Published in final edited form as:
Environ Ecol Stat. 2009 March 1; 16(1): 53–62. doi:10.1007/s10651-007-0076-2.

Confidence limits on one-stage model parameters in benchmark risk assessment

Brooke E. Buckley,
Department of Mathematics, Northern Kentucky University, Highland Heights, KY 41099, USA

Walter W. Piegorsch, and
Department of Mathematics, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA, e-mail: piegorsc@math.sc.edu

R. Webster West
Department of Statistics, 3143 TAMU, Texas A&M University, College Station, TX 77843, USA, e-mail: websterwest@yahoo.com

Abstract

In modern environmental risk analysis, inferences are often desired on those low dose levels at which a fixed benchmark risk is achieved. In this paper, we study the use of confidence limits on parameters from a simple one-stage model of risk historically popular in benchmark analysis with quantal data. Based on these confidence bounds, we present methods for deriving upper confidence limits on extra risk and lower bounds on the benchmark dose. The methods are seen to extend automatically to the case where simultaneous inferences are desired at multiple doses. Monte Carlo evaluations explore characteristics of the parameter estimates and the confidence limits under this setting.

Keywords

Benchmark dose; Bootstrap; Resampling; Environmental risk analysis; Quantal dose response; Quantitative risk assessment; Simultaneous inferences; Weibull; Dose-response model

1 Introduction: benchmark analysis under a one-stage model

A primary objective in environmental risk analysis is characterization of the severity and likelihood of damage caused by a hazardous agent (Coherssen and Covello 1989). Towards this end, experimental studies are often conducted on laboratory animals where exposure levels of the agent are administered at high doses. Estimation of risk at low doses must be based on this high-dose data, leading to an extrapolation. Of particular interest are inferences on the risk at a specific low dose(s) or inferences on the dose(s) at which a certain risk is achieved.

Within this context, we define risk as the probability that a subject exposed to a specified dose, d_i ($i = 1, \ldots, n$), of a hazardous agent will develop a particular adverse effect. We assume that the risk is a monotone increasing function of d, $R(d)$. At each d_i, the number of subjects exhibiting an adverse effect, Y_i, is recorded. This is commonly referred to as the *quantal response setting*. Many formulations are possible when modeling Y_i. Using what is perhaps the most common construction seen in practice, we assume that the Y_i s are independent

e-mail: buckleyb1@nku.edu.

binomial variates with parameters N_i and $R(d_i)$, where N_i is the number of subjects tested at dose d_i and $R(d_i)$ models the unknown probability that a subject will respond adversely.

To specify $R(d)$, there are a variety of models from which to choose. A popular form from toxicological risk assessment is a two-parameter, "single-stage" version of the well-known Armitage-Doll multistage model for adverse response (Armitage and Doll 1954):

$$R(d) = 1 - \exp\{-\beta_0 - \beta_1 d\}, \tag{1}$$

where we require $\beta_j \geq 0$, $j = 0, 1$, and, of course, $d \geq 0$. In this simple, two-parameter form, the multistage model is also often characterized as a *one-stage model*. An alternative form for $R(d)$ often considered in environmental risk assessment is the Weibull dose-response model (U.S. EPA 2000; Parham and Portier 2005). Here, the one-stage model is also a special case of the Abbott-adjusted Weibull form $R(d) = \theta_0 + (1-\theta_0)(1-\exp\{-\beta_1 d^{\beta_2}\})$, where at $\beta_2 = 1$ were cover (1) with $\beta_0 = -ln(1-\theta_0)$. Noting this, we focus our attention in this paper on the one-stage formulation of Eq. 1. Of course, however, many different models are also applied in benchmark risk assessment, including the multiparameter Weibull and multistage forms mentioned above. Our goal herein is to focus on the two-parameter model in (1) and illustrate how its simplicity can lead to useful inferences within the larger benchmark framework. For guidance on risk estimation and inference under a more-complex multistage form, we refer the reader to the works of Al-Saidy et al. (2003) and Nitcheva et al. (2005).

