature of the problem. But it is also due to the absence of a generally accepted framework for characterizing and evaluating paradigms that purport to measure statistical evidence.

This paper proffers a general framework that enables the comparison and the evaluation of statistical paradigms claiming to measure the strength of statistical evidence in data. The framework is simple and general, consisting of only three key components (i.e., three key definitions). These essential components may, at first, appear obvious. For example, the first component is nothing more than a definition of the mathematical quantity used to measure the strength of evidence in data.[1] Unfortunately, the first component is always obvious and its (critical) definition is sometimes missing.[2] Once defined, however, the behavior of competing measures can be detailed and contrasted. As we will see, more than just a measure of evidence is needed to understand and evaluate a paradigm for measuring statistical evidence. The absence of a well defined framework can lead to controversies like those surrounding ad-hoc adjustments to p-values for multiple looks.[3] or for multiple comparisons.[4]

[1] For example, in classical frequentist inference this measure would be the p-value.

[2] For an example consider Bayesian inference. Is it the Bayes factor or the posterior probability that measures the strength of evidence in data?

[3] Reexamining accumulating data during the course of a study

[4] Evaluating several different scientific endpoints (e.g., overall survival, cause-specific survival, safety) in a single study.

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

1.1 The three evidential quantities

Three essential quantities for assessing and interpreting the strength of statistical evidence in data are

1. the measure of the strength of evidence,
2. the probability that a particular study design will generate misleading evidence,[5]
3. the probability that *observed* evidence is misleading.

For brevity I will sometimes denote these evidential quantities as EQ1, EQ2, and EQ3.[6] The first and the third quantities inform the statistical evaluation of data as scientific evidence. The second quantity informs the data collection process. All three quantities are essential to science and to statistics.[7]

Each evidential quantity represents the answer to a critical scientific question. The first provides an answer to the question, "How strong is the evidence in these data?" The second answers the question, "What is the chance that my study will yield data that are misleading?" The third answers the question, "What is the chance that these observed data are misleading?" The first and third quantities depend on, and pertain to, the observed data; they reflect characteristics of an existing set of data. The second quantity depends on, and pertains to, the study design; it does not inform the interpretation of data because it presupposes that data have not yet been collected.

Each quantity provides unique information regarding the interpretation (EQ1), collection (EQ2) and reliability (EQ3) of statistical evidence. In this paper, a statistical paradigm is said to have a well defined 'evidential framework' if that paradigm clearly defines, and distinguishes between, these three quantities. Once established, evidential frameworks are evaluated and contrasted by examining the statistical performance of these quantities in various scenarios.[8]

1.2 An analogy for the second and third evidential quantities

The second (EQ2) and third (EQ3) evidential quantities are routinely mistaken for one another. Once a set of data is actually collected, the probability of collecting

[5] Misleading as measured by EQ1. It helps to think of misleading evidence as strong evidence in support of a false hypothesis. This allows for the possibility of weak or inconclusive evidence in support of a false hypothesis, which is typically not considered misleading in scientific applications. I will be more precise in upcoming illustrations.

[6] I assert the importance of these quantities based on my experience and their immediate obviousness. Others have alluded to the same thing. For example, see Royall's [1997] preface, [Blume, 2002; 2007], and [Strug and Hodge, 2006a; 2006b] as examples.

[7] The natural temporal order in which these quantities are considered during scientific research is: EQ2, EQ1, EQ3.

[8] For example, the rate at which an EQ1 converges to the identification of the correct hypothesis might be a comparison metric between frameworks. The framework with the quicker rate of convergence would probably require less data and therefore be preferable.

some other set of data that turns out to be misleading (i.e., EQ2) is irrelevant. The observed data are either misleading or not, and we know not which. What is of interest is the potential for the observed data to be misleading (i.e., EQ3). After observing data, it makes perfect sense to ask "What is the probability that the data I just collected are misleading?" EQ2 characterizes the chance that the study will yield a misleading result. EQ3 characterizes the chance that the observed result is misleading.

A simple example may help illuminate the distinction. Jena and Jamie each play the lottery.[9] Jena buys one ticket. Jamie buys ten tickets, but one of the tickets is identical to Jena's ticket. If Jena wins, so does Jamie. But Jamie can also win with one of her other nine tickets. Therefore, Jamie's chance of winning is ten times that of Jena's, but still very small at 10 out of 195,249,054. Here the probability of winning is analogous to EQ2 (e.g., "What is the chance of winning with these tickets?").[10] Both Jena and Jamie have different chances of winning due to their different game playing strategies.

The next morning Jena and Jamie look in the newspaper for the winning lottery numbers. Unfortunately, the one of the numbers is smudged and unreadable.[11] The remaining numbers, however, match Jena's ticket. And because Jamie also has an identical ticket, both women may have a winning ticket. At this point, Jamie's and Jena's chance of winning is identical and is equal to 2.5%.[12] Here the (updated) probability of winning is analogous to EQ3 (e.g., "What is the chance of winning after matching all but one number?"). The fact that Jamie had a ten-fold chance to win the lottery yesterday is completely irrelevant. Her ticket is just as valid as Jena's, she stands to win just as much money, and she has the exact same chance of doing so. The insight is this: Once data are collected, EQ2 becomes irrelevant. What 'might have been' becomes irrelevant. It is EQ3, what 'might be', that is the relevant quantity once data are collected.

1.3 Absence of an evidential framework

The absence of a well defined evidential framework can lead to irresolvable controversies, such as those surrounding the proper use and interpretation of p-values or those concerning adjustments to p-values for multiple comparisons and multiple looks at data [Blume 2003, Goodman and Royall 1988, Goodman 1998, and Royall 1986, 1997]. I will revisit this later, but it suffices to note that the lack of clarity on which evidential quantity the p-value represents causes considerable confusion.[13]

[9] By lottery I mean Powerball. To win the jackpot, players must correctly match 5 white balls (numbered 1 through 59) and one red ball (numbered 1 to 39) that are drawn randomly without replacement. The order in which the white balls are drawn does not matter. The odds of winning the jackpot are 1 in 195,249,054.

[10] Here winning the lottery is akin to collecting misleading evidence and the tickets are akin to the study design.

[11] I assume that the red ball number is unreadable, but the details are ancillary to this analogy.

