2. The probability densities on the number of spills and size of a spill

Following the above approach, within each category for a particular hypothetical offshore development we will be dealing with two variables:

1. The number of spills, n, of this category, which will occur in a given time period from this development.

2. The amount of oil which will be spilled, x, from an individual spill of this category emanating from this development.

One thing is immediately obvious. There is no way we can be sure of what values these two variables will take on. When one is faced with a variable which one cannot predict with certainty, such as n or x, one characterizes this variable by a probability density. A probability density is an assignment of likelihoods to each of the possible values of the variable. A sample assignment to the variable n is shown in Figure 2.1, which indicates that n can take on any of the values 0, 1, 2, 3 etc. with probability p(0), p(1), p(2), etc. The height of each arrow is proportional to the likelihood assigned to that value. When likelihoods are represented by probabilities, 0.00 represents the probability of an event which we are sure will not occur and 1.00 represents the probability of an event which is certain to occur. Since we are certain that n will take on at least one of its possible values, the probabilities p(0), p(1), p(2),... must sum to 1.00.

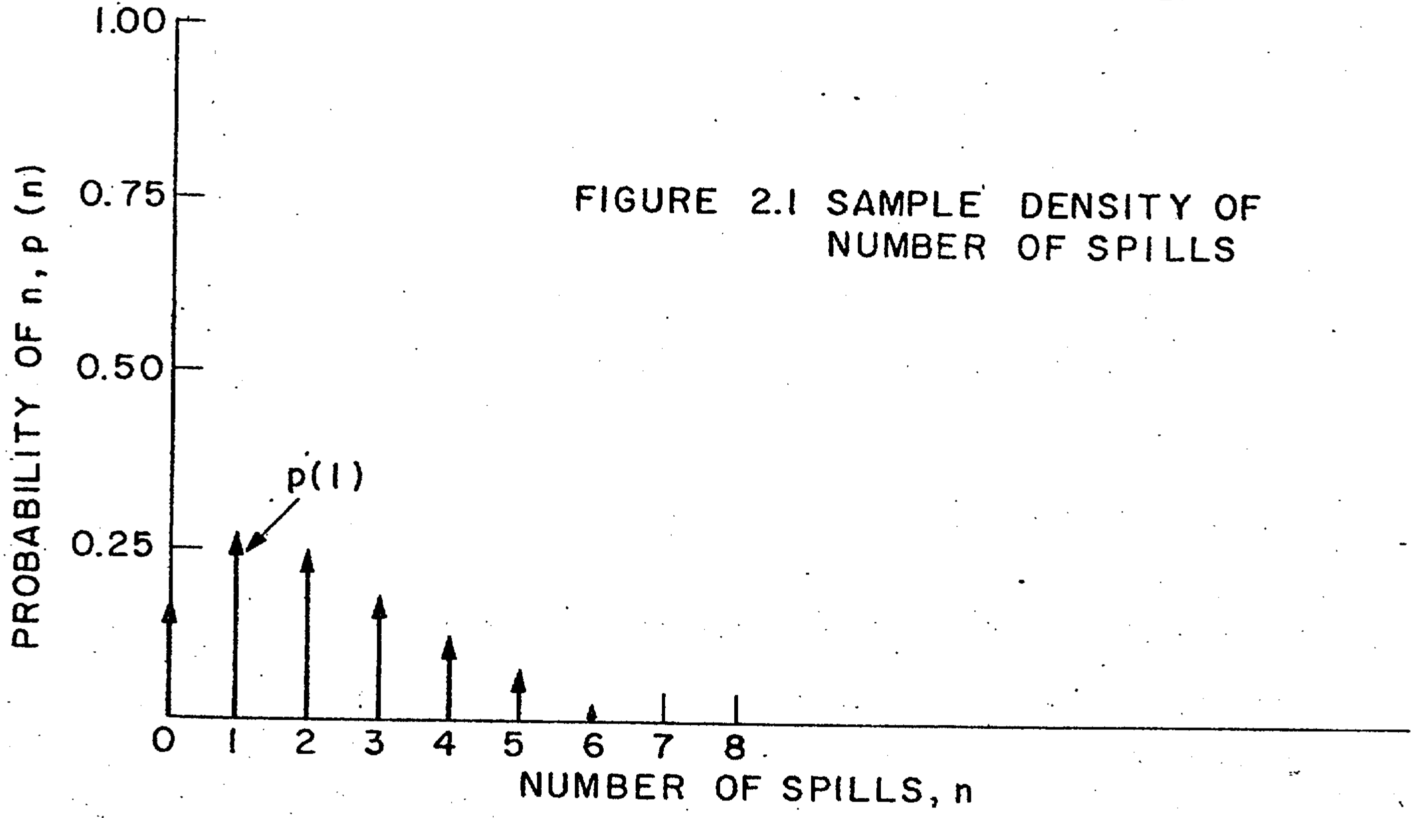

FIGURE 2.1 SAMPLE DENSITY OF NUMBER OF SPILLS

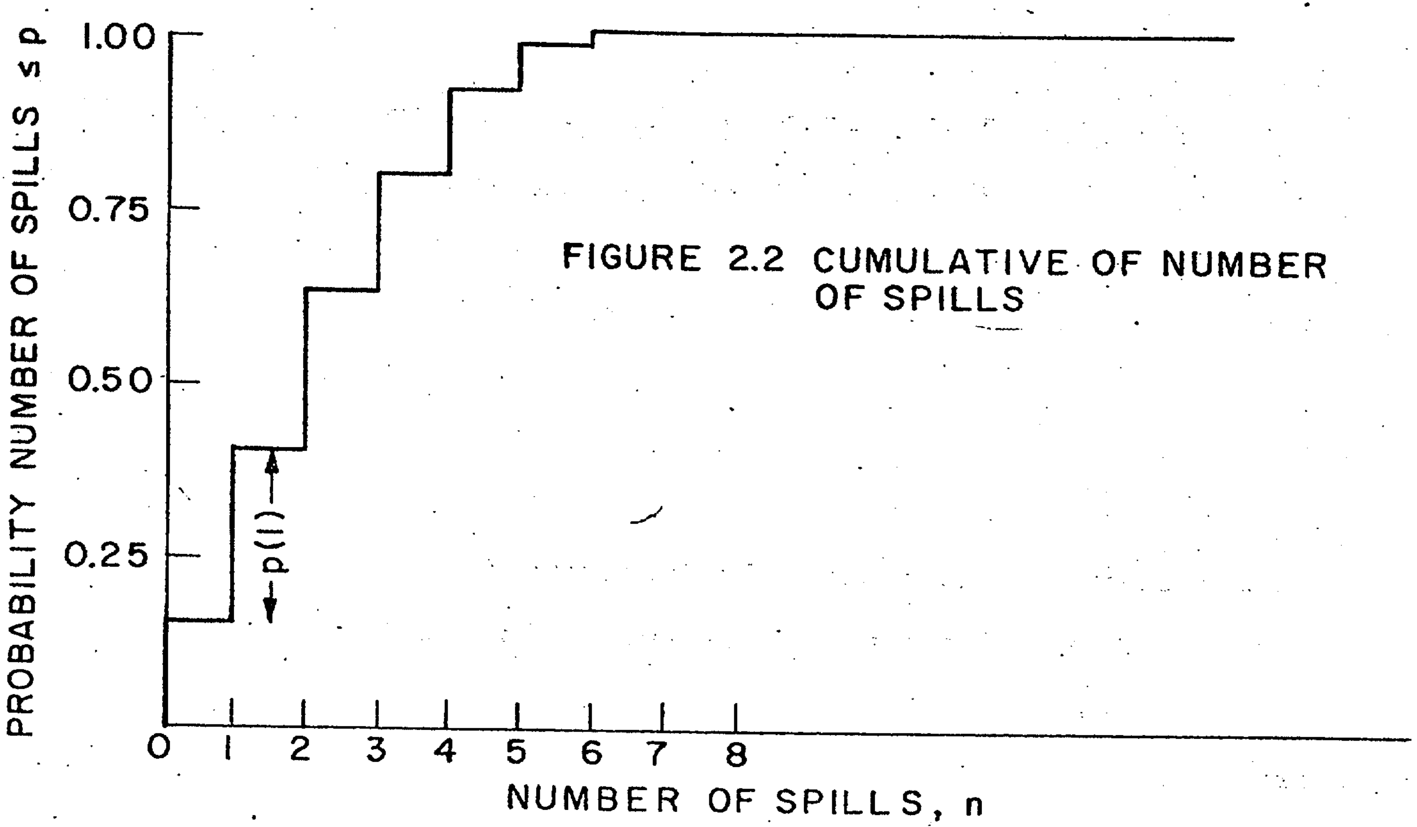

FIGURE 2.2 CUMULATIVE OF NUMBER OF SPILLS

