NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 03 2016

ISSN: 1500-4066

January 2016

Discussion paper

Log-normal creaming and the likelihood of discovering additional giant petroleum fields

BY

Jostein Lillestøl AND **Richard Sinding-Larsen**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Log-normal creaming
and the likelihood of discovering additional giant petroleum fields

Jostein Lillestøl[1]

Norwegian School of Economics

Richard Sinding-Larsen[2]

Norwegian University of Science and Technology

Jan. 20 , 2016

Abstract

This paper considers sampling proportional to expected size from a partly unknown distribution. The applied context is the exploration for undiscovered resources, like oil accumulations in different deposits, where the most promising deposits are likely to be drilled first, based on some geologic size indicators ("creaming"). A Log-normal size distribution turns out to have nice analytical features in this context, and fits available data reasonably well. The theoretical and practical consequences for the accumulation of knowledge on the underlying distribution based on this scheme, named Log-normal creaming, are explored in some detail. The theory is applied on the prediction of remaining oil accumulations to be found on the Norwegian Continental Shelf.

Keywords: Log-normal distribution, sampling proportional to size, resource prediction

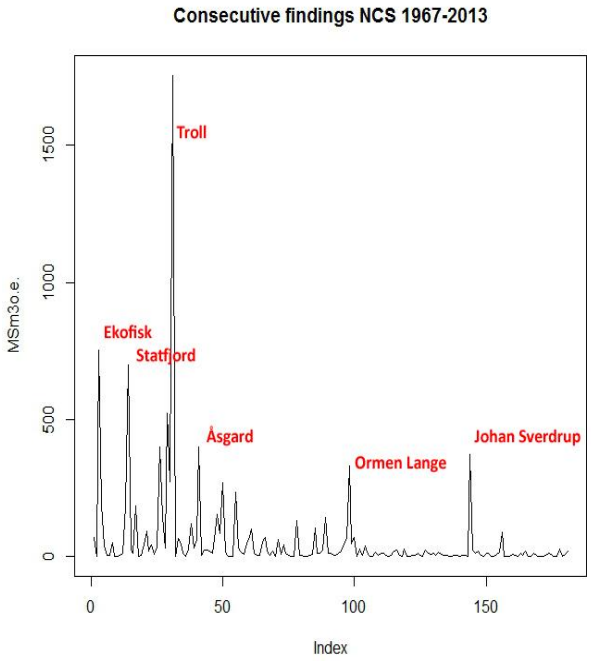
[1] Department of Business and Management Science, Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway; e-mail: jostein.lillestol@nhh.no

[2] Department of Geology and Mineral Resources Engineering, Norwegian University of Science and Technology, Sem Sælands veg 1, N-7491 Trondheim; e-mail: richard.sinding-larsen@ntnu.no

Log-normal creaming and the likelihood of discovering additional giant petroleum fields

1. Introduction

This work is dealing with the opportunities offered by the Log-normal distribution for predicting remaining oil deposit sizes. It is well known that observed size distributions from various regions (e.g Mexican Gulf and Norwegian Continental Shelf) are well fitted by a Log-normal distribution. However, the fitted distribution cannot be used directly for prediction purposes, since the observations are not randomly sampled from the assumed unknown parent distribution. By plotting the size of discoveries in sequence, we typically see a declining pattern, consistent with the behavior of exploring the most promising fields first, so-called creaming. As a representative example take the following graph showing the sizes of all 181 fields and discoveries in resource categories 0-9 on the Norwegian Continental Shelf, as reported in RNB 2014.[3]



The six largest discoveries are named in the graph, among them the first Ekofisk in 1969, the largest Troll in 1979 and the latest Johan Sverdrup in 2010.[4] They are all so-called giant fields, of size well above the borderline given by Halbouty (2001), which corresponds to about 79 million Standard cubic meter oil equivalents (MSm3o.e). [5]

[3] The Norwegian Petroleum Directorate is acknowledged as the source for the data used in this paper. The data is updated values from the petroleum resource account as of 31.12. 2013, prepared as input to the Revised National Budget (RNB 2014). For resource classifications see: www.npd.no/global/engelsk/5-rules-and-regulations/guidelines/ressursklassifisering_e.pdf.

[4] Johan Sverdrup (1816-1892), Norwegian prime minister (1884-1889), named "father of parlamentarism".

[5] More specifically a giant oil field contains more than 500 million barrels of oil recoverable, and a giant gas field contains more than 3 trillion cubic feet of gas.

In general this means that, at any point in the exploration history of a region, the empirical size distribution obtained so far, will give a too optimistic representation of the true underlying distribution, and therefore provides too optimistic probabilities of new discoveries of large sizes. The declining feature may possibly be well represented by sampling with probabilities increasing with size. We will study our problem within the context of independent Log-normal variables, which may be imagined as sampling with replacement from an infinite population. This model turns out to have interesting features in relation to the creaming issue. This was first noted and explored by Meisner and Demirmen (1981). The decline over time may of course be due to the fact that we are really sampling without replacement from a finite population, which gets depleted over time, with mostly smaller sizes left due to creaming. The problem of finite population proportional to size sampling has previously been studied in the exploration context by a two-stage model due to G. M. Kaufman, see Kaufman, Balcer and Kruyt (1975), and Lee and Wang (1985). Other contributions to this area are Barouch and Kaufman (1967), Schuenemeyer and Drew (1983), Andreatta and Kaufman (1986), Kaufman (1992) and Chen and Sinding-Larsen (1994). Charpentier et.al. (1995) provides a bibliography.

Our view is that the infinite population context offers the best framework for understanding some of the basic inference and prediction issues related to creaming. These issues are previously studied within the context of Beta-models, see Lillestøl and Sinding-Larsen (2015). Creaming models of Log-normal and Beta type have different features, and both deserve to remain in the toolbox. Note that our aim is to obtain reasonable predictions of remaining resources and not to provide the best possible explanation of the exploratory behavior itself.

