

Piece-wise Defined Distributions: Discrete Mixtures

Finite (Discrete) Mixtures

Following [Bean], we define a **Finite (Discrete) Mixture** as a random variable whose (cumulative probability) distribution function is a finite weighted average of finitely many distribution functions of other random variables:

$$CDF_{\mathbf{X}}(x) = w_1 CDF_{\mathbf{X}_1}(x) + w_2 CDF_{\mathbf{X}_2}(x) + \cdots + w_k CDF_{\mathbf{X}_k}(x)$$

That is,

$$\Pr(\mathbf{X} \leq x) = w_1 \Pr(\mathbf{X}_1 \leq x) + w_2 \Pr(\mathbf{X}_2 \leq x) + \cdots + w_k \Pr(\mathbf{X}_k \leq x)$$

where, as always, the weights, w_1, w_2, \dots, w_k are positive numbers that sum to 1.

One example of a finite mixture is a **mixed random variable** (previously defined) whose (cumulative probability) distribution function is a weighted average of two *cdf*'s, one discrete (providing jump discontinuities) and one continuous (providing the smooth, increasing segments).

While you are unlikely to have studied mixed random variables in an introductory statistics course, they are extremely important in even simple insurance applications. In this short paper we will look at a number of insurance examples that lead naturally to mixed random variables. The source of most of this material is [Bean]. Bean gives a very clear account of the subject.

Deductibles (A Threshold or Floor for Claims)

It is common for health and accident insurance to have a threshold amount, called a **deductible**, which the insured must pay out-of-pocket before insurance starts to pay.

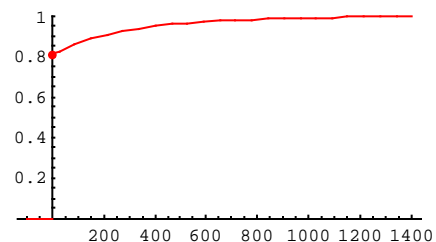
Example 1: A home-owners insurance policy with a \$500 deductible will pay the amount of damages for any adverse event less \$500 if the total loss exceeds \$500. If the loss is less than \$500 the policy pays nothing. Suppose the total loss for a typical adverse event has an exponential distribution with mean \$300.

In analyzing this situation, it is important to distinguish between two related random variables: \mathbf{X} , the amount of loss and \mathbf{Y} , the amount of the claim paid for this loss. We see easily that

$$\mathbf{Y} = \begin{cases} 0 & \text{if } x \leq 500 \\ \mathbf{X} - 500 & \text{if } x > 500 \end{cases} = \max\{0, \mathbf{X} - 500\}$$

We can derive the distribution function of \mathbf{Y} from the known distribution of \mathbf{X} using the **distribution function technique**:

$$\begin{aligned} CDF_{\mathbf{Y}}(y) &= \Pr(\mathbf{Y} \leq y) = \Pr(\max\{0, \mathbf{X} - 500\} \leq y) \\ &= \begin{cases} 0 & y < 0 \\ 1 - e^{-(500+y)/300} & y \geq 0 \end{cases} \end{aligned}$$



This is a mixed distribution with weights $w_1 = Pr(\mathbf{X} \leq 500) = 1 - e^{-500/300} = 1 - e^{-5/3}$ and $w_2 = Pr(\mathbf{X} > 500) = 1 - w_1 = e^{-500/300} = e^{-5/3}$.

- The discrete component, \mathbf{Y}_1 , takes the value 0 with probability 1 (providing the jump); its *cdf* is $CDF_{\mathbf{Y}_1}(y) = \begin{cases} 0 & y < 0 \\ 1 & y \geq 0 \end{cases}$. \mathbf{Y}_1 distinguishes between losses that will result in a claim being paid and those that do not.
- The continuous component, \mathbf{Y}_2 , is equal to $\mathbf{X} - 500$ for values y greater than 0 (corresponding to values x greater than 500); the portion of its *cdf* on positive values of y is the conditional *cdf* of $\mathbf{X} - 500$ given $\mathbf{X} > 500$.

$$\begin{aligned} CDF_{\mathbf{Y}_2}(y) &= \begin{cases} 0 & y < 0 \\ Pr(\mathbf{X} - 500 \leq y | \mathbf{X} > 500) & y \geq 0 \end{cases} \\ &= \begin{cases} 0 & y < 0 \\ Pr(\mathbf{X} \leq 500 + y | \mathbf{X} > 500) & y \geq 0 \end{cases} \\ &= \begin{cases} 0 & y < 0 \\ \frac{e^{-500/300} - e^{-(500+y)/300}}{e^{-500/300}} & y \geq 0 \end{cases} \end{aligned}$$

The weighted average of these two *cdf*'s is the *cdf* of the claim amount:

$$\begin{aligned} CDF_{\mathbf{Y}}(y) &= w_1 CDF_{\mathbf{Y}_1}(y) + w_2 CDF_{\mathbf{Y}_2}(y) \\ &= (1 - e^{-500/300}) CDF_{\mathbf{Y}_1}(y) + e^{-500/300} CDF_{\mathbf{Y}_2}(y) \\ &= \begin{cases} (1 - e^{-500/300}) \cdot 0 + e^{-500/300} \cdot 0 & y < 0 \\ (1 - e^{-500/300}) \cdot 1 + e^{-500/300} \cdot \frac{e^{-500/300} - e^{-(500+y)/300}}{e^{-500/300}} & y \geq 0 \end{cases} \\ &= \begin{cases} 0 & y < 0 \\ (1 - e^{-500/300}) + (e^{-500/300} - e^{-(500+y)/300}) & y \geq 0 \end{cases} \\ &= \begin{cases} 0 & y < 0 \\ 1 - e^{-(500+y)/300} & y \geq 0 \end{cases} \end{aligned}$$

This was our original description of the *cdf* of the claim amount \mathbf{Y} .

The next example illustrates methods for calculating probabilities and expected values for insurance policies with deductibles.

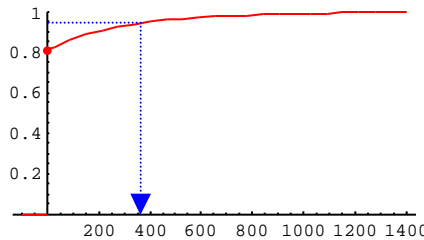
Example 2: Let X be the loss incurred and Y , the amount of the claim paid for this home-owner's insurance policy with a \$500 deductible where the loss per year, X , has an exponential distribution with mean 300. Determine

- the probability that the claim exceeds \$1000
- the 95th percentile of the claim amounts.
- the expected claim amount.