In practice, the *risk above background* is often employed for purposes of assessing and managing exposure risks. To quantify this, we use the *extra risk function*, defined as the risk above the background or control level after correcting for non-response in the unexposed population: $R_E(d) = \{R(d) - R(0)\}/\{1 - R(0)\}$. Clearly, under our one-stage model $R_E(d) = 1 - \exp\{-\beta_1 d\}$. Often, interest exists in estimating the extra risk and from this the particular dose level at which a certain benchmark risk (BMR) is achieved. This level is known as a *Benchmark Dose*, or BMD (Crump 1984). To find the BMD, one sets the given value of BMR equal to the extra risk, and finds the smallest positive solution (if it exists) to this relationship. For purposes of estimation, we employ maximum likelihood estimators (MLEs) and substitute the MLEs of any unknown parameters into the expression for BMD; we denote this ML point estimator as BMˆD.

In passing, we should also note that a simplified formulation of our one-stage structure can be constructed from a simple Taylor-series expansion of the extra risk function. Known as the *linearized multistage* (*LMS*) *model*, the construction in effect employs a first-order Taylor approximation to $R_E(d) = 1 - \exp\{-\beta_1 d\}$ as $d \to 0$, producing $R_E(d) \approx \beta_1 d$. The BMD is then approximated as BMD $\approx$ BMR/β_1, from which easy-to-construct point estimators may be developed. Since the nonlinear structure of the multistage model is not trivial, the simplicity afforded by this LMS approximation is a critical factor in support of its use. With the advent of modern, high-speed, computing technologies, however, it is no longer difficult to perform a model fit and construct pertinent inferences under the model in (1). Indeed, Nitcheva et al. (2005) found that use of the LMS approximation could produce very unstable inferences in selected instances, and cautioned against its use. Coupled with other concerns raised previously over the LMS approach (Lovell and Thomas 1996), these considerations suggest that the practical need for the LMS approximation is waning. As such, we will not study it in any formality.

Formal inferences on the extra risk and/or the BMD are available by manipulating the large-sample properties of the MLE. For instance, an environmental risk assessor would typically be interested in placing upper bounds on $R_E(d)$ at one or more dose levels, or in deriving lower

confidence bounds on the BMD at specified levels of risk, BMR. In the latter case, a lower bound on the benchmark dose is known as a Benchmark Dose Lower Limit, or BMDL (Crump 1995). Modern practice employs BMDLs as *points of departure* in quantitative risk assessment in order to arrive at acceptable levels of human or ecosystem exposure to the hazardous agent, or to otherwise establish practical low-exposure guidelines (Gaylor and Kodell 2002). As such, these quantities serve an important purpose within the larger realm of environmental risk management. Our focus herein concerns statistical inferences on quantities such as R_E or BMD, in order to more precisely refine these important points of departure for the environmental risk analyst. We focus on practical ways to construct 1 - α confidence bounds on $R_E(d)$ and, as a consequence, to find BMDLs. Section 2 gives more formal details on estimation and inferences for R_E under the two-parameter model in (1), while Sect. 3 addresses the computation of BMDs/BMDLs. Section 4 presents results from a short Monte Carlo simulation study on the small-sample features of the parameter estimates and the proposed confidence bounds.

2 Risk estimation

Under our one-stage model the MLEs, $\mathbf{b} = [b_0\ b_1]^T$, of the unknown parameters, $\boldsymbol{\beta} = [\beta_0\ \beta_1]^T$, are found by constrained optimization. The operations can be programmed in the software package **R** (R Development Core Team 2005) using its optim function, in SAS (SAS Institute Inc. 2000) via PROC NLMIXED, or using the U.S. EPA's Benchmark Dose Software (U.S. EPA 2001) (also see Falk Filipsson and Victorin 2003). Convergence is usually attained in 5-15 iterations. With these, the MLE of the extra risk is simply $\hat{R}_E(d) = 1 - \exp\{-b_1 d\}$.

When studying the extra risk function, only the detrimental extent of an adverse outcome is typically of subject-matter concern; this translates to interest in only upper confidence limits. Under the model in (1), we see that the extra risk is a monotone increasing function of β_1 so that bounding $R_E(d)$ simplifies to bounding β_1. Suppose a valid 100(1 - α)% upper limit on β_1, say b_u, satisfies $P[\beta_1 \le b_u] \approx 1 - \alpha$. Equivalently, since we assume $d \ge 0$, $P[\beta_1 d \le b_u d,\ \forall d \ge 0] \approx 1 - \alpha$. An approximate 100(1 - α)% upper bound on $R_E(d)$ is then

$$R_E(d) \le 1 - \exp\{-b_u d\} \tag{2}$$

In fact, since the operation leading to this upper bound is valid $\forall d \ge 0$, (2) represents a simultaneous 100(1 - α)% upper confidence *band* on $R_E(d)$.

