HKU STAT3906 Study Notes

Chiu Ka Long (Leo)*

Last Updated: 2025-02-26

[Note: Related SOA Exam: FAM (short-term)]

Contents

*email: leockl@connect.hku.hk; personal website: https://leochiukl.github.io

1 Basic Distributional Quantities

1.0.1 Here we will introduce some distributional quantities related to a random variable. (Some of them are discussed in STAT2901.)

1.1 Raw and Central Moments

1.1.1 The **kth raw moment** (or kth moment) of a random variable X, denoted by μ'_k, is $\mathbb{E}\left[X^k\right]$.

[Note: The 1st raw moment of a random variable X is the *mean* of X, and is commonly denoted by μ.]

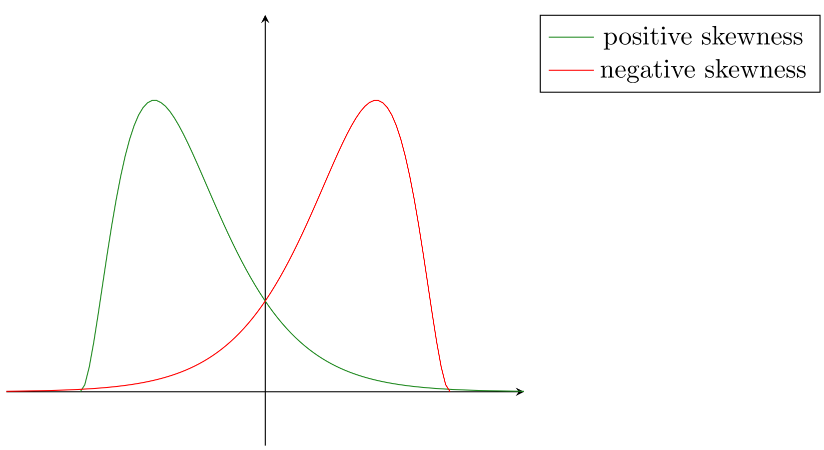
1.1.2 The **kth central moment** of X, denoted by μ_k, is $\mathbb{E}\left[(X-\mu)^k\right]$.

1.1.3 Some quantities of interest related to central moment are as follows.

Quantity	Definition	Notation
variance	μ_2	σ^2
standard deviation	$\sqrt{\mu_2}$	σ
coefficient of variation	σ/μ	/
skewness	μ_3/σ^3	γ_1
kurtosis	μ_4/σ^4	γ_2

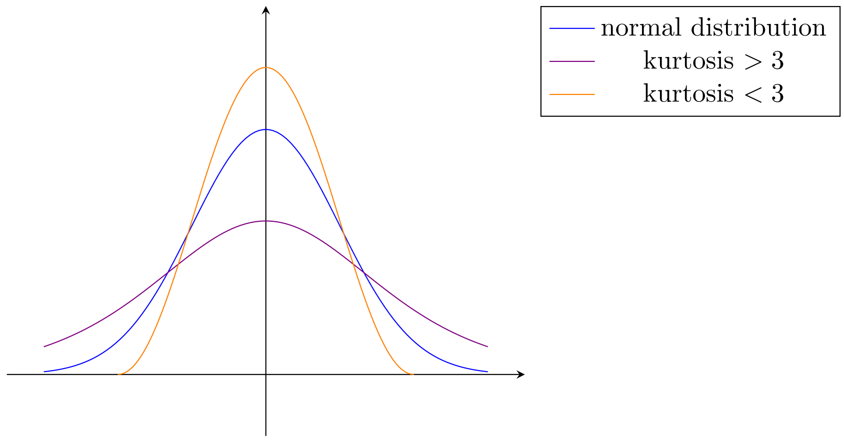
1.1.4 Here we give interpretations of the quantities not covered in STAT2901:

- *Coefficient of variation*: It is "relative" standard deviation (standard deviation per unit mean).
- *Skewness*: It is a measure of *asymmetry*. A symmetric distribution has a skewness of zero. [Warning: This does not mean a distribution with zero skewness is necessarily symmetric!] Positive (negative) skewness indicates that the right (left) tail is *longer* and the mass of distribution is concentrated on the left (right).



[Intuition: Skewness can be written as $\mathbb{E}\left[(X-\mu)/\sigma^3\right]$. Since the term inside is raised to power 3, long right (left) tail contributes very positively (negatively) to skewness value.]

- *Kurtosis*: It measures "flatness" of the distribution relative to a *normal distribution*, which has a kurtosis of 3. When kurtosis is above (below) 3, it is "flatter" ("less flat") than normal distribution, in the sense that more mass of distribution is located away from mean relative to a normal distribution, keeping standard deviation constant.

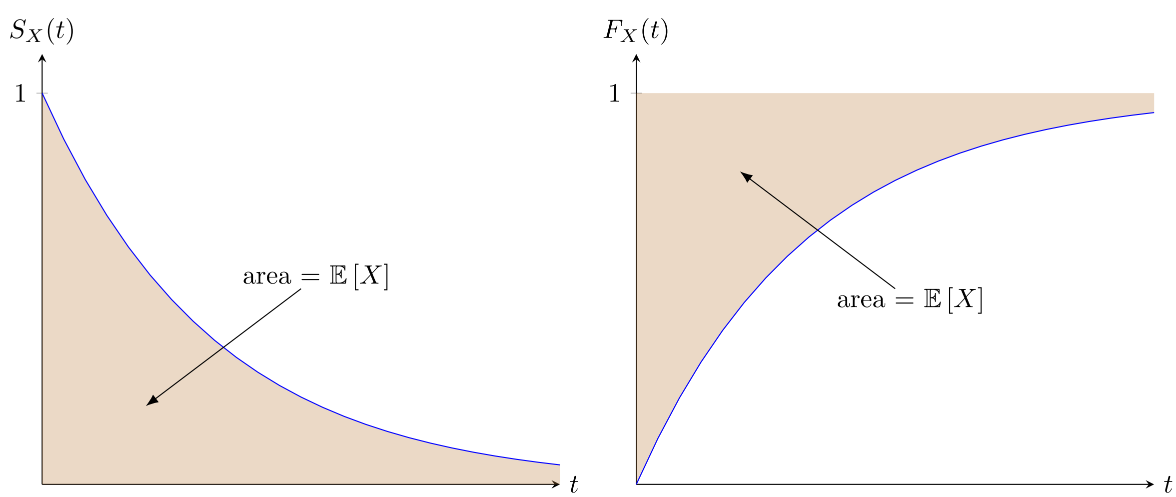


[Intuition: Kurtosis can be written as $\mathbb{E}\left[((X-\mu)/\sigma)^4\right]$. Since the term in the expectation is raised to power 4, mass near mean (with $(X-\mu)/\sigma < 1$) contributes very little to the kurtosis value, while mass away from mean (large $(X-\mu)/\sigma$) contributes a lot to the kurtosis value.]

1.1.5 The following result provides an useful formula for computing mean.

Proposition 1.1.a. Let X be a nonnegative random variable with finite mean (i.e., $\mathbb{E}[X] < \infty$), and let $S_X(x) = \mathbb{P}(X > x)$ be its survival function. Then,

$$\mathbb{E}[X] = \int_0^\infty S_X(t)\,\mathrm{d}t\,.$$



Proof. Since $X \geq 0$, we have

$$X = \int_0^X 1\,\mathrm{d}t = \int_0^\infty \mathbf{1}_{\{t<X\}}\,\mathrm{d}t$$ [1]

Thus,

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{t<X\}}\,\mathrm{d}t\right] = \int_0^\infty \mathbb{E}\left[\mathbf{1}_{\{X>t\}}\right]\mathrm{d}t = \int_0^\infty \mathbb{P}(X>t)\,\mathrm{d}t\,.$$

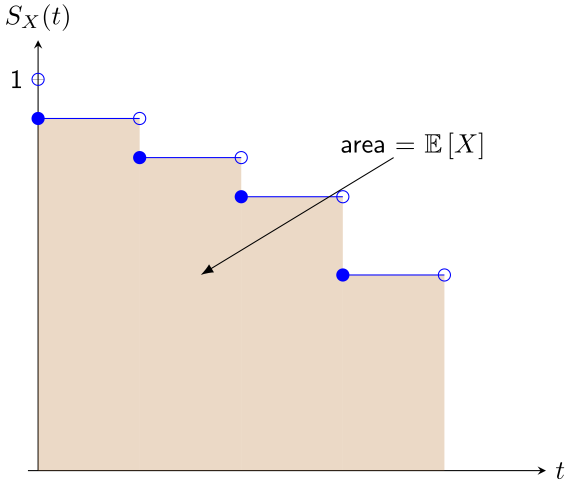
□

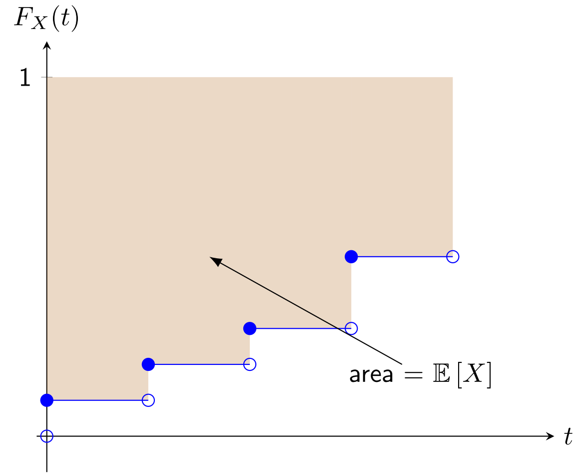
Remarks:

- This result holds no matter X is discrete or continuous.

