STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

AFDELING MATHEMATISCHE STATISTIEK

Report S 316

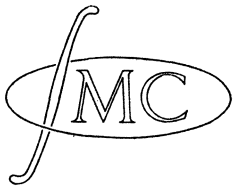
(revised edition)

Statistical Estimation of the Distribution of Sea Level

by

J. Kriens

(Presented at the N.A.T.O. Seminar on Decision Problems in connection with Dike Construction and Administration of Water Reservoirs, Copenhagen, September 19-29, 1963)



October 1963

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.),by the Municipality of Amsterdam and by several industries.

1. Introduction

The title of this report suggests it being evident, that sea levels can be regarded as a statistical phenomenon. It may be, that at present most people are convinced this to be the case, but only a few decades ago, one usually had another point of view. E.g. in The Netherlands, people, including most engineers of the Public Works Department, concentrated attention on the highest level of high tide, known from the past in making decisions about the heights of the dikes.

The question is of course, whether a statistical model can be made of the situation, which on one hand leads to results, reasonably in agreement with the observations and giving us an instrument to adequately study certain problems on the other hand. It has been shown that in many situations a satisfactory statistical model can be constructed. (cf. section 1.0.3 of [7] and [12]). Therefore, we now turn to the subject proper: Statistical Estimation of the Distribution of Sea Level.

2. Some statistical notions and methods

Suppose $\underline{x}$ is a random variable [1]. The probability that $\underline{x}$ equals a value $\leq x$ will be denoted by $P[\underline{x} \leq x]$. The function

$$F(x) = P[\underline{x} \leq x] \tag{2.1}$$

is the distribution function of $\underline{x}$. If $F(x)$ is differentiable with a continuous derivative almost everywhere, we call $F(x)$ a continuous distribution function and

$$f(x) = \frac{d}{dx} F(x) \tag{2.2}$$

the density function or frequency function of $\underline{x}$.

There are many different continuous distribution functions, among which, we here mention:

---

[1] In this report random variables will be underlined.

the normal distribution function:

$$F_{N(\mu,\sigma)}(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt \qquad (2.3)$$

$$-\infty<x<+\infty$$

the lognormal distribution function:

$$F_{\log N(\mu,\sigma)}(x) = 0 \quad \text{if } x \leq 0$$

$$F_{\log N(\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{0}^{x}\exp\left[-\frac{(\log t-\mu)^2}{2\sigma^2}\right]\frac{dt}{t} \qquad (2.4)$$

$$\text{if } x>0$$

and

the exponential distribution function:

$$F_{exp}(x) = 0 \quad \text{if } x \leq 0$$

$$F_{exp}(x) = 1-e^{-\lambda x} \quad \text{if } x > 0 \qquad (2.5)$$

The mean value

$$\alpha_\nu = \int_{-\infty}^{+\infty} x^\nu f(x)dx \qquad (2.6)$$

is called the ν^{th} moment of $\underline{x}$. Furthermore

$$\mu_\nu = \int_{-\infty}^{+\infty} (x-\alpha_1)^\nu f(x)dx \qquad (2.7)$$

is the ν^{th} central moment of $\underline{x}$.

If a sample of n observations is given and we suppose these observations to be random and mutually independent drawings from a distribution function, a natural question is: "From what kind of distribution function may these observations be a sample"?

There exists several general methods to examine whether a given set of independent observations can be regarded as random drawings from a given distribution function. We mention some of these methods briefly, the first one being a graphical method.

A graphical method to test whether observations fit a given distribution function

If we plot a normal distribution function with parameters $\mu=0$, $\sigma=1$ as a function of x on graphpaper with two linear scales, we get a graph as shown in fig. 2.1.

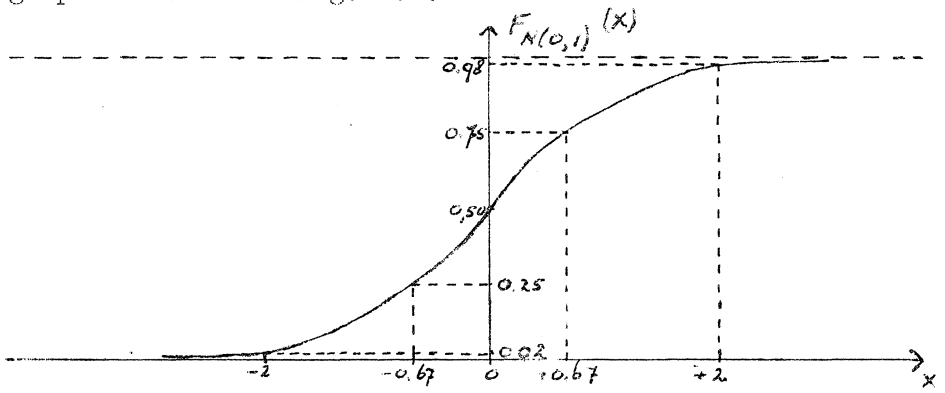


fig. 2.1

The normal distribution with $\mu=0$, $\sigma=1$ plotted on graph paper with two linear scales

