Parma, September 27 and 28th

EFSA Guidance on dermal absorption

Statistical re-analysis to derive default values

Ludwig A. Hothorn
retired Leibniz University Hannover, Germany
e-mail: hothorn@biostat.uni-hannover.de
Biostatistics consultant for ECPA

August 31, 2017

The problem I

- Main question: **are these default values appropriate?**

Table B13: Default values to be used in absence of experimental data

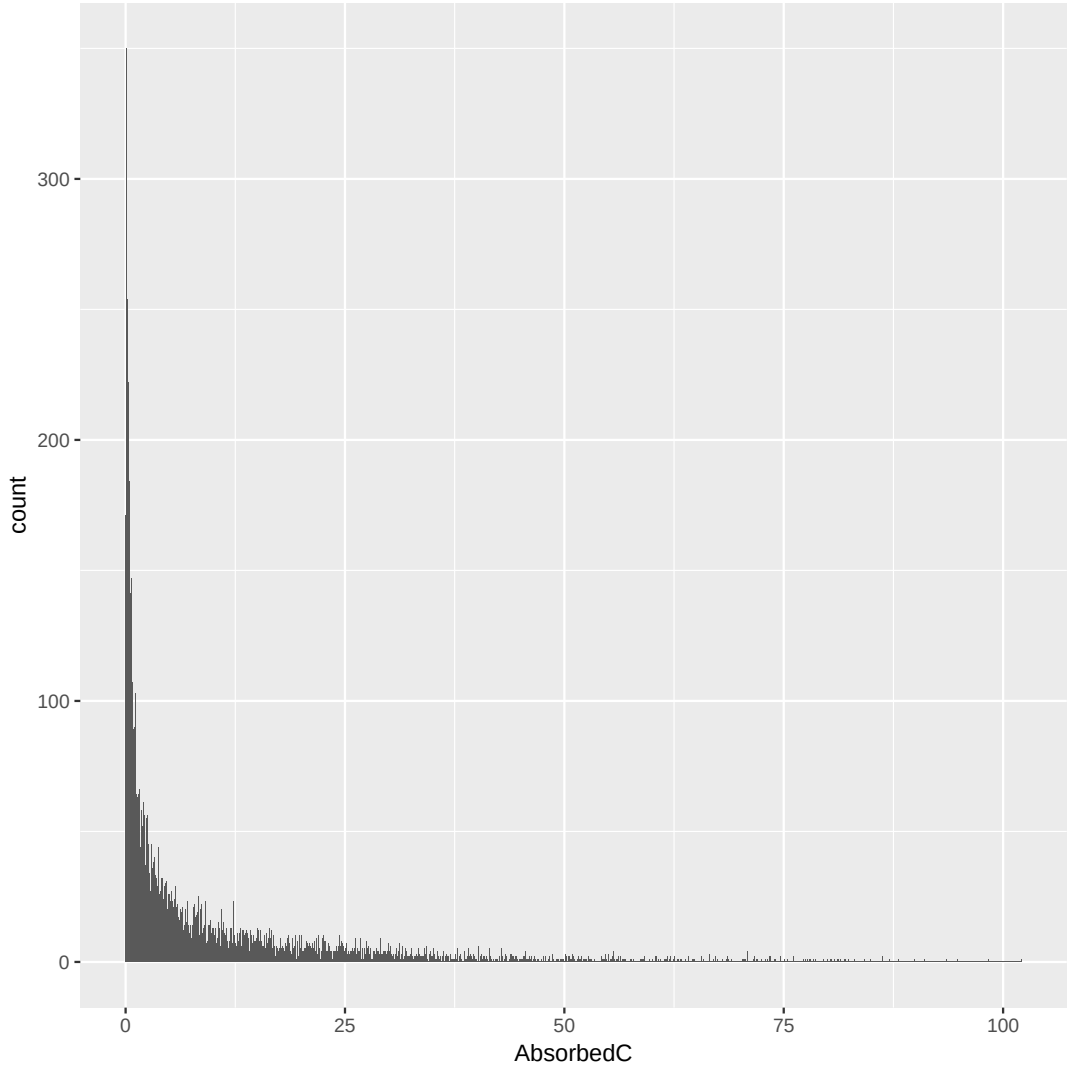
Formulation category	Concentration status	Default value (%)
Organic solvent-based[(a)] and others[(b)]	Concentrate	25
	Dilution	70
Water-based/dispersed[(c)] or solid[(d)]	Concentrate	10
	Dilution	50

- Purely from a (bio)statistical perspective
- All other interesting issues, e.g. has molecular weight an influence will be not considered today
- Sorry, we run into high-sophisticated statistical issues- I try to simplify for 20 mins

Basic considerations I

- The appropriate method is a prediction interval (here the upper limit) for a single future observation, estimated from x donors (with y technical replicates) in a complex, hierarchical design
- However, such intervals can be calculated up to now for strict normal distributed historical data using a specific mixed effects model approach (Schaarschmidt 2016)
- But, the absorption data are extreme skewed- quite naturally for absorption values

Basic considerations II



Basic considerations III

- Assuming that the variances between donors (and between replicates) are small with respect to the main factors (e.g. between substances, between concentrated/diluted, between data base parts, etc), a modified experimental unit is defined as the median over donors (and over replicates).
- I.e. sample size is substantial reduced : from 6323 to 778.