[12] Because only the red ball is in question, the chance of winning is now 1 in 39 or approximately 2.5%.

[13] The p-value is often interpreted as each of the three quantities, sometimes at the same time.

Bayesians are not immune to this criticism either, as it remains unclear which of the evidential quantities a posterior probability[14] is intended to represent.

The advantages of a well defined evidential framework will be illustrated by examining the Likelihood paradigm in the context of multiple examinations of data and multiple comparisons. Among the three[15] prominent statistical paradigms for measuring statistical evidence, only the Likelihood paradigm has a well developed evidential framework. In fact, the decoupling of the three evidential quantities allows Likelihood to obey the Likelihood principle[16] and retain good frequentist properties without having to use ad-hoc adjustments or prior probabilities. This is because its EQ1 may adhere to the Likelihood principle without forcing an adjustment to EQ2, which naturally depends on the study design.

Within a well defined evidential framework it is not a contradiction for two studies to yield identical data, and therefore equivalent statistical evidence, when each study initially had a different chance of generating misleading evidence (due to differing study designs). Moreover, we will see that neither set of *observed* data is more likely to be misleading than the other.[17] In terms of evidential quantities, EQ2 may differ between studies, but those studies may generate identical data such that EQ1 and EQ3 are themselves identical (e.g., see the above lottery example). The often noted controversial case of this is when two different studies, each with a different stopping rule, yield the exact same data.[18]

2 THE LIKELIHOOD PARADIGM

The Likelihood paradigm is based on the Law of Likelihood, which explains when the data represent statistical evidence for one hypothesis over another. Simply put, the data better support the hypothesis that does a better job of predicting the observed events and the likelihood ratio measures the degree to which one hypothesis is better supported over the other [Hacking, 1965; Edwards, 1971; Royall, 1997]. Likelihood ratios are non-negative. Suggested benchmarks of 8 and 32 signify a transition from a weak to moderate level of evidence and from

[14]A posterior probability is the probability of the hypothesis given the observed data. This uses the prior probability of the hypothesis which is set before the data were observed.

[15]The three paradigms are Bayesian, Likelihood and frequentist.

[16]The likelihood principle states that two studies yielding the same data and using the same probabilistic model must have equivalent measurements of the strength of the evidence in those data. P-values violate the likelihood principle.

[17]It may seem obvious to assert that identical sets of data have exactly the same propensity to be misleading. The lottery example would certainly support this point of view. However just the opposite is the current dogma in statistics. It is commonly believed that two identical sets of data generally have different propensities to be misleading when the study designs from which they came differ. This dogma is a consequence of the failure to distinguish between EQ2 and EQ3.

[18]The controversy arises because the classical approach does not yield the same strength of evidence in each study (despite the fact that they generated identical datasets). This is because the p-value is adjusted differently based on the study design (i.e. stopping rule). The validly of this approach has been debated for decades.

a moderate to strong level evidence, respectively. Introductory material on this approach and extensions are readily available (e.g., [Blume, 2002; 2005; 2007; 2008; Goodman and Royall, 1988; Strug and Hodge, 2006a; 2006b; Royall and Tsou, 2003; Tsou and Royall, 1995; Van der Tweel, 2005]).

The Law of Likelihood provides a measure of the strength of evidence between two hypotheses.[19] In contrast, the Likelihood principle sets forth the conditions under which two experiments yield equivalent statistical evidence for two hypotheses of interest. This condition is met when their likelihood functions are proportional [Birnbaum, 1962; Barnard, 1949]. Acceptance of the law implies acceptance of the principle: if all the likelihood ratios between the two experiments are identical, then their likelihood functions must be proportional [Royall, 2000].

2.1 Illustration of the three evidential quantities

The most accessible illustration comes from Royall's [1997] diagnostic test example. Suppose we use a diagnostic test to generate evidence about the disease state of an individual. The properties of this test are given in Table 1. Its sensitivity is $0.94 = P(T+|D+)$ and specificity is $0.98 = P(T-|D-)$.[20] According to the law of likelihood a positive test result would be evidence supporting H_+ (disease is present) over H_- (disease is absent) because the likelihood ratio (LR) would be $47[= 0.94/0.02 = P(T+|D+)/P(T+|D-)]$. Likewise, a negative result would represent evidence supporting H_- over H_+ because the likelihood ratio would be $16.3[= 0.98/0.06 = P(T-|D-)/P(T-|D+)]$. The likelihood ratio is EQ1; it measures the degree to which the data support one hypothesis over another.

Table 1. Properties of the diagnostic test

		Test Result (T)	
		Positive (+)	Negative (−)
Disease Status (D)	Yes (+)	0.94	0.06
	No (−)	0.02	0.98

However, this test may generate evidence that is misleading. A positive result is correctly interpreted as evidence for H_+ over H_-, but positive results can occur when the disease is absent. This happens only 2% of the time in people without the disease, but when it does happen the test is said to have generated misleading evidence.[21] Of course, the test can also generate misleading negative

[19]This representation of evidence is essentially comparative between simple hypotheses. Simple hypothesis specify a single probability distribution. Composite hypothesis specify a family of distributions.

[20]The vertical bar '|' is read as 'given'. This is a conditioning argument, as events after the bar are considered fixed.

[21]Statistical tests (i.e., tests that are not deterministic) must be able to generate misleading

results and this happens 6% of the time in people who have the disease. These two probabilities are second evidential quantities (EQ2); they are the probabilities of observing misleading evidence under this study design. They are analogous to the error rates of hypothesis testing and they play an important role in defining the quality of the diagnostic test and the data collection process. A good diagnostic test maximizes sensitivity and specificity, which here is the same as minimizing the second evidential quantities (i.e., minimizing potential to observe misleading positive and negative tests).

Upon observing a certain test result, the strength of the evidence will be clear from the likelihood ratio (e.g., a positive tests yields a LR of 47). We will never know if the observed test result is misleading or not. However, it is sometimes possible to determine the propensity for the observed test result to be misleading.[22] For example, an observed positive result is misleading *if and only if* the subject does not have the disease. The probability that the subject does not have the disease is $P(D-|T+)$. By Bayes theorem, this is $P(D-|T+) = 1/[1+47\pi_+/\pi_-]$, where $\pi_+ = P(H_+)$ and $\pi_- = P(H_-) = 1-\pi_+$ are the prior probabilities of the individual's disease state. By identical reasoning, an observed negative result is misleading *if and only if* the subject has the disease. Here that probability is $P(D+|T-) = 1/[1+16.3\pi_-/\pi_+]$. So we see that the probability that the observed evidence is misleading, EQ3, is nothing more than the posterior probability.