If one multiplies each probability p(n) with the number of spills to which it has been assigned and sums these products, one obtains a measure of the central value of the density of the number of spills. This measure is called the mean of the density, MEAN(n). In symbols:

$$\text{MEAN}(n) = p(0)\cdot 0 + p(1)\cdot 1 + p(2)\cdot 2 + p(3)\cdot 3 + \ldots$$

The mean corresponds roughly to the average of all the possible values of n.

Another useful measure of a probability density is the variance. The variance is the sum of the squared difference between each possible number of spills and the mean where each difference is weighted by the probability of that value in the sum. In symbols:

$$\text{VAR}(n) = p(0)\cdot(0 - \text{MEAN}(n))^2 + p(1)\cdot(1 - \text{MEAN}(n))^2 + p(2)\cdot(2 - \text{MEAN}(n))^2 + \ldots$$

The variance is a measure of how spread out the density is. The larger the variance, the less likely the actual value of n will be close to the mean. A baseball team made up of 50% .200 hitters and 50% .400 hitters will have a much larger variance than a baseball team composed entirely of .300 hitters. Both teams would have the same mean.

Sometimes it is useful to represent the density of the number of spills in a slightly different form, the cumulative. The cumulative of the number of spills is simply a graph which indicates the likelihood that the number of spills will be less than n for all possible n. It is obtained by successively summing up the arrows as one moves to the right, increasing n as indicated by Figure 2.2, resulting in a staircase-like figure. The cumulative is convenient in that it is possible to read off the probability that the actual number of spills will be between any two specified values by simply subtracting the cumulative associated with the higher value from the cumulative associated with the lower. For example, in the cumulative indicated in Figure 2.2, the probability that n will be between 2 and 4 is .80 - .40 = .40. Often in drawing cumulatives we will simply fair a curve through the high points in the steps, a lazy practice which will cause no difficulty as long as we remember that the number of spills must be an integer.

Our other random variable, the amount of oil which will be spilled in an individual spill, x, like n is inherently uncertain. However, the description of our uncertainty about x is somewhat complicated by the fact that x can at least conceptually take on any value between 0 and some very large number. We are no longer limited to integers. In this case, it is meaningless to ask what the probability of a spill of exactly 42,032.39567... gallons is, for one can always make this probability zero by using enough decimal places in asking the question. However, it is not meaningless to ask what is the probability of a spill being larger than, say, 42,000.00... gallons and smaller than 42,100.00... gallons. Therefore, when we are dealing with continuous variables such as spill size, we assign a probability density such as the two shown in Figure 2.3. In these densities, the probability of a spill