Since we have a simple expression for the *cdf* of Y from the previous example, it would be most efficient to use the *cdf* to calculate probabilities:

a) $\Pr(Y > 1000) = 1 - CDF_Y(1000) = 1 - \left(1 - e^{-(500+1000)/500}\right) = e^{-1500/500} = e^{-3} \approx 0.04979$

- b) The 95th percentile of the claim amounts is the solution, y , of the equation $CDF_Y(y) = 0.95$; graphically, it is the value on the horizontal axis, in this case the y -axis representing claim amount, that is mapped to 0.95.



In finding percentiles for a non-continuous distribution, we must determine if given percentage (in this case 0.95) on the vertical axis is a value attained by the *cdf*, or if it falls within the span of one of the vertical jumps. In this case, the graph jumps from height 0 up to height $e^{-3/5} \approx 0.811$, and is continuous above this level. Consequently, the 95th percentile is the solution of the equation $0.95 = CDF_Y(y)$ (shown above). Solving:

$$0.95 = CDF_Y(y) = 1 - e^{-(500+y)/500} \quad e^{-(500+y)/500} = 0.05$$

$$y = -500 \ln(0.05) - 500 \approx 997.87$$

The 95th percentile is \$997.87. Note that this is consistent with the answer in part a: $\Pr(Y > 1000)$ is slightly less than 0.05, $\Pr(Y \leq 1000)$ is slightly greater than 0.95, so we should expect the 95th percentile to be slightly less than \$1000.

If 0.95 had fallen in the vertical gap created by the jump, the 95th percentile would be 0.

We need the *pmf* (*probability mass function*) and *pdf* (*probability density function*) of the component variables, Y_1 and Y_2 , to calculate the expected value of Y . these functions can be derived from the *cdf*'s found in Example 1.

- The *pmf* of Y_1 is $p(y) = \begin{cases} 1 & y = 0 \\ 0 & \text{elsewhere} \end{cases}$

- The pmf of \mathbf{Y}_2 is $f(y) = \begin{cases} \frac{e^{-(500+y)/300}}{e^{-500/300}} & y > 0 \\ 0 & \text{elsewhere} \end{cases} = \begin{cases} \frac{1}{300} e^{-y/300} & y > 0 \\ 0 & \text{elsewhere} \end{cases}$

Consequently,

$$E(Y) = w_1 \int_0^y yp(y) + w_2 \int_0^y yf(y)dy = w_1(0 \ 1) + w_2 \int_0^y \frac{1}{300} e^{-y/300} dy = w_1(0 \ 1) + w_2 \cdot 300 = w_2 \cdot 300 = 300e^{-500/300} \approx \$56.66$$

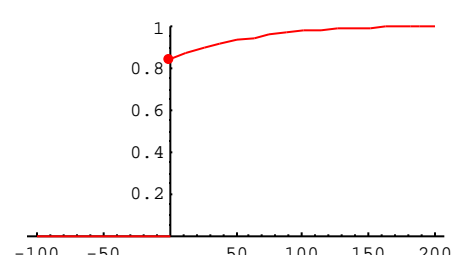
Notice that we are taking a weighted average of the expected values of the component variables. The justification is not completely obvious since the random variable \mathbf{Y} is not itself a weighted average of the components \mathbf{Y}_1 and \mathbf{Y}_2 ; a complete justification will be given later.

Example 3 Answer the questions from Examples 1 and 2 for a home-owners policy whose total loss for any adverse event is normally distributed with mean \$400 and standard deviation \$100.

Let $\tilde{\mathbf{X}}$ be the amount of the loss and $\tilde{\mathbf{Y}}$ the amount of the claim paid for this accident. Then, as in Example 1,

$$\tilde{\mathbf{Y}} = \begin{cases} 0 & \text{if } \tilde{x} \leq 500 \\ \tilde{\mathbf{X}} - 500 & \text{if } \tilde{x} > 500 \end{cases} = \max\{0, \tilde{\mathbf{X}} - 500\}$$

The distribution function of $\tilde{\mathbf{Y}}$ is

$$\begin{aligned} CDF_{\tilde{\mathbf{Y}}}(\tilde{y}) &= \Pr(\tilde{\mathbf{Y}} \leq \tilde{y}) = \Pr(\max\{0, \tilde{\mathbf{X}} - 500\} \leq \tilde{y}) \\ &= \begin{cases} 0 & \tilde{y} < 0 \\ \Pr(\tilde{\mathbf{X}} \leq 500 + \tilde{y}) & \tilde{y} \geq 0 \end{cases} \\ &= \begin{cases} 0 & \tilde{y} < 0 \\ \frac{1}{100\sqrt{2\pi}} \int_{-\infty}^{500+\tilde{y}} e^{-(x-400)^2/20000} dx & \tilde{y} \geq 0 \end{cases} \end{aligned}$$


$$= \begin{cases} 0 & \tilde{y} < 0 \\ \frac{500 + \tilde{y} - 400}{100} & \tilde{y} \geq 0 \end{cases} = \begin{cases} 0 & \tilde{y} < 0 \\ \frac{100 + \tilde{y}}{100} & \tilde{y} \geq 0 \end{cases} = \begin{cases} 0 & \tilde{y} < 0 \\ 1 + \frac{\tilde{y}}{100} & \tilde{y} \geq 0 \end{cases}$$

where $\tilde{\Phi}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$ is the cdf of a standard normal random variable.

Note that since the total loss cannot be negative, it cannot truly be normally distributed; as with many purportedly normal random variable, we should interpret the claim of normality as an acceptably accurate approximation to the true distribution. Note that in

this example $\Pr(\tilde{\mathbf{X}} < 0) = \Phi(-4)$ is very close to 0 as we would expect if the normal approximation is reasonable.

One feature that distinguishes this example from the Example 2 is that there is no elementary closed form for $\tilde{\mathbf{Y}}$; its values will typically have to be determined from a standard normal probability table rather than by direct analytic calculation.