The obvious (non-computational) difficulty with EQ3 is the specification of the prior probabilities, which are inherently subjective.[23] However diagnostic tests are a rare example in which there exists a broad consensus regarding the prior probabilities. Typically, the disease prevalence is used to set the prior when it is reasonable to assume that the subject was selected at random from the population. Suppose the disease prevalence is $\pi_+ = 0.015$. Then $P(D-|T+) = 0.583$ and $P(D+|T-) = 0.0009$. This means that positive results are not as reliable as negative results. In fact, observed positive results are misleading more than half the time *in this population*. This does not mean we are wrong when we interpret a positive result as evidence that the disease is present.[24] It just means that *in this population* the strength of evidence provided by a positive result is not strong enough to outweigh our (very strong) prior knowledge about the presence or absence of disease. Before the test is given, the probability that the individual is diseased is only 1.5%. This probability increases to 41.7% after observing a positive test result.[25] However, it remains more likely (58.3%) that the individual does not have the disease.

Because EQ3 depends on prior probabilities, it remains context based. This is

evidence. If not, then only a single observation would be needed to correctly identify the true hypothesis.

[22] We need only be willing to make certain assumption about the prior probability of the hypotheses or disease state.

[23] Admittedly there are varying degrees of subjectivity, as priors come from many different sources.

[24] Surely a positive result from this test is not evidence the disease is absent!

[25] This large increase is due to having a large likelihood ratio — strong evidence — of 47.

why strong evidence in one study may not be strong enough evidence to provide the same sense of reliability in another study.[26] From the expressions for EQ3, we see that larger likelihood ratios are more reliable in the sense that they are less likely to be misleading.[27] EQ1 and EQ3 have an inverse relationship; as the evidence gets stronger, its potential to be misleading decreases. Lastly, it is worth reiterating that the likelihood principle indicates that EQ2 is irrelevant once the data have been collected. At that point, EQ1 and EQ3 are the only quantities of interest.

2.2 Hypothesis testing and significance testing

The cornerstone of mainstream statistical methods is an ad-hoc union of hypothesis testing and significance testing [Blume 2003, Goodman 1998]. Experiments are designed under the hypothesis testing framework and analyzed under a significance testing one. Because this union was unplanned, the tail area probability[28] and what it represents is a source of continuing confusion. In the hypothesis testing framework the tail area probability represents EQ2 (i.e., the Type I error), but in the significance testing framework it represents EQ1 (i.e., the p-value). Moreover, there is no EQ1 in the hypothesis testing framework and there is no EQ2 in the significance testing framework. So it seems reasonable (and even quite natural) to many non-statisticians to merge the two approaches.[29] Scientists will always look for each of the three evidential quantities; they represent important concepts and they have distinct roles in the scientific process.

To impress the point, let's consider a hypothesis test and a significance test in the diagnostic testing example from the previous section. A hypothesis test sets the null hypothesis as H_0: disease absent and the alternative as H_1: disease present. We 'reject the null hypothesis' when a positive result is observed and 'fail to reject H_0' when a negative result is observed. The test would have a Type I error rate of 2% and a Type II error rate of 6%.[30] According to various decision theoretic standards this is a good test.

But problems arise if we try to interpret the results of the test as statistical evidence. First, failure to reject the null hypothesis cannot be taken as evidence for the null hypothesis. In fact, a negative result (i.e., failure to reject the null hypothesis) can never be interpreted as evidence that the disease is absent[31]; instead it is always interpreted as statistically inconclusive. Second, there is no

[26] Context affects the prior probabilities that are used in the calculation of EQ3.

[27] Here the prior probabilities are considered fixed.

[28] Tail area probability is a probabilistic term representing the core calculation in a p-value and in a Type I error.

[29] Fisher was well aware of this and warned: "In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence" [Fisher, 1959, p93].

[30] A Type I error rate is the probability of rejecting a true null hypothesis and a Type II error rate is the probability of failing to reject a false null hypothesis. The Type II error rate is calculated under a pre-specified simple alternative hypothesis.

[31] Absence of evidence is not evidence of absence.

strength of evidence to report (i.e., no EQ1). The best we can do is report our decision to 'reject' or 'fail to reject', along with the error rates of our decision rule.

But this is largely unsatisfactory from a scientific viewpoint, especially if the desire is to report the strength of evidence for the hypotheses of interest (e.g., weak, moderate or strong). This is why significance testing is employed at the end of the study. Significance testing involves calculating and reporting the p-value, as a measure of the strength of evidence against the null hypothesis. In this case, the p-value is 2% which would be considered strong evidence against the null because it is less than the common benchmark of 5%. Here too, it is not possible to generate evidence in favor of the null hypothesis; large p-values are interpreted as being inconclusive. For example, a negative test result, which yields a p-value of one[32], cannot be interpreted as evidence that the disease is absent. So we see from this example that while significance testing provides an EQ1 (i.e., the p-value), it does not provide an EQ2 or EQ3. EQ2, which would be calculated before the study is conducted, might be the probability of observing a small p-value when the null hypothesis is true (but EQ2 is not identically the p-value itself).[33]

Consider also a Bayesian approach that focuses on the posterior probabilities. In this case, after observing a positive result, the posterior probability of disease, $P(D+|T+)$, is only 0.417. It remains unclear how that positive result ought to be interpreted if the posterior probability is used as the EQ1. This is because even after a positive test result the subject is more likely to be disease free (i.e., 41.7% $< 50\%$). So should the positive result be considered evidence that the disease is absent? If yes, then this test can never generate statistical evidence that the disease is present (because the posterior probability of disease after a negative test result, $P(D+|T-) = 0.0009$ is very small). If no, then by what scale and context are we to interpret the posterior probability?[34] EQ1 needs to be explicitly defined and we need to be told how to use it. Also, does is make sense to define EQ2 in a way that is dependent on the prior probabilities?[35] So here too, the lack of an evidential framework leaves the approach so opened ended that its utility is not clear.

