The system analytical approach to the study of hypotheses

Hans-Rolf Gregorius

Institut für Forstgenetik und Forstpflanzenzüchtung
Universität Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

In experimental sciences, the analysis of real systems usually follows the steps diagrammed in Figure 1. Within this framework, the analytical activities are governed by the concepts and methods that are needed (*i*) to describe the **characteristics of the system** to be studied (including characteristics of a system's state or dynamics and input-output relations), and (*ii*) to detect the **causal mechanism** that produces these characteristics.

≪ **Illustration 1:** In studies of plant mating systems, the proportion of self-fertilization among the seeds of an ovule or pollen parent or of a population frequently constitutes an important *system characteristic*. The *causal mechanisms* may be governed by floral structure, pollen dispersal mechanisms, population density, incompatibility, etc. ≫

Item (*i*) employs a *method of investigation* (abbreviated MI) of the system characteristics to be analyzed. A MI includes methods of identification, observation, sampling, and data transformation. Item (*ii*) concerns the development of a *model system* for the purpose of presenting a hypothesis about the causal mechanisms of the real (actual) system and allowing a test of the discrepancy between the hypothesis and the observations made in the actual system. MIs may involve models that specify the techniques of observation or sampling, for example. When these models include hypotheses on causal mechanisms, they must be considered as part of item (*ii*). The epistemological principle consists in the **falsification of hypotheses**. Although failure to falsify a hypothesis leads to its acceptance, the possibility is not ruled out that other hypotheses and their associated models may exist which likewise would not be falsified by the observations.

With reference to Figure 1, these analytical objectives mainly range on the levels "model validation" and "methods of investigation". The remaining levels are assumed to have been specified with due precision. Therefore, a *method of system analysis* basically consists of

- ▷ a model system that reflects the hypotheses on the causation of real system characteristics,
- ▷ a MI that allows appropriate observation of the system characteristics to be studied, and
- ▷ a method of measuring the discrepancy between the (actual) observations on the system characteristics and the modeled observations that represent the hypothesis.

A method of system analysis becomes *qualified* with respect to the epistemological principle, if these three constituents guarantee **falsifiability** of the associated hypothesis. Thus, the method of system analysis is unqualified if its MI rules out the existence of observations that would contradict the characteristics of the model.

Figure 1

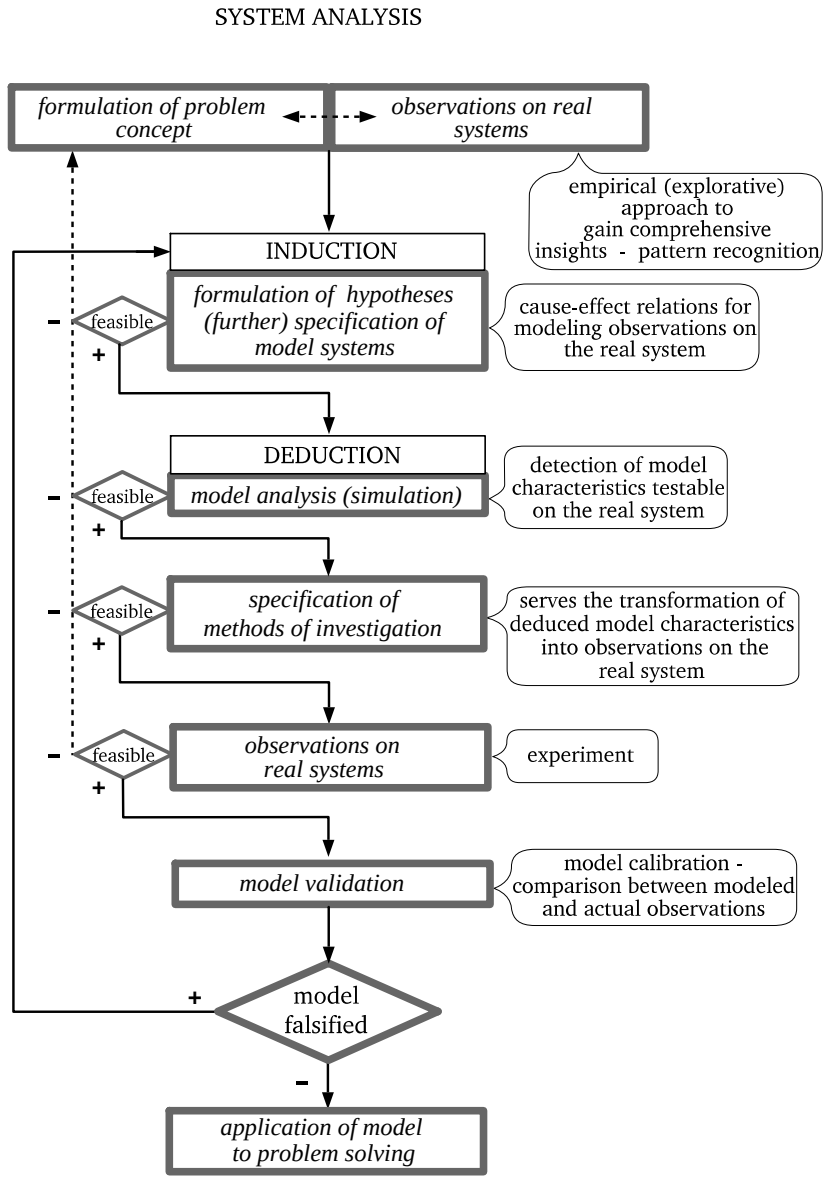


Diagram of the major successive steps involved in the analysis of real systems (modified after Berg and Kuhlmann 1993, p. 13)

