Fortnightly meeting November 19th 2019

Mark Nelms

19/11/2019

Contents

Questions from last time

After our last meeting we had some questions to answer/actions to perform:

- What is the variability of LD50 values in -$\log_{10}$(mol/kg)?
  - Replicate what Agnes did in $\log_{10}${mg/kg}
- Are there chemotypes that seem to consistently underestimate LD50?

1 Acute toxicity

1.1 What is the variability of LD50 values in -$\log_{10}$(mol/kg)?

To be able to calculate the variability in -$\log_{10}$(mol/kg) the first thing I had to do was convert the LD50 from $\log_{10}$(mg/kg) to -$\log_{10}$(mol/kg).

First, I associated the MWs from the CompTox Dashboard to the right chemical. As we already know we, unfortunately, don't have a MW in the Dashboard for every chemical in the Acute Toxicity Working Group (ATWG) list; so I can't fully "replicate" Agnes' work and get a sense of the variability of the LD50s across the chemicals with $\geq$ 3 LD50 values[1].

However, I can calculate the variability for the subset of chemicals that have $\geq$ 3 LD50 values and for which we have MWs. Because this subset of chemicals is also the set we were able to retrieve QSAR-ready structures from the Dashboard, it means our calculation of variability will be representative for our subset of chemicals that we ran through both TEST and TIMES.

1.1.1 Spread of standard deviations across dataset

Figure 1, shows the spread of the standard deviations (in log~10[mg/kg]) for each chemical with $\geq$ 3 LD50s values. The overall standard deviation does not increase by much when we only take into consideration those chemicals for which we have MWs (i.e. the chemicals we can convert to -log~10[mol/kg]). This is good news because it means that the variation in the subset of data we are using doesn't differ too much from the total set.

[1] Agnes' slides says the ATWG list contains 1,120 chemicals with $\geq$ 3 LD50s, I get 1,209 chemicals - the global standard deviation is basically the same (0.8299 $\log_{10}$[mg/kg])