The paper is organized as follows: In section 2 we present the main features of the Log-normal distribution with respect to sampling proportional to size, or any power of size. The result is then confronted with data from the Norwegian Continental Shelf, and some practical implications for population inference and prediction of remaining discoveries are discussed. In section 3 we deal with the sequential nature of our problem and the shortcomings of models with constant creaming factor for the infinite population case. We suggest and explore a model with exponential decaying creaming factor. The model is then illustrated with some probability calculations concerning future discoveries. In section 3 we explore the fact that the data, despite the good fit to Log-normal, seemingly contains an excess of small discoveries. This may be better explained by a mixture of two processes, a creaming process and a pure noise process. For this we suggest a split Log-normal model. This is studied extensively with our data, from which various prognostic measures are calculated. In two appendices we provide some theoretical results that may come to use in connection with creaming populations (or samples obtained by creaming) assumed to be Log-normal.

2. Facing creamed Log-normal data

The Log-normal distribution is frequently used as model for the distribution of petroleum deposit sizes in a population of deposits, the reason being that it has nice theoretical properties and fits available data reasonably well. The probability density is given by

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{x} \cdot e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2} \; ; \; x > 0$$

We will consider i.i.d. sampling, justified by a large population and where it does not matter whether the sampling is with or without replacement. Some prefer to imagine an infinite population. However, we will assume so-called creamed sampling, where larger sizes are more likely to be selected than smaller. In particular we assume sampling with probabilities proportional to x^k for some k>0, where k is named the creaming factor. In this case the observations will appear as coming from a distribution proportional to $x^k f_0(x)$, which means that it can be written as

$$f_k(x) = x^k f_0(x) \cdot e^{-\left(k\mu+\frac{1}{2}k^2\sigma^2\right)} \; ; \quad x > 0$$

The normalizing factor follows from the log-normal moment formula with wrt. $f_0(x)$:

$$E_0(X^k) = \int_0^\infty x^k f_0(x)dx = e^{k\mu+\frac{1}{2}k^2\sigma^2}, \text{ k real}$$

Having a distribution with exponential tail gives sense to proportional to size sampling even if the size space is (0, ∞). Writing $x^k = e^{k \cdot log(x)}$ and collecting all exponential terms in $f_k(x)$, we see that the exponent again becomes a quadratic function of log(x), and thus $f_k(x)$ has to be a Log-normal distribution as well. By completing the square we see that

$$f_0(x) \text{ Log-normal}(\mu, \sigma) \rightarrow f_k(x) \text{ Log-normal}(\mu + k \cdot \sigma^2, \sigma)$$

This means that the nth moment with respect to $f_k(x)$ is

$$E_k(X^n) = e^{n(\mu+k\sigma^2)+\frac{1}{2}n^2\sigma^2}$$

so that $E_k(X^n) = E_0(X^n) \cdot e^{nk\sigma^2}$. In particular note that $\mathrm{k} = \frac{1}{\sigma^2}\log(E_k(X)/E_0(X))$.

Note that our result holds for all real k, positive or negative. This means that we can use the result in both directions, forwards (theoretical) and backwards (inferential):

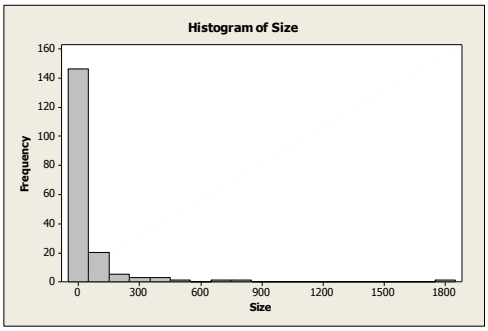
For k=1: If we have assumed proportional to size sampling from a Log-normal population, then the sampled result will appear as regular sampling from a Log-normal population with expected log-size μ increased by σ^2.

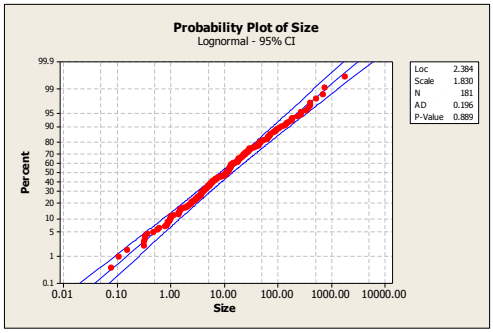
For k=-1: If we have sampled and the result is consistent with a Log-normal distribution, and we assume that the sampling was performed proportional to size, then we can infer that it came from a log-normal population with expected log-size μ decreased by σ^2.

In the latter case we may use estimates of (μ, σ) derived from the sample, which provides estimates for the parent population as well.

Note also that if we have observations under both regimes, k-creaming and 0-creaming, we may use the formula for *k* above. Just replace the two expectations by their respective observed averages and replace the common variance by a joint estimate.

Example: Consider the size of all 181 discoveries and fields s on the Norwegian Continental Shelf as reported by year 2014. The histogram shows a long-tailed distribution and the probability plot indicates a very good fit to the Log-normal distribution



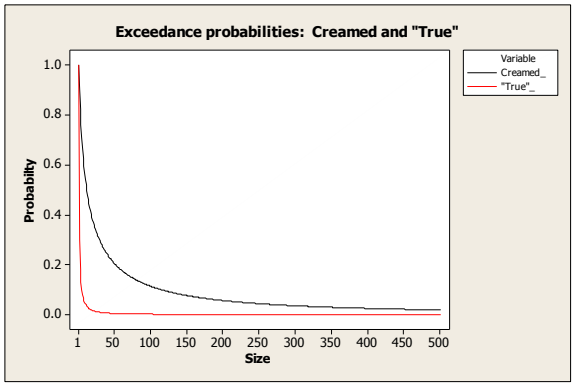


Descriptive Statistics: Size; Logsize

Variable	Count	Mean	StDev	Variance	Minimum	Maximum
Size	181	56.0	165.9	27510.0	0.0800	1755.7
Logsize	181	2.384	1.830	3.350	-2.526	7.471

We see a Log-normal (μ, σ) with $\mu = 2.383$ and $\sigma = 1.830$. Assuming sampling proportional to size, we infer that the parent distribution is Log-normal(μ, σ) with $\mu = 2.384 - 3.350 = -0.966$ and $\sigma = 1.830$. The standard errors of these two estimates are 0.376 and 0.096 respectively.