General procedure for an analysis of causal mechanisms

In the following, unspecified values of variables which appear as independent variables in a model will be referred to as *free parameter values*. The system characteristics observable by application of the MI, i.e. the *actual observations*, will be distinguished from their reproductions by the model, which will be termed the *modeled observations*. Actual and modeled observations must be comparable but need not be of the same type. The system characteristics of interest are called *target characteristics*, and they may or may not be actually observable. If they are not, other model characteristics must be observable, and the target characteristics must be functions of the observable characteristics or of the free parameter values on which the observable characteristics depend. All model characteristics are thus considered as functions of the free parameter values if the model

contains such values. Having specified the model system, the general procedure of the analysis of the hypothesis consists of the two basic steps specified in Table 1.

Table 1

- *Analysis of hypotheses*
  - ▷ *calibrate* the model by adjusting its free parameters values so as to minimize the discrepancy between the modeled and the actually observed system characteristics; the thus obtained modeled observation will be referred to as the *prediction*; constraints on the parameter values are part of the model and must be obeyed in the calibration; if the model contains no free parameter values, the modeled observation is unique and equals the prediction (for an illustration see Figure 2);
  - ▷ reject the causal hypothesis (the model) in the case of excessive discrepancy between prediction and actual observation.

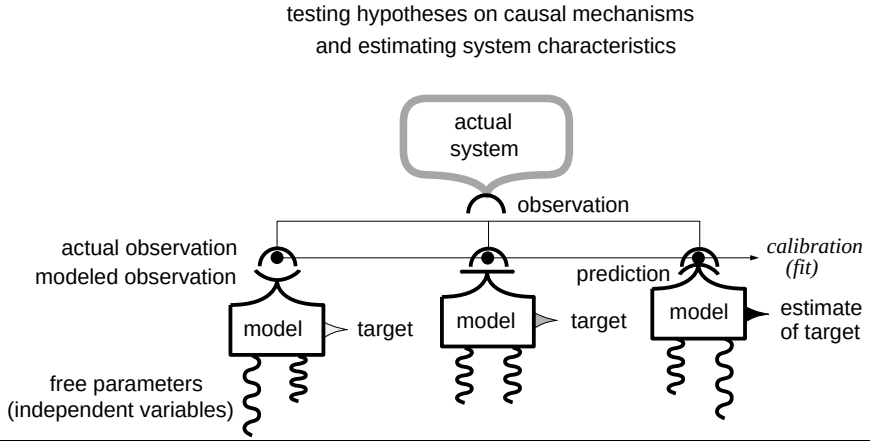
Notes: Calibration of the model requires a measure of discrepancy that is consistently defined for all relevant pairs of modeled and actually observed system characteristics.

≪ **Illustration 2:** In the mixed mating model (random cross-fertilization, constant ovule selfing rate), the observations to be modeled are genotypic frequencies among the offspring of a population; for a single gene locus let P_{ij} be the frequency of parental individuals carrying the i-th and j-th allele, and let P'_{ij} be this frequency among their offspring; the allele frequencies in the parental population are denoted by p_i, and s is the proportion of offspring from selfing. The model system then implies $P'_{ii} = p_i^2 + s(\frac{1}{2}(p_i + P_{ii}) - p_i^2)$ and $P'_{ij} = 2p_ip_j - s(2p_ip_j - \frac{1}{2}P_{ij})$ for $i \neq j$. Given that observations can be carried out only among the offspring, the P'_{ij} constitute the modeled observations (observable system characteristics). The P_{ij} and s constitute the free parameters if they are not specified for given reasons. Apparently, this mating system model does not provide a hypothesis on a causal mechanism of selfing. Instead, the selfing rate appears as a free parameter value that can be used in the calibration of the model. The calibration proceeds by adjusting the selfing rate and parental genotypic frequencies such that the actual observations $\tilde{P}_{ij}$, say, are approached by the modeled observations P'_{ij} as closely as possible. ≫

Testing models

The *principle of testing models* addressed above is based on the stimulus-response concept of systems. On the level of model systems, stimuli (independent variables, inputs) appear as free parameter values, and responses (dependent variables, outputs) are functions of these values. In causal analyses, observable system characteristics may refer to responses and free parameter values, or to responses alone, but never to free parameter values alone (in Illustration 2 only the responses P'_{ij} were considered as observable system characteristics; the free parameters P_{ij} can, however, also be observable when genotypic frequencies can be studied among parents). Free model parameter values are adjusted so as to minimize the discrepancy between modeled and actually observed system characteristics (model calibration). The adjustment may thus concern free parameter values as determinants of observable system responses and these values themselves when they are observable.

Figure 2



Testing hypotheses on causal mechanisms and estimating system characteristics through adjustment of free model parameter values (model calibration) in order to minimize the discrepancy between the modeled and the actual observations.

By this it is guaranteed that the stimulus-response (input-output) relationships of the model system, which specify its structural (relational) characteristics, mirror those of the real system as closely as possible under the limitations of the observable system characteristics. The hypothesis that the modeled system correctly maps the actual system is rejected if the minimized discrepancy between them is excessively high. Thus rejection concerns the model as such and therefore applies to its structural (relational) characteristics as well as to those parameter values not involved in the model calibration (thus excluding the free model parameter values). In other words, falsification refers to the causal mechanisms that are employed in the model system to produce its response (output) from its stimulus (input).

