Science without (Parametric) Models: The Case of Bootstrap Resampling

Jan Sprenger[†]

Abstract

Scientific and statistical inferences build heavily on explicit, parametric models, and often with good reasons. However, the limited scope of parametric models and the increasing complexity of the studied systems in modern science raise the risk of model misspecification. Therefore, I examine alternative, data-based inference techniques, such as bootstrap resampling. I argue that their neglect in the philosophical literature is unjustified: they suit some contexts of inquiry much better and use a more direct approach to scientific inference. Moreover, they make more parsimonious assumptions and often replace theoretical understanding and knowledge about mechanisms by careful experimental design. Thus, it is worthwhile to study in detail how nonparametric models serve as inferential engines in science.

Keywords: models, data, inductive inference, nonparametric statistics, bootstrap resampling

1 Probabilistic Modeling

Modeling plays a key role in empirical science, especially when overarching theories cannot be applied. Many efforts in science focus on constructing, comparing and revising models of physical entities, phenomena and processes. Bohr's model of the atom, Volterra's model of predator-prey populations and the random walk model for the motion of molecules in a fluid are among the most popular ones. Models enable us to recognize fundamental relations between physical quantities, to understand the effects of causal interventions and to generalize observed effects to more complex and realistic cases. Often, their construction is triggered by concrete puzzles: For instance, Volterra (1926) developed his mathematical model of predator-prey population dynamics in response to the surprising shortage of adriatic fish after World War I. Volterra's model started from abstract considerations, but its predictions were found to be in stunning agreement with reality (see Weisberg's (2007) case study for more details). The way the Volterra model has been developed, refined and transferred to other scientific inquiries exemplifies a general strategy: to set up mathematically tractable models which capture fundamental mechanisms of the underlying system, and to gradually amend and refine them in order to account for the complexity of large-scale systems in the real world. In other words, models allow us to discover characteristic regularities (e.g. cycles in the population dynamics) as well as to explain concrete phenomena, such as "why does a disruption in fishing activity increase the predator/prey ratio?".

Hence, it is not surprising that philosophers of science have been spending a lot of paper on the various features of model-building. In particular, they studied

[†]Contact information: Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

the function of models as explanatory engines and the differences to straightforward descriptions and empirical generalizations. Here, it has been pointed out that modelers make *indirect* inferences about the target system: they study a (mathematical) model and hope that the results, when transferred to the target system, remain approximately valid (cf. Weisberg 2007, 2009). Moreover, constructing definite models *presupposes* knowledge about mechanisms and causal interactions within a system, but on the other hand, the technique of indirect inference also *improves* our structural understanding, and leads to more reliable predictions (cf. Godfrey-Smith 2006). Volterra's predator-prey model exemplifies both of these features, as shown above.

This mechanistic ideal of modeling ceases to apply whenever data obey apparently random patterns or when observations are disturbed by noise. In such cases we replace deterministic relationships by *probabilistic models* that are tailor-made to reasoning under uncertainty. The remainder of my paper focuses on probabilistic models that have invaded almost all natural and social sciences, but I see no obstacles to generalize my conclusions to deterministic models.

Definition 1 *A* parametric statistical model[1] *is an ordered triple* $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *where* $(\mathcal{X}, \mathcal{A})$ *is a measurable space (usually called the* sample space*) and* $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ *is a family of probability measures on* $(\mathcal{X}, \mathcal{A})$.[2]

In this definition, the sample space $\mathcal{X}$ corresponds to the set of possible observations whereas the σ-field $\mathcal{A}$ has only technical meaning, defining the set of 'measurable' subsets of the sample space. Crucially, $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ gives a family of probability distributions on $(\mathcal{X}, \mathcal{A})$, which describe one and the same *sampling mechanism* (i.e. the mechanism that generates the data). The parameter ϑ that calibrates the sampling mechanism is, however, unknown. By restricting the set of sampling distributions to a parametric family, the uncertainty about the *structure* of the underlying sampling mechanism is removed. Instead, all uncertainty attaches now to the value of the unknown parameter ϑ which we can try to infer from the observations we make: "this restriction and [...] parametrization should aid one in understanding and efficiently estimating the [true] distribution."[3]

But how do those mathematical constructions that we call statistical models connect to real systems? To what extent do they serve inferential tasks and indirect inference? Well, in the same way that deterministic models do. Morgan and Morrison state that formalizing a sampling process by means of a statistical model

> "[...] provides a model for a certain type of situations thought to exist in the real world and for which statisticians have well worked-out theories."[4]

In other words, statistical models are the crucial link between stochastic theory and the real world. One of the best-known illustration for Morgan and Morrison's claim is the *coin flip model*. A coin is either fair or biased towards one of

[1] Often, the terms 'parametric (statistical) model' and 'statistical model' are used interchangeably.

