

***Probabilistic Aspects  
of  
Exploration Risk***

# ***Nozer D. Singpurwalla***

The George Washington University,  
Washington, D. C. 20052, USA

# *Preamble*

- Exploration Risk, means an assessment of risks associated with decisions about the extraction of resources from the earth's crust.
- Here we limit attention to oil and gas.
- This topic poses *challenges* because :
- a) The underlying phenomenon is a blending of geological, economic, engineering, and behaviorist issues, and
- b) Each issue is endowed with its own paraphernalia of uncertainties.

# Preamble (Contd.)

- We are concerned here with the treatment of uncertainties using probabilistic and statistical techniques.
- My talk is *expository*; I do not claim expertise in this topic.
- I discuss it all the same, because the topic is important and there seems to be an opportunity for **invoking** modern stochastic methodologies here.

# *Background*

- Prospecting for oil and gas is a high risk business because it involves elements of large uncertainties and the odds of success are relatively small.
- Indeed oil exploration is referred to as the *greatest gamble on earth*.
- This is because exploration problems involve a mixture of *subjective* and objective information, and their blending based on feelings.

## *Background (Contd.)*

- A key variable that drives decision making in oil exploration is the *deposit size*, given that there is a deposit.
- But the actual deposit size is rarely, if ever, known so what matters most is the *estimated or reported deposit size*.

## *Background (Contd.)*

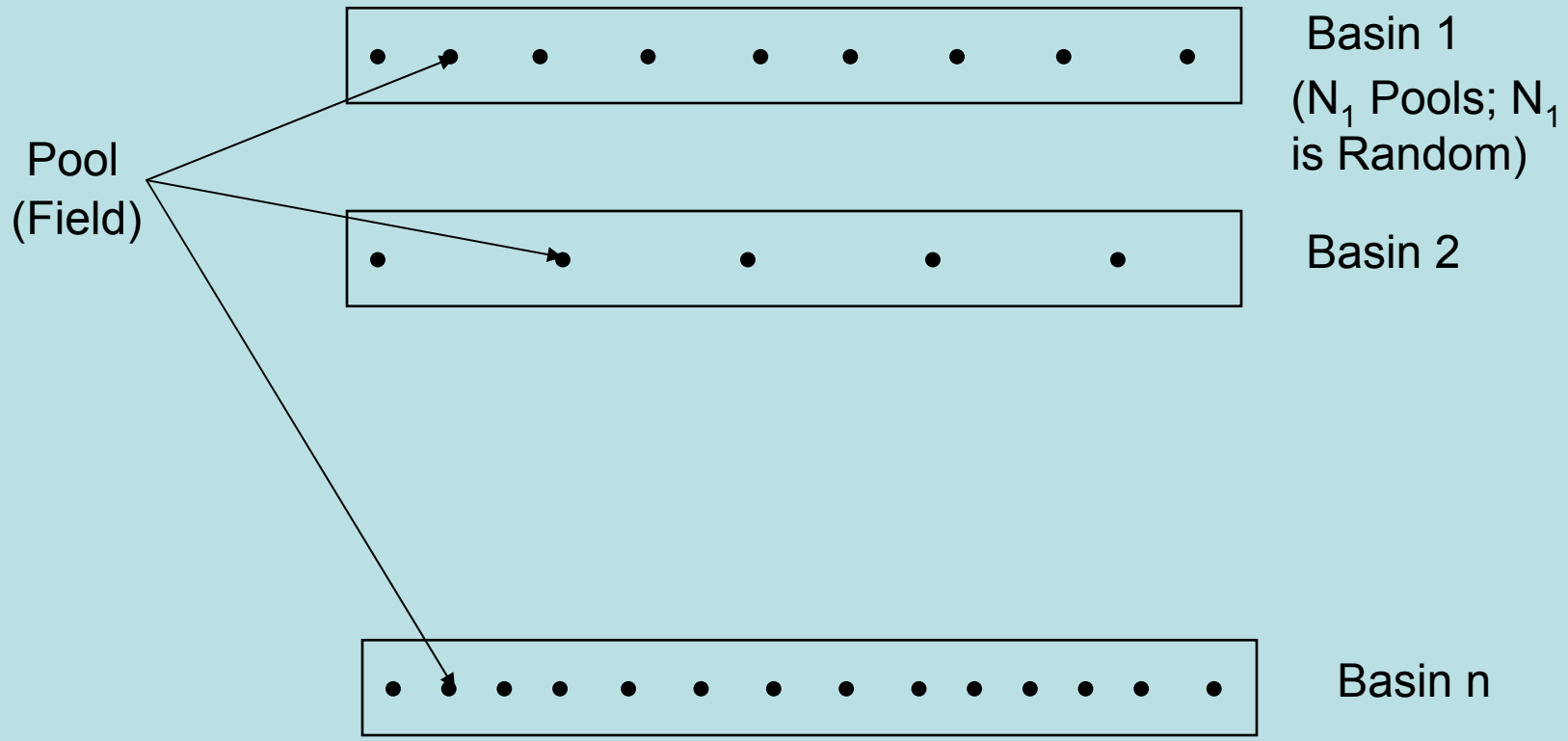
- However, the reported deposit size is *contaminated* due to many features:
  - a) Large “fields” tend to be discovered earlier in the discovery process than smaller fields leading us to the statistical issues of *size biased sampling, and weighted distributions*.
  - b) *Economic Filtering*, wherein the number of fields reported depends on the prevailing price.