[1] $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function.

- This result suggests a geometrical interpretation of mean of (nonnegative) X: area under the graph of S_X or area between the graph of $y = F_X(t)$ and the line $y = 1$. This still holds when X is discrete!



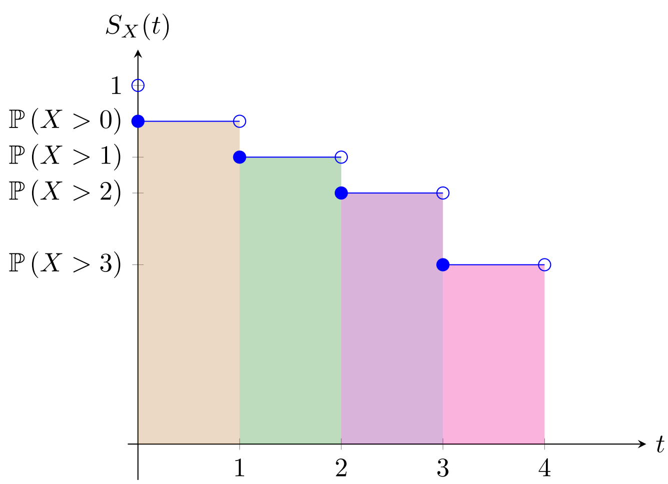


1.1.6 As a corollary, we have the following the result for *discrete* random variable.

Corollary 1.1.b. Let X be a nonnegative discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}(X > n).$$

Proof. $\mathbb{E}[X]$ is the area under S_X:



By summing up the areas of the rectangles, the area under S_X is

$$\mathbb{P}(X > 0) \cdot 1 + \mathbb{P}(X > 1) \cdot 1 + \mathbb{P}(X > 2) \cdot 1 + \mathbb{P}(X > 3) \cdot 1 + \cdots = \sum_{n=0}^{\infty} \mathbb{P}(X > n).$$

□

1.2 Stop Loss Variables

1.2.1 Fix any real number d, and consider a loss random variable X (positive value $\leftrightarrow$ positive loss). Then, the **stop loss variable** is

$$(X - d)_+ = \begin{cases} X - d & \text{if } X > d; \\ 0 & \text{if } X \le d \end{cases}$$

where $x_+ = \max\{x, 0\}$ is the **positive part** of x.

1.2.2 For a practical interpretation of stop loss variable, consider the following. Suppose that the insurer insures a loss X with a **deductible** of d dollars, i.e., the policyholder suffering the loss X is responsible for first d dollars of loss, and insures the remaining portion (if exists). Then, the stop loss variable represents the payment made by :

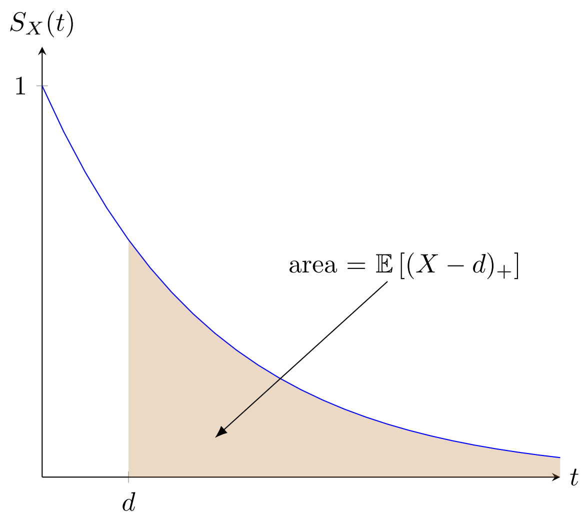
- If the loss $X \leq d$, then there is no payment.
- If the loss $X > d$, then the payment amount is $X - d$.

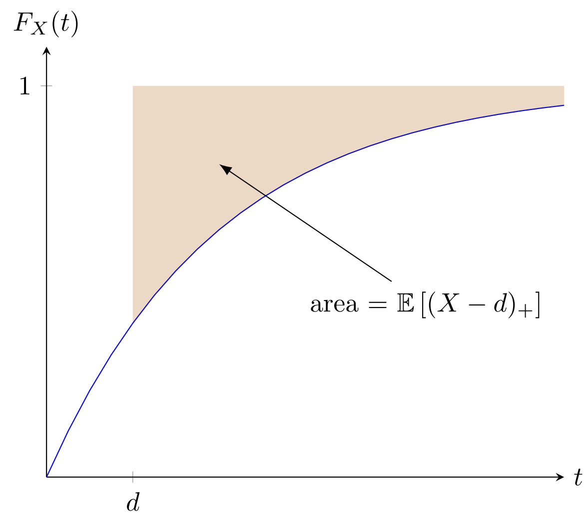
[Note: By having such insurance (called **stop-loss insurance**), the policyholder suffers *at most* d dollars of loss, so the insurance "stops" the loss suffered by from d dollars onwards, hence "stop loss".]

1.2.3 We have the following result for the stop loss variable, which is similar to Proposition 1.1.a.

Proposition 1.2.a. Let X be a random variable with finite mean and let $S_X(x)$ be its survival function. Then, for any $d \in \mathbb{R}$,

$$\mathbb{E}\left[(X-d)_+\right] = \int_d^\infty S_X(x)\,\mathrm{d}x\,.$$

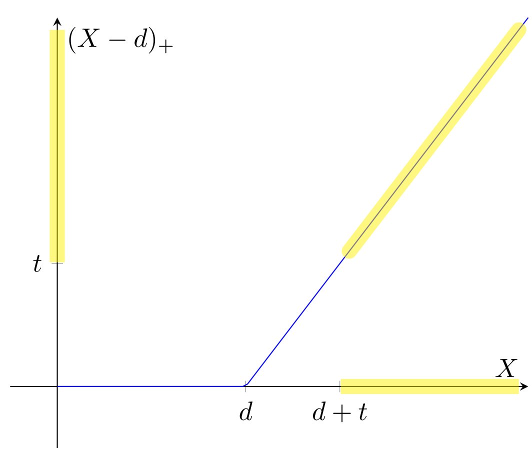




Proof. Since $(X-d)_+ \geq 0$, by Proposition 1.1.a we have

$$\mathbb{E}\left[(X-d)_+\right] = \int_0^\infty \mathbb{P}\left((X-d)_+ > t\right)\mathrm{d}t = \int_0^\infty \mathbb{P}\left(X > d+t\right)\mathrm{d}t \overset{x=t+d}{=} \int_d^\infty S_X(x)\,\mathrm{d}x$$

where the second equality holds as $(X-d)_+ > t \iff X > d+t$ for any $t \geq 0$.



$\square$

1.2.4 Alternatively, one can compute the expectation by definition. However, Proposition 1.2.a is often more handy.

If X is continuous with pdf f_X, then

$$\mathbb{E}\left[(X-d)_+\right] = \int_{-\infty}^{\infty} (x-d)_+ f_X(x)\,\mathrm{d}x = \int_{-\infty}^{d} 0\,\mathrm{d}x + \int_{d}^{\infty} (x-d) f_X(x)\,\mathrm{d}x = \boxed{\int_{d}^{\infty} (x-d) f_X(x)\,\mathrm{d}x}.$$

On the other hand, if X is discrete, then

$$\mathbb{E}\left[(X-d)_+\right] = \sum_{j} (x_j - d)_+ \mathbb{P}\left(X = x_j\right) = \boxed{\sum_{x_j > d} (x_j - d) \cdot p_j}.$$

where $p_j = \mathbb{P}\left(X = x_j\right)$.

1.3 Excess Loss Variables

1.3.1 Consider again a loss random variable X and fix any $d \in \mathbb{R}$ with $\mathbb{P}\left(X > d\right) > 0$. Then, the **excess loss variable** (or **residual lifetime**/"future lifetime random variable for a life aged d" (in life contingencies)) is

$$Y = \begin{cases} \text{undefined} & \text{if } X \le d; \\ X - d & \text{if } X > d \end{cases} = (X - d | X > d).$$

Remarks:

- The excess loss variable gives the amount of loss in excess of the deductible, *given that* such excess of loss exists. If there is no such excess of loss, we *do not define* excess loss variable.
- More precisely, the excess loss variable Y has the same distribution as the *conditional* distribution of $X - d$ given $X > d$.