[2] Cf. Cox 2006.

[3] Spirtes, Glymour and Scheines 1993, 4.

[4] Morgan and Morrison 1999, 33.

the two sides, and we represent the probability that it comes up 'heads' by a parameter ϑ. When the coin is tossed repeatedly (a sequential *Bernoulli trial*), this is supposed to tell us something about the model parameter ϑ. And if we represent the number of heads by, let's say, the Binomial distribution, this codes our causal knowledge that there are just two possible outcomes in each trial, that the trials are independent from each other, etc. As we will soon see, extensions of that very simple model can represent a wide number of complex processes in science.

A extension of the coin flip model is the *random walk* – a discrete stochastic process that describes 'random' wandering on a rectangular grid. In genetics, those random walks are used to describe the variation of allele frequencies on a gene in a given population, e.g. for simulating genetic drift. But actually, random walks also serve as simplified models of the *Brownian motion* in hydrodynamics that describes a molecule's motion in a fluid. The Brownian motion is one of the most important diffusion processes in physics and has numerous (mathematically) beautiful attributes: It is a martingale with quadratic variation, has Normally distributed increments, the Markov property, etc. The abundance of mathematical tools and analytical results that can be used in working with the Brownian motion underlines the power of this model and explains why it is often applied outside hydrodynamics, too, e.g. for modeling financial markets.

Now, it is crucial to note that the scale limit of the random walk (if steps are made infinitely small) is just the Brownian motion. So parametric models do not only facilitate our analysis of real systems: there are also beautiful connections between different parametric models (e.g. extended coin flip models and Brownian motions) that enhance our understanding of physical phenomena and helps us to see how they are related. So it does not come as a surprise that statistics textbooks in empirical sciences as well as philosophers of science stress "the explicit need of a [parametric] model in analyzing the significance of empirical data"[5].

In particular, sophisticated testing procedures such as the F- and t-test have been designed for specific parametric models (here: the Normal distribution) and are widely used for testing scientifically relevant hypotheses. Intuitive and conceptually sound measures of evidence, such as the likelihood ratio and Bayes factors, are motivated by parametric assumptions (cf. Hacking 1965, Royall 1997). Hence, parametric models play a distinguished role in inductive inference.

In spite of all these virtues, I believe that the significance and indispensability of explicit parametric models in science has been overestimated. True, parametric modeling facilitates structural understanding as well as closed form computations and quantifying statistical evidence, but I argue that in scientific practice, valid conclusions can often be attained by means of *nonparametric, data-based inference*. Those methods do not require understanding of some fundamental mechanisms in the system and are therefore immune to pitfalls of parametric modeling. *Bootstrap resampling* gives a convincing case study. Finally, I discuss to what extent nonparametric techniques can serve as inferential engines, contrast them to traditional, parametric approaches in statistical modeling and argue that the former deserve more attention in the philosophical debate.

[5]Suppes [1962] 1969, 33. Cf. Mayo 1996 and Cox 2006.

Survival times	1	2	3	4	5	6	7	8	9	Median
Treatment group	94	197	16	141	38	99	23	—	—	94
Control group	46	30	52	146	40	10	104	27	52	51

Table 1: Survival times (in days) in the treatment and the control group after the test surgery.

2 Problems of Parametric Modeling

There is a central problem for a parametric modeler: to specify the right family of models for the studied system. A parametric inference that is based on an inadequate family of distributions will easily go astray. This is the problem of *model misspecification*. To specify the right model, Ian Hacking (1965) gives three guidelines: *analogy* to other, relevantly similar questions of inquiry, *scientific theory*, i.e. the implications of our physical, biological, etc. background knowledge – in particular knowledge about causal mechanisms – and finally, *simplicity*: the mathematical analysis must be feasible.[6] When model specification fails, the entire inference is usually worthless, and this is the reason why so much literature has addressed model misspecification, even within the philosophical community. For instance, Mayo and Spanos (2004) extensively discuss techniques for detecting misspecification.[7]

In general, correct model specification requires a lot of insights into the target system. The more complex the processes we deal with and the scarcer our theoretical understanding, the less we can be certain to have chosen the right model. When we analyze complex systems, model specification is often not sufficiently guided by theoretical understanding. Causal relations between model variables may be unclear, the entire system may be too complex to model, no mathematically tractable distribution fits the specific values which the observations take, etc. Time series in econometrics and meteorology provide salient examples. We have to account for the uncertainty about the nature of the true distribution, and we cannot expect – as parametric models often do – nature to behave according to our wishes for mathematical convenience and structural simplicity. An example illustrates the point.