The two distributions are illustrated by their exceedance probabilities (i.e. one minus the cumulative distributions) as follows:

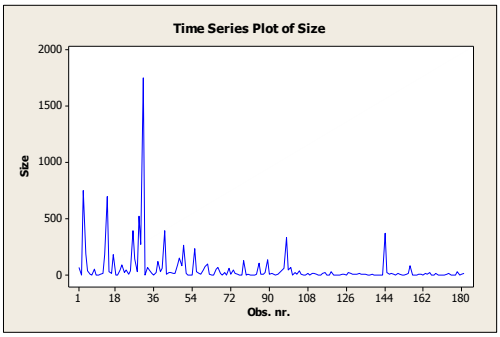


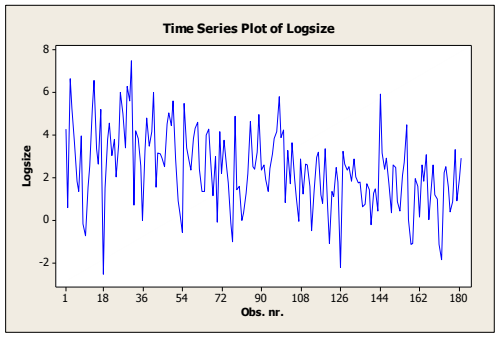
We see that with the derived «true» distribution, it is very unlikely to observe what we actually did, like Ekofisk (753.39), Statfjord (699.39), Troll (1755.66) and even Johan Sverdrup (373.84). From this we may seemingly conclude that the creaming process is not likely to be as strong as proportional to size. However, we will see that this result may be due to the assumption of constant *k* over the whole observation period.

In the probability plot the slope of the line corresponds to the standard deviation σ of the log-sizes. The creamed and the true distribution line will therefore theoretically be parallel with displacement corresponding to $k\sigma^2$.

In some cases a truncated log-normal distribution will give better fit to the data. It turns out that a conclusion similar to that above also holds for log-normal truncation in general, i.e. for the conditional distribution of X given $a< X \leq b$ (see Appendix).

The above analysis does not take advantage of possible time series information in the data. The existence of time series information is apparent from a time series plots of size and logsize.





We see that the level is declining over time. This may also be illustrated by splitting the data in four quarters, here named Q1, Q2, Q3 and Q4 with 45+45+45+46=181 observations. The descriptive statistics are

Descriptive Statistics: Sizes Q1, Q2, Q3, Q4

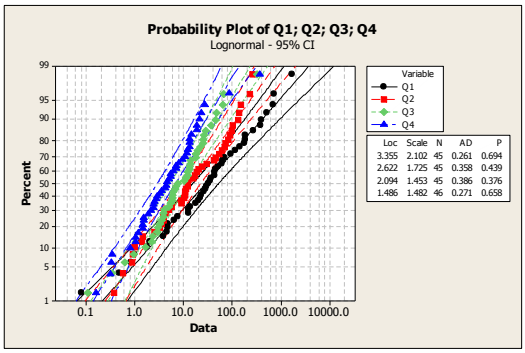
Variable	Count	Mean	StDev	Variance	Minimum	Maximum
Size-Q1	45	143.6	303.6	92201.8	0.0800	1755.7
Size-Q2	45	42.87	61.75	3813.48	0.370	268.16
Size-Q3	45	21.59	50.04	2504.27	0.110	332.00
Size-Q4	46	16.63	55.56	3087.13	0.160	373.84

Descriptive Statistics: LogSizes Q1,Q2,Q3,Q4

Variable	Count	Mean	StDev	Variance
LogSize-Q1	45	3.355	2.102	4.418
LogSize-Q2	45	2.622	1.725	2.976
LogSize-Q3	45	2.094	1.453	2.111
LogSize-Q4	46	1.486	1.482	2.197

We see a decay in the mean (log)sizes. The standard deviation of (log)sizes is higher for the first quarter Q1 compared with the other three quarters, which are fairly similar. A formal test for equal variances of the logsizes shows that the difference is just, but not strongly, significant, with P=0.038 for Bartlett's test, although Levenes's test gives P=0.067.

Log-normal probability plots of each of the quarters follows:



Here we see that a Log-normal distribution is justified for each of the quarters (high P-values), and that the line for the first quarter is sloping somewhat different from the other three.

3. Modelling the time pattern: Exponential decaying creaming

A variety of opportunities exist for modelling the effect of the sequence order, for convenience referred to here as time. If we believe that the data for the first quarter of observations disrupt our model, a possibility may be to leave them out and start with the second quarter and use the model from then on. This will not be pursued here.

The decline over time may of course be due to the fact that we are really sampling without replacement from a finite population, which gets depleted over time, with mostly smaller sizes left due to creaming. We have made efforts to reproduce pattern similar to our exploration sequence by sampling without replacement from simulated finite populations. By varying the population size N and fixed creaming factor k, we were not able to mimic the pattern. Let us therefore continue within the theoretically more convenient i.i.d. framework, where the main issues are more easily exposed.

