THE LOGNORMAL MODEL FOR THE DISTRIBUTION OF ONE CLAIM

LARS-GUNNAR BENCKERT,
Stocksund (Sweden)

The most important property of a distribution function to be used as a model for the distribution of one claim is of course that it fits the data well enough. If there is no natural truncation point in the data a more formal demand is that all the moments of the distribution function exist. Further, to be of a real value to the statistician, the chosen d.f. ought to be reasonably handy to use. As all the moments of the lognormal d.f. exist the first point to be checked is whether the lognormal d.f. fits the data. The other points on the list below are the qualities that I think are of the greatest value when using a distribution function, i.e. they reflect the handiness of the d.f.

Does the lognormal d.f.

1. Fit the data?
2. Give an unbiased and efficient estimate of the mean?
   It is important that this estimate is not too difficult to compute.
3. Give a practicable confidence interval of the mean?
4. Give a known distribution function of the estimate of the risk premium?

This paper is an attempt to give an affirmative answer to these questions. As the lognormal distribution function has been treated in the monograph "The lognormal distribution" by J. Aitchison and J. A. C. Brown (Cambridge University Press) the theory of this distribution function will not be dealt with more than necessary for the context.

DOES THE LOGNORMAL DISTRIBUTION FUNCTION FIT THE DATA

This question can of course not be answered in general, but I will try to shed light upon it by giving some examples from widely

different fields of non-life insurance. The examples are given in the figures 1-6.

Fig. 1. Insurance against loss of profit due to fire 1948-1952. The claims are measured by the period of interruption. *Industrial risks.*

Fig. 2. Insurance against loss of profit due to fire 1948-1952. The claims are measured by the period of interruption. *Non industrial risks.*

Fig. 3. Fire insurance 1948-1951. Industrial and non industrial risks.

Fig. 4. Accident insurance. Number of days of illness. Occupancy Group A.

Fig. 5. Accident insurance. Number of days of illness. Occupancy Group B.

Fig. 6. Motor third party insurance.

The figures 3-6 are given in the form of probit diagrams. Thus the observed d.f : s are plotted on a normal paper, where the scale on the horizontal axis is logarithmic. The lognormal d.f. is then a straight line which cuts the 50 percent-line in μ and the 84 percent-line in $\mu + \sigma$.

The lognormal d.f. is

$$(1)\quad L(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^x \frac{1}{y} e^{-\frac{1}{2}\left(\frac{\log y - \mu}{\sigma}\right)^2} dy$$

It must always be kept in mind that μ and σ^2 are *not* the mean and the variance of x (but of $\log x$).

As seen from the figures the lognormal d.f. is of surprisingly wide applicability especially for the great and most important values of the variable. For the smaller values certain modifications of the data or the d.f. have been made in some cases.

Fig. 1 and 2. Insurance against loss of profit due to fire. Claims in Sweden 1948-1952.

The claims are measured by the period of interruption. There are two censoring points in the data; one at 180 days and one at 360

days. These two censoring points correspond to the periods of indemnity represented in the data.

The figures show no true frequency function as the number of claims is not proportional to the area but only to the height of the line above the interval in question.

Industrial risks

Number of claims 217.
D.f. used is $L(x; \mu, \sigma)$ truncated at 10 days
$\mu^* = 1{,}85$; $\sigma^* = 2{,}02$
$\chi^2 = 3{,}0$ with 5 degrees of freedom

Non industrial risks

Number of claims 307.
D.f. is the same as for industrial risks
$\mu^* = 1{,}9$; $\sigma^* = 1{,}75$
$\chi^2 = 4{,}2$ with 5 degrees of freedom.

Fig. 1. Insurance against loss of profit due to fire 1948-1952
The claims are measured by the period of interruption. *Industrial risks.*