## *Terminology: Fields and Basins*

- Geologically, it is assumed that at some prehistoric time, nature caused petroleum to be created in deposits of various sizes throughout the “basins” of the world.
- Petroleum is organic and can occur as a liquid, gas, or semi-solid.
- It occurs in minute pore spaces, called **traps**, between grains of sand in sandstone or shale.

## *Terminology (Contd.)*

- A *well* is a hole drilled in a trap; it causes petroleum to escape to the surface.
- A deposit is called a *pool or a field*.
- A collection of pools is called a *basin*, and
- A collection of basins, a *petroleum province*.
- For example, Qatar is a petroleum province.

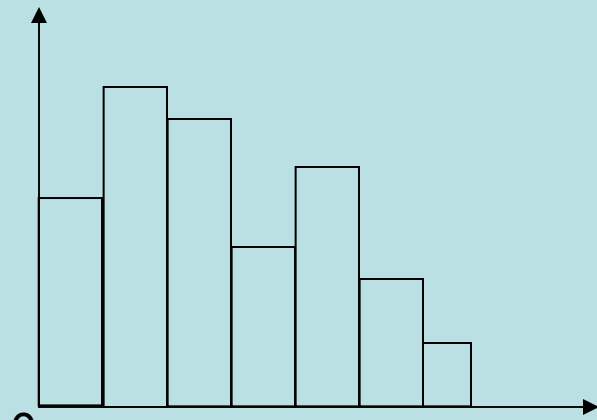


**Schemata of a Petroleum Province**

# *The Discovery Process*

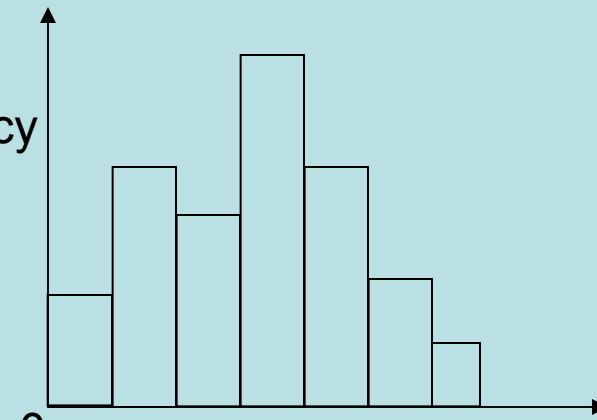
- The discovery process is the act of conducting reconnaissance and surveys, drilling wells and reporting an estimate of the deposit size in a field.
- The estimate is **dynamic**, because it changes with production experience.
- Thus a basin's estimated deposit size history is portrayed as a series of histograms over time, each histogram depicting the estimated deposit size in the basin at a particular point in time.