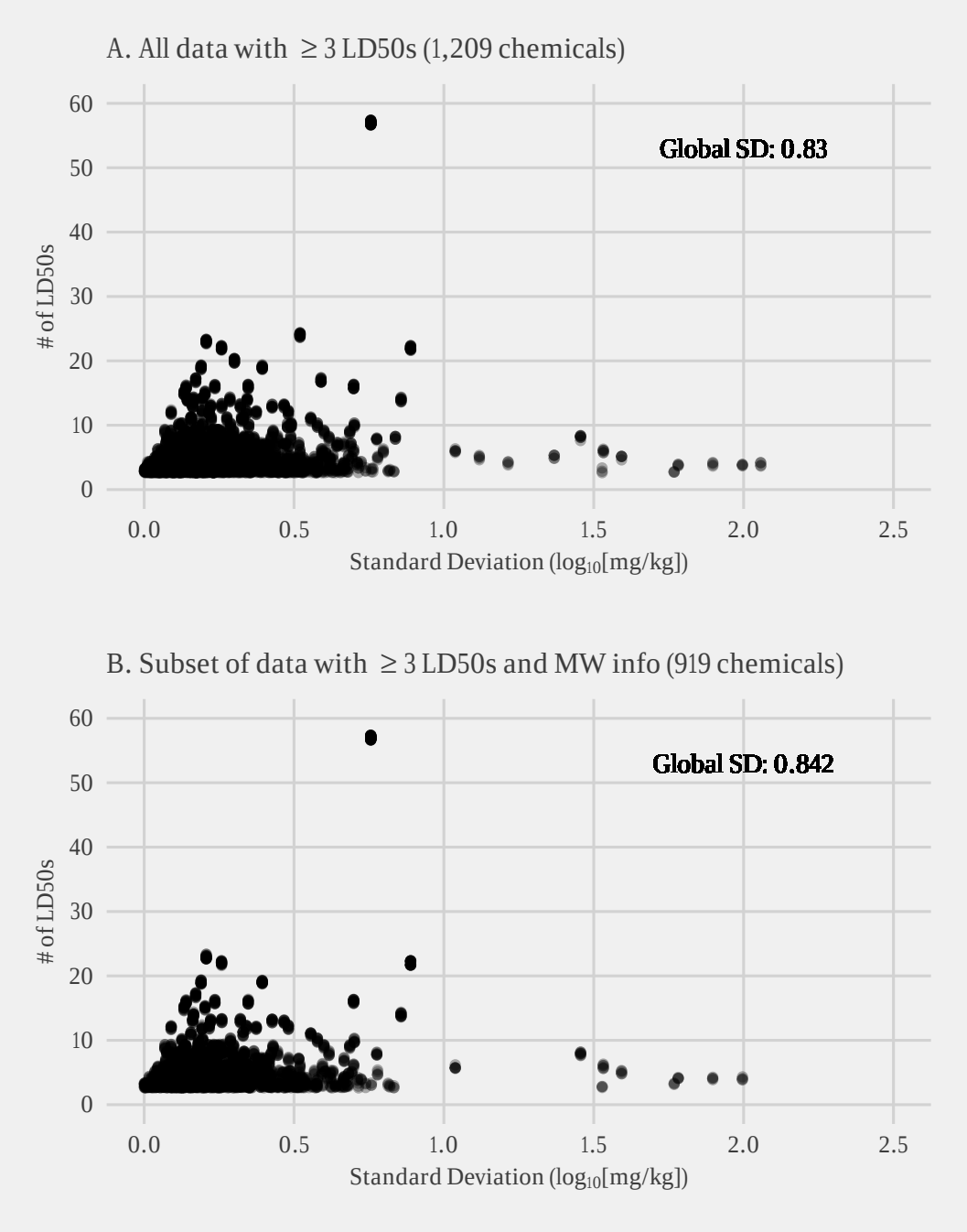

Figure 1: Standard deviations against number of LD50s (in log(mg/kg)

By comparing the global standard deviations for the subset of data with MWs, we can see that the standard deviation is smaller when calculated using the $\log_{10}$(mg/kg) values (Figure 1B.) than when we use the -$\log_{10}$(mol/kg) values (Figure 2. Although, there isn't too much of a difference between the two.

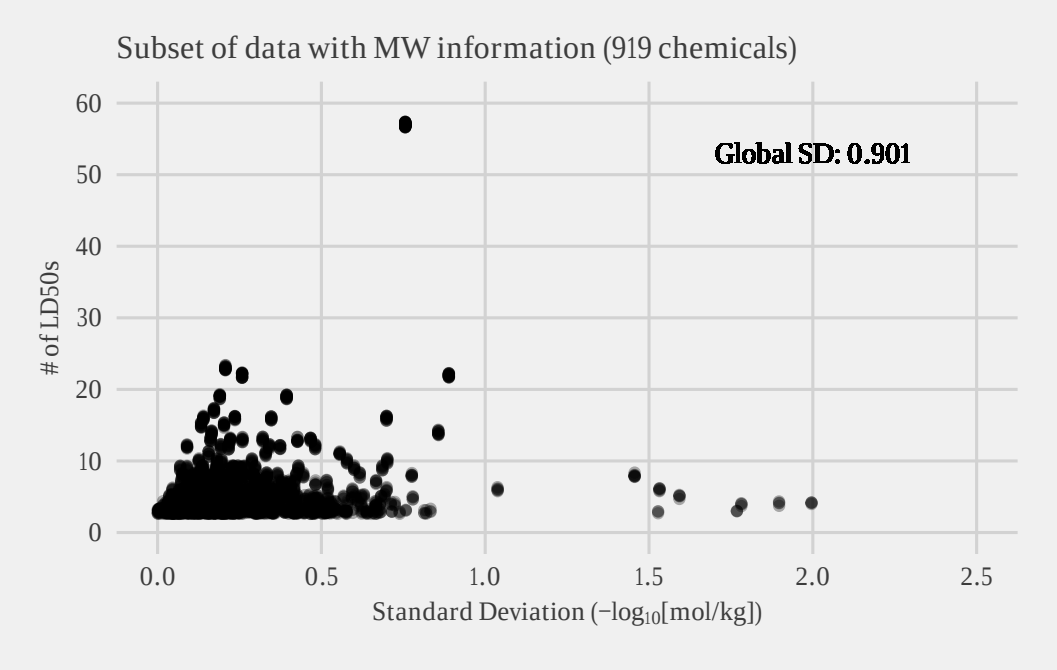

Figure 2: Standard deviations against number of LD50s (in -log[mol/kg])

1.1.2 Identification of 95% confidence intervals

After looking at the overall standard deviations, and seeing there wasn't a difference between the complete dataset and the subset of chemicals we have MW information for, I followed Agnes' workflow to calculate the universal 95% confidence interval. This involved:

