The Variance of the Number of Effects in an Epidemiological Cohort - The Role of Dose Uncertainty

Guthrie Miller*

Los Alamos National Laboratory, 1619 Central Avenue, MS A117, Los Alamos, NM 87545, USA

Abstract: Two basic formulas, for the mean and variance of the number of effects in an epidemiological cohort, are derived. The formula for variance shows "extra-binomial variation" or "overdispersion" when there is correlated uncertainty of the probability of an effect. The formulas were validated by a numerical Monte Carlo study. The method of including "epistemic" uncertainty discussed by Hofer (E. Hofer, Health Physics, 2007) is generalized to include separately uncertainty from a Bayesian posterior distribution when the prior is known, and uncertainty of the prior.

1. INTRODUCTION

In this note, two basic formulas, for the mean and variance of the number of effects in an epidemiological cohort, are derived. It is conventionally assumed that the number of effects has either a Poisson or binomial distribution [1]. The formula for the variance given here reduces to the binomial result in the case of no correlations of the probability of an effect, but shows "extra-binomial variation" or "overdispersion" when there are correlations.

In practice, the variance could be calculated using the method advocated by Hofer [2], where, for each individual in an epidemiological cohort, some number $j = 1 \ldots M$ of alternate realizations of the dose and hence the probability of an effect, taking into account possible correlations, are generated using Monte Carlo. This method is generalized here to include separately uncertainty from a Bayesian posterior distribution when the prior is known, and uncertainties caused by lack of knowledge of the prior. This is because, within a linear dose-effect response model, the average number of effects is proportional to the posterior-average- collective dose, and the important uncertainty is that of the posterior-average-collective dose caused by lack of knowledge of the prior.

2. EFFECT OF DOSE UNCERTAINTIES ON EPIDEMIOLOGY

Individuals in the cohort are denoted by $i = 1, N$. The effect of interest is a function e_i, taking the values 1—an effect is observed, or 0—no effect is observed. The total number of effects, denoted by n, is given by

$$n = \sum_{i=1}^{N} e_i \,, \tag{1}$$

The average number of effects for person i is obtained by performing (conceptually) a large number $t = 1, T$ of identical trials, and is, by the frequency definition of probability, the probability of an effect for person i.

$$\langle e_i \rangle = \frac{1}{T} \sum_{t=1}^{T} e_{it} \equiv p_i \,. \tag{2}$$

Now, it is assumed that p_i itself is uncertain, through it's dependence on uncertain parameters, for example radiation dose. The average related to this uncertainty is denoted by an average over $j = 1, M$, where M is a large number of trials where the probabilities are variable, using a notation similar to that of Hofer [2], who discusses this approach to "epistemic" uncertainty. Thus, one thinks of the total number of independent trials $T \times M$ as being factored into two pieces, T trials where the probabilities are constant and M trials were the probabilities are variable.

The mean number of effects in the cohort is then given by

$$\langle n \rangle = \frac{1}{T\,M} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} e_{itj} = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} \equiv \sum_{i}^{N} \bar{p}_i \,, \tag{3}$$

where

$$\bar{p}_i \equiv \frac{1}{M} \sum_{j=1}^{M} p_{ij} \,, \tag{4}$$

defining a general notation for an average over epistemic uncertainty. Equation (3) is the first of the formulas sought.

The variance of n is defined as

$$Var(n) \equiv \left\langle \left(n - \langle n \rangle \right)^2 \right\rangle = \left\langle n^2 \right\rangle - \langle n \rangle^2 \,, \tag{5}$$

and requires the calculation of the mean of n^2. Returning to the definition,

$$n^2 = \sum_{i,i'} e_i e_{i'} = \sum_{i} e_i + 2 \sum_{i>i'} e_i e_{i'} \,, \tag{6}$$

because $e_i^2 = e_i$. Then,

$$\left\langle n^2 \right\rangle = \sum_{i} p_i + 2 \sum_{i>i'} p_i p_{i'} \,. \tag{7}$$

The second summation term in Eq. (7) follows from the second summation term in Eq. (6) because, for fixed prob-

