Establishing upper reference limits

For left-censored and contaminated data

Niels Henrik Bruun

Unit of Clinical Biostatistics, AaUH

Section 1

Introduction

In collaboration with

- **Stine Linding Andersen**,
  Department of Clinical Biochemistry, Aalborg University Hospital, Aalborg, Denmark,
  Department of Clinical Medicine, Aalborg University, Aalborg, Denmark
- **Nanna Maria Uldall Torp**,
  Department of Clinical Biochemistry, Aalborg University Hospital, Aalborg, Denmark,
  Department of Clinical Medicine, Aalborg University, Aalborg, Denmark
- **Peter Astrup Christensen**,
  Department of Clinical Biochemistry, Aalborg University Hospital, Aalborg, Denmark,
  Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

The initiating case

1. How do you find upper an reference limit when more than 50% of data has values at a lower limit of detection, i.e., are non-detectable?
   - The classical methods if working lead to too high upper reference limit
   - The use of limits of detection (LOD) is not recommended, Hewett and Ganser (2007) and Helsel (2010)
2. What if the data is also contaminated the right by a distribution of extreme values?
3. What if data isn't normal?

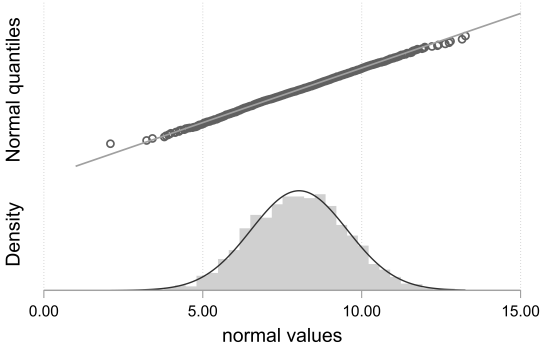
This is a work in progress!!

A stata command is to appear soon on the SSC: -ssc install ros-

Normal quantile plots

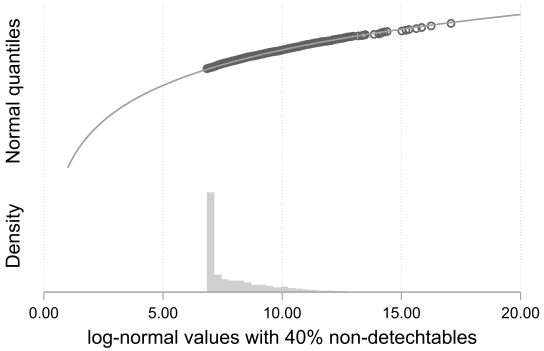
- Drawing normal quantiles against observations
  - If the oberservations are normally distributed, the curve is linear
  - The slope of the line is $1/\sigma$
  - The intercept is μ/σ
- One could do regressions instead of drawing
- The regression goodness of fit is measured by R^2_{adj}
  - number of parameters are always 2
  - AIC and BIC behaves strange with Box-Cox transformations
- The Box-Cox transformation *bct* used is:

$$bct(x;\theta) = \begin{cases} x^{\theta}/\theta & if\ \theta \neq 0 \\ log(x) & if\ \theta = 0 \end{cases}$$



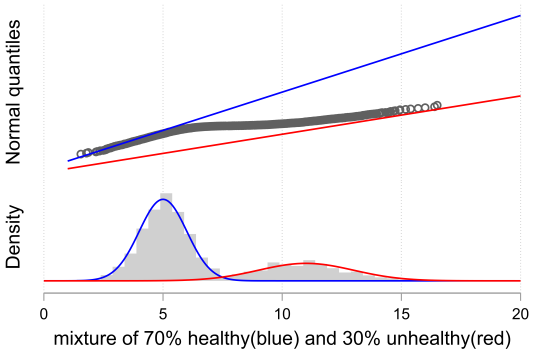
Non-detectable data, environmental statistics

- The regression on order statistics (ROS) is recommended if there is a high proportion of non-detectables, see Helsel (2010), Huston2009, and Hewett and Ganser (2007)
- The principle of ROS is that log-normal quantile plots are extrapolated to get quantiles
  - The curvature on the graph indicates the need for a log transformation of the data