1.3.2 The "form" of excess loss variable is similar to that of stop loss variable. Indeed, it also describes the "payment" made by the insurer for a stop-loss insurance, by from 's perspective. When $X > d$, then there is no difference from the stop loss variable: The payment is of amount $X - d$.

However, when $X \le d$, since in practice the policyholder *would not even report the loss to* , as needs to bear the full responsibility of it anyway.

Hence, such losses are *not found* in 's record (database , spreadsheet , etc.), meaning that they *do not exist* from 's perspective. Thus, we have an "undefined" payment amount.

1.3.3 We are usually interested in studying the *expected value* of the excess loss variable, which is called the **mean excess loss function** (or **mean residual lifetime** (MRL)/"complete expectation of life" in life contingencies):

$$e_X(d) = \mathbb{E}\left[Y\right] = \mathbb{E}\left[X - d | X > d\right].$$

[Note: In life contingencies, the actuarial notation is $\mathring{e}_d$.]

1.3.4 The MRL $e_X(d)$ and the expected stop loss variable $\mathbb{E}\left[(X-d)_+\right]$ can be related as follows.

$$e_X(d) = \frac{\mathbb{E}\left[(X-d)\mathbf{1}_{\{X>d\}}\right]}{\mathbb{P}\left(X > d\right)} = \frac{\mathbb{E}\left[(X-d)\mathbf{1}_{\{X>d\}} + 0 \cdot \mathbf{1}_{\{X \le d\}}\right]}{\mathbb{P}\left(X > d\right)} = \boxed{\frac{\mathbb{E}\left[(X-d)_+\right]}{\mathbb{P}\left(X > d\right)}}.$$

[Note: Using Proposition 1.2.a, we can further write

$$e_X(d) = \boxed{\frac{\int_{d}^{\infty} S_X(x)\,\mathrm{d}x}{\mathbb{P}\left(X > d\right)}}.$$

]

1.4 Limited Loss Variables

1.4.1 Fix any $u \in \mathbb{R}$ and consider a loss random variable X. Then, the **limited loss variable** is

$$X \wedge u = \min\{X, u\} = \begin{cases} X & \text{if } X \leq u; \\ u & \text{if } X > u. \end{cases}$$

1.4.2 For a practical interpretation of limited loss variable, consider the following. Suppose that the insurer 🏢 insures a loss X with a policy limit of u dollars, i.e., the maximum loss insured is u dollars. Then, the limited loss variable represents the payment made by 🏢:

- If the loss $X \leq u$, then 🏢 pays the full amount u to the policyholder 👤.
- If the loss $X > u$, then 🏢 only pays u dollars to 👤.

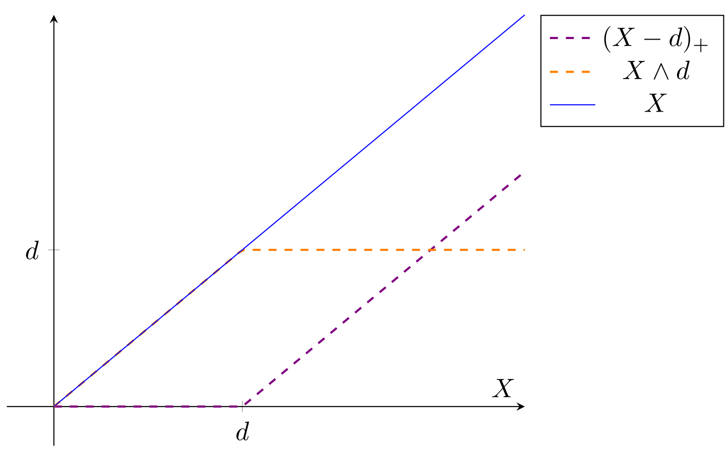
There is a cap of u dollars to the payment made by 🏢.

1.4.3 An important relationship between stop loss and limited loss variables is as follows. For any $d \in \mathbb{R}$ and any random variable X,

$$(X - d)_+ + (X \wedge d) = X.$$

Proof. We have

$$(X - d)_+ + (X \wedge d) = \begin{cases} X - d + d & \text{if } X > d; \\ 0 + X & \text{if } X \leq d \end{cases} = X.$$



□

Remarks:

- A practical interpretation of this result is that combining an insurance with deductible d and another insurance with policy limit d gives an insurance with full coverage.
- We can also show that $(d - X)_+ + (d \wedge X) = d$ using a similar proof. This provides a helpful way to find probabilistic quantities about the "swapped" stop-loss variable $(d - X)_+$.

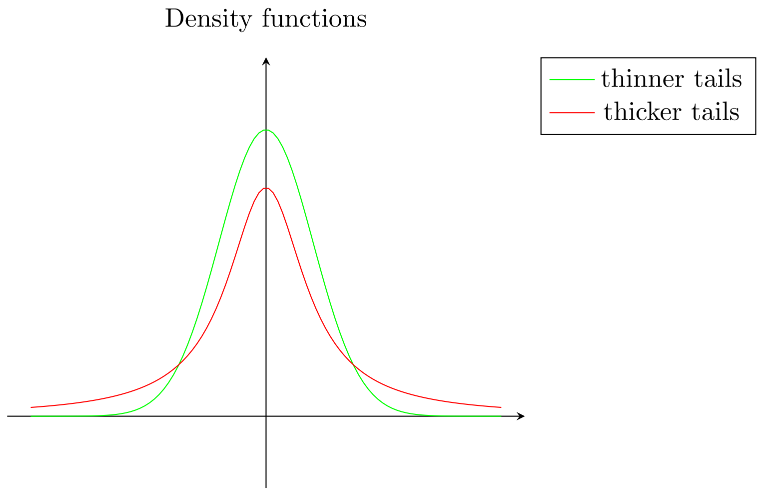
1.4.4 Due to the relationship in [1.4.3], we can obtain the following formula for computing expected limited loss, using Propositions 1.1.a and 1.2.a, assuming X has finite mean and is nonnegative:

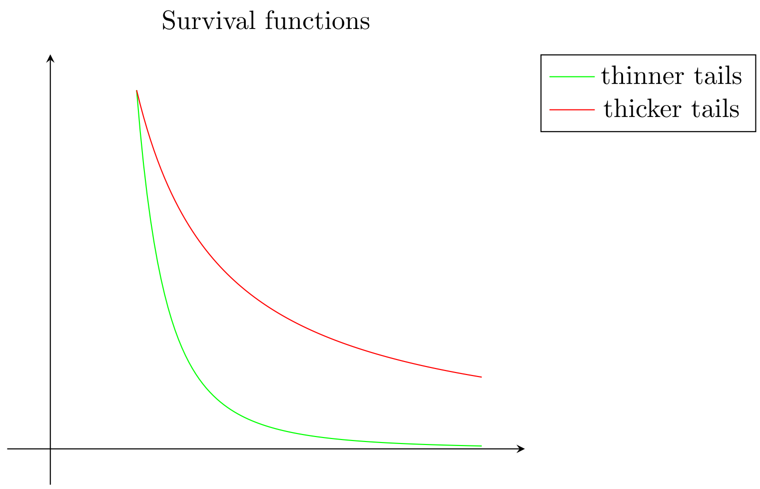
$$\mathbb{E}[X \wedge u] = \mathbb{E}[X] - \mathbb{E}[(X - u)_+] = \int_0^\infty S_X(x)\,\mathrm{d}x - \int_u^\infty S_x(x)\,\mathrm{d}x = \boxed{\int_0^u S_X(x)\,\mathrm{d}x}$$

where S_X is the survival function of X.

1.5 Comparing Tail Thickness of Distributions

1.5.1 An important consideration in risk management for an insurer 🏢 is to properly quantify the "thickness" of the (right) tail of distribution of loss X since it can greatly impact the financial position of 🏢. The higher probability assigned to extremely large values, the "thicker"/"heavier" the right tail.





[Note: One can view the notion of "tail thickness" from the perspective of (right) tail of density function or (right) tail of survival function:

- *tail of density function*: thicker tail ➔ more probability "assigned" to extremely positive values
- *tail of survival function*: thicker tail ➔ given a fixed extremely large value, higher probability for the random variable to exceed it

It turns out that these two approaches of comparing tail thickness are consistent (based on the "limit of ratio" method).]

1.5.2 To *compare* tail thickness, we can consider the following methods:

(a) comparison based on existence and non-existence of moments

(b) comparison based on limit of ratio of survival functions

(c) "comparison" based on hazard rate function ("force of mortality" in life contingencies)

(d) "comparison" based on MRL (or mean excess loss function)

Remarks:

- Here we focus on continuous random variables. But the methods can also be used for discrete random variables in a similar manner.
- More properly, the last two methods are indeed *classification* methods.

