QUARTERLY OF APPLIED MATHEMATICS
VOLUME LVI, NUMBER 4
DECEMBER 1998, PAGES 797–805

A PLACE FOR PHILOSOPHY? THE RISE OF MODELING IN STATISTICAL SCIENCE

BY

PERSI DIACONIS

Department of Statistics, Stanford University, Stanford, CA

Abstract. Large statistical models seem to have reached epidemic proportions. I will document the harm they are currently causing and contrast it with some success stories of smaller modeling efforts. A case study involving a model for flipping coins shows how "adding a few bells and whistles" can lead to nonsense. Finally, an effort to list the "real uses" of models is offered as a start to making sense of current modeling efforts.

1. Two ideal examples.

EXAMPLE 1. *The mathematical formulation of statistical problems.* A clear description of mathematical statistics can be simply put: One is given a family of probability measures $\{P_\theta(dx)\}_{\theta\in\Theta}$. One observes $x \in X$ drawn from $P_\theta(dx)$ and is required to make a guess $\hat{\theta}(x) \in \Theta$ for the true θ. Usually, the quality of a guess $\hat{\theta}$ is quantified by a loss function $L(\hat{\theta}, \theta)$. Then, one tries to choose $\hat{\theta}$ so that $\int L(\hat{\theta}(x), \theta) P_\theta(dx)$ is small. In a Bayesian version of the set up, an apriori distribution $\pi(d\theta)$ is specified and one chooses $\hat{\theta}$ so that $\int L(\hat{\theta}(x), \theta) P_\theta(dx)\pi(d\theta)$ is small. With $\{P_\theta\}$, L, and π specified, this is usually straightforward.

Of course, in any real problem, $P_\theta(dx)$ is only an approximation and there may be problems in specifying L and π. This is further grist for the mathematical mill: How does $\hat{\theta}$ vary with changes in $\{P_\theta\}_{\theta\in\Theta}$, L, and π? The above formulation has had great success at codifying statistical practice and in suggesting new estimates (and new math problems). The clearest treatment of modern mathematical statistics is in Lehmann (1983, 1986).

EXAMPLE 2. *Gauss and the linear model.* Here is a more applied example of mathematical statistics: Gauss' discovery of the asteroid Ceres. The year is 1801. The 21-year-old Gauss has just finished his monumental Disquisitiones Mathematica; he notices he does not have a job. At this time people believed there were only five planets; then, the astronomer Patuzzi spotted the asteroid Ceres and took readings on it for 60 days. Finally, he lost Ceres in the sun. His observations caused great excitement in

Received March 2, 1998.
1991 *Mathematics Subject Classification.* Primary 62A99; Secondary 62J99.

scientific circles and far beyond. The community searched diligently for Ceres with no luck. Finally, the observations were published.

The position of a body orbiting the earth is determined by six parameters $\theta = (\theta^1\theta^2\cdots\theta^6)$, the elements of the orbit. These are basically its position and velocity at some fixed time. Thus the observations were $f_i(\theta)$; here f_i is a known, nonlinear function depending on i through time of day and type of observation. Of course one observes $Y_i = f_i(\theta) + \varepsilon_i$ with ε_i an observational error. If the errors are assumed to be independent disturbance terms from the bell-shaped curve (or Gaussian distribution) one is led to estimate θ by minimizing the sum of squares

$$\sum_i (Y_i - f_i(\theta))^2.$$

Gauss linearized this problem by writing

$$f_i(\theta) = f_i(\theta_0) - (\theta - \theta_0)D_i f(\theta_0) + \cdots$$

with θ_0 a preliminary guess at θ. This approximation is now in the familiar form $Y = X\beta + \varepsilon$, with $\beta = (\theta - \theta_0)$ a vector of length 6 and X the 60 by 6 matrix of partial derivatives evaluated at θ_0. Gauss introduced Gaussian elimination to solve the least squares problem. He iterated, using the solution $\theta_1 = \theta_0 + \widehat{(\theta - \theta_1)}$, and so on.

The final piece of the story is that Gauss announced his results and on the first clear night the astronomers looked where he said, and there was Ceres.

The above is only slightly romanticized. It is a wonderful success story for modeling. At this time, the exact form of the f_i was not perfectly understood. This was one of his contributions. Of course, the model with Gaussian errors is only an approximation as well; here, linearity is approximately correct because of a Taylor series approximation.