  Remember: averaging stabilizes estimators in extreme hierarchical and unbalanced designs
- This approximation allows the estimation of a prediction interval in a **simple fixed effects** model. (**Assumption I**).

Basic considerations IV

- Furthermore, the prediction interval is a function of i) the number of the historical data, ii) the number of predicted values of a group (here simplified to just a single future observation), iii) the $(1 - \alpha)$ probability and iv) some more.
- For $N > 50$, it can be approximated by well known 2σ interval) (Hothorn, 2015, chapter 2.3) (**Assumption II**)
- For medium to large sample sizes the 2σ rule is approximated by the 97.5% upper quantile for normal distribution. E.g. for N=100 2σ/Quantile: 14.5, 14.2 (**Assumption III**)

Basic considerations V

- Dermal absorption data are naturally heavily left-skewed for many non- to weak absorptions. The primary endpoint is the median absorption rate (Abs), a proportion between 0 and 1, (nominator and denominator of this proportion is not available p=number of events/number of cases, just the proportion).
- Notice, the median absorption rate is also the single future observation, ie. **the single value per assay**

Model selection I

- Statistical analysis of skewed proportions is a delicate statistical problem at all
- Appropriateness of the assumed underlying distribution is key for the estimation of a prediction interval, much more than related 2- or k sample tests (dominating in literature)
- (at least) Four approaches can be used:
  1. beta regression (Zeileis 2009) EFSA confirms lacking fit
  2. logit regression (Appendix B) EFSA approach
  3. most likely transformation **library(mlt)** (Hothorn 2017)
  4. naive transformation (eg. Arcsine-root) (not used, because mlt available)
  5. ...

Model specification I

- Model-based prediction intervals represent the best approach, ie. fitting a joint model for all data taking the classification factors (substances, data base, concentrated,..) into account and predict for the related subclassifications.
- In the recent version of mlt, this is available for simple factor structure. Notice, for hierarchical data structure with massive missing values may be problematic. Not available for designs with random factors
- Therefore, the analysis here is focusing on the estimation of less-biased upper prediction intervals for
  1. factor `Typeconcentration` with two levels `Concentrated, Diluted`
  2. the pre-selected formulation classifications by simple cutting the entire data set into subsets according to Tab B13
  3. Notice: can be improved later

Model specification II

- Most remaining sample sizes are almost > 50 (and therefore the interval is well estimated with 2σ,) and by the robust 97.5 percent quantile of the mlt-estimated distribution (**Assumption IV**)
- No interactions between classification can be estimated, but the error structure within a subset is unbiased for any transformation model (**Assumption V**).
- Unfortunately, no unique criteria of "best fit" for these four models exists (such as BIC, LL, ...) (*But remember Box: all models are wrong, some are helpful*)
- The result of the calculations are upper prediction limits for certain classifications. Again, no "absolute" quantitative criteria exist which limits differ (ie the classification give sense) or do not (ie the classification is not appropriate).

Model specification III

- Important: default values are only valid for exactly this data set. In other word, adding, say 20 new substances with higher absorptions would shift the default values to the right. Notice: BFR 0.42, ECPA 0.19

Default values by most likely transformation approach I

- Quite recently (Hothorn,2017) the R-library(mlt) was proposed for maximum likelihood estimators in the class of conditional transformation models.
- Models for the unconditional or conditional distribution function of **any univariate response variable** can be estimated by choosing an appropriate transformation function and parameterisation.
- Also for **discrete responses**, the asymptotic normality of the proposed estimators was shown.
- Particularly simple count regression models are proposed, specially for such count data suffering from over-dispersion or **excess zeros**.

Default values by most likely transformation approach II

- The mlt-model in R-code:

```
library(mlt)
avar <- numeric_var("Abs", bounds = c(0, 1), support = c(0, .99))
m <- ctm(Bernstein_basis(avar, order= 5, ui = "increasing"), todistr = "Normal")
mt <- mlt(m, data = MDat, scale = TRUE) # including zeros
#plot(mt, newdata = data.frame(1), type = "density", col = "black")
pto<-predict(mt, newdata = data.frame(1), type = "quantile", p = c(.025, .975))
```

- The upper 97.5% quantile of the estimated distribution function for all data without a subclassification is given by 0.2885
- **Extension I:** By means of an **additive shift**, the quantiles for the two levels of the factor `Typeconcentration`, namely concentrated, diluted can be estimated assuming an additive effect. Additive shift is similar to a factor in ANOVA, but not the same

Default values by most likely transformation approach III

```
B_x <- as.basis( ~ Typeconcentration, data=MDat)
ms <- ctm(Bernstein_basis(avar, order= 5, ui = "increasing"), todistr = "Normal", shift
mtT <- mlt(ms, data = MDat, scale = TRUE) # including zeros
lex<-expand.grid(Typeconcentration=c("Concentrate", "Diluted")) # lexigrafisch geordnet
plot(mtT, newdata = data.frame(lex), type = "density", col = c("black", "blue"))
```

