Chapter 10

Remarks on the Use of Probabilities in Demography and Forecasting

Juha M. Alho

10.1 Introduction

The concept of “probability” is used as a step in life table construction to get the expected number of survivors in a cohort. However, in traditional texts on demographic methods (*e.g.*, Shryock and Siegel 1976), variance in the number of survivors plays no role. Similarly, concepts of estimation, estimation error, and bias are routinely used, but standard error and sampling distribution are not (except in connection with sample surveys). Although statistically satisfactory accounts of the life table theory have existed for a long time (*e.g.*, Chiang 1968; Hoem 1970), a reason for neglecting population level random variability, and statistical estimation error, has been that the populations being studied are so large that random error must be so small as not to matter, in practice.

In contrast, when statistical methods started to become used in population forecasting in the 1970s, 1980s and 1990s, some of the resulting prediction intervals have been criticized as being implausibly wide. This view has not often been expressed in print, but Smith (2001, pp. 70–71) provides an example. Others, especially sociologically minded critics have gone further and argued that due to the nature of social phenomena, the application of probability concepts in general, is inappropriate. On the other hand, demographers coming with an economics background have tended to find probabilistic thinking more palatable.

The purpose of the following remarks is to review some probability models, and show how the apparent contradiction arises. We will see that the basic principles have been known for decades. The basic cause of the difficulties – and disagreements – is that there are several layers of probabilities that can be considered. Consequently, it is essential to be explicit about the details of the model.

J. M. Alho (✉)
University of Joensuu, Joensuu, Finland
e-mail: juha.alho@helsinki.fi

10.2 Binomial and Poisson Models

As emphasized by good introductory texts on statistics (*e.g.*, Freedman et al. 1978, p. 497), the concept of probability can only be made precise in the context of a mathematical model. To understand why one often might ignore other aspects of random variables besides expectation, let us construct a model for the survival of a cohort of size n for 1 year. For each individual $i = 1, \dots, n$, define an indicator variable such that $X_i = 1$, if i dies during the year, and $X_i = 0$ otherwise. The total number of deaths is then $X = X_1 + \dots + X_n$. We assume that the X_i 's are random variables (*i.e.*, their values are determined by a chance experiment). Suppose we make an assumption concerning their expectation

$$E[X_i] = q, i = 1, \dots, n, \quad (10.1)$$

and assume that

$$X_1, \dots, X_n \text{ are independent.} \quad (10.2)$$

It follows that X has a binomial distribution, $X \sim \text{Bin}(n, q)$. As is well known, we have the expectation $E[X] = nq$, and variance $\text{Var}(X) = n(q - q^2)$. Therefore, the coefficient of variation is $C = ((1 - q)/nq)^{1/2}$.

Now, in industrialized countries the probability of death is about 1% and population size can be in the millions, so relative variation can, indeed, be small. For example, if $q = 0.01$ and $n = 1000,000$, we have that $C = 0.01$. Or, the relative random variability induced by the model defined in (10.1) and (10.2) is about 1%. Equivalent calculations have already been presented by Pollard (1968), for example.

One might object to the conclusion that relative variability is negligible on the grounds that (10.1) does not hold: surely people of different ages (and of different sex, socio-economic status etc.) have different probabilities of death. Therefore, suppose that

$$E[X_i] = q_i, \text{ with } q = (q_1 + \dots + q_n)/n. \quad (10.3)$$

In this case

$$\text{Var}(X) = nq - \sum_{i=1}^n q_i^2. \quad (10.4)$$

However, it follows from the Cauchy-Schwarz inequality that

$$n \sum_{i=1}^n q_i^2 \geq n^2 q^2. \quad (10.5)$$

Therefore, we have that the variance (10.4) is actually *less* than the binomial variance. The naive argument based on population heterogeneity simply does not hold.

Before proceeding further, let us note that, apart from substantive factors, heterogeneity of the type (10.3) is imposed on demographic data, because vital events are typically classified by age, so individuals contribute different times to the “rectangles” of the Lexis diagram. This is one reason why the basic data are typically collected in terms of rates, and a Poisson assumption is invoked. There is some comfort in the fact that if the assumptions (10.2) and (10.3) are complemented by the following assumptions: suppose the q_i 's depend on n and as $n \rightarrow \infty$, (i) $nq = \lambda > 0$, and (ii) $\max\{q_1, \dots, q_n\} \rightarrow 0$, then the distribution of X converges to $Po(\lambda)$ (Feller 1968, p. 282). The Poisson model is of interest, because under that model $E[X] = \lambda$ as before, but $\text{Var}(X) = \lambda > n(q-q^2)$. In other words, the Poisson model has a larger variability than the corresponding binomial models. Quantitatively the difference is small, however, since now $C = \lambda^{-1/2}$. If $n = 1000,000$ and $q = 0.01$, then $\lambda = 10,000$, and $C = 0.01$, for example. Or, the relative variability is the same as that under the homogeneous binomial model, to the degree of accuracy used.