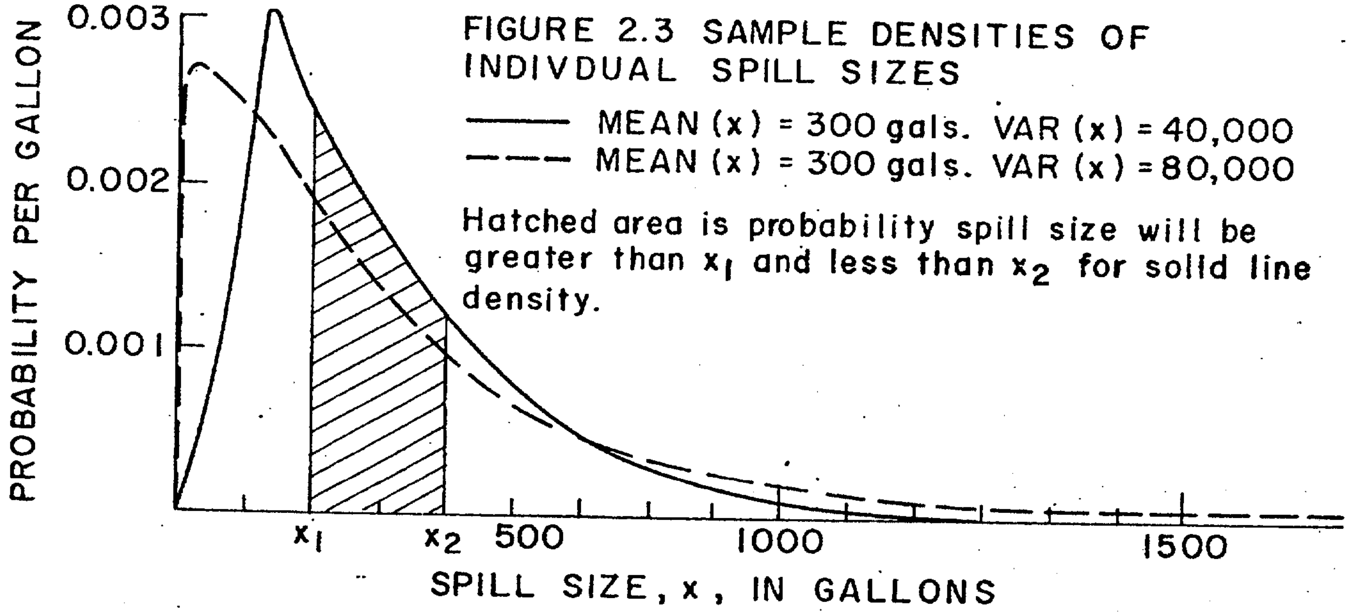

FIGURE 2.3 SAMPLE DENSITIES OF INDIVDUAL SPILL SIZES

Hatched area is probability spill size will be greater than x_1 and less than x_2 for solid line density.