Frequency



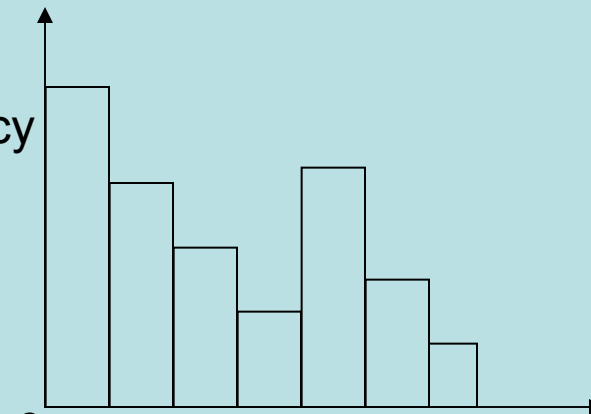
Reported Size (at time  $t = 1$ )

Frequency



Reported Size (at  $t = 2$ )

Frequency



Reported Size (at  $t = n$ )

An Evolving Histogram of Reported Deposit Size of a Petroleum Basin

# *Summary of Discovery Process*

- It results in the following:
- The estimated number of pools (or fields) in a basin  $i$ , say  $N_i$ .
- The estimated deposit size in each pool, say  $S_i$ .
- A histogram of the estimated deposit sizes over all the pools in the basin.
- This histogram changes over time because new pools get discovered and/or the estimated sizes per pool get revised.

# *Probabilistic Models for Deposit Size*

- The unknown deposit size in a field may be described probabilistically via two steps:
  - a) The probabilistic modeling of the physical phenomenon that generates the deposit.
  - b) Estimating the parameters of the model using all the available information, objective and subjective.

## *Probabilistic Modeling (Contd.)*

- However, with b) above we have a problem because we do not have at hand data on actual deposit size. We have data that is truncated (Type I censored), and/or *expert testimonies* about deposit sizes.
- Such testimonies could be based on geological prospecting techniques, or *judgments* of wildcatters.

# *The Extraction Process*

- This process follows the discovery process and is fundamentally an engineering exercise wherein oil gets extracted from the earth's crust.
- The extraction process commences only if the discovery process warrants it; the process takes a long period of time.
- The process terminates with the depletion of a field's reserve or abandonment of a field due to economic considerations.

# *Candidate Probability Distributions.*

- The *lognormal*, Yule, and Pareto have been discussed by Kauffman (1963) as deposit size distributions.
- Kondolf and Adhikari (2000) advocate the Weibull, whereas
- Seyedghasemipour and Bhattacharya (1990) make a case for the *loghyperbolic distribution*.

# *Genesis of the Lognormal*

- The basic ideas can be traced to Kolmogorov's work on the size distribution of coal and dust particles.
- Here one supposes that at some prehistoric time, nature caused a petroleum deposit to be created in a field.
- The size of the deposit changes with time due to geochemical actions. These can be seen as "*shocks*" which cause the deposit size to grow randomly.

# Genesis (Contd.)

- The growth is miniscule, and assumed to be governed by the *law of proportional effect*.
- Specifically, if  $\theta_t$  denotes the size of any particular field at time  $t$ , then

$$\Theta_t - \theta_{t-1} = \delta_t \theta_{t-1},$$

- where  $\{\delta_t\}$ ,  $t=1, 2, 3, \dots$ , is a sequence of i.i.d. random variables, denoting the percent increase in size from time  $(t-1)$  to time  $t$ .

