European Journal of Epidemiology (2005) 20: 205–211
DOI 10.1007/s10654-005-0540-y

REVIEW

Repeated looks at accumulating data: To correct or not to correct?

Ingeborg van der Tweel
Centre for Biostatistics, Utrecht University, 3584 CH Utrecht, The Netherlands

Accepted in revised form 12 January 2005

Abstract. Sequential analysis is a statistical way of analysing cumulative data. Its goal is to come to a decision as soon as enough evidence is reached for one or another hypothesis. In this article three different statistical approaches, the frequentist, the Bayesian and the likelihood approach, are discussed in relation to sequential analysis. In particular, the less known likelihood approach is elucidated.

Key words: Bayesian approach, evidence, frequentist approach, likelihood approach, *p*-value, sequential analysis

Abbreviations: LI = likelihood interval; LR = likelihood ratio; MLE = maximum likelihood estimate; SPRT = sequential probability ratio test; TT = triangular test

Introduction

The history of the development of statistical theory has shown two main schools: the Bayesian and the frequentist school. Fundamental differences between these two schools have divided statisticians in the past into Bayesians and frequentists. Recent developments and publications have drawn attention again to the less known likelihood approach and its applications in a particular statistical way of analysing cumulative data, sequential testing. In the following I will describe the three approaches and their main features and differences. The differences culminate when sequential methodology is applied. Therefore, especially this aspect will be discussed.

The frequentist approach: Neyman–Pearson theory

Sequential testing theory was developed by Abraham Wald during the Second World War to try to minimize cost of industrial experiments. Later sequential testing was adopted in clinical trial settings for ethical reasons: we want to stop a trial early if a new drug or treatment is especially beneficial or harmful. Wald (1902–1950) and also Neyman (1894–1981) were the most influential exponents of the frequentist philosophy [1]. Central in the Neyman–Pearson theory is the *likelihood ratio* (LR), that is defined as $LR = P(x \mid \theta_1) / P(x \mid \theta_2)$, the ratio of the probability distribution or likelihood of the observed data, summarized by x, given that hypothesis H_1: $\theta = \theta_1$ is true and the probability distribution or likelihood of the observed data given hypothesis H_2: $\theta = \theta_2$ is true, where θ is the parameter of interest. If the LR is large, the observed data contain evidence favouring θ_1, if the LR is small, the observed data contain evidence favouring θ_2 and if the LR equals 1, there is no evidence for either θ_1 or θ_2. Wald's Sequential Probability Ratio Test (SPRT) is based on the Neyman–Pearson theory. It says that one should continue collecting data as long as $B < LR < A$, to stop data collection and decide for H_2 as soon as $LR \leqslant B$ and to stop data collection and decide for H_1 as soon as $LR \geqslant A$ [2]. A and B are functions of the type I error α and the type II error β. Thus based on the LR, Wald's SPRT chooses between the two hypotheses using a *stopping rule* and a *decision rule*.

The frequentist approach: Significance testing

While Neyman–Pearson theory is concentrated on hypothesis testing and decision-making, another frequentist approach (usually ascribed to Fisher [3]) concentrates more on significance testing using critical values or *p*-value procedures. A statistical test is performed under the assumption that the null hypothesis is true. Based on the observed data, one rejects or accepts the null hypothesis. There is no explicit role for an alternative hypothesis. Repeated significance testing procedures as introduced and fully explained by Armitage [4] can be viewed in this light. The statistical test is repeated after each new group of observations. To maintain an overall type I error α of, say, 0.05 and to avoid 'chance capitalization' or 'inflation of the error rate', each interim analysis is performed at a lower nominal value for α. Pocock and O'Brien and Fleming, amongst others, have developed different ways to 'divide' α for a fixed number of

interim analyses [5, 6]. DeMets and Lan describe how the overall α can be 'spent' more flexible according to the amount of information (time) used [7].