Let us first address the level change, and ignore the possible change in variance brought to our attention above. One modelling opportunity is the following: Assume that the creaming factor k is decaying over time. A tractable model is

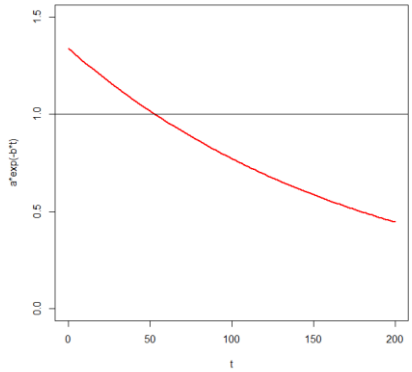
$$k(t) = \alpha \cdot e^{-\beta t}$$

This means that when we at time t sample proportional to $x^{k(t)}$, the observation is as if it is coming from a Lognormal($\mu + k(t) \cdot \sigma^2, \sigma$) distribution, where μ is the true mean log-size of the Log-normal parent distribution. The data give the following maximum likelihood estimates with corresponding standard errors in parenthesis

$$\hat{\mu} = -0.020 \ \ (1.131)\ , \hat{\sigma} = 1.688\ (0.089), \hat{\alpha} = 1.338 \ \ (0.429), \hat{\beta} = 0.0055\ (0.0025)$$

We see that the estimate of β is significant, confirming the decaying pattern of k(t). Note that the estimate of α, which corresponds to the initial k, is above one, but the standard error does not preclude α being less than one. We see that the estimate of μ is higher than for the model with fixed k=1, but its standard error is now inflated. With these estimates, the occasional large discoveries in the data become less unlikely than with the fixed k=1 model, but still fairly unlikely. The estimate of σ is slightly reduced compared to the fixed k=1 model.

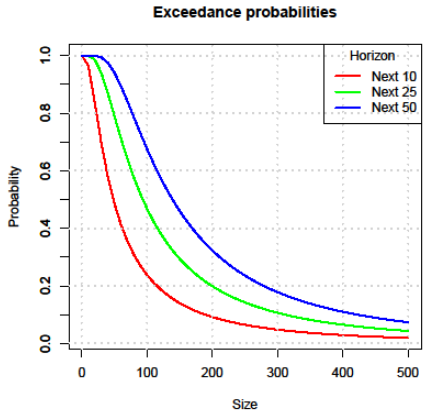
The estimated exponentially decaying function is shown in the following graph:



The graph is consistent with sampling with probabilities higher than proportional to size up about t=50, i.e. less than 1/3 of the observation period. This explains the previous seemingly conflicting evidence about k when judged in the fixed k context.

In connection with the division of observations into quarters above, we made the following crude judgment: Imagine that the first quarter was a k-creamed sample, while all observations are taken as representative of the true population. Taking the size averages in each of the two (partly overlapping) groups and using the derived formula for constant k above, we get k=0.3 which, given its construction, is expected to be a lower estimate. We see that this conforms reasonably well with the graph.

The above estimates may be extrapolated to give the probabilities of exceeding specific size levels for consecutive future discoveries. In practice we may be interested in the probabilities that the next h discoveries contain at least one discovery over a specific size level x. These are shown in the following graph for horizon h=10, 25, 50, for the size range up to 500.



For at least one beyond Johan Sverdrup (373.84 MSm3o.e.) the probabilities are about 0.03, 0.07 and 0.12 respectively. The corresponding probabilities of being beyond the (somewhat arbitrary) borderline of being a giant field (79 MSm3o.e.) are 0.31, 0.64, 0.84 respectively.

Another opportunity for modelling the time is the following: A closer analysis of the data indicates that the data-generating process possibly generates two types of data: Deposits consistent with the creaming and some extra small deposits, due to nearby opportunities and convenience. These may be separated and the true creaming effect revealed through some time series analysis. This may be done in essentially two different ways:

1. Assume for each discovery that it is (not-creamed, creamed) with probabilities (1-p,p), where the creaming factor k(t) may be time-dependent.
2. Choose a low size level and model discoveries below and above this level separately: the low discoveries are coming from non-creaming (k=0) and the above discoveries are coming from time-dependent creaming (k(t)>0).

Both schemes are able to mimic the declining sequence pattern of our data fairly well. We have found the second opportunity the most promising one for practical use, and it will be explored in some detail in the next section. The two schemes may also be used as basis for exploration of the finite population case. For this the population size N has to be chosen large enough, so that sizes of

magnitudes of those observed are likely to be included. For our data N=1000 is likely to be insufficient. It turned out that N=2000 and p=0.9 may give patterns similar to those observed. However, repeated simulations showed a variety of visual expressions.

In practice there may also be some correlation in the data, due to concurrent exploration of prospects with similar geological indicators. This may have some relevance to risk calculations, and can be studied by introducing bivariate Log-normal creaming. Our results in this direction will be reported elsewhere.

It is a challenge to really understand and possibly separate the effects of creaming, correlation and finiteness of population, when we actually do not know the population size. However, we may ask: Do we have to separate, when the aim is just to obtain reasonable predictions of remaining resources, and not to provide the best possible explanation of the exploratory behavior itself?

4. Modelling the time pattern: A two-regime model

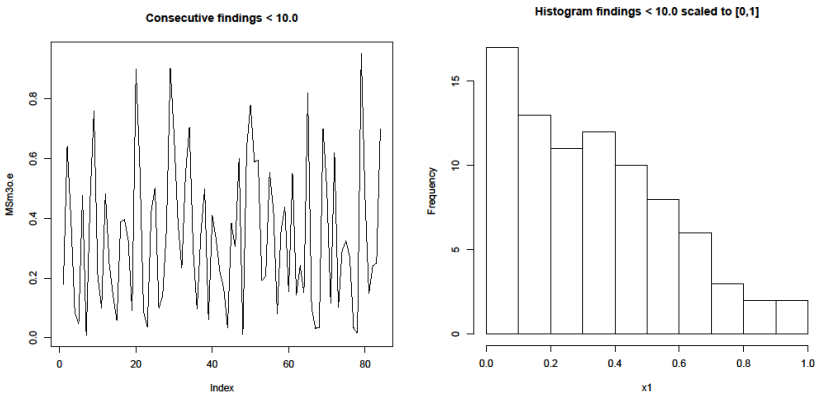
We have seen a pattern consistent with creaming, i.e. explore the most promising prospects first. However, for some reasons related to opportunity and convenience locations with no particular superior indications are quite often explored as well. We therefore expect to see a mixed pattern that is not likely to be explained by a simple statistical model. Despite the fact that a Log-normal model seemingly gave a good fit for all the observations, we have seen indications that improvements can be made. We will therefore examine the possibility of modelling the pattern as a mixture of two processes, one for small size discoveries stable over time and one representing the remaining, where also the large surprises are included. This is done by splitting the data in two groups according to a split point c, and keeping the order of observations within each group. The stakeholders may have asked the question: Which size c is the lowest we want to explore? This may or may not be comparable to our c, which is entirely determined by the data-features. [6]

For each group we scale down the observations, the first group from the interval (0, c] to (0,1], the second group from (c, ∞) to (0, ∞). This invites us to look for a distribution model of Beta-type for the first group and one of Log-normal type for the second group.

