Species Sensitivity Distribution Estimation from Uncertain (QSAR-based) Effects Data

Tom Aldenberg and Emiel Rorije

National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

Summary — In environmental risk assessment, Species Sensitivity Distributions (SSDs) can be applied to estimate a PNEC (Predicted No-Effect Concentration) for a chemical substance, when sufficient data on species toxicities are available. The European Chemicals Agency (ECHA) recommendation is 10 biological species. The question addressed in this paper, is whether QSAR-predicted toxicities can be included in SSD-based PNEC estimates, and whether any modifications need to be made to account for the uncertainty in the QSAR-model estimates. This problem is addressed from a probabilistic modelling point of view. From classical analysis of variation (ANOVA), we review how the error-in-data SSD problem is similar to separation into between-group and within-group variance. ECHA guidance suggests averaging similar endpoint data for a species, which is consistent with group means, as in ANOVA. This exercise reveals that error-in-data reduces the estimation of the between species variation, i.e. the SSD variance, rather than enlarging it. A Bayesian analysis permits the assessment of the uncertainty of the SSD mean and variance parameters for given values of mean species toxicity error. This requires a hierarchical model. Prototyping this model for an artificial five-species data set seems to suggest that the influence of data error is relatively minor. Moreover, when neglecting this data error, a slightly conservative estimate of the SSD results. Hence, we suggest including (model-predicted) data as model point estimates and handling the SSD as usual. The Bayesian simulation of the error-in-data SSD leads to predictive distributions, being an average of posterior spaghetti plot densities or cumulative distributions. We derive new predictive extrapolation constants with several improvements over previous median uncertainty $\log_{10}$HC5 estimates, in that they are easily calculable from spreadsheet Student-*t* functions and based on a more realistic uniform prior for the SSD standard deviation. Other advantages are that they are single-number extrapolation constants and they are more sensitive to small sample size.

Key words: *Bayesian hierarchical model, extrapolation constants, predictive uncertainty, probabilistic risk assessment, QSAR and read-across prediction, species sensitivity distribution.*

Address for correspondence: *Tom Aldenberg, National Institute for Public Health and the Environment (RIVM), Antonie van Leeuwenhoeklaan 9, 3721 MA, Bilthoven, The Netherlands.*
E-mail: *tom.aldenberg@rivm.nl*

Introduction

In environmental risk assessment, the general approach to determine the Predicted No-Effect Concentration (PNEC) of a given chemical substance is to apply a safety factor to the lowest ecotoxicity value(s) observed in a species toxicity data set pertaining to the substance (1). The safety factor varies with the size of the data set and the uncertainty of the toxicity endpoint to which it refers. When the data set is large enough, however, statistical extrapolation can be used, on the basis of a model called a Species Sensitivity Distribution (SSD; 2, 3).

The SSD model assumes a parametric distribution for a set of chronic ecotoxicity data, e.g. No Observed Effect Concentrations (NOECs), measured for a number of species, taken to be representative for some type of target ecosystem, e.g. aquatic fresh water species. SSDs are also built for acute data, such as lethal concentrations (EC50/LC50 values). In either case, the SSD has to be fitted to the appropriate data, which technically is a matter of parameter estimation for a given model.

The PNEC can be derived from the fitted distribution, often — especially for chronic data — via estimating the fifth percentile of the SSD, called the Hazardous Concentration for 5% of the species (genera). The uncertainty of this HC5 value, due to the use of a finite data set, can be estimated and summarised through several measures, such as confidence limits (3).

Different parametric models are possible for SSDs. Commonly, the Normal distribution is fitted over common log-transformed toxicity data (3, 4). Earlier models addressed the Logistic distribution over similar data (2, 5). As two-parameter models, these distributions are good first choices, because of their general applicability and known mathematical properties, which allow percentile and model fit uncertainties to be evaluated. Also, in the European Chemicals Agency (ECHA) Guidance, the Normal and Logistic distributions are recommended as a first choice (1; p. 22). In order to

select, or pre-process the data, the ECHA recommends on this same page that: "The test data applicable to the most sensitive endpoint should be taken as representative for the species."

The objective is to obtain a single representative toxicity value for a species. What should be done with *multiple* values is stated like this: "Multiple values for the same endpoint with the same species should be investigated on a case-by-case basis, looking for reasons for differences between the results. For equivalent data on the same end-point and species, the geometric mean should be used as the input value for the calculation."

Taking the *geometric* mean of raw data corresponds to applying the *arithmetic* mean on log-transformed data. Considering means, when appropriate, introduces a new element in SSD modelling, as multiple data for the same endpoint of one species may not be identical: they exhibit variability, which can be translated into *uncertainty* of the mean estimate. In retrospect, a single species endpoint value, although not showing variability by itself, often derives from, say, a dose–response curve, which is estimated from dose–response data. Hence, the single species estimate is uncertain as well, and on many occasions, e.g. in many EU Risk Assessment Reports, its uncertainty is reported.

In the present EU FP7 CADASTER project that made this research possible, the main question with regard to effect modelling is whether species toxicity data for a certain chemical, which are estimated from QSARs, can be used in SSD assessments. QSAR models, like dose–response models, are estimated from (physicochemical) data, which makes their predictions uncertain. Acknowledging this fact, the introduction of QSARs to predict values in an SSD is essentially no different from using 'real' endpoint data in an SSD, or combining both types of measurement. However, we are forced to more-systematically address this predictive data uncertainty, despite the fact that SSD-practitioners have largely neglected this uncertainty in hitherto 'real' endpoint data cases.

Measurement data error may result when the error is decoupled from how it was originally estimated, e.g. dose–response curve fitting. As outlined in the ECHA technical guidance (1), data error can also arise from combining similar measurements on the same chemical, species, and endpoint, coming from different sources. A third type of measurement data error may be due to *read-across estimation* from a small set of chemicals sharing similarities with the chemical in focus.

Modelled data error is prediction error quantified from a model incorporating covariates. This obviously applies when the model and model input data on the covariates are available, but in principle could be handled as *given* prediction error, not very different from measurement error. In both cases, detached prediction error requires information on what it means and how it was calculated.

The present paper tries to address some of these issues. We recap how classical analysis of variation (ANOVA) treats *within-group* data error *versus between-group* data error, taking multiple estimates of one toxicological endpoint for one species as a *group*. This seems in line with the ECHA guidance to take *means* of similar species data and consider the SSD as a distribution of species means. The important message is that taking within-group error into account reduces the between-group error. Thus, re-interpreting SSD species toxicity variation as variation of species toxicity means, should, in the case of data error, eventually lead to reduced SSD standard deviation estimates. One obviously wants to know by how much, in relation to data or prediction error, and what the consequences are for estimates of hazardous concentrations. We study the simplest extension of the basic Normal SSD, in which all data, or prediction, errors are also distributed Normally with the same standard deviation for each species. Given the value of the data (prediction) error, we consider an SSD of species toxicity means, with an uncertain SSD-mean parameter and an uncertain SSD-standard deviation parameter.

