Uncertainty Modelling and Analysis: Theory and Applications
B.M. Ayyub and M.M. Gupta (Editors)

29. Uncertainty Estimates in Scientific Models: Lessons from Trends in Physical Measurements, Population and Energy Projections

Alexander I. Shlyakhter

Department of Physics and Northeast Regional Center for Global Environmental Change
Harvard University, Cambridge, MA 02138 USA

Results of a systematic analysis of actual *vs.* estimated uncertainty in scientific models are presented. Data sets include: i) time trends in the sequential measurements of the same physical quantity; ii) national population projections; iii) projections for the United States' energy sector. Probabilities of large deviations from the true values are parametrized by an exponential distribution with the slope determined by the data. An alternative parametrization by Levy stable distributions, based on the fractal model for the distribution of errors, is described. In practice, one can hedge against unsuspected uncertainties by inflating the reported uncertainty range by a default safety factor determined from the relevant historical data sets. This empirical approach can be used in the uncertainty analysis of the low probability/high consequence events (such as risk to public health from exposure to electromagnetic fields or risk of extreme sea-level rise resulting from global warming).

1. INTRODUCTION: THE USE OF PAST ERRORS TO PREDICT FUTURE ONES

It is well known that there is a strong tendency for researchers to underestimate uncertainties in results, thus decreasing their reliability and increasing the probability of "surprises" (Parrat 1961; McDonald 1972; Lichtenstein *et al.* 1982; Henrion and Fischoff 1986; Morgan and Henrion 1990; Cooke 1991). In this chapter, I present an overview of recent systematic analysis of actual errors in physical measurements, energy and population projections (Shlyakhter and Kammen 1992a,b; 1993, Shlyakhter *et al.* 1993, Kammen *et al.* 1993, Shlyakhter *et al.* 1994, Shlyakhter 1994). Standard uncertainty analysis is often plagued with its own uncertainties, which can have a profound effect on the tails of probability distributions. In particular, the commonly used 95 pe-cent confidence intervals are determined by the tails of distributions which are very sensitive to the underestimation of the true uncertainty. The history of natural and social sciences contains a wealth of data about reliability of uncertainty estimates in measurements and models.

Empirical methods of building confidence intervals around point estimates are widely used in weather, population, and economic forecasting (Murphy and Winkler 1977; Williams and Goodman 1971; Stoto 1983; Keilman 1990; Zarnowitz 1992). They rely on the assumption that the distribution of errors in future forecasts is the same as the distribution of these errors in past forecasts. In engineering, the importance of empirical control of experts' probability assessments is also well recognized (Cooke 1991). Science policy often hinges on reliable assessment of the uncertainty in predictions derived from various models. For example, uncertainty analysis of the low probability/high consequence events (such as estimating the probability of extreme sea-level rise resulting from global warming) is crucial for decision making in global climate change problem. It is the goal of this chapter to show how historic data on past overconfidence can be used to develop safety factors that can be applied to uncertainty estimates in current models.

2. PHYSICAL MEASUREMENTS

2.1 RANDOM ERRORS, SYSTEMATIC ERRORS, AND BLUNDERS

The general concept of error in physical measurements can be conveniently subdivided into three broad types: random errors, systematic errors, and blunders (Parrat 1961, ISO 1993). In general, the reported *experimental* error is some additive function of all three. Reported random errors mostly come from statistical fluctuations of the mean values obtained with finite number of trials (such as measurements of the length of a wire). These fluctuations are assumed to occur around the "true" values that would be obtained if the number of trials were infinite. Another source of random errors is the inherent variability of the system under study. In different trials, random errors can be positive or negative with equal probability.

Systematic errors (such as changes in dimensions due to thermal expansion) are common to a system and usually have the same algebraic sign in different trials. Various sources of systematic errors become known gradually with time as results obtained by independently working groups are compared. Therefore, some fraction of unsuspected systematic errors is always present in published uncertainty estimates. Blunders are outright mistakes (such as errors in transcription of the data). They are supposed to be identified and discarded before the uncertainty of the result is evaluated, but this is not always feasible. Routine scientific data sets contain 5-10% of gross errors (Hampel *et al.* 1986).

Uncertainties associated with random and systematic errors ("type A" and "type B" uncertainties) are combined using first-order Taylor series. This "combined standard uncertainty" (ISO 1993) then serves as the basis for calculating intervals corresponding to the required level of confidence. If the combined standard uncertainty is not dominated by type B uncertainty, then by the Central Limit Theorem (CLT) the distribution of the arithmetic mean of many observations around the true value is asymptotically normal.

Combined uncertainty is usually reported as an average of many trial measurements (corrected for all recognized systematic effects), A, and an associated standard deviation, Δ.

If the actual value of this quantity is a then the normalized deviation $x = (a - A)/\Delta$ follows the standard normal distribution. In that distribution, the range $A \pm 1.96\Delta$ has a 95 percent probability of including a. The presence of systematic errors, however, violates the assumptions necessary for use of the CLT. If most of the uncertainty comes from systematic errors, the usual justification for normal distribution does not apply. Despite this fact, the normal distribution is often a reasonable approximation for small deviations and remains implicit when researchers report measured values and their corresponding uncertainties.

2.2 ANALYSIS OF TRENDS IN PHYSICAL MEASUREMENTS

The first attempts to quantify overconfidence in physical measurements come from the work of Bukhvostov (1973) and Henrion and Fischoff (1986). They compared elementary particle properties and fundamental constants in early compilations with much more accurate values taken from a more recent compilation. A convenient measure of the deviation of "new" values from the "old" values is the normalized deviation $x = (a - A)/\Delta$, with a the exact value, A the measured value, and Δ the old standard deviation.

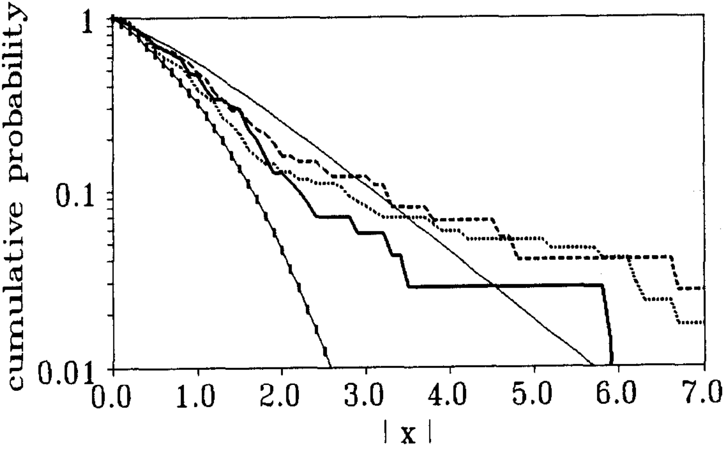

Figure 1. Probability of unexpected results in physical measurements. The plots show the cumulative probability, $S(x) = \int_x^\infty p(t)dt$, that new measurements (a) will be at least $|x|$ standard deviations (Δ) away from the old results (A); $x = (a - A)/\Delta$ as defined in the text. The cumulative probability distributions of $|x|$ are shown for the three data sets: particle data (LBL (1991); heavy solid line); magnetic moments of excited nuclear states (Avotina (1982); dotted line), neutron scattering lengths (Koester *et al.* (1991); heavy dashed line). Also plotted is a cumulative normal distribution, $\mathrm{erfc}(x/\sqrt{2})$ (thin solid line with markers), and compound exponential distribution with parameter $u=1$ from Figure 5 (solid line).