≪ **Illustration 3:** If in the mixed mating model in Illustration 2 only the genotypic frequencies among the offspring are observable and all other parameters are free, then rejection of this model is equivalent to rejection of at least one of its assumptions: random cross-fertilization, equal ovule selfing rates, or random fusion of gametes in selfing.

On the other hand, if the hypothesis of a mixed mating model at equlibrium is to be tested, then the relevant model results from equating in Illustration 2 the P'_{ij} with the P_{ij}, which yields $P_{ii} = ((1-s)p_i^2 + \frac{1}{2}sp_i)/(1-\frac{1}{2}s)$ for the homozygotes and $P_{ij} = 2p_ip_j(1-s)/(1-\frac{1}{2}s)$ for the heterozygotes. The observable model characteristics are now the P_{ij}, and the free parameters are the allelic frequencies p_i and the selfing rate s. Hence, rejection of this equilibrium model includes as an additional cause the assumption that the population is at equilibrium. ≫

Critical regions of observations

An alternative and occasionally preferred way of formulating the principle of testing is based on sets of potential observations of the actual system that, if realized, would falsify a given hypothesis. Such sets are called "critical regions" or "regions of rejection", and they can be determined before making observations. More precisely,

- for a given model, the *critical region* (region of rejection) consists of all po-

tential observations on the actual system which show excessive discrepancy from their model predictions on a given level; the complement of the critical region is termed *non-critical region* (region of acceptance).

Here, the model prediction again equals the modeled observation which results from calibration of the model, and which thus yields minimum discrepancy between modeled and potentially actual observation. Non-critical regions show direct duality with confidence regions via the measure of discrepancy between modeled and actual observations and via the given level of discrepancy.

Estimation of target characteristics and determination of confidence regions

In many (probably most) cases MIs are not available which allow for direct and complete observation of the target characteristics. Observations based on random samples are *incomplete observations*, for example. *Indirect observations* employ models to infer target characteristics from other directly observable characteristics. The lowest degree of precision is realized for indirect and incomplete observations. If any of these situations (incomplete or/and indirect observability) holds true, target characteristics can by definition only be estimated but not determined. The pertaining procedure is called *estimation* and is carried out as presented in Table 2.

Table 2

- *Estimation of not directly or incompletely observable target characteristics*
  - ▷ design a model of the actual system, (1) in which the target characteristics appear as a function of the free parameter values (including the case of identity with some of these parameters), (2) for which further system characteristics exist that can be directly and completely observed under the MI, and (3) that depend on the same free parameter values as the target characteristics;
  - ▷ on the basis of these observations, continue with the general procedure of the analysis of hypotheses as in Table 1;
  - ▷ in the case of not having to reject the model, the target characteristics associated with the adjusted (calibrated) free parameter values of the model system constitute *estimates* of these characteristics of the actual system (for an illustration see Figure 2).

Notes: *Direct* are distinguished from *indirect estimates* according to whether the estimate is based on direct or indirect observations. Model testing and estimation refer to the same calibrated parameter values.

The dependability of an estimate is determined by (a) the discrepancy between its associated model predictions and the actual observations (which is the minimum discrepancy between the modeled and the actually observed system characteristics) and by (b) the set of alternative target characteristics, the associated observable characteristics of which show a discrepancy from the actual observations that is sufficiently small not to justify rejection of the model. The latter set is commonly referred to as confidence region. More precisely

- the *confidence region of an estimate* is defined as the set of all target characteristics with discrepancy between the associated observable model characteristics and the actual observation that does not exceed a level effecting rejection of the model.

Empty confidence regions are equivalent to rejection of the model, non-empty confidence regions contain the estimate, and the larger the confidence region, the less precise is the estimate.

≪ **Illustration 4:** With the exception of some very rare situations, MIs allowing direct observation of selfing rates are not available, but these rates are frequently the target characteristics. In the mixed mating model of Illustration 2 the ovule selfing rate is a free parameter which could be used for indirect observation and thus for estimation of this rate as target characteristic of an actual mating system. This can be achieved, if a MI exists that allows observation of genotypic frequencies among the offspring. These observations are completely determined in the model by the genotypic frequencies among the parents and the selfing rate as free parameters. Hence, the observable characteristics depend on the same free parameter values as the target characteristic (which is indeed identical to one of the free parameters). Model calibration then yields an indirect estimate of the selfing rate. It has to be considered, however, that other models that also include ovule selfing rates as free parameters or as functions of these (see e.g. Gregorius et al. 1987) may yield different estimates. ≫

Statistical aspects of system analysis – incomplete observations

From a system analytic perspective, two aspects are of basic importance in studies involving MIs that can only yield incomplete observations (the realm of statistics): the measurement of discrepancy between actually observed and modeled system characteristics, and the limited dependability of this measurement as a consequence of randomness in the MI. Clearly, any reasonable planning of an experiment attempts to design the MI to ensure the sufficient dependability of the observations under the condition that the hypothesis being analyzed is true. Based on this condition, a measured degree of discrepancy between actual and modeled observations that exceeds some threshold discrepancy would be conceived of as a safe indication of inadequacy of the hypothesized model. Hence, if the model correctly maps the actual system, then the MI is required to produce with "sufficiently" low probability observations that show "unacceptably" large discrepancy from the model characteristics. More precisely, this involves

- ▷ determination of a threshold discrepancy δ between the actually observed characteristics and the model characteristics under consideration, below which observations are considered to be *representative* of the model characteristics and thus give no reason to reject the model. Conforming with common statistical terminology, δ will be called the *critical discrepancy* for representativity;
- ▷ utilization only of MIs for which the probability of obtaining observations with discrepancy from the model characteristics that is greater than or equal to the critical discrepancy δ, does not exceed a given *level of significance* ε. This exceedence probability is called the *critical probability of a MI*, again

conforming with statistical terminology. MIs which fulfill this prerequisite will be called *qualified* on the levels δ and ε for the analysis of the model characteristics. (This terminology reflects that introduced earlier for methods of system analysis.)