In order to compare our observations

$$x_1, x_2, \ldots, x_n$$

with this distribution function, we rank them in increasing magnitude; the ranked sample is denoted by

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \tag{2.8}$$

The probability

$$F(x_{(i)}) = P[x \leq x_{(i)}] \tag{2.9}$$

can be estimated by $\frac{i}{n}$. Plotting the points $(x_{(i)}, \frac{i}{n})$ in the graph of figure 2.1 should then give a series of points, lying approximately on the curve of $F_{N(0,1)}(x)$ if $x_1, \ldots, x_n$ are observations from a normally distributed random variable x with $\mu=0$ and $\sigma=1$.

Mostly it is rather difficult to see whether a series of points fit reasonably to a curved line, or not. Therefore a kind of paper - "normal probability paper" - has been constructed on which the function

$F_{N(0,1)}(x)$ is transformed into a straight line, as shown in fig. 2.2. On this paper the points $(x_{(i)}, \frac{i}{n})$ should approximately ly on the

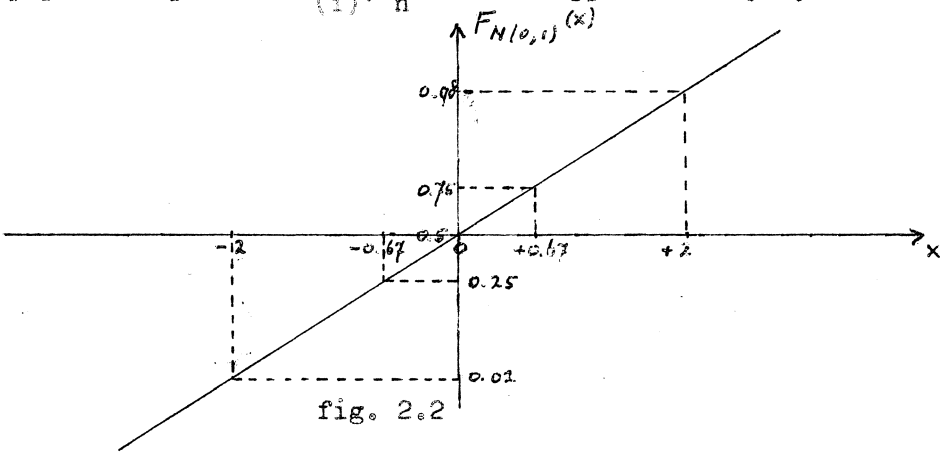


fig. 2.2

The normal distribution function with μ=0 and σ=1 plotted on normal probability paper

straight line drawn in fig. 2.2.

Normal distribution functions with arbitrary values of the parameters μ and σ also become straight lines on normal probability paper, because

$$F_{N(\mu,\sigma)}(x) = F_{N(0,1)}\left(\frac{x-\mu}{\sigma}\right) \qquad (2.10)$$

Probability paper transforming other types of distribution functions into straight lines can be constructed as well. Just one other example: to see whether observations fit an exponential distribution function, we plot the points

$$\left(x_{(i)}, \; 1 - \frac{i}{n}\right) \qquad (2.11)$$

on paper with the ordinate on a linear scale and the abscissa on a logarithmic scale, for

$$\log P[\underline{x} \geq x] = \log[1-F_{exp}(x)] = -\lambda x \qquad (2.12)$$

is a straight line on this type of paper.

In practice, plotting the points $(x_{(i)}, \frac{i}{n})$ or $(x_{(i)}, 1-\frac{i}{n})$ has some drawbacks. Therefore plotting the points

$$\left(x_{(i)}, \frac{i-0,3}{n+0,4}\right), \tag{2.13}$$

or

$$\left(x_{(i)}, 1-\frac{1-0,3}{n+0,4}\right) \tag{2.14}$$

may be recommended if we want to fit a distribution function [2]. This method has the property, that with a very good approximation

$$P\left[F(x_{(i)}) \leq \frac{i-0,3}{n+0,4}\right] = \frac{1}{2} \tag{2.15}$$

implying that the probability of finding a point below the straight line equals the probability of finding a point above this line. This property is independent of the distribution function we are sampling from (cf [1]).

Graphical methods have the advantage that they learn us a lot about an unknown distribution function, but they have the disadvantage that one always can dispute, whether a given fit is good or not. This difficulty can be avoided by using an analytical method; the oldest one is due to K. PEARSON.