The usual demographic application of the Poisson model proceeds from the further identification $\lambda = \mu K$, where K is the person years lived in the population, and μ is the force of mortality. The validity of this model is not self-evident, since unlike n , K is also random. At least when λ is of a smaller order of magnitude than K , the approximation appears to be good, however (Breslow and Day 1987, pp. 132–133). As is well-known, the maximum likelihood estimator of the force of mortality is $\hat{\mu} = X/K$ with the estimated standard error of $X^{1/2}/K$. Extensions to log-linear models that allow for the incorporation of explanatory variables follow similarly.

Since the Poisson distribution has the variance maximizing property, and it provides a model for both the independent trials and occurrence/exposure rates, we will below restrict the discussion primarily to the Poisson case.

10.3 Random Rates

Since (10.1) is not the cause of the low level of variability in the number of deaths, we need to look more closely at (10.2). A simple (but unrealistic) example showing that there are many opportunities here is the following. Suppose we make a single random experiment with probability of success $= q$, and probability of failure $= 1 - q$. If the experiment succeeds, define $X_i = 1$ for all i . Otherwise define $X_i = 0$ for all i . In this case we have, for example, that $X = nX_1$, so $E[X] = nq$ as before, but $\text{Var}(X) = n^2(q-q^2)$, and $C = ((1-q)/q)^{1/2}$ independently of n . For $q = 0.01$ we have $C = 9.95$, for example, indicating a huge (nearly 1000%) level of variability.

More realistically, we may think that dependence across individuals arises because they may all be influenced by common factors to some extent, at least. For example, there may be year to year variation in mortality around a mean that is

due to irregular trends in economics, epidemics, weather etc. If the interest would center on a given year, the model might still be $X \sim \text{Po}(\mu K)$, but if several years are considered jointly, then the year-to-year variation due to such factors would have to be considered. In this case, we would entertain a hierarchical model of the type

$$X \sim \text{Po}(\mu K) \text{ with } E[\mu] = \mu_0, \text{Var}(\mu) = \sigma^2. \quad (10.6)$$

In other words, the rate μ itself is being considered random, with a mean μ_0 that reflects the average level of mortality over the (relatively short) period of interest, and variance σ^2 that describes the year to year variation.

In this case we have that

$$\begin{aligned} \text{Var}(X) &= E[\text{Var}(X|\mu)] + \text{Var}(E[X|\mu]) \\ &= \mu_0 K + \sigma^2 K^2. \end{aligned} \quad (10.7)$$

It follows that $\text{Var}(\hat{\mu}) = \sigma^2 + \mu_0/K$. This result is of fundamental interest in demography, because if K is large, then the dominant part of the error is due to the annual variability. If the interest centers (as in the production of official population statistics) on a given year, with no regard to other years, we would be left with the pure Poisson variance μ_0/K , which is often small. An exception is the oldest-old mortality, where Poisson variation is always an issue, because for ages high enough K will always be small and μ_0 large.

However, when the interest centers on the time trends of mortality, and eventually on forecasting its future values, then the year to year variation σ^2 must be considered. Under model (10.6) this is independent of population size K . This is a realistic first approximation, but we note that model (10.6) does not take into account the possibility that a population might consist of relatively independent subpopulations. In that case, populations having many such subpopulations would have a smaller variance than a population with no independent subpopulations.

10.4 Handling of Trends

Consider now two counts. Or assume that for $i = 1, 2$, we have that

$$X_i \sim \text{Po}(\mu_i K_i) \text{ with } E[\mu_i] = \mu_{0i}, \text{Var}(\mu_i) = \sigma_i^2, \text{Corr}(\mu_1, \mu_2) = \rho. \quad (10.8)$$

Repeating the argument leading to (10.7) for covariances yields the result

$$\text{Corr}(X_1, X_2) = \rho / \{(1 + \mu_{01}/\sigma_1 K_1)(1 + \mu_{02}/\sigma_2 K_2)\}^{1/2}. \quad (10.9)$$

Or, the effect of Poisson variability is to decrease the correlation between the observed rates. We note that if the K_i 's are large, the attenuation is small. However, for the oldest old the K_i 's are eventually small, and the μ_{0i} 's large, so attenuation is expected.