## Genesis (Contd.)

- If we set  $\varepsilon_t = 1 + \delta_t$ , then

$$\theta_t = \theta_0 \prod_{i=1}^t \varepsilon_i$$

- where  $\theta_0$  is the size of the prehistoric deposit.
- Taking logarithms to base e, we have

$$\ln \theta_t = \ln \theta_0 + \sum_{i=1}^t \ln \varepsilon_i$$

## Genesis of Lognormal (Contd.)

- Suppose that  $E(\text{Ln } \varepsilon_i) = \mu < \infty$ , and that  $\text{Var}(\text{Ln } \varepsilon_i) = \sigma^2 < \infty$ ,

- Then, as  $t \uparrow \infty$ ,

$$\text{Ln } \theta_t \sim N(\text{Ln } \theta_0 + t\mu, t\sigma^2)$$

- Or that

$$\theta_t \sim \Lambda(\text{Ln } \theta_0 + t\mu, t\sigma^2).$$

- Here N is the normal distribution and  $\Lambda$  is the lognormal distribution

# *Inferential Issues*

- In order to make the above result concrete, so that drilling decisions can be made, we need to pin down  $\ln \theta_0$ ,  $\mu$ ,  $\sigma^2$ , and also  $t$ , the time since inception.
- The above of course are problematic, since one does not have direct observations on  $\theta_t$ , and thus standard statistical methods cannot be invoked.
- In view of the above we consider Bayesian techniques of elicitation and coding expert testimonies.

# *Eliciting Expert Testimonies*

- Suppose that a subject matter expert, say E, who could be a geologist, a wildcatter, or a simulation/engineering modeler declares at time T two quantities,  $m_T$  and  $s_T$ , as his/her reported deposit size  $\theta_t$  at time t, and a measure of uncertainty about  $\theta_t$ , respectively..
- We interpret  $m_T$  as expert testimony of E about  $\mu_t = \text{Ln } \theta_0 + \mu t$ , the mean of  $\text{Ln } \theta_t$ .

# *Expert Testimony (Contd.)*

- With the expert testimony inputs at hand, our task is to assess:
- $\Lambda(\mu_t; m_T, s_T)$ , our uncertainty about  $\mu_t$  in the light of E's expert testimony  $(m_T, s_T)$ .
- This can be done via Bayes' law as:
- $\Lambda(\mu_t; m_T, s_T) \propto \Lambda(m_T, s_T | \mu_t) \cdot \Lambda(\mu_t)$   
 $= \Lambda(m_T | s_T, \mu_t) \cdot \Lambda(s_T | \mu_t) \cdot \Lambda(\mu_t)$ ,  
where  $\Lambda(\mu_t)$  is a decision maker D's prior on  $\mu_t$ .

# *Modulating Expert Testimonies*

- To proceed further, D may make the following assumptions:
- $\Pi(\mu_t)$  is effectively constant on  $[0, \infty)$ ;
- $S_t$  by itself is not informative about  $\mu_t$ ; i.e.  $\Pi(s_T | \mu_t)$  is effectively constant on  $[0, \infty)$ ;
- For some specified constants  $\alpha$ ,  $\beta$ , and  $v$   
 $\Pi(m_T | s_T, \mu_t) \sim N(\alpha + \beta \mu_t, v s_T)$ .

## *Modulating Testimony (Contd.)*

- The above assumptions imply that when:
- $\beta = 1$ ,  $\alpha$  encapsulates the extent of E's **bias** in overestimating  $\mu_t$ , as seen by D.
- $\beta \neq 1 \rightarrow$  bias increases linearly in  $\mu_t$ .
- If E overestimates by 10%, then D sets  $\alpha$  as 0 and  $\beta = 1.1$ .
- With  $v > (<) 1$ , E tends to exaggerate (is overcautious) about in specifying the precision  $s_T$ .

## *Modulating Testimony (Contd.)*

- With  $\alpha$ ,  $\beta$ , and  $v$  so specified, we can see using standard arguments that:

$$\Pi (\mu_t; \alpha, \beta, v, m_T, s_T) \sim N \left( (m_T - \alpha)/\beta, (vs_t/\beta)^2 \right)$$

- Setting  $\alpha = 0$ , and  $\beta = v = 1$ , tantamount to D adopting E's testimony as is.