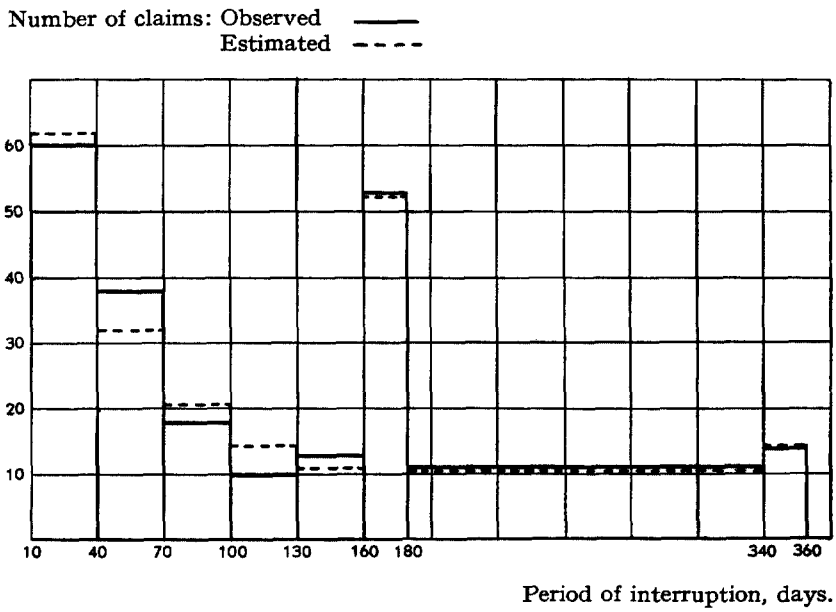


Fig. 2. Insurance against loss of profit due to fire 1948-1952.
The claims are measured by the period of interruption. *Non industrial risks.*

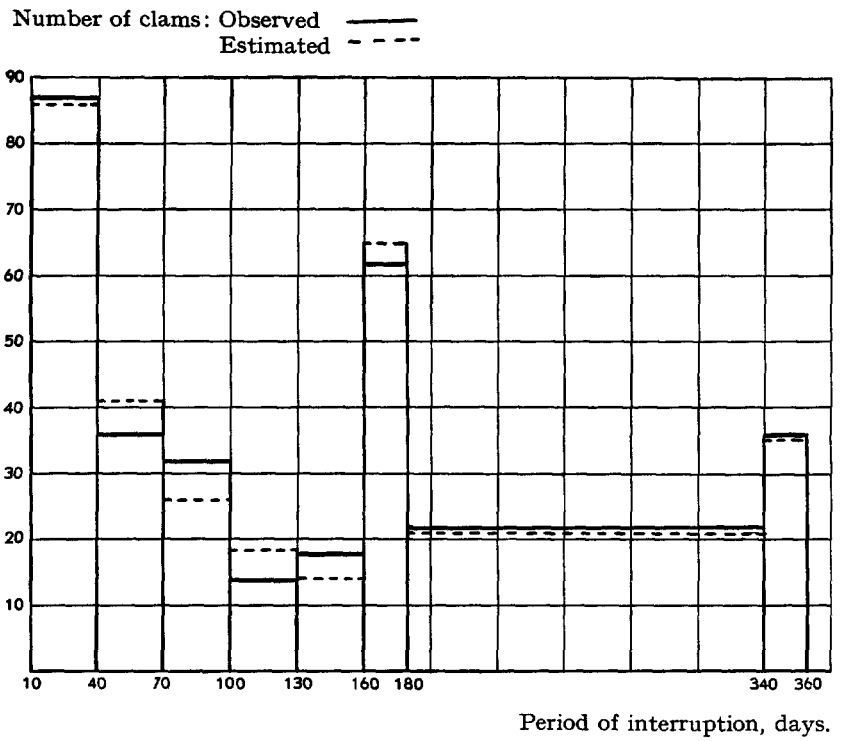


Fig. 3. Fire insurance.
Claims from some Swedish companies 1948-1951.

The d.f. used is $L(x - a; \mu, \sigma)$ where a is small. Of course a is the smallest value possible for x i.e. the observed values of x smaller than a should be excluded. This has not been done which is the same as treating these values as greater then they are.