PEARSON's method to fit a distribution function

PEARSON observed that many density functions

$$y = f(x) = \frac{d}{dx} F(x) \tag{2.16}$$

satisfy the differential equation

$$\frac{dy}{dx} = \frac{x+a}{b_0+b_1x+b_2x^2} \cdot y \tag{2.17}$$

and started from this equation to build up a system of density functions. The nature of the roots of the denominator in the right

---

2) If one has another aim, other plotting - positions may be preferable (cf. section 1.2.7 of [7]).

hand side of (2.17) determines the main types of the system, to which a.o. belong the distribution functions (2.3) and (2.5).

We can fit a curve from the PEARSON-system by estimating the constants a, b_0,b_1 and b_2 from the observations, which is done by equating the moments (2.6) and the corresponding sample moments

$$\alpha'_\nu = \frac{1}{n} \sum_{i=1}^{n} x_i^\nu \qquad (2.18).$$

The method is extensively described by W.P. ELDERTON in 1906 (cf. [6]). As drawbacks of the method, we mention:

1. one limits oneself a priori to a specific class of distribution functions, i.e. the class satisfying (2.17),

2. if one takes only a few constants, there is a very limited choice, whereas if one takes many constants, the estimations for them, are very sensitive for the actual observations, because the higher moments are,

3. no attention is paid to problems of extrapolation.

One of the most frequently used analytical methods is:

<u>The χ^2 - method to test whether observations fit a given distribution function</u>

In applying this method, we first divide the range of possible values of the observations into disjunct intervals (= classes or cells), suppose r intervals. If the distribution function F is completely specified, we can find the probability p_i, that a random drawing from this distribution will have a value in the i^{th} class, whereas in a sample of n observations, the expected number in this class will be $n\,p_i$. The actual number of observations in a class is a random variable, denoted by f_i for the i^{th} class.

As statistic to test the goodness of fit between the distribution function F of our hypothesis and the observations, we use

$$\underline{t} = \sum_{i=1}^{r} \frac{(\underline{f}_i - n\, p_i)^2}{n\, p_i} \qquad (2.19).$$

It is clear that if our hypothesis is true, the $\underline{f}_i$ will have values f_i not deviating too much from the expected values $n\, p_i$ and if this is the case, we find a not too large value t for our test-statistic $\underline{t}$.

More precise, one can proof the following theorem: if we are sampling from a distribution function F, having a probability p_i for an observation in class i, then for $n \to \infty$, the density function of $\underline{t}$ tends to

$$f(t) = \frac{1}{2^{\frac{r-1}{2}}\, \Gamma(\frac{r-1}{2})}\, t^{\frac{r-3}{2}}\, e^{-\frac{t}{2}} \qquad (2.20),$$

which is the density function of the so called χ^2 distribution with r-1 degrees of freedom. For finite n, (2.20) is a good approximation of the density function of $\underline{t}$ e.g. as long as $r > 2$, all $n\, p_i > 1$ and some $n\, p_i > 5$ (cf [11, §56] and [2]).

The procedure is as follows: one computes the value t of $\underline{t}$ for the sample and finds the probability

$$P[\underline{t} \geq t] = \frac{1}{2^{\frac{r-1}{2}}\, \Gamma(\frac{r-1}{2})} \int_t^{\infty} x^{\frac{r-3}{2}}\, e^{-\frac{x}{2}}\, dx \qquad (2.21).$$

If this is smaller than a pre-assigned value α, e.g. $\alpha = 0{,}01$ or $\alpha = 0{,}05$, the hypothesis that the sample is drawn from the distribution function F is rejected. Otherwise we conclude that the observations are compatible with our hypothesis.

In the case our hypothesis is not a completely specified distribution function, but only states it to be a distribution function of a certain class, e.g. the class of normal distribution functions, then the p_i are functions of the unknown parameters $\pi_1, \ldots, \pi_s$. (μ and σ for the class of normal distribution functions).

So in stead of (2.19), we have

$$\underline{t} \sum_{i=1}^{r} \frac{(f_i - n\, p_i(\pi_1,\ldots,\pi_s))^2}{n\, p_i(\pi_1,\ldots,\pi_s)} \qquad (2.22).$$

If we estimate $\pi_1,\ldots,\pi_s$ in such a way from the sample, that (2.22) is minimized, then we can proceed in the same way as outlined above and only have to replace the number of degrees of freedom in (2.20) by r-s-1, s being the number of parameters estimated from the sample.

In practice minimizing (2.22) leads to very complicated equations in almost all cases. Therefore this method is replaced very often by the method of equating moments, giving not too bad approximations usually.

This analytic method of testing is more objective than the graphical method, mentioned before. However, we should realize that fixing the value of α and choosing the number of classes as well as the boundaries of the classes introduce rather arbitrary elements.

