Confidence regions for hypotheses on system characteristics

Hans-Rolf Gregorius

Institut für Forstgenetik und Forstpflanzenzüchtung
der Universität Göttingen, Büsgenweg 2, D-37077 Göttingen

Abstract – The principle of hypothesis falsification forms the epistemological basis of systems analysis. It is realized by specification of a method of investigation suitable for testing a set of hypotheses on system characteristics. From this perspective, the principle has a special position in statistics, in that its observance calls for controlling the chances of erroneous falsification for each of the considered hypotheses. Falsification by improbability thus replaces strict falsification. Regions of rejection of hypotheses are then solely determined by the levels of significance set to the probability of not obtaining a result more probable than the observed under the respective hypothesis. Being a function of hypothesis and observation, this probability is termed the *confidence* ζ in a hypothesis supported by an observation, and it establishes the relation to the theory of exact testing. Taking advantage of the analogy to hypothesis testing, "exact" confidence regions can then be determined as the set of hypotheses that is supported by the observation with confidence ζ above a given level α of significance. The problem of hypothesis (including parameter) estimation is introduced as a problem of selecting hypotheses from confidence regions, and the roles of maximizing likelihood and confidence for obtaining and testing estimates is discussed. Maximization of confidence over subsets of hypotheses is also shown to yield unconditional tests of composite hypotheses. Application to the determination of "exact" confidence regions for the frequency parameter of the binomial distribution yielded results that are in many cases closely approximated by the direct method of Clopper and Pearson but also show distinct and unfamiliar features.

Introduction

In experimental sciences, systems analysis is an indispensable tool for acquiring dependable knowledge. Its epistemological principle revolves around the modelling of experimentally testable hypotheses on real systems. Hypotheses may concern type of model (e.g. all models of positive assortative mating), a particular model (of partial selfing with random cross-fertilization), or special values of the parameters of a model (concerning the rate of selfing in the latter model; note that it is common in statistics to restrict usage of the term "hypothesis" to model parameters (see e.g. Edwards 1972, p.3)). Insufficient conformity of the experimental results with the hypothesis is considered as falsification of the hypothesis, where failure of falsification is no proof of correctness of the hypothesis (see e.g. Popper 1968). In a strict sense this principle excludes the possibility of incorrect falsification and thus of erroneous rejection of a hypothesis (known as type I error in mathematical statistics). The fact that reality is accessible only in samples, however, enforces consideration of exactly this possibility in the analysis of real systems. Adherence to the principle of falsification thus requires us to control the chances of erroneous falsification.

This requirement largely determines the common (Neyman-Pearson) methods of statistical decision making. Yet, the system analytical objective places an emphasis that differs from many decision theoretical methods in statistics with respect to the kind of risk valuation. In the planning of an investigation, correct falsification is given unequivocal priority over other considerations of risk minimization or profit maximization that are significant in game theory or economics, for example. Such considerations chiefly concern the risk of accepting a false hypothesis (type II error), the valuation of which is based on a designated set of alternative hypotheses. In contrast with the "null hypotheses", the alternative hypotheses are themselves not a matter of falsification in this valuation. Yet, in a system analytical context, each hypothesis is an object of unconditional falsification, which leaves us with rejecting hypotheses if they provide the observations with sufficiently low probability. Rejection therefore amounts to falsification by improbability.

Thus, if there are no reasons compelling distinction between null and their alternative hypotheses as objects of falsification, the problem of hypothesis testing extends to the whole space of hypotheses and therefore calls for the specification of confidence regions (the relation between testing of hypotheses against alternatives with the involved type II errors and the specification of confidence regions has attracted some attention; see e.g. Wellek and Michaelis 1991). At first sight, the concept of confidence regions, which suggests reliability on nonfalsified hypotheses, also appears problematic. However, via complete specification of a confidence region, i.e. specification of an *exact* confidence region, the principle of falsification could be accounted for. In this case, the complement of a confidence region would consist of *all* hypotheses falsifiable by improbability of an observation obtained under the applied method of investigation.

