university of groningen

University Medical Center Groningen

University of Groningen

Distributional inference

Albers, Casper Johannes

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2003

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Albers, C. J. (2003). *Distributional inference: the limits of reason*. s.n.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 30-09-2022

Part III

Applications

Chapter 5

Analyzing water-quality data

> *'To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.'*
>
> Sir R.A. Fisher[1]

The Advancement of Science is largely a matter of combining rational argument and empirical evidence. In areas like philosophy and pure mathematics rationalization is in the forefront whereas the empirical sciences deal with the design of experiments, collection of empirical evidence, etcetera. In mathematical statistics we try to combine these arguments. In principle the development of theory should follow the needs from practice while experiments should follow the needs of theory. Fisher's statement is therefore also justified in the reversed case where *practice is called in after the theory has been developed*. Interaction between theory and practice is needed. That is why the author of this thesis spent two months at the Centre for Quantitative Methods in Eindhoven (The Netherlands) adapting the (original) theory of Chapter 3 such that it can be applied to water quality data available at CQM.

5.1 Description of the data

The Institute for Inland Water Management and Waste Water Treatment , RIZA[2], is the research and advisory body for inland water for the Dutch Directorate-General for Public Works and Water Management, *Rijkswaterstaat*. The Institute for Coast and Sea, RIKZ[3], is its research and advisory body on the subject of sea and coastal water. These institutes collect data on and conduct research into quality and quantity of water. On the basis of the data, RIZA and RIKZ make recommendations concerning the management of water in The Netherlands and abroad. Both institutes participate in Rijkswaterstaat's MWTL-project[4], a large-scale long-running project concerning the topics mentioned.

RIKZ and RIZA are interested in concentrations of physical, chemical and biological substances in inland and sea water. For that purpose measurements of concentrations

[1] *Indian Statistical Congress*, Sankhyā, ca. 1938
[2] RIZA: Rijksinstituut voor Integraal Zoetwaterbeheer en Afvalwaterbehandeling
[3] RIKZ: Rijksinstituut voor Kust en Zee
[4] MWTL: Monitoring Waterstaatkundige Toestand des Lands

in surface water are made. Those concentrations include chemicals, metals, radioactivity levels, and biological entities. A considerable number of observations falls below a given detection-, determination- or report-threshold Δ.

The level of the threshold depends on the measured substance. This censoring to the left generates difficulties, at least when concentrations below Δ were encountered in the sample. If only very few values are below Δ, then different methods will not lead to too different results. For example, any 'reasonable' method to generate a median estimate will not be effected if (much) less than 50% of the data are censored. However, in the data collected by RIZA and RIKZ, these fractions are often high and a reliable method for dealing with such observations is needed.

The MWTL data set used in the analysis is of size $N = 47850$. Measurements are performed for 35 different substances, and the data spread a range of approximately 10 years. All kinds of water-systems (sea water, small and large rivers, lakes, etcetera) in The Netherlands are covered; measurements were taken at 64 locations (at most locations only a few of the 35 substances were measured). For the different substances, the locations and time period of measurement vary. Each entry in the data set consists of the following parameters: the name of the substance, chemical or quantity measured, the location where this is measured, the date of measurement (only for RIKZ-measurements), and, of course the measured concentration.

On average, 1350 measurements have been performed per substance. The MWTL data set is a sample in space (64 locations) and time (a span of approximately ten years). It is a questionable approach to regard the 1350 measurement sessions as a representative sample of the distribution of concentrations of the specified substance. Nevertheless, we will start out by using this assumption in our analysis. With a more sophisticated time-series analysis, the time factor might be taken into account, see Section 5.5 for more considerations.

As stated above, concentrations of 35 different 'substances' are measured. In 34 cases real concentrations were involved. One case deals with the presence of bacteriological entities, and the measured values are frequencies. This case will not be analyzed. The 34 cases can be classified according to whether

1. All measurements are above Δ (Section 5.2);
2. Some measurements are below Δ (Section 5.3);
3. Two or more thresholds are involved (Section 5.4).

The third situation occurs for instance when, during the measurement period, the measuring equipment is improved, thus lowering the threshold. Another possibility is that different equipment is used at different locations for measuring the same substance, not all equipments being equally sensitive.