*Address correspondence to this author at the Los Alamos National Laboratory, 1619 Central Avenue, MS A117, Los Alamos, NM 87545, USA;
Tel: +1 505 667 5547; Fax: +1 505 665 2052; E-mail: guthrie@lanl.gov

1874-2971/08

abilities, the effects in different persons are independent. This is the same as saying that for a large number of trials, the fraction of all trials that have effects for person k that also have effects for person i is p_i. Then,

$$\left\langle n^2 \right\rangle = \sum_i^N \overline{p}_i + 2\sum_{i>i'} \frac{1}{M}\sum_j p_{ij} p_{i'j} \,. \tag{8}$$

By elementary algebra,

$$Var(n) = \left\langle n^2 \right\rangle - \langle n \rangle^2 = \sum_i \overline{p}_i(1-\overline{p}_i) + 2\sum_{i>i'} \frac{1}{M}\sum_j (p_{ij}-\overline{p}_i)(p_{i'j}-\overline{p}_{i'}) \,. \tag{9}$$

Equation (9) can also be written in terms of the covariance of uncertainty between individuals i and i' defined by

$$C_{i,i'} = \frac{\frac{1}{M}\sum_j \left(p_{ij}-\overline{p}_i\right)\left(p_{i'j}-\overline{p}_{i'}\right)}{\sigma_{p_i}\sigma_{p_{i'}}}$$

$$\sigma_{p_i} = \sqrt{\frac{1}{M}\sum_j \left(p_{ij}-\overline{p}_i\right)^2} \,. \tag{10}$$

In terms of the correlation matrix,

$$Var(n) = \sum_i \overline{p}_i(1-\overline{p}_i) + 2\sum_{i>i'} C_{i,i'}\sigma_{p_i}\sigma_{p_{i'}} \,, \tag{11}$$

which is the second of the formulas sought. One can show algebraically that $|C_{i,k}| \leq 1$. Equation (11) shows that the usual expression for the variance for a sum of binomial distributions is modified by the addition of the covariance term. When the average probability for all individuals in the cohort is the same (as in the numerical examples that follow), and the probabilities are uncorrelated ($C_{i,i'} = 0$),

$$Var(n) = N\left(\overline{p}(1-\overline{p})\right). \tag{12}$$

When the average probability and probability uncertainty for all individuals in the cohort is the same and the probabilities are completely correlated ($C_{i,i} = 1$),

$$Var(n) = N\left(\overline{p}(1-\overline{p})\right) + N(N-1)\sigma_p^2 \,. \tag{13}$$

3. NUMERICAL VALIDATION

For this Monte Carlo study it is assumed there are 30 individuals in the cohort. The probability of an effect is 0.1 plus a lognormally distributed term with mean 0.05 and standard deviation 0.05 (coefficient of variation = 1, standard deviation of logs = 0.833). Effects are generated by Monte Carlo starting with the probability of an effect. The variable probability term is either considered to be uncorrelated from individual to individual or completely correlated (all individuals have the same probability, variable from realization to realization). Table **1** shows results of averages over a large number of trials (10^7) for the two terms on the right-hand-side of Eq. (11). The column labeled "ratio" is the ratio of the numerically observed variance to that given by Eq. (11).

Table 1. Numerical Test of Formula for Variance in the Number of Effects

	Var(n)	Ratio	$\langle n \rangle$	Var-Binomial	Covariance
uncorrelated	3.825	1.0009	4.5	3.825	$1.7\ 10^{-5}$
correlated	6.000	0.9998	4.5	3.825	2.175

The probability of an effect is 0.1 plus a lognormally distributed term with mean 0.05 and standard deviation 0.05 (coefficient of variation = 1). The column labeled "ratio" shows the ratio of the numerically observed variance to that given by the formula, which consists of the usual binomial variance plus a covariance term shown in the right-hand-side two columns.