Here, we study five methods for obtaining the upper limit, b_u, in (2). The first is a simple Wald-type upper bound based on appeal to the large sample normality of the MLE. In particular, Guess and Crump (1976) showed that **b** has an asymptotic normal distribution for our model when $\beta_j > 0$, for all j and with at least $n > 2$ dose levels. Thus, we can construct an asymptotic 1 - α upper confidence bound on β_1 as the Wald limit

$$b_{uW} = b_1 + z_\alpha se(b_1), \tag{3}$$

where b_1 is the MLE of β_1, $se(b_1)$ is its large-sample standard error, and z_α is an upper-α critical point from the standard normal distribution. For use in (2), simply substitute (3) for b_u to build Wald-type confidence bounds (and bands) on $R_E(d)$. This is essentially the approach suggested by Krewski and Van Ryzin (1981),Crump et al. (1977), and Crump and Howe (1985) for building confidence limits on functions such as $R_E(d)$ with the multistage model.

As part of a larger exposition on simultaneous confidence bands in multistage modeling, Al-Saidy et al. (2003) studied use of (3) for building confidence bounds on $R_E(d)$. They found that for large samples the method operated in a nominal fashion, but at smaller sample sizes

(such as $N = 25$ or sometimes $N = 50$) the coverage characteristics were somewhat variable, sometimes moving above nominal coverage levels and sometimes dropping below them. We will follow up on Al-Saidy et al.'s study in Sect. 4, below.

Our second method for finding an upper confidence limit on β_1 appeals to the asymptotic features of the likelihood ratio (LR) test. For model (1) and under a set of regularity conditions that can be shown to hold in most cases when this model is employed, the LR will possess large-sample χ^2 characteristics (Krewski and van Ryzin 1981) so that by inverting the LR test, one can derive approximate confidence bounds on the model parameters (Crump and Howe 1985;Bailer and Smith 1994). For our problem, we obtain the upper bound $b_{uLR}=\inf\left\{2\left(\ln L\left(\boldsymbol{b}\right)-\ln L\left(\boldsymbol{b}^*\right)\right)\geq\chi^2_{2\alpha}\left(1\right)\right\}$ over the set $\beta_1^*\geq 0$. Here, $\boldsymbol{b}^*=\left[b_0^*\beta_1^*\right]^T$ such that b_0^* is the maximum likelihood estimator of β_0 for some fixed value, β_1^*, of β_1 and L is the binomial likelihood function under model (1). For use in (2), simply substitute b_{uLR} for b_u to build LR confidence bounds (and bands) on $R_E(d)$.

Notice that the LR test is by nature two-sided; to obtain a one-sided upper bound, we apply an adjustment by doubling the significance level of the test and then ignoring the lower limit. Although admittedly ad hoc, this adjustment has been seen to exhibit reasonable operating characteristics for the sorts of risk-analytic calculations we study here (Crump and Howe 1985; Nitcheva et al. 2005). We will investigate the coverage characteristics of b_{uLR} as part of our Monte Carlo study in Sect. 4.

Our remaining methods for finding an upper confidence limit on β_1 employ bootstrap-based approaches, in the spirit of Crump and Howe (1985) and Bailer and Smith (1994). These authors noted that the small-sample stability of likelihood-based confidence limits could be in question. Bootstrap resampling provides a natural alternative for building confidence limits on pertinent model parameters such as β_1; the resampling process uses the observed data to generate pseudo-replicates of the experiment, from which pseudo-confidence limits may be derived based on percentiles of the bootstrap distribution (Dixon 2002).