Whitehead further elaborated Wald's SPRT in his 'boundaries approach' to sequential testing [8]. He developed continuous stopping boundaries such that the type I error is maintained and power requirements are satisfied. The use of these boundaries is a very flexible way of performing interim analyses. Test statistics are the efficient score statistic Z and Fisher's information V. Z is a cumulative measure for the effect size, V is a measure for the amount of information about the parameter θ contained in Z. The parameter θ that is to be tested can be standardized such that it is always equal to zero under the null hypothesis. Under the null hypothesis the distribution of Z is Normal with mean θV and variance V. Whitehead's approach is thus close to Fisher's: no choice between two hypotheses is made, but the null hypothesis is rejected or accepted based on the cumulative observed data. When the sequential test leads to the decision to stop further data collection, the *p*-value has to be adjusted for the multiple looks at the data.

The Bayesian approach

Already in the 18th century, Bayes (1701–1761) developed his theory on probability which was published (posthumously) in 1763. Bayesians express their prior knowledge, ideas, theories, ... in a *prior* distribution function for the parameter of interest. Subsequently they observe data as result of an experiment. The product of the prior distribution function and the information about this parameter contained in the data and expressed in the likelihood, leads to the *posterior* distribution function. This posterior distribution can thus be viewed as an update of the prior information or the way belief is altered by data. If the two hypotheses H_1 and H_2 are to be distinguished, the posterior probability ratio can be expressed as the product of the LR and the prior probability ratio:

$$\frac{P(\theta_1|x)}{P(\theta_2|x)} = LR \cdot \frac{P(\theta_1)}{P(\theta_2)}.$$

When cumulative data from an experiment are analysed sequentially following the Bayesian approach, the posterior distribution describes the currently available information about the parameter of interest. This information can be used to decide whether to stop the experiment because enough evidence is already gathered or whether additional evidence is needed. In a Bayesian sequential setting no adjustment is necessary for interim looks at accumulating data [8]. The fact that test results following the Bayesian approach depend to a large extent on the choice of the prior distribution makes the approach less attractive.

The likelihood approach

Over decades statisticians have divided themselves into two, often controversial, groups: the frequentists and the Bayesians. The frequentists try to answer the question: "What should I do?", while the Bayesians ask: "What should I believe?". Neither of these approaches explicitly answers the question: "What do the data say?" [3, 9] Perhaps a *third way*, the likelihood approach, deserves more attention than it has got until now. The concept of likelihood can be ascribed to Fisher (1890–1962). Fisher was against the use of prior probability distributions, but also rejected the idea that probability can only be interpreted in a long-run frequency way. For example: one can state, that, if the null hypothesis is true, the probability that we observe a specific test result, is smaller than, say, 0.05. We mean to say that, if we would repeat our experiment a very large number of times, we would observe this test result or one more extreme in less than 5% of the experiments. Fisher's ideas [1] are formulated as

- whenever possible to get exact results we should base inference on probability statements, otherwise it should be based on the likelihood;
- the likelihood can be interpreted subjectively as a rational degree of belief, but it is weaker than probability, since it does not allow an external verification, and
- in large samples there is a strengthening of likelihood statements where it becomes possible to attach some probabilistic properties ('asymptotic approach to a higher status').

Fisher's view differs, however, from the 'pure likelihood' view as supported by, amongst others, Royall [3] and Blume [9]. This 'pure likelihood' view, or 'evidentialism' as Vieland and Hodge coined it [10], tries to answer the question "What do the data say?" by the use of a methodology based only on the likelihood function. The Likelihood Principle states that the likelihood function contains all of the information in an experiment relevant for statistical inference about the parameter θ [11] According to the Law of Likelihood, as formulated by Hacking, the observed data are evidence supporting one hypothesis *over* another hypothesis and the LR measures the strength of that evidence [3]. Note that no choice is made is for one or the other hypothesis.