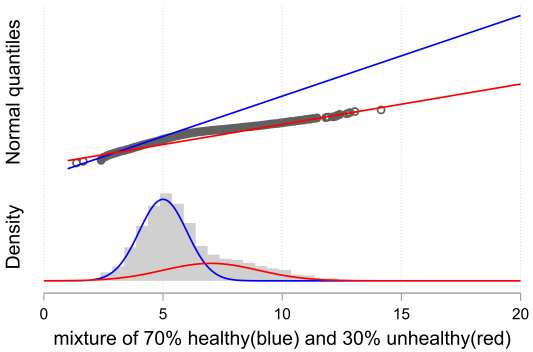
Contaminated (mixed) data, laboratory medicine, clinical biochemistry

- The Hoffmann method, Hoffmann (1963) and Jensen et al. (2006), uses normal quantile plots to separate the "healthy" from the "unhealthy".
- Non-detectables (in a large scale) are not considered.
- The log-transformation or the Box-Cox transformations are often used



When the mixed distributions are close

- the separation of the mixed parts still works
- A Box-Cox transformation might be equally good



The command -ros-

- A command for finding upper reference bounds when the data has non-detechtable values and possibly are contaminated to the right by an extreme value distribution
- Contamination of data is visually identified from a normal quantile plot
- Data are assumed to be from a Box-Cox distribution
  - Data is normally distributed after a Box-Cox transformation
- Optimal Box-Cox transformation is chosen by selecting a theta with a high adjusted R squared
  - The Box-Cox command gives biased estimates when there are non-detectables
  - Always only two parameters in the regressions, so adjusted R squared is an acceptable measure
  - AIC and BIC does not work with the current Box-Cox transformation formulas
- Regressing (ROS) the observed values on the emperical (normal) z-values
  - Estimation of the mean (the intercept) and standard deviation (the slope)
  - The mean and standard deviation are used to estimate the quantiles

Real life applications

- The -ros- command has been applied in Andersen et al. (2022) and Uldall Torp et al. (2022)
  - for finding upper reference bounds
  - Follow-up in Danish nationwide registers
  - Empirical quantile method lead to too high upper reference bounds

Section 2

The savona example dataset, Huston and Juarez-Colunga (2009)

Description

- Provided by the British Columbia Ministry of Environment
- Information on orthophosphate concentrations taken at the Thompson River from Savona
- Contains 32 oberservations of four variables:
  - The date of the measurement taken
  - Indication on whether a measurement is below the detection limit
  - Concentration is the level of orthophosphate observed
  - Censored also indicates whether an observation is below the detection limit or not

The -ros- command and output

Some of the options for -ros-

- **censor** A variable indicating whether a value is censored (1) or not (0)
- **scatter** Generates a qnorm scatterplot as model control
- **rsqrtheta** Generates a line plot of adjusted R squares by thetas (Box-Cox transformation)
- **theta(#)** for the choice of Box-Cox transformation, default = 1 (No transformation)

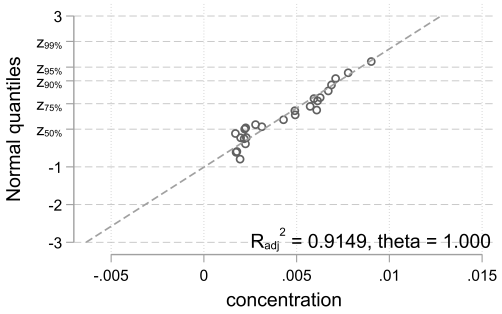
```
use concentration censored using "savona (NADA).dta", clear
```

```
ros concentration, censor(censored) rsqrtheta scatter
Adjusted Rsquared is 0.9149
Percentiles
  50.00%    0.0032
  75.00%    0.0054
  90.00%    0.0073
  95.00%    0.0085
  99.00%    0.0106
```

The -ros- diagnostics, the scatter plot

Look for contamination and/or transformation

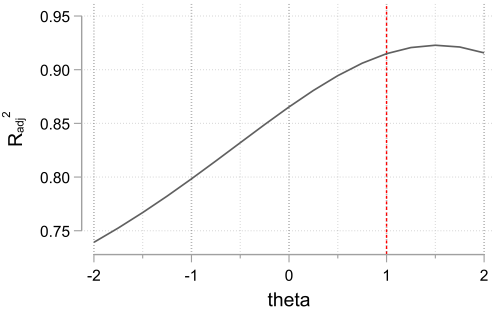
- Works for a small sample, N = 32
- Contamination is (probably) not detectable
- No transformation seems necessary