Shlyakhter *et al.* (1992a,b; 1993) expanded original studies by following trends in data sets derived from nuclear and particle physics: masses and lifetimes of elementary particles, magnetic moments and lifetimes of excited nuclear states, and neutron scattering lengths. All data sets were first converted into a standard format. Successive measurements of the same quantity comprised a block of data; a data set typically consisted of several hundred such blocks. In order to limit the effects of "noise" in the data on final results, two selection criteria were applied: i) new stated uncertainty had to be much smaller than the old one: $\Delta_{old}/\Delta_{new} \geq 4$; ii) only cases in which deviation from the true value did not exceed ten standard deviations were included in the analysis (in this way most blunders were excluded). The results confirm the earlier findings that a normal distribution grossly underestimates the probability of large deviations from the expected values. A new finding is that the *pattern* of overconfidence is similar in different kinds of measurements.

One can also look at the trends in the measurements of the same quantity in order to see whether experts become less overconfident with time. This is shown in Figure 2, where the results of successive measurements of a fundamental quantity, neutron lifetime, are presented together with the corresponding x values (data compiled by Yerozolimsky 1993). The "true" value a is calculated as the weighted average of the most recent measurements. Absolute errors decrease with time, but normalized errors, x, do not decrease.

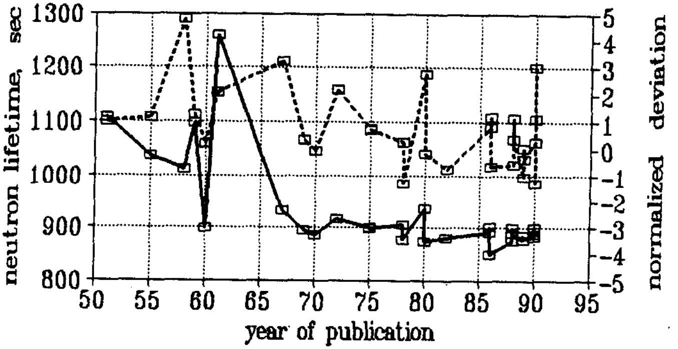

Figure 2. Trends in the measured values of the neutron lifetime. The error, a-$A(t)$, decreases with time and the measured values converge to a "true" value (heavy solid line; see scale on the left Y axis). Here a is the "true" value calculated as the weighted average of four measurements done in 1990; A(t) is the value measured at time t. Trends in the normalized deviations from the true value (heavy dashed line; see scale on the right Y axis). Plotted is the quantity $x(t)=[a-A(t)]/\Delta(t)$ *vs.* time; $\Delta(t)$ is the standard error estimated at time t. Normalized errors do not decrease with time and many deviations exceed two standard errors.

3. MODELS OF SOCIAL PARAMETERS: POPULATION AND ENERGY PROJECTIONS.

3.1 UNCERTAINTY IN FUTURE FORECASTS

Uncertainty in future forecasts is defined less formally than uncertainty in physical measurements. In this section an algorithm for analysis of uncertainty in historical forecasts is presented. One can estimate the standard deviation Δ of an equivalent normal distribution and then draw the empirical probability distributions of the deviations of the old forecasts from the true values normalized by Δ. Experts may not necessarily imply the normally distributed error terms, however, the users of the results tend to base their decisions on the assumption that deviations exceeding several uncertainty ranges are improbable. Comparison of errors in historical data sets with those predicted by the normal distribution provides a useful measure of the credibility of current uncertainty estimates.

Uncertainty in the forecasts is usually presented in the form of "reference," "lower" and "upper" estimates (R, L, and U respectively) that are obtained by running a model with different sets of exogenous parameters (e.g. the annual rate of growth). The range of scatter around the reference value R does not formally define a Gaussian standard deviation because the fundamental uncertainties involved (e.g. the rate of future economic growth) are frequently not stochastic. However, it is reasonable to assume that the range of parameter variation presented by a forecaster represents a subjective judgment about the probability that the true value $T \in [L, U]$. Generally, lower and upper bounds present what is believed to be an "envelope" most likely to bracket the true value and include the majority of possible outcomes.

Note that using the bounded distributions (such as triangular) assigns zero probability to large deviations. Historical data presented below, however, suggests that deviations far exceeding the expected uncertainty range are not uncommon. Therefore, using a normal (unbounded) distribution as a frame of reference *underestimates* true overconfidence.

The standard deviation of the equivalent normal distribution is calculated as follows:

a) Specify the subjective probability α that the true value will lie between the low (L) and high (U) estimates. I assume $\alpha=68\%$; larger values of α increase the discrepancy between the Gaussian model and that calculated by this method.

b) Draw an equivalent normal distribution that would have a specified cumulative probability α between L and U. For $\alpha = 68\%$ the standard deviation of the equivalent normal distribution is $(U - L)/2$ so that $x=2 \cdot (T-R)/(U-L)$. Therefore this choice of α corresponds to the usual practice of splitting the uncertainty range in half and using it as a surrogate of standard deviation.

c) If the reference value (R) is not in the middle of the (L, U) interval, x is defined using the uncertainty range on the same side of R as T: $x=(T-R)/(R-L)$ for R > T and $x=(T-R)/(U-R)$ for R < T .

3.2 POPULATION PROJECTIONS

The history of population projections provides an opportunity to test the reliability of uncertainty estimates in demographic models. Shlyakhter and Kammen (1992a,b; 1993) analyzed United Nations population projections, made in 1972, for the year 1985, census data for which can serve as the set of "exact" values, a.

The population data base includes projections from 164 nations with population exceeding 100,000 presented in the form of "high" and "medium" and "low" variants for each nation (UN 1991). Data for 31 countries were excluded due to extreme errors resulting for example from unanticipated international migration and cases of politically motivated reporting bias. Data for 133 nations satisfying the criteria $|x| < 10$ were included in the analysis.