- Filtering for chemicals with ≥ 3 LD50s (in -log[mol/kg]),
- Computing the standard deviation across the LD50s, on a per chemical basis,
- Generating 10,000 bootstrap resamples,
- Calculating the standard deviation for each bootstrap resample,
- Using the `quantile` function to calculate the 95% confidence intervals
  - I also calculated the CIs by multiplying the mean standard deviation across the 10,000 resamples by 1.96 as a comparison

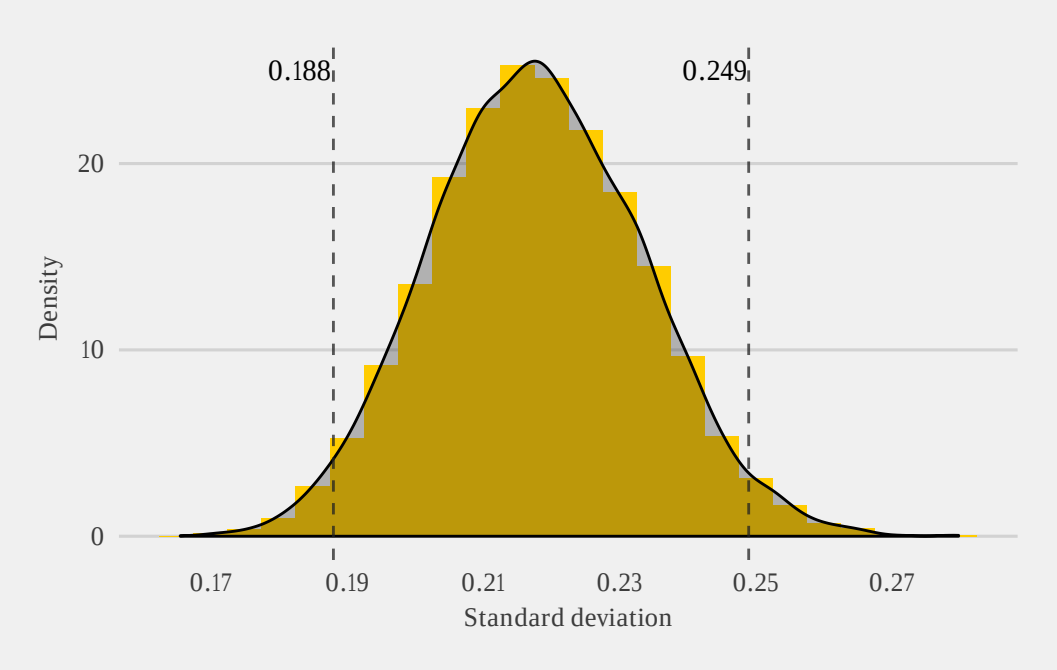

Figure 3: Standard deviations of 10,000 bootstrap resamples (dashed lines are 95% confidence intervals)

Above, we can see that the standard deviations for the 10,000 bootstrap resamples are normally distributed (Figure 3). This is what we'd expect, but it also means we are justified being able to use these data to calculate the 95% confidence intervals (due to the central limit theory).

Figure 4, below, shows a comparison of the 95% confidence intervals calculated using the two methods, either: 1) using the `quantile` function (in red), or 2) multiplying the mean standard deviation by 1.96 (in blue). As you can see, using the `quantile` function generated narrower 95% confidence intervals than multiplying the mean standard deviation by 1.96.

The "benefit" of the second method is that it provides a single number we can reference when talking about the 95% confidence interval (i.e. ±0.427 -$\log_{10}$[mol/kg]). This value is slightly smaller than the default value of ±0.5 -$\log_{10}$(mol/kg) I used for our last meeting. It's also slightly larger than what Agnes originally calculated (i.e. ±0.31 $\log_{10}$[mg/kg]). However, as we have more chemicals with $\geq$ 3 LD50 values, our data seems to be slightly different from the data Agnes used[2].