$\tilde{\mathbf{Y}}$ has a mixed distribution with weights

$$w_1 = \Pr(\tilde{\mathbf{X}} \leq 500) = \Phi(1) = 0.8413 \text{ and}$$

$$w_2 = \Pr(\tilde{\mathbf{X}} > 500) = 1 - w_1 = 0.1587:$$

The distribution function of $\tilde{\mathbf{Y}}$ is $CDF_{\tilde{\mathbf{Y}}}(\tilde{y}) = w_1 CDF_{\tilde{\mathbf{Y}}_1}(\tilde{y}) + w_2 CDF_{\tilde{\mathbf{Y}}_2}(\tilde{y})$ where

- The *cdf* of the discrete component, $\tilde{\mathbf{Y}}_1$, is $CDF_{\tilde{\mathbf{Y}}_1}(\tilde{y}) = \begin{cases} 0 & \tilde{y} < 0 \\ 1 & \tilde{y} \geq 0 \end{cases}$.
- The *cdf* of the continuous component, $\tilde{\mathbf{Y}}_2$, is 0 for values \tilde{y} less than 0, and is equal to the conditional *cdf* of $\tilde{\mathbf{X}} - 500$ given $\tilde{\mathbf{X}} > 500$ for on positive values of \tilde{y} .

$$CDF_{\tilde{\mathbf{Y}}_2}(y) = \begin{cases} 0 & y < 0 \\ \Pr(\tilde{\mathbf{X}} - 500 \leq y | \tilde{\mathbf{X}} > 500) & y \geq 0 \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{1 + \frac{\tilde{y}}{100} - \Phi(1)}{1 - \Phi(1)} & y \geq 0 \end{cases}$$

The weighted average of these functions is

$$CDF_{\tilde{\mathbf{Y}}}(y) = \begin{cases} \Phi(1) \cdot 0 + (1 - \Phi(1)) \cdot 0 & y < 0 \\ \Phi(1) \cdot 1 + (1 - \Phi(1)) \cdot \frac{1 + \frac{\tilde{y}}{100} - \Phi(1)}{1 - \Phi(1)} & y \geq 0 \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ 1 + \frac{\tilde{y}}{100} & y \geq 0 \end{cases}$$

This is the expression we derived earlier without considering the components separately .

We can use the *cdf* of \tilde{Y} to calculate probabilities and percentiles:

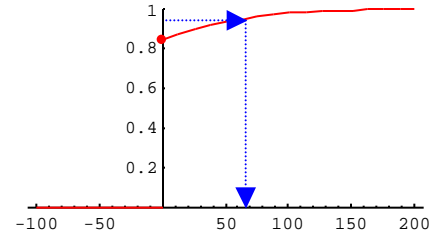
- a) $\Pr(\tilde{Y} > 1000) = 1 - CDF_{\tilde{Y}}(1000) = 1 - \left(1 + \frac{1000}{100}\right)^{-1} = 1 - \frac{1}{11} \approx 0.91$
- b) The 95th percentile of the claim amounts is the solution, y , of the equation $CDF_{\tilde{Y}}(y) = 0.95$.

Solving analytically:

$$0.95 = \left(1 + \frac{\tilde{y}}{100}\right)^{-1}$$

$$1 + \frac{\tilde{y}}{100} = \frac{1}{0.95} \approx 1.0526$$

$$\tilde{y} = 64.5$$



Thus, 95th percentile is approximately \$64.50.

Contrast these answers with those in Example 2 where more than 5% of the claims are expected to exceed \$1000. Although the exponential distribution of losses in Example 2 has a smaller mean than the normal distribution in Example 3, it has a larger standard deviation (\$300 versus \$100), and it is skewed toward large losses.

The more compact normal distribution of damages also produces a smaller expected claim amount than the exponential distribution:

$$E(\tilde{Y}) = w_1 \int_0^y \tilde{y} p(\tilde{y}) d\tilde{y} + w_2 \int_0^{\infty} \tilde{y} f(\tilde{y}) d\tilde{y}$$

$$= (1)(0) + (1 - (1)) \int_0^{\infty} \tilde{y} \frac{d}{d\tilde{y}} \left(\frac{1 + \frac{\tilde{y}}{100}}{1 - (1)} \right) d\tilde{y}$$

$$= 0 + \frac{1 - (1)}{1 - (1)} \int_0^{\infty} \tilde{y} \frac{1}{\sqrt{2}} e^{-1 + \frac{\tilde{y}}{100}} e^{-\tilde{y}^2 / (2 \cdot 100^2)} \frac{1}{100} d\tilde{y} = \int_0^{\infty} \tilde{y} \frac{1}{100\sqrt{2}} e^{-(\tilde{y}+100)^2 / (2 \cdot 100^2)} d\tilde{y}$$

$$= \int_{100}^{\infty} (t - 100) \frac{1}{100\sqrt{2}} e^{-t^2 / (2 \cdot 100^2)} dt$$

$$= \int_{100}^{\infty} t \frac{1}{100\sqrt{2}} e^{-t^2 / (2 \cdot 100^2)} dt - \int_{100}^{\infty} 100 \frac{1}{100\sqrt{2}} e^{-t^2 / (2 \cdot 100^2)} dt$$

$$= 100 \int_{1/2}^{\infty} \frac{1}{\sqrt{2}} e^{-u} du - 100 \int_1^{\infty} \frac{1}{\sqrt{2}} e^{-w^2 / 2} dw = \frac{100}{\sqrt{2}} e^{-1/2} - (\text{table value})$$

$$24.20 - 15.87 = 8.33$$

We can generalize the above examples to develop a method for problems involving insurance with a deductible. Suppose the deductible amount for the insurance is ,

and the distribution of loss amounts per insured event, \mathbf{X} , has continuous distribution function $cdf_{\mathbf{X}}(x) = F(x)$ and $pdf f(x)$. The claim amount per accident is

$$\mathbf{Y} = \begin{cases} 0 & \text{if } x \leq \delta \\ \mathbf{X} - \delta & \text{if } x > \delta \end{cases} = \max\{0, \mathbf{X} - \delta\}$$

Its cdf is

$$CDF_{\mathbf{Y}}(y) = \begin{cases} 0 & y < 0 \\ F(y + \delta) & y \geq 0 \end{cases}$$

This function is a weighted average of the cdf of a discrete random variable \mathbf{Y}_1 ,

$$CDF_{\mathbf{Y}_1}(y) = \begin{cases} 0 & y < 0 \\ 1 & y \geq 0 \end{cases}$$

and the cdf of a continuous random variable \mathbf{Y}_2 ,

$$CDF_{\mathbf{Y}_2}(y) = \begin{cases} 0 & y < 0 \\ \Pr(\mathbf{Y}_2 \leq y \mid y > 0) = \frac{F(y + \delta) - F(\delta)}{1 - F(\delta)} = \frac{\int_{\delta}^{y+\delta} f(t) dt}{\int_{\delta}^{\infty} f(t) dt} & y \geq 0 \end{cases}$$