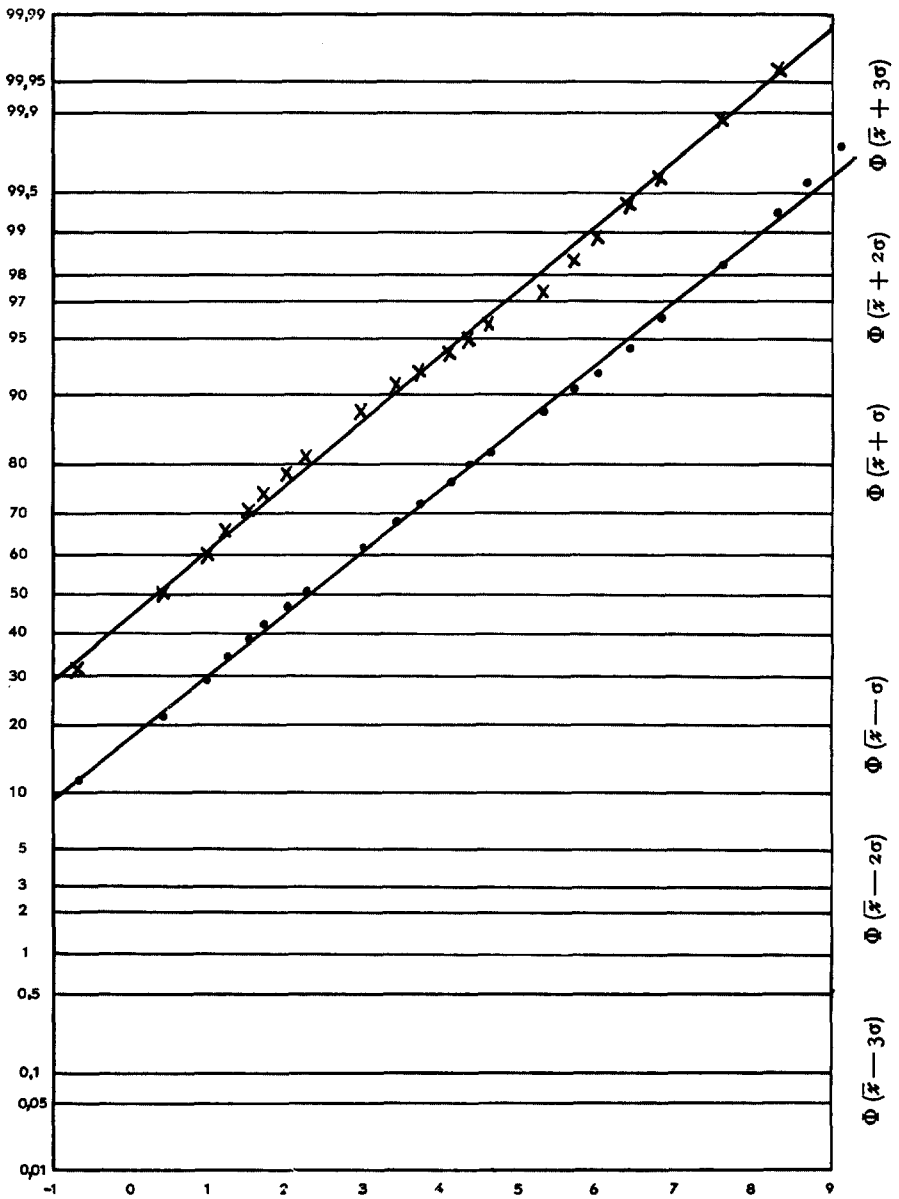
Industrial risks

Number of claims 7.711.
$\mu^* = 2{,}3$
$\sigma^* = 2{,}5$

Non industrial risks

Number of claims 38.075.

Fig. 3. Fire insurance 1948-1951.
. Industrial
x Non industrial



$\mu^* = 0{,}3$
$\sigma^* = 2{,}4$

The difference between the d.f. of industrial and non industrial risks lies almost entirely in μ while σ is practically constant. Thus the difference is the same as it should have been if the both types of risks had had the same d.f. but were measured by different monetary units.

χ^2 has not been calculated as it seems evident that it will be great. Still the lognormal curve seems to give a good fit over a remarkably wide interval — from 1.000 Sw. cr. to more than 1.000.000 Sw. cr.

Fig. 4 and 5. Accident insurance.

The data have been censored at 7,5 days. The variable is not quite the number of days of illness but

$$\frac{\text{amount paid}}{\text{amount insured per day}}$$

which in some cases is smaller than number of days of illness but the difference is not great.
The d.f. used is $L(x - a;\ \mu,\ \sigma)$

Occupancy group A.

Number of claims 2.840.
$\mu^* = 2{,}77$
$\sigma^* = 1{,}14$
$\chi^2 = 25{,}47$
Degrees of freedom 12
$P_{\chi^2} = 1{,}3\ \%$

Occupancy group B.

Number of claims 2.291.
$\mu^* = 2{,}68$
$\sigma^* = 1{,}11$
$\chi^2 = 25{,}21$
Degrees of freedom 11
$P_{\chi^2} = 0{,}8\ \%$

Fig. 6. Motor third party insurance.

The data come from two geographical areas. They are severely

Fig. 4. Accident insurance.
Number of days of illness.
Occupancy Group A.