As classical ANOVA is limited to point estimates of both variance components, and, moreover, subtraction of the within-mean square from the between-mean square may yield a negative point estimate of the SSD-standard deviation, a Bayesian analysis is needed to address the uncertainties of the estimates. The model is essentially a Bayesian hierarchical model, where the SSD mean and standard deviation are so-called *hyperparameters*. The theory and calculations are explained by Gelman *et al.* (6; pp. 134–138). The Bayesian analysis for a sample of $n = 5$, reveals, to our surprise, that the influence of data error seems relatively small.

Although more simulations are needed to confirm this, the fact that accounting for data error yields less conservative SSD fits — it follows that by *not taking data error into account*, as in the current guidance, one stays on the safe side. Intuitively, one would think that data error makes things worse, but that is not the case. Better measurement data, or better data predictions, reduce this conservativeness. This leads to the advice to base SSDs on measurement data, or QSAR point estimates, as we have always done.

The Bayesian analysis suggests two important improvements of current extrapolation constants as developed by Aldenberg & Jaworska (3). This leads to a revised set of — as we call them — *predictive extrapolation constants* that do more justice to small sample size, are easier to calculate, and are based on sample mean and standard deviation of species toxicity data, as before.

Variance Components in Classical ANOVA: Between Species and Within Species Variation

In the previous section, we have discussed some methods and assumptions of Normal SSD data fitting, supposing the data to be without error.

However, everyone involved with SSD fitting knows that input data are not without error in general. In some EU risk assessment reports, these errors are reported. As mentioned in the *Introduction*, the ECHA Guidance recommends the averaging of multiple estimates of one toxicological endpoint for one species, by taking the geometric mean (1; p. 22). On the logarithmic scale, this implies arithmetical averaging. Current SSD fitting proceeds, however, without taking the uncertainty of the species mean(s) into account. Also, for single-number species toxicities, no methods are proposed to modify the SSD fitting for point uncertainty.

When some or all of the input data must come from models, e.g. QSARs, we cannot ignore the incorporation of data error into the SSD-based risk assessment, although in no way does this differ fundamentally from taking measurement uncertainty into account.

Clearly, there are two components in variation: *between* species and *within* species. The archetypical example of competing variance components is classical ANOVA.

Suppose y_{ij} is the ith $\log_{10}$ toxicity of species j for the same endpoint and chemical. We have n species, and for simplicity, we assume that we have the same number $i = 1,\ldots, m$ replicates for each species. We denote μ as the mean of the SSD of species means, τ as the standard deviation of SSD of species means, and σ as the standard deviation within species. Again simplifying, σ is taken as identical for all the species.

Table 1 presents the classical ANOVA table layout to calculate *between* species and *within* species sums of squares and mean squares (6; p. 133, 7; p. 247).

The SSD mean is estimated from the grand mean:

$$\hat{\mu} = \bar{y}_{\bullet\bullet}$$

One computes the SS and MS columns from the data and uses MS_B and MS_W to estimate both σ, the within species standard deviation, as well as τ, the SSD species means standard deviation:

$$\begin{cases} \hat{\sigma} = \sqrt{MS_W} \\ \hat{\tau} = \sqrt{\dfrac{(MS_B - MS_W)}{m}} \end{cases}$$

The τ point estimate can become negative, which is considered a defect of basic classical ANOVA by Bayesians (6; p. 138, 7; pp. 247–248). We will only use this calculation to illustrate what happens with τ for increasing data error, σ. This is best demonstrated with a numerical example (Tables 2a, 2b, and 2c). Later on, we will analyse the same data from the perspective of a hierarchical Bayesian model.

Suppose we have five species with the toxicity values shown in row 2 of Table 2a. These are chosen as the inverse *standard* Normal cdf values (quantiles) at probabilities: [0.10, 0.30, 0.50, 0.70, 0.90], leading to the raw data: [–1.282, –0.524, 0.000, 0.524, 1.282]. The *sample mean* and *sample standard deviation* are 0.000 and 0.979, respectively. We rescale the sample by the sample statistics to get the standardised data: [–1.309, –0.536, 0.000, 0.536, 1.309]. This is our $n = 5$ SSD base data set (row 2 in Tables 2a, 2b, and 2c), with sample mean 0.0 and sample standard deviation 1.0. These rows are copied unmodified to rows 1 and 3 in Table 2a, representing error-free species data: $\sigma = 0$. The point estimates of the variance components become: $\hat{\sigma} = 0$, $\hat{\tau} = 1.000$. The SSD standard deviation estimate of species means is equal to the standard deviation of the base data set in row 2. This can be interpreted in the sense that estimating the SSD standard deviation from unreplicated data, like those in row 2, makes the silent assumption that the data are error-free.

In Table 2b, we subtract a small data error $\sigma = 0.1$ from row 2 to obtain row 1, and add the same amount to row 3. It follows from Table 2b that $\hat{\sigma} = 0.100$, $\hat{\tau} = 0.998$. Note that $\hat{\tau}$ is a little smaller

Table 1: A classical ANOVA table to calculate *between* species and *within* species sums of squares (SS) and mean squares (MS)

	df	SS	MS	Expected (MS)
Between species	$n - 1$	$SS_B = \sum_{i,j}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2$	$MS_B = SS_B/(n-1)$	$\sigma^2 + m \cdot \tau^2$
Within species	$(m-1) \cdot n$	$SS_W = \sum_{i,j}(y_{ij} - \bar{y}_{\bullet j})^2$	$MS_W = SS_W/((m-1) \cdot n)$	σ^2
Total	$m \cdot n - 1$	$SS_T = \sum_{i,j}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	$MS_T = SS_T/(m \cdot n - 1)$	

The last column shows the expected MS, given the variance components.