# *Testimony on Variance*

- Suppose that E were also to report, in addition to  $m_T$  and  $s_T$ , two quantities  $M_T$  and  $S_T$ , as E's measures of location and scale for the unknown  $\sigma_t^2 = t \sigma^2$ , of the assumed lognormal distribution for  $\theta_t$ .
- Then, by Bayes' Law,' D's uncertainty about  $\sigma_t^2$  can be shown to be prescribed via  $\Pi(\sigma_t^2; M_T, S_T)$ , as:

## *Testimony on Variance (Contd.)*

$$\left( \frac{c^2 r}{M_\tau S_\tau} \right)^{1/2} \sigma_t^2 \sim \chi \left( \frac{\gamma}{S_\tau} + 1 \right)$$

# *Interpretation of Result*

- The constants  $c$  and  $r$  reflect  $E$ 's biases and precision in specifying  $M_T$  and  $S_T$  as seen by  $D$ .
- $c = 1 \rightarrow$  suggests the absence of a bias,
- $c < (>) 1 \rightarrow$  that  $E$  underestimates (overestimates)  $\sigma_t^2$  when specifying  $M_T$ .
- When  $r < (>) 1/2$ ,  $\rightarrow$   $E$  exaggerates (is overcautious) in specifying  $S_T$ .
- Setting  $c = 1$  and  $r = 1/2$  tantamount to  $D$  accepting  $E$ 's testimony, as is.

## *Field Size Estimation Based on Fusion*

- With the above at hand, we may obtain an assessment at time  $T$  of the field size  $\theta_t$  with  $t$  being the time since inception, based on a fusion of probabilistic modeling and expert testimonies.
- Since  $E$ 's testimonies are indexed by  $T$ ,  $\theta_t$  should also be indexed by  $T$  as  $\theta_t(T)$ .
- Let  $Y_t(T) = \text{Ln}(\theta_t(T))$ . Then its pdf at  $y$  will be of the form:

# *Final Assessment of Deposit Size*

$$\int_{-\infty}^{\infty} \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sigma_t} \exp \left[ -\frac{1}{2} \left( \frac{y - \mu_t}{\sigma_t} \right)^2 \right] \Pi(\mu_t; \cdot) \Pi(\sigma_t^2; \cdot) d\mu_t d\sigma_t^2$$

# Monte Carlo Integral

- To evaluate the integral given before we may use the method of Monte Carlo Integration:
- For this, we generate  $n$  pair of observations say  $(\mu_t(i), \sigma_t^2(i))$ ,  $i = 1, 2, \dots, n$ , from  $\Pi(\mu_t; \cdot)$  and  $\Pi(\sigma_t^2; \cdot)$ , and plug these in the exponential term of the integral, to obtain

$$\exp \left[ - \frac{1}{2} \left( \frac{y - \mu_t(i)}{\sigma_t(i)} \right)^2 \right], \quad i = 1, \dots, n.$$

# *Predictive Density for Deposit Size*

- With the above so generated, we have as a Monte Carlo approximation of the predictive density of  $Y_t(T)$  at time  $T$ , as:

$$\frac{1}{n} \sum \frac{1}{\sqrt{2\pi} \sigma_t(i)} \exp \left[ -\frac{1}{2} \left( \frac{y - \mu_t(i)}{\sigma_t(i)} \right)^2 \right]$$

# General Comments

- Generating Monte Carlo samples is straightforward because we need concern ourselves only with the Gaussian and the Chi-Square distributions.
- Once the pdf of  $Y_t(T)$  is at hand, we may induce the pdf of  $\theta_t(T)$ , the deposit size at time  $\tau$  via the relationship:

$$\theta_t(\tau) = \exp(Y_t(T)).$$

## *Loghyperbolic Distribution for Deposit Size*

- Whereas the lognormal distribution for deposit size has its genesis in probabilistic modeling, the genesis of the loghyperbolic, is primarily empirical.
- This is because histograms of actual deposit sizes are positively skewed with fat right hand tails.
- One such functional form is the loghyperbolic distribution, henceforth LHC, motivated by the empirical work of Bagnold (1937), and formalized by Barndorff-Nielsen (1977).