Based on this elementary supposition, an attempt will be made in the present paper to demonstrate the strong interplay among the concepts of hypothesis testing, confidence region, and estimation, as dictated by consistent application of the system analytic principle of falsification by improbability. In so doing, the essence of any testing procedure, the specification of regions of rejection, will be shown to follow cogently for each method of investigation from the idea underlying the level of significance. This idea amounts to basing the decision to reject a hypothesis on the probability of not obtaining a result more probable than the current observation. The theory of "exact" testing (as described e.g. in Weerahandi 1995, Preface, p.*viii*) then presents itself as a consequence of this approach. These results will be used to illustrate the close relation of the concept of confidence regions to the problem of hypothesis estimation.

The question as to the most appropriate among available or newly developed methods of investigation (tests) is of direct relevance to the improvement of systems analysis. Yet, again the objective differs from that of established mathematical statistics, where the quality of a test is measured in terms of various properties of type II errors and thus of the acceptance region under a certain set of alternative hypotheses. The system analytical perspective instead focusses on the efficiency of two methods of investigation applied to the same system characteristics in falsifying hypotheses. Moreover, by applying additional methods of investigation, the non-falsified region of hypotheses shrinks until either the whole space of considered hypotheses is falsified – which may require new hypotheses – or until the remaining hypotheses allow for a mathematical proof of their correctness (which is rarely possible in reality). Even though questions of optimal choice of methods of investigation cannot be treated in this paper, emphasis will be put on stating its results in a way that may help in making these questions more precise.

The exact confidence region

The concept of hypothesis testing

Under a system analytical perspective the significance of the concept of hypothesis testing lies in its capacity to provide a consistent and generally applicable (canonical) specification of the region of rejection for each of the considered hypotheses (models) on incompletely observable system characteristics. This requires a method which allows investigation of the concerned system characteristics under each hypothesis, where the possible outcomes from application of the method form a space of which each region of rejection is a part. A method of investigation thus includes details of experimental design, primary observations and their method of sampling, transformation of primary observations (statistic), etc.. Its possible outcomes are viewed as a sample space, the elements of which are realizations of a sampling variable. Which distributions of the sampling variable (or a statistic, which may replace the sampling variable in various applications) are taken into consideration is determined by the hypotheses on the system characteristics and the method of sampling as part of the method of investigation. This lays the basis for testing the conformity of the hypothesis with the outcome of the investigation (experiment).

If the investigation yields a result (an observation, a sample) that has low probability under the hypothesis to be tested, this hypothesis will be rejected in accordance with the principle of falsification by improbability. The measure of conformity underlying the decision whether to reject a hypothesis is thus specified by the measure chosen to quantify the probability of the observation.

The set of all results with sufficiently low probability (falling below a "critical" value) then defines the region of rejection of the hypothesis, where it has to be kept in mind that this region depends on the method of investigation and the hypothesis. In contrast with the conventional definition in mathematical statistics, the region of rejection is completely specified by the hypothesis and by the critical value above which the hypothesis is to be falsified. The absence of effects of alternative hypotheses on the determination of regions of rejection reflects the system analytical principle of unconditional falsification by improbability.

Table 1: *Notation*

$S :=$ sample variable with values taken from the sample space $\mathcal{S}$;

$\mathcal{A}(\mathcal{S}) :=$ σ-algebra on the sample space $\mathcal{S}$;

$\mathcal{H} :=$ space of hypotheses;

$R_\alpha : \mathcal{H} \to \mathcal{A}(\mathcal{S})$, regions of rejection of individual hypotheses at the level α of significance;

$\mathcal{P}(\mathcal{H}) :=$ power set of $\mathcal{H}$;

$C_\alpha : \mathcal{S} \to \mathcal{P}(\mathcal{H})$, confidence regions of individual samples on the level α of significance;

$P_h :=$ probability measure resulting for the method of investigation under the hypothesis $h \in \mathcal{H}$ on the system characteristics; the notation $P_h(S = s)$ for the distribution of S will be used interchangeably as the probability or the probability density of sample s, depending on whether discrete of continuous sample variables are considered.

By definition, any region of rejection of a hypothesis $h \in \mathcal{H}$ (compare the notation and conventions in Table 1) must be of the form

$$R(h, \lambda) := \{s \in \mathcal{S} \,|\, P_h(S = s) \leq \lambda\},$$

where the non-negative number λ represents a potential critical value. Yet, since probabilities of individual samples tend to be exceedingly small or, as is the case with continuous traits, are given as probability densities, λ itself is a poor indicator of improbability. In fact, improbable samples are expected to occur with low probability, which implies that $P_h(S \in R(h, \lambda))$ be controlled in the first place. The level of this control is called the *level of significance* and is usually denoted by α.