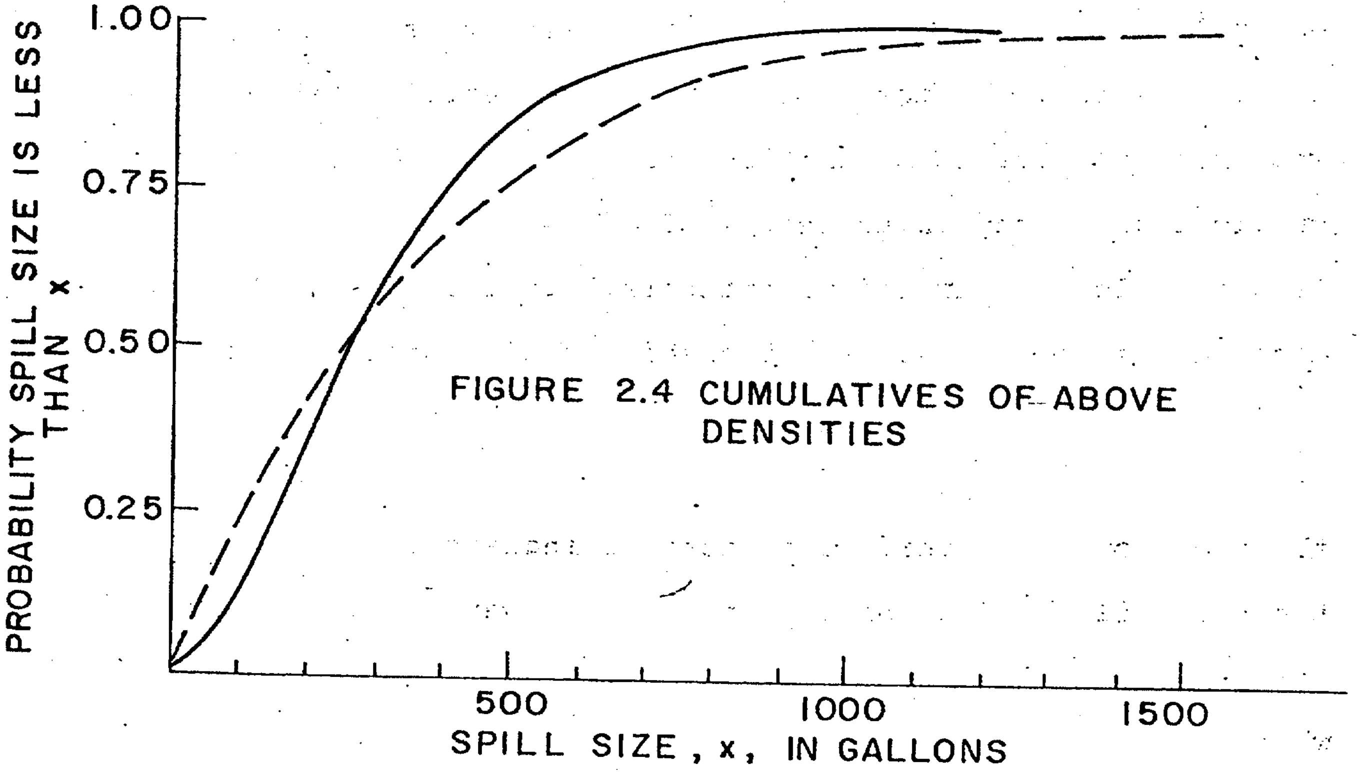

FIGURE 2.4 CUMULATIVES OF ABOVE DENSITIES

being larger than x_1 and smaller than x_2 is represented by the area under the density between x_1 and x_2. Thus, in intervals where the curve is high, there is more chance of the corresponding spill sizes than where it's low. If it is quite likely that the spill will be within a narrow range of sizes, then one obtains a sharply peaked, narrow density such as the solid density. If one is quite unsure of how large a spill will be, one will obtain a low, broad distribution, such as the dotted curve.

By summing the area above each small spill size interval multiplied by the corresponding spill size over all possible spill sizes, one obtains the mean of the spill size density, which once again is a measure of the average spill size. By summing the mean area about each interval multiplied by the square difference between the corresponding spill size and the mean, over all possible spill sizes, one obtains the variance of the density, which is a measure of the dispersion of the density. Both the densities shown in Figure 2.3 have the same mean, but the dotted density has a larger variance, implying that for this density the probability that a spill will be close to the mean in volume is much lower.

Our assignment of likelihoods to x can be represented by the cumulative of the density of x as well as by the density itself. Like the case of n, the cumulative is simply a graph indicating the probability that the actual spill size x will be less than x for all possible x. The cumulative of x is obtained by simply summing up the areas under the density as one moves to the right increasing x. Figure 2.4 shows the cumulatives for the densities shown in Figure 2.3. Notice how a tight density leads to a cumulative which rises sharply over a relatively narrow range while a widely dispersed density leads to a cumulative which rises more gradually over a much wider range. As in the case of n, one can obtain the probability that the spill size will be between any two given spill sizes by subtracting the value of the cumulative at the lower spill size from the value of the cumulative at the higher size.

Given that we are going to characterize inherently uncertain variables such as number of future spills of a particular category emanating from a particular hypothetic development and the amount spilled in such a future spill by probability densities, the key question becomes: how are we to assign these probabilities? At least conceptually, there are any number of ways one might go about assigning these likelihoods. We believe it is insightful to assign these probabilities in a manner which is consistent with the following ground rules:

1. The assignment will depend only on the available statistics. That is, we will not let our judgments about future improvements, changes in tanker size, and any non-quantitative experience we may have had relevant to spillage affect our assignment of likelihoods.

2. For each spill category, there is an underlying process generating spill occurrences and another generating spill size. These underlying processes have been constant over the period over which we have spill data and the same processes will govern future spillage.

3. These processes generate spills independently, that is, the fact that a spill occurs does not change the chances of the next spill occurring.

4. With respect to spill incidence, we will assume that the probability of a spill's occurring in a particular short exposure interval is proportional to the amount of exposure in this interval. Together with (3), this implies that spill occurrence is governed by a Poisson process.

5. With respect to spill size, we will assume that this variable is governed by a Gamma process. The Gamma process is a rather general set of processes which has some attractive analytical properties.

6. Consistent with (1), we will assume that before looking at the spill data, we have no idea which Poisson process is generating spill occurrence or which Gamma process is generating spill sizes. We are, in effect, tabulas rasas. We therefore

assign densities to the unknown parameters governing these processes, beginning with completely blank densities in which any of the possible values of these parameters is, for all practical purposes, equally likely.