In concentrating on Poisson variability that is primarily of interest in the assessment of the accuracy of vital registration, demographers have viewed annual variation as something to be explained. Annual changes in mortality and fertility are analyzed by decomposing the population into ever finer subgroups in an effort to try to find out, which are the groups most responsible for the observed change. Often partial explanations can be found in this manner, but they rarely provide a basis for anticipating future changes (Keyfitz 1982). To be of value in forecasting, an explanation must have certain robustness against idiosyncratic conditions, and persistence over time. This leads to considering changes around a trend as random.

One cause for why some demographers find statistical analyses of demographic time-series irritating seems to lie here: what a demographer views as a phenomenon of considerable analytical interest, may seem to a statistician as a mere random phenomenon, sufficiently described once σ^2 is known. [This tension has counterparts in many parts of science. Linguists, for example, differ in whether they study the fine details of specific dialects, or whether they try to see general patterns underlying many languages.]

In forecasting, the situation is more complex than outlined so far. In mortality forecasting one would typically be interested taking account of the nearly universal decline in mortality, by making further assumptions about time trends. For example, suppose the count at time t is of the form $X_t \sim \text{Po}(\mu_t K_t)$, such that

$$\log(\mu_t) = \alpha + \beta t + \xi_t, \text{ where } E[\xi_t] = 0, \text{Cov}(\xi_t, \xi_s) = \sigma^2 \min\{t, s\}. \quad (10.10)$$

If the ξ_t 's have normal distributions, under (10.10) we would have that $E[\mu_t] = \exp(\alpha + \beta t + \sigma^2 t/2) \equiv \mu_{0t}$ (This model is closely related to the so-called Lee-Carter model.)

One reason that makes (10.10) more complicated than (10.8), is that μ_{0t} involves parameters to be estimated, so standard errors become an issue. Especially, if $\text{Var}(\hat{\beta})$ is large, this source of error may have a considerable effect for large t , because it induces a quadratic term into the variance of error, whereas the effect of the random walk via σ^2 is only linear.

The way $\text{Var}(\hat{\beta})$ is usually estimated from past data assumes that the model specified in (10.10) is correct. Therefore, probabilistic analyses based on (10.10) are conditional on the chosen model. What these probabilities do not formally cover is the uncertainty in model choice itself (Chatfield 1996).

One should pay attention to model choice because it is typically based on iteration, in which lack of fit is balanced against parametric parsimony (*cf.*, Box and Jenkins 1976). One would expect error estimates calculated after a selection process to be too small, because of potential overfitting. Yet, a curious empirical fact seems to be that statistical time-series models identified and estimated in this

manner, for example demographic time-series, often produce prediction intervals that are rather wide, and even implausibly wide in the sense that in a matter of decades they may include values that are thought to be biologically implausible.

A possible explanation is that the standard time-series models belong to simple classes of models (*e.g.*, (10.10) can be seen as belonging to models with polynomial trends with once integrated, or I (10.1), errors) and the identification procedures used are tilted in favor of simple models within those classes. This shows that although judgment is certainly exercised in model choice, it can be exercised in a relatively open manner that tends to produce models that are too simple rather than too complex. When such models are estimated from the data, part of the lack of fit is due to modeling error. Therefore, the estimated models can actually incorporate some aspects of modeling error.

Modeling error can sometimes be handled probabilistically by considering alternative models within a larger class of models, and by weighting the results according to the credibility of the alternatives (*e.g.*, Draper 1995). Alho and Spencer (1985) discuss some minimax type alternatives in a demographic context. A simpler approach is to use models that are not optimized to provide the best possible fit obtainable. In that case the residual error may capture some of the modeling error, as well.

10.5 On Judgment and Subjectivity in Statistical Modeling

“One cannot be slightly pregnant”. In analogy, it is sometimes inferred from this dictum that if judgment is exercised in some part of a forecasting exercise, then all probabilistic aspects of the forecast are necessarily judgmental in nature. In addition, since judgment always involves subjective elements, then the probabilities are also purely subjective. I believe these analogies are misleading in that they fail to appreciate the many layers of probabilities one must consider.