## LHC (Contd.)

- Indeed, Seyedghasemipour and Bhattacharrya (1990) analyzed several sets of data on the *actual size* of oil deposits from the Denver Basin oil fields and concluded that the LHC provided a better fit than the lognormal.
- Furthermore, these authors support this advocacy by appealing to the following mixtures property of the normal distribution:
- If  $(x|\mu, \sigma) \sim N(\mu + \beta\sigma^2, \sigma^2)$ , and if  $\sigma^2 \sim \text{IG}$  then

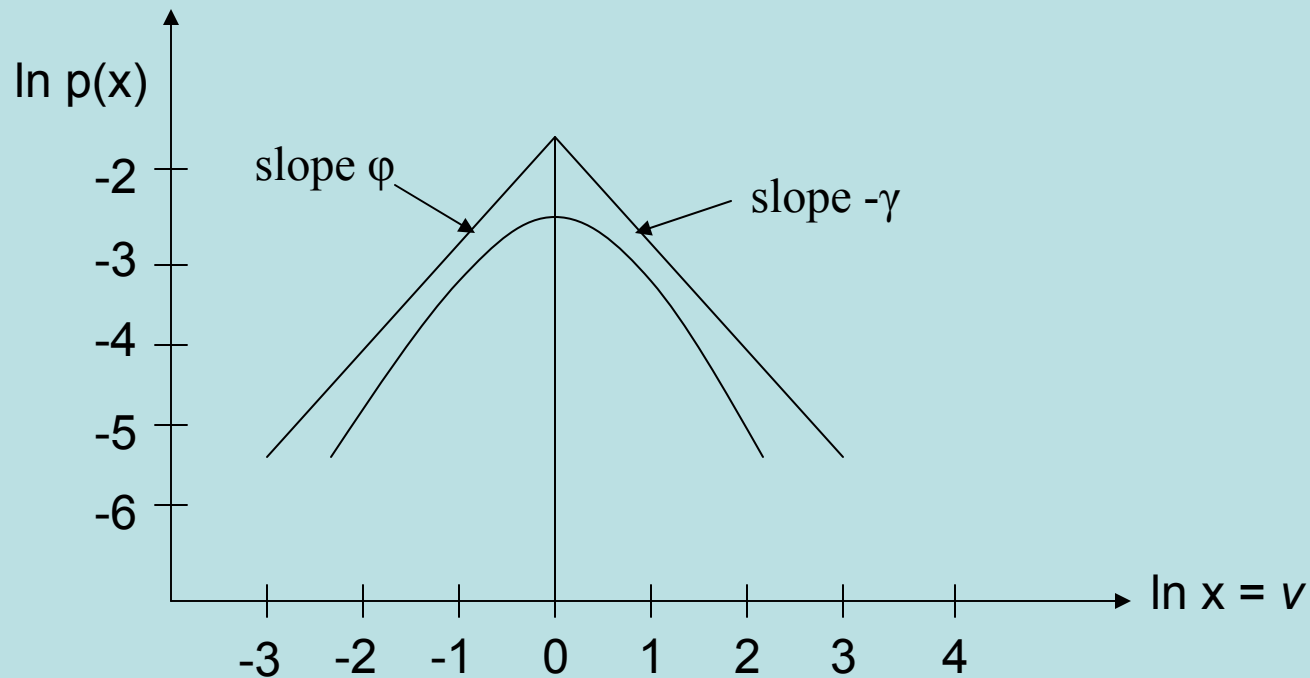
$$(x|\mu) \sim \text{LHC}.$$

## LHC (Contd.)

- Thus in a reasonably large oil basin if we conceptualize a lognormal distribution for the size of each field with the parameters of the lognormal varying from field to field then the deposit size for any randomly chosen field in the basin can be judged as being a LHC.
- Another way to describe the LHC is via the claim that if  $\text{Ln}(x)$  has a *hyperbolic distribution*, then  $x$  has a LHC distribution.