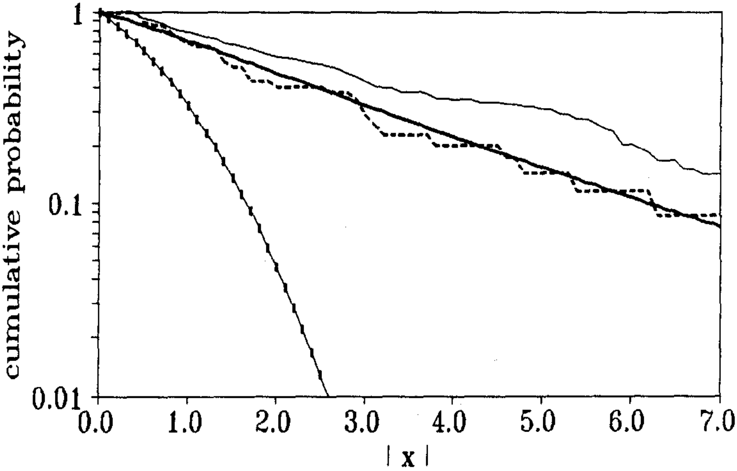

Figure 3. Population projections. The plots depict the cumulative probability, $S(x)=\int_x^\infty p(t)dt$, that true values (T) will be at least $|x|$ standard deviations (Δ) away from the reference value of old projections (R). The population data base is described in the text. The cumulative probability distributions of $|x|$ are shown for the total dataset of 133 countries (solid line) and for a subset of 37 industrialized countries (heavy dashed line). Also shown are the normal distribution (solid line with markers) and the compound distribution with $u=3$ from Figure 5 (heavy solid line).

The results are shown in Figure 3. Because all the population estimates come from an authoritative source - namely, the United Nations - it might be expected that systematic errors would be small, representing a well-calibrated model. The unsuspected uncertainty, however, is very large. Data for 37 industrialized countries (where data are generally more reliable) show a little less surprise, but probability of large errors is still grossly underestimated by the normal distribution.

3.3 ENERGY PROJECTIONS

Forecast of future energy consumption is a prerequisite for many major economic and policy decisions, such as how best to reduce carbon dioxide emissions to alleviate global warming, or how best to stimulate the pace of development of alternate sources of energy. Analysis of credibility of uncertainty estimates was performed using the largest coherent set of US energy forecasts for the year 1990 A.D., the Annual Energy Outlook (AEO 1992; Kammen *et al.* (1993) and Shlyakhter *et al.* (1994)).

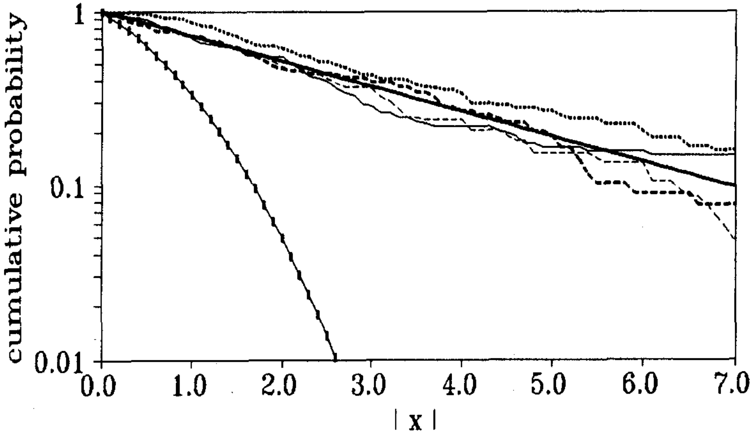

Figure 4. Annual Energy Outlook projections. The presentation is as in Figure 3: 1983 to 1990 (solid line); 1985 to 1990 (dashed line); 1987 to 1990 (heavy dotted line), aggregated sectors of economy (heavy dashed line); normal distribution (solid line with markers); distribution with $u=3.4$ (heavy solid line).

AEO projections for 1990 made in 1983, 1985, and 1987 consist of 182, 185, and 177 energy producing or consuming sectors of the U.S. economy respectively. The variation in the number of sectors resulted because the low and high projections coincided in some cases, and no corresponding uncertainty range could be derived.

In 47, 50, and 47 cases respectively, the *x* values (calculated as described in section 3.1) exceeded 100; such cases were omitted as they apparently could not be due to parametric uncertainty of the AEO model. For all remaining cases the *x* values were calculated and the frequency distributions analyzed. The distribution of signed *x* values is approximately symmetric with respect to zero. There is no large systematic bias (e.g. a gross underestimation of energy consumption in all or many sectors) and no strong trends in the scattergrams of *x* values; this indicates that the forecasts are generally independent.

Figure 4 shows the cumulative probability distributions of $|x|$ for the projections made for 1990 in 1983, 1985, and 1987 together with the Gaussian and exponential distributions. The three empirical distributions are strikingly similar. Although the absolute error in forecasts made in 1987 for 1990 is somewhat smaller than that made in 1983 for 1990, the range of uncertainty is also smaller so that the probability of "large" deviations relative to the observed uncertainty is roughly the same as for the other two years. One would expect that energy forecasts for aggregated sectors of economy would be more reliable than forecasts for individual sectors. However, this appears not to be the case (Figure 4, heavy dashed line).

4. PARAMETRIZATION OF THE OBSERVED DISTRIBUTION OF ERRORS

4.1 EXPONENTIAL PARAMETRIZATION

Bukhvostov (1973) and Shlyakhter and Kammen (1993) suggested simple heuristic arguments to describe how an exponential distribution of errors might arise. Let us assume that the estimate of the mean, *A*, is unbiased but that the estimate of the true standard deviation, Δ', is randomly biased with a distribution $f(t)$ where $t=\Delta'/\Delta$. Here Δ is the estimated standard deviation. In other words, I assume that the deviations normalized by Δ', $x'=(a-A)/\Delta'$, follow the standard normal distribution while the deviations normalized by Δ, $x=(a-A)/\Delta$, follow a normal distribution with a randomly chosen standard deviation *t*:

$$p_t(x) = \frac{1}{\sqrt{2\pi}\,t} e^{-\frac{x^2}{2t^2}} \qquad (1)$$

Integrating over all values of *t* gives a compound distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{dt}{t} f(t)\, e^{-\frac{x^2}{2t^2}} \qquad (2)$$