Along the way, Gauss proved the Gauss-Markov Theorem showing that the least squares estimate is best (minimum variance) along all linear unbiased estimators. This is an early example of statistical theory. Gauss also justified the bell-shaped curve by showing that it is the unique error distribution having the average as the maximum aposteriori distribution. The really wonderful part of the story is that Gauss found Ceres where others had failed. For more details, Gauss (1963, 1995) himself is highly readable. The second volume, a recent translation by Stewart, contains pointers to the standard literature.

The two examples of this section, the elegant clarity of the decision-theoretic formulation of statistics and the practical and theoretical successes of young Gauss are things we dream about. Unfortunately, much of the current work I see is neither interesting mathematics nor useful in practice. Worse, in its complexity, modern model building seems to drive away from the truth into a fantasy land beyond objective reality. These are strong words and need justification.

2. The bad and the ugly.

EXAMPLE 1. *An energy model.* Let me begin by describing a typical medium size model—RDFOR—used by the Department of Energy to forecast energy demand. For

ten regions, the model is

$$Y_{rt} = a_r + b_r p_{rt} + c_r z_{rt} + d_r h_{rt} + e_r c_{rt} + f_r Y_{rt-1} + c'_r z_{rt-1} + d'_r h_{rt-2} + e'_r c_{rt-1} + \varepsilon_{rt}.$$

Here $r = 1, 2, \ldots, 10$ indexes the regions, Y_{rt} is log fuel consumption in region r at time t, p_{rt} is log fuel price, h_{rt} is log "heating days", z_{rt} is log income, c_{rt} is log cooling degree days. The parameters to be estimated are a_r, b_r, c_r, d_r, e_r, f_r, c'_r, d'_r, e'_r. The "disturbance term" is specified as $\varepsilon_{rt} = \lambda_r \varepsilon_{rt-1} + \varepsilon'_{rt}$ with ε'_{rt} independent noise terms and λ_r yet another parameter. The parameters are constrained to be equal or constant over various regions in a rather complex pattern. For more detail, see Freedman et al (1983).

There were 16 years of data on the ten regions. This gives 160 data points. After adjusting for equalities, there were 57 parameters to be estimated. For use in forecasting, the exogenous variables p_{rt}, z_{rt}, etc. have to be conjured up. These come from other models. When they are not available, the relevant variables are dropped from the equations.

One can only marvel that someone was brave enough to write down such a model with its log terms, linearity, prespecified parameter equalities and other details. The story of how such a model comes to be used is complex. One of its protagonists said "the RDFOR is the descendant of a modeling effort that began with simple forecasting models containing little detail. The policy makers' demand for detail led to what was considered at the time to be an enormous effort to collect more data and faced with the imperfect data that result, nearly all the ad hoc adjustments or assumptions were imposed to avoid even greater problems in the models when estimates using a naive application of 'economics ±' theory."

The RDFOR model turned out to be worse than useless; giving the illusion of knowledge, soaking up a large amount of money for development and support, and in the end giving wrong forecasts.

It is to be emphasized that this is a typical medium-sized model developed over years of work to try to make useful forecasts; there are thousands of larger, woolier models in routine use today. These models are used by policy makers and affect all of our lives. It is difficult to find open discussion of these issues. Statisticians, like doctors, tend not to criticize each other lest someone should look at their applied work!

EXAMPLE 2. *From mouse to man.* Low-dose extrapolation from mouse to man is a very prevalent business. On its face this is unbelievably delicate; one takes specially bred laboratory mice (if you breathe on them, they fall over). These mice are subjected to extreme conditions, and then one attempts to extrapolate to man. This last involves fitting a variety of curves and extrapolating far into the tails. It is obvious that the shape of the family of curves used will make a huge difference. As above there are a variety of models used to adjust for various things and so a number of parameters need to be estimated.

The history and sociology of the development of these models would be fascinating. The problems of environmental pollution are clearly important and one can imagine a sequence of steps leading to the present state. A highly critical review of the present

state may be found in Freedman and Zeisel (1988). They argue, more carefully than I can here, that the present state is a morass, with made-up equations, unverifiable wild assumptions, and unjustified conclusions. Their critique is accompanied by a reply from the modelers and a rebuttal. Non-statistical readers may not know that the statistics literature is filled with "discussion" articles. The debate is often surprisingly honest.