Hence, the smaller the critical probability becomes for a given critical discrepancy δ and different MIs, the smaller ε can be chosen and, consequently, the larger is the system analytic *dependability* on the MI of the model characteristics to be analyzed. Increasing the sample sizes may be a means of raising this dependability (decreasing ε) in some cases. Determination of the qualification of a MI is of considerable concern in classical statistical applications (see e.g. Sokal and Rohlf 1981, p.262).

≪ **Illustration 5:** Consider the situation of testing a hypothesis about the relative frequency p of a phenomenon in a population. The MI consists of randomly sampling n times with replacement and recording the relative frequency of the phenomenon in the sample. Then the (incompletely) observable system characteristic is p, the actual observation equals the relative frequency of the phenomenon in the sample, and the requirement of randomness of sampling in the MI is modeled in terms of a binomial distribution with parameters p and n (which constitute the model parameters). Furthermore, let the measure of discrepancy be given by the absolute difference between the relative frequency in the sample and in the population, and assume that the *levels of qualification of the MI* are $\delta = 0.08$ and $\varepsilon = 0.05$ for the model characteristic p. Then, for $p = 0.1$ and $n = 70$, the critical probability equals 0.03. This probability is below the level $\varepsilon = 0.05$ of significance, which renders the MI with $n = 70$ qualified. For a smaller sample size, $n = 30$ say, the critical probability equals 0.12, which is distinctly above the level of significance $\varepsilon = 0.05$ and thus reveals that a MI based on this sample size is unqualified for testing the hypothesis $p = 0.1$. ≫

Testing models without free parameter values

Consider a model that contains no free parameter values, and for which a MI exists that is qualified on the levels δ (critical discrepancy) and ε (significance level) for an analysis of given model characteristics. If an actual observation had discrepancy from the model characteristics of at least δ, the model would be rejected. The justification of the rejection would result from the fact that if the model were valid, this observation were not representative of the model characteristics (discrepancy $\geq \delta$) and were rendered very unlikely (critical probability $\leq \varepsilon$) by the qualification of the MI for the analysis of the hypothesized model (the first statement in Table 3 includes this situation of model testing as a special case). Hence, to allow model rejection *both the critical discrepancy and the level of significance need to be controlled*.

This method of testing hypotheses differs from that common in classical statistics, where only the level of significance is controlled, and where the critical discrepancy is completely determined by the level of significance. For each level of significance ε, a critical discrepancy δ is determined as the minimum discrepancy for which the critical probability does not exceed ε. In this assignment, δ increases with decreasing ε, which reflects the obvious fact that an increase in the demands on the dependability of the MI can only be compensated by a change of the MI, if the level of representativity of the observa-

tions is to be maintained. In accordance with the above principle of testing, hypotheses are rejected for actual observations with discrepancy from the model characteristics of at least the ε-dependent (!) critical discrepancy.

The consequences of controlling one vs. two levels can be demonstrated more clearly with the help of the *significance probability of an actual observation* which is defined as the probability of obtaining an observation from the MI that shows discrepancy from the model characteristics that is not smaller than that of the actual observation (the terminology follows common usage in statistics, see e.g. Barnett 1982, p.130). Controlling only the level of significance leads to rejection of a hypothesis if the significance probability of the actual observation does not exceed ε. In this case, no criteria are specified, which allow to judge the quality of the MI underlying the test.

On the other hand, controlling both the significance level and the critical discrepancy allows judgement of the quality of the MI before it is accepted as a decisive basis for testing. Under the premise of a qualified MI, rejection on the basis of excessive discrepancy ($\geq \delta$) of an actual observation implies for this observation a significance probability $\leq \varepsilon$. However, a non-critical discrepancy ($< \delta$) of an actual observation does not rule out the possibility that the pertaining significance probability is $\leq \varepsilon$. This would indicate a quality of the MI that is higher than demanded for the detection of non-representative observations. Yet, given an observation with significance probability $\leq \varepsilon$ and discrepancy $< \delta$, the MI is seen to be qualified, and the hypothesis is not rejected under the control of both δ and ε; it is, however, rejected if only ε is controlled. Otherwise, if the significance probability exceeds ε and the discrepancy of the observation is $\geq \delta$, then the hypothesis is not rejected when controlling ε only, but the MI turns out to be unqualified when both δ and ε are controlled.