The -ros- diagnostics, the R2 adj vs theta plot

Look for an optimal transformation

- Optimal transformation is around 1.5, $R^2_{adj} = 0.923$
- Little gain from the chosen transformation 1, $R^2_{adj} = 0.915$
- The log transformation (default in Helsel (2010) and Hewett and Ganser (2007)) is relatively bad, $R^2_{adj} = 0.865$



Section 3

The TGAb eaxmple dataset

Description

- A random subsample of 1000 TGAb values from the North Denmark region pregnancy cohort
  - variable *tgab*
- Non-detectables are set at 7 (59.4%)
  - variable *tgab_c*

The -ros- command and output

- Data are assumed to be contaminated when TGAb > 20
  - see why at next slide
- Compare emperical percentiles (from -sumat-) with -ros- percentiles
- The 75% percentiles are the almost same
  - See the yellow line on the next slide
- The 95% percentile differ from 33 (-ros-) to 140 (empirical)

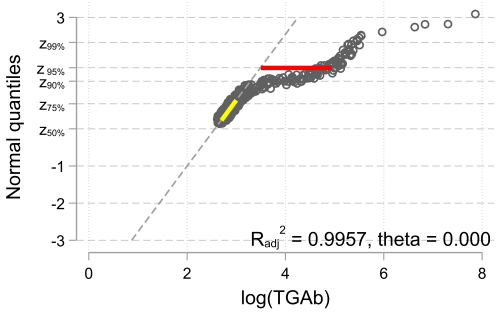
```
sumat tgab, statistics(p50 p75 p90 p95 p99)

------------------------------------------------
              p50    p75    p90     p95     p99
------------------------------------------------
TGAb(IU/ml)  7.00  19.10  50.60  140.18  233.54
------------------------------------------------
```

```
ros tgab if tgab < 20, censor(tgab_c) scatter ///
  rsqrtheta theta(0)
Adjusted Rsquared is 0.9957
Percentiles
  50.00%   12.9897
  75.00%   19.0034
  90.00%   26.7636
  95.00%   32.8506
  99.00%   48.2492
```

The -ros- diagnostics, the scatter plot

Look for contamination and/or transformation

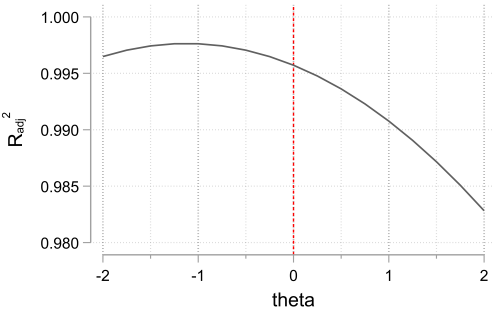
- Data are assumed to be contaminated when TGAb > 20
  - straight line only present at TGAb < 20
- **The distribution of the contaminated is ignored (ignore extreme values??)**
- Percentiles are extrapolated from around 160 (out of 1000) measurements (the yellow line)
- The red line is the change in 95% upper bound for TGAb from 140 UI/ml to 33 UI/ml
- At the yellow line the ROS upper bound estimates are the same as the emperical upper bound estimates



The -ros- diagnostics, the R2 adj vs theta plot

Look for an optimal transformation

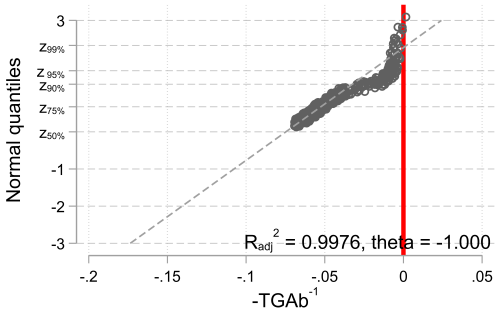
- Data are contaminated when TGAb > 20
- Optimal Box-Cox transformation is around -1, $R^2_{adj} = 0.998$
  - negative thetas does not behave well
  - negative value for the 99% upper bound
  - negative thetas implies negative values
  - asymptotic vertical behavior for values close to zero
- Little gain from the chosen log transformation (theta $= 0$), $R^2_{adj} = 0.996$
- No transformation (theta $= 1$) would be acceptable too, $R^2_{adj} = 0.991$