Since with increasing λ, the sets $R(h,\lambda)$ form an ascending sequence, the probability $P_h(S \in R(h,\lambda))$ also increases with λ, so that there exists a unique value $\lambda_{h,\alpha}$ as the smallest λ for which

$$P_h(S \in R(h,\lambda)) \leq \alpha \text{ and for all } \lambda' > \lambda \text{ either}$$
$$P_h(S \in R(h,\lambda')) = P_h(S \in R(h,\lambda)) \text{ or } P_h(S \in R(h,\lambda')) > \alpha.$$

The slight complication of the condition arises from the fact that $P_h(S \in R(h,\lambda))$ need neither be a strictly increasing nor a continuous function of λ. By this, the *critical value* $\lambda_{h,\alpha}$ of the test is unambiguously specified, and the *region of rejection* $R_\alpha(h)$ of the hypothesis h at a level α of significance has the canonical representation

$$R_\alpha(h) = \{s \in \mathcal{S} \,|\, P_h(S = s) \leq \lambda_{h,\alpha}\}. \tag{1a}$$

For given level α of significance, R_α constitutes a mapping assigning hypotheses to subsets from the sample space as stated in Table 1. The complement $\mathcal{S} \setminus R_\alpha(h)$ of $R_\alpha(h)$ is frequently referred to as the *region of acceptance* of the hypothesis h. As pointed out above, adherence to the system analytical principle of falsification forces us to regard this region as the set of sample outcomes that do not lead to rejection of the hypothesis.

At this point, the difference of the conventional approach from the specification of regions of rejection becomes apparent. In this approach any subset R from the sample space fulfilling the condition $P_h(S \in R) \leq \alpha$ and a variable number of conditions concerning type II errors connected with a designated set of alternative hypotheses could function as region of rejection. In contrast, following the above principle leads to a specification that is unique for the hypothesis to be tested.

The thus defined regions $R_\alpha(h)$ of rejection can now be shown to form the basis for the theory of *exact tests*. The relationship to this theory is drawn by the sets

$$R(h, P_h(S = s)) = \{s' \in \mathcal{S} \,|\, P_h(S = s') \leq P_h(S = s)\}$$

of samples with probability equal to or lower than that of the obtained sample s. On the basis of the observation $S = s$, the exact test rejects the hypothesis h on a level α of significance if

$$\zeta(h,s) := P_h(S \in R(h, P_h(S = s))) \leq \alpha,$$

i.e. if *the probability of not obtaining a sample more probable than* s does not exceed α. This verbal characterization of $\zeta(h,s)$ emphasizes its significance as an indicator of trustworthiness of the hypothesis. To reflect this characteristic of the probability $\zeta(h,s)$ as the central decision variable in the theory of exact testing, be referred to it will in the sequel as the *confidence* in the hypothesis h supported by the observation $S = s$. In the technical literature, ζ or analogously defined quantities are occasionally referred to as "p-value" or "significance probability" (see e.g. Barnett 1982, p.130, where $P_h(S = s)$ plays the role of the author's test statistic, however, for one hypothesis h only; this emphasizes the intrinsic difference between "confidence" and "significance probability"). However, because of the variable usage and poor expressivity of these terms, "confidence" will be preferred for the designation of ζ.

As is easily realized, the region $R_\alpha(h)$ of rejection can be equivalently stated as

$$R_\alpha(h) = \{s \in \mathcal{S} \,|\, \zeta(h,s) \leq \alpha\}, \tag{1b}$$

which, in accordance with the theory of exact tests, reveals the region of rejection as the set of all sample outcomes, the confidence of which does not exceed the level α of significance. Since the supremum of the confidence taken over the total sample space always equals 1, the region of rejection can never extend to the whole sample space for $\alpha < 1$.