1.5.3 First we consider comparison based on moments. Here we focus on a nonnegative loss X. Recall that the kth raw moment of X with pdf f_X is

$$\mu_k' = \mathbb{E}\left[X^k\right] = \int_{-\infty}^{\infty} x^k f_X(x)\,\mathrm{d}x$$

If $f_X(x)$ is relatively large for large x, it indicates that X has a relatively thick right tail. Hence, a method for assessing tail thickness is to check the *speed* of $f_X(x) \to 0$ as x rises.

When the kth moment $\mathbb{E}\left[X^k\right]$ exists/is finite (i.e., $\mathbb{E}\left[X^k\right] < \infty$), it suggests that $f_X(x) \to 0$ *much faster* than the speed at which $x^k \to \infty$ ➔ thinner right tail.

On the other hand, if it is infinite (i.e., $\mathbb{E}\left[X^k\right] = \infty$), then it indicates that $f_X(x) \to 0$ *much slower* than the speed at which $x^k \to \infty$ ➔ thicker right tail.

1.5.4 Regarding the existence/non-existence of moments, we have the following result.

Proposition 1.5.a. Suppose that X is a nonnegative random variable. Fix any $r, k > 0$ with $0 < r < k$. Then,

$$\mathbb{E}\left[X^k\right] < \infty \implies \mathbb{E}\left[X^r\right] < \infty.$$

That is, existence of kth moment implies existence of all smaller positive moments.

[Note: Equivalently,

$$\mathbb{E}\left[X^r\right] = \infty \implies \mathbb{E}\left[X^k\right] = \infty.$$

That is, non-existence of rth moment implies non-existence of all larger positive moments.]

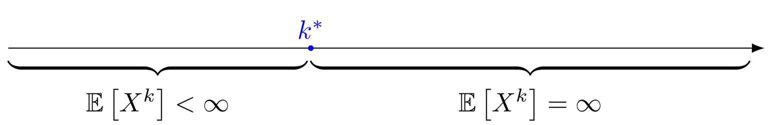
Proof. Assume that $\mathbb{E}\left[X^k\right] < \infty$. Then,

$$\begin{aligned}
\mathbb{E}\left[X^r\right] &= \mathbb{E}\left[X^r \mathbf{1}_{\{0 \le X \le 1\}}\right] + \mathbb{E}\left[X^r \mathbf{1}_{\{X > 1\}}\right] \\
&\le \mathbb{E}\left[\mathbf{1}_{\{0 \le X \le 1\}}\right] + \mathbb{E}\left[X^k \mathbf{1}_{\{X > 1\}}\right] && (X^r \le 1 \text{ when } 0 \le X \le 1, \text{ and } X^r \le X^k \text{ when } X > 1) \\
&\le \mathbb{E}\left[\mathbf{1}_{\{0 \le X \le 1\}}\right] + \mathbb{E}\left[X^k\right] && (\mathbf{1}_{\{X > 1\}} \le 1 \text{ and } X^k \ge 0) \\
&= \mathbb{P}\left(0 \le X \le 1\right) + \mathbb{E}\left[X^k\right] \\
&< \infty.
\end{aligned}$$

□

Hence, if it is not the case that X has finite kth moment for any $k > 0$, then we can find a "turning point" $k^* > 0$ at which the moment changes from finite to infinite, i.e.,

$$\mathbb{E}\left[X^k\right] < \infty \ \forall 0 < k < k^* \quad \text{and} \quad \mathbb{E}\left[X^k\right] = \infty \ \forall k \ge k^*.$$



1.5.5 Thus, we have the following indicators for tail thickness:

- Existence of kth moment for any $k > 0$ ➔ "rather thin" right tail.

- Between two nonnegative losses X and Y, if the "turning point" of X is smaller than that of Y, then X has a *thicker* right tail than Y. [Note: To see this, note that in this case we can find a $k > 0$ such that $\mathbb{E}\left[X^k\right] = \infty$ and $\mathbb{E}\left[Y^k\right] < \infty$, which means that $f_X(x) \to 0$ *much slower* than $x^k \to \infty$, while $f_Y(x) \to 0$ *much faster* than $x^k \to \infty$ ➔ $f_X(x) \to 0$ *much slower* than $f_Y(x) \to 0$ ➔ X has a *thicker* right tail than Y.]

1.5.6 Now we consider comparison based on limit of ratio of survival functions. To compare the tail thickness between X and Y, we consider the limit

$$\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)}$$

where S_X and S_Y are survival functions of X and Y respectively.

Suppose the limit is c (which may be ∞). Then, we can compare the tail thickness based on c:

- $c = 0$: $S_X(t) \to 0$ *much faster* than $S_Y(t) \to 0$ as $t \to \infty$ ➔ Y has a *thicker* right tail than X
- $0 < c < \infty$: $S_X(t) \to 0$ at a "similar" speed to $S_Y(t) \to 0$ as $t \to \infty$ ➔ X and Y have "similar" right tail thickness
- $c = \infty$: $S_X(t) \to 0$ *much slower* than $S_Y(t) \to 0$ as $t \to \infty$ ➔ X has a *thicker* right tail than Y

[Note: When $S_X(t)$ is relatively large for large t, it suggests that relatively high probability is assigned to large values ➔ relatively thick right tail.]

1.5.7 To compute the limit $\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)}$, we can apply the L'Hôpital's rule:

$$\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)} = \lim_{t\to\infty} \frac{S_X'(t)}{S_Y'(t)} = \boxed{\lim_{t\to\infty} \frac{f_X(t)}{f_Y(t)}}.$$

where f_X and f_Y are pdfs of X and Y respectively.

[Note: This suggests that for this method, comparing the right tails of survival functions is the same as comparing the right tails of density functions.]

1.5.8 Next, we consider the comparison based on hazard rate function. The **hazard rate function** of a random variable X is

$$h_X(x) = \frac{f_X(x)}{S_X(x)}$$

where f_X and S_X are pdf and survival function of X respectively.

[Note: It is "force of mortality" in life contingencies.]

To interpret the hazard rate, note that

$$\mathbb{P}\left(x < X \le x + \Delta x | X > x\right) = \frac{\mathbb{P}\left(x < X \le x + \Delta x\right)}{S_X(x)} \approx \frac{f(x)\Delta x}{S_X(x)} = h_X(x)\Delta x$$

for small Δx. Thus, in infinitesimal ($\Delta x \to \mathrm{d}x$), $h_X(x)\,\mathrm{d}x$ gives the probability that $X \in (x, x + \mathrm{d}x]$ ("failing instantaneously") given $X > x$ ("surviving"). Hence, the hazard rate can be regarded as *instantaneous failure rate*.

1.5.9 To assess the tail thickness based on hazard rate function, consider the following.

- $h_X(x)$ is a decreasing (or non-increasing) in x ➔ less likely to fail/"die" as the "lifetime" increases ➔ more likely to have a long "lifetime" ➔ X has a thick right tail.
- $h_X(x)$ is a increasing (or non-increasing) in x ➔ more likely to fail/"die" as the "lifetime" increases ➔ less likely to have a long "lifetime" ➔ X has a thin right tail.

1.5.10 A random variable X has **decreasing failure rate** (DFR) if $h_X(x)$ is decreasing in x, and has **increasing failure rate** (IFR) if $h_X(x)$ is increasing in x.

Hence, X is *classified* into a thick-tail (thin-tail) category if X has DFR (IFR).

1.5.11 Lastly, we consider the comparison based on mean residual lifetime (MRL).

- $e_X(d)$ is decreasing in d ➔ shorter "life expectancy" as the "lifetime" increases ➔ less likely to have a long "lifetime" ➔ X has a thin right tail.
- $e_X(d)$ is increasing in d ➔ longer "life expectancy" as the "lifetime" increases ➔ more likely to have a long "lifetime" ➔ X has a thick right tail.

1.5.12 A random variable X has **decreasing mean residual lifetime** (DMRL) if $e_X(d)$ is decreasing in d, and has **increasing mean residual lifetime** (IMRL) if $e_X(d)$ is increasing in d.

Hence, X is *classified* into a thick-tail (thin-tail) category if X has IMRL (DMRL).

1.5.13 Naturally, one would then be interested in the relationship between DFR/IFR and DMRL/IMRL. This is given as follows.

Proposition 1.5.b. Let X be a random variable. Then,

$$X \text{ has DFR} \implies X \text{ has IMRL} \quad \text{and} \quad X \text{ has IFR} \implies X \text{ has DMRL}.$$

Proof. Firstly, since

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = -\frac{\mathrm{d}}{\mathrm{d}x}\ln S_X(x),$$

we have

$$S_X(x) = \exp\left(-\int_{-\infty}^{x} h_X(y)\,\mathrm{d}y\right),$$

so

$$\frac{S_X(x+t)}{S_X(x)} = \exp\left(-\int_{x}^{x+t} h_X(y)\,\mathrm{d}y\right) = \exp\left(-\int_{0}^{t} h_X(x+y)\,\mathrm{d}y\right).$$

[Note: The formula corresponding to this in life contingencies is $S_x(t) = \exp\left(-\int_0^t \mu_{x+s}\,\mathrm{d}s\right)$.]