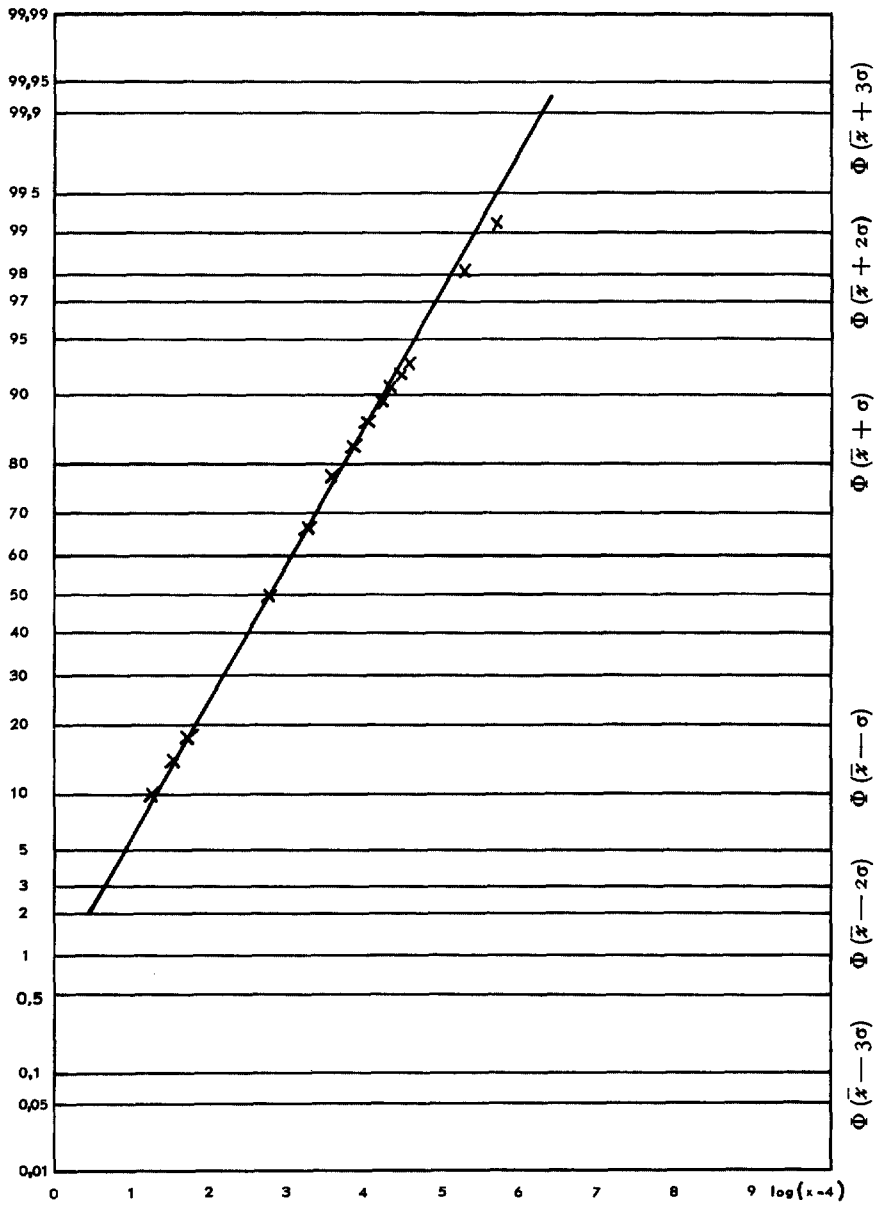
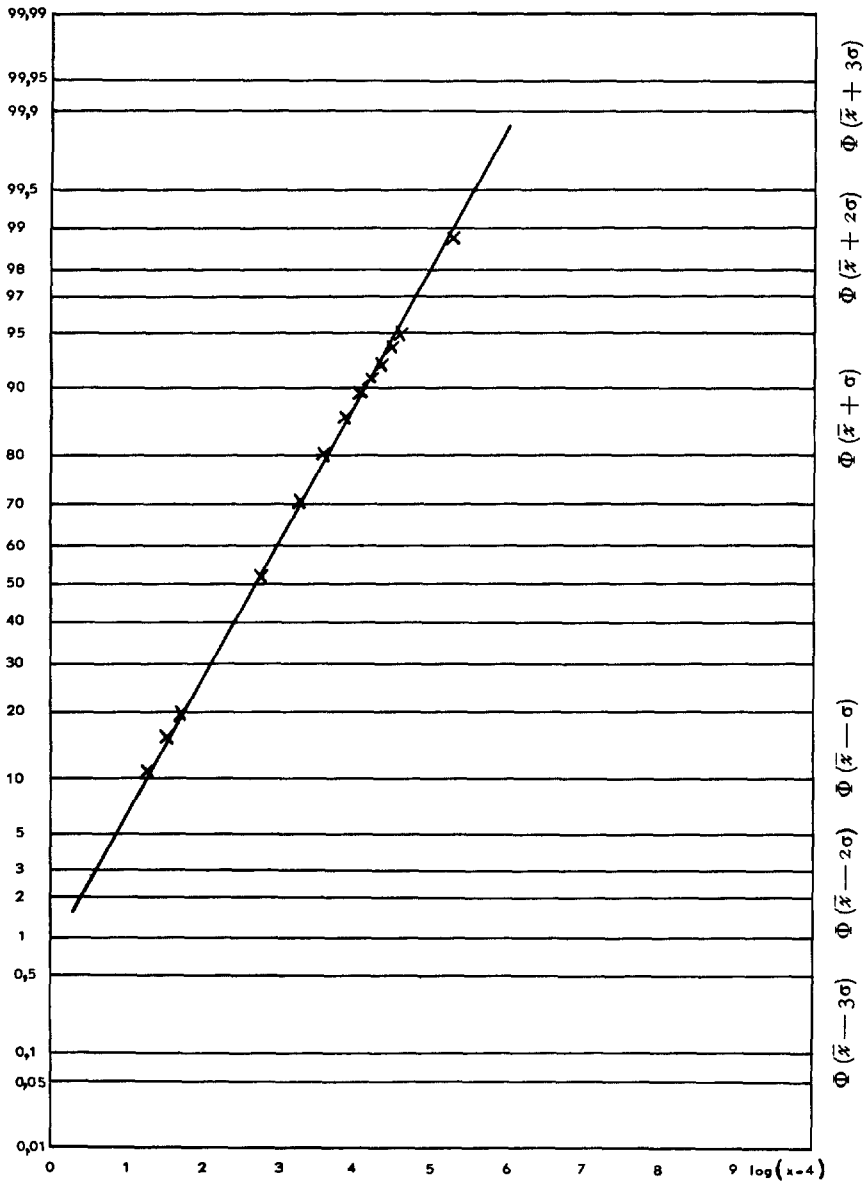