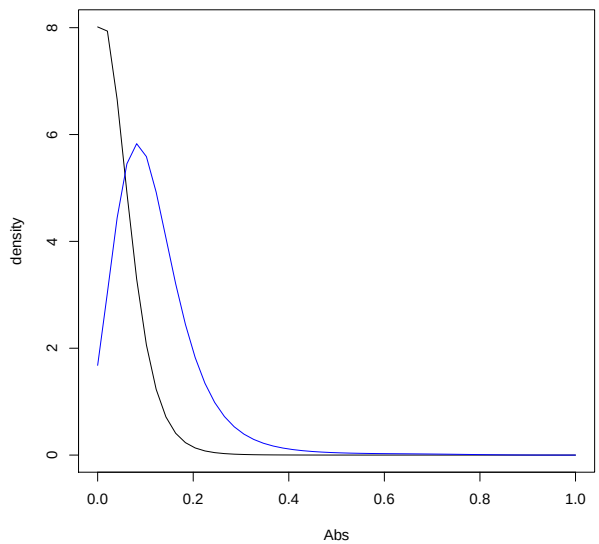


Figure: Estimated densities for concentrated, diluted

```
ptT<-predict(mtT, newdata = data.frame(lex), type = "quantile", p = c(.025, .975))# 2 s
```

Default values by most likely transformation approach IV

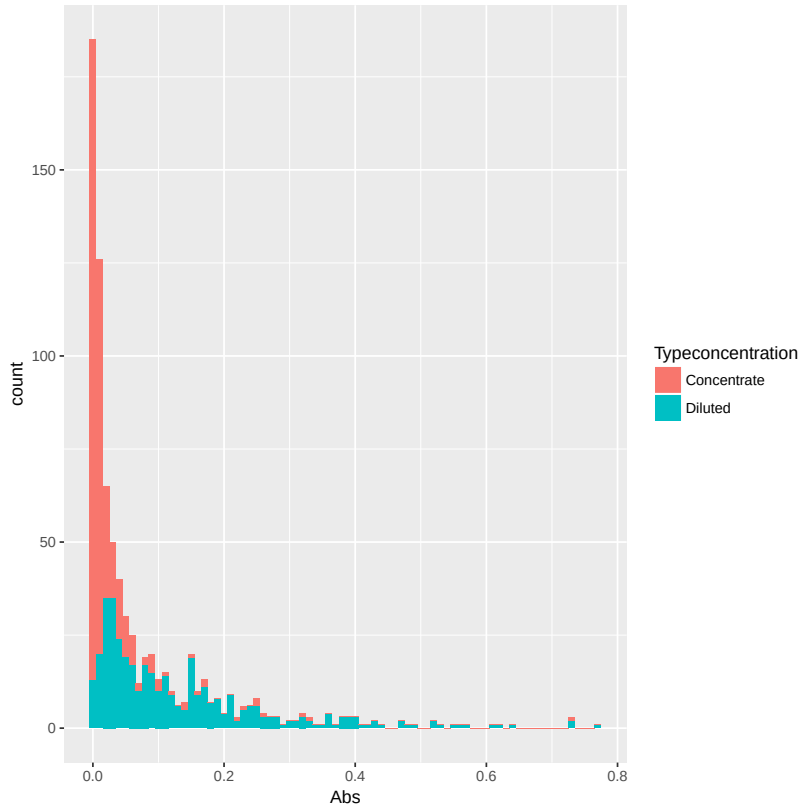


Figure: Histogram for concentrated and diluted additive shifted effect

Default values by most likely transformation approach V

- These upper limits 0, 0.3399 fits to the plotted sub-populations `"Concentrate", "Diluted`, probably the major classification in these data
- Notice, how extreme different the default values are

Proposed default values by mlt (I) - Subclassifications I

- Selection B.10 subpopulations as different populations

	Concentrated	Diluted
Organic solvent-based	0.15	0.39
Water-based	0.13	0.26
Solid	0.14	0.35
Others	0.26	0.70

- Discuss the differences, but consider different sample sizes: 303/305/132/32. Maybe for all chemicals - but not *others*- simple default value is about 0.14 for concentrated

Proposed default values by mlt (II) - Comparison to EFSA values Tab B.13 I

Main result:

	Concentrated	Diluted
Organic solvent-based and others	0.16	0.42
Water-based/ dispersed or solid	0.13	0.29

Compare

Table B13: Default values to be used in absence of experimental data

Formulation category	Concentration status	Default value (%)
Organic solvent-based[(a)] and others[(b)]	Concentrate	25
	Dilution	70
Water-based/dispersed[(c)] or solid[(d)]	Concentrate	10
	Dilution	50

Take home message I

- Estimating a prediction limit for skewed hierarchical data is not a trivial issue, particularly in complex designs
- Proposed EFSA methodology seems to be not optimal/appropriate
- Consequence: different default values
- This re-analysis should be considered as an alternative and should be compared with other approaches (e.g. beta regression using new R libraries) and published together with all parties
- All default values are conditional to the used data base containing certain chemicals
- Question: which classification is really needed for prediction?