Now, assume X has DFR and fix any $t \geq 0$. Then, from the equation above we know $\dfrac{S_X(x+t)}{S_X(x)}$ is increasing in x. Thus, by [1.3.4], for any $d_1 \leq d_2$,

$$e_X(d_1) = \frac{\int_0^\infty S_X(d_1+t)\,\mathrm{d}t}{S_X(d_1)} = \int_0^\infty \frac{S_X(d_1+t)}{S_X(d_1)}\,\mathrm{d}t \leq \int_0^\infty \frac{S_X(d_2+t)}{S_X(d_2)}\,\mathrm{d}t = e_X(d_2),$$

meaning that $e_X(d)$ is increasing in d, so X is IMRL.

Proof of another implication is similar. □

[⚠ Warning: The converse implications are <u>not</u> true. That is,

$$X \text{ has IMRL} \not\Longrightarrow X \text{ has DFR} \quad \text{and} \quad X \text{ has DMRL} \not\Longrightarrow X \text{ has IFR}.$$

]

2 Mixing and Conditional Expectation

2.1 Mixing

2.1.1 Consider n random variables $X_1, \dots, X_n$ which are assumed to be all continuous or all discrete. We can create a new distribution by *mixing* them, whose probability function is

$$f_X(x) = p_1 f_{X_1}(x) + \cdots + p_n f_{X_n}(x)$$

where f_{X_i} is the probability function of X_i for each $i = 1, \dots, n$, $p_1, \dots, p_n \geq 0$, and $p_1 + \cdots + p_n = 1$. The resulting probability function is a weighted average of the probability functions of the n random variables.

2.1.2 Another important point of view on mixing is through *conditioning*. By introducing another discrete random variable Λ with support $\{\lambda_1, \dots, \lambda_n\}$ such that

$$X_i \stackrel{d}{=} (X|\Lambda = \lambda_i) \quad \text{for any } i = 1, \dots, n,$$

we can associate $f_{X_i}(x)$ with $f_{X|\Lambda}(x|\lambda_i)$ and p_i with $\mathbb{P}(\Lambda = \lambda_i)$:

$$f_X(x) = \sum_{i=1}^{n} f_{X|\Lambda}(x|\lambda_i)\mathbb{P}(\Lambda = \lambda_i) = \boxed{\sum_{i=1}^{n} f_{X|\Lambda}(x|\lambda_i)p_i}.$$

[Note: $f_{X|\Lambda}(x|\lambda_i)$ is the conditional probability function of X given $\Lambda = \lambda_i$.]

The equality holds since each summand is the joint probability function $f_{X,\Lambda}(x, \lambda_i)$, so summing them up gives the marginal probability function $f_{X_i}(x)$. This induces mixing of finitely many (all continuous/all discrete) random variables using (continuous or discrete) X and discrete Λ.[2]

2.1.3 Practically, this kind of mixing can arise when we try to model loss X for a randomly selected policyholder from a pool.

For example, suppose that the pool of all policyholders can be classified into several groups, say (i) good policyholder and (ii) bad policyholder. Then, in this context, Λ can represent the group indicator for the randomly selected policyholder.

For instance, we can let $\Lambda = 1$ for the good group and $\Lambda = 0$ for the bad group. [Note: The values assigned are not important.] Then Λ is a discrete random variable taking values 0 or 1. The probability that $\Lambda = 1$ ($\Lambda = 0$) can be regarded as the "proportion" of good (bad) policyholders in the pool.

Then, we would like to model the losses for a good policyholder and a bad policyholder using different distributions. This can be done by choosing different distributions for $X|\Lambda = 1$ (distribution for good policyholder) and $X|\Lambda = 0$ (distribution for bad policyholder).

2.1.4 To induce mixing of countably infinitely many random variables, we can introduce a discrete random variable Λ with support $\{\lambda_1, \lambda_2, \dots\}$. Then, the pdf/pmf of X can similarly be written as

$$f_X(x) = \boxed{\sum_{i=1}^{\infty} f_{X|\Lambda}(x|\lambda_i)p_i}$$

where $p_i = \mathbb{P}(\Lambda = \lambda_i)$ for any $i = 1, 2, \dots$.

2.1.5 To induce mixing of uncountably infinitely many random variables, we can introduce a continuous random variable Λ. Then, the pdf/pmf of X can be written as

$$f_X(x) = \boxed{\int_{-\infty}^{\infty} f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda}.$$

[2] Here the notion of conditional distribution includes also the case where one of the two random variables involved is discrete while another is continuous. This may not be defined in elementary probability theory, but this can be allowed through more advanced probability theory. So one may take the results for this kind of case here as given.

Here the integrand is the joint probability function $f_{X,\Lambda}(x,\lambda)$, so integrating it gives the marginal probability function $f_X(x)$.

2.2 Conditional Expectation

2.2.1 This sections serves as a review on the concept of conditional expectation, which is covered in STAT2901.

2.2.2 Consider again the random variables X and Λ. Recall that the conditional expectation of $g(X)$ given $\Lambda = \lambda$ is

$$\mathbb{E}\left[g(X)|\Lambda=\lambda\right] = \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x) f_{X|\Lambda}(x|\lambda)\,\mathrm{d}x & \text{if } X \text{ and } \Lambda \text{ are both continuous;} \\ \displaystyle\sum_i g(x_i) f_{X|\Lambda}(x_i|\lambda) & \text{if } X \text{ and } \Lambda \text{ are both discrete} \end{cases}$$

where the sum is taken over all i where $f_{X|\Lambda}(x_i|\lambda) > 0$.

Write $h(\lambda) = \mathbb{E}\left[g(X)|\Lambda=\lambda\right]$. Then, the conditional expectation of $g(X)$ given Λ is denoted by $\mathbb{E}\left[g(X)|\Lambda\right]$ and is a *random variable* (as a function of Λ):

$$h(\Lambda) = \mathbb{E}\left[g(X)|\Lambda\right]$$

which takes the value $h(\lambda)$ when $\Lambda = \lambda$.

[Note: Practically, to obtain an expression of $\mathbb{E}\left[g(X)|\Lambda\right]$, we can first find an expression for $\mathbb{E}\left[g(X)|\Lambda=\lambda\right]$ and replace every λ by Λ.]

2.2.3 Next, recall that the conditional variance of $g(X)$ given $\Lambda = \lambda$ is

$$\mathrm{Var}\left(g(X)|\Lambda=\lambda\right) = \mathbb{E}\left[\left[g(X) - \left(\mathbb{E}\left[g(X)|\Lambda=\lambda\right]\right)\right]^2|\Lambda=\lambda\right] = \mathbb{E}\left[g(X)^2|\Lambda=\lambda\right] - \left(\mathbb{E}\left[g(X)|\Lambda=\lambda\right]\right)^2.$$

[Note: Similarly, we can write $h(\lambda) = \mathrm{Var}\left(g(X)|\Lambda=\lambda\right)$, and $\mathrm{Var}\left(g(X)|\Lambda\right)$ is the random variable $h(\Lambda)$. Thus,

$$\mathrm{Var}\left(g(X)|\Lambda\right) = \mathbb{E}\left[(g(X))^2|\Lambda\right] - \left(\mathbb{E}\left[g(X)|\Lambda\right]\right)^2.$$

]

2.2.4 Two remarkable results related to conditional expectation and conditional variance are *law of total expectation* and *law of total variance*.

2.2.5 Law of total expectation relates unconditional and conditional expectations.

Theorem 2.2.a (Law of total expectation)**.** For any function g and any random variables X and Λ where $\mathbb{E}\left[g(X)\right]$ is finite,

$$\mathbb{E}\left[\mathbb{E}\left[g(X)|\Lambda\right]\right] = \mathbb{E}\left[g(X)\right].$$

Proof. We only prove for the case where X and Λ are both continuous. (The case where both are discrete can be proved similarly.[3]) Let $h(\lambda) = \mathbb{E}[g(X)|\Lambda = \lambda]$. Consider:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} g(x) \int_{-\infty}^{\infty} f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} f_\Lambda(\lambda) \int_{-\infty}^{\infty} g(x) f_{X|\Lambda}(x|\lambda)\,\mathrm{d}x\,\mathrm{d}\lambda \\
&= \int_{-\infty}^{\infty} h(\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda \\
&= \mathbb{E}[h(\Lambda)] \\
&= \mathbb{E}[\mathbb{E}[g(X)|\Lambda]].
\end{aligned}$$

□

2.2.6 An interesting application of law of total expectation is to derive the following results regarding the expression of cdf in mixing.

Proposition 2.2.b. Let X and Λ be random variables. Denote the cdf of X by F_X and the conditional cdf of X given $\Lambda = \lambda$ by $F_{X|\Lambda}(x|\lambda)$. Then,

- (Λ is continuous) $F_X(x) = \displaystyle\int_{-\infty}^{\infty} F_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda$.
- (Λ is discrete) $F_X(x) = \displaystyle\sum_i F_{X|\Lambda}(x|\lambda_i) p_i$ where $p_i = \mathbb{P}(\Lambda = \lambda_i)$.