Another drawback is, that in fact we do not test whether the sample is from a distribution F, but only whether it may be a sample from a distribution, having probabilities $p_1,\ldots,p_r$ for the classes $1,\ldots,r$. So, we never can differentiate between distribution functions with the same values of the p_i. Of course, if the p_i of the F of our hypothesis are not correct, we want to reject this hypothesis. The freedom in choosing the number of classes can be used to maximize the probability of rejecting an hypothesis which is not correct (cf [2] and [9]). In [9] some suggestions to standardize the procedure are given also.

There exist general tests, alternative to the χ^2 - test, but these too have many drawbacks.

3. Fitting a distribution function to levels of high tide

In decision problems connected with dike construction, we are interested in the distribution function of levels of high tide and more particularly in the distribution function of high high tides.

In this section, which is mainly an account of work done for the so called "Delta - Commissie" (cf. [4]), we shall therefore limit our discussion to this problem.

Usually the data will consist of the levels of high tides, observed during a number of years, say N years. For each level h we can find the number N(h) of high tides, which exceeded h during these N years. Then

$$n(h) = \frac{N(h)}{N} \qquad (3.1)$$

is the mean number of exceedances of the level h per year.

Already in 1939 WEMELSFELDER drew attention to the fact, that if we compute the values n(h) for Hook of Holland for the period 1888-1937 and then plot log n(h) as a function of h, we get a curve which in the middle of the range of observations fits very good to a straight line (cf [12]). The same holds true if we take the longer period 1888-1956; see fig. 3.1.
There is a strong deviation from this line for low values of h, because for sufficient low h, almost all high tides will exceed h and n(h) tends to a constant value. Furthermore the straight line becomes vaguer for high values of h.

In solving decision problems, we are more interested in the probability that a level h of high tide will exceed a value h, or in the probability that a value h will be exceeded during a certain period, than in the mean number of exceedances per year. Therefore we shall not examine n(h), but the probability distribution F(h) of h. If we know F(h), the exceedance probability of a given level h during a certain period can be found if we know the number of random drawings from F(h) during that period. Moreover it can easily be proved, that for high values of h, n(h) is approximately equal to the probability of a value of h larger than h during a year.

In most cases fitting a distribution function to observations can best be started by plotting points on different kinds of probability paper. If we are not too unlucky, the behaviour of the