Table 2: ANOVA tables showing the effects of species toxicity error on estimated SSD variance

a) artificial log-toxicity data for five species, each measured without error

	1	2	3	4	5
1	–1.309	–0.536	0.000	0.536	1.309
2	–1.309	–0.536	0.000	0.536	1.309
3	–1.309	–0.536	0.000	0.536	1.309
$y_{\bullet j}$	–1.309	–0.536	0.000	0.536	1.309
$y_{\bullet\bullet}$	0.000				

	df	SS	MS	Estimates
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 0.000$	$MS_W = 0.000$	$\hat{\sigma}^2$
Total	14	$SS_T = 12.000$		

b) as in Table 2a, with 0.1 unit standard deviation in the three replicates per species

	1	2	3	4	5
1	–1.409	–0.636	–0.100	0.436	1.209
2	–1.309	–0.536	0.000	0.536	1.309
3	–1.209	–0.436	0.100	0.636	1.409
$\bar{y}_{\bullet j}$	–1.309	–0.536	0.000	0.536	1.309
$\bar{y}_{\bullet\bullet}$	0.000				

	df	SS	MS	Estimates
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 0.000$	$MS_W = 0.010$	$\hat{\sigma}^2$
Total	14	$SS_T = 12.100$		

c) as in Tables 2a and 2b, with 1.0 unit standard deviation in the three replicates per species

	1	2	3	4	5
1	–2.309	–1.536	–1.000	–0.464	0.309
2	–1.309	–0.536	0.000	0.536	1.309
3	–0.309	0.464	1.000	1.536	2.309
$\bar{y}_{\bullet j}$	–1.282	–0.524	0.000	0.524	1.282
$\bar{y}_{\bullet\bullet}$	0.000				

	df	SS	MS	Estimates
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 10.000$	$MS_W = 1.000$	$\hat{\sigma}^2$
Total	14	$SS_T = 22000$		

than 1.0, i.e. the sample standard deviation of the base data set.

When the data error $\sigma = 1.0$, we obtain Table 2c. Note that the data error now competes with the SSD variability. It follows from Table 2c that $\hat{\sigma} = 0.100$, $\hat{\tau} = 0.816$. Now, $\hat{\tau}$ is substantially smaller than 1.0, the error-free sample standard deviation of the base data set. If we take $\sigma = 1.5$ (not tabulated), $\hat{\tau}$ drops further down: $\hat{\sigma} = 1.500$, $\hat{\tau} = 0.500$.

Indeed, when the data error is taken equal to $\sigma = \sqrt{MS_B} = \sqrt{3.000} = 1.732$, the SSD would shrink to a point probability mass, as $\hat{\sigma} = 1.732$, $\hat{\tau} = 0$. Bigger σ-values would lead to a negative point estimate of τ, without altering the method.

In the next section, we will explore the Bayesian solution, to see if what ANOVA tells us — SSD shrinkage with increasing data error — comes out of the Bayesian analysis as well.

Normal SSD with Normal Data Error: Bayesian Analysis

In this section, we present a Bayesian analysis of the five-point SSD base sample of the previous section, explicitly assuming data error. This technique is called *hierarchical modelling*. The theory is explained in Gelman *et al.* (6; p. 134–138).

When the data error of each species is distributed Normally with standard deviation σ_j, the equation for the posterior density of SSD mean μ and SSD standard deviation τ becomes:

$$p(\mu,\tau\,|\mathrm{y}) \propto p(\mu,\tau)\prod_j \mathrm{Normal}\left(\bar{y}_j\,|\mu,\sqrt{\sigma_j^2+\tau^2}\right)$$

This is Equation 5.18 in Gelman *et al.* (6). We assume σ_j to be given and equal for each species j. These are interpreted as error of the species mean, or prediction estimate. We suppress the dot in the former ANOVA notation, regarding $\bar{y}_j$ as a point estimate, from either data, or a model. In hierarchical modelling parlance, μ and τ are called *hyperparameters*. A Bayesian method that includes estimation of the hyperparameters is called *full-Bayesian*.

The prior distribution $p(\mu,\tau)\propto 1$ is taken to be uniform. Gelman *et al.* (6) report that taking $\log(\tau)$ uniform, i.e. the reference (non-informative) prior for the error-free data SSD (3, 8), leads to trouble in the hierarchical model. Since the prior has influence with small sample sizes, the results cannot be directly compared with error-free SSD fitting based on the non-informative prior. The data vector **y** only enters the posterior via the species means, $\bar{y}_j$, for each species. These are the sufficient statistics.

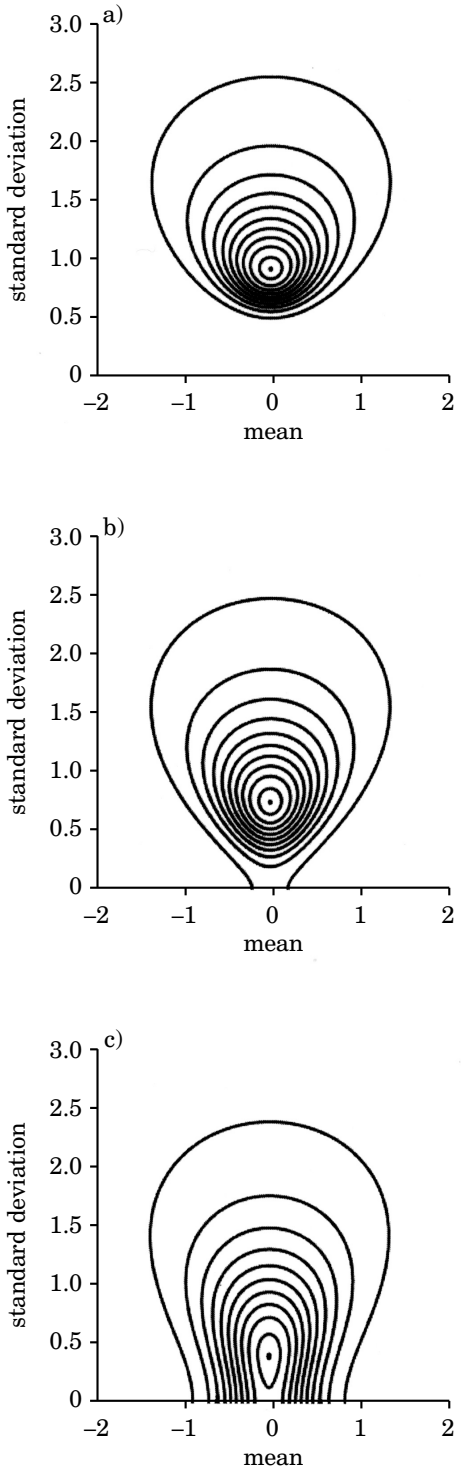
The five-point artificial data set is [–1.309, –0.536, 0.000, 0.536, 1.309], taken from Table 2a–2c. We interpret them as data averages, or model estimates, much in line with the ECHA Guidance. The sample mean is 0.0, and the sample standard deviation is 1.0.