7. As samples of spill occurrence and spill size become known, *we change our feelings about these unknown parameters according to the laws of probability theory*.

Now this is a rather long list of assumptions, and all of them, except, perhaps, the last, are open to question. For example, (3) can be challenged on the grounds that when a large spill occurs, there is generally an intensification of vigilance and care which will decrease the probability of a spill's occurring in the future from what it would have been. And it is certainly questionable whether the processes generating spills in the recent past are the same as those which will be generating spills in the future. It is even doubtful that the processes generating spills in the recent past were completely unchanged over the period during which the data was collected.

Nonetheless, let's accept these assumptions for the moment as working hypotheses and see where they lead us. We believe the results will be of great use even if they are regarded only as baselines from which modifications should begin. The list of assumptions underlying classical statistical analysis is at least as long and for at least certain of our spill categories involves such presumptions as: the next "large" spill has a significant probability of being negative in size. The above set of assumptions will at least avoid building on such imaginative foundations.*

*Classical statistical analysis also involves making assumptions 1, 2, 3 and 7. Assumptions 4 and 5 are usually replaced by assuming that the random variables in question are governed by Normal processes. Most classical statisticians would be unwilling to assign probability densities to the parameters governing the unknown random processes, assumption 6, even densities which give no weight to whatever feelings we had about these processes before looking at the data. Strictly speaking, this unwillingness prohibits one from making probabilistic statements about the variables under analysis. In practice, such statements are often made anyway. When they are made anyway, from a logical point of view, the analyst is acting as if he accepts assumption 6.

3. Quantitative implementation of the assumptions

3.1 Spill incidence

As indicated in Section 2, our procedure is to assume that spill incidence for a particular category is governed by a Poisson process. Under this assumption, if we know the intensity of the Poisson process, λ, the density of the number of spills is given by

$$p(n|\lambda,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

where t is the amount of exposure contemplated in the hypothetical development currently under analysis and λ is the mean spill rate in spills per unit exposure.

This assumption leads to two problems: what should we use for t, the exposure variable, and what should we use for λ, the mean spill rate? With respect to t, we will assume that the exposure variable in the Poisson process governing spill incidence is volume of oil handled. Some empirical support for this presumption is offered in the next section for tanker spillage. Similar support in the other categories has not yet been developed - at present it is simply a working hypothesis, albeit an obvious and natural starting point for spill analysis. It is also a hypothesis which underlies, usually tacitly, almost all spillage analysis which has taken place to date. Nonetheless, other hypotheses, such as "the exposure variable is number of landings" or "number of platforms" or "number of wells" or "number of pipeline miles" certainly deserve attention and should be examined.

When we turn to the choice of λ, the mean spill rate, things become still more complicated. Under our basic ground rules, the only information we allow ourselves on λ is the spill data. This implies two things:

1. Even after observing, say, ν spills in τ volume handled, we cannot be certain about the value of λ. Such data does not necessarily imply that $\lambda = \nu/\tau$ for other λ could easily have resulted in the same experimental outcome. Of course, the more data we have, the larger ν and τ, the more likely it is that λ is "close" to ν/τ. In short, λ is an uncertain quantity and, therefore, we must describe our knowledge about this quantity by a probability density.

2. Before having observed any spill data under our ground rules we have essentially no feelings about λ other than that it's somewhere between 0 and ∞. This implies that however we describe our feelings about λ before observing any-data, these prior feelings must be completely overwhelmed by whatever data we then observe.

We can meet requirements 1 and 2 and at the same time save ourselves some computational travail by assuming that our density on λ before observing any data is a Gamma in which the parameters are both zero.

Assumption 8 and some elementary probability analysis then reveals that, after having observed ν spills in τ volume handled, our density on λ is:

$$f(\lambda|\nu,\tau) = e^{-\lambda\tau}(\lambda\tau)^{\nu-1}\tau/(\nu - 1)!$$

The density on λ thus is the inlet through which our past spill experience enters the analysis.