(Mis)interpretation of the *p*-value

Controversies arise between the frequentist and the likelihood approach when it comes to statistical

inference. The controversies arise because of the way *p*-values are used and interpreted in the frequentist approach. A *p*-value is the probability that the null hypothesis is rejected erroneously. It is, however, also interpreted as a measure of strength of evidence against the null hypothesis: ''the smaller the *p*-value, the stronger the evidence''. Several authors show that data from different experiments can have the same likelihood, but do not necessarily lead to the same *p*-value [1, 3, 12]. As an example, suppose we observe 8 successes in 10 experiments with probability of success for each experiment equal to 0.5, a typical example of a binomial experiment. The one-sided *p*-value corresponding to 8 or more successes in 10 experiments is equal to 0.055. If we had not planned beforehand to do exactly 10 experiments, but to continue until 2 failures (8 successes) were observed, the sampling scheme is a different one. We then would have a negative binomial experiment. In that case the one-sided *p*-value corresponding to 8 or more successes is equal to 0.0195. While these two experiments have the same likelihood and thus the same evidence about the parameter of interest, the *p*-values are different and even lead to different conclusions with regard to the rejection of the null hypothesis. The (strong) likelihood principle states that two data sets that produce proportional likelihoods should lead to identical conclusions and thus should also carry the same evidence about the parameter of interest. People are inclined to give different interpretations to the *p*-value depending on the sample size of the experiment. If two experiments that are identical except for their sample sizes produce results with the same *p*-value, these results do not represent equally strong evidence against the null hypothesis. Some statisticians will argue that the evidence is stronger in the smaller experiment, while others will state that the results of the larger experiment give stronger evidence [3]. Thus, the α-postulate as formulated by Cornfield: ''All hypotheses rejected at the same critical level have equal amounts of evidence against them'' or in other words ''Equal *p*-values represent, at least approximately, equal amounts of evidence'' is wrong [3].

Correction of the type I error

The discrepancy between frequentist and likelihood inference culminates in the use of sequential methodology. When, for example, accumulating results of a clinical trial are monitored applying a frequentist method, at each interim analysis part of the overall α is spent. Smaller nominal values for α must be used at each interim analysis to guarantee that the overall value of α is not inflated at the end of the trial. The consequence is that an experiment cannot be extended beyond its planned sample size, because the preset level of α is already 'spent' [12]. The decision to continue or to stop further data collection depends not only on the information obtained so far but also on a stopping rule. This stopping rule is a function of the type I error α.

The type I error is not part of the likelihood approach. Here the evidence in the data is entirely independent of the type of sampling, be it sequential or fixed. It has been shown that the likelihood function in sequential experimentation ignores the stopping rule and thus that the evidence from an experiment is independent of the stopping rule [1, 13]. So, in a sequential experiment multiple looks at the data do not affect the likelihood function. (Note that also Armitage remarks in passing: ''In fact, the likelihood function is unaffected, apart from a constant multiplier, by the stopping rule under which the data were collected.'')[4]

Strength of the evidence

Pawitan, however, also emphasizes that the likelihood principle states something about evidence, but not about any particular course of action [1]. Then the question can arise when 'enough' evidence is obtained in favour of one of two hypotheses. One could argue that 'enough' evidence is a very subjective matter. Nevertheless, several likelihood supporters have searched how to quantify the amount of evidence. The LR measures the strength of the evidence. Let us denote the value of the LR by k. Royall suggested benchmark values of $k = 8$ and $k = 32$ to distinguish between weak, moderate or fairly strong, and strong evidence [3]. If the LR $\geqslant k$, the data show (*fairly*) *strong* evidence in favour of H_1, if the LR $\leqslant 1/k$, the data show (*fairly*) *strong* evidence in favour of H_2 and if $1/k < \text{LR} < k$, the data show *weak* evidence. (Of course this is just a crude categorization of a continuous measure.) We can see what the data tell us by graphing the likelihood function and by calculating $1/k$ likelihood intervals. An $1/k$ likelihood interval (LI) encloses all values for the parameter of interest θ for which $L(\theta)/L(\hat{\theta}) \geqslant 1/k$, where $L(\theta)/L(\hat{\theta})$ is the normalized or standardized likelihood function and $\hat{\theta}$ is the maximum likelihood estimate (MLE) for θ. Or, in other words, it consists of all values that are ''consistent with the observed data''. Any θ within the likelihood interval is supported by the data because the best-supported value, the MLE $\hat{\theta}$, is only better supported by a factor k or less. One could notice a similarity between likelihood intervals and confidence intervals. An 1/8 likelihood interval corresponds to a 95.9%-confidence interval and an 1/32 likelihood interval corresponds to a 99.1%-confidence interval. (An 1/6.67 likelihood interval corresponds to a 95%-confidence interval.) Nevertheless, likelihood intervals should not be interpreted as identical to confidence intervals. Furthermore, a (frequentist) confidence interval depends

also on the number of interim looks at the data, while likelihood intervals depend only on the data itself.