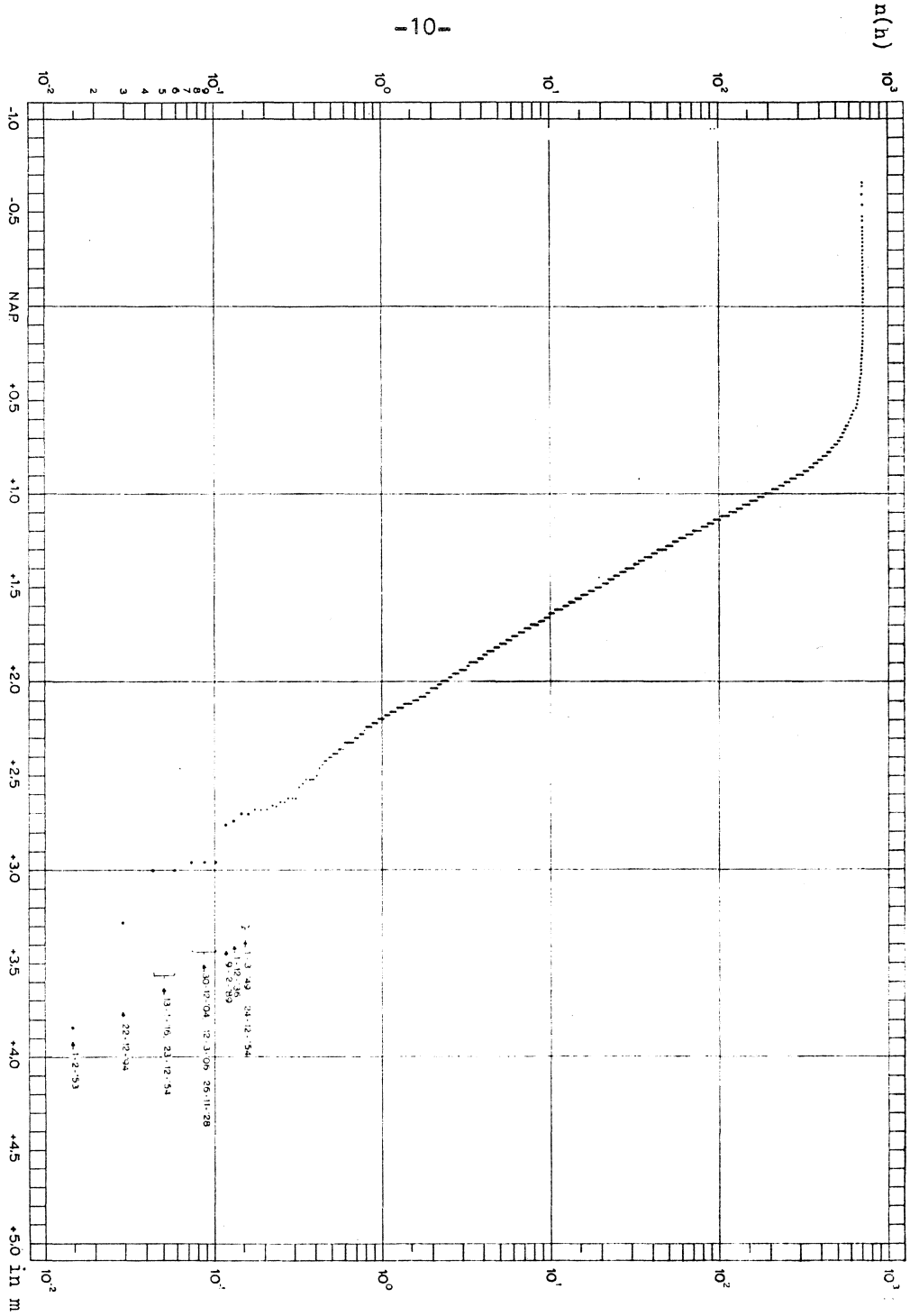


fig. 3.1

The mean number of exceedances per year of the level h in Hook of Holland during the period 1888-1956

different graphs will suggest what types of distribution functions may be relevant.

The results of WEMELSFELDER's work suggest that in Hook of Holland, levels of high tide agree with an exponential distribution function for high values of h. Therefore, after ranking all observations, larger or equal to a level h_0, in decreasing order

$$h_{(1)} \geq h_{(2)} \geq \ldots \geq h_{(n)}, \tag{3.2}$$

we plot the points

$$(h_{(i)}, \frac{i-0,3}{n+0,4}) \tag{3.3}$$

on paper with the ordinate on a linear scale and the abscissa on a logarithmic scale. If the observations $\geq h_0$ are random drawings from an exponential distribution, we should find, denoting the number of high tides, higher than h by i(h),

$$\log \frac{i(h)-0,3}{n+0,4} \approx a\ h + b, \tag{3.4}$$

or

$$\frac{i(h)-0,3}{n+0,4} \approx e^{a\ h + b} \tag{3.5}.$$

For $h=h_0$, we have

$$e^{a\ h_0 + b} = 1,$$

thus

$$a\ h_0 + b = 0$$

and

$$b = - a\ h_0 \tag{3.6}.$$

For $h \uparrow \infty$, we have

$$e^{a(h-h_0)} \downarrow 0,$$

so

$$\lim_{h \uparrow \infty} a(h-h_0) = - \infty$$

and a must be negative. If we put

$$a = -\lambda \quad (\lambda > 0) \tag{3.7}$$

and substitute (3.6) and (3.7) in (3.5), we find as an estimation of the conditional probability $P[\underline{h} \geq h | \underline{h} \geq h_0]$:

$$e^{-\lambda(h-h_0)}. \tag{3.8}$$

Whether the observations $\geq h_0$ really fit an exponential distribution function, can be tested with the χ^2-method, described in section 2. The unknown parameter λ is estimated by the so-called maximum likelihood estimation

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n}(h_j - h_0)} = \frac{1}{\overline{h} - h_0} \tag{3.9}.$$

If the range of values $\geq h_0$ is divided into r intervals, then the test-statistic $\underline{t}$ is χ^2 - distributed with r-1-1=r-2 degrees of freedom.

It turns out that not only in Hook of Holland, but also in many other stations along the Dutch coast, exponential distribution functions fit quite satisfactory to the observations, though with different values of the parameters λ (cf. [13]). Higher up in the estuaries the situation is different (cf also [13]).

There is some arbitrariness in choosing the point h_0. However, one can easily proof that if the observations $\geq h_0$ are drawings from an exponential distribution function, the same is true for the observations $\geq h_0'$ if $h_0' > h_0$. In practice one should choose the level h_0 not too low, because of the bending off of the curve for low values of h and on the other hand not too high in order to include as many observations as possible.

In the literature one finds many publications in which not exponential, but logarithmic normal distributions are fitted to

the observations in problems, we are discussing here (cf. a.o. [8]). One may therefore ask the question, whether the observations are also compatible with a logarithmic normal distribution, or not.

Since we concentrate our attention to high levels of high tide, it is obvious to fit a logarithmic normal distribution, which is truncated on the left, let us say at the value a:

$$F_{\log N(\mu,\sigma)}(h|h\geq a)=\left[1-F_{\log N(\mu,\sigma)}(a)\right]^{-1}\frac{1}{\sigma\sqrt{2\pi}}\int_{a}^{a+h}\frac{1}{t}e^{-\frac{(\log t-\mu)^2}{2\sigma^2}}dt \qquad (3.10).$$

Comparing the observations in Hook of Holland with this distribution, after estimating the parameters a,μ and σ also gives a very satisfying result.