The concept of confidence regions

Fiducial or confidence regions consist of hypotheses that in some sense comply with an observation. In system analytical terms, a confidence region summarizes all hypotheses from the space under consideration which are not falsified by the information obtained from a sample. This characterizes confidence regions in a manner analogous to the acceptance regions in hypothesis testing. Thus having specified regions of rejection as a mapping R_α for all hypotheses, a family of confidence regions $C_\alpha(s)$ for samples $S = s$ can be derived as sets of hypotheses $h \in \mathcal{H}$ which an observation $S = s$ does not allow to reject at the level α of significance, i.e. for which $s \notin R_\alpha(h)$ (see e.g. Ferguson 1967, p.258). The confidence region for $S = s$ at the level α of significance thus obtains the representation

$$C_\alpha(s) := \{h \in \mathcal{H} \,|\, s \notin R_\alpha(h)\}, \tag{2}$$

where C_α again appears as a mapping assigning to each element of the sample space a set of hypotheses (see Table 1). The two representations of the region of rejection given in equations (1a) and (1b) can now be used to express the definition of the confidence region $C_\alpha(s)$ in terms of the "critical value" and the "confidence":

$$C_\alpha(s) = \{h \in \mathcal{H} \,|\, P_h(S = s) > \lambda_{h,\alpha}\}, \tag{2a}$$

$$C_\alpha(s) = \{h \in \mathcal{H} \,|\, \zeta(h,s) > \alpha\}. \tag{2b}$$

The latter representation, which is directly derived from the theory of exact testing, justifies reference to $C_\alpha(s)$ as an *exact confidence region.* It furthermore makes apparent that the confidence region is empty (does not exist) if the supremum of the confidences in the sample does not exceed the level of significance. This is a relevant case, since special conditions on the space of hypotheses may imply that the supremum taken over this space does not reach a given value smaller than 1. It points at the possibility of inappropriate choice of the space of hypotheses for explaining the observation $S = s$. As will be returned to later, it may, however, also serve as a means for exact testing of composite hypotheses. In any case, the analogy between testing hypotheses and specifying confidence regions is now clearly seen to be established by the confidence $\zeta(h, s)$, which, for a specified level of significance, yields regions of rejection when varying observations for a fixed hypothesis h and yields regions of confidence when varying hypotheses for a fixed observation s.

The concept of hypothesis estimation

So far hypothesis testing and the dual specification of confidence regions have been treated without explicit reference to problems of estimating hypotheses. This reflects the system analytical perspective by drawing attention to details of the method of investigation (including testing procedures) as primary objects of epistemological improvement. Yet, it is inherent in this approach that any attempt to distinguish among the considered hypotheses (one of which is to be considered as an estimate) must be confined to those hypotheses which are not falsified by the observation. In other words, a meaningful estimation procedure must thus be minimally required to yield a result in which one can have confidence, i.e. that belongs to the confidence region.

Again several procedures are conceivable. The decision in favour of those hypotheses that assign to the observation highest probability has high plausibility and refers to one of the most popular methods of estimation, the maximum likelihood method. In this method, the probability $L(h, s) := P_h(S = s)$ forms the basis for selecting a hypothesis. To be an acceptable method in the above sense, it has to be made sure, however, that the hypotheses yielding maximum L for given s over the whole space of hypotheses $\mathcal{H}$ actually belong to the confidence region $C_\alpha(s)$. It is not self-evident that this condition is met in all cases, particularly if different spaces of hypotheses can be relevant in an analysis of the same system. Thus, until proven otherwise, one has to consider the possibility that the maximum likelihood method of estimation of hypotheses (parameters) may not be compatible with the specification of exact confidence regions for given levels of significance.

This possible restriction does not apply *a priori* when replacing the likelihood L by the confidence ζ. In the estimation procedure, maximization of confidence would then replace maximization of likelihood. Provided the confidence region for an observation is not empty, the supremum of ζ taken over the total space of hypotheses exceeds the level of significance by definition. Any estimate arrived at by this procedure thus belongs to the confidence region and, instead of assigning maximum probability of occurrence to the observation, it minimizes the probability of obtaining a result more probable than the observed. This underlines the direct orientation of ζ at the system analytical principle of falsification and distinguishes its position from that of L in procedures for the estimation of hypotheses.

Despite the elementary difference between ζ and L there exists an important relation between their maxima. This relation results from the fact that $\zeta(h,s)=1$ if and only if $P_h(S=s)=\max\{P_h(S=s')\,|\,s'\in\mathcal{S}\}$, i.e. if under the hypothesis h no observation has higher probability than the observation $S=s$. Consequently, if the likelihood of a hypothesis h is maximal for an observation s, and if under this hypothesis s has highest probability, then $\zeta(h,s)=1$, and the maximum likelihood estimate h is also a maximum confidence estimate. This condition may help to identify classes of probability distributions for which both methods of estimation yield the same results. In general, however, the choice of the space of hypotheses may imply that the maximum confidence supported by an observation does not reach a value of 1, so that the above condition has no basis.