If f(t) has a sharp peak near t=1, Eq. (2) reduces to the normal distribution. If f(t) is broad, however, the result is different. For simplicity, let us assume that for large *t*, *f(t)* follows the Gaussian distribution with the standard deviation *u*: *f(t) ~ exp[-t²/(2u²)]*. The main contribution to the integral in Eq.(2) comes from the vicinity of the saddle point where the exponential term reaches a maximum (for $t=t_{max}$: $t^2_{max}=u\ |\ x\ |$). It is straightforward to show that, for large values of *x*, the probability distribution *p(x)* is not *Gaussian* but *exponential*: *p(x) ~ exp(-|x|/u)*. In order to reflect the fact that experts are mostly overconfident ($\Delta' \geq \Delta$), I use a truncated normal form of *f(t)*:

$$f(t)=0,\ t\leq 1$$
$$f(t)=\sqrt{\frac{2}{\pi}}\frac{1}{u}e^{-\frac{(t-1)^2}{2u^2}},\ t>1 \tag{3}$$

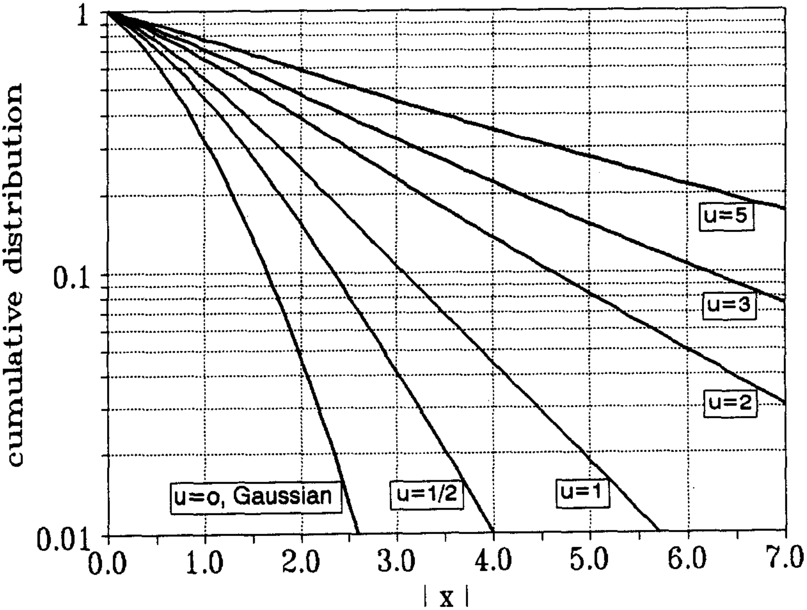

Figure 5. One-parameter family of compound distributions. Parameter *u* is defined in Eq. (3); it is a measure of uncertainty in the standard deviation Δ). The values of *u* are indicated in the figure. The curves demonstrate the continuum of probability distributions, from Gaussian (u=0) to exponential (*u* > 1).

Integrating Eq. (2) with *f(t)* from Eq. (3) gives the cumulative probability *S(x)* of deviations exceeding $|x|$:

$$S(x) = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{u} \int_1^\infty e^{-\frac{(t-1)^2}{2u^2}} \operatorname{erfc}\left(\frac{|x|}{t\sqrt{2}}\right) dt \tag{4}$$

The normal ($u = 0$) and exponential distributions ($u > 1$) are members of a single-parameter family of curves (Figure 5). For quick estimates for $u \geq 1$, $x \geq 3$, one can use the approximation $e^{-|x|/(0.7u + 0.6)}$. In this framework, the parametric uncertainty can be quantified by analyzing the record of prior projections and estimating the value of u. Data presented in Figures 1,3,4 show that $u \sim 1$ for physical constants and $u \sim 3$ for population and energy projections.

Parametrization with compound distributions described above is not the only one possible. For the data sets of physical measurements shown in Figure 1, formal tests for exponentiality based on Shapiro-Wilk W-statistics (Shapiro and Gross 1981) cannot reject exponential parametrization at the 95% level, but only if the data set is limited to $|x| < 4$. Further work on parametrizations for different types of data is needed. One possibility is discussed below.

4.2 LEVY DISTRIBUTION AND FRACTAL MODEL FOR ANALYSIS OF ERRORS

The total error in a physical measurement or in the value of a model parameter is a sum of many random variables. Each single source of error is represented by one term in the sum. It is important to realize that there is no upper limit for possible errors. In fact, there is a wide spectrum of uncertainties extending from negligible systematic uncertainties to gross errors caused by the use of a wrong model or by a blunder. An analyst usually has a scale in mind for "important" uncertainties: smaller uncertainties are assumed to be negligible and are excluded from the detailed analysis. "Important" uncertainties are carefully evaluated and combined to produce the final estimate of the combined standard uncertainty. However, errors of larger scales, particularly those arising from the unrecognized uncertainties, are also possible. Although such gross errors are much less frequent than small errors, their effect on the total error can be large. This heuristic model is an attempt to describe the human thought processes that are responsible for the observed pattern of overconfidence.

Levy generalized the Central Limit Theorem for the case of sums of random variables which may have infinite second moments. Consider a random walk where each jump length is chosen from the distribution *p(x)*. Levy asked when the distribution of the sum of n steps $p_n(x)$ will have the same functional form as *p(x)*. This is the basic question of the theory of fractals: when does the whole (the sum) look like any of its parts? Levy discovered the general solution to this problem (Mandelbrot 1983; Shlesinger *et al.* 1993).

A stable Levy distribution is a long-tail generalization of a normal distribution. It is the only possible limiting distribution for sums of independent identically distributed random variables (Feller 1966; Fama and Roll 1968,1971). For symmetric stable Levy distributions, the characteristic function is $exp(-|ct|^{\alpha})$. Probability density, $p(x)$, has two parameters: c, scale and α, characteristic exponent.

$$p(x)=\frac{1}{\pi}\cdot\int_0^{\infty}e^{-(ct)^{\alpha}}\cdot\cos(tx)\,dt \qquad (5)$$

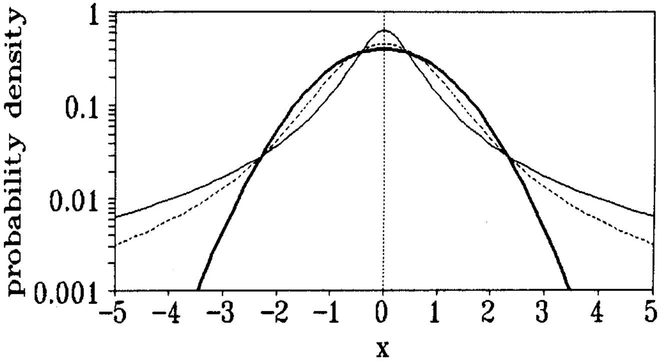

Figure 6. Probability densities for Levy distribution with $c=1/\sqrt{2}$: $\alpha=2$ (normal, heavy solid line); $\alpha=1.5$ (dotted line); $\alpha=1$ (Cauchy, solid line).