2.3 Background and notation

In order to consider more complex situations it is helpful to establish some notation, if only to help the reader distinguish between the three evidential quantities.

[32]The p-value is the probability of observing the data or data more extreme. Thus, the p-value is the probability of observing a negative or positive test result when the null hypothesis is true. This probability is one.

[33]The corresponding EQ3 would be the probability that the disease is absent given $p = 0.02$.

[34]If the answer is "They should be interpreted in relation to how much they have changed from the prior probability" then this is nothing more than the likelihood paradigm. See [Royall, 1997].

[35]One example is "What is the probability that the posterior will be as large as 0.95 under various hypotheses?" The subtle caveat here is that we are providing long run frequency properties that are dependent on the priors and this may or may not rely on the 'validity' of the specification of the priors.

Every effort is made to skip unnecessary mathematical complexities in the examples that follow.

Suppose observations $X_1, \ldots, X_n$ are independent and identically distributed according to some density $f(X_i; \theta)$. For a fixed sample size of n, let the likelihood function be $L_n(\theta) \propto \Pi_i f(X_i; \theta)$. There are two hypotheses of interest: the null, $H_0 : \theta = \theta_0$, and the alternative, $H_1 : \theta = \theta_1$. According to the Law of Likelihood, the strength of the evidence for H_1 over H_0 is measured by $LR = L_n(\theta_1)/L_n(\theta_0)$. An observed likelihood ratio will fall into one of three regions: $LR \in [0, 1/k]$ indicating strong evidence for H_0 over H_1, $LR \in (1/k, k)$ indicating weak evidence in either direction, and $LR \in [k, \infty)$ indicating strong evidence for H_1 over H_0 for some $k \geq 1$. By convention, we use $k = 8$ or 32 to denote the transition from weak to moderate and from moderate to strong evidence.[36]

2.4 Misleading evidence

Misleading evidence, by definition, is strong evidence in favor of the incorrect hypothesis (e.g., observing $LR = 47$ when H_0 is true or $LR = 1/47$ when H_1 is true).[37] We never know if observed evidence is misleading or not, but we can know the probabilities that a particular study will generate misleading evidence. They are represented, in general, by $\text{mis}_0 = P(LR \geq k|H_0)$ and $\text{mis}_1 = P(LR \leq 1/k|H_1)$ for some $k > 1$, typically 8 or 32. Here mis_0 is the probability of observing misleading evidence under the null hypothesis and mis_1 is the probability of observing misleading evidence under the alternative hypothesis. Both are EQ2s. Misleading evidence is seldom observed in the likelihood paradigm. Under quite general conditions, mis_0 and mis_1 are each bounded by $1/k$, the so-called universal bound, and their average is bounded by $1/(k+1)$ [Royall, 1997; 2000].

The universal bound is a crude device for controlling the probability of observing misleading evidence. The actual probability is often much less. As the sample size grows, both probabilities of observing misleading evidence (i.e., mis_0 and mis_1) converge to zero.[38] Under some mild parametric conditions and in moderate to large samples, the probability of observing misleading evidence is approximately $\Phi[-\ln k/c - c/2](\approx \text{mis}_0)$, where $c = |\theta_1 - \theta_0|/(nI(\theta_0))^{-1/2}$, $\Phi[\cdot]$ is the standard normal cumulative distribution function, and $I(\theta_0)$ is the Fisher information[39] [Royall, 2000]. From this we get an asymptotic maximum[40] probability of observ-

[36] A likelihood ratio of 4 would be weak evidence in favor of the H_1 and a likelihood ratio (LR) of 1/47 would be strong evidence in favor of H_0. The weak evidence region — LRs between 1/8 and 8 — is important to future discussions.

[37] Weak evidence — LRs between $1/k$ and k or typically 1/8 and 8 — are not considered to be misleading by definition. This is because weak evidence is inconclusive and has little force. In practice, Likelihoodists tend to ignore the directionality of weak evidence, characterizing it as inconclusive. See Royall [1997] for more on this topic.

[38] This is an interesting and useful property for EQ2. Contrast this with hypothesis testing in which one of its EQ2s (the Type I error) remains constant regardless of the sample size. Thus errors are made even with an infinite sample size.

[39] Under a normal model with known variance this approximation is exact.

[40] The maximum is over all alternatives

ing misleading evidence of only $\Phi[-\sqrt{2\ln k}](\leq 1/[2k\sqrt{(\pi \ln k)}])$, which is typically much less than the universal bound, $1/k$. Understanding the behavior of EQ2 is important in evaluating study designs. Technical extensions are recently available [Royall, 2000; 2003; Blume, 2002; 2007].

EQ2 and EQ3 have different mathematical representations. In general, the third evidential quantities are $P(H_0|\text{data}) = [1 + k\pi_1/\pi_0]^{-1}$ and $P(H_1|\text{data}) = [1+\pi_0/k\pi_1]^{-1}$ where $\pi_0 = P(H_0), \pi_1 = 1-\pi_0$, and $k > 1$ is the observed likelihood ratio in support of H_1 over H_0. So, as the sample size increases, these probabilities are driven to zero or one by the likelihood ratio, which itself convergences to 0 or ∞ in support of the correct (true) hypothesis. Thus, for a fixed prior probability, larger observed likelihood ratios are more reliable and the degree of reliability can always be improved by increasing the sample size.

3 REEXAMINATION OF ACCUMULATING DATA ('MULTIPLE LOOKS')

Scenarios involving repeated examination of accumulating data (i.e., multiple looks) provide an excellent opportunity to illustrate the value of having an evidential framework. A common misconception is that there is no 'penalty' for multiple looks at data in the Likelihood paradigm.[41] It is true that likelihood ratios (EQ1) are not modified by the number of looks at the data. But it is also true that there is a 'penalty'. The caveat is that the penalty applies to EQ2 and not to EQ1 as it would with p-values. The probability of observing misleading evidence, EQ2, is indeed affected by the number of examinations of the data. Every additional examination of the data increase EQ2. The stopping rule is indeed relevant for determining how often a study design will yield misleading evidence. But once data are collected, the stopping rule is irrelevant; it does not affect the measure of evidence nor the potential for observed evidence to be misleading (EQ3).