Misleading evidence

Strong evidence, however, can be misleading evidence. Observations can hold strong evidence supporting H_1 over H_2, while in fact H_2 is true. However, although evidence can be misleading, the probability of observing strong misleading evidence is small and limited by a *universal upper bound* $P(LR \geqslant k) \leqslant 1/k$, when the *true* distribution of the data is according to H_2 [14]. This important fact implies that it is difficult to collect, deliberately or not, strong misleading evidence [9]. As a devil's advocate, one could plan to continue sampling until enough evidence is gathered for one's favourite hypothesis although it is an erroneous one compared to the rival hypothesis. The probability that one will be successful in the end is always smaller than $1/k$ even if the number of observations is unlimited.

The probability of misleading evidence can be compared with the type I error α. Both point into the direction of H_1, when in fact H_2 is true. The type II error β can be compared to the probability of failing to find strong evidence in favour of H_1, i.e., the probability of finding only weak evidence plus the probability of finding misleading evidence in favour of H_2.

Under a sequential design the probability of observing misleading evidence is greater than that under a fixed sample size design. Although this probability increases with each look at the data, it remains bounded because the amount by which it increases converges to zero as the sample size grows [15].

An example

As an example, let us look at a simple sequential experimental design [11, 16]. A clinical trial compared remission times for two treatments for acute leukaemia: 6-mercaptopurine (treatment A) and placebo (treatment B). The original study was stopped after the analysis of 21 pairs of patients. For each pair of patients a preference was recorded for A or B according to which therapy resulted in a longer remission time. Under the null hypothesis H_0 the probability of a preference for A was 0.50, under the alternative hypothesis H_A this probability was thought equal to 0.75. In the first two columns of Table 1 the data for the 21 pairs of patients are shown. A sequential triangular test (TT), according to Whitehead [8], was designed with $\alpha = 0.05$, power $1-\beta = 0.95$ and parameter $\theta = \log(OR) = \log(3) = 1.0986$. In the 3rd and 4th column of Table 1 the test statistics Z and V for the TT are given. Z is equal to the difference of the observed total number of preferences for A and the expected number under H_0 ($= n*0.5$); V is equal to the variance of Z under H_0 i.e. $n/4$ ($= n*0.5*(1-0.5)$). In the last column of Table 1 the LR is given for the accumulating data, i.e. the likelihood function for the data under H_A: $\theta = 0.75$ divided by the likelihood function for the data under H_0: $\theta = 0.50$.

Table 1. Data for the clinical trial as described [11, 16]

Number of pairs, N	Preference	Z	V	LR
1	A	0.5	0.25	1.50
2	B	0	0.50	0.75
3	A	0.5	0.75	1.12
4	A	1.0	1.00	1.69
5	A	1.5	1.25	2.53
6	B	1.0	1.50	1.27
7	A	1.5	1.75	1.90
8	A	2.0	2.00	2.85
9	A	2.5	2.25	4.27
10	A	3.0	2.50	6.41
11	A	3.5	2.75	9.61
12	A	4.0	3.00	14.42
13	A	4.5	3.25	21.62
14	B	4.0	3.50	10.81
15	A	4.5	3.75	16.22
16	A	5.0	4.00	24.33
17	A	5.5	4.25	36.40
18	A	6.0	4.50	54.75
19	A	6.5	4.75	82.11
20	A	7.0	5.00	123.16
21	A	7.5	5.25	184.74

In Figure 1, Z and V are plotted in a triangular design. After the 16th pair the upper boundary of the triangular test was crossed, which led to the conclusion that the null hypothesis could be rejected. The 90%-confidence interval for θ is (0.59; 0.88).

In Figure 2, the $LR = P(x \mid \theta = 0.75) / P(x \mid \theta = 0.50)$ is plotted against the number of pairs. The LR was greater than 8 after 11 pairs and greater than 32 after 17 pairs of patients.