Fig. 5. Accident insurance.
Number of days of illness.
Occupancy Group B.



truncated inasmuch as about 80 % of the number of claims lie below the truncation point. Still the chosen d.f. covers the most interesting part of the claims as those claims are the source of more than 2/3 of the loss.

The d.f. is of the form $L(x - a;\ \mu,\ \sigma)$
where a is larger than the truncation point.

The observed claims that are less than a have not been taken away from the data used in the figure. For many purposes it will be necessary to make the truncation at a point that is larger than a which causes about 90 % of the number of claims and 50 % of the claims amount to be excluded.

The root of the difficulties is probably the fact that claims arising from personal injuries and other claims have not been separated.

Group A.

Number of claims above truncation point 3.200
$\mu^* = 0{,}9$
$\sigma^* = 2{,}2$

Group B

Number of claims above truncation point 1.286
$\mu^* = 1{,}3$
$\sigma^* = 2{,}1$

Estimation of the mean

The maxlike estimate of the mean is

$$(2)\quad \alpha^* = e^{\mu^* + \frac{\sigma^{2*}}{2}}, \text{ where } \mu^* = \frac{\Sigma \log x_i}{n} \text{ and } \sigma^{2*} = \frac{1}{n}\sum (\log x_i - \mu^*)^2$$