Figure 1 shows the contour line plots of the posterior distribution of μ and τ for increasing values of σ: 0.1, 0.5, and 0.8. Finally, we compare with $\sigma = 0$. The reason to stop at $\sigma = 0.8$ is that at higher values the posterior distribution maximisation has convergence problems, when the optimum is approaching the origin (0, 0). Apparently, this occurs at a much lower data noise level than in the ANOVA.

The posteriors density values are scaled, in such a way that peaks have height 1.0 in each plot. Ten contours are drawn at values: 0.05, 0.15, 0.25,…, 0.95. The dot indicates the location of the maximum. Note that the maxima are located on the vertical line $\mu = 0$, and that τ_{max} *decreases* with *increasing* data error (σ), like in the ANOVA analysis.

The SSD intrinsic μ and τ values with maximum posterior density are, for each σ, as shown in Table 3, where we have also added the case where the species means are error-free ($\sigma = 0$).

Figure 1: Contour plots of posterior densities for Normal SSD mean (μ) and standard deviation (τ) for three given Normal data error standard deviations (σ)



*Artificial data set (*n = *5) from Table 2a (row 2).*

Posterior contours: a) sigma = 0.1; b) sigma = 0.5 and c) sigma = 0.8.

Table 3: Maximum posterior density estimates for the Normal SSD parameters

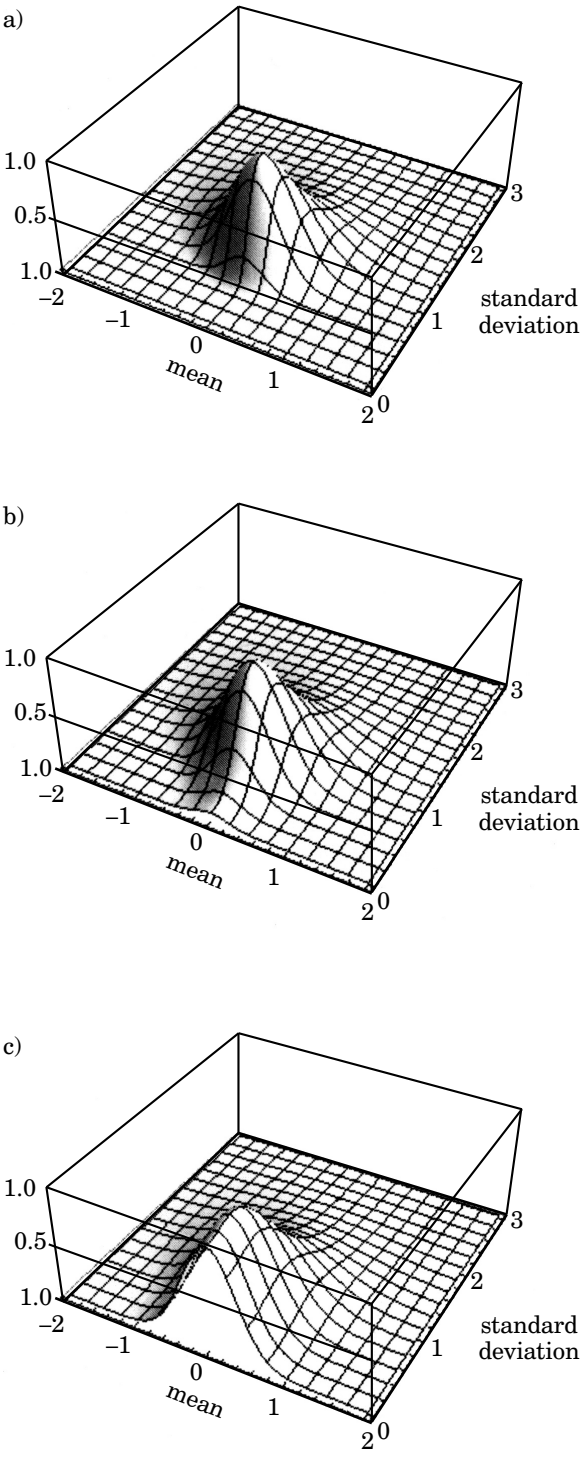
σ	μ_{max}	τ_{max}
0.0	0.0	0.894
0.1	0.0	0.889
0.5	0.0	0.742
0.8	0.0	0.401

μ = *mean and* τ = *standard deviation, for given Normal species error,* σ *equal for all species. Data (*n = *5) are from Tables 2a–2c, second rows respectively.*

Figure 2 presents 3-D plots of the posterior densities for the same settings as in Figure 1. Note that at the near critical $\sigma = 0.8$, the posterior maximum is approaching the boundary at $\tau = 0$. To simulate the predictive SSD distribution, we sample 20,000 points from the posterior distributions, by rejection sampling in the rectangular region delimited by: $\mu = [-4.0, 4.0]$ and $\tau = [0.0, 8.0]$. More sophisticated methods are available, e.g. Markov Chain Monte Carlo (MCMC) techniques, but this suffices for the moment. Each sampled point from the posterior densities defines a SSD, the threads generating a so-called *spaghetti plot*, as in Aldenberg and Jaworska (3). These are averaged point-wise to obtain the respective predictive distributions (both pdf and cdf), which are plotted in Figure 3 for increasing values of data error (σ). The dotted curves have lowest $\sigma = 0.1$; the dashed distribution curves result when the data error standard deviation ($\sigma = 0.5$) is roughly *half* the species data standard deviation. The continuous SSD curves (both pdf and cdf) are obtained when the data error competes with the SSD standard deviation. The cdf data plotting positions, only for visualisation, are those proposed by Tukey: $(i - 0.333)/(n + 0.333)$, which are easy to memorise, and respect median order statistics (8; pp. 88, 91).