[2]Our bootstrapping approach is also slightly different: Agnes took 1,000,000 resamples (with replacement) of the standard deviations and used this to calculate the confidence intervals; whereas, I used 10,000 bootstrap resamples for my calculation

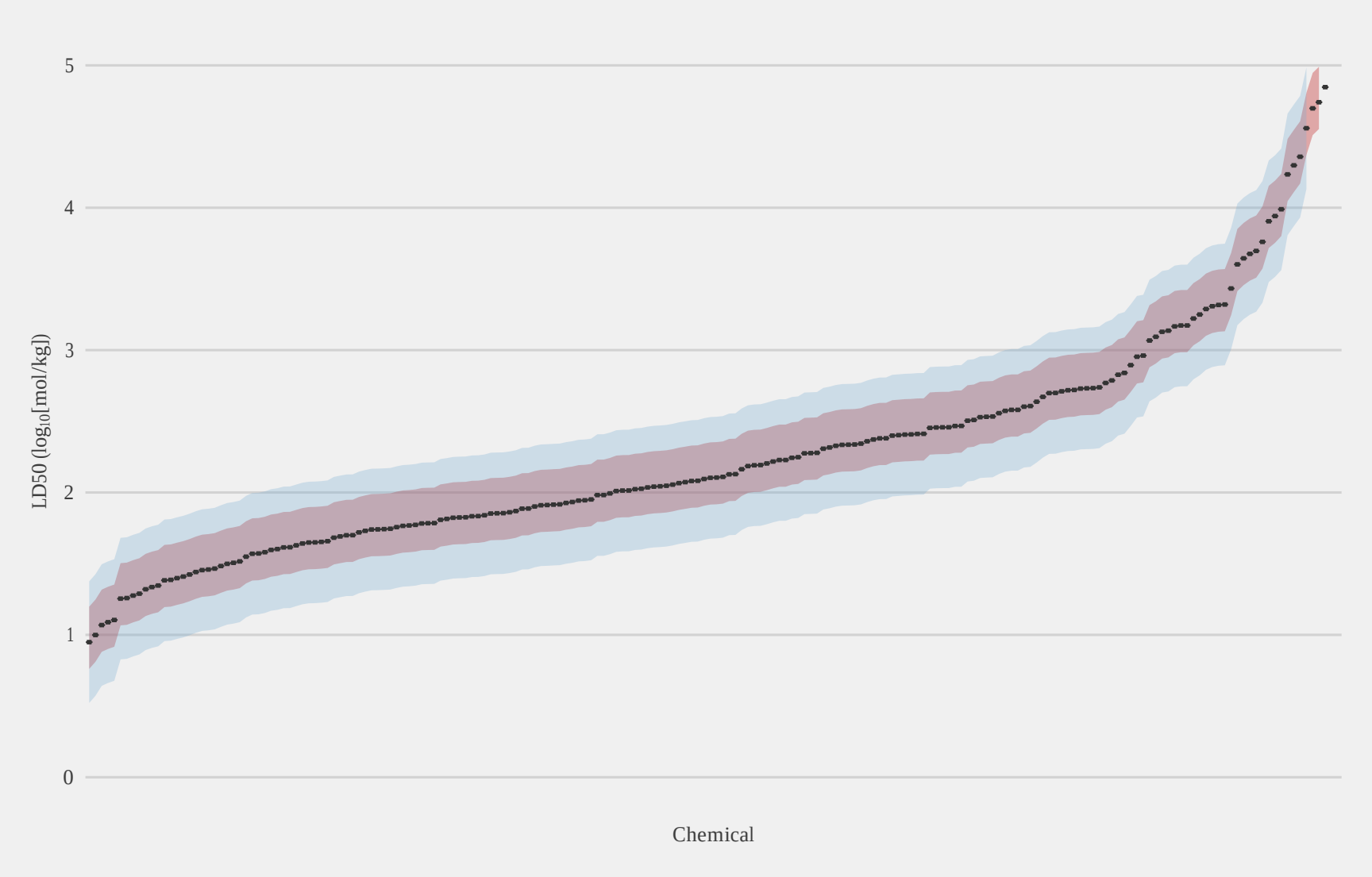

Figure 4: Comparison of methods for calculating 95% confidence intervals (200 chemicals ordered by median LD50)

Figure 5, contains boxplots of the LD50 values for 200 arbitrarily chosen chemicals with ≥ 3 LD50s. These boxplots are overlaid with the 95% confidence interval (in red), calculated using the `quantile` function, surrounding the median LD50 value for each chemical.