The weights are

$$w_1 = \Pr(\mathbf{Y} = 0) = \Pr(\mathbf{X} \leq \delta) = F(\delta)$$

and

$$w_2 = \Pr(\mathbf{Y} > 0) = \Pr(\mathbf{X} > \delta) = 1 - F(\delta).$$

The discrete component has pmf (probability mass function) $p(y) = \begin{cases} 1 & y = 0 \\ 0 & y > 0 \end{cases}$, and

the continuous component has $pdf \frac{1}{1 - F(\delta)} f(y + \delta)$ for $y > 0$. Claim probabilities can be calculated by taking the weighted average of probabilities calculated either from the cdf 's or from the mixed pmf and pdf . The expected claim amount is

$$\begin{aligned} E(Y) &= w_1 \int_0^{\infty} yp(y) + w_2 \int_0^{\infty} yf(y)dy = w_1(0 \cdot 1) + w_2 \int_0^{\infty} y \frac{1}{1 - F(\delta)} f(y + \delta)dy \\ &= F(\delta) (0 \cdot 1) + \frac{1 - F(\delta)}{1 - F(\delta)} \int_0^{\infty} y f(y + \delta)dy = 0 + \int_{\delta}^{\infty} (x - \delta) f(x)dx \end{aligned}$$

Caps (A Ceiling for Claims)

Many insurance policies place a ceiling on claims paid. For simplicity, consider an insurance policy that pays 100% of the loss incurred up to cap, C , and then pays no more. Let \mathbf{X} be the amount of a loss to an insured, and \mathbf{Y} the amount of the claim paid. According to our assumptions, $\mathbf{Y} = \begin{cases} \mathbf{X} & x \leq C \\ C & x > C \end{cases}$. Suppose the loss amount has a continuous probability distribution with *cdf* $F(x)$ and *pdf* $f(x)$. Using reasoning similar to that for insurance with a deductible, we see that

$$cdf_{\mathbf{Y}}(y) = \begin{cases} F(y) & y < C \\ 1 & y \geq C \end{cases}$$

The graph coincides with the graph of F until $y = C$ where it jumps up to height 1 (the claim amount is certain to be C or less. The *cdf* is a weighted average of the discrete distribution with probability 1 at $y = C$, and the continuous distribution with *pdf*

$\frac{1}{F(C)} f(y)$ for $y < C$ and 0 elsewhere. (Since f has been truncated above C , the remaining portion integrates to $\int_0^C f(y) dy = F(C)$. It is necessary to divide by $F(C)$ to force the *pdf* to integrate to 1.) The discrete component of the distribution has weight $\Pr(\mathbf{X} \leq C) = F(C)$, and the continuous component has weight $1 - F(C)$.

Example 4: A dental insurance policy with a \$250 deductible will pay 80% of the annual cost of dental care after the first \$250 up to a maximum payment of \$2000. Suppose the total annual cost has an exponential distribution with mean \$200. Determine structure of the random variable representing the amount of a random claim amount.

Let \mathbf{X} be the annual cost of dental care and \mathbf{Y} the total claim paid for a randomly selected insured. By assumption \mathbf{X} has an exponential distribution with mean 200,

$$\text{and } \mathbf{Y} = \begin{cases} 0 & 0.8(\mathbf{X} - 250) < 0 \\ 0.8(\mathbf{X} - 250) & 0 \leq 0.8(\mathbf{X} - 250) < 2000 \\ 2000 & 0.8(\mathbf{X} - 250) \geq 2000 \end{cases} = \min\{1, \max\{0, 0.8(\mathbf{X} - 250)\}\}. \text{ The}$$

cdf of \mathbf{Y} is

$$CDF_{\mathbf{Y}}(y) = \Pr(\mathbf{Y} \leq y) = \begin{cases} 0 & y < 0 \\ \Pr(0.8(\mathbf{X} - 250) \leq y) & 0 \leq y < 2000 \\ 1 & 2000 \leq y \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \Pr(\mathbf{X} \leq \frac{1}{0.8}y + 250) & 0 \leq y < 2000 \\ 1 & 2000 \leq y \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{1}{0.8} e^{-y/200} & 0 \leq y < 2000 \\ 1 & y \geq 2000 \end{cases} = \begin{cases} 0 & y < 0 \\ 1 - e^{-\frac{1}{0.8} y + 250/200} & 0 \leq y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

This function has two jump discontinuities: a jump of height $1 - e^{-250/200} = 0.7135$ at $y = 0$ and a jump of height $CDF_Y(2000) - \lim_{y \rightarrow 2000^-} CDF_Y(y) = 1 - 1 - e^{-\frac{2000}{0.8} + 250/200}$

$= e^{-2750/200} = 0.0000011$ at $y = 2000$. The discrete component of the distribution must provide two jumps with the same *relative* heights, but the total of the two jumps must be 1. Since the total of the jumps above is $1 - e^{-250/200} + e^{-2750/200}$, we can construct the appropriate discrete distribution by dividing each jump height by the total of the jumps. These new jump heights will then sum to 1, as required for a discrete probability distribution. The discrete component therefore has the following *cdf* and *pmf*:

$$CDF_{Y_1}(y) = \begin{cases} 0 & y < 0 \\ \frac{1 - e^{-250/200}}{1 - e^{-250/200} + e^{-2750/200}} & 0 \leq y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

$$PMF_{Y_1}(y) = p(y) = \begin{cases} \frac{1 - e^{-250/200}}{1 - e^{-250/200} + e^{-2750/200}} & y = 0 \\ \frac{e^{-2750/200}}{1 - e^{-250/200} + e^{-2750/200}} & y = 2000 \\ 0 & \text{elsewhere} \end{cases}$$

The weight of the discrete component is $w_1 = \frac{1 - e^{-250/200}}{1 - e^{-250/200} + e^{-2750/200}}$, the total of all of the jump heights in Y 's distribution.

The continuous component of this mixed random variable is the random variable Y_2 with the following *cdf* and *pdf*:

$$CDF_{Y_2}(y) = \Pr \mathbf{X} \left\{ \frac{1}{0.8} y + 250 \leq X \right\} = \begin{cases} 0 & y < 0 \\ 0 & 0 \leq y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

$$CDF_{Y_2}(y) = \begin{cases} 0 & y < 0 \\ \Pr \mathbf{X} \geq \frac{1}{0.8}y + 250 & 0 < y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{\int_{\frac{1}{0.8}y+250}^{2750} \frac{1}{200} e^{-x/200} dx}{\int_{250}^{2750} \frac{1}{200} e^{-x/200} dx} & 0 < y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{e^{-250/200} - e^{-\frac{1}{0.8}y+250/200}}{e^{-250/200} - e^{-2750/200}} & 0 < y < 2000 \\ 1 & y \geq 2000 \end{cases}$$

$$PDF_{Y_2}(y) = f(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{e^{-250/200} - e^{-2750/200}} \cdot \frac{1}{0.8 \cdot 200} e^{-\frac{1}{0.8}y+250/200} & 0 < y < 2000 \\ 0 & y \geq 2000 \end{cases}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{1}{e^{-250/200} - e^{-2750/200}} \cdot \frac{1}{160} e^{-\frac{1}{0.8}y+250/200} & 0 < y < 2000 \\ 0 & y \geq 2000 \end{cases}$$

The weight of the continuous component is $w_2 = 1 - w_1 = e^{-250/200} - e^{-2750/200}$, the total vertical distance spanned by the increasing portions of the graph of $CDF_Y(y)$.