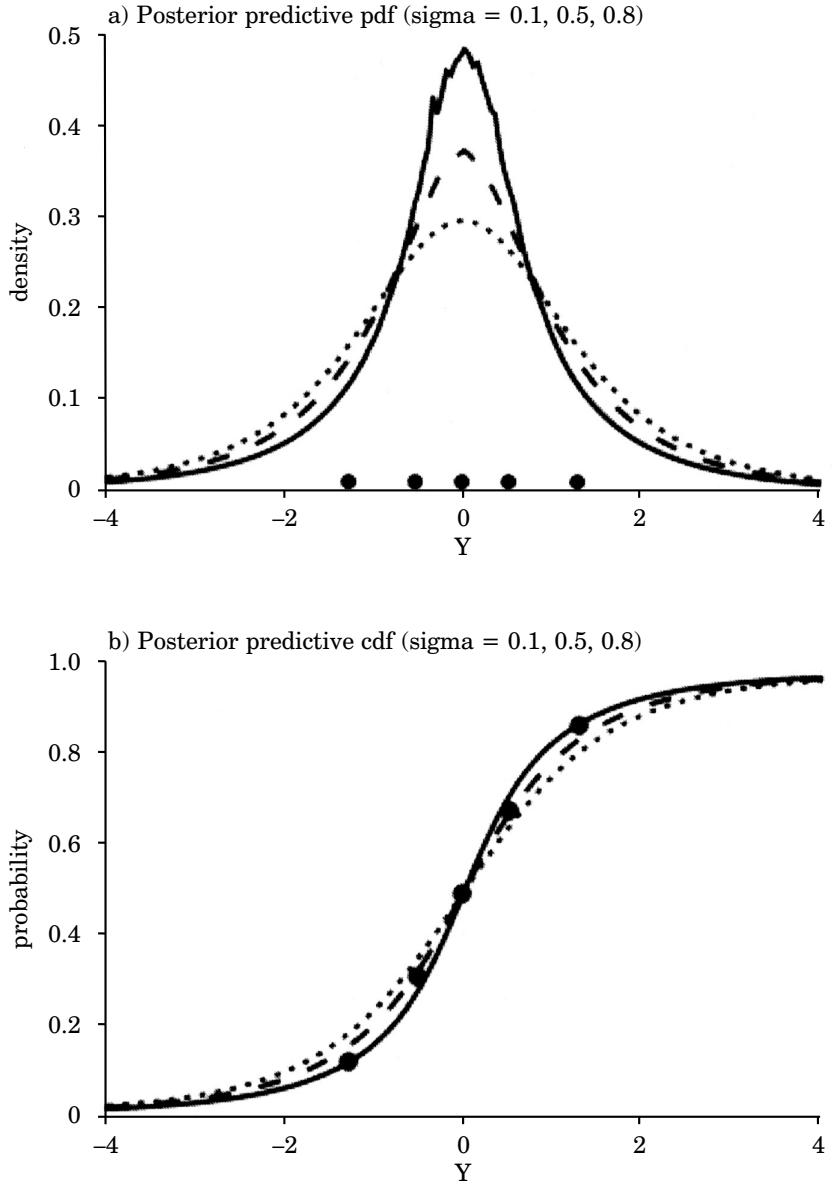
From the hierarchical cdf predictive SSDs, the right panel in Figure 3, we calculate the respective 5th percentiles of each predictive distribution in Table 4, back-calculating the HC5 for sample standard deviation equal to 1. Note that for increasing data error, the predictive SSDs become *less wide*, leading to an increase of the 5th percentile hierarchical predictive hazardous concentration. This conforms to the findings in the ANOVA. However, the increase seems quite modest to us. As in Table 3, we have added the simulation for negligible ('zero') data error (σ) in Table 4. Later on, we will relate this to the conventional non-hierarchical error-free data SSD. To be able to compare the hierarchical zero-error predictive $\log_{10}$HC5 to a similar measure from the traditional error-free

Figure 2: 3D plots of posterior densities for Normal SSD mean (μ) and standard deviation (τ), corresponding to the contour plots in Figure 1



Posterior densities: a) sigma = 0.1; b) sigma = 0.5 and c) sigma = 0.8.

Figure 3: Simulated posterior predictive SSDs fitted to the artificial data set from Table 2a (row 2) for increasing data error



pdf is shown in the upper panel; cdf is shown in the lower panel. In the artificial data set from Table 2a (row 2), n = *5. Increasing data error (0.1, 0.5, and 0.8) is represented by the dotted, dashed, and continuous lines, respectively. In the latter case, data error competes with the SSD standard deviation.*

$\log_{10}$HC5 (which must obviously be equal), we need to surmount some technical hurdles. This is accomplished in the next section.

The preliminary conclusion from this simulation is that data error seems to have a quite moderate influence on the 5th percentile estimation of the predictive SSD, even for a five-point data set. Moreover, neglecting data error yields a slightly conservative estimate, which we are going to exploit to our advantage.

Predictive Hazardous Concentrations: New Extrapolation Constants

The predictive SSD distribution is a model for the probability of where to find a new observation given a sample of size n. Predictive distributions are point-wise averages of pdf and cdf spaghetti plots (Figure 3). Previously (3), we have adopted the non-informative prior for SSD mean and stan-

Table 4: Logarithmic 5th percentile Hazardous Concentrations, and anti-logs, from hierarchical Normal–Normal SSD models for given (mean) species error σ, equal for all species

σ	$\log_{10}$HC5	HC5
0.0	–2.97	1.07×10^{-3}
0.1	–2.95	1.12×10^{-3}
0.5	–2.71	1.95×10^{-3}
0.8	–2.35	4.47×10^{-3}

*Data (*n = *5) are from Tables 2a–2c. We observe a relatively minor influence of data error on the results.*

dard deviation in the Bayesian analysis for error-free data SSDs: $p(\mu,\tau) \propto 1/\tau$.

However, the error-in-data model will not work with the non-informative prior, but needs the uniform prior (6). Over the years, we have developed reservations about the non-informative prior also in the error-free data case (which almost never holds anyway). Despite its historical importance in Bayesian statistics, the non-informative prior tends to favour smaller, quite biased, SSD standard deviations, with no justification at all. And this especially matters in small samples. We currently think that the non-informative prior is unwarranted in small sample SSD fitting. So, we are inclined to switch to the uniform prior for the non-hierarchical model as well, which moreover makes it possible to connect the error-free data SSD and the zero-error error-in-data SSD.

With the uniform prior, $p(\mu,\tau) \propto 1$, the joint maximum posterior point estimate becomes:

$$\begin{cases} \hat{\mu} = \bar{y} \\ \hat{\tau} = \frac{\sqrt{n-1}}{\sqrt{n}} . s_y \end{cases}$$

The τ estimate is essentially the n-based standard deviation known from maximum likelihood arguments. With the uniform prior, the posterior density becomes the normalised likelihood function, which is important, as both Bayesians and non-Bayesians accept the likelihood as the default weighting function of parameter uncertainty.

For the n = 5 artificial data set, with sample mean 0 and sample standard deviation 1, the max posterior estimates becomes:

$$\begin{cases} \hat{\mu} = 0 \\ \hat{\tau} = \frac{\sqrt{n-1}}{\sqrt{n}} = \frac{2}{\sqrt{5}} = 0.894 \end{cases}$$

which conforms to Table 3, for data-error σ = 0.0. So the uniform prior error-free SSD and the hierarchical SSD model, but with zero data error, now yield the same answer.