5.2 All measurements are above the threshold

SWAVING AND DE VRIES (2000) investigated how these data should be dealt with

such that bias and uncertainty in estimates for the first four population moments are smallest. In this section we consider the cases where censoring is absent.

We want to use the estimates $f_n^{(m)}$ introduced in Chapter 3 to make inferences concerning all substances individually. The question now is how to choose the initial guess ψ, and the confidence ν in it. Given these ψ and ν, the smoothing factor m follows from the rule of thumb $m = 5.2\sqrt{n}\, v$ (see Section 3.7), where v is the L_1-distance $||\psi - f||_1$ where optimality is required (and the choice $w = 2$ has been incorporated). Simulations suggested that the method performs well if a choice of $v < \frac{1}{2}$ is justified. Using the approximate standard error

$$n^{-1/4} m^{1/8} \sigma_n^{(m)}(x) \approx \frac{m^{1/8} f_n^{(m)}(x)}{\sqrt[4]{4\pi n F_n^{(m)}(x)(1 - F_n^{(m)}(x))}}$$

confidence bands around $f_n^{(m)}(x)$ will be constructed via

$$f_n^{(m)}(x) \ \pm \ n^{-1/4} m^{1/8} \Phi^{-1}\left(\tfrac{\alpha}{2}\right) \sigma_n^{(m)}(x).$$

SWAVING AND DE VRIES (2000) state that the 34 substances with continuous measurement scale can be described reasonably well with suitable lognormal densities. That is why we will use (log)normal densities as initial guesses and transform the data via $t = \log x$. The parameters μ en σ^2 will be estimated from the data, so $\psi \sim \mathcal{N}(\bar{t}, s_t^2)$. The theory prescribes how to specify ψ a priori. Some data peeping is necessary and allowed if this is taken into account in specifying m (see Section 3.8 for a theoretical motivation). That the (transformed) data follow a normal distribution might be questionable. In many cases, a superposition of (normal) distributions should be expected (e.g. when measurements are performed at different locations, each location having its own distribution). However, as initial guess, a normal distribution is expected to suffice.

Because theoretical arguments are lacking, the specification of the confidence in the initial guesses will be based on the data. In this section, the quantification of v, is based on the following inputs:

1. the sample size n. Section 3.8 provides a contribution $v_* = v_*(n)$;
2. how well the data fits a normal density. The squared correlation coefficient r^2 in a normal probability plot regression (explained in the next section) provides information about the confidence in the initial guess. The higher the correlation, the higher the confidence in the assumption of (log)normality. This is formalized by the contribution $v_{**}(r) = 2(1 - r^2)$. Alternative contributions, mainly decreasing functions of r or another measure of correlation (e.g. $2\sqrt{1 - r^2}$) are also possible, but we consider the specification given as good enough.

In this section the data-dependent choice of v is defined through[5]

$$v = v(n, r) = v_*(n) + v_{**}(r).$$

[5]This is rounded off to the nearest multiple of .05 because of numerical arguments.

Swaving and De Vries were computing estimates of some characteristics, most importantly population moments. Using the estimated densities, alternative estimates for the first four (central) moments as characteristics of the estimated density $f_n^{(m)}$ can be constructed. Calculations for average, variance, skewness, and kurtosis, respectively, were done as follows:

$$\begin{aligned}
\bar{t} &= \int t f_n^{(m)}(t)\,\mathrm{d}t \\
s^2 &= \int (t-\bar{t})^2 f_n^{(m)}(t)\,\mathrm{d}t \\
g_1 &= \int \frac{(t-\bar{t})^3}{(s^2)^{3/2}} f_n^{(m)}(t)\,\mathrm{d}t, \\
g_2 &= \int \frac{(t-\bar{t})^4}{(s^2)^{4/2}} f_n^{(m)}(t)\,\mathrm{d}t.
\end{aligned}$$

Results for the substances with all measurements uncensored, were given in ALBERS (2001). In Figure 5.1 some results are visualized. The figures display the estimated concentration density and the 95%-confidence bands of

`pb50`, the concentration of led in particle dust suspended in water (measured in mg per kg 'dry-weight' particle dust;

`pb60`, the concentration of led in the meat of mussels (measured in mg per kg 'dry-weight' meat);

`Cl`, the concentration of chloride (measured in mg per liter);

`SiO2`, the concentration of silicate (measured in mg per liter).