Let's generalize these results to any insurance product where the insured's loss is a continuous random variable, \mathbf{X} , with cdf $F(x)$ and pdf $f(x)$ and where the insurance will pay the claim, \mathbf{Y} , equal to the proportion r of the excess of losses above the deductible up to a cap of C . That is,

$$\mathbf{Y} = \begin{cases} 0 & r(x - \delta) < 0 \\ r(x - \delta) & 0 < r(x - \delta) < C \\ C & r(x - \delta) \geq C \end{cases} = \min\{1, \max\{0, r(x - \delta)\}\}$$

The *cdf* of \mathbf{Y} is

$$CDF_{\mathbf{Y}}(y) = \Pr(\mathbf{Y} \leq y) = \begin{cases} 0 & y < 0 \\ \Pr(r(\mathbf{X} - \delta) \leq y) & 0 \leq y < C \\ 1 & y \geq C \end{cases} = F \begin{cases} \frac{1}{r}y + \delta & 0 \leq y < C \\ & y \geq C \end{cases}$$

The discrete component of the *cdf* of \mathbf{Y} is

$$\begin{aligned} CDF_{\mathbf{Y}}(y) &= \Pr(\mathbf{X} \leq \delta \mid \mathbf{X} \leq \delta \text{ or } \mathbf{Y} \leq C) = \begin{cases} 0 & y < 0 \\ 1 & 0 \leq y < C \\ & y \geq C \end{cases} \\ &= \Pr(\mathbf{X} \leq \delta \mid \mathbf{X} \leq \delta \text{ or } \mathbf{X} > \frac{1}{r}C + \delta) = \begin{cases} 0 & y < 0 \\ 1 & 0 \leq y < C \\ & y \geq C \end{cases} \\ &= \frac{F(\delta)}{F(\delta) + 1 - F(\frac{1}{r}C + \delta)} = \begin{cases} 0 & y < 0 \\ & 0 \leq y < C \\ 1 & y \geq C \end{cases} \end{aligned}$$

The corresponding probability mass function is

$$\begin{aligned} PMF_{\mathbf{Y}}(y) = p(y) &= \Pr(\mathbf{X} \leq \delta \mid \mathbf{X} \leq \delta \text{ or } \mathbf{X} > \frac{1}{r}C + \delta) = \begin{cases} 0 & y = 0 \\ & y = C \\ & \text{elsewhere} \end{cases} \\ &= \frac{\Pr(\mathbf{X} \leq \delta)}{\Pr(\mathbf{X} \leq \delta) + \Pr(\mathbf{X} > \frac{1}{r}C + \delta)} = \begin{cases} \frac{F(\delta)}{F(\delta) + 1 - F(\frac{1}{r}C + \delta)} & y = 0 \\ \frac{\Pr(\mathbf{X} > \frac{1}{r}C + \delta)}{\Pr(\mathbf{X} \leq \delta) + \Pr(\mathbf{X} > \frac{1}{r}C + \delta)} & y = C \\ 0 & \text{elsewhere} \end{cases} \\ &= \frac{1 - F(\frac{1}{r}C + \delta)}{F(\delta) + 1 - F(\frac{1}{r}C + \delta)} = \begin{cases} \frac{F(\delta)}{F(\delta) + 1 - F(\frac{1}{r}C + \delta)} & y = 0 \\ \frac{1 - F(\frac{1}{r}C + \delta)}{F(\delta) + 1 - F(\frac{1}{r}C + \delta)} & y = C \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

The weight of the discrete component in the *cdf* is $w_1 = F(\delta) + 1 - F\left(\frac{1}{r}C + \delta\right)$
 $= \Pr\left(X \leq \frac{1}{r}C + \delta\right) + \Pr(X > \delta)$.

The continuous component of the *cdf* of \mathbf{Y} is

$$\begin{aligned} CDF_{Y_2}(y) &= \Pr\left(\mathbf{X} \leq \frac{1}{r}y + C \mid \delta < \mathbf{X} < \frac{1}{r}C + \delta\right) && 0 < y < C \\ &0 && \text{elsewhere} \\ &= \frac{\Pr\left(\delta < \mathbf{X} < \frac{1}{r}y + \delta\right)}{\Pr\left(\delta < \mathbf{X} < \frac{1}{r}C + \delta\right)} && 0 < y < C \\ &0 && \text{elsewhere} \end{aligned} = \frac{F\left(\frac{1}{r}y + \delta\right) - F(\delta)}{F\left(\frac{1}{r}C + \delta\right) - F(\delta)} \begin{matrix} 0 < y < C \\ 0 \\ \text{elsewhere} \end{matrix}$$

The corresponding probability density function is

$$\begin{aligned} PDF_{Y_2}(y) &= \frac{1}{\Pr\left(\delta < \mathbf{X} < \frac{1}{r}C + \delta\right)} f\left(\frac{1}{r}y + \delta\right) && 0 < y < C \\ &0 && \text{elsewhere} \\ &= \frac{1}{F\left(\frac{1}{r}C + \delta\right) - F(\delta)} f\left(\frac{1}{r}y + \delta\right) && 0 < y < C \\ &0 && \text{elsewhere} \end{aligned}$$

The weight of the continuous component is $w_2 = 1 - w_1 = F\left(\frac{1}{r}C + \delta\right) - F(\delta)$
 $= \Pr\left(\delta < X \leq \frac{1}{r}C + \delta\right)$.

Heterogeneous Classes

All of these discrete mixtures can be interpreted as mixtures of different classes of insurance policies. The classes partition the collection of policies into non-overlapping groups that behave differently in terms of the claims that they generate. This enables us to sort the policies into homogeneous groups, simplifying calculation.

