Modeling the T2-HT2 concentration data

Rudolf Hoogenveen, 27 June 2024

Table of Contents

[1 Introduction 1](#_Toc168403700)

[2 The likelihood functions of the distributional characteristics 2](#_Toc168403701)

[2.1 Fraction of measurement values above LOQ 2](#_Toc168403702)

[2.2 Maximum value 3](#_Toc168403703)

[2.3 Distributional characteristics of measurement value above LOQ 3](#_Toc168403704)

[2.4 Median value of all measurement values above LOQ 3](#_Toc168403705)

[2.5 Mean value over all measurement values above LOQ 4](#_Toc168403706)

[2.6 Mean value over all measurement values 4](#_Toc168403707)

[2.7 Upper boundary of overall mean value (PM) 4](#_Toc168403708)

[2.8 Resulting log-likelihood function 5](#_Toc168403709)

[3 Fitting the probability distribution function to the data distributional characteristics 6](#_Toc168403710)

[3.1 Different cases of available distributional characteristics 6](#_Toc168403711)

[3.2 Results 6](#_Toc168403712)

[4 Second research question: combining these distributions into an overall distribution 9](#_Toc168403713)

[4.1 General description 9](#_Toc168403714)

[4.2 Calculation of distributional characteristics 9](#_Toc168403715)

[4.3 Results 10](#_Toc168403716)

[4.3.1 Results from current analyses, **on the original scale** 11](#_Toc168403717)

[4.3.2 Results from current analyses, **on the log-scale** 11](#_Toc168403718)

[4.3.3 Results from previous analyses, **on the original scale** 12](#_Toc168403719)

[4.3.4 Conclusion and discussion 12](#_Toc168403720)

[5 Appendix 1: results for all studies from previous analyses 12](#_Toc168403721)

# Introduction

Distributional characteristics of T2/HT2-concentrations in food products are available from literature, see the Excel-file. Examples of these characteristics are mean values, median values, maximum values, fractions with measurement values (given LOQ value), etc. These characteristics were found for different food products from different countries from different publications etc. Since these characteristics are difficult to compare (e.g. how to fit a model on both maximum and mean value?), I developed a maximum likelihood method to estimate the model parameters. The likelihood function describes the likelihood of model parameter values given the type of probability distribution and the data values. The likelihood function is proportional to the probability function. I assumed that the concentrations are log-normally distributed, i.e. the log-transformed concentrations are normally distributed. So 2 research questions arise:

(1) how to assess probability distributions for these characteristics for each study, in order to fit a lognormal distribution to each study separately (separate rows in Excel-file)

2) how to combine these distributions from several studies to an overall ‘uncertain’ probability distribution (for separate foods)

I solved the 1st research question by maximizing the likelihood function. The parameters to be estimated are the μ and σ of a normal distribution. Note these parameters are also the arguments of the R-routines related to log-normal distributions (plnorm etc.).

I interpreted the data in the following way:

- ‘levels average’ means the average of all measurement values

- ‘levels average (pos samples)’ means the average of all measurement values above LOQ

The formulas presented below are the basis for the new analyses. The previous analyses were based on erroneous formulas for the mean values, see below for a more detailed description of these errors.

# The likelihood functions of the distributional characteristics

I start with writing out the likelihood functions of several distributional characteristics. The cumulative probability and density function are written as F(x;μ,σ) and f(x; μ,σ) resp., with: x: measurement value, μ,σ: distributional parameters, n: sample size, k: # measurement values above LOQ. Note: x must be interpreted as the log-transformed concentration value here.

I write the total likelihood function as the product of separate likelihood functions, one for each distributional characteristic being presented, being defined below. The total log-likelihood function is the logarithm of the total likelihood function, and thus the sum of the separate log-likelihood functions. Estimating the parameters μ and σ means maximizing the log-likelihood function over μ and σ. Note: in the code I applied a minimization procedure, so the signs are opposite. Because of the optimization over μ and σ., I omitted all constant terms of the resulting log-likelihood function, i.e. those that do not depend on μ and σ. Note: in some cases with missing LOQ values I estimated this value, next to parameters μ and σ.