Example 1 *(Efron and Tibshirani 1993): A group of seven mice is assigned medical treatment after a test surgery. We would like to study whether this treatment is able to prolong the survival time of the mice, compared to mice which are operated without being assigned the treatment. To this end, we set up a control group of nine mice. The incoming data are displayed in table 1.*

Although the example is very simple, it is not clear how model specification could proceed. Certainly, simplicity might speak for choosing a Normal distribution, but how do we defend that claim? The asymmetry of the data in the control group speaks against the assumption of Normality. Note further that the data points are noted as integer values which speaks for a discrete distribution, instead of a continuous one like the Normal distribution. Hacking's other guidelines also fail: If the medical treatment is a novel one, and we choose a

[6] Cf. Hacking 1965, 83-85.

[7] Cf. Burnham and Anderson 1998 for a practitioner's perspectives.

parametric model by analogy to an old drug, we implicitly impose constraints on our interpretation of the observations, instead of taking an unbiased perspective. Finally, in such a complex process as the effect of chemical drugs on biological organisms, there is no overarching theory which directly links chemical properties of the drug to the survival time of the mice. Thus, model specification is quite difficult and risky. Furthermore, having a specific model of drug efficacy is arguably less important than knowing that the drug is effective at all and that we should administer it in future cases. (Especially if not mice, but humans are assigned medical treatment after a serious surgery.)

The latter goal – predicting future performance – has become especially important in modern science. Geophysical and meteorological models provide paradigmatic examples. The availability of loads of data on the actual weather together with our geophysical theory gives us a sensible idea of the local weather in the next 24 hours. But it would be presumptuous to capture the essential structure of complex systems, such as the Earth's climate, through explicit, parametric models, even if the underlying physics are roughly understood. First, the *scale* of the model is simply too large to warrant that a model parameter can still be meaningfully mapped to a real physical quantity (such as moist convection or surface pressure at particular spot). Second, the number of physical interactions in a system that is as *complex* as Earth's climate are so numerous and messy that

> "we know *a priori* that there is no combination of parametrizations, parameter values and initial conditions which would accurately mimic all relevant aspects of the climate system."[8]

In light of these limitations of parametric statistical modeling, modern science has to make recourse to more parsimonious assumptions. In particular, forecasting techniques that are guided by data (such as mathematical extrapolation techniques) often replace predictions that have been gained by a top-down approach and stipulating an explicit model.

Of course, the statistical techniques for detecting model misspecification become more and more refined. But their power does not keep up with the increasing model complexity in modern science. Thus, shouldn't we better seek for alternative inference techniques? In scientific practice, that conclusion is often drawn and exemplifies a trend away from explicit models and closed form solutions (cf. Humphreys 2004). This trend has even reached statistical physics, one of the most theoretical branches of empirical science. I mentioned the great number of analytic results proven for the Brownian motion. Still, simulation-based methods such as Monte Carlo methods are nowadays omnipresent in studying Brownian motion and related stochastic processes (cf. Sharma and Patankar 2004). While the pioneer work in simulating hydrodynamic processes goes back to the 1950s (e.g. the Metropolis algorithm), it was the advent of fast and efficient computing resources that made simulation-based analyses widely available and practicable. So even branches of physics where parametric modeling achieved its greatest unifying successes have been infiltrated by numerical, simulation-based methods.

Moreover, statistics that are particularly easy to analyze in parametric models, such as the population mean, are notoriously vulnerable to measurement

[8] Stainforth et al. 2007, 2148. Italics in the original. Cf. Sprenger 2009 for a philosophically minded discussion of statistical inference in the face of model uncertainty.

errors, biases and outliers in the data. On the other hand, more robust statistics of interest as the median are hard to analyze in a parametric framework. All these concerns show that working with parametric models does not only have benefits, but also severe drawbacks and that we need alternative techniques for addressing classical questions of statistical inference, such as causal inference and estimating standard errors. In the next section, I illustrate how modeling assumptions can be kept to a bare minimum by means of computer-intensive *resampling methods*. There, the actual sample is taken as a nonparametric model of the population. The *bootstrap strategy* provides a particularly nice illustration of the resampling principle and a template for investigating the inferential virtues of non-parametric models.