Example 5: Let's generalize these results to any insurance product where the insured's loss is a continuous random variable, \mathbf{X} , with *cdf* $F(x)$ and *pdf* $f(x)$ and

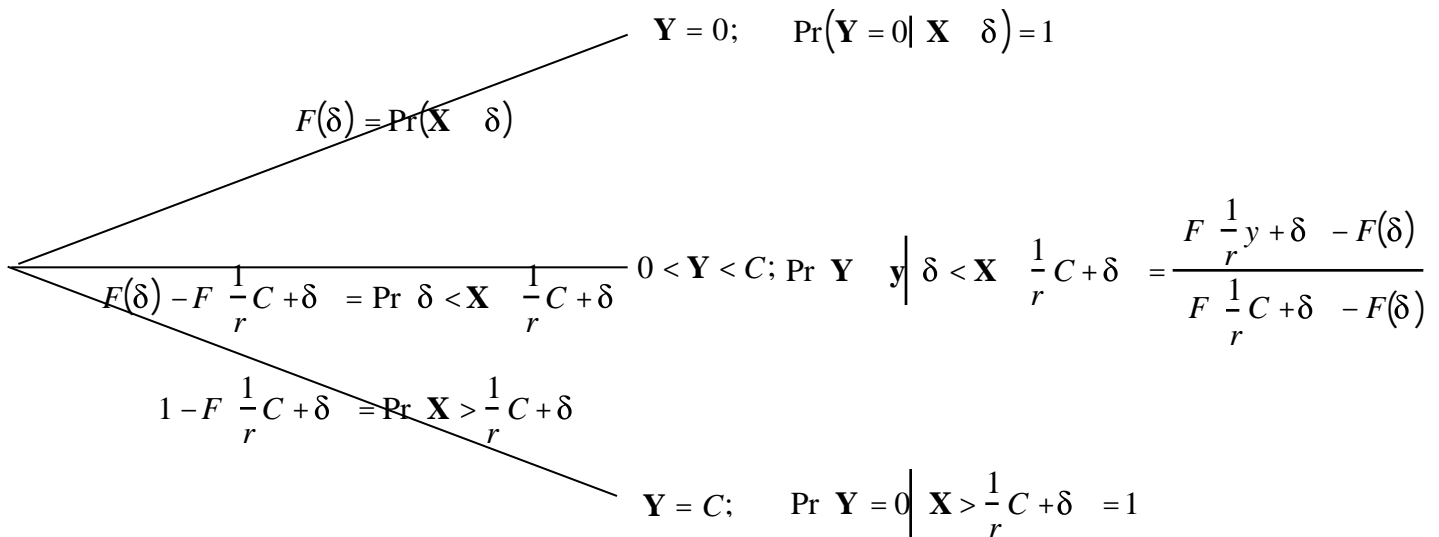
where the insurance will pay the claim, Y , equal to the proportion r of the excess of losses above the deductible up to a cap of C . There are three different policy classes in this example:

- policies whose policy holders who experience losses that are less than the deductible amount; these policy holders receive no payment back
- policies whose policy holders experience losses in excess of the deductible amount, but whose payments do not exceed the claim cap
- policies whose policy holders experience losses large enough to receive the maximum payment.

The classes are defined by the claim payment amount, Y , but since probabilities are determined from the known distribution of the loss incurred, X , conditions on Y should be translated into conditions on X . In this example we sort into classes as follows:

- $Y = 0$ corresponds to $0 < X \leq \delta$; the probability of falling into this class is $\Pr(0 < X \leq \delta) = F(\delta) - F(0) = F(\delta)$
- $Y = r(X - \delta)$ corresponds to $0 < Y = r(X - \delta) < C$ or $\delta < X < \frac{1}{r}C + \delta$; the probability of falling into this class is $\Pr(\delta < X < \frac{1}{r}C + \delta) = F(\frac{1}{r}C + \delta) - F(\delta)$
- $Y = C$ corresponds to $X > \frac{1}{r}C + \delta$; the probability of falling into this class is $\Pr(X > \frac{1}{r}C + \delta) = 1 - F(\frac{1}{r}C + \delta)$

It may be useful use a probability tree to visualize the situation:



The three nodes of the three represent the three policy classes, Each branch is labeled by the proportion of policies in the class. The conditional probability distribution for

each class is indicated to the right of the class node. (Recall that probabilities beyond the first stage of a probability tree are conditioned on the path to that point)

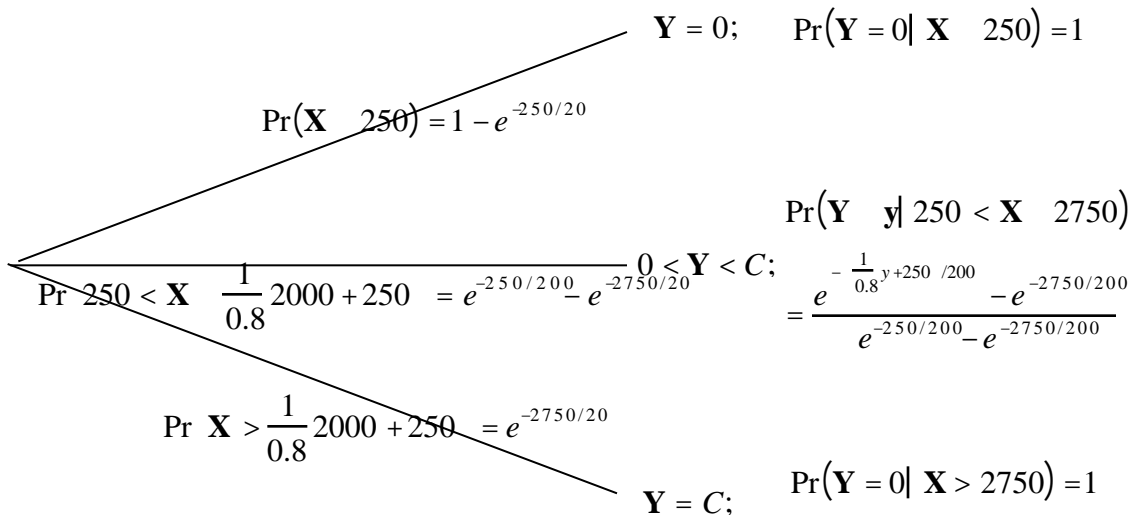
The classes were clear in the initial description of the claim variable, Y , in Example 4:

$$Y = \begin{cases} 0 & r(x - \delta) < 0 \\ r(\mathbf{X} - \delta) & 0 < r(x - \delta) < C. \\ C & r(x - \delta) \end{cases}$$

However, in Example 4, the cdf of Y was decomposed into two parts, the discrete and continuous components. The two-jump discrete component was formed by combining the jumps at $y = 0$ and $y = C$. It might help conceptually to consider the cases $y = 0$ and $y = C$ separately as we have done in the probability tree.