The two graphs for led concentrations show similarities with normal densities, but also clear dissimilarities. This can also be seen from the estimated population characteristics (e.g. the estimated skewness for `pb60` is $-.34$, differing significantly from zero). The extreme skewness in the graph for silicate could indicate either the existence of a physical maximum concentration of `SiO2` in water or a superposition of densities. The latter explanation is the case, since silicate a substance that is needed for growth by a certain kind of fytoplankton, and these micro-algae appear only in spring and summer, hence in winter there is no 'thread' to silicate. The typical, almost bimodal character for chloride is explained by looking at locations separately: the concentration densities at fixed locations are approximately normal whereas their mixture is not (this is visualized in ALBERS, 2001).

5.3 Some measurements below threshold

For the substances with one fixed detection threshold Δ (and some observations below this Δ) the procedure used is as follows. The number of censored observations

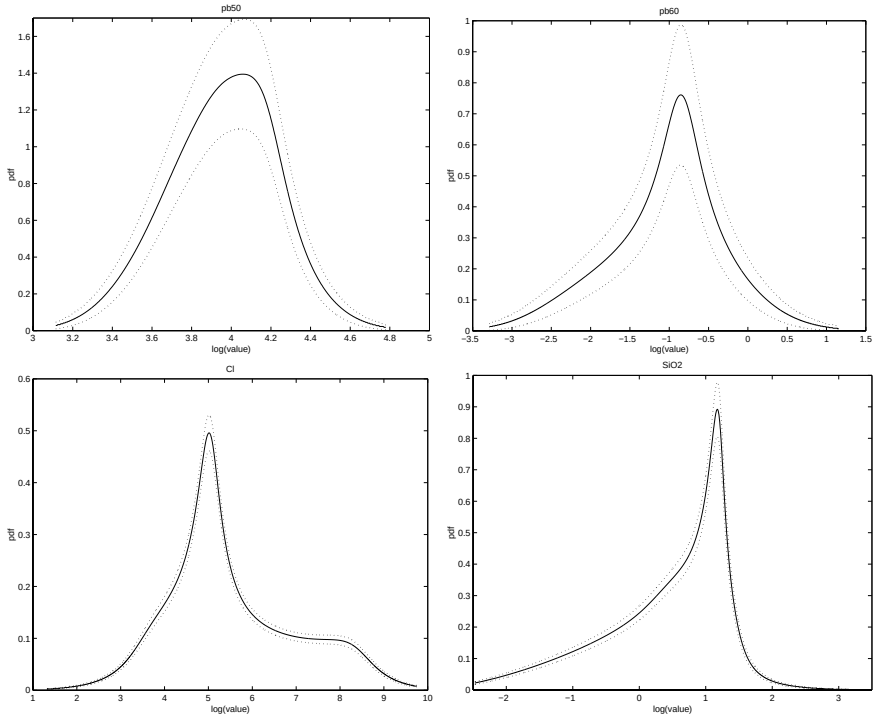

Figure 5.1: Above-left: `pb50` *($n = 213$, $m = 19$), above-right:* `pb60` *($n = 99$, $m = 16$), below-left:* `Cl` *($n = 2827$,$m = 41$), below-right:* `SiO2` *($n = 2206$, $m = 61$) and their 95% confidence bounds.*

k, the height of the threshold Δ, and, of course, the values of the $(n-k)$ uncensored measurements are known. Different methods exist for treating censored values. When the fraction of censored values is low, say $k \leq 2$, using an appropriate method is of course less important than when the fraction is considerable. The fraction of observations below the threshold varies between the substances from approximately 0.1% to 67%, and some sophistication will be necessary.