≪ **Illustration 6:** The example in Illustration 5 can again be chosen to demonstrate the problem of controlling only the level of significance ε when testing hypotheses, instead of controlling both ε and the critical discrepancy δ. In this example, a MI with sample size $n = 30$ was shown to be unqualified for testing the hypothesis $p = 0.1$ on the levels of qualification $\varepsilon = 0.05$ and $\delta = 0.08$. Irrespective of the actual observation, no decision on rejection of the hypothesis can be made on the basis of this MI. On the other hand, if only the level of significance were controlled, and given that the phenomenon of interest is actually observed $k = 7$ times in the sample, the significance probability of this observation would equal 0.026. Since 0.026 is less than ε, this would have implied rejection of the hypothesis on the basis of this observation. Moreover, the discrepancy of the observation from the model characteristics equals $|p - (k/n)| = 0.133$. In order to justify rejection on the basis of a qualified MI, the condition δ for qualification would have to be relaxed such that for some $\delta \leq 0.133$ the critical probability is $\leq \varepsilon$ (=0.05). ≫

Testing models with free parameter values

A different situation arises when a model contains free parameter values. In this case a given MI may be qualified on the levels δ and ε for the model characteristics associated with some parameter values but not with others (recall that model characteristics are a function of the free parameter values or are identical to some or all of these). This follows from the fact that the probability distribution specified by the MI depends on

the respective model characteristics and thus on the free parameter value. Consequently, it may happen that only a restricted set of free parameter values is amenable to system analytic treatment due to the non-comprehensive qualification of the MI. Only these amenable parameter values can be tested with this MI. For the non-amenable parameter values, the model can neither be rejected nor accepted.

As was emphasized above, specification of the ranges of free parameter values is an integral part of the model, so that testing procedures should ideally apply to the entire set of free parameter values. Hence, for fixed levels of qualification δ and ε, a MI would be ideal if it were qualified for all of the model characteristics associated with the entire range of free parameter values. To simplify the wording, usage of the term "qualification" of a MI will be extended in the following to the free parameter values associated with the model characteristics. The testing procedure appropriate for this case is stated in Table 3. In accordance with the specification in Table 1, model calibration again forms the basis of the testing procedure in combination with a defined measure of discrepancy. Incompleteness of observations is accounted for here by an appropriate choice of the MI.

Table 3

• *Analysis of hypotheses for incomplete observations*

Given the levels δ and ε on which the MI is to be qualified, then

- ▷ reject the hypothesized model if its calibration yields predictions (see Table 1) for which the discrepancy from the actual observation exceeds or equals the critical discrepancy δ, and for which the MI is qualified;
- ▷ do not reject the hypothesized model if there exist model characteristics (produced by their underlying free parameter values) which show discrepancy from the actual observation that lies below δ and for which the MI is qualified; the hypothesis is called *qualifiedly non-rejectable* in this case.

Notes: No decision as to rejection or acceptance of the hypothesis can be made in two cases: (1) the MI is not qualified for the model predictions and their discrepancy from the actual observations exceeds or equals δ, and (2) no free parameter values exist for which the associated model characteristics have discrepancy less than δ and for which the MI is qualified. The method of system analysis is unqualified in both cases.

Cases of indetermination arise only if the MI is qualified for some but not all free parameter values. This does not yet imply that the chosen method of system analysis is totally unqualified, as follows from the statements in Table 3. Both rejection and acceptance are possible despite the fact that the MI is not qualified for all free parameter values. On the other hand, even if the discrepancy exceeds or equals the critical discrepancy for all free parameter values for which the MI is qualified, the model is not rejectable on a qualified basis if the qualification of the MI does not extend to the model predictions. The method of system analysis is merely unqualified for such observations. To overcome this situation, either the levels of qualification δ and ε must be relaxed, or a new experiment must be performed on the basis of a different MI.

Confidence regions and estimation

As was demonstrated above, estimation of target characteristics becomes relevant only for non-rejectable models with free parameter values. Estimates can therefore be selected only among those target characteristics which are associated with the free parameter values that make the model qualifiedly non-rejectable. This corresponds to the above general definition of the confidence region, and we can therefore now

- define the *confidence region of estimates* as the set of all target characteristics that are associated with observable model charateristics for which the MI is qualified, and which show discrepancy from the actual observation that lies below the critical value.

Recall that target characteristics are associated with the observable characteristics via the free parameter values on which the target characteristics completely depend and which determine the observable characteristics.

Reasonable (direct or indirect) estimates must consequently belong to the thus qualified confidence region. Applying the model calibration principle, the estimate would consistently result from minimizing the discrepancy between the actual observation and all observable model characteristics with associated target characteristics belonging to the confidence region. Belonging to the confidence region makes sure that the MI is qualified for all of the concerned free parameter values. Yet, since the MI may not be qualified for the model predictions, the corresponding target characteristics may be excluded as estimates. Moreover, minimization of the discrepancy need not go along with maximization of the qualification of the MI in the sense that the critical probability is minimized, since this probability does not depend on the actual observation. Minimization of discrepancy also need not simultaneously maximize the probability (likelihood) of the observation even when normalized for the respective maximum probability. It may therefore be desirable to specify the MI and the measure of discrepancy such that they yield similar results in the calibration procedure.

≪ **Illustration 7:** Consider the equilibrium mixed mating model presented in Illustration 3 for two alleles. The observable model characteristics P_{ij} $(i, j = 1, 2)$ result from variation of the allele frequency p_1 and the selfing proportion s as free parameters and, by this, form a subset of the two-dimensional frequency simplex which is defined by $P_{12} \leq 2p_1p_2$ (set of frequency vectors on or below the Hardy-Weinberg parabola). Furthermore, define the measure of discrepancy by the metric $d(P, P') := \frac{1}{2}\sum_{i\leq j} |P_{ij} - P'_{ij}|$ on the frequency simplex. Model calibration for the free parameters p_1 and s then leads to a predicted vector of genotypic frequencies which equals the actually observed vector $\tilde{P}$, say, if $\tilde{P}_{12} \leq 2\tilde{p}_1\tilde{p}_2$. Otherwise, if $\tilde{P}_{12} > 2\tilde{p}_1\tilde{p}_2$, minimization of the d-distance of the actual observation from the set of observable model characteristics yields $s = 0$ so that the model predicition equals Hardy-Weinberg-proportions.