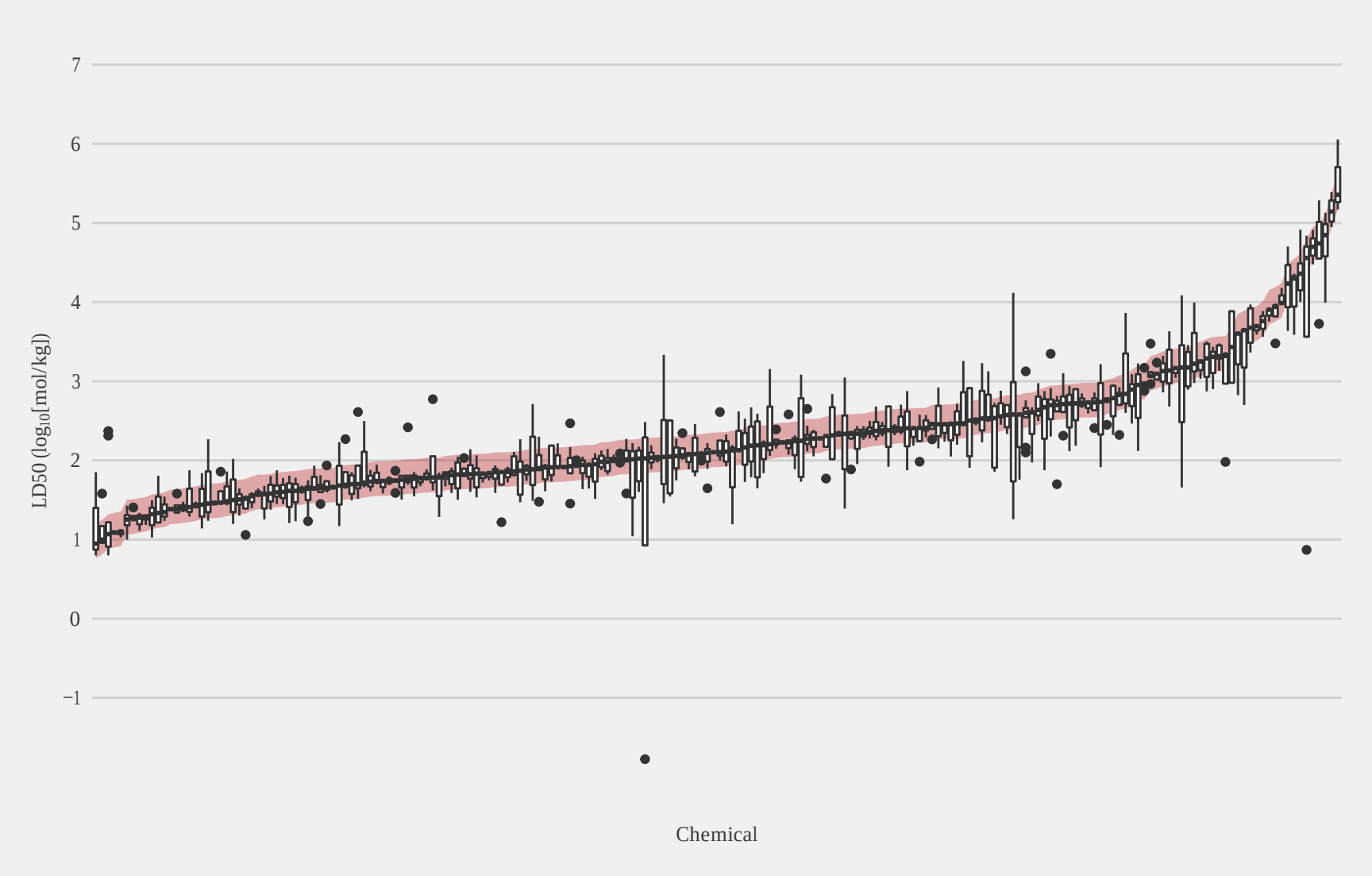

Figure 5: Boxplot of LD50s with 95% CI (in red) for 200 chemicals (ordered by median LD50)

1.2 Are there chemotypes that seem to consistently underestimate LD50?

To begin doing this I had to convert the TEST predictions (in -$\log_{10}$[mol/kg]) into standard mg/kg units; as I've already done this before it was relatively straight forward. I've included the equation below just so we don't have to search back through old meeting notes to find them.

$$LD50(mg/kg) = 10^{-log10(mol/kg)} \times MW \times 1000 \quad (1)$$

Next, I used the LD50s (in mg/kg) to assign each chemical with a prediction to the appropriate EPA category:

- Class I: $\leq$ 50mg/kg
- Class II: > 50mg/kg and $\leq$ 500mg/kg
- Class III: > 500mg/kg and $\leq$ 5000mg/kg
- Class IV: > 5000mg/kg