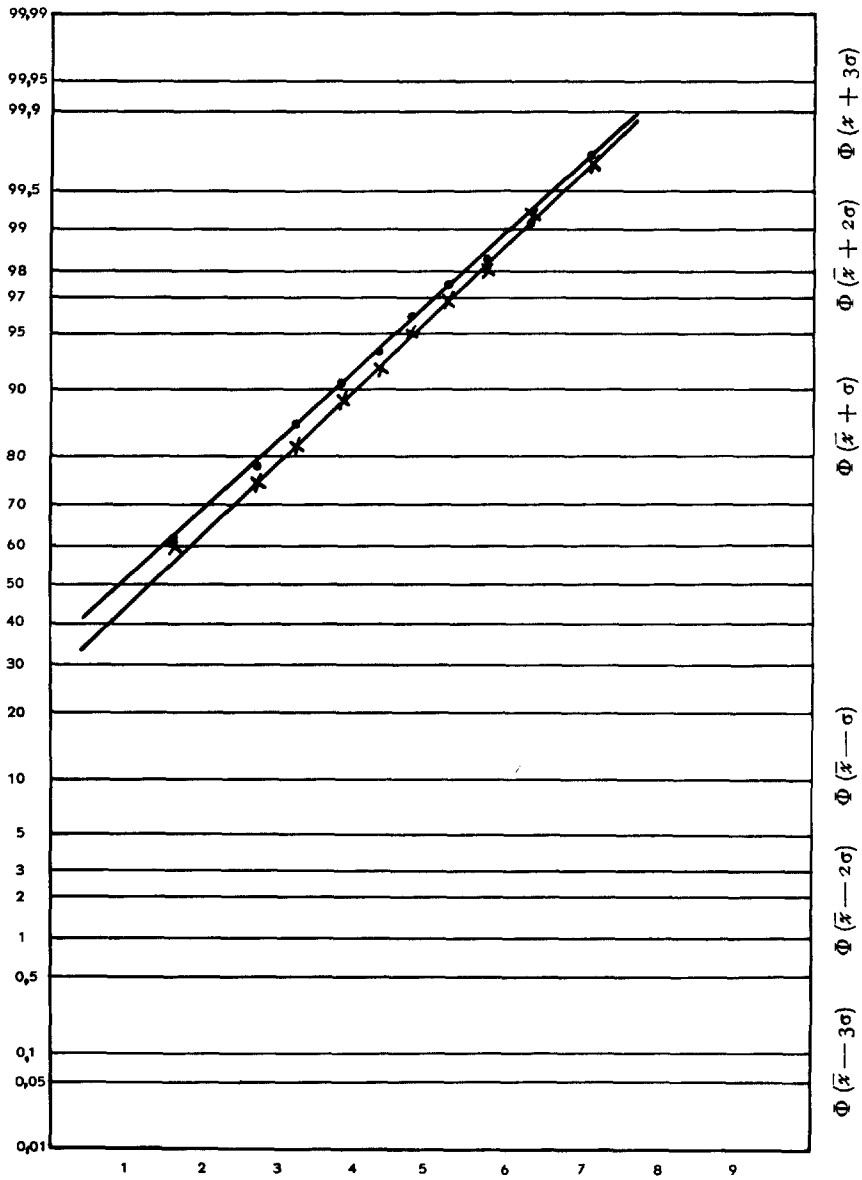
x = the size of the claim
n = number of claims

This estimate is asymptotically unbiased. If the number of claims is small, the estimate

$$(3)\quad \alpha^{**} = e^{\mu^* + \frac{\sigma^{2*}}{2} - \frac{\sigma^{2*2}}{4n}}$$

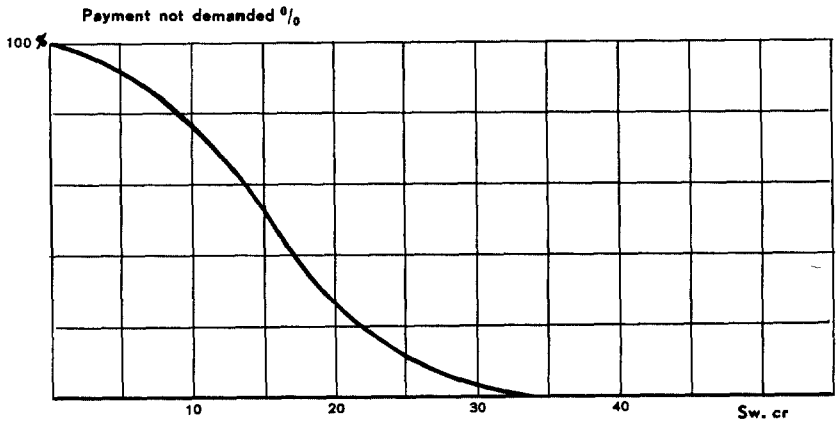
may be used as it converges much faster towards α than α^*.