Given a MI that employs random sampling of $n = 30$ offspring with replacement and records relative frequencies of the three genotypes in the sample as observation. Let $\delta = 0.08$ and $\varepsilon = 0.05$ be the desired levels of qualification of the MI. The set $\mathcal{Q}$, say, of modeled genotypic frequency vectors for which the MI is qualified results from collecting those vectors for which the critical probability is less than or equal to 0.05. If $\tilde{P}$ again designates the actual observation (relative fre-

quencies of genotypes in the sample), the pertaining confidence region results from the intersection of the set $\mathcal{Q}$ with the ball of radius δ and center $\tilde{P}$. If, in addition, $\tilde{P}$ belonged to $\mathcal{Q}$, so that $\tilde{P}_{12} \leq 2\tilde{p}_1\tilde{p}_2$, the estimate of the allele frequency and the selfing proportion equalled $\tilde{p}_1$ and $(2\tilde{p}_1\tilde{p}_2 - \tilde{P}_{12})/(2\tilde{p}_1\tilde{p}_2 - \frac{1}{2}\tilde{P}_{12})$, respectively.

However, computer calculations for the above values of n, δ and ε reveal a very weird structure of $\mathcal{Q}$ comprising coherent as well as strongly disconnected areas and even isolated points in the frequency simplex. The coherent areas are very small and restricted to the extreme allele frequencies. Hence, even if $\tilde{P}_{12} \leq 2\tilde{p}_1\tilde{p}_2$, the δ-ball surrounding the observation $\tilde{P}$ may in many cases show no and in other cases very complex intersections with $\mathcal{Q}$. Thus, the method of system analysis is either unqualified or the resulting confidence regions and estimates are highly sensitive towards changes in the observations. This emphasizes the importance of MIs which are qualified for the whole range of free parameter values. For the present levels of qualification, this can be realized for a sample size of $n = 200$. $\gg$

Measures of discrepancy

In statistics, the measures of discrepancy between actual and modeled observations involved in the model calibration are commonly termed "test statistics" or "test variables" (for a more generalizing definition see e.g. Weerahandi 1995, p.29). It is well known that the outcome of both testing and estimation procedures depends on the applied measure of discrepancy (see e.g. Bishop et al. 1975, chapter 14.7). This becomes especially evident when considering a particular value for the critical discrepancy. This value may be meaningfully specified for a bounded but not for an unbounded measure of discrepancy. Moreover, as was emphasized in the last chapter, it may be desirable to harmonize the MI with the discrepancy measure in a way that associates probabilities of observations with the discrepany of the observations from the model characteristics. This would require that with increasing discrepancy of an observation from the model characteristic the probability (density) of the observation deceases strictly, and that this holds for all probability distributions associated with the model characteristics and their underlying parameter values. As an example of how this harmonization can be completely realized, the exact p-value, which is the basis of many so-called "exact" statistical testing procedures, will be briefly addressed.

The *exact p-value* is commonly specified by the probability of not obtaining an observation from the model that is more probable than the actual observation (which can, in turn be considered as a special case of the definition used in Weerahandi, 1995, p.30). The exact p-value can thus be considered as a function of the potential actual observations, and it equals 1 only for actual observations that have maximum probability (density) under the respective free parameter value that determines the model characteristics. Thus 1 minus the exact p-value constitutes a measure of discrepancy between actual and modeled observations that becomes zero for observations of maximum probability and reaches its maximum value of 1 only for observations of probability zero. The probably undesirable fact that the model characteristics include aspects of the MI (via the probabilities) will not be argued further here.

It is easily realized that the exact p-value then presents itself as a significance probability based on this measure of discrepancy. In fact, this discrepancy measure equals 1

minus the significance probability. Critical discrepancies, which specify the required representativity of observations relative to the most probable observation under the model and MI, can be freely chosen between 0 and 1, and this choice essentially determines the corresponding critical probabilities. Hence, once the critical discrepancy is determined, a lower bound for the level of significance is fixed irrespective of the applied MI. This is a particular feature of the exact p-value method (Gregorius, 1996, provides a detailed analysis of the exact p-value, there called "confidence", under the system analytic perspective).

Model dependence of analysis

It follows from the above that "model independent" analyses are confined to situations where in the actual system the target characteristics are observable or estimable without having to resort to hypotheses on their causal development. These analyses are thus restricted to the description of system characteristics. Occasionally, the involved methods are referred to as "descriptive models" (see e.g. Bossel 1992 or Wissel 1992), particularly if the methods serve the condensation of data in the form of parameters obtained from fitting a function to the observations, for example. Yet, since causal hypotheses are not an explicit matter of consideration here, the term "model" may be a misnomer. This does not rule out the possibility that methods of description can appear as models in other contexts, as was demonstrated for the species-area curve by Wissel (1992).

The same applies to parameter-free methods of data transformation (indices, measures) which are derived as characteristics of special models. Even such conceptual methods of data condensation or transformation, which serve the representation of specific characteristics of broad classes of observations, may be intrinsically model-based, since the underlying concept utilizes cause-effect relations of a very general kind. As was emphasized above, the design of a MI may also have to rely on models if only incomplete observations are possible. In fact, all sample distributions could be viewed as models of the actual sampling procedure.