EXAMPLE 3. *Census adjustment.* Every 10 years the U. S. government takes a census. In recent years, the Census Bureau has veered from attempting a complete enumeration to sampling and modeling. The current plan for the year 2000 involves an ambitious model to attempt to deal with the undercount (perhaps 5% of the population is "missing"). It seems to some of us that the modeling effort is too ambitious and may well do more harm than good. Law makers want estimates of population counts at the city block level, stratified by age, race, income, and other variables. This requires that the model be right in a fair amount of detail.

In the end, the model is a sophisticated version of the linear model that Gauss used "$Y = X\beta + \varepsilon$". Recall that Gauss justified linearity by Taylor series approximations. There is no such justification when dealing with human populations. The additional bells and whistles added to make the model roughly right have an out-of-this-world quality.

The debate over the use and misuse of models for census adjustment is well documented in the discussion of the critical reviews by Freedman and Wachter (1994). It is complex, politicized and important. My belief is that modeling is doing real harm to the census.

The three examples above are meant to show that large linear models are in widespread use for important problems. In some way, one has to be a professional to understand just how made up and wild these models are.

For me, the modeling efforts discussed in this section are elaborate fabrications that are wrong in fundamental ways:

- They do not work on fresh data.
- They do not predict new phenomena or the results of interventions.
- They give the illusion of knowledge.
- They block honest work.

Gauss' example shows that linear models can be useful. Gauss found Ceres; the biological modelers cannot extrapolate from mice to rats let alone man; for much further discussion, I recommend Freedman's "statistics and shoe leather" essay (1991).

3. Tossing a coin: A detailed example. It is easy to make fun of others, especially if they are not around to defend themselves. In this section I recount a development in my own work where "adding bells and whistles" led to nonsense.

As I tried to document in the section above, I have become extremely skeptical about large statistical models. With probability theory itself under attack from a "fuzzy, upper-lower crowd", I wanted some simple success stories. I decided to go back to the primitive images of probability in gambling. I here record some thoughts about the analysis of tossing a coin.

A. *A subjectivist tosses a coin.* Consider repeated tosses of a coin under stable conditions. The outcomes will be called $X_1, X_2, \ldots$ with $X_i = 1$ corresponding to heads and $X_i = 0$ corresponding to tails. Probability statements $P(\cdot)$ refer to *my* best guess. Under

stable conditions, my best guess is symmetric in time $P\{X_1 = 1, X_2 = 0\} = P\{X_1 = 0, X_2 = 1\}$. More generally,

$$P\{X_1 = e_1, X_2 = e_2, \ldots, X_n = e_n\} = P\{X_1 = e_{\pi(1)}, \ldots, X_n = e_{\pi(n)}\}.$$

Above, $e_1, \ldots, e_n$ is any sequence of zeros and ones and π is any permutation. This symmetry is called exchangeability. Under this condition, deFinetti proved his fundamental representation theorem.

THEOREM. Let $P(\cdot)$ be an exchangeable probability distribution on binary sequences. Let $S_n = X_1 + \cdots + X_n$. Then there is a unique probability μ on $[0, 1]$ such that

(1) $P\left\{\frac{S_n}{n} \in A\right\} \to \mu(A)$ for all Borel sets A as n tends to infinity,

(2) $P\{X_1 = e, \ldots, X_n = e_n\} = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\mu(d\theta)$ with $s_n = e_1 + \cdots + e_n$.

deFinetti's theorem shows that under exchangeability a subjectivist believes that long-term frequencies exist and that the probability of an observable sequence must be assigned according to the classical recipe of Bayes and Laplace: There is an a priori distribution μ on the parameter space $[0, 1]$ and a model $P_\theta(e_1, \ldots, e_n) = \theta^{s_n}(1-\theta)^{n-s_n}$. Note that the theorem builds the model from a symmetry assumption.

Further, Bayes theorem says that predictions about the future must be given by an expression like the right side of (2) with μ replaced by the posterior $\mu^{X_1,\ldots,X_n}$. Finally, under mild conditions, Laplace showed that the posterior gets highly peaked about the observed proportion of heads:

$$\mu^{X_1,\ldots,X_n} \Rightarrow \delta_{S_n/n}.$$

Thus a subjectivist will tend to predict that the future will be like the past.

The above sketch of Bayesian probability may be amplified by reading deFinetti (1937, 1972). For a discussion of deFinetti's theorem and its extensions beyond binary events, see Diaconis (1988) and Diaconis and Freedman (1984). For convergence of the posterior, see Diaconis and Freedman (1990) and the references therein.