3 Resampling Methods: A Bootstrap Case Study

One of the most common statistical activities consists in comparing two samples of different populations with respect to a specific characteristic. This is called the *two-sample problem* and it is exemplified in example 1: we would like to test the hypothesis that the medical treatment is not effective at all, i.e. that the two samples (the treatment and the control group) are actually drawn from the same distribution. As argued in the previous section, we have to test that hypothesis in the face of strong model uncertainty and little structural understanding.

A parametric statistical approach would assign a specific family of distributions to the treatment and the control group, for instance a Normal distribution with means μ_1, μ_2 and variances σ_1^2, σ_2^2. Then, we could apply the t-test for testing equality of the means (of Normally distributed populations) and the F-test for testing equality of the variances. Since the t- and F-distributions are well studied, such a procedure would be easy to handle. But we have already argued that the assumption of Normality would be highly contentious in the mice example – neither underlying scientific theory nor simplicity nor analogy recommend the choice of a specific model. I show how *bootstrap resampling* transforms the sample into a nonparametric model of the population. Hence, it can be used to make inferences about the underlying population in the absence of explicit model assumptions, so the modeler 'pulls herself up by her own bootstraps'.[9]

Let $(x_1, \ldots, x_m)$ denote the survival times in the treatment group and let $(y_1, \ldots, y_n)$ denote the survival times in the control group. Let us pool all those data into a single sample $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ and let $\hat{F}$ denote the empirical distribution function (EDF) of the pooled sample. The EDF gives equal probability weight $1/(m+n)$ to any element of the sample and zero to all other points. Under the null hypothesis H_0 that the treatment has no effect, all the x_i and y_j are drawn from the same population. Given H_0, the EDF becomes a non-parametric estimate of the joint distribution of the x_i and y_j.[10] [11] Now, the resampling mechanism evaluates the actual observations under the

[9] In spite of the terminology, there is no analogy to Clark Glymour's (1980) theory of "bootstrap confirmation".

[10] The EDF assigns probability zero to all points which are not in the actual sample. Especially in quite small samples, this assumption is often ruled out by our background knowledge. In such cases, the EDF can be smoothed using adequate techniques, e.g. kernel-dressing.

[11] By estimating the unknown population distribution with the EDF of the sample, the bootstrap generalizes the principle of *maximum likelihood estimation* to the nonparametric

assumption that H_0 is true:[12]

1. Let $b = 1$.

2. Draw with replacement $m + n$ '*bootstrap resamples*' from the distribution $\hat{F}$. Randomly assign them to an ordered $m+n$-tuple $(x_1^b, \ldots, x_m^b, y_1^b, \ldots, y_n^b)$. (So an x_i^b can also be assigned the value of a y_j in the original sample, and vice versa.)

3. Calculate the group means $\bar{x^b}$ and $\bar{y^b}$ for the bootstrap resamples. Then, calculate the value of the discrepancy-measuring statistic

$$t(\bar{x^b}, \bar{y^b}) \quad := \quad \frac{\bar{x^b} - \bar{y^b}}{\bar{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{1}$$

   where

$$\bar{\sigma} \quad := \quad \sqrt{\frac{\sum_{i=1}^{m}(x_i^b - \bar{x^b})^2 + \sum_{j=1}^{n}(y_j^b - \bar{y^b})^2}{(n-1)+(m-1)}}. \tag{2}$$

   The denominator in (1) and the complicated expressions in (2) may trouble the reader, but they have merely technical significance: The distance statistic $\bar{x^b} - \bar{y^b}$ is adjusted by dividing it through an estimate of its standard deviation ('studentization').