The method we will use, based on normal probability plot regression, is in line with SWAVING AND DE VRIES (2000). The censored measurements will be replaced by estimated values. Using ordinary least squares, the ordered observations $t_{[k+1]}, t_{[k+2]}, \ldots, t_{[n]}$ are fitted to quantiles $\Phi^{-1}(\frac{k+1}{n+1}), \Phi^{-1}(\frac{k+2}{n+1}), \ldots, \Phi^{-1}(\frac{n}{n+1})$ from a normal distribution. On basis of the resulting regression line, the estimator for $t_{[i]}$, $(i = 1, \ldots, k)$ is constructed by using the value corresponding to $\Phi^{-1}(\frac{i}{n+1})$, $(i = 1, \ldots, k)$. A disadvantage of this method is that estimated values are sometimes larger than Δ. When this occurs, the estimate is replaced by Δ.

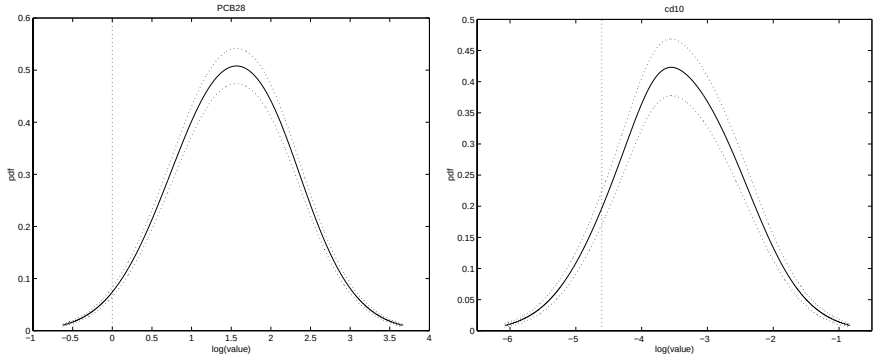

Figure 5.2: `PCB28` *($n = 1558$, $k = 34$, $\Delta = \log(1)$, $m = 10$) (left) and* `cd10` *($n = 706$, $k = 61$, $\Delta = \log(0.01)$, $m = 14$) (right).*

In SWAVING AND DE VRIES (2000) it was assumed that Δ was at a fixed point, and measurement errors were all zero. We continue these assumptions, in Section 5.5 we will reconsider this.

Using this adapted data set as 'the' data set, we continue in the same way as in Section 5.2. Of course, the consistency of the estimator for values below Δ has disappeared (unless $\Psi = F$), and the confidence bands are only exactly of level $(1 - \alpha)$ when $\psi = f$.

More precisely: for small values of concentration t ($t \ll \Delta$), we have replaced the latent observations by 'observations' from $\psi(t)$, and therefore $\mathbf{E}\ f_n^{(m)}(t) \approx \psi(t)$. For values larger than the threshold, the observations constitute a sample from the (unknown) true density f, providing $\mathbf{E}\ f_n^{(m)}(t) \approx f(t)$. Since our estimator is always a continuous estimator, the expected value of $f_n^{(m)}(t)$ with $t \approx \Delta$ will be 'somewhere between $\psi(t)$ and $f(t)$'. 'Where exactly' cannot easily be defined. Of course in the case that $\Psi = F$, practically impossible, consistency still holds. When f is truly distributed normally, ψ is a consistent estimator of f and, eventually, our estimator will be consistent as well. When f is not normal, a systematic error is made. In general, every method will generate systematic errors when the tail assumptions are incorrect. Making tail assumptions is a necessity, since no information about the tail behaviour is present.

Results are given in ALBERS (2001). Some examples are in Figure 5.2. The substances displayed are `PCB28`, a component of poly-chlorinated biphenyls; a type of organic micro-pollution (measured in μg per kg 'dry-weight' particles), and `cd10`, the concentration of cadmium (measured in μg per liter). Again, our method of fine-tuning a normal initial guess seems quite appropriate.