Models are the essence of any causal analysis, which is thus intrinsically "model dependent". The probably most extreme case of model dependence arises in the common situation in which target system characteristics are not accessible to direct observation, so that their study must completely rely on model dependent methods of analysis. The resulting "indirect estimates" of the target characteristics introduced in Table 2 may change considerably when the same observations are analyzed with a different model. It has to be taken into consideration, however, that such an alternative model constitutes a different causal hypothesis and must therefore first be subjected to a test. This is mandatory, since estimates are irrelevant if the test recommends rejection of the model. If more than one non-falsifiable model exists, the dependability of the respective estimates becomes a particularly important criterion of decision. Apparently, the model with the lower discrepancy between actually observed and predicted system characteristics and with the smaller size of the confidence region has the higher dependability. If two models turn out to be equivalent in these two aspects, deduction of additional observable model characteristics and development of appropriate MIs (cf. Figure 1) is required.

This makes it clear that *an estimate always directly reflects model characteristics*; characteristics of the actual system appear only indirectly through the adjustment of the model characteristics. Model aspects of the MI are included in this statement, as becomes apparent from the simple fact that the sample mean, for example, could not be an estimate of an expectation unless it occurred as an expectation in a model fitted to the sample.

Design of model

As can be taken from Figure 1, models ought to be inducible from a precisely formulated problem in order to avoid the danger of testable nonsense constructs (corresponding to the well-known nonsense correlations from statistics). For the same reason, the target characteristics must be deducible from the model in order to establish an explicit cause-effect mechanism, and a qualified MI must exist for validation of the model. The demands on the qualification of a method of system analysis increase with the model's number of free parameter values, since the model can be calibrated more precisely. In order to produce observations which can give rise to rejection of such models, the MIs must be increasingly elaborate and levels of qualification high, which, since difficult to realize, lowers the chances to detect inadequacy of the model. This corresponds to the obvious fact that *a non-falsified model becomes the more dependable the fewer free parameter values it has*, where this is achieved by specifying more and more of the formerly free parameter values by supplementary actual observations. Moreover, this explains the *precedence of simple over complex models* in cases of indecision (c.f. e.g. Wissel 1992, p. 3-5, for a critique of further features of complex models based on problems of their validation).

A special situation arises for conceptual models. These models are usually aimed at the analysis of hypotheses on general characteristics of a system, which must thus directly relate to fundamental conceptual properties of that system and are therefore expected to yield results of more general validity (see also Wissel 1992, p. 4). Irrespective of the number of free parameter values involved in the design of a conceptual model, the requirement of general validity may be expected to imply the possibility of precise calibration. Yet, in conceptual models, the formulation of hypotheses typically deals with constraints on the validity of system characteristics, which, according to the principle of testing, restricts calibration of the model to the free parameter values corresponding to these characteristics (see Table 1). The thus reduced opportunities for adjustment increase the chance of rejection of the hypothesis (compare the following Illustration 8 for mating preferences).

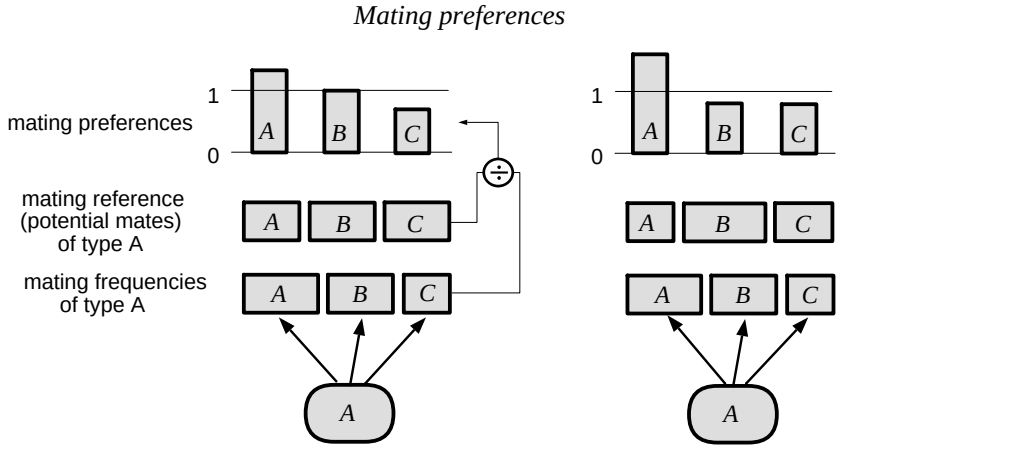
≪ **Illustration 8:** Mating preferences are the conceptual system characteristics of mating systems. Their determination requires specification of a trait, definition of mating events, and actual observations on frequencies of matings among the respective trait types as well as on frequencies of potential mating partners (mating reference, for an example see Figure 3). The causes of the conceptual characteristics can be hypothesized by modeling of suitable mechanisms. Conversely, each mating system model yields its particular mating preferences and can thus be characterized by special features and patterns of these preferences (for an overview see Gregorius

1989).

Model parameters that have no effect on mating preferences are irrelevant to the mechanisms hypothesized in the model mating system. Hence, the analysis of mating systems centers upon the description or estimation of special system characteristics (such as proportion of selfing), all of which must be shown to affect the mating preferences. Even in cases where no prior information suggests a hypothesis on an actual mating system, the description of mating preferences will help to recognize classes of mating systems which may give rise to the formulation of more precise hypotheses on the underlying mechanisms.