4. Let $b := b + 1$ and go back to step 2 until $b = B$, the number of bootstrap resamples, is attained.

5. Calculate the fraction of times where the actually observed discrepancy exceeds the discrepancy in the bootstrap replications:

$$p_{\text{obs}} := \frac{1}{B}\,\#\{t(\bar{x^b}, \bar{y^b}) \leq t(\bar{x}, \bar{y}),\, b \leq B\} \tag{3}$$

The rationale of bootstrapping is quickly explained. We would like to test (and possibly to reject) the null hypothesis that the medical treatment does not have any effect, i.e. that treatment and control sample are drawn from the same population. To this end, we pool both samples into a single sample and draw simulated resamples out of this pooled sample (step 2). For each of these resamples, we check whether the discrepancy between the two resampled groups exceeds the discrepancy in the original data (step 3). Under quite mild conditions, the bootstrap is asymptotically consistent (see Efron 1979, Bickel and Freedman 1981), i.e. for increasing sample size ($m, n \to \infty$) and an increasing number of resamples ($B \to \infty$), the bootstrapped distribution of the distance statistic t will mimic the real distribution of t. Thus, we repeat the process a large number of times in order to get a reasonably high number of resamples (step 4). At the end, we count the fraction of times where the actual discrepancy between the group means exceeds the discrepancy in the resamples (step 5). If that happens very often, it is unlikely to be a result of pure chance, and the

case. The Glivenko-Cantelli theorem guarantees that in the limit, the EDF of the sample converges uniformly against the population distribution.

[12] Cf. Efron and Tibshirani 1993.

result (i.e. a large value of p_{obs}) will *significantly* speak against the null hypothesis that the two populations are equally distributed.[13] In the actual example of table 1, we obtain a p-value (or actual significance level) of $p_{\text{obs}} = .866$ for a value of $B = 1000$. This is clearly not enough to reject the hypothesis that the medical treatment is just a placebo since in 13% of all cases, such a high result would have been obtained by chance.

Note that the consistency results for the bootstrap crucially turn on the assumption that the single data points are independent and identically distributed. Thus, the bootstrap is not free of modeling assumptions (cf. Rubin 1981). But the assumptions are of a quite different type – they are *qualitative* and can be warranted with the help of a careful experimental setup, controlling that the trials were really screened off from each other, etc.. Thus, such assumptions are much easier to defend than specific parametric assumptions, and they replace theory- and parameter-based inference by *design-based inference*. In other words, the responsibility for model adequacy lies with the experimenter's practical skills rather than with his theoretical understanding. This allows a more direct approach to testing scientifically relevant claims, without setting up a refined (and possibly misspecified) parametric model.

Actually, the two-sample problem is characteristic of any situation where two samples are compared with respect to some characteristic, as the mean, the variance, etc. For instance, we could ask whether the average height of 10-year-old boys equals the average height of 10-year-old girls. Or we could ask whether a simulation-based model of a physical process is indeed a faithful model of the target process and compare the two data sets to this end. It is a distinctive feature of the bootstrap that it does not only apply to the two-sample problem discussed above (equality of two distributions) – it can be applied to almost all statistical inference problems.

For instance, in the above example, we have tested a causal hypothesis (does the treatment have effect?), but Demiralp, Hoover and Perez (2009) use the bootstrap as well for assessing the confidence in the result of a search for causal dependencies. Equally, the bootstrap provides a nice means of quantifying the confidence that we put into a statistical estimate, as measured by the *standard error* of an estimate. Analytic formulas for estimating standard errors are in general only available for specific statistics, such as the sample mean. In the example of table 1, we might be more interested in the median ('the treatment effect for the average mouse') than in the mean since outliers in the data easily bias the sample mean. Thus, we estimate the population medians for the treatment and the control group by the respective sample medians, leading to estimates of $\tilde{x} = 94$ and $\tilde{y} = 51$. This is apparently a large effect. However, we should accompany that estimate by an estimation of the standard error in order to quantify how much of that difference may be due to random sampling. To this end, we draw bootstrap resamples $(x_1^b, \ldots, x_m^b)$ from the treatment group and estimate the standard error of the sample median by the standard deviation

[13] The use of p-values in testing point null hypotheses has been subject to severe criticism (Berger and Sellke 1987), but it is not necessary to rehearse the Bayesians vs. frequentists debate since the bootstrap can be equally applied in a Bayesian framework (Rubin 1981).

B	50	100	250	500	1000	∞
Median	32.21	36.35	34.46	36.72	36.48	37.83

Table 2: Bootstrap estimates of the standard error of the sample median for the treatment group in table 1, as a function of the number of replications B.

of the median in the bootstrap replications:[14]

$$\hat{\text{se}} := \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{x^b} - \frac{1}{B}\sum_{j=1}^{B}\tilde{x^b}\right)^2}. \tag{4}$$

Here, $\tilde{x^b}$ denotes the median in the b-th bootstrap resample. In other words, we draw a large number of resamples from the original sample and look to what extent the replicated medians diverge from each other under the assumption of independent sampling. Then, $\hat{\text{se}}$ is supposed to give a reasonable approximation of the standard error of the sample median. Note that bootstrap resampling squeezes out all available information from the sample (by stipulating the sample as a nonparametric model) whereas parametric estimation focuses on selected aspects of the data.