We considered three bootstrap approaches, one fully parametric, the second fully non-parametric, and the third a mixture of the previous two. For the parametric bootstrap, the approach is straightforward: generate B independent pseudo-random samples from a binomial population with parameters N_i and $\hat{R}(d_i)$, where $\hat{R}(d_i)$ is the ML-estimated risk function from the observed data set. For each jth bootstrap data set, compute a new MLE for β_1, denoted as b_{1j}^*. This produces B bootstrap estimates $b_{11}^*, b_{12}^*, \ldots, b_{1B}^*$. The 100(1 - α)% upper confidence limit, b_{uPB}, is then taken as the 100(1 - α)th percentile of the B b_{1j}^*s. (We will investigate the coverage characteristics of b_{uPB} as part of our Monte Carlo study in Sect. 4.)

For the non-parametric and semi-parametric bootstraps, we start instead with the sample proportions, Y_i / N_i. Let $\left\{Y_{1j}^*, \ldots, Y_{nj}^*\right\}$ for $j = 1, \ldots, B$ denote a sequence of B independent bootstrap resamples taken with replacement from the observed data. Note that each Y_{ij}^* is a binomial pseudo-random variable with sample size parameter N_i and success probability Y_i / N_i. The fully non-parametric bootstrap then computes the maximum likelihood estimate of β_1 from each bootstrapped resample.

Now, note that if none or all of the responses at a particular dose level are adverse, i.e., if $Y_i = 0$ or $Y_i = N_i$, the observed proportions are exactly zero or one, respectively. In either case, there will be no variability in the non-parametric bootstrap resamples at this dose value. Bailer and Smith (1994) noted a similar concern with the non-parametric approach, and suggested that some correction was necessary to give the bootstrap results greater practical stability. Our

solution to this problem takes on a semi-parametric flavor: we operate in general under a non-parametric strategy, but in the special cases where $Y_i = 0$ or $Y_i = N_i$ we replace the sample proportion of adverse responses with the estimated risk $\hat{R}(d_i)$ from the model at that d_i. In either case (non-parametric or semi-parametric) we generate B independent bootstrap resamples and again obtain B bootstrap ML estimates of the unknown parameter β_1. The 100 (1 - α)% upper confidence limit, b_{uNB} or b_{uSB}, respectively, is then defined as the 100(1 -α)th percentile of these B statistics. As with the other methods we discuss above, we will study the operating characteristics of all the bootstrap upper bounds in Sect. 4, next.

In passing, it is important to note that none of our bootstrap strategies can guarantee a non-degenerate bootstrap sample when $d = 0$. For instance, under the parametric approach if the MLE of β_0 is 0, then a degenerate sample will always occur at $d = 0$. Likewise under the semi-parametric approach, if the MLE of β_0 is 0 and there are no adverse responses at $d = 0$, then a degenerate sample will always occur at $d = 0$.

3 Benchmark dose estimation

As introduced above, under the model in (1) the benchmark dose is the value of d that solves $R_E(d) = 1 - \exp\{-\beta_1 d\} = \text{BMR}$ at a given benchmark risk, BMR$\in$(0,1). To help clarify at which specific BMR this is determined, we use the notation $\text{BMD}_{100\text{BMR}}$, $\hat{\text{BMD}}_{100\text{BMR}}$, $\text{BMDL}_{100\text{BMR}}$, etc. Clearly, solving $1 - \exp\{-\beta_1 d\} = \text{BMR}$ for d gives

$$\text{BMD}_{100\text{BMR}} = \frac{-\ln(1 - \text{BMR})}{\beta_1}. \tag{4}$$

The MLE, $\hat{\text{BMD}}_{100\text{BMR}}$, is found by substituting b_1 for β_1 in the denominator of (4) and appealing to the ML invariance property (Casella and Berger 2002, Sect. 7.2).

To compute a BMDL, one simply mimics the BMD construction and inverts the upper confidence band on $R_E(d)$. That is, given the relationship $R_E(d) \leq 1 - \exp\{-b_u d\}$, where b_u satisfies $P[\beta_1 \leq b_u] \approx 1 - \alpha$, set BMR=1 - $\exp\{-b_u d$ and solve for d. The result is

$$\text{BMDL}_{100\text{BMR}} = \frac{-\ln(1 - \text{BMR})}{b_u}. \tag{5}$$

Any valid 1 - α upper limit b_u may be employed in (5), including the likelihood-based bounds b_{uW} or b_{uLR}, or the three bootstrap-based bounds from Sect. 2; see, e.g., Sand et al. (2002) for (pointwise) illustrations with the LR-based approach.