For Levy distributions, the generalized form of the reproductive property holds: for any two independent quantities, X_1 and X_2, each following the Levy distribution with parameter α, the sum $X=X_1+X_2$ also follows the Levy distribution with parameter α and average value of X^{α} is the sum of the average values of X_1^{α} and X_2^{α}. The case $\alpha=2$ is normal distribution with the standard deviation $c\cdot\sqrt{2}$, $p(x)=1/(2c\cdot\sqrt{\pi})exp(-x^2/4c^2)$. For comparison with the standard normal distribution, I consider Levy distributions with $c=1/\sqrt{2}$ and one free parameter, α. For $\alpha=1$, Levy distribution is reduced to Cauchy distribution, $p(x)=c/[\pi(c^2+x^2)]$. For $2>\alpha\geq 1/2$, Levy distributions are tabulated (Fama and Roll 1968).

A special example of a random walk described by a Levy distribution is provided by Weierstrass random walks, in which jumps of size ±1, $\pm b$, $\pm b^2$ and so on can occur but jumps an order of magnitude longer in base b occur an order of magnitude less often in base a (Shlesinger *et al.* 1993). The characteristic exponent (fractal dimension of the random walk path) is given in this case by $\alpha=ln(a)/ln(b)$. When $b^2\leq a$, this random walk will produce a Gaussian distribution; when $b^2>a$, it will produce a Levy distribution.

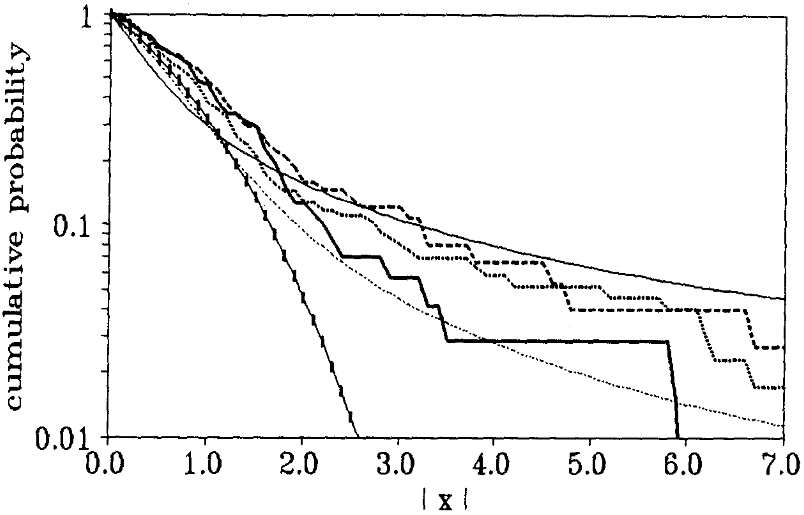

Figure 7. Cumulative Levy stable distributions with $c=1/\sqrt{2}$ for three values of parameter α and three data sets shown in Figure 1: $\alpha=2$ (normal, solid line with markers); $\alpha=1.5$ (dotted line); $\alpha=1$ (Cauchy, solid line); particle data (heavy solid line); magnetic moments of excited nuclear states (heavy dotted line), neutron scattering lengths (heavy dashed line).

Cumulative Levy stable distributions for three values of parameter α are shown in Figure 7 together with the empirical distributions for three data sets: magnetic moments, neutron scattering and particle data. It appears that Levy stable distributions suggest a useful tool for modeling the observed distributions of errors.

5. APPLICATIONS: LOW PROBABILITY/HIGH CONSEQUENCE EVENTS

Choosing appropriate safety factors as a hedge against unsuspected errors is particularly important in the uncertainty analysis of many situations in public policy. These describe events with low probability but high consequences that are determined by the tails of the probability distributions. I illustrate possible applications using two open questions derived from risk analysis: estimates of risk to public health from exposure to electromagnetic fields and the risk of extreme sea-level rise resulting from global warming. For applications of the inflated confidence intervals to population and energy projections see Shlyakhter and Kammen (1993), Kammen *et al.* (1993), and Shlyakhter *et al.* (1994).

5.1 HOW CONVINCING ARE OBSERVED ASSOCIATIONS OF LEUKEMIA CASES WITH THE EXPOSURE TO ELECTROMAGNETIC FIELDS?

Epidemiologic studies provide the basis for many public health decisions. Results of such studies are usually presented in the form of the 95 percent confidence interval (CI) for relative risk, *RR* (or odds ratio, *OR*, for case-control studies), which accounts for the uncertainty caused by the finite sample size. The result is termed "statistically significant positive finding" if the lower bound of the confidence interval lies above one. The trouble here is that possible sources of bias are only taken into account on the basis of plausible assumptions which cannot be independently tested for the population under study. Heterogeneity of case and control groups, misclassification, and confounding in the observational studies are the analogues of systematic uncertainties in physical measurements (Armstrong *et al.* 1992). One can present the results of many epidemiological studies of the same outcome as a probability distribution of normalized deviations from $RR=1$. Comparison with distributions of errors in other situations can help us to better understand how convincing the evidence of elevated risk really is.

To illustrate this point, consider the occupational studies of the effects of electromagnetic fields (EMF) on leukemia compiled in a recent study (ORAU 1992, Table V-10). These are reported as risk ratios with 95% confidence intervals and are shown in Figure 8. The data set consists of 31 studies for which *RR* values for all types of leukemia were reported and a subset of 15 studies for which *RR* values for acute myelogenous leukemia were also reported. In 20 out of 31 studies, were *RR* values above one; in 13 of the set of 15 were *RR* values above one. Four studies out of 20 and another four studies out of 15 can be considered statistically positive findings. By chance alone, one would expect only 2.5% such findings when real risk is not elevated. However, when compared with physical measurements, this is less surprising. In order to perform such a comparison I assume that the true relative risk is $RR=1.0$ and plot the probability distribution of the normalized deviations of *ln(RR)* from zero.

For example, in the study of leukemia among electricians exposed to EMF (shown as study #12 in Fig. 8; Stern *et al.* 1986), an *RR* of 3.0 was reported. The 95% CI reported was 1.3-7.0. Since relative risk can take any value from zero to infinity, transformation to the natural logarithm is used to make the range of *RR* values symmetric around $RR=1$. For large samples, *ln(RR)* follows a normal distribution (Rothman 1986). The confidence interval of *ln(RR)* in the study #12 is 0.26-1.95. The standard deviation of *ln(RR)* is equal to *ln(GSD)* where *GSD* is the geometric standard deviation of *RR*; $ln(GSD) = (1.95-0.26)/2/1.96=0.43$. The middle of the confidence interval, $ln(RR)=1.10$, is $x=1.10/0.43=2.56$ standard deviations away from the postulated true value, $ln(RR)=0.0$.