For any Normal model, e.g. SSD or read-across, the classical *non-informative prior* leads to a predictive distribution (see Appendix 1) that is Student-t with $n - 1$ degrees of freedom, location parameter $\bar{y}$ (the sample mean), and scale parameter proportional to the sample standard deviation:

$$s_y \cdot \sqrt{1 + \frac{1}{n}}$$

The posterior uncertainty of the mean, μ, has a reduced scale parameter:

$$s_y \cdot \sqrt{\frac{1}{n}}$$

The *uniform prior* changes only the parameters of the Student-t distributions involved (7; p. 102). In this case, the predictive distribution becomes a Student-t with $n - 2$ degrees of freedom (one df less!), the same location parameter $\bar{y}$ (the sample mean), but enlarged scale parameter proportional to the sample standard deviation:

$$s_y \cdot \sqrt{1 + \frac{1}{n}} \cdot \sqrt{\frac{n-1}{n-2}}$$

In fact, Box & Tiao (7) note that this predictive scale parameter can be re-interpreted as a modified estimate of the SSD standard deviation:

$$\bar{s}_y = \sqrt{\sum (y_i - \bar{y})^2 / (n-2)}$$

Percentiles (e.g. the 5th percentile) of the predictive distribution can be used as an estimate of $\log_{10}$HC5, doing more justice to small sample size than the current median extrapolation constants, as developed by Aldenberg and Jaworska (3). The new extrapolation constants are easily calculated in MS Excel® as:

= T.INV(FRAC,N−2)*SQRT(1+1/N)*SQRT((N−1)/(N−2))

Similar T-functions would allow the plotting of predictive distributions (pdf and cdf), as well as those of posterior μ, i.e. by leaving out '1+' from SQRT(1+1/N). At sample size n = 5, the predictive extrapolation constant becomes (from the above spreadsheet function) –2.997, which quite well matches the 5th percentile predictive *hierarchical* error-in-data simulated value at σ = 0 in Table 4. Hence, for the fitting of Normal SSDs, and the development of new extrapolation constants, we propose two modifications: from median percentile uncertainty extrapolation estimates toward new *predictive* extrapolation constants, and from the too-optimistic, non-informative prior toward the more-realistic uniform prior.

In Table 5, we present extrapolation constants k_{pred} for a selection of fractions of species affected. For unstandardised samples, the extrapolation formula becomes: $\bar{y}+k_{\mathrm{pred}}\cdot s_y$

We have extended the percentages in Table 5, in order to cover extrapolation of both chronic/NOEC and acute data, cf. Aldenberg and Luttik (9). The ECHA-recommended sample size (of 10) is shown in bold-face, and the extrapolation constants for very large sample size, which derive from the standard Normal distribution, are shown in italics. The major reasons why we like these predictive extrapolation constants are four-fold:

1. They are based on the uniform prior and therefore a limiting case of the error-in-data model.
2. They are easy to calculate from the predictive distribution.
3. They are more sensitive to sample size, compared to the median $\log_{10}$HC5 uncertainty estimators.
4. There is just one number, instead of three.

Note that at sample sizes 3, 4, and 5, the classical median extrapolation constants (3, 8) are –1.94, –1.83, –1.78, which are not dramatically different from the 'infinite' sample size value: –1.65. In current practice, the problem is that the lower and upper extrapolation constants, *which do reflect sample size*, are reported only occasionally.

The predictive extrapolation constants at sample sizes 3, 4, and 5, when rounded, are: –10, –4, –3, becoming –2 at the ECHA-recommended sample size of 10, and likewise settling at –1.6 for large samples. This is due to the predictive distribution to be the *average SSD* from the spaghetti plot, as well as the more realistic uniform prior for the SSD standard deviation.

Discussion: The Proposal to Not Correct for Data Error

Although the basic ANOVA point estimation procedure is insufficient as a general procedure to correct SSD and hazardous concentration estimation for data error, it points in the right direction that an estimate of SSD species toxicity variation should decrease for increasing data noise.

In this paper, we have prototyped the full-Bayesian solution of the simplest hierarchical model to address the interaction between two Normal variance components: variation *between* expected species toxicity values, and uncertainty *within* a species toxicity estimate. The Bayesian analysis confirms the ANOVA findings: the more noise within the individual species estimates, the less information remains for the SSD variance of expected endpoint values of different species.

In the ECHA Guidance on SSD (1), one is advised to *average* multiple data for the same endpoint and species. This implies that, conceptually, one strives for stable species numbers approximating expected values. In particular, an SSD should not be a mixed bag of raw multiple predictions, or measurements, for each species. The ECHA guid-

Table 5: Predictive extrapolation constants for given fraction of species affected for samples of size *n*, standardised by sample mean and sample standard deviation

n	0.10%	0.25%	0.50%	1%	2.5%	5%
3	*	*	*	*	*	–10.310
4	*	*	*	–9.537	–5.892	–3.998
5	*	–9.428	–7.388	–5.744	–4.026	–2.977
6	–8.662	–6.760	–5.560	–4.525	–3.353	–2.574
7	–6.902	–5.590	–4.722	–3.941	–3.010	–2.360
8	–5.966	–4.946	–4.247	–3.600	–2.803	–2.226
9	–5.392	–4.541	–3.943	–3.378	–2.665	–2.135
10	**–5.007**	**–4.263**	**–3.733**	**–3.222**	**–2.565**	**–2.069**
12	–4.523	–3.910	–3.460	–3.017	–2.432	–1.979
15	–4.128	–3.615	–3.229	–2.841	–2.315	–1.898
20	–3.801	–3.365	–3.030	–2.687	–2.212	–1.826
30	–3.526	–3.152	–2.859	–2.552	–2.119	–1.760
50	–3.336	–3.003	–2.737	–2.456	–2.052	–1.711
100	–3.208	–2.901	–2.653	–2.389	–2.005	–1.677
200	–3.148	–2.853	–2.614	–2.357	–1.982	–1.661
∞	*–3.090*	*–2.807*	*–2.576*	*–2.326*	*–1.960*	*–1.645*

**Entries smaller than –12. The ECHA-recommended sample size (of 10) is shown in bold-face. The extrapolation constants for very large sample size, which derive from the standard normal distribution, are shown in italics.*

ance does not recommend correcting for data error, whatever its source. In this paper, we have observed that implementing such a correction would be likely to *reduce* estimates of species toxicity variance. This means that not doing so yields a *conservative estimate of the SSD*. A very nice consistency coming out of the mathematics is that the current SSD methodology, implicitly assuming error-free data, yields the same results as the error-in-data hierarchical model for negligible error.