4. APPLICATION TO EPIDEMIOLOGY-LINEAR DOSE-RESPONSE RELATIONSHIP

A linear model is assumed for p, namely

$$p_i = \alpha + \beta\xi_i \,, \tag{14}$$

where ξ_i is the true value of some measurable dose quantity (e.g. radiation dose) suspected of causing the effect, α is the known background probability of occurrence of the effect, and β is the dose-effect coefficient. Equation (3) then becomes approximately

$$n \cong \langle n \rangle = N\alpha + \beta\sum_{i=1}^N \overline{\xi}_i \,. \tag{15}$$

One can then solve for β, obtaining

$$\beta \cong \frac{n - N\alpha}{\Sigma} \,, \tag{16}$$

where n is the observed number of effects, and Σ is the average collective dose for the cohort given by

$$\Sigma = \sum_{i=1}^N \overline{\xi}_i = \sum_{i=1}^N \frac{1}{M}\sum_{j=1}^M \xi_{ij} \,, \tag{17}$$

Because of the assumed linear dose-response relationship, Eq. (16) involves only the cohort average collective dose, which is the sum of average doses for the N individuals. Equation (16) is an important relationship constraining β and its uncertainty.

Using linearized uncertainty analysis of Eq. (16), the coefficient of variation of β is given by

$$C_V(\beta) \equiv \frac{\sqrt{Var(\beta)}}{\beta} = \sqrt{\frac{Var(n)}{(\beta\Sigma)^2} + \left(C_V(\Sigma)\right)^2} \,, \tag{18}$$

assuming no variation of α and no correlation between variations of the numerator and the denominator. The first assumption is purely to avoid inessential complexity and may be reasonable in some cases. The last assumption will be discussed subsequently.

At this point the average over j representing epistemic uncertainty needs to be considered more carefully. Using a Bayesian analysis, this average would represent averaging over the posterior distribution when the prior probability

distribution is correct, meaning that the prior accurately gives the distribution of true values of the parameters of interest in the population before any measurements. However, the prior probability distributions are rarely known with certainty. Thus, instead of averaging over j, two other averages are introduced: averaging over $l = 1,\ldots L$ equally likely prior probability distributions and averaging over $k = 1,\ldots K$ realizations of the posterior distribution, given prior l. In the formula for $Var(n)$ given by Eq. (9), the averaging over j is replaced by averaging over k, for $l = l^*$, where l^* is the "correct" prior probability distribution. Because the correct prior is not known, it is reasonable to approximate this using an average over all values of l instead,

$$Var(n) = \langle n^2 \rangle - \langle n \rangle^2 = 2\frac{1}{L}\sum_{l=1}^{L}\sum_{i>i'}\frac{1}{K}\sum_{k=1}^{K} p_{ikl}p_{i'kl} + \langle n \rangle - \langle n \rangle^2$$
$$\langle n \rangle = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K} p_{ikl} \quad . \quad (19)$$

In the formula for the collective dose Σ, however, the sum over j is replaced by a sum over k, and the posterior mean collective dose becomes a function of the choice of prior l.

$$\Sigma_l \equiv \sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\xi_{ikl} \ , \quad (20)$$

In Eq. (16) we make the identification

$$\Sigma \to \langle \Sigma \rangle \equiv \frac{1}{L}\sum_{l=1}^{L}\Sigma_l \ , \quad (21)$$

using the average over different priors as the best estimate of collective mean dose. At this point the factorization $M = K \times L$ makes no difference. However, the variance of the collective dose is given by

$$Var(\Sigma) = \frac{1}{L(L-1)}\sum_{l=1}^{L}\left(\Sigma_l - \langle \Sigma \rangle\right)^2 = \frac{1}{L-1}\left(\langle \Sigma_l^2 \rangle - \langle \Sigma \rangle^2\right). \quad (22)$$

Because the numerator in equation (16) may be imagined to be evaluated at $l = l^*$, and the denominator is only a function of l, there is no correlation between the numerator and the denominator.

5. NUMERICAL EXAMPLE RELEVANT TO EPIDEMIOLOGY--LINEAR DOSE-RESPONSE RELATIONSHIP

This hypothetical example is similar to the example discussed in the Appendix of Hofer [2]. The effect is thyroid cancer in persons living near a nuclear reactor possibly caused by releases of radioactive iodine from the reactor. As done by Hofer, the radiation dose is calculated for a cohort of $N = 3000$ individuals. The true value of the radiation dose to individual i is denoted by ξ_i and is imagined to be obtained from a Bayesian calculation. The Bayesian method gives the joint posterior probability that each ξ_i is in some interval $d\xi_i$ given the data and the assumed prior, and this joint posterior distribution can be approximately represented by a finite discrete sample,