What many classical statisticians seem to find unsettling (or at least unfamiliar) about the likelihood paradigm is the separation of EQ1 from EQ2. EQ1 is the likelihood ratio; EQ2 is the probability that the likelihood ratio will be misleading. This is unfamiliar because the p-value is essentially used as both EQ1 and EQ2 indiscriminately. For example, consider two studies, one of which examines the data as it accumulates. Both studies happen to generate the same data set. Despite having identical data, they will report different p-values and claim to have different amounts of evidence. This is because the calculation of a p-value depends on the study design; it is calculated as if it were an EQ2, but interpreted as if it were an EQ1.

The underlying problem is that just one number, the p-value, has to function as both EQ1 and EQ2 (and maybe even the EQ3). This problem is avoided when a well defined evidential framework exists. For example, under a likelihood approach these two studies would report the same likelihood ratio and therefore the same

[41] P-values are penalized (i.e., discounted/inflated) for each look at accumulating data and for each planned look at the data that has not yet been executed. Likelihood ratios are not.

observed strength of evidence in the data. What is different is their study design; one study had a much larger chance of generating misleading evidence before any data were collected. But this probability is irrelevant now that data have been observed. What might have been is irrelevant. All that is relevant is what was observed (EQ1) and the propensity for the observed data to be misleading (EQ3).

Let's consider a numerical example to fix ideas. Suppose I design a study to collect evidence about the average change in systolic blood pressure due to a new medication. Assume that these changes are normally distributed with a standard deviation of $\sigma = 16$ mm Hg. My null hypothesis is that the there is no change in mean blood pressure ($H_0 : \mu = 0$) and my alternative hypothesis is that the mean change is 8 mm Hg or one-half of the population standard deviation ($H_1 : \mu = 8$ or 0.5σ).[42] With this information we can calculate how often certain study designs will generate misleading evidence in favor of the alternative hypothesis.

First, consider a fixed sample size design. The study enrolls 25 people, collecting their before and after blood pressure measurements, and the data are examined after all observations have been collected. The probability of observing misleading evidence of strength 8 or more in this design is only 0.0187, which decreases to 0.0042 for (misleading) likelihood ratios of 32 or more.[43]

In contrast, consider instead a 'truncated sequential' design, where the data are examined after every observation is collected and the study is stopped as soon as strong evidence for the alternative hypothesis is observed or when data on 25 people is collected, whichever comes first. When the null hypothesis is true, the probability that this design generates misleading evidence of strength 8 or more is 0.0717, which decreases to 0.0123 for (misleading) likelihood ratios of 32 or more.[44]

Lastly, suppose it was possible to continue collecting data forever, so we could plan to collect observations until we observed evidence in support of the alternative hypothesis. This is a very biased design. Data are collected until we find strong evidence in support of the alternative hypothesis while strong evidence in favor of the null hypothesis is ignored. Thus, this design provides the best chance of collecting misleading evidence, because even if the null hypothesis is true we continue sampling until we find misleading evidence. But even in this biased design the probability of observing misleading evidence is only to 0.0934 (for evidence of strength 8 or more) and 0.0233 (for evidence of strength 32 or more).[45] Notice that there is a very large probability that misleading evidence will never be generated. This is an important scientific safeguard; an investigator searching for evidence to support his pet hypothesis over the correct hypothesis is likely never to find such

[42] σ is the standard deviation, μ is the mean, and H_0 and H_1 are the null and alternative hypotheses. For convenience I expressed the alternative hypothesis in units of standard deviations. This makes the upcoming probability calculations simpler and adds to the generality of this example in the sense that any dependence on the stated value of the standard deviation is removed.

[43] This probability is simply $\Phi[-\ln k/c - c/2]$, noted in an earlier section, with $c = 0.5$ and $k = 8$ or 32.

[44] See [Blume, 2008] for the formula for this calculation.

[45] This probability is $\exp(-0.583 * c)/k$ with $c = 0.5$ and $k = 8$ or 32 [Blume, 2002; 2007].

evidence when using the likelihood paradigm.

It should be clear that even within the Likelihood paradigm a price is exacted for each examination of the data. That price is the inflation in the probability of observing misleading evidence, EQ2. One important characteristic of this probability is that the amount by which it increases converges to zero as the sample size grows [Blume, 2008]. Thus the probability of generating misleading evidence remains bounded even with an infinite number of examinations of the data. Simply put, the chance of observing misdealing evidence at a single point in time is less than the chance of observing misleading evidence at any point in time, although the latter remains bounded. More precisely we can write [Robbins. 1970]:[46]

$$P(LR_n \geq k|H_0) < P(LR_n \geq k \text{ for any } n = 1, 2, \ldots|H_0) \leq 1/k$$

As we have just seen, the probability of observing misleading evidence increases as the number of examinations increases. The chance of observing misleading evidence after collecting 25 participants is 0.0187, which increased to 0.0717 when the data were continuously monitored up to the 25 subjects, which increased to 0.0934 when the data were continually monitored until misleading evidence was obtained (i.e., there was no limit on the sample size). While the behavior of EQ2 is interesting and clearly depends on the study design, this behavior is independent of the observed EQ1 which does not depend on the study design.

4 MEASURING EVIDENCE ABOUT SEVERAL ENDPOINTS SIMULTANEOUSLY ('MULTIPLE COMPARISONS')

An evidential framework is also helpful when dealing with multiple comparisons for reasons similar to that just discussed in the context of reexamination of accumulating data. Consider a study in which 4 different measurements or endpoints are collected from each subject before and after a certain medical intervention. The goal of the study is to see if any of these endpoints change after the medical intervention. Our study will collect data from 6 participants.[47] For simplicity and without loss of generality, we will also assume that the 4 endpoints are independent and normally distributed with a known variance.[48]

The null hypothesis is that of no change, $H_0 : \mu = 0$. Because our sample size is small, we set our alternative to be a change of one standard deviation, $H_1 : \mu = \mu_1 = \sigma$ (i.e., $(\mu_1 - \mu_0)/\sigma = 1$).[49] The likelihood probabilities of observing misleading evidence are $\text{mis}_0 = \Phi[-\ln k/\sqrt{n} - \sqrt{n}/2]$ and $\text{mis}_1 = \Phi[-\ln k/\sqrt{n} -$

[46] A general approximation is $P(LR_n = k \text{ for any } n = 1, 2, \ldots|H_0) \equiv \exp(-0.583 * c)/k$ [Blume 2002,2007].