In Figure 3, the standardized likelihood function is plotted together with the 1/8 and 1/32 likelihood intervals for the data from the trial (14 preferences for A in 17 pairs of patients). The value $\theta = 0.50$ is

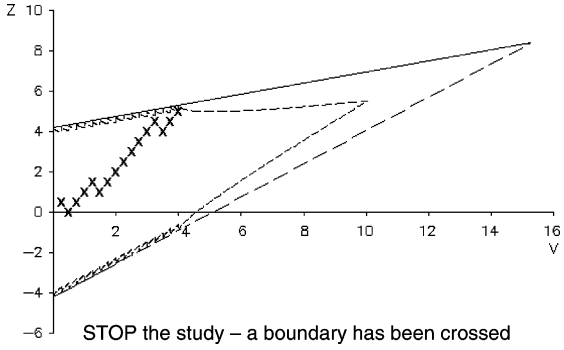

Figure 1. Results of the trial plotted as test statistics Z vs. V in a triangular test.

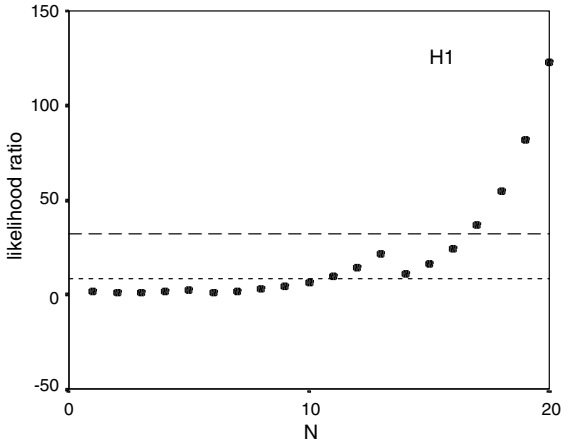

Figure 2. Results of the trial plotted as the Likelihood Ratio (LR) vs. the number of pairs N. The dotted line corresponds to a LR of 8, the dashed line to a LR of 32.

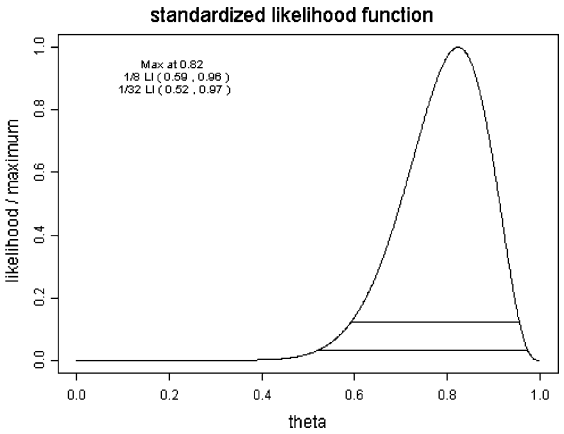

Figure 3. Results of the trial plotted as the standardized likelihood function vs. θ, together with the 1/8 and 1/32 likelihood intervals (LI).

not consistent with the data, the value $\theta = 0.75$ is included in both likelihood intervals.

To compare results from a trial that rejected the null hypothesis with those from an experiment that led to the acceptance of the null hypothesis I simulated preference data under the null hypothesis H_0: $\theta = 0.50$. Results of the simulated data are presented in Figures 4, 5 and 6. After the 17th pair the lower boundary of the triangular test was crossed, which led to the conclusion that the null hypothesis could be accepted (Figure 4). The 90%-confidence interval for θ is (0.30; 0.69). The LR was smaller than 1/8 after 16 pairs and smaller than 1/32 after 18 pairs of patients (Figure 5).

In Figure 6, the standardized likelihood function is plotted together with the 1/8 and 1/32 likelihood intervals for the simulated data (8 preferences for A in 18 pairs of patients). The value $\theta = 0.50$ is consistent with the data, the value $\theta = 0.75$ is not consistent with the data.