Proof. The key idea is to recognize that $F_X(x) = \mathbb{E}\left[\mathbf{1}_{\{X \le x\}}\right]$ and apply the law of total expectation on this expectation, i.e., write $F_X(x) = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{X \le x\}}|\Lambda\right]\right]$. Depending on whether Λ is continuous or discrete, the evaluation of this double expectation would differ. Here we only prove the continuous case and the discrete case can be proven similarly.

Firstly, note that $h(\lambda) = \mathbb{E}\left[\mathbf{1}_{\{X \le x\}}|\Lambda = \lambda\right] = F_{X|\Lambda}(x|\lambda)$. Thus,

$$F_X(x) = \mathbb{E}[h(\Lambda)] = \int_{-\infty}^{\infty} h(\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda = \int_{-\infty}^{\infty} F_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\,\mathrm{d}\lambda.$$

□

2.2.7 Law of total variance relates unconditional variance and conditional mean & variance.

Proposition 2.2.c (Law of total variance)**.** For any random variable X with finite variance and random variable Λ,

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|\Lambda)] + \mathrm{Var}(\mathbb{E}[X|\Lambda]).$$

Proof. Note that

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}[\mathbb{E}\left[X^2|\Lambda\right]] - (\mathbb{E}[\mathbb{E}[X|\Lambda]])^2 \\
&= \mathbb{E}[\mathrm{Var}(X|\Lambda) + (\mathbb{E}[X|\Lambda])^2] - (\mathbb{E}[\mathbb{E}[X|\Lambda]])^2 \\
&= \mathbb{E}[\mathrm{Var}(X|\Lambda)] + \mathbb{E}\left[(\mathbb{E}[X|\Lambda])^2\right] - (\mathbb{E}[\mathbb{E}[X|\Lambda]])^2 \\
&= \mathbb{E}[\mathrm{Var}(X|\Lambda)] + \mathrm{Var}(\mathbb{E}[X|\Lambda]).
\end{aligned}$$

□

[3] A general proof where both random variables can be arbitrary (as long as $\mathbb{E}[g(X)]$ is finite) is out of scope.

3 Basic Frequency Models

3.1 Preliminaries

3.1.1 Suppose that an insurer sells a type of insurance to policyholders which provides some payment on each *claim* filed by policyholders. Then, the number of claims filed would be of interest to the insurer, and here several basic models of such number of claims will be introduced.

3.1.2 Let N be the random variable representing the number of *claims* filed in a certain time period (*claim frequency*). Since the number of claims is an nonnegative integer, N is a nonnegative discrete random variable.

3.1.3 Let $p_k = \mathbb{P}(N = k)$ be the pmf of N. Then, the **probability generating function** (pgf) of N is

$$P_N(t) = \mathbb{E}\left[t^N\right] = \sum_{k=0}^{\infty} t^k p_k.$$

Remarks:

- Note that $P_N(1) = \mathbb{E}[1] = 1$ always.
- In this context, we set $0^0 = 1$. So, $P_N(0) = p_0$.
- Probability generating function is only defined for nonnegative discrete random variable.

3.1.4 The following result suggests the "probability generating" property of pgf:

Proposition 3.1.a. Let P_N be the pgf of a nonnegative discrete random variable N. Then, the pmf of N is given by

$$p_m = \frac{P_N^{(m)}(0)}{m!}$$

for any $m = 0, 1, \ldots$, where $P_N^{(m)}$ denote the mth derivative of P_N with $P_N^{(0)} = P_N$.

Proof. Fix any $m = 0, 1, \ldots$. Firstly, if $m = 0$, we have $p_0 = P_N(0) = P_N(0)/0!$. So, henceforth consider the case where $m = 1, 2, \ldots$. Now, note that for any $k = 0, 1, \ldots$,

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m} t^k p_k = \begin{cases} k(k-1)\cdots(k-m+1)t^{k-m} & \text{if } k \geq m; \\ 0 & \text{if } k < m. \end{cases}$$

Thus,

$$P_N^{(m)}(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} \sum_{k=0}^{\infty} t^k p_k = \sum_{k=0}^{\infty} \frac{\mathrm{d}^m}{\mathrm{d}t^m} t^k p_k = \sum_{k=m}^{\infty} [k(k-1)\cdots(k-m+1)t^{k-m} p_k].$$

Now, since $0^0 = 1$, we have

$$P_N^{(m)}(0) = m(m-1)\cdots(m-m+1)(1)p_m = m!p_m \implies p_m = \frac{P_N^{(m)}(0)}{m!}.$$

□

3.1.5 As a corollary of Proposition 3.1.a, probability generating function gives a sufficient condition for equality in distribution:

Corollary 3.1.b. Let M and N be two nonnegative discrete random variables with pgf P_M and P_N. If they have the same pgf, then they have the same distribution.

Proof. By assumption, we have $P_M^{(m)}(0) = P_N^{(m)}(0)$ for any $m = 0, 1, \ldots$. Thus, by Proposition 3.1.a, for any $m = 0, 1, \ldots$,

$$\mathbb{P}(M = m) = \mathbb{P}(N = m),$$

which means that M and N have the same distribution. □

3.1.6 Probability generating function also has a "moment generating" property as follows.

Proposition 3.1.c. Let P_N be the pgf of a nonnegative discrete random variable N. Then,

$$P'_N(1) = \mathbb{E}[N] \quad \text{and} \quad P''_N(1) = \mathbb{E}[N(N-1)].$$

Proof. Note first that

$$P'_N(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[t^N\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}t^N\right] = \mathbb{E}\left[Nt^{N-1}\right]$$

and

$$P''_N(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}\left[t^N\right] = \mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}t^N\right] = \mathbb{E}\left[N(N-1)t^{N-2}\right].$$

Then, putting $t = 1$ gives the desired result. □

3.1.7 The pgf of a sum of independent nonnegative discrete random variables can be obtained by the following formula.

Theorem 3.1.d. Let $N_1, \ldots, N_m$ be independent nonnegative discrete random variables, and let $S = N_1 + \cdots + N_m$. Then, the pgf of S is given by

$$P_S(t) = P_{N_1}(t) \cdots P_{N_m}(t)$$

where P_N denotes the pgf of N.

Proof. First note that $t^{N_1}, \ldots, t^{N_m}$ are also independent. Thus,

$$P_S(t) = \mathbb{E}\left[t^S\right] = \mathbb{E}\left[t^{N_1+\cdots+N_m}\right] = \mathbb{E}\left[t^{N_1}\cdots t^{N_m}\right] = \mathbb{E}\left[t^{N_1}\right]\cdots\mathbb{E}\left[t^{N_m}\right] = P_{N_1}(t)\cdots P_{N_m}(t).$$

□

Theorem 3.1.d is often useful for deducing the distribution of a random variable given a complex pgf P. By *factorizing* the pgf P into products of several more well-known pgf's (to be introduced in the following sections), we can deduce the underlying distribution. This is often the key 🔑 to analyze a seeming complex pgf.

3.1.8 Starting from here, we will discuss several kinds of probability distributions for modelling the number of claims N (*frequency models*):

(a) Poisson distribution
(b) mixed Poisson distribution
(c) negative binomial distribution
(d) geometric distribution
(e) binomial distribution

3.2 The Poisson Distribution

3.2.1 A random variable N follows the **Poisson distribution** with parameter $\lambda > 0$ (denoted by $N \sim \mathrm{Poi}(\lambda)$) if its pmf is

$$p_k = \mathbb{P}(N = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for any $k = 0, 1, \ldots$.

3.2.2 The pgf of $N \sim \mathrm{Poi}(\lambda)$ is

$$P_N(t) = \mathbb{E}\left[t^N\right] = \boxed{e^{\lambda(t-1)}}.$$

Proof. Note that

$$P_N(t) = \sum_{k=0}^{\infty} t^k p_k = \sum_{k=0}^{\infty} t^k \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!} = e^{-\lambda}e^{\lambda t}.$$

(Recall that $e^x = \sum_{k=0}^{\infty} x^k/k!$.) □

3.2.3 We can then deduce its mean and variance based on pgf:

$$\mathbb{E}[N] = \mathrm{Var}(N) = \boxed{\lambda}.$$

Proof. Since $P_N(t) = e^{\lambda(t-1)}$, we have

$$P_N'(t) = \lambda e^{\lambda(t-1)} \quad \text{and} \quad P_N''(t) = \lambda^2 e^{-\lambda(t-1)}.$$

Therefore, by Proposition 3.1.c,

$$\mathbb{E}[N] = P_N'(1) = \lambda \quad \text{and} \quad \mathbb{E}[N(N-1)] = P_N''(1) = \lambda^2.$$

Since $\mathbb{E}[N(N-1)] = \mathbb{E}\left[N^2\right] - \mathbb{E}[N] = \mathbb{E}\left[N^2\right] - \lambda$, it follows that

$$\mathrm{Var}(N) = \mathbb{E}\left[N^2\right] - (\mathbb{E}[N])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

3.2.4 Here we introduce two remarkable results for a Poisson distribution: *convolution* and *thinning* (or *decomposition*).