Once one has the density on λ given the spillage we have observed, it is a simple matter to obtain the density on the number of future spills which will occur in a particular period given that we are going to handle t units of oil in that period. For each n, it is the probability that we will have n spills given each possible λ times that λ summed over all possible λ:

$$p(n|t,\nu,\tau) = \int_0^\infty p(n|\lambda,t)f(\lambda|\nu,\tau)d\lambda$$

After some algebra, the resulting density on n spills in a contemplated exposure of t units of oil handled given that we have already observed ν spills in our past exposure of τ units can be shown to be

$$p(n|t,\nu,\tau) = \frac{(n + \nu - 1)!t^n\tau^\nu}{n!(\nu - 1)!(t + \tau)^{n+\nu}}$$

which is known as the negative binomial density.

3.2 Spill size

We have adopted the same basic philosophy in obtaining densities of the size of an individual spill of given category. First we must hypothesize a random process which governs the size of a spill given that it occurs. A priori, we know only that a spill will not be negative in size. Thus, such commonly used processes as the Normal are out. We have chosen to assume that spill sizes are samples of a Gamma density.

$$f(x|\rho,\omega) = \frac{e^{-\omega x}(\omega x)^{\rho-1}\omega}{(\rho - 1)!}$$

The Gamma family of densities has two parameters, ρ and ω, and by varying these two parameters a complete range of means and variances can be obtained. In fact, for a given ρ and ω,

$\bar{x} - 104$

$$\text{MEAN}(x|\rho,\omega) = \rho/\omega$$

Platforms: $\omega = .000144$ $\rho = .0149$

$$\text{VAR}(x|\rho,\omega) = \rho/\omega^2$$

Thus, by making ρ small, a high ratio of variance to square of the mean can be obtained - a widely spread out density. By making ρ large, a relatively small ratio of variance to mean squares--a tight density--can be obtained. All the Gamma densities have only one peak and apply only to $x \geq 0$. In fact, by varying ρ and ω it is possible to obtain a reasonable approximation of any single-peaked density over the interval 0 to ∞. Thus, if one believes that the density of spill sizes is single-peaked, one loses very little generality by assuming that this density is a Gamma.*

*There is no a priori reason for believing that the spill size density is single-peaked. Spills occasioned by different causes almost certainly have different most likely sizes. In

Having assumed that the density of spill size is a Gamma, the next question is what are the values of the parameters ρ and ω. The obvious answer is we don't know so we must specify a density over these two random variables. In so doing, we desire a density which

a. fits well with the Gamma in an analytical sense in order to keep our computational travail within reason;

b. depends only on the sample of spill sizes of the category in question.

Stewart [6] has shown that having observed on spills of volumes $(x_1, x_2, x_3, \ldots x_m)$ respectively a density which fits these requirements is the so-called Gamma-hyperpoisson:

$$f(\omega,\rho|m,s,p) = e^{-s\omega}\omega^{m\rho-1}/(\Gamma(\rho)^m S(m,s,p)) \quad \times \rho^{\rho-1}$$

where $S(m,s,p)$ is a normalizing constant and

m = number of spills observed

$s = \Sigma x_i$ = total amount spilled

$p = \Pi x_i$ = product of all the individual spill sizes.

One can then obtain the density on x by multiplying the density on x given ω and ρ times the density on ω and ρ and then running over all possible values of ω and ρ. The result is

$$f(x|m,s,p) = \int_0^\infty \frac{(xp)^{\rho-1}\Gamma[(m+1)\cdot\rho]\,d\rho}{\Gamma(\rho)^{m+1}S(m,s,p)(x+s)^{(m+1)\cdot\rho}}$$

our actual analysis, we take a first step toward multiple peaks by dividing all spills into spills less than 42,000 gallons and spills greater than 42,000 gallons.

This is the density whose cumulative is shown in the spill size probability figures. Its mean is

$$\mathrm{MEAN}(x) = \frac{s/m}{S(m,s,p)} \int_0^\infty \frac{p^{\rho-1}\Gamma(m\rho)}{\Gamma(\rho)^m s^{m\rho}} \left\{\frac{\rho}{\rho - (1/m)}\right\} d\rho$$

which for large m tends quickly to the sample mean, s/m. For small m, the mean is higher than the sample mean.