Fig. 6. Motor third party insurance.
. Group A
$\times$ Group B



The influence of the smallest claim on the estimates (2) and (3) is approximately the same as the influence of the largest claim. It is therefore often desirable to diminish this great influence of the small claims as they are of no economic importance. Statistically the small claims are of a special nature as the insured often does not bother to demand payment for a small loss. This is easily seen by comparing two sets of statistical data where one set has no excess but the other set has a certain excess. The following fig. 7 shows how much the Swedes are willing to sacrifice rather than ask the company for payment.

Fig. 7. Small claims for which payment is not demanded.



There are several ways of diminishing the influence of the small claims on the estimates.

A natural way is to censor the sample at a value that is somewhat greater than zero. If the censoring point is denoted c and $L(c) \leq 0{,}2$ the estimates of μ and σ^2 are approximately

$$\sigma^{2*} = \frac{Y_2}{r} - \frac{Y_1 \cdot \bar{Y}_1}{r} + 0{,}4 \sqrt{\frac{Y_2}{r} - \frac{Y_1 \cdot \bar{Y}_1}{r}} \cdot \frac{m}{r} (\bar{Y}_1 - \log c)$$

$$\mu^* = \bar{Y}_1 - \frac{m \cdot \sigma^* \cdot 0{,}4}{n}$$

where $Y_1 = \Sigma \log x_i + m \log c$ for $x_i > c$

$Y_2 = \Sigma \log^2 x_i + m \log^2 c$ for $x_i > c$

$$\overline{Y}_1 = \frac{Y_1}{n}$$

m is the number of claims $\leq c$
r is the number of claims $> c$
$m + r = n$

Efficiency

The efficiency of α^* is denoted $e(\alpha^*)$.

$e(\alpha^*) \underset{n \to \infty}{\to} 1$, α^* is thus asymptotically efficient

The efficiency $e(\bar{x})$ of $\bar{x} = \frac{1}{n}\Sigma x_i$ is given by

$$e(\bar{x}) = \frac{\sigma^2 + \frac{\sigma^4}{2}}{e^{\sigma^2} - 1} \text{ if } n \text{ is not small}$$

As seen from the examples given in the fig. 1-6 $\sigma = 2$ is a rather common value. $e(\bar{x}) = 0{,}2$ for $\sigma = 2$ i.e. α^* and $\bar{x}$ give the same information if $\bar{x}$ is based upon five times as many claims as α^*. The point is illustrated in fig. 8, where it is seen that $\bar{x}$ varies much more with the different years than α^* in spite of the fact $e(\bar{x})$ for the data in fig. 8 is as great as 0,8.

The confidence interval of α

$$\ln \alpha^* = e^{\mu^* + \frac{\sigma^{2*}}{2}}$$

μ^* and σ^{2*} are independent and

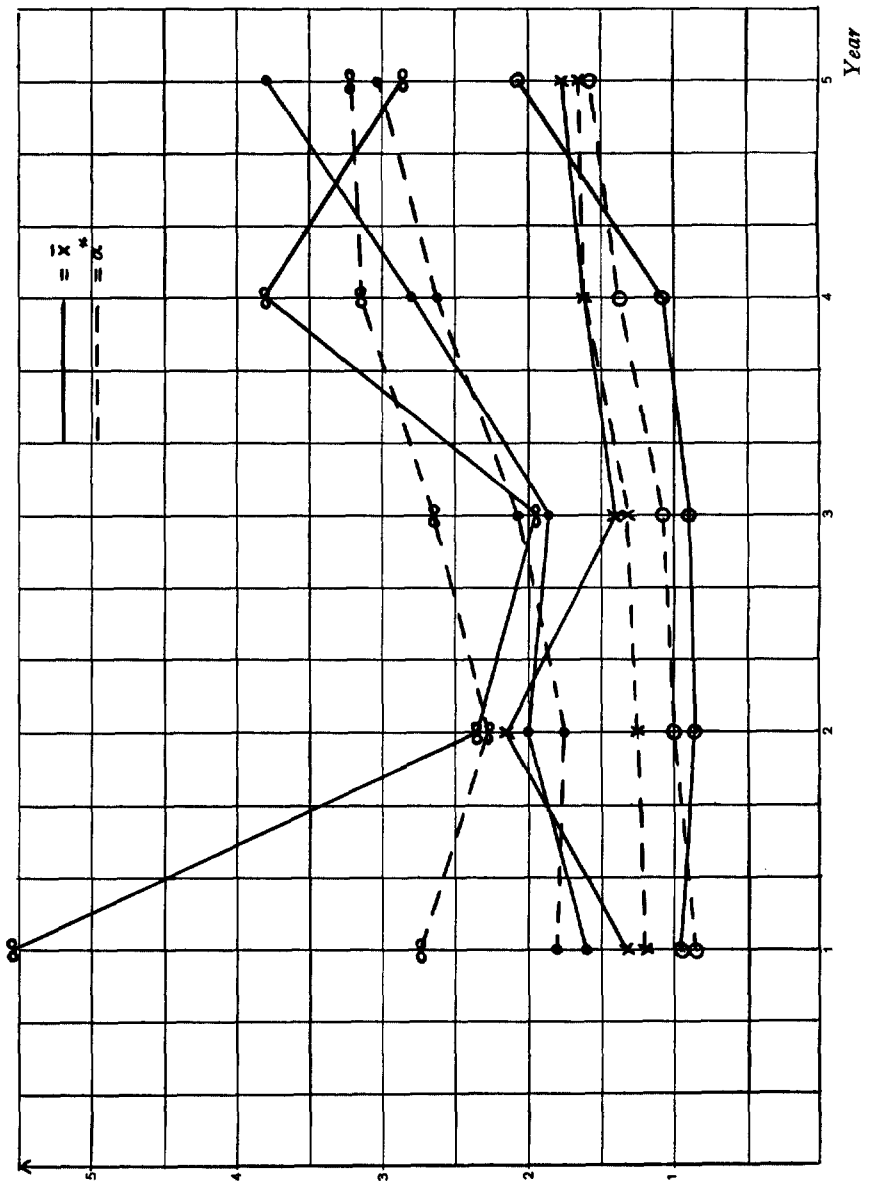
μ^* is normally distributed $N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$