3.2.5 The convolution result concerns *sum* of independent Poisson random variables. It turns out that such sum is *also* Poisson distributed.

Theorem 3.2.a. Let $N_1, \ldots, N_k$ be k independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_k$ respectively. Then, the sum $S = N_1 + \cdots + N_k$ follows the Poisson distribution with parameter $\lambda_1 + \cdots + \lambda_k$.

Proof. We prove this using pgf. We denote pgf of N by $P_N(t)$. By Theorem 3.1.d, we have

$$P_S(t) = P_{N_1}(t) \cdots P_{N_k}(t).$$

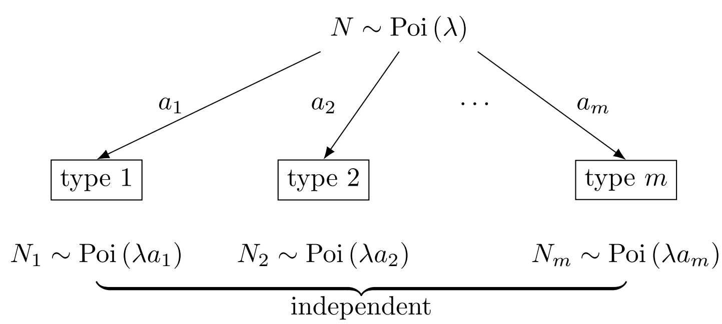
Now using [3.2.2], we can further write

$$P_S(t) = e^{\lambda_1(t-1)} \cdots e^{\lambda_k(t-1)} = e^{(\lambda_1 + \cdots + \lambda_k)(t-1)}.$$

But note that this is exactly the same as the pgf of (a random variable following) $\mathrm{Poi}(\lambda_1 + \cdots + \lambda_k)$ distribution. Hence, by Corollary 3.1.b, we conclude that $S \sim \mathrm{Poi}(\lambda_1 \cdots + \lambda_k)$. □

3.2.6 The thinning/decomposition result concerns the "finer pieces" obtained from "slicing"/"decomposing" a Poisson random variable ("thinning"). It turns out that by performing the "slicing" in a certain way, the resulting "finer pieces" of random variables are also Poisson distributed.

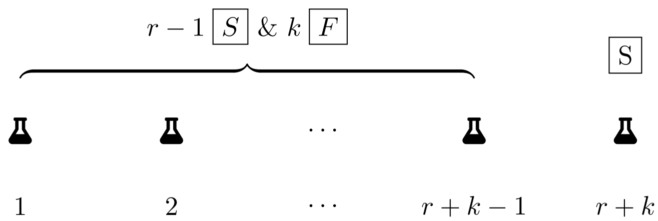
Theorem 3.2.b. Let $N \sim \mathrm{Poi}(\lambda)$ be the number of claims. Suppose that each claim is independently classified into exactly one of type $1, \ldots, m$ with probabilities $a_1, \ldots, a_m$ respectively (where $a_1 + \cdots + a_m = 1$). Let $N_1, \ldots, N_m$ be the numbers of claims with types $1, \ldots, m$ respectively. Then, $N_1, \ldots, N_m$ are independent Poisson random variables with parameters $\lambda a_1, \ldots, \lambda a_m$ respectively.



3.3 The Negative Binomial Distribution

3.3.1 In elementary probability course, a negative binomial random variable $N \sim \mathrm{NB}(r, p)$ is defined as the number of failures before rth success in a sequence of independent Bernoulli trials[4] with success probability $p \in (0, 1)$ (neither 0 nor 1 so that there is randomness). Under this definition, the pmf of N is

$$p_k = \mathbb{P}(N = k) = \binom{r+k-1}{k} p^r (1-p)^k.$$



The conditions on the parameters r and p are $r = 1, 2, \ldots$ and $0 < p < 1$.

3.3.2 Here we consider a more general definition of negative binomial distribution. A random variable N follows the **negative binomial distribution** with parameters $r > 0$ and $\beta > 0$ (denoted by $N \sim \mathrm{NB}(r, \beta)$) if its pmf is given by

$$p_k = \mathbb{P}(N = k) = \binom{r+k-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k \tag{1}$$

for any $k = 0, 1, \ldots$.

<u>Remarks</u>:

- In this notes, when we use the notation $\mathrm{NB}(\cdot, \cdot)$, it carries the meaning of $\mathrm{NB}(r, \beta)$ here instead of $\mathrm{NB}(r, p)$ from the definition in [3.3.1].
- To move from the definition in [3.3.1] to the definition here, we allow r to be any positive real number and reparametrize the success probability $p \in (0, 1)$ by $\dfrac{1}{1+\beta}$ where $\beta > 0$.
- Here the binomial coefficient $\dbinom{x}{k}$ has the *general* definition (which permits x to be any real number and k be any nonnegative integer):

$$\binom{x}{k} = \begin{cases} \dfrac{\overbrace{x(x-1)(x-2)\cdots(x-k+1)}^{k \text{ terms}}}{k!} & \text{if } k \neq 0; \\ 1 & \text{if } k = 0. \end{cases}$$

3.3.3 The pgf of $N \sim \mathrm{NB}(r, \beta)$ is given by

$$P_N(t) = \boxed{[1 - \beta(t-1)]^{-r}}.$$

[4] i.e., experiments with two possible outcomes: "success" and "failures"

Proof. Note that

$$
\begin{aligned}
P_N(t) &= \mathbb{E}\left[t^N\right] \\
&= \sum_{k=0}^{\infty} t^k \binom{k+r-1}{k} \left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)^r \\
&= \left(\frac{1}{1+\beta}\right)^r \sum_{k=0}^{\infty} \binom{k+r-1}{k} \left(\frac{t\beta}{1+\beta}\right)^k \\
&= \left(\frac{1}{1+\beta}\right)^r \left(1-\frac{t\beta}{1+\beta}\right)^{-r} && \text{(negative binomial series formula)} \\
&= \left(\frac{1}{1+\beta}\right)^r \left(\frac{1+\beta}{1-(t-1)\beta}\right)^r \\
&= [1-\beta(t-1)]^{-r}.
\end{aligned}
$$

□

3.3.4 Based on the pgf, we know

- $P_N'(t) = r\beta[1-\beta(t-1)]^{-r-1}$.
- $P_N''(t) = r(r+1)\beta^2[1-\beta(t-1)]^{-r-2}$.

We can thus obtain the mean and variance of $N \sim \text{NB}(r, \beta)$ as follows.

- $\mathbb{E}[N] = P_N'(1) = \boxed{r\beta}$.
- $\mathbb{E}[N(N-1)] = P_N''(1) = r(r+1)\beta^2 \implies \text{Var}(N) = \mathbb{E}\left[N^2\right] - (\mathbb{E}[N])^2 = r(r+1)\beta^2 + r\beta - (r\beta)^2 = \boxed{r\beta(1+\beta)}$.

3.4 The Geometric Distribution

3.4.1 The geometric distribution is simply the negative binomial distribution with $r = 1$. More explicitly, a random variable N follows the **geometric distribution** with parameter $\beta > 0$ (denoted by $N \sim \text{Geom}(\beta)$) if its pmf is given by

$$p_k = \mathbb{P}(N = k) = \left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)$$

for any $k = 0, 1, \ldots$, i.e., $\text{Geom}(\beta) \equiv \text{NB}(1, \beta)$.

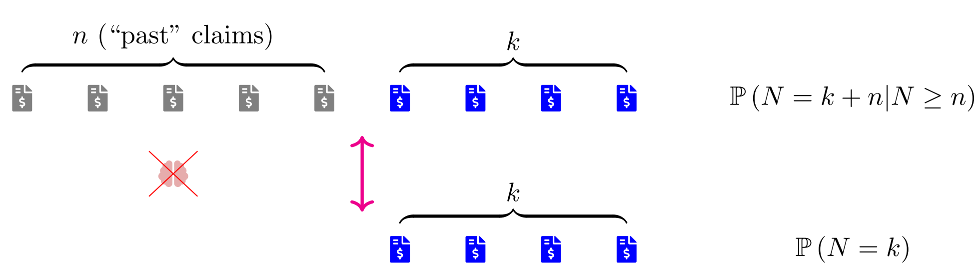
[Note: As a special case of negative binomial distribution, the random variable $N \sim \text{Geom}(\beta)$ can be interpreted as the number of failures in a sequence of independent Bernoulli trials before the *first* success, with success probability being $1/(1+\beta)$.]