[47] I picked this for ease of calculation. Small sample sizes like this are, unfortunately, not unusual in genomic or in proteomic studies, although the number of genes or proteins being examined is often much larger than 4.

[48] These assumptions allow us to focus on the conceptual underpinnings of the problem instead of its mathematics. No generality is lost.

[49] In the example from the previous section the alternative was set to $1/2$.

$\sqrt{n}/2]$ with $n = 6$. When $k = 8$, these probabilities are both only 2%. The family-wise[50] probability — the probability of observing misleading evidence on at least one endpoint — is 7.8% (=1-(1-0.02)^4) under both the null and alternative hypotheses. When $k = 20$, the probability of observing misleading evidence on a single endpoint is 0.72% and the family-wise probability is 2.85%. We can see from this example that the sample size plays an important role. With n=12 observations and k=8, the probabilities of observing misleading evidence are only 1%, and the family-wise error rate drops to 3.9% from 7.8%.

Note that in the Likelihood paradigm we do not choose a cutoff for 'k' to denote which likelihood ratios are significant or not. Likelihood ratios are descriptive. If we want to control the probability of observing misleading evidence, then this is done completely through the sample size. Also, it should be clear that the probability of observing misleading evidence on at least one endpoint will always increase as the number of endpoints increase. However, this probability is also controlled through the sample size and can be driven to zero with a large enough sample.[51]

Likelihood is different from hypothesis testing in that there are effectively three evidence regions: strong evidence for the null over the alternative, strong evidence for the alternative over the null, and weak evidence supporting either hypothesis over the other.[52] In this example, there is a 33% chance of observing weak evidence.[53] Strong evidence, when it is observed, tends to be reliable but one-third of the time we will be left with weak inconclusive evidence. In contrast, hypothesis testing has only two zones: 'reject H_0' and 'fail to reject H_0'.[54]

For comparison purposes, it is instructive to consider what happens when we handicap the likelihood approach by removing the weak evidence zone. We do this by using $k = 1$, so that any amount of evidence (not just strong evidence) is potentially misleading. When $k = 1$, the probabilities are symmetrical and $\text{mis}_0 = \text{mis}_1 = \Phi[-\sqrt{n}/2]$. The two corresponding design probabilities (e.g., the probability of observing any evidence in support of the alternative when the null is true) increase from 2% to 11% and the family-wise rates increase from 7.8% to 37%. This happens because weak evidence is naturally less reliable.[55]

Let's consider how this situation would be handled with a hypothesis test. The Type I and II errors of a hypothesis test are α and $\beta = \Phi[Z_{1-\alpha} - \sqrt{n}]$. With a one-sided Type I error rate of 5% and 6 observations, we get a Type II error rate

[50]The family-wise error rate is defined as the probability of observing misleading evidence on at least one endpoint.

[51]This is not so with hypothesis testing, as we will see later.

[52]This inferential refinement is important in many respects and it should not be overlooked; it is the weak evidence that is most often misleading and should not be taken too seriously.

[53]It is the same under either hypothesis.

[54]Allowing for three zones is a major plus in my opinion. It completely resolves the asymmetry in hypothesis testing and significance testing that prevents them from generating evidence in support of the null hypothesis.

[55]Judging from this, it seems best to call weak evidence inconclusive rather than over-interpreting it in practice. However, the upcoming comparison to hypothesis testing is insightful and worth pondering for academic purposes.

of 21%. But when there are 4 endpoints with this identical structure, a Bonferroni adjustment is required to control the overall Type I error.[56] This adjustment results in a Type I error rate of 1.25% and a Type II error rate of 42% for each endpoint. The family-wise Type I error rate (i.e., the probability of making at least one Type I error) is now controlled at 5% (=1-0.9875^4), but the family-wise Type II error rate balloons to 89%.

Remember that the comparative family-wise rates for likelihood were both 37%. Certainly, 37% is far from 5%, but it is even further from 89%. For a small increase in one rate there is a large drop in the other, and this occurs without the complexity of post-hoc adjustments to the p-value. For example, the adjusted error rates of hypothesis testing are now dependent on how many other endpoints are being considered in the study. Thus, changing the number of endpoints in the study affects the accuracy (i.e., EQ2) with which you can test other endpoints. In contrast, the likelihood paradigm leaves the chance of observing misleading evidence (EQ2) on a single endpoint the same regardless of how many other endpoints are being considered. A price is still paid for multiple comparisons (the family-wise probabilities increase), but each endpoint remains consistent within itself because there is a clear evidential framework.

One way to balance the tradeoff between the different error rates is to use a metric like the average rate. The average error rate even has a nice interpretation — as the probability of making either error — if we are willing to assume that the null and alternative hypotheses are equally likely.[57] With Likelihood, the average rate increases from 2% to 11% when we eliminate the weak evidence region. This yields a family-wise average rate of 37%, which is the probability of making at least one error, in any direction, over all the endpoints. In comparison, the hypothesis test has an average error rate of 13% for a single endpoint, which increases to 22% when that endpoint is properly adjusted for the multiple comparisons, and this yields an average family-wise error rate of 62% (0.62=1-(1-0.22)^4).

This simple example is a good illustration of the general state of nature; there is no reason to assume that a lack of adjustment necessitates an uncontrollable or outrageous increase in observing misleading evidence. In fact, the worst case average rate at which likelihood ratios are misleading is 40% less than the same rate of a hypothesis test that adjusts for the multiple comparisons. Perhaps more importantly, the likelihood rates can always be driven to zero by increasing the sample size (unlike the Type I error), so a large enough sample size virtually guarantees there will not be any misleading evidence, regardless of the number of comparisons.

Figure 1, displays the family-wise average error rate for 1 and 4 endpoints under both likelihood and hypothesis testing (with multiple endpoints and a Bonferroni

[56] Adjustments other than Bonferroni are available, but this is the most common. The type of adjustment is irrelevant to the thrust of this example in any case. All adjustments simply trade Type II errors for Type I errors.

[57] These two types of errors can be weighted unequally in the likelihood paradigm when it makes sense to do so. But this is beyond the scope of this paper.

correction). Notice that for every sample size, the average error rate is smaller under Likelihood (when $\alpha = \beta$ in hypothesis testing the average error rates are equal). Notice also that the Likelihood average error rate (solid lines) converges to zero, instead of some fixed threshold, and hence can be driven as low as desired by increasing the sample size.