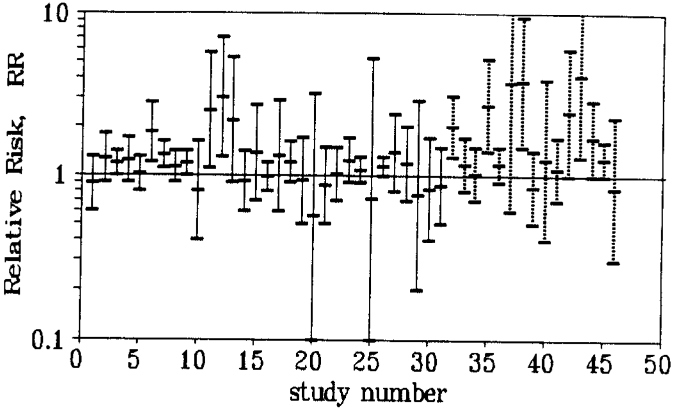

Figure 8. Risks of leukemia from occupational exposure to EMF. Data for 31 studies (1 to 31) in which combined *RR* for all leukemia was reported (solid error bars) and a subset of 15 studies (32 to 45) in which *RR* for acute myelogenous leukemia was reported (heavy dotted error bars) compiled in ORAU (1992) are shown.

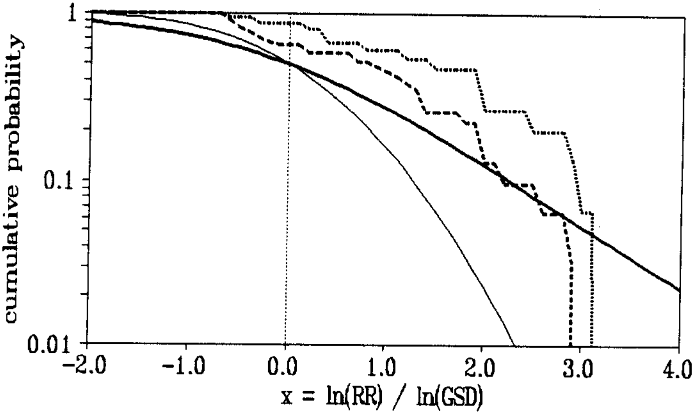

Figure 9. Data from Figure 8 presented in a different format. Distributions of the normalized deviations of *ln(RR)* from zero for 31 studies of all types of leukemia (solid line) and for 15 studies of acute myelogenous leukemia (heavy dotted line) are shown together with the curves for $u=0$ (normal distribution, solid line with markers), and $u=1$ (heavy solid line).

The cumulative probability distribution of x values is shown in Figure 9 together with the curves for $u=0$ (normal distribution) and $u=1$ (for both positive and negative x). The observed distributions of x have longer tails than the normal distribution and are better described by the curve $u=1$. This could be due to a truly elevated risk, a positive bias (such as increased chance for a positive finding to be published) or a combination of the above.

The fraction of statistically significant positive findings is similar to the fraction of large deviations from the true values in physical measurements. Since the quality of data used in epidemiological studies is lower than the quality of data used in experimental science (Feinstein 1988), one is tempted to conclude from Figure 9 that the distribution of *RR* values is compatible with the null hypothesis of *RR=1*. However, such a conclusion would be premature. Epidemiologists may argue that most of the unaccounted systematic uncertainties (such as non-random misclassification of exposure status) move *RR* closer to the null value *RR*=1 (Rothman 1986) so that the observed fraction of statistically significant positive findings must be caused by a truly elevated risk.

In order to find the answer, it is necessary to analyze in a similar fashion the distribution of *RR* values in several sets of observational studies where it is known that the true relative risk is not elevated. A comparison with the normal distribution will show if a default safety factor for the 95% confidence intervals could be derived. One possible source of data for such analysis is provided by the numerous studies of the effects of very low doses of radiation (Shihab-Eldin *et al.* 1992). Interestingly, the distribution of AML studies exibits longer tails than the distribution for all types of leukemia, although the systematic errors for these types of study should be similar; this issue deserves further analysis.

Note that for large case-control studies that produce tight statistical confidence intervals, systematic errors are relatively more important than random errors. For a weakly positive finding even a small inflation of the confidence intervals can push the lower bound of the confidence interval for *RR* below one and make the conclusions of a study much less convincing (Shlyakhter *et al.* 1993). Epidemiologists have been generally aware of this and the authors of the ORAU (1992) report cited above *do not* consider the collection of occupational studies as convincing; I merely confirm and describe this statement. Nonepidemiological scientists, who are the consumers of such results need more than just reported confidence intervals. These users must retain their own common sense in evaluating how convincing the reported evidence is. My recommended procedure for presenting the results of epidemiological studies may help in this.

5.2. PROBABILITY OF EXTREME SEA-LEVEL RISE

Estimating the probability of extreme sea-level due to greenhouse warming is a natural application of the proposed technique of uncertainty characterization. The causal sequence leading to sea-level rise is as follows: population -> energy production -> CO_2 emissions -

> greenhouse warming -> sea-level rise. One can present the sea-level rise as a product of five factors which are roughly independent:

$$h=Population \cdot \left(\frac{energy}{person}\right) \cdot \left(\frac{CO_2}{energy}\right) \cdot \left(\frac{\Delta T}{CO_2}\right) \cdot \left(\frac{h}{\Delta T}\right) \qquad (6)$$

The first factor is the world population; the second factor is energy production *per capita*; the third factor is CO_2 emissions per unit energy production; the fourth factor is temperature increase ΔT *per* unit rise in CO_2; the fifth factor is sea-level rise *per* unit temperature increase ΔT. For each factor there are uncertainties in its respective model. In particular, the last two factors include uncertainties in physical models of climate system and sea-level rise. Shlyakhter and Kammen (1992a,b; 1993) applied inflated confidence intervals to the results of Oerlemans (1989), who assumed a normal distribution of uncertainties in the physical model for sea-level rise. He used a simple fit for the temperature rise based on a "Business-as-Usual" scenario (Houghton *et al.* 1990): $T = \alpha(t - 1850)^3$, where t is time (yrs), $\alpha = 27 \times 10^{-8}$ °K yr^{-3} and assumed that Δ for each parameter was 35% of the mean value.