There will of course be a formal inconsistency between this assumption and the one in the previous section, in that the original data and the grouped downscaled data cannot both be Lognormal, and the proportional to power of size conception now applies to the downscaled data. However, seen as an exercise in prediction based on a well-fitted model, this is not worrisome. What we need here is that future observations still appear as coming from the Log-normal distribution with declining expectations due to creaming, and they are going to be modelled and estimated from data, with no need to go via the true expectation of the parent distribution. There is an alternative achieving consistency: With the original Log-normal assumption, the upper group will have a so-called truncated Log-normal distribution, which turns out to have the same creaming property as the Log-normal distribution itself (see Appendix). The calculations here get somewhat more involved, but do not lead to substantially different results.

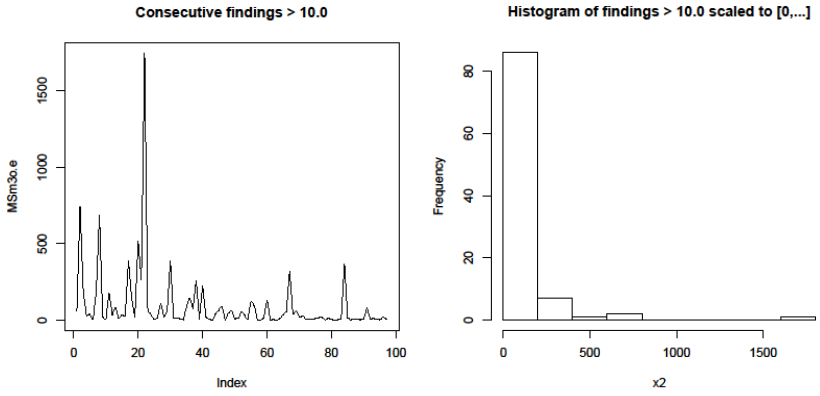
[6] In reality the range of the two processes may overlap, but taking this into account just adds complexity, and will typically not improve predictions.

Judged from the key characteristics above a reasonable tentative splitting point is c=10.0. Below are histograms and time series plots for the scaled observations of consecutive discoveries in each group. The aim is to reveal some kind of stable features that can be used for extrapolation to the following years.

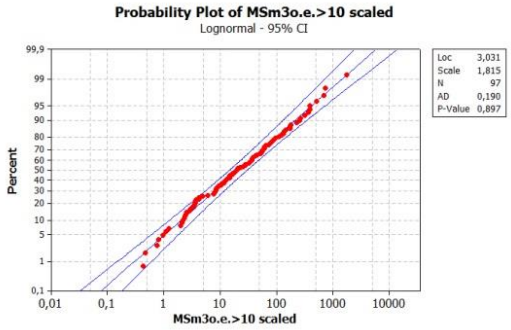


For the low group (<10.0) we see above a stable random pattern similar to background noise. The size distribution is close to triangular, i.e. Beta(1,2). In fact, the maximum likelihood estimates of the two shape parameters are

shape1	shape2
0.9870958	1.8771440



For the high group (>10.0) the main feature above is the peaks that seemingly occur less frequent and diminish in magnitude over time. The size distribution has a long right tail, and among the distributions tried out (Gamma, Weibull and Log-logistic), the Log-normal turned out best. The probability plot below shows that the fit is very good, and that the distribution explains the high observations as well.



The maximum likelihood estimates of the corresponding normal parameters are

```
 meanlog    sdlog
3.031424 1.805332
```

Note that this lognormal fit is before we have taken into account the possible decaying size pattern, but we expect that the lognormal will fit just as well after correction for this.

Now we address the estimation of the two time decaying patterns (frequency and size). We will use year as explanatory variable in both cases. We could alternatively use the number in the sequence.[7] Neither choice is without problems, related to the different number of discoveries within each year and the assigned order of discoveries within each year respectively. For the decaying frequency we employ a logistic regression model explaining the probability that a discovery in a particular year will belong to the high group (Y=1) or the low group (Y=0). The maximum likelihood regression estimates turned out to be

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 101.30710   23.72662   4.270 1.96e-05 ***
year         -0.05070    0.01189  -4.265 2.00e-05 ***
```

This shows that we approximately have

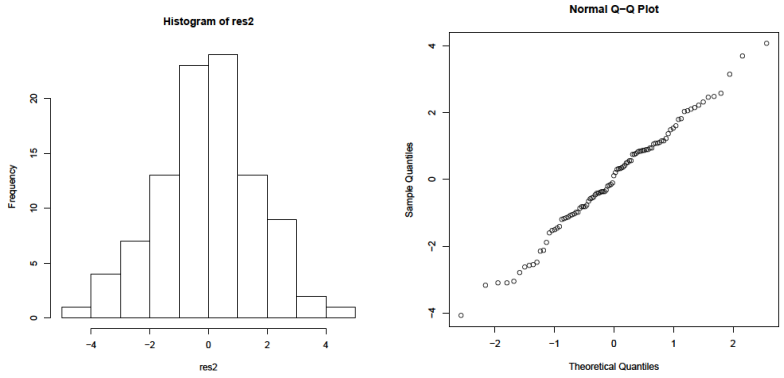
$$P(Y=1)=\frac{e^{-0.05\cdot(year-2000)}}{1+e^{-0.05\cdot(year-2000)}}$$

For the decaying expected sizes in the high group we use an ordinary linear regression for log(size) with year as explanatory variable. This gave the following

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 127.23656   25.11995   5.065 2.00e-06 ***
year2        -0.06239    0.01262  -4.945 3.28e-06 ***
```

[7] Note that there are some formal inconsistencies between the assumption made here and in the previous section, in that the original data and the grouped and scaled down data cannot be both lognormal, and the proportional to power of size conception is now applied to the downscaled data. However, seen as an exercise in prediction based on good model fit, this is not worrisome, as we use direct estimation of the expected log-sizes at each stage, with no need to go via the true expectation of the parent distribution. What we need is just that future observations still appear as coming from the log-normal distribution.