This is not surprising, because of the following property. Suppose

$$F_{exp}(h) = 1 - e^{-\lambda h} \quad (h>0) \qquad (3.11)$$

is a given exponential distribution. Then one can prove, that for every pre-assigned value ϵ and for every finite interval $0 \leq h \leq 1$, there exist values a_0, μ_0 and σ_0 of the parameters a,μ and σ in (3.10), such that

$$\left|F_{exp}(h) - F_{\log N(\mu_0,\sigma_0)}(h|h \geq a_0)\right| < \epsilon \qquad (3.12)$$

for every h satisfying $0 \leq h \leq 1$ (cf [4]). An arbitrary exponential distribution can thus be approximated by a truncated lognormal distribution to any degree of precision. The approximation is at its worst for high values of h.

As the class of truncated lognormal distributions has three parameters to be chosen, it is of a more general nature than the class of exponential distribution functions. We can therefore always find a truncated lognormal distribution fitting a finite number of observations as least as well as an exponential distribution. This

implies for the case in which an exponential distribution gives a good fit, it being impossible to distinguish statistically between these two distribution and thus both distributions can be used from the statistical point of view. However, the exponential distribution is much more tractable mathematically and this may be a strong argument in favour of using this distribution function.

The most important property with respect to extrapolation problems, which can be proved is the following one. Suppose

$$\underline{h}_{(1)} \geq \underline{h}_{(2)} \geq \ldots \geq \underline{h}_{(n)}$$

are ordered random drawings $\geq h_0$ from an exponential distribution function with parameter λ. We define $\underline{v}_i$ by

$$\underline{v}_i = i(\underline{h}_{(i)} - \underline{h}_{(i+1)}) \qquad (i=1,\ldots,n) \tag{3.13}$$

with $\underline{h}_{(n+1)}$ being equal to h_0 with probability one. Then the $\underline{v}_i$ are also independent random drawings from an exponential distribution function with parameter λ (cf. [3]) and the statistic

$$\underline{B} = \frac{\underline{v}_1 + \ldots + \underline{v}_k}{\underline{v}_1 + \ldots\ldots\ldots + \underline{v}_n} \tag{3.14}$$

has a β-distribution with parameters k and n-k, i.e. its density functions equals

$$f(x) = \frac{(n-1)!}{(k-1)!(n-k-1)!} \; x^{k-1}(1-x)^{n-k-1} \tag{3.15}.$$

If we fit a distribution function to observations of high water levels, then the fitted distribution function is mainly determined by the bulk of not so very high observations. The statistic $\underline{B}$ gives us a method to test, whether the highest observations are deviating too much from the fitted distribution function in comparison with all observations, if we assume the observations to come from an exponential distribution function.

Remark 3.1.

In this section we supposed $h_1,\ldots,h_n$ to be independent observations. The cause of dangerous high tides are depressions and as more than one high tide may occur during a single depression, the observations are not necessarily independent if we use all high tides. Being especially interested in high high tides, the best one can do seems to choose the highest high tides during the separate depressions as observations. This procedure assumes the depressions to be statistically independent, which is not completely true (cf [10]).

Remark 3.2.

The level of high tide in Hook of Holland on February 1, 1953 was 0,57 m higher than the highest level observed before in Hook of Holland. This large difference is in agreement with the properties of an exponential distribution. According to (3.8) the distribution of $\underline{h}$, subject to the condition that $\underline{h}$ is larger than a certain value h_0 is

$$P[\underline{h} \leq h \mid \underline{h} \geq h_0] = 1 - e^{-\lambda(h-h_0)} \qquad (3.16).$$

The difference between the highest high tide observed and a new, higher high tide is thus again exponentially distributed with the same parameter λ and the distribution is independent of the value of h_0. For Hook of Holland $\hat{\lambda} = 2,97$; the probability of a difference of 0,57m or more equals $e^{-2,97 \cdot 0,57} \approx 0,18$, which certainly is not extremely small.

Remark 3.3.

Fitting PEARSON-curves, as described in section 2, to the observations in Hook of Holland gives rather unrealistic results in some cases. E.g. if we fit a curve with 4 parameters to the high tides $\geq$ 200 cm in the period 1901-1950, we get a density function of $\underline{h}$ with as a range $199 \leq h \leq 331$, so with an upperlimit, being 54cm lower than the high tide of 385cm in February 1953!

4. Extrapolation of a fitted frequency curve

In decision problems in connection with dike construction, we are not only interested in the range of levels of high tide in which we have observations. More important are questions concerning the possibility of high tides, higher than observed during the past.

The statistical study of this question can be done along two different lines. First we can investigate problems of extreme values and analyse the distribution function of e.g. the highest observation of a series of n observations, or the distribution function of the highest observation during a given period. A second line of attack is trying to extrapolate the distribution function, fitted to the observed values. In this section, we shall only make a few remarks about the latter problem, referring for the first one to [7].