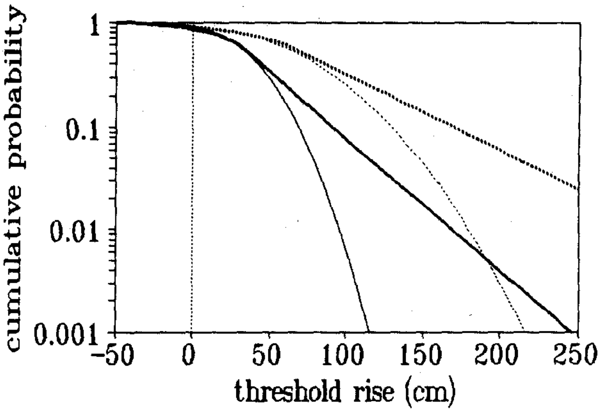

Figure 10. Projections of sea-level rise for 2050 A.D. and 2100 A.D. The probability of a sea-level rise greater than a given threshold is plotted for the normal probability (2050: thin solid line; 2100: thin dashed line) and for distribution with $u=1$ in Figure 5. (2050: heavy solid line; 2100: heavy dashed line). Note that a fall in sea-level is also possible.

An important assumption is that the uncertainty in individual contributions to changes in sea-level is characterized by combining independent normal probability distributions, hence: $\Delta^2 = \Delta^2_{glac} + \Delta^2_{ant} + \Delta^2_{green} + \Delta^2_{wais} + \Delta^2_{expa}$ + internal variability. The subscripts refer to the effect of glaciers, the Antarctic, Greenland and West Antarctic ice sheets and thermal expansion of sea water. Uncertainties that are not reflected in the combined standard uncertainty in Oerlemans' model include possible feedbacks linking the 5 factors in Eq.(6), and uncertainties in future emissions of greenhouse gases which are determined by population growth and energy demand. Since no historical datasets of actual errors in the predictions of sea-level rise are available, I use the value $u=1$ derived from physical measurements as a lower estimate of unsuspected uncertainty.

Oerlemans (1989) projects a sea-level rise with errors comparable to the estimates themselves: 33 ± 32 cm in 2050 and 65 ± 57 cm in 2100. Extreme sea-level rise, of perhaps 150 cm in 50 years, is of prime regulatory concern. A comparison of Gaussian and exponential threshold probabilities for sea-level rise by 2050 and 2100 A.D. is presented in Figure 10. The probability of sea level rise greater than 150 cm by the year 2050 is $5.1 \cdot 10^{-6}$ according to Oerlemans' model, but with the inflated standard deviations, the probability is 1.8 percent (3,400 times greater). In my view, any policy decision should be based on the second, rather than the first number.

6. SUMMARY

Empirical analysis of actual uncertainties in scientific models can provide valuable information about the credibility of current uncertainty estimates. Data sets for such analysis can be derived, for example, from time trends in sequential measurements of the same physical quantity (for models used in natural sciences) or comparison of energy and population projections with actual values that became available later (for models of social parameters). For all data sets analyzed so far, distributions of deviations from the true values show the same pattern: long tails that do not follow normal distribution but can be pragmatically parametrized by exponential distribution, with the slope determined by the data. A more promising description is based on Levy distribution.

The additional component of uncertainty derived from such analyses can be viewed as a safety factor accounting for overconfidence of the experts. It therefore incorporates the possibility of human error into the framework of uncertainty analysis. Although data on past misunderstanding of a given situation cannot prevent our current misunderstanding of a significantly different situation, statistical analysis of the frequency of past underestimates of uncertainty can provide useful clues to the choice of the appropriate safety factors.

A legitimate concern about the use of the default inflation factors for the confidence intervals is that this procedure ignores the specifics of particular studies. Some of the studies may be of much higher quality than an average study in the data set from which the inflation coefficient was derived. Unfortunately, elicitation of expert opinions about each study is rarely feasible. The user can hedge against unsuspected uncertainties, multiplying the reported uncertainty range by a safety factor. As recommended in the ISO (1993) report, this factor should be clearly specified and applied only *after* the uncertainty has been determined by a standard method, so that the operation may be easily reversed.

Another interesting question is how to truncate the long tails of the inflated probability distributions in order to avoid absurd conclusions. This involves the imposition of the constraints external to the model itself. If the constraints are sharp, it is easy to truncate the tails of the probability distribution (and renormalize it accordingly). However, in many cases available additional evidence is not sharp (such as upper limits on health risks resulting from the negative epidemiological studies). Such "fuzzy" external restrictions should be reflected in the uncertainty estimates. This question is important both for the standard uncertainty analysis and the improved version proposed here.

Interestingly, the u values derived from the data sets presented here cover a rather narrow interval: $u \sim 1$ for physical constants and $u \sim 3$ for current models of population growth and energy projections. Although there are many different scientific models with specific sources of uncertainty, it may be possible to combine them in several distinct groups according to reliability of past uncertainty estimates. This would allow the use of default inflation factors when no historical data sets are available.

ACKNOWLEDGEMENTS

This research was partially funded by the US Department of Energy's (DOE) National Institute for Global Environmental Change (NIGEC) through the NIGEC Northeast Regional Center at Harvard University (DOE Cooperative Agreement No. DE-FC03-90ER61010) and by the Biological Effects Branch of the U.S. Department of Air Force through the contract F33615-92-C-0602. I am deeply grateful to J.S.Evans, J.D.Graham, J.K.Hammit, D.M.Kammen, S.L.Sakharov, I.A.Shlyakhter, M.S.Raykin, B.G.Yerozolimsky, and R.Wilson for discussions and comments, and to B.N.Taylor for providing a copy of the ISO (1993) report. Electronic mail address: shlyakhter@huhepl.harvard.edu.

REFERENCES

AEO (1992) *Annual Energy Outlook with projections to 2010*, Energy Information Administration, US Department of Energy, Washington, DC 20585, DOE/EIA-0383 (92).

Armstrong, B.K., White E., and Saracci, R. (1992) *Principles of Exposure Measurement in Epidemiology*, Oxford University Press.

Avotina M.P., Kondurov I.A., and Sbitneva O.N. (1982) "Tables of nuclear moments and deformation parameters of atomic nuclei," Leningrad Nuclear Physics Institute Report.

Bukhvostov, A.P. (1973) "On the statistical meaning of experimental results," Leningrad Nuclear Physics Institute Preprint, LNPI-45 (in Russian).

Cooke R.M. (1991) *Experts in Uncertainty: Opinion and Subjective probability in Science*, Oxford University Press.

Fama E.F. and R. Roll (1968) "Some Properties of Symmetric Stable Distributions," *Journal of the American Statistical Association*, **63**, 817-836.