Residuals plots show that residuals from this regression are approximately normal, thus justifying a Log-normal model for sizes also after correction for the time decay.

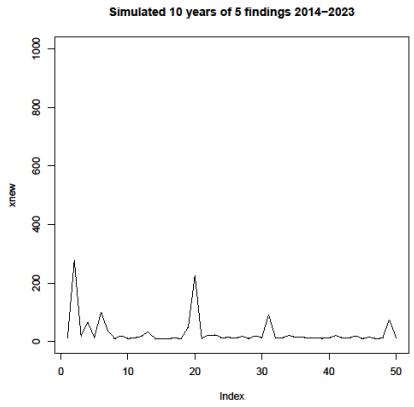


Equipped with this set of models and our estimates based on the data from the period 1967-2013 we can now simulate outcomes for the following years, say 5 discoveries a year over 10 years, totaling 50 observations. This is done in the following steps:

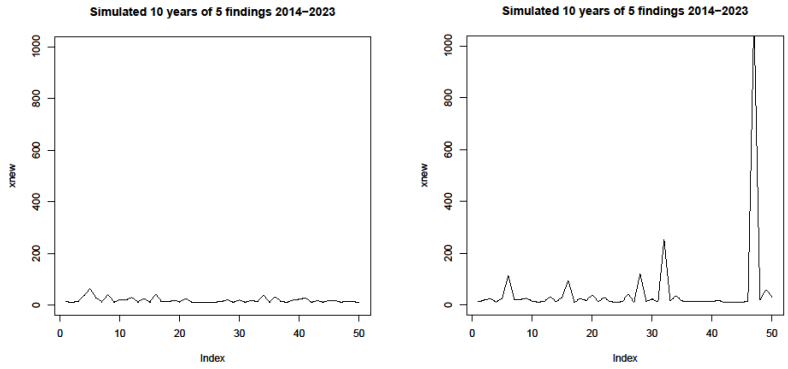
1. Extend the variable “year” to the given horizon
2. Estimate the mean log(size) for the high group up to the given horizon
   - keeping the standard deviation fixed (may be modified as well)
3. Generate the sequence of Y’s according to the estimated logistic model
4. For each Y=0 generate X according to the estimated Beta-model for the low group, and scale up accordingly (i.e. multiply by c)
   For each Y=1 generate X according to the lognormal model with meanlog, sdlog) found in pt.1, and scale up accordingly (i.e. add c)

This algorithm can in principle be implemented by vector computation without iterations.

Here is a typical simulated outcome based our models, using maximum likelihood estimates:



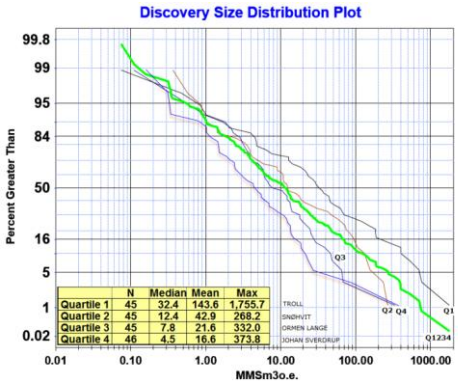
However, our modelling and the data do not rule out scenarios like the following:



Consequently many runs are necessary to be able to judge the uncertainty.

Some comments on the choice of split point c and alternative modelling:

We have here chosen c so that the sequence of observations below c look like a stationary sequence, i.e. ideally one with constant level and spread with no autocorrelation. The choice could be done by graphics alone, or supported by formal testing. An alternative to this may be to first split the data according to time, say in four quartiles, and then judge the (dis)similarity of the left tail of the four distributions. This is illustrated in the log-scale plot below, which tells the observations in each quartile is approximately Log-normal with a close to common left tail, and that the distributions separate slightly above 1.0. This may suggest that c=10.0 is too high. With a choice closer to 1.0 we will assign more observations to the time-dependent part of the model, with some risk of adding just noise to that one. This plot also exposes the opportunity of modelling the quartiles separately. For future predictions it may then seem reasonable to disregard the distant past and use only the last quartile of data. However, by doing this we base our prediction on scarce data, possibly missing the inherent time-dependent outlier features (surprises) that is revealed only by looking at all data.



We will now valuate the chances of one or more large discoveries in years to come. As level of a large discovery we take the Johan Sverdrup field of size JSv=373.84. Based on our estimated model we have simulated 25 and 50 discoveries. With assumed 5 discoveries each year this corresponds to 5 and 10 years respectively. With 10 discoveries a year the latter case corresponds to 5 years. The following table shows estimates of the following for the given period:

- probability of at least one discovery of size larger than JSv
- expected number of discoveries of size larger than JSv
- expected median discovery within the period

- expected maximum discovery within the period
- expected total discoveries for the period regardless of size
- expected median discovery, given that at least one discovery of size larger than JSv
- expected maximum discovery, given that at least one discovery of size larger than JSv
- expected total discoveries, given that at least one discovery of size larger than JSv

The calculations are based on c=10.0 as split point and 100 000 repeated simulations.