First, the assumption of binomial or Poisson type randomness is the basis of grouped mortality analyses, and as such implicitly shared by essentially all demographers. It takes some talent to see how such randomness could be viewed as subjective.

Second, although models of random rates are not currently used in descriptive demography, they are implicit in all analyses of trends in mortality. Such analyses use smooth models for trends, and deviations from trends are viewed as random. The validity of alternative models can be tested against empirical data and subjective preferences have relatively little role.

On the other hand, models used in forecasting are different in that they are thought to hold in the future, as well as in the past. Yet, they can only be tested against the past. However, even here, there are different grades. In short term forecasting (say, 1–5 years ahead), we have plenty of empirical data on the

performance of the competing models in forecasting. Hence, there is an empirical and fairly formal basis for the choice of models. In medium term forecasting (say, 10–20 years ahead), empirical data are much more scant, and alternative data sets produce conflicting results of forecast performance. Judgment becomes an important ingredient in forecasting. In long-term forecasting (say, 30+ years ahead), the probabilities calculated based on any statistical model begin to be dominated by the possibility of modeling error and beliefs concerning new factors whose influence has not manifested itself in the past data. Judgment, and subjective elements that cannot be empirically checked, get an important role. Note that the binomial/Poisson variability, and the annual variability of the rates, still exist, but they have become dominated by other uncertainties.

In short, instead of viewing probabilities in forecasting as a black and white subjective/objective dichotomy suggested by the “pregnancy dictum”, we have a gradation of shades of gray.

10.6 On the Interpretation of Probabilities

A remaining issue is how one might interpret the probabilities of various types. Philosophically, the problem has been much studied (*e.g.*, Kyburg 1970). It is well-known that the so-called frequency interpretation of probabilities is not a logically sound basis for defining the concept of probability. (For example, laws of large numbers presume the existence of the concept of probability for their statement and proof.) However, it does provide a useful interpretation that serves as a basis of the empirical validation of statistical models we have discussed from binomial/Poisson variation to short and even medium term forecasting. For long term forecasting it is less useful, since we are not interested in what might happen if the history were to be repeated probabilistically again and again. We only experience one sample path.

It is equally well-known that there is a logically coherent theory of subjective probabilities that underlies much of the Bayesian approach to statistics. This theory is rather more subtle than is often appreciated. As discussed by Savage (1954), for example, the theory is prescriptive in the sense that a completely rational actor would behave according to its rules. Since mere mortals are rarely, if ever, completely rational, the representation of actual beliefs in terms of subjective probabilities is a non-trivial task.

For example, actual humans rarely contemplate all possible events that might logically occur. If a person is asked about three events, he might first say that A is three times as likely as B, and B is five times as likely as C; but later say that A is ten times as likely as C. Of course, when confronted with the intransitivity of the answers, he could correct them in any number of ways, but it is not clear that the likelihood of any given event would, after adjustment, be more trustworthy than before.

Actual humans are also much less precise as “computing machines” than the idealized rational actors. Suppose, for example, that a person P says that his uncertainty about the life expectancy in Sweden in the year 2050 can be represented by a normal distribution $N(100, 8^2)$. One can then imagine the following dialogue with a questioner Q:

- Q: So you think the probability that life expectancy exceeds 110 years is over 10%?
 P: I don’t know if I think that. If you say so.
 Q: Why can’t you say for sure?
 P: Because I can’t recall the 0.9 quantile of the standard normal distribution right now.

(Upon learning that it is 1.2816, and calculating $100 + 1.2816 \cdot 8 = 110.25$ P then agrees.)

Both difficulties suggest that any “subjective” probability statements need to be understood in an idealized sense. To be taken seriously, a person can hardly claim that he or she “feels” that some probabilities apply. Instead, careful argumentation is needed, if one were to want to persuade others to share the same probabilistic characterization (Why a mean of 100 years? Why a standard deviation of 8 years? Why a normal distributional form?).

10.7 Eliciting Expert Views on Uncertainty

Particular problems in the elicitation of probabilistic statements from “experts” are caused by the very fact that an expert is a person who should know how things are.

First, representing one’s uncertainty truthfully may be tantamount to saying that one does not really know, if what he or she is saying is accurate. A client paying consulting fees may then deduce that the person is not really an expert! Thus, there is an incentive for the expert to downplay his or her uncertainty.