Limitations of the Box-Cox transformation

- The transformed values becomes negative with the negative thetas
  - can not cross the red line (jitter is used here)
  - the predictions can giving opposite signs

```
ros tgab if tgab < 20, censor(tgab_c) scatter theta(-1)
Adjusted Rsquared is 0.9976
Percentiles
  50.00%    13.3796
  75.00%    19.0435
  90.00%    30.7648
  95.00%    48.7057
  99.00%   -5.2e+02
```



Section 4

The final

Summary

We combine two classical graphical methods into a simple one

- to identify and handle possible contamination in data
- to identify a Box-Cox transformation to yield a better fit
  - and a better estimation
- to estimate upper bounds for reference intervals
  - to estimate the mean and standard deviations are means to an end, not the goal
- most times external validation of the bounds is necessary
- using simple regression estimates instead of using quantile plots
- "All models are wrong, but some are useful", so there might be more than one acceptable solution
  - in most cases with similar bounds
- This approach can handle mixtures of data with different limits of detection

TO DO

- How to predict confidence intervals for the estimated upper bounds
- To explore GLS estimation for the location-scale models, Rinne (2010)
- To explore the relation to the truncated finite mixed models (FMM)
- To explore the relation to the Tobit regression
- To explore the relation to parametric quantile models, Bottai (2021)
- To find a better class of distributions than the Box-Cox distributions
  - A challenge is negative data values and negative exponentiation
- Investigate whether regression modeling be done in this setup

references I

Andersen, Stine Linding, Niels Henrik Bruun, Peter Astrup Christensen, Simon Lykkeboe, Aase Handberg, Annebirthe Bo Hansen, Maja Hjelm Lundgaard, et al. 2022. "Cut-Offs for Thyroid Peroxidase and Thyroglobulin Antibodies in Early Pregnancy." *European Thyroid Journal* 1 (aop).

Bottai, Matteo. 2021. "Understanding and Estimating Conditional Parametric Quantile Models." https://www.stata.com/symposiums/biostatistics-and-epidemiology21/slides/Bio21_Bottai.pdf.

Helsel, Dennis. 2010. "Much Ado About Next to Nothing: Incorporating Nondetects in Science." *The Annals of Occupational Hygiene* 54 (April): 257–62. https://doi.org/10.1093/annhyg/mep092.

Hewett, Paul, and Gary Ganser. 2007. "A Comparison of Several Methods for Analyzing Censored Data." *The Annals of Occupational Hygiene* 51 (November): 611–32. https://doi.org/10.1093/annhyg/mem045.

Hoffmann, Robert G. 1963. "Statistics in the Practice of Medicine." *JAMA* 185 (11): 864–73. https://doi.org/10.1001/jama.1963.03060110068020.

Huston, C., and E. Juarez-Colunga. 2009. "Guidelines for Computing Summary Statistics for Data-Sets Containing Non-Detects." https://bvcentre.ca/files/research_reports/08-03GuidanceDocument.pdf.

Jensen, Esther A., Per Hyltoft Petersen, Ole Blaabjerg, Pia Skov Hansen, Thomas H. Brix, and Laszlo Hegedüs. 2006. "Establishment of Reference Distributions and Decision Values for Thyroid Antibodies Against Thyroid Peroxidase (Tpoab), Thyroglobulin (Tgab) and the Thyrotropin Receptor (Trab)." *Clinical Chemistry and Laboratory Medicine (CCLM)* 44 (8): 991–98. https://doi.org/doi:10.1515/CCLM.2006.166.

references II

Rinne, Horst. 2010. *Location-Scale Distributions : Linear Estimation and Probability Plotting*. Justus-Liebig-Universität. http://geb.uni-giessen.de/geb/volltexte/2010/7607.

Uldall Torp, Nanna Maria, Niels Henrik Bruun, Peter Astrup Christensen, Aase Handberg, Stig Andersen, and Stine Linding Andersen. 2022. "Thyrotropin Receptor Antibodies in Early Pregnancy." *The Journal of Clinical Endocrinology & Metabolism* 107 (9): e3705–e3713. https://doi.org/10.1210/clinem/dgac383.