$$P(\{\xi_i\} \mid data, l)\prod_{i=1}^{N} d\xi_i \approx \xi_{ikl} \ , \quad (23)$$

where *data* represents any measurement data that may be used, and l enumerates the assumed prior probability distribution. As suggested by Hofer, it is convenient to replace the 3000 dimensional joint probability distribution with a finite sample consisting of some large number of values drawn from this joint probability distribution. In practice this sample would be obtained from forward Monte Carlo modeling from stack release at the reactor to inhalation or ingestion and impartation of dose to the individuals. Many different probability distributions would be incorporated into this modeling calculation, from variability and uncertainty of the release, to atmospheric dispersion, to variability and uncertainty of inhalation or ingestion, etc. In the language of Bayesian statistics, the probability distribution of dose so obtained would be termed a prior probability distribution, and one that involves quite a lot of expert judgment. If there are measurements, for example, of the amount of iodine in a person's thyroid, this prior would be multiplied by the measurement likelihood function to obtain the updated posterior using Bayes theorem. If the measurement data are weak or nonexistent, the posterior is the same as the prior. The suggestion by Hofer is that the joint posterior distribution of doses to the individuals in the cohort be characterized by M samples drawn from it, where Hofer suggests that $M = 100$ might be sufficient. It is suggested here that this method be used with $M = K \times L$, where $k = 1 \ldots K$ samples are drawn from the posterior assuming prior l and $l = 1 \ldots L$ different priors are considered. The reason for this is that, from Eq. (15), the number of effects is proportional to the average collective dose for the cohort, irrespective of the breadth of the posterior distribution. The uncertainty important for epidemiology in this context is the uncertainty of the posterior-average-collective dose caused by lack of knowledge of the "true" prior, given by Eq. (20).

Let us assume that the background probability of thyroid cancer is 0.005 (the average of Hofer's 0.003 for men and 0.007 for women) and $\beta = 0.02$ Gy^{-1} (instead of Hofer's 0.0169 Gy^{-1}). The true dose to the cohort is assumed to be a lognormal with mean value 1 Gy and coefficient of variation of 1. Two possibilities are considered: that the doses to individuals are perfectly correlated (for example, the variability and uncertainty of the amount released from the reactor dominates all other uncertainty), or that the doses to individuals are completely uncorrelated (for example, variability and uncertainty of atmospheric dispersion and other factors are most important). The mean and variance of the number of effects for these two extreme cases are as given in Table **2**.

Let us assume that the effect of the uncertainty regarding the prior is to multiply the true dose by another lognormal distribution, with mean value 1 and coefficient of variation 1. In practice one can only determine the variations caused by different prior assumptions without ever knowing which prior is correct. One might optimistically hope that by averaging over all the different assumed priors obtained using expert judgment by independent experts, one would end up with something closer to the true prior.

Table 2. The Variance of the Number of Thyroid Cancers in a Cohort of 3000 Individuals Exposed to a Lognormally Distributed Radiation Dose with Mean 1 Gy and Standard Deviation 1 Gy

Mean Number of Effects $\langle n \rangle$	$3000 \times (0.005 + 0.02 Gy^{-1} \times 1 Gy) = 75$
Background Number of Effects	$3000 \times 0.005 = 15$
Poisson Variance	$3000 \times 0.025 = 75$
Binomial Variance—No Correlation	$3000 \times 0.025 \times (1 - 0.025) = 73.1$
Additional Variance—Complete Correlation	$3000 \times 2999 \times (0.02)^2 = 3600$

The Poisson variance is equal to the average number of effects and applies when the probabilities of effects are small and uncorrelated. The binomial variance applies when the doses are uncorrelated. An additional variance term must be added when the doses are correlated.

The dose assessment provides for each of the $i = 1,\ldots N$ individuals in the cohort the quantities ξ_{ikl}, which are a set of dose values containing $k = 1,\ldots K$ samples from the joint posterior distribution for prior l and $l = 1,\ldots L$ choices of prior. Table **3** shows the dependence on K and L of various quantities of interest in estimating the uncertainties of the dose-effect coefficient β using Eq. (18) evaluated using the quantities ξ_{ikl}. The variance of the number of effects may be calculated from ξ_{ikl} using Eq. (19) or (Table **3** values) by directly simulating effects starting with the probability of an effect from Eq. (14).