In the mixed mating model of Illustration 2 (which assumes natural mating references), the mating preference pattern is characterized by identical heterotypic preferences that equal the proportion of cross-fertilization and are smaller than the homotypic preferences. All classical models of assortative mating, in which heterotypic mating is assumed to occur at random, share this kind of pattern (compare the example on the right in Figure 3). The effect of these preferences on genotypic frequencies of the offspring is observable for certain gene markers. ≫

Figure 3



Mating preferences of a type (A) for three types including its own; left: homotypic preference, random mating with type B, and negative preference for type C; right: preference pattern resulting from selfing and random cross-fertilization.

If a model is rejected, either the desired result is obtained or further analysis of potential causes for the rejection is required. In the latter situation one again has to start with the induction of these causes from the results of the preceding analysis (see Figure 1). Since causes are to be found within the model, its submodel structure has to be explicated in order to allow identification of the model features which can be held responsible for the rejection. The subsequent procedure is identical to the one depicted in Figure 1 and employs the methods and techniques of analysis presented above. Concerning the choice of the MI for an analysis of the hypothesized submodel, two basic opportunities exist. The more convenient opportunity arises from the possibility to deduce submodel characteristics which are – possibly after some reorganisation and transformation – observable by application of the original MI. This would allow utilization of the original observations for the analysis of the submodel. Otherwise, if it is not possible to deduce such submodel

characteristics, additional or new observations, possibly combined with newly developed MIs, are required for an analysis.

Acknowledgements – The author wishes to thank E. Gillet for many intense discussions of the presented concepts and M. Ziehe for commenting on earlier drafts of this paper.

References

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

Berg E, F Kuhlmann 1993. Systemanalyse und Simulation für Agrarwissenschaftler und Biologen. Verlag Eugen Ulmer, Stuttgart

Bishop YMM, SE Fienberg, PW Holland 1975. Discrete Multivariate Analysis. MIT Press, Cambridge etc.

Bossel H 1992. Real-structure process description as the basis of understanding ecosystems and their development. Ecological Modelling 63: 261-276

Gregorius H-R 1989. Characterization and Analysis of Mating Systems. Ekopan Verlag, Witzenhausen

Gregorius H-R 1996. Confidence regions for hypotheses on system characteristics. Göttingen Research Notes in Forest Genetics 21: 1-14

Gregorius H-R, M Ziehe, MD Ross 1987. Selection caused by self-fertilization. I. Four measures of self fertilization and their effects on fitness. Theor. Pop. Biol. 31: 91-115

Sokal RR, FJ Rohlf 1981. Biometry. 2nd edition. WH Freeman, New York

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

Wissel C 1992. Aims and limits of ecological modelling exemplified by island theory. Ecological Modelling 63: 1-12

Glossary

calibration of a model – adjustment of a model's free parameter values so as to minimize the discrepancy between the modeled and the actually observed system characteristics.

confidence region of model characteristics for an actual observation – the set of all target characteristics that are associated with observable characteristics (via free parameter values) which have discrepancies from the actual observation that are sufficiently small to prevent rejection of the model. Thus, confidence regions specify sets of target characteristics which the model could realize without being falsified by a given actual observation.

critical discrepancy – threshold discrepancy (δ) between the actually observed and the model characteristics, below which the actual observations are considered representative of the model characteristics and thus do not effect rejection of the model.

critical probability of a MI – the probability of obtaining an observation with discrepancy at least as large as the critical discrepancy.

critical region of a model – set of potential observations of the actual system that, if realized, would falsify a given hypothesis.

estimate of target characteristics – the target characteristics associated with those free parameter values of the model that result from calibration of the model on the basis

of indirect or incomplete actual observations in a qualifiedly non-rejectable model.

exact p-value – the probability of obtaining an observation from the model that is not more probable than the actual observation.

free parameter values – the values of model parameters which are not specified in a model and vary independently.

levels of qualification of a MI – the chosen critical discrepancy (δ) and level of significance (ε)

level of significance – level (ε) which must not be exceeded by the critical probability in order to qualify the pertaining MI.

method of investigation (MI) – includes identification, observation, sampling (and its modelling), and transformation of the system characteristics to be analyzed.

method of system analysis – consists of three components: a model system reflecting a hypothesis on the causation of real system characteristics, a MI that allows appropriate observation of the system characteristics to be studied, and a method of measuring discrepancy between the actual observations (on the system characteristics) and the modeled observations (representing the hypothesis).

prediction of modeled observation – the modeled observation that results from calibration of the model.

qualified MI – a MI for which the critical probability does not exceed a given level of significance for the model characteristics to be studied.

qualified method of system analysis – a method of system analysis that guarantees that observations exist that could falsify the associated hypothesis.

qualifiedly non-rejectable model – for a given MI, a model for which free parameter values exist, such that the MI is qualified for these parameters and the associated model characteristics show discrepancy from the actual observations that lies below the critical discrepancy.

representative observations – see "critical discrepancy".

significance probability of an actual observation – the probability of obtaining an observation from the MI that shows discrepancy from the model characteristics at least as large as that of the actual observation.

target characteristics of a model – system characteristics of interest which may or may not be observable. If they are not observable, they must be functions of observable model characteristics or of the free parameter values on which the observable characteristics depend; such target characteristics are termed *indirectly observable* in contrast with directly observable characteristics.