In the mice example, we obtain the numbers shown in table 2. For $B \geq 500$, the asymptotics work fine, and in terms of computation time, the effort for the resampling analysis is negligible. The observed difference between the sample medians is greater than the estimated standard error (to be precise, 1.14 estimated standard errors), but again, a difference of that magnitude may still be due to chance alone.

Again, we see the simplicity and efficiency of the bootstrap at work. And even on a theoretical level, the bootstrap may fare better than a classical, parametric approach. Under a large set of conditions, the bootstrap approximation of the standardized sample mean outperforms an asymptotic analysis based on the central limit theorem.[15] So the bootstrap does not only replace parametric approaches whenever their application would be problematic or too cumbersome, as in the case of median estimation – it actually has theoretical virtues on its own. Given all these successes, it is now time that the philosophy of statistical inference acknowledges those developments and integrates resampling methods into a unified scheme of data analysis and inductive inference.

4 Summary and Discussion

For a long time, parametric modeling has been the unchallenged paradigm for inductive inference in the sciences. As explained in section 1, parametric modeling requires some theoretical understanding or knowledge about causal mechanism, but it often yields high explanatory power and mathematical convenience. The coin flip model and one of its extensions, the random walk, provide a salient example. Hence, a parametric framework is also the natural context for debating principal issues in statistical methodology (cf. Mayo 1996, Royall 1997).

[14] See Chapter 2 in Efron and Tibshirani 1993.
[15] Cf. Singh 1981.

However, the increasing complexity of statistical analysis requires us to focus on nonparametric techniques: The less we know about a target system, the greater the scale of the model, or the more opaque the interactions between the modeled quantities, the less can parametric assumptions be justified and the more likely is our inference to be led astray. Under such circumstances, guidelines for correct model specification, such as background theory and analogical reasoning, cease to apply and the goal of adequate modeling may be hard or impossible to achieve.

These criticisms, made explicit in section 2, triggered the question how classical statistical inquiries may be addressed without contentious modeling assumptions. Section 3 drew attention to a particular non-parametric technique: bootstrap resampling. Due to its parsimonious presuppositions and its versatile applicability, it deserves special attention. Bootstrap methods draw simulated resamples from the actual data and work under much milder conditions, replacing the choice of a particular family of distributions by the assumption that the observed random variables are independent and identically distributed. This constraint can be satisfied by means of qualitative understanding or careful experimentation. In other words, the bootstrap uses the sample as a model of the population and exemplifies *design-based data analysis*, instead of *theory- or mechanism-based data analysis* that is typical of explicit parametric modeling. This has general implications for theory testing in science: If a model is rejected in a parametric hypothesis test, does this negative result transfer from the *statistical* model to the *scientific* thesis which we wanted to test? Actually, scientists often avoid that conclusion (cf. Keuzenkamp and Magnus 1995). One reason for this reluctancy is certainly model uncertainty. This worry might, however, be addressed by resampling techniques which offer a more direct way to address scientifically relevant questions of inquiry.

Traditional parametric modeling starts with an easily understandable model, such as the coin flip model. The properties of such a model are studied and we hope that some of the insights we gain transfer to the target system. This is an *indirect top-down approach* – we study a stipulated model before we make inferences about the real system. Nonparametric models, however, work *bottom-up* and combine strategies of direct and indirect inference: On the one hand, the model is directly constructed from the visible elements of the target population, namely the actual sample. No mediation via a toy model or an imagined system is required. On the other hand, we can simulate further experiments within our sparse data model, and by drawing on the results of those simulations, we can make reliable scientific inferences. In other words, we derive our inferences about real-world phenomena from studying simulated resamples that have been generated by a mathematical model. Thus, resampling inferences are neither straightforward descriptions nor mere generalizations of observed data – they combine direct and indirect inference techniques (cf. Weisberg 2007).