Example 6: In example 4 we analyzed the distribution of reimbursement, Y , for a dental policy. Recall that the policy will pay 80% of the annual cost of dental care after the first \$250 up to a maximum payment of \$2000. Determine the $P(Y = 1000)$, the 95th percentile and the expected value of Y .

There are three types of policies: those where expenditures do not exceed the deductible, those where expenditures are high enough to trigger the payment cap and those in between. The associated sorting tree is shown below:



Calculating Probabilities: Recall that to calculate probabilities using a probability tree, we must multiply probabilities along the relevant path. Remember when calculating the probability of an event that crosses over two or more categories to consider all cases.

$$\begin{aligned}
\Pr(\mathbf{Y} \leq 1000) &= \Pr(\mathbf{Y} = 0) + \Pr(0 < \mathbf{Y} \leq 1000) \\
&= \Pr(\mathbf{Y} = 0) + \Pr(0 < \mathbf{Y} \leq 2000) \Pr(\mathbf{Y} \leq 1000 | 0 < \mathbf{Y} \leq 2000) \\
&= \Pr(\mathbf{X} \leq 250) \Pr(\mathbf{Y} = 0 | \mathbf{X} \leq 250) + \Pr(250 < \mathbf{X} \leq 2750) \Pr(\mathbf{X} \leq 1500 | 250 < \mathbf{X} \leq 2750) \\
&= \Pr(\mathbf{X} \leq 250) \frac{\Pr(\mathbf{X} \leq 250)}{\Pr(\mathbf{X} \leq 250)} + \Pr(250 < \mathbf{X} \leq 2750) \frac{\Pr(250 < \mathbf{X} \leq 1500)}{\Pr(250 < \mathbf{X} \leq 2750)} \\
&= \Pr(\mathbf{X} \leq 250) + \Pr(250 < \mathbf{X} \leq 1500) \\
&= \Pr(\mathbf{X} \leq 1500) = 1 - e^{-1500/200} = 0.99945
\end{aligned}$$

In general, for values of y greater than 0 but less than 2000,

$$\begin{aligned}
\Pr(\mathbf{Y} \leq y) &= \Pr(\mathbf{Y} < 0) + \Pr(\mathbf{Y} = 0) + \Pr(0 < \mathbf{Y} \leq y) \\
&= \Pr(\mathbf{Y} = 0) + \Pr(0 < \mathbf{Y} \leq 2000) \Pr(\mathbf{Y} \leq y | 0 < \mathbf{Y} \leq 2000) \\
&= \Pr(\mathbf{X} \leq 250) \frac{\Pr(\mathbf{X} \leq 250)}{\Pr(\mathbf{X} \leq 250)} + \Pr(250 < \mathbf{X} \leq 2750) \frac{\Pr(250 < \mathbf{X} \leq \frac{1}{.08}y + 250)}{\Pr(250 < \mathbf{X} \leq 2750)} \\
&= \Pr(\mathbf{X} \leq 250) + \Pr(250 < \mathbf{X} \leq \frac{1}{.08}y + 250) \\
&= \Pr(\mathbf{X} \leq \frac{1}{.08}y + 250) = 1 - e^{-\frac{1}{.08}y + 250 / 200}
\end{aligned}$$

Note that we could have made this calculation easily from the definition of \mathbf{Y} without using the tree:

$$\mathbf{Y} = \begin{matrix} 0 & 0.8(x - 250) & 0 & 0 & x - 250 \\ 0.8(\mathbf{X} - 250) & 0 < 0.8(x - 250) < 2000 & = & 0.8(\mathbf{X} - 250) & 250 < x < 2750 \\ 2000 & 2000 & 0.8(x - 250) & 2000 & 2750 & x \end{matrix}$$

so for values of y between 0 and 2000,

$$\begin{aligned}
\Pr(\mathbf{Y} \leq y) &= \Pr(\mathbf{Y} = 0) + \Pr(0 < \mathbf{Y} \leq y) = \Pr(\mathbf{X} \leq 250) + \Pr(250 < \mathbf{X} \leq \frac{1}{.08}y + 250) \\
&= \Pr(\mathbf{X} \leq \frac{1}{.08}y + 250) = 1 - e^{-\frac{1}{.08}y + 250 / 200}
\end{aligned}$$

Percentiles and Other Inverse Probabilities: Since .95 does not lie within a vertical interval corresponding to a jump on the *cdf* of \mathbf{Y} , we find the the 95th percentile of \mathbf{Y} can be found as follows:

$$.95 = \Pr(\mathbf{Y} \leq y) = 1 - e^{-\frac{1}{.08}y + 250 / 200} = e^{-\frac{1}{.08}y + 250 / 200} = 0.05 \quad y = -1.601 \ln(0.05) - 200 = 279.32$$

Expected Value: The expected reimbursement (in dollars) is

$$\begin{aligned}
 E(\mathbf{Y}) &= \Pr(\mathbf{Y} = 0) \cdot 0 + \Pr(250 < \mathbf{X} < 2750) \int_0^{2000} y \frac{d}{dy} \frac{F\left(\frac{1}{0.8}y + 250\right) - F(250)}{\Pr(250 < \mathbf{X} < 2750)} dy + \Pr(\mathbf{Y} = 2000) \cdot 2000 \\
 &= 0 + \int_0^{2000} y \frac{d}{dy} \left[F\left(\frac{1}{0.8}y + 250\right) - F(250) \right] dy + \Pr(\mathbf{Y} = 2000) \cdot 2000 \\
 &= 2000 \left[F\left(\frac{1}{0.8}(2000) + 250\right) - F(250) \right] - \int_0^{2000} F\left(\frac{1}{0.8}y + 250\right) dy + \Pr(\mathbf{Y} = 2000) \cdot 2000 \\
 &= 2000 \left[F\left(\frac{1}{0.8}(2000) + 250\right) - F(250) \right] - \int_0^{2000} F\left(\frac{1}{0.8}y + 250\right) dy + \left[1 - F\left(\frac{1}{0.8}(2000) + 250\right) \right] \cdot 2000 \\
 &= 2000 - \int_0^{2000} F\left(\frac{1}{0.8}y + 250\right) dy = 2000 - \int_0^{2000} \left[1 - e^{-\frac{1}{0.8}y + 250 / 200} \right] dy = \int_0^{2000} e^{-\frac{1}{0.8}y + 250 / 200} dy \\
 &= -0.8 \cdot 200 e^{-\frac{1}{0.8}y + 250 / 200} \Big|_0^{2000} = 160 \left(e^{-250/200} - e^{-2750/200} \right) = 45.84
 \end{aligned}$$