Application to the space of binomial distributions

The concept developed above and the implied method for the specification of exact confidence regions will now be applied to the classical situation of analyzing hypotheses on the frequency of a defined phenomenon in a population. The associated method of investigation consists of taking random samples of fixed size with replacement from the population and observing the frequency of the phenomenon (sample frequency) in the sample. Sampling thus follows a binomial distribution, and the space of hypotheses is given by the parameters of these distributions (for further notational details see Table 2).

To simplify computation of confidences in hypotheses p supported by the observation $S_n=k$, it should be recalled that $P_p(S_n=i)$ increases with i as long as $i\le np$, and it decreases as soon as $i\ge np$. One therefore obtains for $k\ge np$ the confidence

$$\zeta(p,k)=P_p(S_n\in R(p,P_p(S_n=k)))=P_p(S_n\ge k)+P_p(S_n\le k'),$$

where k' is the greatest $i<np$ for which $P_p(S_n=i)\le P_p(S_n=k)$. The summand $P_p(S_n\le k')$ is dropped if $P_p(S_n=0)>P_p(S_n=k)$, since then an appropriate k' does not exist. In the reverse case where $k\le np$, the confidence equals

$$\zeta(p,k)=P_p(S_n\le k)+P_p(S_n\ge k'),$$

Table 2: *Notation*

$S =$ $S_n :=$ sampling variable specifying the frequency of a defined phenomenon (sample frequency) in a sample of size n.

$\mathcal{S} =$ set of non-negative integers $\leq n$ representing the sample frequencies in a sample of size n.

$\mathcal{H} =$ set of all binomial distributions for given sample size n and distinguished by the relative frequency (probability) p of the defined phenomenon. $\mathcal{H}$ will be identified with the closed interval $[0, 1]$. The probability distribution of the sample variable S_n for $p \in [0, 1]$ is described by

$$P_p(S_n = i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i}$$

and k' is the smallest $i > np$ with $P_p(S_n = i) \leq P_p(S_n = k)$. Again the summand $P_p(S_n \geq k')$ is dropped for $P_p(S_n = n) > P_p(S_n = k)$.

These representations show that as p moves from 0 to $\frac{k}{n}$, the confidence increases from 0 to 1 and then decreases again to 0 as p moves on to 1. During this change of p, k' may also change, which leads to discontinuity of ζ as a function of p. As can be taken from the examples presented in Figure 1, the initial increase and subsequent decrease of the confidence takes place in an essentially monotonic manner. This monotonicity is also expected from the corresponding behaviour of $P_p(S_n = k)$ as the maximum probability of the sum of probabilities in ζ. Therefore, there exist unique values p' und p'' of p such that $p' \leq \frac{k}{n} \leq p''$ and