Notice that the operation leading to this BMDL is a one-to-one inversion of the upper simultaneous confidence band represented in (2). Hence, as in Sect. 2, (5) can be viewed as a simultaneous 100(1 - α)% lower confidence band on the BMD which varies as a function of BMR$\in$ (0,1). From this, various multiplicity-adjusted inferences may be derived (Al-Saidy et al. 2003).

4 Small sample performance

All of the methods described in Sects. 2-3 for finding upper limits on $R_E(d)$ or BMDLs are based on either asymptotic or bootstrap approximations. Hence in large samples we expect the simultaneous limits to contain the true value of $R_E(d)$ or the true BMD approximately 100(1 - α)% of the time. In small samples, however, their coverage characteristics may be less certain. To evaluate this, we undertook a Monte Carlo study of the small-sample simultaneous coverage

associated with each of the methods over a variety of one-stage quantal response models. Specifications for $\boldsymbol{\beta}$ in each model were taken from parameterizations used in previous studies of low-dose risk estimation described by Bailer and Smith (1994), Kodell and Park (1995), and Al-Saidy et al. (2003). The five parameterizations considered are given in Table 1.

As we have noted throughout, only β_1 appears in the expression for R_E and BMD and so the inferences under our two-parameter model will depend solely on the coverage quality of the confidence limit for β_1; i.e., the coverage for the simultaneous upper bound on R_E based on (2) and for the simultaneous BMDL based on (5) will be identical to that for the pointwise limit on β_1 and, in the latter case, will be independent of BMR. As such we display our results on empirical coverage only as a function of N. [In effect, while motivated from a risk-analytic perspective, our simulations study the small-sample quality of these various confidence limits for making inferences on β_1 under the choice of (1) to describe the "link function" in a binomial regression.]

We report results at $\alpha = 0.05$. Four dose levels, $d = 0, 0.25, 0.5, 1$, with equal numbers of subjects, $N_i = N$, per dose-group were used in the simulations, corresponding to a common design in cancer risk experimentation (Portier 1994). We selected values of N that ranged between 25 and 500. For each model configuration, 2,000 pseudo-binomial data sets were simulated, and the empirical simultaneous coverage of each method was computed. Notice then that the approximate standard error of the estimated coverage near $\alpha = 0.05$ is $\sqrt{(0.05)(0.95)/2000} \approx 0.005$, and it never exceeds $\sqrt{(0.5)(0.5)/2000} \approx 0.011$.

In order to gain a stronger understanding of the various procedures' operating characteristics, we also calculated the average separation the one-sided limits exhibited relative to the true values they were intended to bound: Separation = |Bound - True Value|. This measure was also employed by Nitcheva et al. (2005) in their Monte Carlo study of large-sample BMDLs for a more complex multistage risk model. We use the separation measure to represent a form of 'width' for the one-sided bounds when bounding extra risk: large positive differences suggest poor performance in that the bound is typically far from the true quantity of interest. Large negative differences are similar, except of course that any negative difference also corresponds to a coverage failure. (Viewed in terms of bounding BMD, the reverse is true.) In either case, however, values close to zero may be useful for regulatory purposes, since they indicate that the bound is close to the regulatory parameter being studied. Thus this separation measures summary performance: given two or more methods with similar coverage characteristics, those with smaller separations would be preferred for practical use. For each procedure, we also computed these separations in our simulations. For summary purposes, we report the median separation over each set of 2,000 simulated samples.

Results of our Monte Carlo calculations appear in Table 2. The empirical coverage rates displayed under each sample size in the table were computed by determining the number of times out of the 2,000 simulations that the upper confidence limit on β_1 was above the true value of β_1. In most cases, the coverage probabilities across methods are close to the nominal 95% level, at least within Monte Carlo sampling error. Indeed, Al-Saidy et al. (2003) obtained similar coverage probabilities when studying (only) the Wald approach with these five models. The most notable difference we observe relative to their results occurs in Model 3: we find coverage probabilities much closer to nominal than those reported by Al-Saidy et al. One possible explanation for this anomaly is that we have refined slightly the ML fitting algorithm that those authors used. Like Al-Saidy et al. we used the **R** system with its optim function for the constrained optimization, but for our initial estimates, we shrunk the observed proportions towards 1/2 by adding +2 to each Y_i and +4 to each N_i. This mimics a shrinkage estimator employed by Agresti and Coull (1998) for building confidence intervals on binomial parameters, and appears to stabilize extreme observations used in initializing the constrained

optimization. Model 3 tends to generate very small observed proportions, and the shrinkage may have some effect in these cases. (For other models that generate data not as extreme, the initial shrinkage does not appear to make a substantial difference.)