6. DISCUSSION

In a recent analysis of the Japanese A-bomb survivor data [3], which takes into account dosimetry uncertainties, "models are fitted by Poisson maximum likelihood", and there is no discussion of the effect of possible correlations of dosimetry uncertainties. As shown here, in the case of dose correlations, the conventional Poisson variance assumptions are completely overturned in the limit of large sample size, and the coefficient of variation of the number of effects approaches a constant rather than decreasing to zero.

The finite-sample calculational method proposed by Hofer [2] offers a practical way to handle dose correlations, as long as sample sizes large enough to guarantee convergence are used. For a linear dose-response relationship, the average number of effects is proportional to the posterior-average-collective dose (PACD), and the important uncertainty is that of the PACD caused by lack of knowledge of the prior. Therefore it makes sense to generalize Hofer's method to include separately uncertainty from a Bayesian posterior distribution when the prior is known, and uncertainties caused by lack of knowledge of the prior. An example of uncertainty of the PACD caused by lack of knowledge of the prior is given in a recent publication [4], where alternate calculations of PACD are done using different prior assumptions.

ACKNOWLEDGMENTS

The author gratefully acknowledges very stimulating and helpful conversations with I. Apostaoi, O. Hoffman, and D. Pawel. The author would also like to thank the referees for suggestions that were helpful in clarifying the presentation of this material.

Table 3. The Dependence on K and L of Various Quantities of Interest in Estimating the Uncertainties of the Dose-Effect Coefficient β in the Hypothetical Epidemiological Study

L	5	10	20	∞
K	15	30	60	∞
Σ	1373 Gy	4179 Gy	2971 Gy	3000 Gy
C_V **(Σ)**	0.219	0.232	0.139	0
Var(n)-uncorr	181	3446	1389	73
Var(n)-corr	931	10987	9177	3673
C_V **(β)-uncorr**	$\sqrt{181/(0.02\times 1373)^2 + (0.219)^2} = 0.537$	0.740	0.642	0.142
C_V **(β)-corr**	$\sqrt{931/(0.02\times 1373)^2 + (0.219)^2} = 1.133$	1.275	1.618	1.01

The quantity K is the number of samples from the joint posterior probability distribution for a single prior, and L is the number of choices of prior. The quantity Σ is the posterior-average-collective dose, C_V refers to the coefficient of variation (standard deviation divided by mean), n is the number of thyroid cancers, and the dose-effect coefficient β is assumed to be 0.02 Gy^{-1}. Two extreme situations are considered, where the doses to individuals in the cohort are completely uncorrelated (uncorr) and where they are completely correlated (corr).

The last column of Table **3** shows that the uncertainty of the dose-effect coefficient can vary from 14% to 100% depending on whether the doses are correlated or not. For the finite sample sizes shown, there is not yet convergence of the variance of the number of effects. One must, of course, remember the ad hoc nature of the dose uncertainty assumptions in this example, which is intended to be didactic rather than realistic.

REFERENCES

[1] Rothman K, Greenland S. Modern epidemiology. 2nd ed. Lippincott-Raven; 1998.

[2] Hofer E. Hypothesis testing, statistical power, and confidence limits in the presence of epistemic uncertainty. Health Phys 2007; 92(3): 225-35.

[3] Little MP, Hoel DG, Molitor J, Boice JD, Wakeford R, Muirhead CR. New Models for the evaluation of radiation-induced lifetime cancer risk. Radiat Res 2008; 169: 660-76.

[4] Miller G, Guilmette R, Bertelli L, Waters T, Romanov SA, Zaytseva YV. Uncertainties in Internal Doses Calculated for Mayak Workers—A Study of 63 Cases. Radiation Protection Dosimetry. 2008 [advance access published August 8, 2008] Available from: http://rpd.oxfordjournals.org/cgi/content/full/ncn181v1

Received: March 4, 2008 Revised: August 9, 2008 Accepted: August 14, 2008