$$\zeta(p', k) = \zeta(p'', k) = \alpha.$$

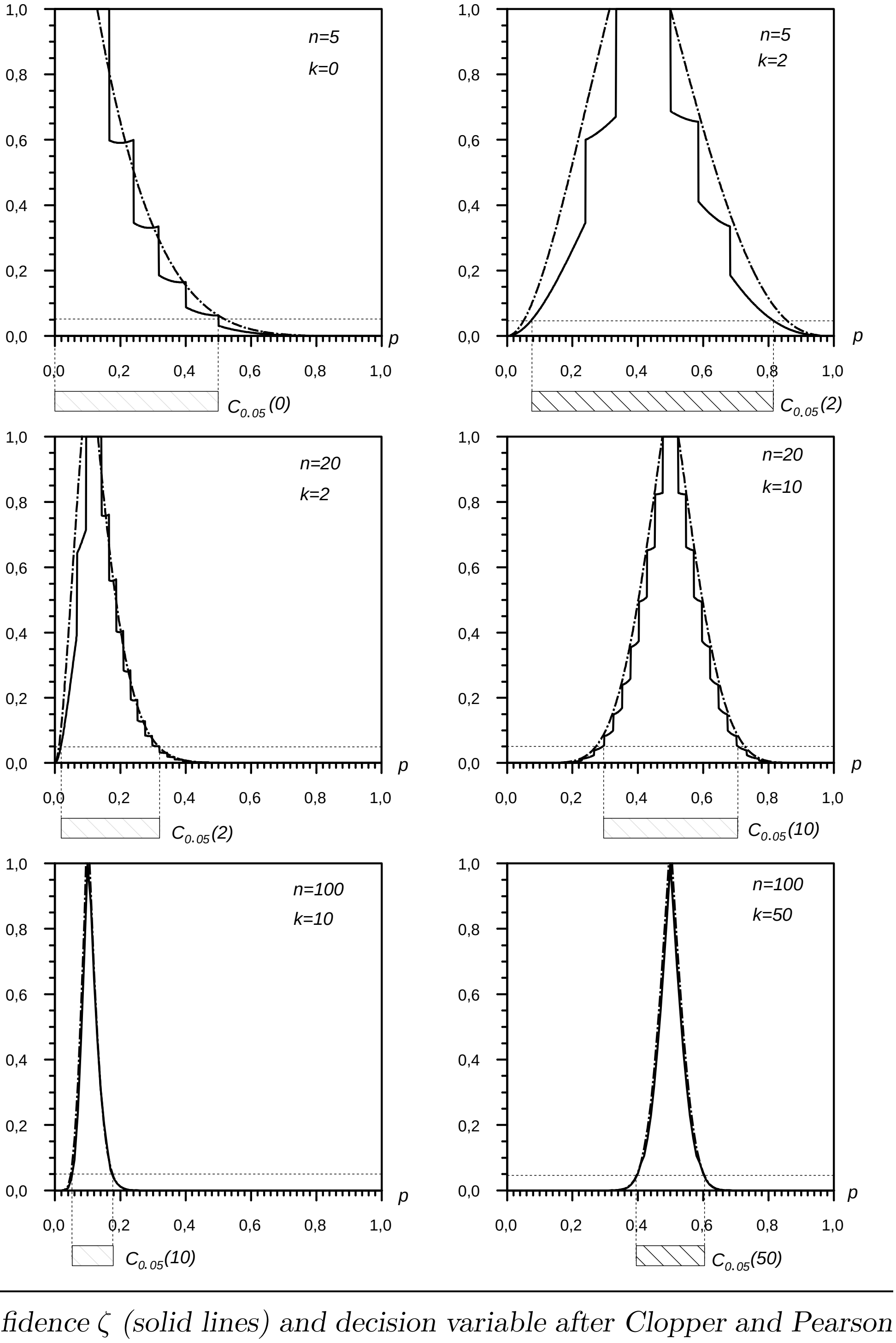
Since in the interior of the interval $[p', p'']$ the strict inequality $\zeta(p, k) > \alpha$ must hold, the confidence region for the observation $S_n = k$ attains the form of the open interval

$$C_\alpha(k) =]p', p''[.$$

However, as is indicated in the upper left example of Figure 1 ($n = 5$, $k = 0$), there is a chance of non-monotonicity at least for larger confidences. Hence, it cannot be ruled out that for large levels of significance confidence regions emerge with interspersed gaps.

The maximum confidence of 1 is reached for $p = \frac{k}{n}$. Hence, given a level of significance $\alpha < 1$, the method of parameter (hypothesis) estimation specified

Figure 1



Confidence ζ (solid lines) and decision variable after Clopper and Pearson (dashed lines) as function of the frequency parameter p of the binomial distribution. $n :=$ sample size, $k :=$ sample frequency, $C_\alpha(k) :=$ confidence region at level α of significance.

by maximization of the confidence yields the relative sample frequency $\frac{k}{n}$ as an estimate. Since maximization of the likelihood arrives at the same estimate, it is in this case compatible with the specification of exact confidence regions. Yet, as can be seen from the plateaus at $\zeta = 1$ extending around $p = \frac{k}{n}$ in Figure 1, maximum confidence estimates for p could cover a more or less small interval. Additional criteria, such as supplied by the maximum likelihood method, could thus be used to arrive at a single estimate which, besides belonging to the exact confidence region, fulfills the basically reasonable condition of maximizing confidence in the estimate.