From Table 2, we also observe that the bootstrap approaches appear to exhibit slight instabilities in empirical coverage at small sample sizes with Models 1 and 3. These models are problematic in that their true β_1 is close to zero, which tends to generate many instances of $Y_i = 0$ across multiple doses. This response pattern is difficult to fit (under any method!) and apparently causes the bootstrap methods' coverages to drive too far above (Model 1) or below (Model 3) the nominal level. As we suggested above, and as Bailer and Smith (1994) also noted, bootstrap methods appear difficult to put into practice when many cases of $Y_i = 0$ are encountered in the data. For the remaining models, however, the bootstrap methods produced somewhat less conservative results.

Overall, Table 2 suggests that all our methods operate reasonably at large sample sizes, substantiating the asymptotic arguments underlying their use. In practice, however, sample sizes near $N_i = 25$ or $N_i = 50$ are more common, and in this case the LR method appears to operate with the greatest level of stability, at least among the five models we considered. Its measure of separation is much larger than that for the Wald method, however, and so with larger sample sizes we can recommend the latter for use along with the LR approach.

Acknowledgments

This research was initiated while all the authors were with the University of South Carolina. Thanks are due Drs. Obaid M. Al-Saidy, Ralph L. Kodell, and Daniela K. Nitcheva for their invaluable support to the project, and two external reviewers for their helpful comments. This work was funded under grant #R01-CA76031 from the U.S. National Cancer Institute, grant #RD-83241901 from the U.S. Environmental Protection Agency, and as part of the research arm of the U.S. Department of Homeland Security's Center of Excellence for the Study of Terrorism and Responses to Terrorism (START). Its contents are solely the responsibility of the authors and do not necessarily reflect the official views of these various agencies.

Author Biographies

Brooke E. Buckley received a Ph.D. degree in Statistics from the University of South Carolina. In the Fall of 2006, she joined the faculty in the Department of Mathematics at Northern Kentucky University as an Assistant Professor of Statistics. Her primary research interest is quantitative risk estimation, with a focus simultaneous inference under Abbott-adjusted models.

Walter W. Piegorsch is the Director of the Graduate Interdisciplinary Program in Statistics at the University of Arizona, Tucson, AZ, and a member of the Research Faculty of the University's BIO5 Institute. He studies modeling and analysis for environmental data, with emphasis on environmental hazards and risk assessment. He also has research interests in geo-spatially referenced disaster informatics, simultaneous inferences, generalized linear models, and the historical development of statistical thought as prompted by problems in the biological and environmental sciences. Among other activities, he has served on the Board of Scientific Counselors for the U.S. National Toxicology Program (2000-2004), on the Council of the International Biometric Society (2002-2005), and as Chairman of the American Statistical Association Section on Statistics & the Environment (2004). He earned his Ph.D. in Statistics at the Biometrics Unit, Cornell University, Ithaca, NY in 1984, after which he spent nine years as a practicing statistician with the U.S. National Institute of Environmental Health Sciences in Research Triangle Park, NC, and then 13 years as a faculty member with the Department of Statistics at the University of South Carolina.

R. Webster West is a Professor in the Department of Statistics at Texas A&M University. He received his Ph.D. in Statistics from rice University in 1994. Since that time, he has been actively developing new statistical methods for application to dose response models used in toxicology.

References

Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52:119–126.

Al-Saidy OM, Piegorsch WW, West RW, Nitcheva DK. Confidence bands for low-dose risk estimation with quantal response data. Biometrics 2003;59:1056–1062. [PubMed: 14969485]

Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954;8:1–12. [PubMed: 13172380]

Bailer AJ, Smith RJ. Estimating upper confidence limits for extra risk in quantal multistage models. Risk Anal 1994;14:1001–1010. [PubMed: 7846307]

Casella, G.; Berger, RL. Statistical inference. 2nd edn.. Pacific Grove; Duxbury: 2002.