The $n = 5$ trial example shows that — to our surprise — the effects of the within species estimation error seem quite modest, even in cases where this error approaches the SSD species variation, although other sample sizes and other distributions need to be studied to confirm this.

If SSD guidance requires correction for estimation error of expected species toxicities, then this will not simply be a matter of making the hierarchical SSD model available. In practice, it will be very difficult to get reliable species error estimates, as the species data derive not only from different models, but from different statistical estimates, asymmetric *versus* symmetric intervals, small sample *versus* asymptotic confidence limits, varying experimental conditions, and so on. In addition, the raw data may not be available to allow every estimate to be harmonised.

As the influence of individual species estimation error seems rather limited and more data error leads to less conservative estimation, we think that *not correcting for species estimation noise* is in fact defensible, at least for the time being. We assume that, if species toxicity estimates were to become more precise, then conservativeness would decrease. Mathematical evidence suggests that this is a persistent pattern. Noise in the data reduces the SSD variability of estimated species toxicities.

When including QSAR-based species estimates, the situation is very similar to that for estimated species data, or for dose–response derived species data. The ECHA guidance is very clear on the point that multiple data on the same endpoint and species must be averaged. The implication of averaging multiple measured data for advice on how to deal with modelled data is that the *model estimate* goes into the SSD — not the model estimate uncertainty, nor the predictive uncertainty of new data, given the model. Hence, it makes little sense to perform error propagation on SSD results from either of these two uncertainty sources.

The final recipe is embarrassingly simple: collect your average data or model point estimates; then use the revised predictive extrapolation constants, as if these estimates were error-free, and you're done. The hierarchical model paradigm strongly suggests that your extrapolation is on the safe side — without extreme over-protecting.

In read-across estimation (10), we can follow a similar procedure. Read-across is much like SSD estimation, since no covariates are involved. One can either hunt for the prediction of the expected value, or the predictive uncertainty of where similar data are to be found (see Appendix 1). By neglecting the error in the data, one obtains a conservative estimate of the variation, i.e. on the high side.

If one feels the need to make adjustments for data or model error, then the appropriate analysis is a hierarchical model that corrects for estimation errors of *expected values*, by separating variance components, instead of by propagating errors and compounding variance components.

Since species data are often model estimates from modelled dose–response data, there seems little difference between species toxicity data from dose–response models and species toxicity estimates from QSARs based on predictor covariates. We do not see any fundamental problem in combining the two types of estimates into a single SSD.

The present paper exploits the Bayesian predictive distribution concept for estimating a single curve SSD, which operationally is a mean curve of Bayesian spaghetti plots. The ubiquitous median $\log_{10}$HC5 estimator is found to be quite robust for varying sample size — in fact, too much so. The skewness of the $\log_{10}$HC5 uncertainty distribution is often unaddressed, yielding an over-optimistic impression of insensitivity of the median estimator. The same is true for the ubiquitous non-informative prior on the SSD standard deviation, which we now consider to be too optimistic, and in need of modification in the error-in-data model anyway.

Deliberating these thoughts, we have presented a set of revised *predictive extrapolation constants* (Table 5), to be used in the same way as the median percentile estimator: i.e. by multiplying by the sample standard deviation and subtracting the result from the sample mean. In Table 5, these predictive extrapolation constants vary over sample sizes and protection levels. The predictive SSD distribution that defines the revised extrapolation constants is much less complicated to evaluate than the former percentile uncertainty distributions, and has more realistic sample size dependence built in.

Although the ECHA recommendation of minimum sample size of 10, preferably 15, is very reasonable advice, we have tabulated the predictive extrapolation constants for lower sample sizes as well. Not only do lower sample sizes occur for some branches of ecotoxicology (sediments, terrestrial data, etc.), but the improved sample size dependence relaxes the minimum sample size requirements to some extent.

Extrapolation is done on both NOEC, or chronic, data, as well as (sub-) acute data. Although a

plethora of existing safety factors is one way of addressing the different kinds of toxicological effect, Table 5 presents an enlarged range of target species protection levels to adapt to the needs of acute *versus* chronic risk assessments.

Note that, if the SSD model has any predictive value, then safety factors should always depend on (estimated) SSD variance.

Acknowledgements

This research was funded by the European Seventh Framework Programme through the CADASTER project FP7-ENV-2007-1-212668. The authors wish to thank the partners in the CADASTER project for their cooperation and stimulating discussions, especially Mojca Kos Durjava, Laura Golsteijn, Mark Huijbregts, Muhammad Sarfraz Iqbal, Willie Peijnenburg, Ullrika Sahlin, and Igor Tetko. Wout Slob and Ad Ragas suggested that error-in-data would lead to reduced SSD variance, many years ago. Christian Damgaard advocated use of the predictive distribution for estimating the HC5, decades ago.

References

1. ECHA (2008). *Guidance on information requirements and chemical safety assessment. Chapter R.10: Characterisation of dose [concentration]–response for environment*, 65pp. Helsinki, Finland: European Chemicals Agency.
2. Aldenberg, T. & Slob, W. (1993). Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology & Environmental Safety* **25**, 48–63.
3. Aldenberg, T. & Jaworska, J.S. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology & Environmental Safety* **46**, 1–18.
4. Wagner, C. & Løkke, H. (1991). Estimation of ecotoxicological protection levels from NOEC toxicity data. *Water Research* **25**, 1237–1242.
5. Van Straalen, N.M. & Denneman, C.A.J. (1989). Ecotoxicological evaluation of soil quality criteria. *Ecotoxicology & Environmental Safety* **18**, 241–251.
6. Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd edn, 668pp. Boca Raton, FL, USA: Chapman & Hall/CRC.
7. Box, G.E.P. & Tiao, G.C. (1973/1992). *Bayesian Inference in Statistical Analysis*, 588pp. New York, NY, USA: Wiley.
8. Aldenberg, T., Jaworska, J.S. & Traas, T.P. (2002). Normal species sensitivity distributions and probabilistic ecological risk assessment. In *Species Sensitivity Distributions in Ecotoxicology* (ed. L. Posthuma, G.W. Suter, II, & T.P. Traas), pp. 49–102. Boca Raton, FL, USA: CRC/Lewis Publishers.
9. Aldenberg, T. & Luttik, R. (2002). Extrapolation factors for tiny toxicity data sets from Species Sensitivity Distributions with known standard deviation. In *Species Sensitivity Distributions in Ecotoxicology* (ed. L. Posthuma, G.W. Suter, II, & T.P. Traas), pp. 103–118. Boca Raton, FL, USA: CRC/Lewis Publishers.
10. Rorije, E., Aldenberg, T. & Peijnenburg, W.J.G.M. (2013). Read-across estimates of aquatic toxicity for selected fragrances. *ATLA* **41**, 77–90.
11. Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd edn, 706pp. New York, NY, USA: Wiley.
12. Helsel, D.R. & Hirsch, R.M. (2002). *Statistical Methods in Water Resources*, 524pp. Reston, VA, USA: US Geological Survey. Available at: http://pubs.usgs.gov/twri/twri4a3/#pdf (Accessed 02.02.12).
13. Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd edn, 363pp. New York, NY, USA: Springer.