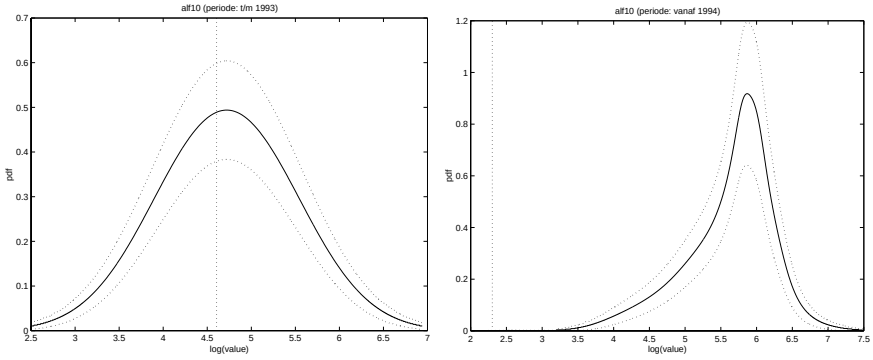

Figure 5.3: `alf10` *(left: until 1993, right: from 1994). A significant time effect is clearly visible.*

5.4 Substances involving multiple thresholds

Multiple thresholds may occur because of technological improvements, measuring with different types of equipment at different locations, etcetera. Two cases can be distinguished: those with and those without recording of the measurement-dates[6]. The case without date recording is not of interest to us.

After analyzing the data it is possible to classify each observation in a group with one Δ. Classification can be performed according to time, location, etcetera. When there are d thresholds, the classification will provide d groups, each with one threshold that matters, and the previous theory can be applied. Finally, with their sample sizes as weights, the separate estimates can be joined to construct one density estimate. This is, of course, only relevant if there are no significant differences between the d groups. Figure 5.3, for example, displays density estimates corresponding to the substance `alf10`, the amount of radiation in the water (measured in mili-Becquerel per liter). Until 1993, measurements were performed with equipment with a threshold of $\Delta_{d=1} = \log(100 \text{ mBq}/l)$, after that measurements were performed with a corresponding threshold of $\Delta_{d=2} = \log(10 \text{ mBq}/l)$. Kolmogorov's goodness of fit test (see Section 4.1 and KOLMOGOROV, 1933) clearly showed that the two (estimated) densities were significantly different. Joining them is unreasonable: it is better to infer about the two sub-populations separately.

5.5 Complications

Discrete measuring
Our density estimation method is used to estimate the distribution of the concen-

[6]It is unclear why the dates for RIKZ-measurements are missing, since they actually were recorded.

trations as measured. It is assumed that this is a continuous distribution, though measurements cannot be performed with infinite precision. When the precision is good enough, as is the case for most substances, the violation of the continuity assumption has almost no effect. However, for a few substances the precision is so low, that our method does not generate usable estimates, and a histogram seems the best possible thing.

Measurement errors
Until now, emphasis was put on descriptive statistics: the data were used to describe the population's behaviour. Focus was on the distributions of the measurements, not on the 'true' underlying concentration values (though even their definition is not more than operational).

The measurements are compositions of the true values, and systematic and random disturbances. To make inferences about the distribution of the true values y instead of the observed t, information has to be gathered about the measurement error $e = t - y$. The distribution $\mathcal{L}\ (e|Y = y)$ has to be known in order to transform $\mathcal{L}\ t$ into $\mathcal{L}\ Y$, which is the distribution of true interest.

With repeated measurements, information can be gathered about the measurement errors. Sometimes this error (or its variance) is neglectable, and the previously presented theory is applicable. This is not generally the case, especially for low concentrations the measurement error can be (relatively) high.

Point mass at zero
As mentioned earlier, we assumed continuity. It is not impossible that a point mass occurs for concentrations of zero. This would be the case for substances that only occur in water when the water is polluted, but not in general. Then, the concentration is zero with unknown probability λ, and larger than zero with probability $1-\lambda$. Before using the density estimation theory, an estimate $\hat{\lambda}$ has to be constructed.

Time effects
One of the major aspect of the data is unaccounted for in the analysis up till now: the measurements are performed in a time-frame. Reasons behind not discussing this earlier are that in our data set there was no record of date of measurement for RIZA-measurements while the density estimation method of Chapter 3 is not suitable, in its current form, for time series analysis.

In the analyses of this chapter the data were interpreted as being measured at one fixed time-point. This wrong assumption mainly influences the trustworthiness of the inferences for substances with a structural trend in its concentration, because the true density involved is constantly shifting and changing. Effects of the interpretation for substances that are assumed to behave fairly stable in time are (much) smaller. Previous experience (SWAVING AND VRIES, 2000) shows that most substances could be described quite well by lognormal distributions, this gives us confidence in the assumption that the time-effect is usually small compared to other effects.