We thus have a neat formulation of statistical problems associated to binary outcomes. All is reduced to specifying an a priori distribution on the parameter space $[0, 1]$. Further, with a moderate amount of data, the prior washes away. The next section shows how things are modified if one builds in some physics.

B. *A Newtonian coin toss.* Consider a real toss of a real coin. If we knew how fast the coin was traveling when it leaves the flipping hand and how many times per second the coin was revolving, Newton's $F = ma$ law tells us how the coin will land. Put this way, it is obvious that coin tossing is physics, not probability!

It is relatively easy to do the physics under some simplifying assumptions—neglect air resistance, suppose the coin spins around a line through its plane (instead of precessing), and suppose the coin lands in sand (or on your hand) without bouncing. Now, all is determined by the initial conditions, say (V, ω); here V is the upward velocity at the time of release (say in feet per second) and ω is the angular velocity (say in revolutions per second). A given flip now corresponds to a point in the velocity/spin plane. There is a region of initial conditions where the coin only turns over once. Consider the point $*$ shown in Fig. 1—this has high velocity but low spin, so the coin goes up like a pizza

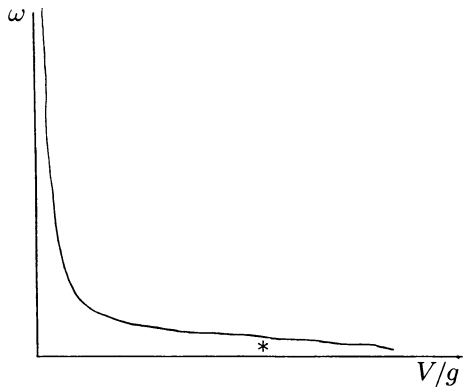

FIG. 1. Phase space for coin tossing

without turning over. The curve shows the bounding region for the coin landing without turning. Similarly, there are regions where the coin turns over once, twice, and so on. These regions are shown in Fig. 2.

Notice that the regions in Fig. 2 get close together as one moves away from $(0,0)$. Thus, small changes in initial conditions make for the difference between heads and tails.

To be more quantitative about this, one needs to know the popularity of different regions of phase space for typical coin tosses. For a typical one-foot toss, experiments show that coins go up at about 5 m.p.h. and turn over 35–40 revolutions per second. In the units of the picture the velocity is concentrated at about 0.2 on the velocity axis. This is close to zero in the picture. Fortunately, the spin is concentrated at about 40 units up on the ω axis. We thus see that the wonderful picture says nothing about actual flips of a coin. However, the math behind the picture shows that the partition of phase space in a neighborhood of $(0.2, 40)$ is very fine; building in what we know about variability about $(0.2, 40)$ there is a real sense that coin tossing is fair to two decimal places but not to three.

The discussion above is based on joint work with J. Keller (1986). This physics-based analysis of randomness dates back to Poincaré. For a review of this along with work of Hopf, and much that is new, see Engel (1992).

In the next section we attempt a combination of the subjectivist and physical analysis.

C. *A Bayes-Newtonian synthesis.* After the careful study of the physics of coin tossing it is natural to try to incorporate this into the subjectivist analysis reported in Section A. Note that while that section dealt with zero binary outcomes, the refined analysis of Section B is based on the real-valued pairs (V, ω). To describe a sequence of tosses involves a sequence of initial conditions $(V_1, \omega_1), (V_2, \omega_2), \ldots$. deFinetti's theorem is in force in the following form.

THEOREM. If $\{V_i, \omega_i\}_{i=1}^{\infty}$ is an exchangeable sequence then there is a unique probability measure μ such that

$$P\{(V_1, \omega_1) \in A_1, (V_2, \omega_2) \in A_2, \ldots, (V_n, \omega_n) \in A_n\} = \int_{\mathcal{P}} \prod P(s_i) \mu(dP).$$