and σ^{2*} Γ — distributed, with $n - 1$ degrees of freedom.

The d.f. of log α^* is then the convolution of a normal and a Γ — d.f.

I have not been able to do this convolution and there is small need for doing so as the Γ — d.f. very soon becomes normal as n grows. Log α^* is then approximately normal. A confidence interval of log α can thus be constructed by the Students t — d.f. and the confidence interval of α then follows

Fig. 8. Comparison between α^* and $\bar{x}$ as estimates of the mean claims amount. α^* an $\bar{x}$ are calculated from accident insurance data.



$$e^{\mu^* + \frac{\sigma^{2*}}{2}} \pm t_p \cdot \sqrt{\frac{\sigma^{2*}}{n-1} + \frac{\sigma^{2*2}}{2(n+1)}}$$

where t_p is the p-percent value of the t — d.f.

The d.f. of the estimate of the risk premium

The basis of the estimate of the risk premium is the d.f. of the estimate of the total loss R.

$R^* = v \cdot e^{\mu^* + \frac{\sigma^{2*}}{2}}$. R^* has the following frequency function

$$f_R(x) = \sum_1^\infty \frac{nv}{v!} \cdot e^{-n} \cdot \frac{1 \cdot \sqrt{v}}{\sqrt{2\pi}\, S \cdot x} \cdot e^{-\frac{(\log x - \log v - M)^2}{2S^2/v}}$$

where v = number of claims

$$n = E(v)$$

$$M = \mu + \frac{\sigma^2}{2}$$

$$S^2 = \sigma^2 + \frac{\sigma^4}{2}$$

where again $\mu^* + \frac{\sigma^{2*}}{2}$ is presumed to be normal.

To make the formula more lucid $n - 1$ and $n + 1$ have been replaced by n.

The formula for $f_R(x)$ is not very pleasing but it does at least not contain any convolutions and it might not be too difficult to find a good approximation.

When it comes to convolutions the lognormal d.f. has very nasty qualities. It is not even possible to find an explicit expression for the characteristic function.

Summary

The lognormal d.f. shows a good fit—especially for large values of the variable—in many branches of non life insurance.

This d.f. has many nice qualities of which the following are of great value in the practical work.

1. The estimate of the mean is efficient and reasonably easy to use.
2. It is possible to construct a confidence interval for the estimate of the mean.
3. The frequency function of the total claims amount may be expressed as an infinite series, including only simple and wellknown functions.

The lognormal d.f. has less nice qualities too. The following seem to be the most annoying.

4. The integral in the characteristic function is not solvable and the convolution cannot be expressed explicitly.
5. The less good fit for the small values of the variable calls for adjustments which often causes a good deal of work.