# *Anatomy of the LHC*

- Let  $x > 0$  be a R.V. of interest, and let  $V = \text{Ln}(x)$ .
- Suppose that  $v$  has pdf  $p(v)$ .
- Sp. That a plot of  $\text{Ln}(v)$  vs.  $\log[p(v)]$  reveals the shape of a linear increase followed by a linear decrease.
- Sp. that the asymptotes for  $\log [p(v)]$  have slopes  $\Phi$  and  $\gamma$ , with  $\Phi = \alpha + \beta$ , &  $\gamma = \alpha - \beta$ .



- The simplest mathematical function displaying the above form is a hyperbola for which the ordinate at  $v$  is

$$f(v) = -\alpha\sqrt{1+v^2} + \beta v$$

- The p.d.f. of  $v$  corresponding to  $f(v)$  is

$$c(\alpha, \beta) \exp\left[-\alpha\sqrt{1+v^2} + \beta v\right],$$

where  $c(\cdot)$  is a normalizing constant.

- Reparameterizing in terms of  $\phi$  and  $\gamma$  the p.d.f. becomes

$$c(\phi, \gamma) \exp\left[\frac{1}{2} \{\phi(\sqrt{1+v^2} - v) + \gamma(\sqrt{1+v^2} + v)\}\right]$$

where  $c(\phi, \gamma)$  is the norming constant.

$$\frac{1}{c(\phi, \gamma)} = kK_1(k) / w, \text{ where } k = \sqrt{\phi\gamma} \text{ and } w = (\phi^{-1} + \gamma^{-1})^{-1}$$

where  $K_1(\cdot)$  is the modified Bessel Function of the third kind with index 1.

- To introduce a location parameter  $\mu$  and a scale parameter  $\delta > 0$ , we may write  $\phi / \delta = \hat{\phi}$  and  $\gamma / \delta = \hat{\gamma}$  to obtain the p.d.f.

$$f(v; \hat{\phi}, \hat{\gamma}, \mu, \delta)$$

$$= \frac{w}{\delta k K_1(\delta k)} \exp \left[ -\frac{1}{2} (\hat{\phi} + \hat{\gamma}) \sqrt{\delta^2 + (v - \mu)^2} + \frac{1}{2} (\hat{\phi} - \hat{\gamma})(v - \mu) \right]$$

- $\hat{\phi}$  and  $\hat{\gamma}$  are the slopes of the two asymptotes of  $\ln f(v; \bullet)$  versus  $\log v$ .
- The asymptotes intersect at  $\mu$ , and the ordinate at  $\mu$  is

$$\ln \frac{w}{\delta k K_1(\delta k)}.$$

- The mode of  $f(v; \hat{\phi}, \hat{\gamma}, \mu, \delta)$  is  $m = \mu + \delta(\hat{\phi} - \hat{\gamma}) / 2k$  and the abscissa at  $m$  is

$$f(m; \cdot) = \frac{w}{\delta k K_1(\delta k)} \exp(-\delta k).$$

- The distribution with density  $f(v; \cdot)$  is called the *hyperbolic density*, and if a RV  $V$  has a hyperbolic distribution then  $X = \exp(V)$  is said to have a *loghyperbolic distribution*.
- Also note, that if  $X$  and  $Y$  are independent and gamma distributed, then  $(X|X.Y)$  has a loghyperbolic distribution. Similarly for  $(Y|X.Y)$ .

# Observed Size: Selection Bias

- The observed size distribution of deposit size in a *basin* has often been described by a lognormal, a Weibull, or a LHC.
- However, to justify such claims based on observations alone, the data obtained should be based on *random sampling*.
- This however, is not the case because large fields tend to be found early on in the discovery process.