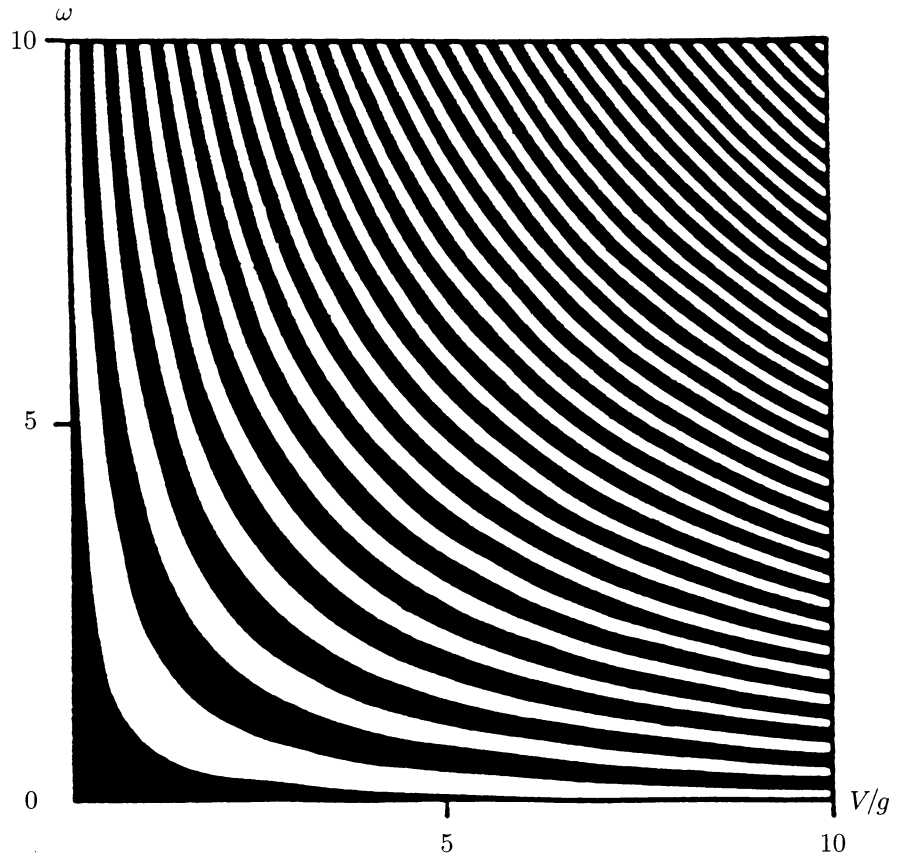

FIG. 2. Partition of phase space induced by heads and tails

On the left, n is arbitrary and $A_1, A_2, \ldots$ are Borel sets in the plane. On the right, $\mathcal{P}$ is the set of all probability measures on the plane and μ is a probability measure on the Borel sets of $\mathcal{P}$.

Let me pause to note the trouble we are in. Starting from an innocent coin-tossing problem, incorporating basic physical analysis, has led to the need for putting an apriori distribution on a huge space $\mathcal{P}$. While this is not impossible, it is not a simple thing to do in a meaningful way. See Diaconis and Freedman (1986) for discussion and illustrations of the strange things that can happen.

I find the above problem deeply disturbing. I call it the problem of thinking too much. It shows that the rational incorporation of information is not something to be undertaken lightly. The coin-tossing problem is quite a simple one. The extra information was clean and unambiguous as well. The possibilities for trouble in messier problems boggle the imagination.

4. What should we do? The sections above attempt to document the troubles of routine statistical analysis using big models. There is such a wealth of modeling in the theoretical and applied arenas that I feel a sense of alarm. What are all these papers doing?

A. *Cynical answers.* Models are used for

- *Publishing papers*—Research, promotion, and reputation seem tied to number of publications. Further, publishing has become easier (thanks to the computer).
- *Covering behinds*—In complex real-world problems, it is natural to order a study. This often results in impossible demands but at least a "state of the art analysis" might diffuse criticism.
- *Mathematical beauty*—At its best, the theory of statistics has elegant, telling results. The study of properties of models can be fascinating mathematics to say nothing of the intriguing mysteries of "correct" statistical inference.

B. *More temperate answers.* Models are used for

- *Data summary*—In exploratory data analysis a model may be a straw man allowing an in-depth analysis, examination of residuals, leading to further models that come close to being useful and correct.
- *Communication*—The simple summaries of a model may help to give an understandable picture of a complex situation.
- *Cleaning up data*—Often big data sets have missing data. If these parts are thrown out, very little may remain. Interpolation or imputation, usually model based, allows standard analyses such as graphical viewing or simple summaries to proceed.
- *Checking computer code*—In my day-to-day work I see many simulations. Roughly one third of these are wrong (every number on the page is "off"). Having simple models with well-understood solutions run along with the real problem helps check the output of complex computer code.
- *Understanding black-box algorithms*—Complex procedures such as neural nets or Box-Jenkins time series forecasts or worse are increasingly in use. One way to calibrate these is to use models with known properties or parameters. From these, one can generate data, feed it to the black box, and compare the output with the known "truth".