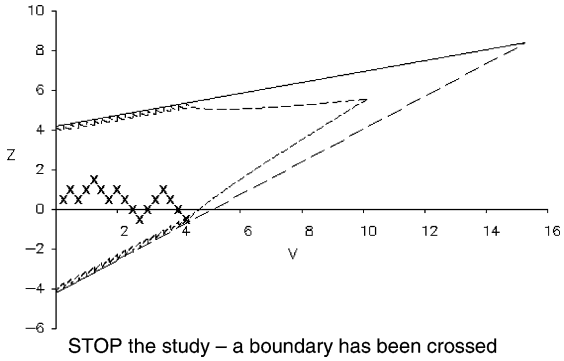

Figure 4. Results of the simulated data plotted as test statistics Z vs. V in a triangular test.

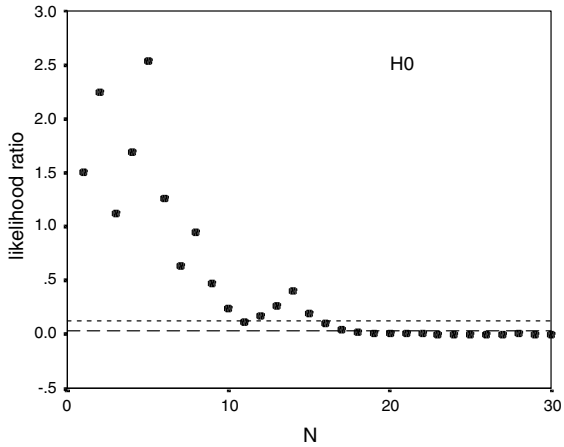

Figure 5. Results of simulated data plotted as the Likelihood Ratio (LR) vs. the number of pairs N. The dotted line corresponds to a LR of 1/8, the dashed line to a LR of 1/32.

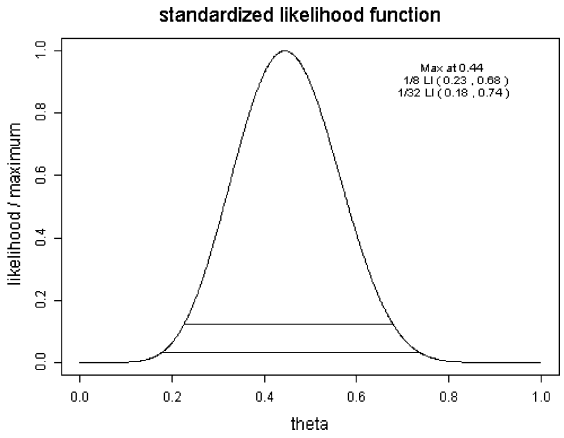

Figure 6. Results of the simulated data plotted as the standardized likelihood function vs. θ, together with the 1/8 and 1/32 likelihood intervals (LI).

(Note that the LR is invariant to the choice of the parameter i.e. it makes no difference whether θ is used or $\psi = \log(OR) = \log(\theta/(1-\theta))$.)

Blume gives an approximation to the probability P(M) that a sequential design will generate

misleading evidence [15]. For Bernoulli-data, like the example above, $P(M \mid k = 8) \approx 0.0922$, while $P(M \mid k = 32) \approx 0.0230$. A 10,000 simulations for these data under H_0 resulted in $P(M \mid k = 8) = 0.1002$ and $P(M \mid k = 32) = 0.0238$; a 10,000 simulations under H_A resulted in $P(M \mid k = 8) = 0.0874$ and $P(M \mid k = 32) = 0.0224$. (The universal upper bound for P(M) is equal to 0.125 for $k = 8$ and equal to 0.03125 for $k = 32$.)

Summary

In a Bayesian sequential setting, as in the likelihood approach, no adjustment is necessary for repeated looks at accumulating data. The practical problem in the Bayesian approach of statistical testing lies in the choice of an appropriate prior distribution and the amount of (subjective) belief that is assigned to it. In large sample problems the data will dominate the prior distribution and thus determine the posterior distribution so that the Bayesian approach becomes the likelihood approach. This is also the case when a non-informative prior distribution is used. Furthermore, the invariance property that holds for the likelihood approach does not hold in a Bayesian setting.

The Neyman–Pearson approach indeed makes use of the LR, but its numerical value is not interpreted as a measure of the strength of evidence. Only its extremeness is compared to critical boundaries to make a decision. The *p*-value was added to have a measure of the strength of the evidence after all. This use of the *p*-value comes into conflict with the likelihood principle. (Data sets with the same (or proportional) likelihood carry the same evidence and should thus lead to the same *p*-value.)