Second, experts typically share a common information basis, so presenting views that run counter to what other experts say, may label the expert as an eccentric, whose views cannot be trusted. This leads to expert flocking: an expert does not want to present a view that is far from what his or her colleagues say. An example (pointed out by Nico Keilman) is Henshels (1982, p. 71) assessment of the U.S. population forecasts of the 1930s and 1940s. The forecasts came too high because according to Henshel the experts talked too much to each other. Therefore, a consideration of the range of expert opinions may not give a reasonable estimate of the uncertainty of expert judgment.

Economic forecasting provides continuing evidence of both phenomena. First, one only has to think of stock market analysts making erroneous predictions with great confidence on prime time TV, week after week. Second, one can think of think-tanks making forecasts of the GDP. Often (as in the case of Finland in 2001), all tend to err on the same side.

Of course, an expert can also learn to exaggerate uncertainty, should it become professionally acceptable. However, although exaggeration is less serious than the underestimation of uncertainty, it is not harmless either, since it may discredit more reasonable views.

A third, and much more practical problem in the elicitation of probabilities from experts stems from the issues discussed in Sect. 10.7. It is difficult, even for a trained person, to express one's views with the mathematical precision that is needed. One approach that is commonly used is to translate probabilities into betting language. (These concepts are commonly used in the Anglo-Saxon world, but less so in the Nordic countries, for example.) If a player thinks that the probability is at least p that a certain event A happens, then it would be rational to accept a p : $(1-p)$ bet that A happens. (I.e., if A does not happen, the player must pay p , but if it does happen, he or she will receive $1-p$. If the player thinks the true probability of A occurring is $\rho \geq p$, then the expected outcome of the game is $\rho(1-p) - (1-p)p = \rho - p \geq 0$.)

This approach has two problems, when applied in the elicitation of probabilities from experts. First, it is sometimes difficult to convince the experts to take the notion of a gamble seriously when they know that the "game" does not really take place. Even if the experts are given actual money with which to gamble, the amount may have an effect on the outcome. The second problem, in its standard form, the gambling argument assumes that the players are risk neutral. This may only be approximately true if the sums involved are small. If the sums are large, people tend to be more risk adverse (Arrow 1971). Moreover, experimental evidence suggests (e.g., Kahneman et al. 1982) that people frequently violate the principle of maximizing expected utility.

The betting approach has been used in Finland to elicit expert views on migration (Alho 1998). In an effort to anchor the elicitation on something empirical, a preliminary time series model was estimated, and the experts were asked about the probability content of the model based prediction intervals for future migration. Experts had previously emphasized the essential unpredictability of migration, but seeing the intervals they felt that future migration is not as uncertain as indicated. The intervals were then narrowed down using the betting argument. In this case the use of an empirical bench mark may have lead to a higher level of uncertainty than would otherwise have been obtained.

References

- Alho, J. M. (1998). A stochastic forecast of the population of Finland. Reviews 1998/4. Statistics Finland.
- Alho, J. M., & Spencer, B. D. (1985). Uncertain population forecasting. *Journal of the American Statistical Association*, 80, 306–314.
- Arrow, K. J. (1971). *Essays in the theory of risk-bearing*. Chicago: Markham Publishing Co.
- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis* (Rev ed.). San Francisco: Holden-Day.

- Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research* (Vol. II). Lyon: I.A.R.C.
- Chatfield, C. (1996). *The analysis of time series*. London: Chapman and Hall.
- Chiang, C. L. (1968). *Introduction to stochastic processes in biostatistics*. New York: Wiley.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society. Series B*, 57, 45–97.
- Feller, W. (1968). *An introduction to probability and its applications* (Vol. I, 3rd ed.). New York: Wiley.
- Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics*. New York: Norton.
- Henshel, R. L. (1982). Sociology and social forecasting. *Annual Review of Sociology*, 8, 57–79.
- Hoem, J. (1970). *Grunnbegreper i formell befolkningslære*. Oslo: Universitetsforlaget.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Keyfitz, N. (1982). Can knowledge improve forecasts? *Population and Development Review*, 8, 719–751.
- Kyburg, H. E. (1970). *Probability and inductive logic*. London: Macmillan.
- Lee, R. D., & Auerbach, A. J. (2001). *Demographic change and fiscal policy*. Cambridge: Cambridge University Press.
- Pollard, J. H. (1968). A note on multi type Galton-Watson processes with random branching probabilities. *Biometrika*, 55, 589–590.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Shryock, H. S., & Siegel, J. S. (1976). *The methods and materials of demography*. New York: Academic.
- Smith, J. P. (2001). Comment, in Lee and Auerbach, pp. 69–72.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.