C. *What should we do?* The problems addressed in this essay have all the complexity of Hume's problem of induction or the creation of a believable philosophy of science. Nonetheless, I will offer some remedies that I feel would alleviate much of the problems in practice.

- *Do good work*—A good remedy for shoddy work is a collection of careful analyses where theory and modeling are honestly tested on fresh data. It actually does not seem so impossible.
- *Do not tolerate bad work*—I see an alarming tendency to be nice—to see or imagine some good piece in a poor data analysis or theoretical development. It is hard to do otherwise, but crucial in hiring, promotion, and elsewhere. I feel that the believability of statistical analyses is at an all time low. If we do not stand up and say something the field will vanish.

- *Defensive statistics as research*—Reading a complex study outside one's area is hard work. I have cited many efforts of David Freedman and his coworkers (my favorite is Freedman's (1997) decisive debunking of the efforts of Spirtes, Glymour, and Scheines to build an automated causality machine). This debunking work deserves tremendous respect; we should teach defensive statistics courses (like defensive driving). The reader who looks will find much to do.
- *Look for new foundations*—The measurement model of statistics as outlined in the first section above is a poor structure on which to hang statistical research. Statisticians are called upon to do many things. Perhaps if we can figure out what is good and useful about the huge modeling efforts, we can try to make a theory that optimizes this good rather than the parody "estimating θ based on independent repetitions".

Less alarmist treatments of the role of modeling in statistics can be found in Cox (1990) and Lehmann (1990).

REFERENCES

[1] D. Cox, *Role of models in statistical analysis*, Statistical Science **5**, 169–174 (1990)

[2] B. deFinetti, *Foresight: Its logical laws, its surjective sources*, Ann. l'Institut H. Poincaré **7**, translation in H. Kyburg, H. Smokler (Eds.), *Studies in Subjective Probability*, Krieger, Huntington, N.Y., 1937, pp. 55–131

[3] B. deFinetti, *Probability, Induction and Statistics*, Wiley, N.Y., 1972

[4] P. Diaconis, *Recent progress in deFinetti's notions of exchangeability*, in *Bayesian Statistics* (J. Bernardo, et al, eds.), vol. 3, Oxford Press, Oxford, 1988, pp. 111–125

[5] P. Diaconis and D. Freedman, *Partial exchangeability and sufficiency*, Proc. Indian Statist. Inst. Golden Jubilee Internat. Conf. Stat., Applications and New Directions, J. K. Ghosh, J. Roy (Eds.), Indian Statist. Institut., Calcutta, 1984, pp. 205–236

[6] P. Diaconis and D. Freedman, *On the consistency of Bayes estimators*, Ann. Statist. **14**, 1–67 (1986)

[7] P. Diaconis and D. Freedman, *On the uniform consistency of Bayes estimates for multinomial probabilities*, Ann. Statist. **18**, 1317–1327 (1990)

[8] E. Engel, *A Road to Randomness in Physical Systems*, Springer Lecture Notes in Statistics, No. 71, Springer-Verlag, N.Y., 1992

[9] D. Freedman, *Statistical models and shoe leather*, in *Sociological Methodology* (Peter Marsden, eds.), Amer. Sociol. Assoc. Washington, 1991

[10] D. Freedman, *From association to causation via regression*, Adv. Appl. Math. **18**, 59–110 (1997)

[11] D. Freedman, T. Rothenberg, and R. Such, *On energy policy models*, Jour. Business and Economic Statistics **1**, 24–36 (1983)

[12] D. Freedman and K. Wachter, *Heterogeneity and census adjustments for the intercensal base*, Statist. Sci **9**, 476–537 (1994)

[13] D. Freedman and H. Zeisel, *From mouse to man: the quantitative assessment of cancer risks*, Statist. Sci. **3**, 3–56 (1988)

[14] C. F. Gauss, *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*, Translated by C. H. Davis, Dover, N.Y., 1963

[15] C. F. Gauss, *Theory of the combination of observations least subject to errors*, translated by G. N. Stewart, SIAM, Philadelphia, 1995

[16] J. Keller, *The probability of heads*, Amer. Math. Monthly **93**, 191–196 (1986)

[17] E. Lehmann, *Theory of Point Estimation*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, N.Y., 1983

[18] E. Lehmann, *Testing Statistical Hypotheses*, second edition, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, N.Y., 1986

[19] E. Lehmann, *Model specification: The views of Fisher and Neyman, and later developments*, Statistical Science **5**, 160–168 (1990)