The likelihood approach is a simple and elegant *third way* to deal with evidence in experimental data. It makes a clear distinction between the degree of uncertainty and the strength of the evidence. Other favourable qualities of the likelihood approach are:

- Two hypotheses of equal importance are compared instead of focusing on the acceptance or rejection of the null hypothesis.
- No correction for interim looks at accumulating data is necessary, so there is also no problem in extending an already obtained sample.
- The MLE can be used for the parameter of interest without adjustment, while following a frequentist sequential test it is biased [8].
- For a sequence of observations the universal upper bound applies, i.e. the probability of finding strong misleading evidence of strength k or greater cannot exceed, and often is much less than, the value $1/k$ [14].

Of course there are topics that call for further investigation:

- The Law of Likelihood is restricted to the comparison of simple hypotheses and does not apply to most composite hypotheses [3], although Blume suggests a transformation such that the Law can be applied [9].
- In multi-parameter models there is no general way to eliminate nuisance parameters. Royall and Blume suggest some *ad hoc* methods, of which the use of profile likelihoods looks the most satisfactory [3, 9]. This has to be further investigated, especially in the context of sequential likelihood testing.
- Simulations of sequential designs using the LR will have to show their characteristics, like the efficiency, the average sample size to come to a decision, the probabilities of weak and of misleading evidence, ... for different outcome variables.

Conclusion

Recent developments and publications on the likelihood approach and especially its application in sequential designs [3, 9, 14, 15] prompted me to go into this '*third way*' and compare it with the frequentist and the Bayesian approach. Although there are still topics to investigate further before a definite recommendation can be made to turn into this way, I would like to end with the following conclusion.

Because the number of interim looks at accumulating data does not affect the LR, sequential designs based on the LR are a very natural way of monitoring the strength of evidence in the observations. A sample of observed data can be enlarged without worrying about the effect on the type I error, and thus without any adjustments. For (simple) sequential testing problems in observational, epidemiological studies on (matched) case-control data, where the goal is to achieve evidence for one hypothesis over another and where it is important to make efficient use of available resources, the likelihood approach is an objective answer to the question: "What do the data say?"

References

1. Pawitan Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford: Clarendon Press, 2001.
2. Wetherill GB, Glazebrook KD. Sequential methods in statistics. 3rd edn. New York: Chapman & Hall, 1986.
3. Royall R. Statistical evidence. A likelihood paradigm. London: Chapman & Hall, 1997.
4. Armitage P. Sequential medical trials. Oxford: Blackwell Scientific Publications, 1975.
5. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64: 191–199.
6. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549–556.

7. DeMets DL, Lan KKG. Interim analysis: The alpha spending function approach. Stat Med 1994; 13: 1341–1352.
8. Whitehead J. The design and analysis of sequential clinical trials. Rev. 2nd edn. Chichester: Wiley, 1997.
9. Blume JD. Likelihood methods for measuring statistical evidence. Stat Med 2002; 21: 2563–2599.
10. Vieland VJ, Hodge SE. Statistical evidence: A likelihood paradigm. Am J Hum Genet 1998; 63: 283–289 (Book review).
11. Berry DA. Interim analysis in clinical trials. The role of the likelihood principle. Am Stat 1987; 41: 117–122.
12. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. Am Stat 1966; 20: 18–23.
13. Anscombe FJ. Sequential medical trials. JASA 1963; 58: 365–383.
14. Royall R. On the probability of observing misleading statistical evidence (with discussion). JASA 2000; 95: 760–780.
15. Blume JD. On observing misleading evidence in sequential trials. http://alexander.stat.brown.edu/~jblume/slides/, 2003.
16. Freireich EJ, Gehan EA, Frei E, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remission in acute leukemia: A model for evaluation of other potentially useful therapy. Blood 1963; 21: 699–716.

Address for correspondence: Ingeborg van der Tweel, Centre for Biostatistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands
Phone : +31-30-253-3903; Fax: +31-30-252-1105
E-mail: i.vandertweel@bio.uu.nl