The common direct method for the specification of confidence intervals (not mentioning the approximation by the normal distribution) we owe to Clopper and Pearson (1934), and the interval limits p' and p'' obtained by this method satisfy the equation

$$P_{p'}(S_n \geq k) = P_{p''}(S_n \leq k) = \tfrac{1}{2}\alpha.$$

Hence, the decision variable corresponding to ζ can be stated as $2{\cdot}P_p(S_n \geq k)$ for $p \leq \frac{k}{n}$ and $2{\cdot}P_p(S_n \leq k)$ for $p \geq \frac{k}{n}$. Apparently, the decision variable after Clopper und Pearson is not identical to ζ, which implies the possibility that they yield different specifications of the confidence region. As is illustrated in Figure 1, this could indeed happen for very small sample sizes, however, for larger sample sizes both variables differ only negligibly due to the dwindling effect of discontinuities in ζ. Thus, with the exception of very small sample sizes, the Clopper-Pearson method of specifying confidence intervals can be considered as a satisfactory approximation to the exact confidence region.

Concluding remarks

Even though the demonstration on binomial distributions pointed at the possibility that the system analytical perspective may lead to results that differ only negligibly from those of the common methods of statistical decision making, the difference in approach may cause basic problems to be viewed from different angles and, as a consequence, unfamiliar solutions may be reached. An example of this was provided by the subordination of the estimation problem to the problem of specifying confidence regions.

In particular, the decision to allow only for estimates of maximum confidence avoids problems of circularity arising from the fact that both estimation of a hypothesis and its testing refer to information derived from the same sample (see e.g. chapter 9.4 in Bishop et al. 1975). Confidence is the joint measure of conformity between observation and hypothesis (decision variable) for exact testing, specification of confidence regions and estimation of hypotheses. Therefore, if the maximum confidence supported by a given observation over a specified space of hypotheses exceeds the level of significance, all estimates of maximum confidence yield hypotheses that are not rejected by exact testing. Following from the conceptual relationship between critical value and level of significance, these findings apply to any type of probability distribution, including non-symmetrical and multi-modal.

Another important application of the method of maximizing confidence over special spaces of hypotheses is found in the field of exact testing of composite hypotheses, which can be any subset of a given space of hypotheses. For example, in the space of joint distributions of two random variables, stochastic independence between the variables defines a subset of hypotheses which is of particular interest in binary tests of association. If the method of investigation is again based on random sampling without replacement, ζ sums up probabilities of a multinomial distribution, which is represented as a vector of population frequencies in the space of hypotheses. The thus specified space $\mathcal{H}$ of hypotheses contains independent association as a subset $\mathcal{H}_I$ of vectors, the components of which result from products of frequencies from the two marginal distributions.

Depending on the observed vector s of sample frequencies, maximization of ζ over the subset $\mathcal{H}_I$ of independent association may then yield confidence values which fall below a specified level α of significance, and which thus lead to rejection of the hypothesis of independent association. On the other hand, if the maximum of ζ taken over $\mathcal{H}_I$ exceeds α, then exact confidence regions for the marginal distributions defining $\mathcal{H}_I$ can be obtained by consideration of $\{h \in \mathcal{H}_I \mid \zeta(h,s) > \alpha\} = C_\alpha(s) \cap \mathcal{H}_I$. This method avoids the problems associated with "conditional" tests such as Fisher's (1935) exact test, in which the conditioning hypotheses on the marginal distributions are extracted from the sample frequencies. Many other problems, such as those concerning one-sided or certain equivalence and sequential tests, can be stated in terms of testing composite hypotheses and can thus be treated via maximization of confidence in the same manner. For a one-sided test of a frequency parameter p, the subset of hypotheses forming the composite hypothesis could consist of all $p > p_0$, for example, and, given the observation s, $\zeta(\cdot, s)$ would have to be maximized over all $p > p_0$.

From this point of view, one-sided tests appear as a special case of equivalence testing, the essence of which consists in falsifying the null hypothesis that the truth lies outside of a range $\mathcal{H}_e$, say, of hypotheses that are considered to be "equivalent". Obviously, the composite null hypothesis is falsified if all single hypotheses outside $\mathcal{H}_e$ have confidence less than or equal to α, so that $C_\alpha(s) \subseteq \mathcal{H}_e$. In fact, this is the test for rejection of the null hypothesis of nonequivalence, the idea of which seems to be due to Westlake (1972). Here the close general relationship between hypothesis testing and confidence regions becomes especially clear, in that falsification of a (composite) hypothesis turns out to be tantamount to the confidence region being included in the complement of that hypothesis.