Both, computations about extreme levels of high tide and extrapolation of a fitted distribution function, start from the hypothesis that the mathematical model and the estimations made of the parameters will remain correct in the future, anyhow during a not too remote future. As long as there are no arguments against it, this is a generally accepted procedure in applying mathematical models in empirical sciences.

Extrapolation of a curve fitted to observations is always a very delicate task. In this case, extrapolation means, making a prediction about the curve we would fit to the observations, if we have not only observations for a period of 69 years, but for a much longer period. Therefore we should work as carefully as possible.

The most dangerous thing we can do is extrapolating, which is not the curve from one distribution function, but the curve of a mixture of distribution functions. Let us suppose that the observations are generated by two different exponential distribution functions I and II. Separating the observations and plotting them as explained in section 2 would then give two different straight lines; cf fig 4.1. These two lines may approximately coincide, but nevertheless give highly different results in extrapolating them.

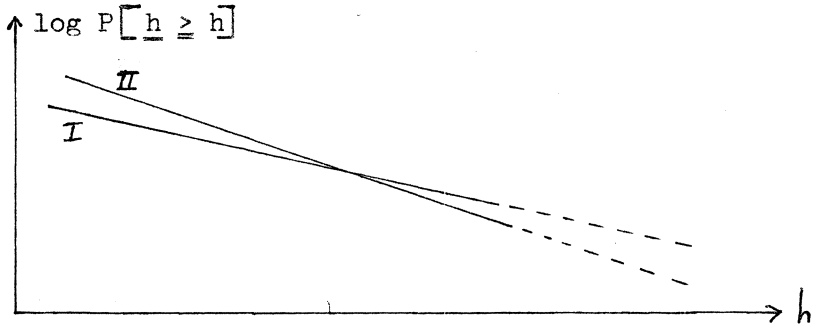


fig. 4.1

Two exponential frequency curves and their extrapolations

If we mix both distributions, we get another line, giving again different results in extrapolating it. In the case I and II are really straight lines on semi-logpaper, theoretically the line representing the mixture can never be a straight line.

One may raise the question, whether there are arguments for expecting the observations in Hook of Holland, being generated from two different populations. In fact the high tides in summer are much lower high tides than those in winter. If we split the observations of high tide in let us say "winterobservations"(e.g. during the months November, December and January) and "summerobservations", we we get two different straight lines, suggesting that both the "winterobservations" and the "summerobservations" are random drawings from exponential distribution functions, though with different values of the parameters. Also getting a straight line if we mix both populations (cf fig. 3.1) is a consequence of the fact that for the actual values of the parameters, the curvature of the new line is too small to be perceptable. The slope of the new line is somewhere between those of the two old ones and extrapolation of this line gives too optimistic estimations of the exceedance probabilities of high tides. As to be expected the line fitted to the winterobservations is the one with the smallest slope.

After this splitting one can ask the question again, whether we are extrapolating from homogeneous observations, i.e. observations

from one distribution function, or whether the data are still heterogeneous. Dutch meteorologists argued the observations still to be heterogeneous: high high tides are caused by dangerous depressions and one can make a differentiation in all depressions in winter between potentially dangerous depressions from a meteorological point of view and the other ones (cf [4]. Splitting the winterobservations anew into those belonging to dangerous depressions and the other ones (cf. also remark 3.1), leads again to two different straight lines on semi-logpaper. The line estimated from the levels of high tide, belonging to the observations during dangerous depressions from the meteorological point of view was the one, used ultimately. Evidently one can never be completely sure, whether a given set of observations is homogeneous or not. An argument in favour of this assumption lies in the fact, that the higher observations now fit much better to the fitted straight line than to the line fitted to all observations.

Remark 4.1

We tried making new splittings in the material, but did not get lines, differing significantly. These splittings were based on a) years with great activity of sun spots and years with low activity, and b) on potentially dangerous years and other years, respectively.

There is one more argument, supporting the hypothesis that the observations are now homogeneous. This argument is based on the theory of extreme values.

If we are sampling from a distribution function F(h) and take a sample of n observations, then the distribution function $G_{(1)}(h)$ of the highest observation $\underline{h}_{(1)}$ is

$$G_{(1)}(h) = \{F(h)\}^n \ , \qquad (4.1)$$

or, denoting

$$1 - F(h) = \phi(h) \ , \qquad (4.2)$$

$$G_{(1)}(h) = \{1 - \phi(h)\}^n \quad . \tag{4.3}$$

For high values of h, (4.3) is approximately equal to

$$G_{(1)}(h) \approx e^{-n\phi(h)} \quad . \tag{4.4}$$

In the case of F(h) being an exponential distribution function, we find

$$G_{(1)}(h) \approx e^{-n\, e^{-\lambda h}} \quad , \tag{4.5}$$

which is the so-called limiting distribution function of the first type (cf the classification in chapter 5 of [7]). As $\lambda > 0$, (4.5) can be approximated for large values of h as follows:

$$G_{(1)}(h) \approx 1 - n\, e^{-\lambda h} + \frac{n^2 e^{-2\lambda h}}{2!} \ldots \approx 1 - n\, e^{-\lambda h} \tag{4.6}$$