## Fraction of measurement values above LOQ

The probability that k measured values (on the original scale) are above LOQ, and thus n-k values below, is:

with: k = np, p: fraction above log-transformed LOQ (lLOQ), F(lLOQ;μ,σ): probability that the log-transformed concentration value is below the log-transformed LOQ-value, given parameters μ and σ. The log-transformed value of the likelihood function is, omitting the constants:

(n-k) log(F(lLOQ;μ,σ)) + k log(1-F(lLOQ;μ,σ)) = n(1-p) log( F(lLOQ;μ,σ) ) + np log( 1 - F(lLOQ;μ,σ) )

## Maximum value

The probability that all values do not exceed the reported maximum value is the product of each value not exceeding the maximum value:

with: X: log-transformed reported maximum value. The probability density function of X being the measured maximum value is the 1st order derivative of the cumulative probability function:

The log-transformed value of the likelihood function is, omitting the constants:

(n-1) log( F(X; μ,σ) ) + log( f(X; μ,σ) )

## Distributional characteristics of measurement value above LOQ

If the log-transformed measurement value is N(μ,σ) distributed, then the conditional probability density function of a log-transformed measurement value above log-transformed LOQ is:

g(x; μ,σ) = f(x|x>lLOQ; μ,σ) = f(x;μ,σ) / ( 1 – F(lLOQ; μ,σ) )

The distributional characteristics μ’ and σ’2 (on the log-scale) of this conditional distribution are:

μ’ = E( x|x>lLOQ; μ,σ ) = = ,

E( x2|x>lLOQ; μ,σ ) =

-> σ’2 = V( x|x>lLOQ; μ,σ ) = E( x2|x>lLOQ; μ,σ ) - E( x|x>lLOQ; μ,σ )2

The related cumulative distribution function is written as G( x;μ,σ ) to be read as G( x; μ’(μ,σ), σ’(μ,σ) ). I assumed that the log-transformed values above lLOQ are also normally distributed.

## Median value of all measurement values above LOQ

Given the distribution G of all measurement values above lLOQ and the total number of measured values, the probability that half of the values are below and half of the values are above X is:

with: X: log-transformed reported median value. The log-transformed value of the likelihood function is, omitting the constants:

½k log( G(X; μ,σ) ) + ½k log( 1 - G(X; μ,σ) ) = ½np log( G(X; μ,σ) ) + ½np log( 1 - G(X; μ,σ) )

## Mean value over all measurement values above LOQ

The distributional characteristics of any value above LOQ are:

mean μ’’ = ∫≥lLOQ exp(x) f(x;μ,σ) dx / ( 1 – F(lLOQ; μ,σ) )

variance σ’’2 = ∫≥lLOQ exp(x)2 f(x;μ,σ) dx / ( 1 – F(lLOQ; μ,σ) ) – (μ’’)2

Following statistical theory, the mean value of a sufficiently large number of values above LOQ can be assumed normally distributed, with characteristics:

mean μ’’

variance σ’’2 / k

Then the likelihood function has the form of the probability density function of a normal distribution. The log-transformed value of the likelihood function is, omitting the constants, with X the reported untransformed mean value:

## Mean value over all measurement values

See also the previous section §2.5. The distributional characteristics of any value are:

mean μ’’’ = ∫ exp(x) f(x;μ,σ) dx

variance σ’’’2 = ∫ exp(x)2 f(x;μ,σ) dx – (μ’’’)2

Again, assuming the overall mean value being normally distributed, the distributional characteristics of this overall mean value are:

mean μ’’’

variance σ’’’2 / n

Then the log-transformed value of the loglikelihood function is, omitting the constants, with X the reported untransformed overall mean value:

Note: in case of <LOQ values the reported mean value depends on the proportion <LOQ, and the way the authors dealt with these <LOQ values: assumed being 0, or imputed by extrapolating from known values. My calculation method assumed the latter method.

Note: in the previous analyses I assumed that the expected value of the reported mean concentration value equals parameter μ, and that the variance equals parameter σ2/n. Which is wrong. As could be expected, the main effects were on the calculated values of parameter σ.

## Mean value lower than LOQ

See §2.6, the probability becomes:

F(LOQ; μ’’’, σ’’’ / √n)

## Upper boundary of overall mean value (PM)

Since I could not interpret this value well at second thoughts, I leave this section as PM, meaning not applied in the new analyses. Note: I did use this characteristics in the old analyses.

The overall mean value is assumed normally distributed with characteristics (see previous section):

μ’’’ and σ’’’2

The probability of the mean value being below the reported upper boundary is:

F( UB0; )

with: UB0: untransformed upper boundary value. The log-transformed value of the likelihood function is:

log( F( UB0; ) )

## Resulting log-likelihood function

The resulting log-likelihood function is the sum of the log-likelihood functions for all distributional characteristics being reported:

n(1-p) log( F(LOQ;μ,σ) ) + np log( 1 - F(LOQ;μ,σ) ) # measurement values > LOQ

(n-1) log( F(X; μ,σ) ) + log( f(X; μ,σ) ) maximum measured value

½np log( G(Xmed; μ,σ) ) + ½np log( 1 - G(Xmed; μ,σ) ) median of all values > LOQ

+ mean of all values > LOQ

mean of all values

This log-likelihood function is maximized to assess the optimal values of μ,σ for each study (row of the Excel-file) separately. Some notes:

(1) in practice not all distributional characteristics are reported, so terms may not be included when fitting specific studies (rows of Excel file)