Discoveries ahead	25: 5 years of 5	50: 10 years of 5	50: 5 years of 10
P(at least one > JSv)	0.050	0.075	0.097
Expected no. > JSv	0.051	0.077	0.102
Expected Median	4.89	4.55	4.11
Expected Maximum	117	154	183
Expected Total	294	514	588
Exp. Med conditional	5.23	4.66	4.23
Exp. Max conditional	851	820	841
Exp. Tot. conditional	1103	1282	1364

We see an estimated chance of 5% of hitting a new deposit of size at least Johan Sverdrup in the next 5 years, which increases to about 7.5% for 10 years. This assumes 5 discoveries a year. With a more intensive activity of 10 discoveries a year the chance is increased to 9.7% in 5 years. Note also that the expected total in the case of 5 year of 10 is twice that of 5 years of 5 discoveries.

One may argue that the latter two cases both may correspond to the same 50 discoveries compressed on the time scale, and therefore should have the same probability and expectations. This will be so if we had used sequence number instead of year as explanatory variable, but then the estimates will be affected by the order of the data within each year. A specific reporting practice, say monotone, increasing or decreasing sizes, will bias the estimated probabilities and expectations. There are no apparent patterns of this kind in the data. However, by using sequence number as time variable instead of year, we get in the case of 5 years of 5 discoveries that P(at least one > JSv)=0.023, i.e. down from 0.05 in the case of year as time variable.

It may be of some interest to compare the results with the ones obtained by the unordered data, i.e. neglecting the decaying pattern over time. The corresponding probabilities are given in the table for the case of no level decay (keeping the frequency decay) and for the case of both decays left out.

Discoveries ahead	25: 5 years of 5	50: 10 years of 5	50: 5 years of 10
P(at least one > JSv)	0.050	0.075	0.097
with no level decay	0.336	0.530	0.551
No level and freq. decay	0.533	0.782	0.782

For the time-independent model we must have $0.551=1-(1-0.336)^2$ in the second row of the table, since this corresponds to no larger than JSV in two independent groups of 25 discoveries. This is so in the third row as well, and here the cases “10 years of 5” and “5 years of 10” of course become equal.

We see that there is a considerable sensitivity to modelling assumptions. This is not surprising. Probabilities of type “at least” will in theory involve products of factors, here 25 or 50, and even small differences in the base probability will then magnify. We will explore some aspects of this sensitivity.

The sensitivity of results with respect to the choice c is shown for the case of the next 25 discoveries (5 years of 5) in the following table:

Split point c	1.0	2.0	3.0	5.0	**10.0**
P(at least one > JSv)	0.145	0.100	0.096	0.077	**0.050**
Expected Maximum	241	184	179	153	**117**
Exp. Max conditional	919	838	852	824	**851**

Reducing c from the 10.0, which we have argued is the reasonable choice given the data, we see that the probabilities increases, so that the probability given for c=10.0 also will be the conservative (least optimistic) one among them. Data have told us that there is no need to be more conservative by taking c>10.0. The increase in probability may be understood as follows: The probability is determined jointly by frequency and size. With a reduction of c we get a larger fraction of observations under the high regime, which provides more opportunities for a large one to occur in future predictions. This is not fully compensated by the left-shift of the estimated predictive distribution, due to the inclusion of more small observations in the high group.

The main justification for our split approach was linked to stationary for the low group and a reasonable decaying model for a high group of less noisy observations. The choice of c was then an empirical question on how to balance this, based on looking at the data only. However, note that the choice of c also affects the parameter estimation. With a very small or very large c, we get disparate number observations in the two groups and possibly bad estimates of the distribution of the minority group and the probability of belonging to either group. The model fit may then be sensitive to group transfer of observations around the chosen c.

We have used the size of the last found giant field on the Norwegian Continental Shelf JSv=373.84 MSm3o.e as borderline for the probability and expectation calculations, mainly because the public and politicians can relate to it. We have, in a sense, tried to answer the question: “Is there another Johan Sverdup or better?” The calculations may of course be repeated for different sizes. The table below gives the probabilities using the borderline for the naming of “giant field” mentioned earlier, i.e. Gborder=79 MSm3o.e.

Time variable	Year		Index	
Discoveries ahead	5 years of 5	10 years of 5	5 years of 5	10 years of 5
P(at least one > Gborder)	0.45	0.72	0.24	0.35

Here we see a considerable discrepancy between the two modes of calculation. This may appear somewhat surprising, given that there was no apparent biasing element due to size ordering within years. However, as mentioned above, this is the way it may turn out: slight differences in the individual probabilities are magnified by many multiplications in “at least”-type calculations. Despite such discrepancies, the numbers give some ideas on the probability level to expect. For decision support it makes a difference to know whether the probability is likely to be less than 10% or more than 30%. At this point we should recall that the corresponding probability calculated from the exponential decaying creaming function $k(t)$ gave the probability 31%, a number we now feel is fair.

5. Comments and concluding remarks

Our probability estimates are based on models involving just sequence and size. This may seem very naïve. At a specific point in time there may of course be other information that might improve the predictions. A new area may be opened for exploration, which has been kept aside for years for other reasons than its promises to deliver oil or gas. Moreover, experts may have specific opinions on the immediate future. There are numerous things to take into account. They are hard to model, and experts may disagree.

Our use of the NCS-data deserves some comments. The 181 discoveries/fields reported in RNB 2014 are from all resource classes with positive reserves. In some cases this may be a mixture of several nearby discoveries which are developed together under a common name, one being Snøhvit. Using just public data we cannot separate out each of them. It is not clear how much complete knowledge will affect the probability estimates. In some cases a new discovery near a producing field may be added to that one, and not be registered separately. This may cause some distortion of the data. What we really try to predict is therefore future additions to the public reports, regardless of any exploration models used in the industry. This is relevant, since such reports typically are used as decision support for the politicians.

Experience from other fields indicates that there is not much to gain by adding complexity or adding explanatory variables, with additional parameters which have to be estimated. In many cases the data itself is carrier of the more hidden information, which is not available to us anyway. We believe that our probability estimates are fair, and cannot be considerably improved, unless very specific and agreed upon information is available. Anyway, the found estimates may be used as a benchmark, where adjustments up or down have to be justified by relevant arguments.