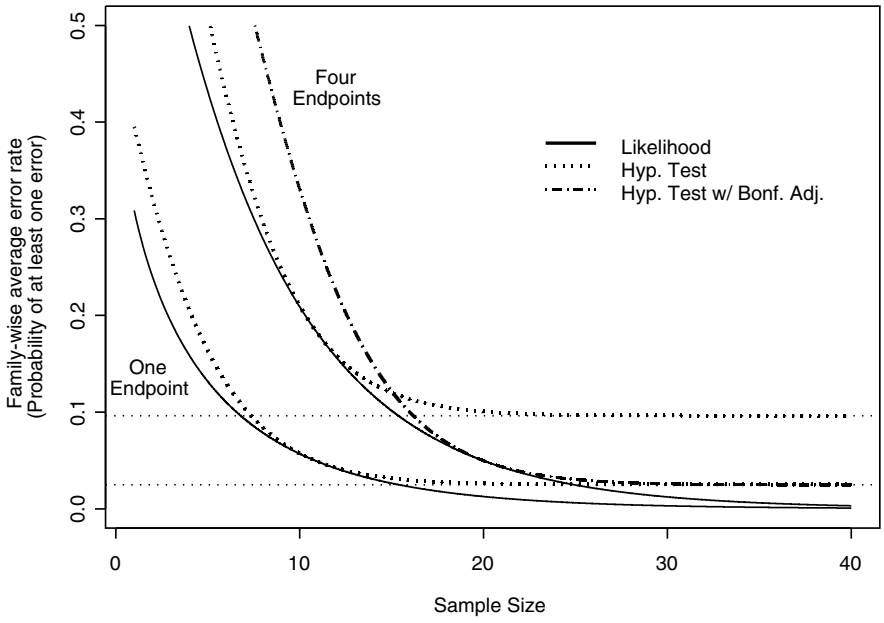

Figure 1. Average 'error' rates and family-wise average 'error' rates

Figure 2, displays the individual rates themselves. The single solid line represents both likelihood rates. Unlike hypothesis testing, how often misleading evidence is observed on a single endpoint does not depend on the total number of endpoints. For example, consider Fig. 2 at 11 observations, where, for a single endpoint, each of the design probabilities is 5% (all dashed and dotted lines cross at 5% with 11 observations). However, the hypothesis testing rates change when the three other endpoints are considered (solid and dashed line). What happened is this: the Type I error rate for each endpoint decreased to 1.25% and Type II error rate increased to 14%. As a result, the new family-wise average error rate is 27% (family-wise Type II rate is 45% and family-wise Type I rate is 5%) instead of 19% (=1-0.95^4) for likelihood, where the individual probabilities remain at 5% regardless of the number of endpoints.

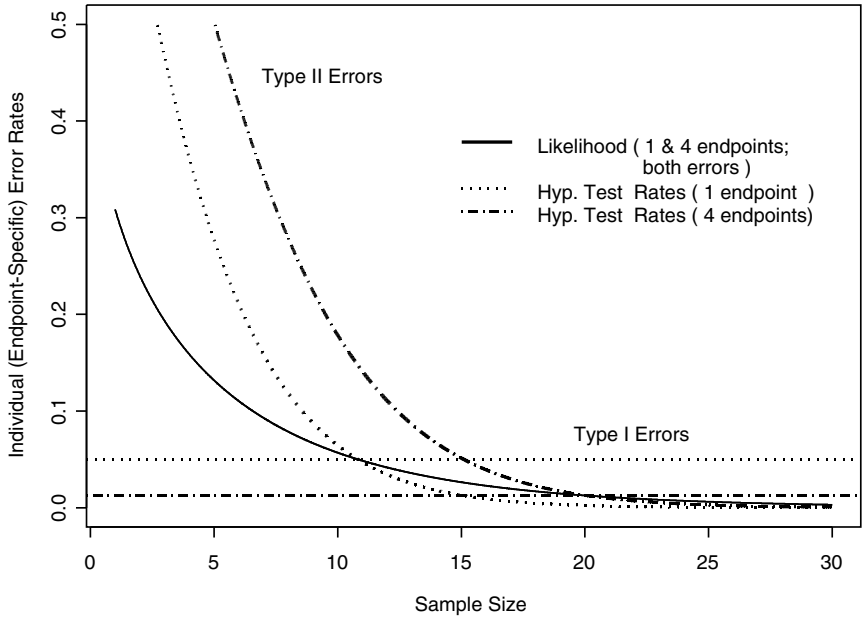

Figure 2. Type I and II error rates and probabilities of observing any evidence in the wrong direction (i.e., k=1) for Likelihood

Thus, even in the Likelihood paradigm there is a traditional frequentist price exacted for having multiple endpoints in a single study: The probability of observing misleading evidence on at least one endpoint increases with each additional endpoint. The important difference is that this price is not exacted on EQ1, so that the evaluation of the evidence from any individual endpoint does not depend on how many other endpoints are under consideration.[58] Moreover, the probability of observing misleading evidence on any single endpoint (EQ2) does not change with the number of endpoints (unlike hypothesis testing), thus conferring some internal consistency.

4.1 The potential for observed data to be misleading

The focus to this point has been on the probability of generating misleading evidence (EQ2) and how that varies with the addition of endpoints to the study

[58]The only minor exception to this is when the underlying statistical model allows endpoints to be correlated, but this not the dependence that we are trying to avoid.

in question. While EQ2 is an important piece of the puzzle, it is irrelevant once observations are collected. At the end of a study, EQ2 is often confused with the probability that the *observed* evidence is misleading (EQ3). Once we collect data, the design probability is irrelevant — the data are either misleading or not and we do not know which. So what we really want to know is how likely it is that *the observed* data *are* misleading.

Unlike EQ2, the probability that the observed evidence is misleading, EQ3, does not depend on the study design. An example is helpful to fix ideas and I'll continue with the multiple comparisons example from the previous section. In that example, we planned to collect six observations under one of two different designs: Study A has just a single endpoint, while Study B has three additional endpoints (4 total). Table 2 displays their design probabilities which have already been discussed at length.

Table 2. Two study designs and their performance characteristics. The interior of table is expressed as percentages. *based on the average error rate.