(2) the distributional characteristics being reported may not be independent, so I am not sure whether the resulting log-likelihood is indeed a formal log-likelihood; I think it is not, but the effects are probably small

(3) since I got some problems when optimizing the log-likelihood function, I also included an alternative optimization procedure, i.e. minimizing the absolute relative differences; however, the problems turned out to result from some implementation errors, so I indeed applied the log-likelihood procedure, but still included the relative difference procedure in the code (so far, removal to be done)

(4) note that most, but not all, terms are linear w.r.t. the sample size n

# Fitting the probability distribution function to the data distributional characteristics

## Different cases of available distributional characteristics

I distinguished several cases of sets of distributional characteristics available (known) that enabled fitting the probability distribution function to the data. Case 5 was not applied in the new analyses. The numbers refer to the case numbers in the code.

1 overall mean level, max level, # pos cases, LOQ

2 mean level > LOQ, max level, # pos cases, LOQ

3 median level, max level, # pos cases, LOQ

4 mean level overall and > LOQ, max level, # pos cases, LOQ

6 mean level > LOQ, max level, # pos cases, unknown LOQ

7 overall mean level, max level, # pos cases, unknown LOQ

8 mean level overall and > LOQ, max level, # pos cases, unknown LOQ

For each case I selected the studies (rows) that matched the selection criterium, and fitted the log-normal distribution function to the data by maximizing the log-likelihood function (see §3.1) using the r-routine optim. The variances of the parameter estimations were calculated using the Hessian matrix, following the relation:

where the matrix with 2nd order derivatives is the Hessian matrix, and f is the log-likelihood function. I applied the (inverse of the ) variances as weights when combining the results (see §4).

Some numerical details related to applying the R-routine optim:

- to enforce the parameter σ being positive, I used the parameter log(σ) as the parameter to be estimated; afterwards I applied a back-transformation.

- to prevent taking the logarithm of argument value zero, I took the maximum of the value and a very small constant, whenever necessary.

## Results

I present the results of the new analyses in the table below. The results of the previous analyses are presented in an appendix. The index values refer to the rows of the excel file. The other columns are:

- mu and sigma are the calculated mean and sigma parameter of the log-normal distribution

- the standard error (square-root of variance) of these parameters (see §3.1)

- poscase are the data and estimated fraction of above LOQ values

- mean are the data and estimated mean values over all measurement values

- meanpos are the data and estimated mean values over all above LOQ measurement values

- median are the data and estimated median values over all measurement values

After fitting the log-normal distribution estimations of all distributional characteristics are available, although data values may be missing (NA).

*Table with calculated parameter mu and sigma values of log-normal distribution and the calculated and data distributional characteristics for all studies (rows in Excel file)*

A white grid with many small dots

Description automatically generated with medium confidence

Comparing the new results to those from the previous analyses showed some differences. The main differences are:

- although the estimated parameter values generally did not change that much, the confidence intervals are generally smaller; the reason is that the probability distributions of the different characteristics are now consistent instead of inconsistent

- the largest differences are found for rye; the reason is that for most rows with rye the data values are reported mean concentration values

# Second research question: combining these distributions into an overall distribution

## General description

The 2nd research question is to combine these distributions from several studies to an overall ‘uncertain’ probability distribution. This 2nd step was applied for T2 and HT2, and each raw and processed food (group) separately, and consisted of two sub-steps. The 1st sub-step was to select the studies (rows) for T2/HT2 and food (group) with valid model fitting results. I defined valid as the relative difference between all predicted and reported distributional characteristic values not being larger than a given upper boundary percentage (50%). The 2nd sub-step was to fit a linear regression model on the calculated values for the selected studies (rows) for the parameters μ (mean) and σ (sigma) separately. I applied the R-routines lm in case of studies from 1 country only (fixed study-effect), or lme in case of studies from more than 1 country (random effect). In the latter case the countries are included as random effect in the linear model. I included the inverse of the variances of the parameter estimations as weights to take account of the strength of the model fit for each study (row of Excel-file) included.

## Calculation of distributional characteristics

Based on these results I calculated the distributional characteristics of the concentrations of T2/HT2 in foods for which data were available on the original scale. Note that two issues of uncertainty play a role here: heterogeneity and parameter uncertainty. Heterogeneity means that the concentrations are not fixed, but are (as we assumed) log-normally distributed with parameters μ (mean) and σ (sigma). Parameter uncertainty means that these parameters μ and σ have no fixed value, but instead are stochastic (normally distributed), each of them with their own mean and sigma parameter.

I fitted a linear regression model for the parameters μ and σ separately. However, when fitting the log-normal distribution for each study (row of Excel-file) separately, the parameter estimations turned out to be negatively correlated: the uncertainty of the parameter μ is negatively correlated with the uncertainty of the parameter σ. I calculated the median value of all calculated correlation values, and applied this correlation when calculating the uncertainty of the concentrations on the original scale.