In essence, these examples are based on a model, in which the values of certain parameters are not specified and are thus "free" for adjustment of the model to the observation. The composite hypothesis then consists of the model together with the restrictions posed on the ranges in which the free parameters are allowed to vary. Rejection of this composite hypothesis therefore concerns the model together with the hypothesized parameter range; whether the structural characteristics of the model or the parameter range give rise to the rejection remains undecided. On the other hand, if the observations do not falsify the composite hypothesis, the underlying maximization of confidences justifies acceptance of the model with a restricted parameter range, the latter of which represents a confidence region.

As a final remark, the implication of the present approach for the evaluation of different methods of investigation of the same system characteristics with respect to their capacity for falsification will be briefly addressed. Fixing the system characteristics implies that the object of analysis, the space of hypotheses $\mathcal{H}$, also remains the same under the application of different methods of investigation, whereas these methods can differ in the sample variable and the sample space. A comparison of methods of investigation must thus rely on the space of hypotheses only. In essence, the capacity of a method to falsify hypotheses increases as the size of the confidence region narrows down to the hypothesis to be considered "true" on the basis of the respective observation. This relates to the idea of consistency of an estimator, where increasing precision can be achieved by the expectation that with increasing sample size the estimate ought to converge to the true value.

Hence, higher precision implies smaller confidence regions for the different observations. To avoid having to introduce a set measure on $\mathcal{H}$ in order to measure the size of the regions, it may be more practicable and meaningful to use set inclusion as a means for defining the relative precision of a method of investigation. Thus method A can be considered to be *more efficient* in falsifying hypotheses than method B at a given level of significance α, if for each observation a under method A there exists an observation b under method B, such that the confidence region $C_\alpha^A(a)$ for observation a is included in the confidence region $C_\alpha^B(b)$, i.e. $C_\alpha^A(a) \subseteq C_\alpha^B(b)$ (a and b are elements of the sample spaces defined by method A and B, respectively).

Extending this characterization to all levels of significance, and thus distinguishing A as *uniformly more efficient* than method B, requires that to each observation a under method A there exists an observation b under method B for which $\zeta^A(h, a) \leq \zeta^B(h, b)$ for all $h \in \mathcal{H}$, i.e. for which no hypothesis has confidence of higher support by a than by b. Apparently, and as demonstrated in Figure 1, this situation applies to changes in method obtained from increasing the sample size. Other changes such as between sampling strategies, however, may increase the efficiency of falsification at some levels of significance but not uniformly. Despite certain resemblance in terminology it should be recalled that the object of optimization is here the confidence region rather than the region of rejection usually put in the foreground of statistical test theory.

Acknowledgement – The author owes to E. Gillet, W.J. Grüneberg, J. Saborowski, M. Ziehe and two anonymous reviewers many comments which helped considerably in sharpening the views and formulations in this paper.

References

Barnett V. 1982. Comparative Statistical Inference. John Wiley & Sons, Chichester etc.

Bishop Y.M.M., S.E. Fienberg, P.W. Holland 1975. Discrete Multivariate Analysis. MIT Press, Cambridge, Massachusetts

Clopper C.J., E.S. Pearson 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413

Edwards A.W.F. 1972. Likelihood. Cambridge Univ. Press, Cambridge

Ferguson T.S. 1967. Mathematical Statistics. Academic Press, New York, London

Fisher R.A. 1935. The logic of inductive inference. Journal of the Royal Statistical Society, Series A 98: 39-54

Popper K.R. 1968. Conjectures and Refutations: The Growth of Scientific Knowledge. Harper Torchbooks, Harper & Row, New York

Weerahandi, S. 1995. Exact Statistical Methods for Data Analysis. Springer Series in Statistics, Springer-Verlag, New York etc.

Wellek, S., J. Michaelis 1991. Elements of significance testing with equivalence problems. Meth. Inform. Med. 30: 194-198

Westlake, W.J. 1972. Use of confidence intervals in analysis of comparative bioavailability trials. J. Pharm. Sc. 61: 1340-1341