Appendix 1

Matrix Formulation of Model Prediction Uncertainty

We consider a regression model with k predictors. The predictor variables can be non-linear functions of covariates, although this is not often done. Linear regression assumes linearity of the regression coefficients.

The data vector is y and the design matrix is called X. For QSAR models, each row of the matrix is a $(1 + k)$-dimensional (QSAR) vector of predictors, each for a chemical in the training data set. Numbers and types of predictors can differ between QSARs.

The model estimates, $\hat{y}$, at the training points, are:

$$\hat{y} = X \cdot \hat{\beta}$$

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}}y$$

The mean square error (MSE) due to regression is:

$$s^2 = \frac{1}{n - (1 + k)} (y - X \cdot \hat{\beta})^{\mathrm{T}} (y - X \cdot \hat{\beta})$$

At a given $(1 + k)$-dimensional QSAR input vector, x_0, of a new substance, say, the model point estimate is: $\hat{y}_0 = x_0 \cdot \hat{\beta}$

Assuming Normal errors, the Bayesian uncertainty of the *model estimate* for this substance has a *location-scale* Student-t distribution, with location parameter: $\hat{y}_0$, scale parameter:

$$s \cdot \sqrt{x_0^{\mathrm{T}} \cdot (X^{\mathrm{T}}X)^{-1} \cdot x_0}$$

and $df = n - (1 + k)$ degrees of freedom. This Student-t distribution is analogous to the classical confidence interval, with significance level (two-sided), for the expected response at QSAR input vector x_0:

$$\hat{y}_0 \pm t_{\alpha/2, n-(1+k)} \cdot \sqrt{s^2 \cdot x_0^{\mathrm{T}} \cdot (X^{\mathrm{T}}X)^{-1} \cdot x_0}$$

With only one covariate, the input vector is: $x_0 = (1\ x_{10})$. Then, the model estimate becomes:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{10}$$

The 95% two-sided *model estimate confidence interval* at x_{10} turns out to be:

$$\hat{y}_0 \pm t_{0.975, n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_{10} - \bar{x})^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x})^2}}$$

[cf. Draper and Smith (11; p. 80–81), Helsel and Hirsch (12; p. 241), Dalgaard (13; p. 120).]

As a function of x_{10}, these model estimate confidence bands plot as the familiar hollow shaped curves around the line. Many data points may be outside of these bands.

If the input covariate x_{10} is located in the mean of the x training data, $\bar{x}$, and sample size is very large, then the model estimate uncertainty becomes *nil* at that point. This is the centre of the regression. Outside the centre, the model estimate uncertainty increases with the distance of the input covariate x_{10} from $\bar{x}$, even for very large samples.

Similar to the prediction uncertainty of the model estimate, the Bayesian *predictive distribution* of an *individual observation* of the response variable for a new substance is a Student-t distribution, again with location parameter: $\hat{y}_0$, but now with increased scale parameter:

$$s \cdot \sqrt{1 + x_0^{\mathrm{T}} \cdot (X^{\mathrm{T}}X)^{-1} \cdot x_0}$$

and $df = n - (1 + k)$ degrees of freedom.
Again the *classical predictive limits* are consistent with the Bayesian version. The 95% two-sided predictive limits for a single covariate are:

$$\hat{y}_0 \pm t_{0.975, n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{10} - \bar{x})^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x})^2}}$$

As a function of x_{10}, these bands plot as almost straight lines more or less enveloping the data. In the centre of the regression (see above), and with very large sample size, the predictive limits settle at $\hat{y}_0 \pm 1.96 \cdot s$.

Read-Across and SSD as 'Covariate-less' Regression

The matrix regression formalism of the previous section allows an easy derivation for modelling a Normal distribution of variability without covariates, e.g. $k = 0$, such as in read-across or an SSD. The design matrix of training points becomes: $X = (1, 1, ..., 1)^{\mathrm{T}}$, i.e. a n-size column vector of ones, since we only have an *intercept* (constant) to be estimated.
With $X^{\mathrm{T}} \cdot X = n, (X^{\mathrm{T}} \cdot X)^{-1} = 1/n$, the intercept is estimated by the mean of the data:

$$\hat{\beta}_0 = (X^{\mathrm{T}} \cdot X)^{-1} \cdot X^{\mathrm{T}} \cdot y = \frac{1}{n} \cdot (1, 1, ..., 1) \cdot (y_1, y_2, ..., y_n)^{\mathrm{T}} = \bar{y}$$

With input: $x_0 = 1$, the model estimate is:

$$\hat{y} = \hat{y}_0 = \hat{\beta}_0 = \bar{y}$$

The MSE due to regression is:

$$s^2 = \frac{1}{n - (1 + k)} (y - X \cdot \hat{\beta})^{\mathrm{T}} (y - X \cdot \hat{\beta}) = \frac{1}{n - 1} \sum_i (y_i - \bar{y})^2$$

the usual *sample variance*.

The uncertainty distribution of the *model estimate* is Student-t with location: $\hat{y} = \bar{y}$, and scale parameter:

$$s \cdot \sqrt{x_0^{\mathrm{T}} \cdot (X^{\mathrm{T}} \cdot X)^{-1} \cdot x_0} = s \cdot \sqrt{1 \cdot \frac{1}{n} \cdot 1} = s \cdot \sqrt{\frac{1}{n}}$$

which shrinks to negligible values for very large samples. The degrees of freedom is $n - 1$.

The uncertainty distribution of a *new observation*, or *predictive distribution*, is also Student-*t* with location: $\hat{y} = \bar{y}$, but larger scale parameter:

$$s \cdot \sqrt{1 + x_0^{\mathrm{T}} \cdot (X^{\mathrm{T}} \cdot X)^{-1}} \cdot x_0 = s \cdot \sqrt{1 + \frac{1}{n}}$$

which converges to s in very large samples. The degrees of freedom is also $n - 1$.