Coherssen, JJ.; Covello, VT. Risk analysis: a guide to principles and methods for analyzing health and environmental risks. Executive Office of the President; Washington: 1989.

Crump KS. A new method for determining allowable daily intake. Fundam Appl Toxicol 1984;4:854–871. [PubMed: 6510615]

Crump KS. Calculation of benchmark doses from continuous data. Risk Anal 1995;15:79–89.

Crump KS, Guess HA, Deal KL. Confidence intervals and tests of hypotheses concerning dose response relations inferred from animal carcinogenicity data. Biometrics 1977;33:437–451. [PubMed: 911968]

Crump, KS.; Howe, R. A review of methods for calculating confidence limits in low dose extrapolation. In: Clayson, DB.; Krewski, D.; Munro, I., editors. Toxicological risk assessment, volume I: biological and statistical criteria. CRC Press; Boca Raton: 1985. p. 187-203.

Dixon, PM. Bootstrap resampling. In: El-Shaarawi, AH.; Piegorsch, WW., editors. Encyclopedia of environmetrics. Vol. 1. Wiley; Chichester: 2002. p. 212-220.

Falk Filipsson A, Victorin K. Comparison of available benchmark dose softwares and models using trichloroethylene as a model substance. Regul Toxicol Pharmacol 2003;37:343–355. [PubMed: 12758215]

Gaylor DW, Kodell RL. A procedure for developing risk-based reference doses. Regul Toxicol Pharmacol 2002;35:137–141. [PubMed: 12051999]

Guess HA, Crump KS. Low-dose extrapolation of data from animal carcinogenicity experiments—Analysis of a new statistical technique. Math Biosci 1976;32:15–36.

Kodell, RL.; Park, CN. Linear extrapolation in cancer risk assessment. In: Olin, S.; Farland, W.; Park, C.; Rhomberg, L.; Scheuplein, R.; Starr, T.; Wilson, J., editors. Low-dose extrapolation of cancer risks: issues and perspectives. ILSI Press; Washington: 1995. p. 87-104.

Krewski, D.; van Ryzin, J. Dose response models for quantal response toxicity data. In: Csörgö, M.; Dawson, DA.; Rao, JNK.; Saleh, AKME., editors. Statistics and related topics. North-Holland; Amsterdam: 1981. p. 201-231.

Lovell DP, Thomas G. Quantitative risk assessment and the limitations of the linearized multistage model. Hum Exp Toxicol 1996;15:87–110. [PubMed: 8645508]

Nitcheva DK, Piegorsch WW, West RW, Kodell RL. Multiplicity-adjusted inferences in risk assessment: Benchmark analysis with quantal response data. Biometrics 2005;61:277–286. [PubMed: 15737104]

Parham, F.; Portier, CJ. Benchmark dose approach. In: Edler, L.; Kitsos, C., editors. Recent advances in quantitative methods in cancer and human health risk assessment. Wiley; Chichester: 2005. p. 239-254.

Portier CJ. Biostatistical issues in the design and analysis of animal carcinogenicity experiments. Environ Health Perspect 1994;102(Suppl 1):5–8. [PubMed: 8187725]

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2005.

Sand S, Filipsson AF, Victorin K. Evaluation of the benchmark dose method for dichotomous data: model dependence and model selection. Regul Toxicol Pharmacol 2002;36:184–197. [PubMed: 12460753]

SAS Institute Inc.. SAS/STAT® user's guide, version 8. Vol. 1-3. SAS Institute Inc.; Cary: 2000.

U.S. EPA. Benchmark dose technical guidance document. U.S. Environmental Protection Agency; Washington, DC: 2000. External review draft number EPA/630/R-00/001

U.S. EPA. Help manual for benchmark dose software version 1.3. National Center for Environmental Assessment, U.S. Environmental Protection Agency; Research Triangle Park, NC: 2001. Technical report number EPA/600/R-00/014F

Table 1

Two-parameter one-stage models for the simulation study in Sect. 4 (Al-Saidy et al. 2003)

Model	β_0	β_1	$R(0)$	$R(1)$
1	0.05129	0.11123	0.05	0.15
2	0.05130	0.91630	0.05	0.62
3	0.01005	0.07333	0.01	0.08
4	0.10536	0.25131	0.10	0.30
5	0.35667	1.94592	0.30	0.90