The man and sigma of concentration in food on the original scale are calculated in the following way:

μ' = exp( μ + ½ σ2 )

σ’ = μ' √( exp(σ2) – 1 ) = exp( μ + ½ σ2 ) ( exp(σ2) – 1 )1/2

with: μ, σ: the estimated (fixed) coefficients resulting from fitting the lm (lme) model to all study (row)-specific values being available. The variance of the parameter μ' due to parameter uncertainty was calculated by applying the so-called delta-rule (1st order Taylor-series expansion):

V(μ') ≈ ( δμ'/δμ )2 V(μ) + ( δμ'/δσ )2 V(σ) + 2 δμ'/δμ δμ'/δσ cov(μ,σ)

V(σ') ≈ ( δσ'/δμ )2 V(μ) + ( δσ'/δσ )2 V(σ) + 2 δσ'/δμ δσ'/δσ cov(μ,σ)

with: δμ'/δμ = exp( μ + ½ σ2 ) = μ'

δμ'/δσ = σ exp( μ + ½ σ2 ) = σ μ'

δσ'/δμ = exp( μ + ½ σ2 ) ( exp(σ2) – 1 )1/2 = σ'

δσ'/δσ = σ exp( μ + ½ σ2 ) ( exp(σ2) – 1 )1/2 + exp( μ + ½ σ2 ) ½ ( exp(σ2) – 1 )-1/2 2 σ exp(σ2) =

σ σ’ + σ’ σ exp(σ2) ( exp(σ2) – 1 )-1 = σ σ’ + σ’ σ ( 1 + ( exp(σ2) – 1 )-1 ) =

σ σ’ ( 2 + ( exp(σ2) – 1 )-1 )

V(μ), V(σ) result from fitting the linear model to all selected rows

The term cov(μ,σ) describes the co-variance between μ and σ, and can be written as:

cov(μ,σ) = corr(μ,σ) √V(μ) √V(σ)

with: corr(μ,σ) the median correlation value over all correlation values that resulted from

fitting the log-normal distribution to the data for each row in the Excel-file

Thus:

V(μ') ≈ μ'2 V(μ) + σ2 μ'2 V(σ) + 2 σ μ'2 corr(μ,σ) √V(μ) √V(σ) =

μ'2 ( V(μ) + σ2 V(σ) + 2 σ corr(μ,σ) √V(μ) √V(σ) )

V(σ') ≈ σ’2 V(μ) + σ2 σ’2 ( 2 + ( exp(σ2) – 1 )-1 )2 V(σ) +

2 σ’ σ σ’ ( 2 + ( exp(σ2) – 1 )-1 ) corr(μ,σ) √V(μ) √V(σ) =

σ’2 V(μ) + σ2 σ’2 ( 2 + ( exp(σ2) – 1 )-1 )2 V(σ) +

2 σ σ’2 ( 2 + ( exp(σ2) – 1 )-1 ) corr(μ,σ) √V(μ) √V(σ)

## Results

In the table below the results are presented of fitting the linear (mixed effects) model on the study(row)-specific data for each raw/processed food with data available.

The columns have the following interpretation:

- substance

- food, raw or processed

- mean, sigma: the calculated population-mean parameter values on the original scale

- se(mean), se(sigma): the values that describe the variability of the parameter value over different countries I terms of standard deviation: if a sigma-value 0 is presented, there was only 1 study available, and the variability could not be assessed

### Results from current analyses, **on the original scale**

A table with numbers and text

Description automatically generated

### Results from current analyses, **on the log-scale**

A table with text and numbers

Description automatically generated

Notes:

1 in case only 1 study was available, no confidence bounds could be calculated, resulting in NaN in this table, and value 0 in previous table

2 the correlation between calculated parameter μ and σ values was not applied in this table

### Results from previous analyses, **on the original scale**

A table with numbers and text

Description automatically generated

### Conclusion and discussion

As written before, the differences are not that big, except for rye. Again, for rye most distributional characteristics being reported are mean values, and here the methodology has changed. For most other products the differences are much smaller. The differences in terms of number of different countries with estimated values available can be explained in two ways: (1) I did not include data (rows in Excel file) for case 5 in the new analyses (although the number of studies (rows) is relatively small), (2) results for specific studies (rows) could be accepted in the new analyses and not in the previous analyses, or vice versa.

The inclusion of the correlation between estimated values for parameter μ and σ (on the log-scale) had only small effects on the calculated mean and sigma on the original scale.

# Appendix 1: results for all studies from previous analyses

*Table with calculated parameter mu and sigma values of log-normal distribution and the calculated and data distributional characteristics for all studies (rows in Excel file)*

A grid of lines with small dots

Description automatically generated with medium confidence