Appendix. Truncated Log-normal creaming

In some cases one consider the truncated Log-normal distribution, i.e. the conditional distribution of X given a<X ≤b with density

$$g_0(x) = \frac{1}{\Phi\left(\frac{\log(b)-\mu}{\sigma}\right) - \Phi(\frac{\log(a)-\mu}{\sigma})} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{x} \cdot e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2} \; ; \; a < \; x \leq b$$

where Φ denotes the cumulative standard normal distribution. As shorthand notation we take

$$C_k(a,b,\mu,\sigma) = \left[\Phi\left(\frac{\log(b)-(\mu+k\sigma^2)}{\sigma}\right) - \Phi\left(\frac{\log(a)-(\mu+k\sigma^2)}{\sigma}\right)\right]^{-1}$$

so that, in terms of the non-truncated Log-normal, we have $g_0(x) = C_0(a,b,\mu,\sigma) \cdot f_0(x)$.

If we sample proportional to x^k, the observations appear as coming from a distribution proportional to $x^k g_0(x)$, which means that it can be written as

$$g_k(x) = \; x^k g_0(x) \cdot \frac{C_k(a,b,\mu,\sigma)}{C_0(a,b,\mu,\sigma)} \; \cdot \; e^{-\left(k\mu+\frac{1}{2}k^2\sigma^2\right)} ; \quad a < x \leq b$$

The normalizing factor follows from the Log-normal conditional expectation formula

$$E_0(X^k|\text{a} < \text{X} \; \leq \text{b}) = \int_a^b x^k g_0(x)dx = \frac{C_0(a,b,\mu,\sigma)}{C_k(a,b,\mu,\sigma)} \cdot e^{k\mu+\frac{1}{2}k^2\sigma^2}, \text{ k real}$$

We see that $g_k(x) = C_k(a,b,\mu,\sigma) \cdot f_k(x)$, and since $C_k(a,b,\mu,\sigma) = C_0(a,b,\mu+k\sigma^2,\sigma)$, we get the same k-creaming property for the truncated Log-normal distribution as in the non-creamed case:

$g_0(x)$ (a,b)-Truncated Log-normal$(\mu,\sigma) \rightarrow g_k(x)$ (a,b)-Truncated Log-normal$(\mu + k \cdot \sigma^2, \sigma)$

i.e. we just replace μ by $\mu + k \cdot \sigma^2$.

The nth moment with respect to $g_k(x)$ may then be expressed as follows

$$E_k(X^n|\text{a} < \text{X} \; \leq \text{b}) = \int_a^b x^n g_k(x)dx = \frac{C_k(a,b,\mu,\sigma)}{C_{k+n}(a,b,\mu,\sigma)} \cdot e^{n(\mu+k\sigma^2)+\frac{1}{2}n^2\sigma^2}$$

$$= E_0(X^n|\text{a} < \text{X} \; \leq \text{b}) \cdot \frac{C_n(a,b,\mu,\sigma)}{C_0(a,b,\mu,\sigma)} \cdot \frac{C_k(a,b,\mu,\sigma)}{C_{k+n}(a,b,\mu,\sigma)} \cdot e^{nk\sigma^2}$$

References

Andreatta, G. and Kaufman G.W. (1986) Estimation of finite population properties when sampling is without replacement and proportional to magnitude*, Jour. Amer. Statist. Assoc*., vol. 81, no. 395, 657-666.

Barouch, E. and Kaufman G.W. (1967) Oil and gas discovery modelled as sampling proportional to random size, *Cambridge, Mass: MIT Alfred P. Sloan School of Management.* *http://dspace.mit.edu/handle/1721.1/48701*

Charpentier, D.R., Dolton, G.L. and Ulmishek G.F. (1995) Annotated bibliography of methodology for assessment of undiscovered oil and gas resources, *Nonrenewable Resources*, vol. 4, no. 2, 154-185.

Chen, Z. and Sinding-Larsen R. (1994) Estimating number and field size distribution in frontier sedimentary basins using a Pareto model*, Natural Resources Research vol.3, no.2*, 91-95.

Halbouty, M.T. (2001) Giant oil and gas fields of the 1990s. An introduction. Presentation at Symposium, Giant Oil and Gas Fields of the Decade 1990-2000, AAPG Convention, Denver, CO, June 5, 2001 (http://www.searchanddiscovery.com/documents/halbouty03/, downloaded Jan 12. 2016)

Kaufman, G.M., Balcer, Y. and Kruyt, D. (1975) A probabilistic model of oil and gas discoveries. In Methods of Estimating the Volume of Oil and Gas reserves (ed. J.D. Dunn). *American Assoc. Petroleum Geologists, Studies in Geology* no.1, 113-142.

Kaufman, G. M. (1992) Statistical Issues in the Assessment of Undiscovered Oil and Gas Resources, MIT-CEEPR-92-010WP. http://dspace.mit.edu/bitstream/handle/1721.1/50204/35719963.pdf?sequence=1

Lee,P. J. and Wang P.C.C. (1985) Prediction of Oil and Gas Pool sizes when discovery record is available, *Math Geol*. v.17,No.2,p 95-113.

Lillestøl, J. and Sinding-Larsen, R. (2015) Beta Creaming, Discussion paper FOR 8 2015, Dept. of Management Science, Norwegian School of Economics.

Meisner, J. and Demirmen, F. (1981) The creaming method: A Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces*, Journ. Royal Statist. Soc., ser. A*, vol. 144, no.1, 1-31.

RNB 2014 (2014) Resource account for the Norwegian Continental Shelf as of December 31, 2013. The Norwegian Petroleum Directorate. (Data file: Publiserte-tabeller-RNB2014.xlsx). http://www.npd.no/en/Topics/Resource-accounts-and--analysis/Temaartikler/Resource-accounts/2013/

Schuenemeyer, J. H. and Drew L.J. (1983) A Procedure to Estimate the Parent Population of the Size Oil and Gas Fields as Revealed by a Study of Economic Truncation, *Math. Geol*. v. 15, No.1, p.145-161.