To some extent, the dichotomy between design-based, bottom-up resampling methods and theory-based, top-down parametric methods is blurred in practice. For the debate about models in science, it seems to be a fruitful project to explore if the two approaches can complement each other. In particular, I would like to investigate how the virtues of both approaches – structural understanding in one case; parsimonious, design-based inference in the other case – can be combined without being exposed to the drawbacks of either strategy. This is, however, a project for future research.

Acknowledgements

I would like to thank the audience at the *Models and Simulations 2* conference in Tilburg and the London-Paris-Tilburg Workshop in Philosophy of Science for their helpful criticisms and suggestions. In particular, I would like to thank Roman Frigg, Stephan Hartmann, Kevin Hoover, Jan-Willem Romeijn, Jonah Schupbach, Leonard Smith and Michael Weisberg as well as two anonymous referees of this journal for their detailed and stimulating feedback.

References

Berger, James O. and Thomas Sellke (1987): "Testing a Point Null Hypothesis: the Irreconcilability of P Values and Evidence", *Journal of the American Statistical Association* **82**, 112-122.

Bickel, Peter J. and David A. Freedman (1981): "Some Asymptotic Theory for the Bootstrap", *Annals of Statistics* **9**, 1196-1217.

Kenneth P. Burnham, David R. Anderson (1998): *Model Selection and Inference: a Practical Information-Theoretic Approach*. Springer, New York.

Cox, David R. (2006): *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Demiralp, Selva, Kevin D. Hoover and Stephen J. Perez (2008): "A Bootstrap Method for Identifying and Evaluating a Structural Vector Autoregression", *Oxford Bulletin of Economics and Statistics* **70**, 509-533.

Efron, Bradley (1979): "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics* **7**, 1-26.

Efron, Bradley and Robert Tibshirani (1993): *An Introduction to the Bootstrap*. Chapman & Hall, London.

Glymour, Clark (1980): *Theory and Evidence*. Princeton University Press, Princeton.

Godfrey-Smith, Peter (2006): "The strategy of model-based science", *Biology and Philosophy* **21**, 725-740.

Hacking, Ian (1965): *Logic of Statistical Inference*. Cambridge University Press, Cambridge.

Humphreys, Paul (2004): *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press, Oxford.

Keuzenkamp, Hugo A. and Jan R. Magnus (1995): "On tests and significance in econometrics", *Journal of Econometrics* **67**, 5-24.

Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago & London.

Mayo, Deborah G. and Aris Spanos (2004): "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science* **71**, 1007-1025.

Morrison, Margaret and Mary Morgan (1999): "Models as Mediating Instruments", in: Mary Morgan and Margaret Morrison (ed.), *Models as Mediators. Perspectives on Natural and Social Science*, 10-37. Cambridge University Press, Cambridge.

Royall, Richard (1997): *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Rubin, D.B. (1981): "The Bayesian Bootstrap", *Annals of Statistics* **9**, 130-134.

Sharma, Nitin and Neelesh A. Patankar (2004): "Direct numerical simulation of the Brownian motion of particles by using fluctuating hydrodynamic equations", *Journal of Computational Physics* **201**, 466-486.

Singh, Kesar (1981): "On the Asymptotic Accuracy of Efron's Bootstrap", *Annals of Statistics* **9**, 1187-1195.

Spirtes, Peter, Clark Glymour and Richard Scheines (1993): *Causation, Prediction, and Search*. Springer, New York.

Sprenger, Jan (2009): "Statistics between Inductive Logic and Empirical Science", forthcoming in *Journal of Applied Logic*.

Suppes, Patrick (1969): "Models of Data", in: P. Suppes (ed.), *Studies in the Methodology and Foundations of Science. Selected Papers from 1951 to 1969*, 24-35. Reidel, Dordrecht. Orginally published in Ernest Nagel, Patrick Suppes and Alfred Tarski (eds.): "Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress". Stanford: Stanford University Press, 252-261, 1962.

Stainforth, D. A., M. R. Allen, E. R. Tredger and L. A. Smith (2007): "Confidence, uncertainty and decision-support relevance in climate predictions", *Philosophical Transactions of the Royal Society A* **365**, 2145-2161.

Volterra, V. (1926): "Fluctuations in the abundance of a species considered mathematically", *Nature* **118**, 558-560.

Weisberg, Michael (2007): "Who is a Modeler?", *British Journal for the Philosophy of Science* **58**, 207-233.

Weisberg, Michael (2009): "Models of Modeling", unpublished manuscript, source: *http://www.phil.upenn.edu/~weisberg/Homepage/Papers.html*.