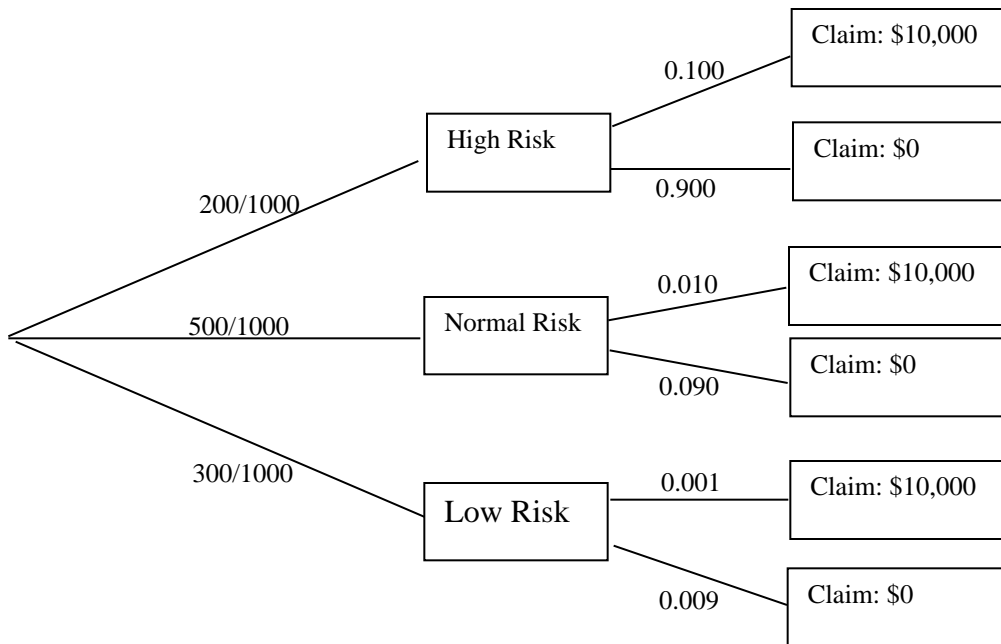
Discrete Mixtures and Mixed Distributions

In all of the examples considered thus far, a discrete mixture produced a mixed (neither discrete nor continuous) distribution. This is because we were mixing a discrete random variable and a continuous random variable. If we took a discrete mixture of two random variables of the same type, the mixture would inherit that type (discrete or continuous). You can see this in insurance examples involving heterogeneous classes whose loss or claim experience are all discrete or all continuous.

Example 7: A supplemental medical policy pays \$10,000 for the loss of a limb. The insured population can be divided into three different risk classes: high risk, normal risk and low risk. The company expects about 10% of the high risk policy holders, 1% of the normal risk policy holders and 0.1% of the low risk policy holders to collect on the policy. If there are 100 people in the high risk group, 500 in the normal risk group and 400 in the low risk group, what is the expected claim per policy holder?

Let \mathbf{X} be the amount of a randomly selected claim. In order to analyze the behavior of \mathbf{W} , we break down the heterogeneous group of insureds into their homogeneous risk classes.

We can summarize this situation in a tree::



Calculating probabilities: \mathbf{X} must take one of two values: \$0 or \$10,000. The probabilities can be determined by applying the law of total probability to a decomposition into risk classes:

$$\begin{aligned} \Pr(\mathbf{X} = 0) &= \Pr(\mathbf{X} = 0 \text{ High}) + \Pr(\mathbf{X} = 0 \text{ Normal}) + \Pr(\mathbf{X} = 0 \text{ Low}) \\ &= \Pr(\text{High}) \Pr(\mathbf{X} = 0 | \text{High}) + \Pr(\text{Normal}) \Pr(\mathbf{X} = 0 | \text{Normal}) \\ &\quad + \Pr(\text{Low}) \Pr(\mathbf{X} = 0 | \text{Low}) \\ &= 0.2 \cdot 0.9 + 0.5 \cdot 0.99 + 0.3 \cdot 0.999 = 0.18 + 0.495 + 0.2997 = 0.9747 \end{aligned}$$

$$\begin{aligned} \Pr(\mathbf{X} = 10000) &= \Pr(\mathbf{X} = 10000 \text{ High}) + \Pr(\mathbf{X} = 10000 \text{ Normal}) + \Pr(\mathbf{X} = 10000 \text{ Low}) \\ &= \Pr(\text{High}) \Pr(\mathbf{X} = 10000 | \text{High}) + \Pr(\text{Normal}) \Pr(\mathbf{X} = 10000 | \text{Normal}) \\ &\quad + \Pr(\text{Low}) \Pr(\mathbf{X} = 10000 | \text{Low}) \\ &= 0.2 \cdot 0.1 + 0.5 \cdot 0.01 + 0.3 \cdot 0.001 = 0.02 + 0.005 + 0.0003 = 0.0253 = 1 - 0.9747 \end{aligned}$$

This gives us a probability table for the discrete mixture \mathbf{X} :

x	$p(x) = \Pr(\mathbf{X} = x)$
0	0.9747
10000	0.0253

Notice that this discrete mixture of discrete random variables, \mathbf{X}_{High} , $\mathbf{X}_{\text{Normal}}$ and \mathbf{X}_{Low} produces a discrete random variable. The weights are, respectively $w_1 = 0.2$, $w_2 = 0.5$ and $w_3 = 0.3$, the relative proportions of the three risk classes.

Expected Value: The expected (average) claim amount can be calculated directly from the probability distribution that we determined above:

$$E(\mathbf{X}) = \$10000 \cdot 0.0253 + \$0 \cdot 0.9747 = \$253.00$$

or we can follow the general rule for calculating the expected value of a discrete mixture (see page 4) by taking the weighted average of the expected values of \mathbf{X}_{High} , \mathbf{X}_{Normal} and \mathbf{X}_L :

$$\begin{aligned} E(\mathbf{X}) &= 0.2E(\mathbf{X}_{High}) + 0.5E(\mathbf{X}_{Normal}) + 0.3E(\mathbf{X}_{Low}) \\ &= 0.2(0.90 \$0 + 0.10 \$10000) + 0.5(0.99 \$0 + 0.01 \$10000) \\ &\quad + 0.3(0.999 \$0 + 0.001 \$10000) \\ &= (0.18 + 0.495 + 0.2997) \$0 + (0.02 + 0.005 + 0.0003) \$10000 \\ &= 0.0253 \$10000 = \$253.00 \end{aligned}$$