From F(h) being an exponential distribution function with parameter λ, it thus follows that $G_{(1)}(h)$ satisfies approximately (4.6). So if we plot the estimations (3.3) of $\log P[\underline{h} \geq h]$ on semi-logpaper on one hand and the yearly maxima on probability paper for the limiting distribution function of the first type on the other hand, we should get series of points, lying approximately on two parallel lines. For Hook of Holland this turns out to be the case for the observations left after the splitting procedure, but not for all observations taken together. This result once more suggests the "selected" observations to be a homogeneous sample.

Remark 4.2

We note that the derivation of $G_{(1)}(h)$, given above does not depend on the stability equation used by GUMBEL. The double - exponential distribution function (4.5) can be shown to be a good approximation for $\underline{h}_{(1)}$ for a large class of distribution functions, containing a.o. the γ-distributions and the normal distributions.

Extrapolation of a frequency curve, fitted to homogeneous observations seams reasonable for a not too long range above the observations. However, we know that the fitted straight line must bend off downwards for some value of h, because very high levels of high tide are impossible physically. Fortunately we are only interested in extrapolation in a range of a few meters to estimate the corresponding probabilities of exceedance. As there are no arguments to expect the fitted line to bend off already for these relatively low values of h and these levels are certainly possible from a physical point of view, the fitted line was extrapolated linearly.

Remark 4.3.

GUMBEL's method of extreme values was not used straightforwardly, because of the following arguments.

1. After the splitting procedure, discussed before, 166 observations were left in a period of 63 years. In applying the method of yearly maxima, only 63 observations would have been used. As a consequence of this neglect of available information, the results would be much less reliable.

2. In this method there is no discussion about the homogeneity of the data, which is an essential element in our treatment.

3. If one considers the decisionproblem of dike building as an economic decision in which costs of dikebuilding are balanced against the present value of future losses, caused by floods, one needs the distribution function F(h) (cf. [5]). This distribution function can never be derived by only studying distribution functions which are approximations of the distributions of the extreme values, as whole classes of distribution functions lead to exactly the same approximating limiting distributions.

4. References

[1] A. BENARD and E.C. BOS-LEVENBACH, Het uitzetten van waarnemingen op waarschijnlijkheidspapier, Statistica Neerlandica 7(1953) 163-173.

[2] W.G. COCHRAN, The χ^2-test of goodness of fit, The Annals of Mathematical Statistics 23(1952) 315-345.

[3] D. VAN DANTZIG, Kadercursus Mathematische Statistiek, Hoofdstuk 6, §1, blz. 282.

[4] D. VAN DANTZIG and J. HEMELRIJK, Extrapolatie van de overschrijdingslijn van de hoogwaterstanden te Hoek van Holland met behulp van geselecteerde stormen, Rapport Delta-Commissie, Bijdrage II.1, Staatsdrukkerij- en Uitgeverijbedrijf, The Hague (1960).

[5] D. VAN DANTZIG and J. KRIENS, Het economisch beslissingsprobleem inzake de beveiliging van Nederland tegen stormvloeden, Rapport Delta-Commissie, Bijdrage II.2, Staatsdrukkerij- en Uitgeverijbedrijf, The Hague (1960).

[6] W.P. ELDERTON, Frequency Curves and Correlation, Harren Press, Washington, (1953, 4th edition).

[7] E.J. GUMBEL, Statistics of Extremes, Columbia University Press, New York (1958).

[8] A. HAZEN, Flood Flows, Wiley and Sons, New York (1930).

[9] H.B. MANN and A. WALD, On the choice of the number of clas intervals in the application of the chi-square test, The Annals of Mathematical Statistics 13(1942) 306-317.

4. References

[10] P.J. RIJKOORT and J.HEMELRIJK, The occurrence of "twin" storms from the North West on the Dutch coast, Statistica Neerlandica 11(1957) 121-130.

[11] B.L. VAN DER WAERDEN, Mathematische Statistik, Springer Verlag, Berlin (1957).

[12] P.J. WEMELSFELDER, Wetmatigheden in het optreden van stormvloeden, De Ingenieur 54(1939) B31-B35.

[13] P.J. WEMELSFELDER, De overschrijdingslijnen van de hoogwaterstanden in het Nederlandse getijgebied, Rapport Delta-Commissie, Bijdrage III.2, Staatsdrukkerij- en Uitgeverijbedrijf, The Hague (1960).