Table 2

Empirical coverage rates and median separations for confidence intervals on β_1 under the multistage model $R(d) = 1 - \exp\{-\beta_0 - \beta_1 d\}$ (rates based on 2,000 simulated data sets, nominal $\alpha = 0.05$)

Model	Method[a]	Coverage					Median separation				
		N = 25	*N* = 50	*N* = 100	*N* = 300	*N* = 500	*N* = 25	*N* = 50	*N* = 100	*N* = 300	*N* = 500
1	Wald	0.9785	0.9345	0.9360	0.9415	0.9605	0.0789	0.0581	0.0425	0.0241	0.0180
	LR	0.9460	0.9410	0.9470	0.9450	0.9615	0.1543	0.1094	0.0766	0.0435	0.0331
	P Boot	0.9875	0.9490	0.9425	0.9445	0.9595	0.1561	0.1091	0.0771	0.0433	0.0331
	N Boot	0.9325	0.9405	0.9435	0.9465	0.9610	0.1472	0.1090	0.0766	0.0434	0.0331
	SP Boot	0.9865	0.9540	0.9445	0.9450	0.9610	0.1518	0.1083	0.0767	0.0431	0.0330
2	Wald	0.9385	0.9395	0.9540	0.9490	0.9525	0.2188	0.1404	0.0898	0.0566	0.0438
	LR	0.9460	0.9450	0.9620	0.9540	0.9540	0.3613	0.2441	0.1689	0.0990	0.0786
	P Boot	0.9485	0.9490	0.9615	0.9535	0.9560	0.3716	0.2491	0.1705	0.1002	0.0785
	N Boot	0.9430	0.9465	0.9640	0.9535	0.9550	0.3646	0.2507	0.1684	0.0995	0.0785
	SP Boot	0.9465	0.9505	0.9595	0.9520	0.9550	0.3695	0.2499	0.1710	0.0995	0.0785
3	Wald	0.9560	0.9310	0.9440	0.9425	0.9525	0.0625	0.0432	0.0300	0.0146	0.0111
	LR	0.9590	0.9405	0.9510	0.9470	0.9580	0.0877	0.0679	0.0493	0.0264	0.0204
	P Boot	0.9020	0.9155	0.9470	0.9440	0.9515	0.0723	0.0591	0.0463	0.0258	0.0203
	N Boot	0.8720	0.9060	0.9490	0.9500	0.9585	0.0742	0.0585	0.0461	0.0258	0.0202
	SP Boot	0.9025	0.9115	0.9475	0.9440	0.9560	0.0723	0.0591	0.0463	0.0258	0.0200
4	Wald	0.9360	0.9450	0.9440	0.9435	0.9455	0.1311	0.0899	0.0647	0.0366	0.0284
	LR	0.9490	0.9535	0.9460	0.9480	0.9490	0.2343	0.1649	0.1143	0.0652	0.0510
	P Boot	0.9490	0.9505	0.9475	0.9430	0.9465	0.2397	0.1642	0.1160	0.0650	0.0509
	N Boot	0.9505	0.9525	0.9475	0.9455	0.9490	0.2375	0.1671	0.1150	0.0651	0.5105
	SP Boot	0.9490	0.9505	0.9495	0.9445	0.9495	0.2397	0.1642	0.1159	0.0651	0.0509
5	Wald	0.9445	0.9455	0.9480	0.9415	0.9485	0.4465	0.3082	0.2114	0.1236	0.0958
	LR	0.9495	0.9490	0.9545	0.9445	0.9500	0.8309	0.5700	0.3936	0.2229	0.1727
	P Boot	0.9625	0.9560	0.9600	0.9495	0.9540	0.9228	0.6050	0.4104	0.2294	0.1769
	N Boot	0.9615	0.9600	0.9615	0.9495	0.9535	0.9012	0.6020	0.4137	0.2287	0.1769
	SP Boot	0.9600	0.9575	0.9565	0.9490	0.9500	0.9247	0.6067	0.4153	0.2290	0.1761

[a]Methods from Sect. 2: Wald: simple Wald with std. error via delta method; LR: likelihood ratio; P boot: fully parametric bootstrap with percentile method; N boot: non-parametric bootstrap with percentile method; SP boot: semi-parametric bootstrap with percentile method