		One endpoint (Study A)			**Four endpoints (Study B)**		
		$\alpha(\mathbf{k})$	$\beta(\mathbf{k})$	**Overall***	$\alpha(\mathbf{k})$	$\beta(\mathbf{k})$	**Overall***
Likelihood with $k = 8$	Primary endpoint	2	2	2	2	2	2
	Any endpoint	-	-	-	7.8	7.8	7.8
Likelihood with $k = 1$	Primary endpoint	11	11	11	11	11	11
	Any endpoint	-	-	-	37	37	37
Hypothesis Testing	Primary endpoint	5	21	13	1.25	42	22
	Any endpoint	-	-	-	5	89	62

Now suppose the observed data yield a mean change on the first endpoint that was 1.01 standard deviations from zero. This evidence favors the alternative hypothesis by a factor of 21.34 ($= L(\mu_1)/L(\mu_0)$) regardless of which study generated the data (EQ1 does not vary by study). But the correct p-value is different depending on the design: $p = 0.013$ under Study A and $p = 0.052$ (=1-(1-0.013)^4) under Study B.[59] This inevitably raises the following question: Do these data

[59] Adjustment for p-values according to [Wright, 1992].

represent stronger evidence when they come from Study A? Or, when these data come from Study A, are they not "less likely to be misleading", "more reliable" in some sense, or don't they warrant more "confidence"?

The answer to both of these questions is a resounding no because EQ3 does not vary by study. These data are equivalent as evidence about the unknown mean regardless of the study design; they are no more likely to be misleading under one design or the other. Here is why: these data are misleading if and only if the null hypothesis is true. And here $(H_0|\text{data})$ is the same regardless of the design. An application of Bayes theorem yields $P(H_0|\text{data}) = 1/[1 + 21.34] = 0.045$, which does not differ by study design.[60] There is only a 4.5% chance that these data are misleading and this probability does not change if there were one, four or 2 million endpoints. Of course, changing the prior will change the degree to which we think these observed data are misleading. But EQ3 will not vary by study design unless the priors do. Thus we can use EQ3 to help assess the reliability of the observed evidence for a given prior, or as a sensitivity analysis across several priors.

5 COMMENTS

The Likelihood approach is used to measure the strength of evidence in observed data. It avoids the pitfalls of other statistical paradigms because it has a well defined evidential framework. The likelihood paradigm is often viewed as the common ground between the frequentists and Bayesians when it comes to measuring the strength of statistical evidence. This is because the likelihood function is often the only thing a frequentist and Bayesian can agree upon. Another reason is that likelihood ratios retain the desirable properties from both paradigms (irrelevance of sample space, good performance probabilities) while shedding the undesirable ones (dependence on prior distributions, ad-hoc adjustments to control error probabilities).

BIBLIOGRAPHY

[Barnard, 1949] G. A. Barnard. Statistical Inference. Journal of the Royal Statistical Society, Series B, 11: 115-149, 1949.

[Birnbaum, 1962] A. Birnbaum. On the foundations of statistical inference (with discussion). Journal of the American Statistical Association 53: 259-326, 1962.

[Blume, 2002] J. D. Blume. Likelihood methods for measuring statistical evidence. Stat Med. Sep 15, 21(17): 2563-99, 2002.

[Blume and Peipert, 2003] J. D. Blume and J. F. Peipert. What your statistician never told you about *p*-values. J AM Assoc Gynecol Laparosc 10(4):439-444, 2003.

[Blume, 2005] J. D. Blume. How to choose a working model for measuring the statistical evidence about a regression parameter. International statistical review, 73(2): 351-363, 2005.

[Blume *et al.*, 2007] J. D. Blume, L. Su, L. Acosta, R. M. Olveda, and S. T. McGarvey. Statistical evidence for GLM regression parameters: a robust likelihood approach Statistics in Medicine 26(15) 2919-36, 2007.

[60] Assumes a non-informative prior, i.e., the hypotheses were equally likely before any data were collected.

[Blume, 2008] J. D. Blume. How often likelihood ratio are misleading in sequential trials. Communications on Statistics-Theory and Methods, 38(8):1193-1206, 2008.
[Cornfield, 1966] J. Cornfield. Sequential Trials, Sequential Analysis and the Likelihood Principle. The American Statistician 29(2): 18-23, 1966.
[Edwards, 1971] A. W. F. Edwards. Likelihood. Cambridge University Press, London, 1971.
[Fisher, 1959] R. A. Fisher. Statistical Methods and Scientific Inference, 2^{nd} ed. New York, Hafner, 1959.
[Goodman and Royall, 1988] S. N. Goodman and R. M. Royall. Evidence and scientific research. American Journal of Public Health 78(12):1568-1574, 1988.
[Hacking, 1965] I. Hacking. Logic of Statistical Evidence. Cambridge University Press, New York, 1965.
[Royall, 1997] R. M. Royall. Statistical Evidence. Chapman & Hall, London, 1997.
[Royall, 2000] R. M. Royall. On the probability of observing misleading statistical evidence (with discussion). Journal of the American Statistical Association 95 (451), 760–767, 2000.
[Royall and Tsou, 2003] R. M. Royall and T. S. Tsou. Interpreting statistical evidence using imperfect models: Robust adjusted likelihood functions. Journal of the Royal statistical society, series B. 65,part2,391-404, 2003.
[Tsou and Royall, 1995] T. S. Tsou and R. M. Royall. Robust likelihoods. Journal of the American Statistical Association 90(419): 316-320, 1995.
[Savage, 1964] L. J. Savage. The foundations of statistics reconsidered. In Studies in Subjective probability with Kyburg HE and Smokler HE (eds). New York: John Wiley and Sons, 1964.
[Strug and Hodge, 2006a] L. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis: 1. Decoupling 'error probabilities' from 'measures of evidence'. Human Heredity 61:166-188, 2006.
[Strug and Hodge, 2006b] L. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis: 1. Implications for multiple test adjustments. Human Heredity 61:200-209, 2006.
[van der Tweel, 2005] I. Van der Tweel. Repeated looks at accumulating data: to correct or not to correct? European journal of epidemiology, 20: 205-211, 2005.
[Wright, 1992] S. P. Wright. Adjusted *P*-Values for Simultaneous Inference. *Biometrics*, vol. 48 1005-1013, 1992.