Fama E.F. and R. Roll (1971) "Parameter Estimates for Symmetric Stable Distributions," *Journal of the American Statistical Association*, **66**, 331-338.

Nuclear Physics Institute Preprint, LNPI-45 (in Russian).

Cooke R.M. (1991) *Experts in Uncertainty: Opinion and Subjective probability in Science*, Oxford University Press.

Fama E.F. and R. Roll (1968) "Some Properties of Symmetric Stable Distributions," *Journal of the American Statistical Association*, **63**, 817-836.

Fama E.F. and R. Roll (1971) "Parameter Estimates for Symmetric Stable Distributions," *Journal of the American Statistical Association*, **66**, 331-338.

Feinstein A.R. (1988), "Scientific Standards in Epidemiologic Studies of the Menace of Daily Life," *Science*, **242**, 1257-1263.

Feller, W.(1966) *Introduction to Probability Theory and its Applications*, Vol. 2. John Wiley.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986) "*Robust Statistics: the approach based on influence functions*," John Wiley & Sons, New York.

Henrion M. and B. Fischoff (1986) "Assessing uncertainty in physical constants," *American Journal of Physics* **54**, 791 - 797.

Houghton, J. T., Jenkins, G. J. & Ephraums, J. J., eds. (1990) Intergovernmental Panel on Climate Change, Climate Change: The IPCC Scientific Assessment, Cambridge University Press, Cambridge.

ISO (1993) "Guide to the Expression of Uncertainty in Measurement," International Organization for Standardization (ISO), Geneva, Switzerland.

Kammen D.M., A.I. Shlyakhter, C.L. Broido and R. Wilson (1993) "Non-Gaussian Uncertainty Distributions: Historical Trends and Forecasts of the United States Energy Sector, 1983-2010," *Proceedings of ISUMA'93 the Second International Symposium on Uncertainty Modeling and Analysis*, University of Maryland, College Park, Maryland, April 25-28, 1993, p. 112-119, IEEE Computer Soc. Press, Los Alamitos, California.

Keilman N.W. (1990) *Uncertainty in National Population Forecasting: Issues, Backgrounds, Analyses, Recommendations*, Amsterdam, Swets and Zeitlinger, Publications of the Netherlands Interdisciplinary Demographic Institute (NIDI) and the Population and Family Study Centre (CBGS), vol. 20.

Koester L., Rauch H., and Seymann E. (1991) "Neutron scattering lengths: a survey of experimental data and methods", *Atomic Data and Nuclear Data Tables*, 49:65-120.

LBL (1991) Data file of elementary particle masses and lifetimes maintained by Nuclear Data Center, Lawrence Berkeley National Laboratory.

Lichtenstein S. and Fischoff B. (1980) "Training for calibration," *Organizational Behavior and Human Performance*, **28**, 149-171.

Lichtenstein, S. B.Fischoff, and L.D. Phillips (1982) "Calibration of Probabilities: The State of the art to 1980," in: Kahneman, D., P. Slovic, and A. Tversky, eds. *Judgment Under Uncertainty*, Cambridge University Press, Cambridge, p.306-334.

Mandelbrot B. (1983) "The Fractal Geometry of Nature," Freeman and Co., New York.

Macdonald J.R. (1972) "Are the data worth owning?" *Science*, **176**, 1377.

Morgan M.G. and M. Henrion (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, New York.

Murphy A.H. and Winkler R.L. (1977) "Reliability of subjective probability forecasts of precipitation and temperature," *Journal of the Royal Statistical Society*, Ser.C, **26**, 41-47.

ORAU (1992) *Health Effects of Low-Frequency Electric and Magnetic Fields*, Prepared by Oak Ridge Associated Universities Panel for the Committee on Interagency Radiation Research and Policy Coordination, Oak Ridge Associated Universities, ORAU 92/F8.

Oerlemans J. (1989) "A Projection of Future Sea-level," *Climatic Change*, 15, 151-174.

Parrat L.G. (1961), *Probability and experimental errors in science*, John Wiley, New York.

Rothman K.J. (1986) *Modern Epidemiology*, Little, Brown & Co., Boston.

Shlesinger M.F., Zaslavsky G.M., Klafter J. (1993) "Strange Kinetics," *Nature*, **363**, 31-37.

Shapiro S.S. and Gross A.J. (1982) *Statistical Modeling Techniques*, Statistics: textbooks and monographs, v.38, Marcel Dekker, Inc., New York and Basel.

Shihab-Eldin A., Shlyakhter A., and Wilson R. (1992), "Is there a large risk of radiation? A critical review of pessimistic claims", *Environment International*, v. 18, 117-151.

Shlyakhter A.I. and D.M. Kammen (1992a) "Sea-level rise or fall?" *Nature*, **357**, 25.

Shlyakhter A.I. and D.M. Kammen (1992b) "Estimating the range of uncertainty in future development from trends in physical constants and predictions of global change," CSIA discussion paper 92-06, Kennedy School of Government, Harvard University, July 1992.

Shlyakhter A.I., I.A. Shlyakhter, C.L. Broido, and R. Wilson (1993a) "Estimating uncertainty in physical measurements and observational studies: lessons from trends in nuclear data," *Proceedings of ISUMA'93 the Second International Symposium on Uncertainty Modeling and Analysis*, University of Maryland, College Park, Maryland, April 25-28, 1993, p. 310-317, IEEE Computer Soc. Press, Los Alamitos, California.

Shlyakhter A.I. and D.M. Kammen (1993b) "Uncertainties in Modeling Low Probability/High Consequence Events: Application to Population Projections and Models of Sea-level Rise," *ibid.* p. 246-253.

Shlyakhter A.I. (1994) "Improved Framework for Uncertainty Analysis: Accounting for Unsuspected Errors, *Risk Analysis*, in press.

Shlyakhter A.I., D.M. Kammen, C.L. Broido and R. Wilson (1994) "Quantifying the Credibility of Energy Projections from Trends in Past Data: the U.S. Energy Sector," *Energy Policy*, in press.

Stern F.B. *et al.* (1986) "A case-control study of leukemia at a naval shipyard," *Am. J.. Epidemiol.*, **123**, 980-992.

Stoto M.A. (1983) "The Accuracy of Population Projections," *Journal of the American Statistical Association*, **78**, 13-20.

UN (1991) United Nations, *World Population Prospects*, Population Studies Paper No. 120.

Williams W.H. and M.L. Goodman (1971) "A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts," *Journal of the American Statistical Association*, **66** (366), 752-754.

Yerozolimsky B.G. (1993) Private communication.

Zarnowitz V. (1992) *Business Cycles: Theory, History, Indicators, and Forecasting*, The University of Chicago Press.