## *Selection Bias (Contd.)*

- Also, economic considerations limit the number of small fields developed and reported; i.e. the reports are filtered. This process is called *economic truncation*.
- To summarize, sampling bias and economic truncation lead us to the point of view that one cannot impute the shape of the deposit size distribution based on observations alone.
- Indeed simulation studies have shown that biased sampling from a Weibull can be well described by a lognormal; see Power (1992).

# *Weighted Distributions and Encountered Data*

- Original work on this topic, including its introduction to the literature, is due to Fisher (1934).
- Suppose that the true deposit size  $X$  has pdf  $f(x|\theta)$ ; thus under random sampling, we will observe an  $x$  with probability  $\approx f(x|\theta)$ .
- If sampling is not random and if  $x$  enters our record with a probability proportional to  $\omega(x;\beta)$ , then the recorded  $x$  is a realization of a random variable  $X^\omega$ , whose pdf is of the form:

# Weighted Distributions

- $f^\omega(x \mid \theta; \beta) = \omega(x; \beta) \cdot f(x \mid \theta) / \omega$ , where  $\omega$  is a norming constant given as  $\Omega = E_f[\omega(X; \beta)]$ .
- $f^\omega(x \mid \theta; \beta)$  is called a **weighted distribution**.
- $\omega(x; \beta)$  need not be restricted to be within  $(0, 1]$ . Indeed, in oil exploration, since larger fields tend to get discovered early on in the discovery process,  $\omega(x; \beta) = x$ .
- When such is the case  $f^\omega(x \mid \theta; \beta) = x f(x \mid \theta) / \mu$ , where  $\mu$  is  $E(X)$ , and the sampling mechanism is **size biased**.

## *Encountered Data & Selection Models*

- A more general choice for  $\omega(x | \beta) = x^\beta$ , where  $\beta$  is unknown.
- Yet another choice would be  $\omega(x | \beta) = 1(0)$ , where  $X$  can be observed only when  $X > (<) \beta$ , with  $\beta$  unknown.
- This choice is germane in the context of economic truncation wherein small fields tend to get overlooked.
- Here  $\omega(x | \beta) = I[\beta, \infty)$ , and the pdf of an observed  $x$  is  $f(x | \theta) / P(x > \beta | \theta)$  for  $x > \beta$ , and the underlying model is called a *selection model*.

## Bayesian Inference in Selection Models

- Consider the selection model,

$$f^x(\omega | \theta, \beta) = \frac{\omega(x | \beta) \cdot f(x | \theta)}{\int \omega(x | \beta) \cdot f(x | \theta) dx}$$

where  $\theta$  and  $\beta$  are unknown, but the probability, under  $f(x|\theta)$ , of encountering an observation in an unknown *selection set*  $S$ , which is a subset of the sample space  $\Omega$ , is assumed known.

- Let this probability be  $\alpha$ . Then, if  $\Omega$  is the real line  $S=[\tau, \infty)$ , where  $\tau = F^{-1}((1-\alpha)|\theta)$ .

# *Bayesian Inference (Contd.)*

- Consequently, the selection model now becomes

$$f^\omega(x|\theta; \alpha) = f(x|\theta)/\alpha, \text{ for } x \geq F^{-1}(1-\alpha|\theta). \\ = 0, \text{ otherwise.}$$

- Consider the case of  $n$  encountered observations  $y_i, i= 1,2, \dots, n$ , with  $y^* = \min(y_i)$ , and  $f(x|\theta) = \theta \exp(-\theta x)$ .
- Let the prior on  $\theta$  be a gamma with scale  $a_0$  and shape  $b_0$ .

# *Bayesian Contd.) Inference*

- Then it can be shown that the posterior of  $\theta$  is of the form
- $C(a_1, b_1, \theta^*) \cdot \Gamma(\theta; a_1, b_1)$ ,  
where  $a_1 = a_0 + n$ , and  $b_1 = b_0 + r$ ,  
 $r = \sum y_i$ , and  $C$  is a (known) function of its arguments;  $\theta^* = -\text{Ln } \alpha/y^*$ .
- From the above we can obtain  $E(X)$ .