3.4.2 A remarkable property of geometric distribution is the *memoryless property*.

Proposition 3.4.a. If $N \sim \text{Geom}(\beta)$, then

$$\mathbb{P}(N = k+n | N \geq n) = \mathbb{P}(N = k)$$

for any $k, n = 0, 1, \ldots$.



Proof. First denote the success probability by $p = 1/(1+\beta)$. Then, $p_k = (1-p)^k p$. Thus,

$$\begin{aligned}
\mathbb{P}(N = k+n | N \geq n) &= \frac{\mathbb{P}(N = k+n \cap N \geq n)}{\mathbb{P}(N \geq n)} \\
&= \frac{\mathbb{P}(N = k+n)}{\mathbb{P}(N \geq n)} \\
&= \frac{(1-p)^{k+n}\cancel{p}}{\sum_{i=n}^{\infty}(1-p)^i\cancel{p}} \\
&= \frac{(1-p)^{k+n}}{(1-p)^n/(1-(1-p))} \\
&= (1-p)^k p \\
&= \mathbb{P}(N = k).
\end{aligned}$$

□

3.5 The Binomial Distribution

3.5.1 Consider m independent Bernoulli trials with the same success probability $q \in (0,1)$. Then, in the m trials, the number of successes N follows the **binomial distribution** with parameters m and q, denoted by $N \sim \text{Bin}(m,q)$. The pmf of $N \sim \text{Bin}(m,q)$ is given by

$$p_k = \mathbb{P}(N = k) = \binom{m}{k} q^k (1-q)^{m-k}$$

for any $k = 0, 1, \ldots, m$.

Remarks:

- $N \sim \text{Bin}(m,q)$ can be more practically interpreted as the number of claims made when there are m moments in a certain time period at which a claim can be filed, independently with the same claim probability q.
- For life insurance, a claim is made when the insured dies, where the letter q is used in the notation of death/claim probability. So, to "match" with this case, we use q to denote the "success"/claim probability here.

3.5.2 Since $N \sim \text{Bin}(m,q)$ has the same distribution as a sum of m independent Bernoulli random variables with $\text{Ber}(q) \equiv \text{Bin}(1,q)$ distribution, and the pgf of a random variable with $\text{Ber}(q)$ distribution is $t^0(1-q) + t^1 q = 1 + q(t-1)$, it follows that the pgf of $N \sim \text{Bin}(m,q)$ is

$$P_N(t) = \boxed{[1+q(t-1)]^m}$$

by Theorem 3.1.d.

3.5.3 Based on the pgf, we know

- $P_N'(t) = mq[1+q(t-1)]^{m-1}$.
- $P_N''(t) = m(m-1)q^2[1+q(t-1)]^{m-2}$.

Thus, we can obtain the mean and variance of $N \sim \text{Bin}(m,q)$ as follows.

- $\mathbb{E}[N] = P_N'(1) = \boxed{mq}$.
- $\mathbb{E}[N(N-1)] = P_N''(1) = m(m-1)q^2 \implies \text{Var}(N) = \mathbb{E}[N^2] - (\mathbb{E}[N])^2 = m(m-1)q^2 + mq - (mq)^2 = \boxed{mq(1-q)}$.

4 The $(a, b, 0)$ and $(a, b, 1)$ Classes

4.0.1 In this section, we discuss two *classes* of probability distributions: $(a, b, 0)$ and $(a, b, 1)$ classes, which contain many useful distributions for modelling claim frequency N.

4.1 The $(a, b, 0)$ Class

4.1.1 Let N be a nonnegative discrete random variable, and let $p_k = \mathbb{P}(N = k)$ for any $k = 0, 1, \ldots$. Then, the distribution of N is in the $(a, b, 0)$ **class** if there exist constants a and b such that

$$p_k = \left(a + \frac{b}{k}\right)p_{k-1}, \quad \text{for any } k = 1, 2, \ldots.$$

[Note: The relationship above starts with the pair p_0 and p_1, so the class is called $(a, b, 0)$.]

4.1.2 For a random variable N in the $(a, b, 0)$ class, once a and b are fixed, the probabilities $p_1, p_2, \ldots$ can all be deduced from p_0. Furthermore, we can deduce the value of p_0 by the fact that $p_0 + p_1 + \cdots = 1$. Thus, all the probabilities $p_0, p_1, \ldots$ are fixed after a and b are fixed.

Hence, the values of a and b can be used to uniquely characterize a distribution in the $(a, b, 0)$ class.

4.1.3 An important result regarding the $(a, b, 0)$ class is as follows.

Theorem 4.1.a. Poisson, negative binomial, and binomial distributions are the *only* distributions in the $(a, b, 0)$ class.

Proof. Omitted. (See, e.g., Sundt and Jewell (1981).) □

4.1.4 Although here we do not show Poisson, negative binomial, and binomial distributions are the *only* distributions in the $(a, b, 0)$ class, we will show that they *are* distributions in the $(a, b, 0)$ class in the following.

4.1.5 For $N \sim \text{Poi}(\lambda)$, for any $k = 1, 2, \ldots$,

$$\frac{p_k}{p_{k-1}} = \frac{e^{-\lambda}\lambda^k/k!}{e^{-\lambda}\lambda^{k-1}/(k-1)!} = 0 + \frac{\lambda}{k}.$$

Hence, $N \sim \text{Poi}(\lambda)$ is in the $(a, b, 0)$ class with $a = 0$ and $b = \lambda$.

4.1.6 For $N \sim \text{NB}(r, \beta)$, for any $k = 1, 2, \ldots$,

$$\begin{aligned}
\frac{p_k}{p_{k-1}} &= \frac{\binom{r+k-1}{k}\left(\frac{1}{1+\beta}\right)\left(\frac{\beta}{1+\beta}\right)^k}{\binom{r+k-1-1}{k-1}\left(\frac{1}{1+\beta}\right)\left(\frac{\beta}{1+\beta}\right)^{k-1}} \\
&= \frac{\beta}{1+\beta}\,\frac{(r+k-1)(r+k-2)(r+k-3)\cdots r/k!}{(r+k-2)(r+k-3)\cdots r/(k-1)!} \\
&= \frac{\beta}{1+\beta}\cdot\underbrace{\frac{r+k-1}{k}}_{1+\frac{r-1}{k}} \\
&= \frac{\beta}{1+\beta} + \frac{(r-1)\frac{\beta}{1+\beta}}{k}.
\end{aligned}$$

Hence, $N \sim \text{NB}(r, \beta)$ is in the $(a, b, 0)$ class with $a = \dfrac{\beta}{1+\beta} > 0$ and $b = (r-1)\dfrac{\beta}{1+\beta}$.

4.1.7 For $N \sim \text{Bin}(m, q)$, for any $k = 1, 2, \dots, n$,

$$\begin{aligned}\frac{p_k}{p_{k-1}} &= \frac{\binom{m}{k} q^k (1-q)^{n-k}}{\binom{m}{k-1} q^{k-1}(1-q)^{n-k+1}} \\ &= \frac{q}{1-q} \cdot \frac{m(m-1)\cdots(m-k+2)(m-k+1)/k!}{m(m-1)\cdots(m-k+2)/(k-1)!} \\ &= \frac{q}{1-q} \cdot \frac{m-k+1}{k} \\ &= -\frac{q}{1-q} + \frac{(m+1)\frac{q}{1-q}}{k}.\end{aligned}$$

Hence, $N \sim \text{Bin}(m, q)$ is in the $(a, b, 0)$ class with $a = -\frac{q}{1-q} < 0$ and $b = (m+1)\frac{q}{1-q}$.

4.1.8 By Theorem 4.1.a, [4.1.5] to [4.1.7] include all possibilities for distributions in the $(a, b, 0)$ class. Notably, if a distribution in the $(a, b, 0)$ class has a *positive* (*negative*) a, it must be *negative binomial* (*binomial*) distributed. Of course, if $a = 0$, it must be Poisson distributed. [⚠ Warning: We do not have $a < 0 \implies$ negative binomial distributed!]

4.1.9 We can rewrite the equation in the definition of $(a, b, 0)$ class to

$$k \cdot \frac{p_k}{p_{k-1}} = ka + b \quad \text{for any } k = 1, 2, \dots,$$

so $k \cdot \frac{p_k}{p_{k-1}}$ is linearly related to k.

Practically, this is helpful for having a quick check on whether the actual claim frequency in practice is in the $(a, b, 0)$ class as follows. First note the following approximated relationship:

$$k \cdot \frac{\widehat{p}_k}{\widehat{p}_{k-1}} \approx ka + b$$

where $\widehat{p}_k = \frac{\text{no. of policies with } k \text{ claims}}{\text{total no. of policies considered}}$ for any $k = 0, 1, \dots$.

Then, by plotting $k \cdot \frac{\widehat{p}_k}{\widehat{p}_{k-1}}$ against k (or other means), we can examine whether a linear relationship is plausible. If such linear relationship is deemed plausible, we can then use some line fitting technique to obtain estimated value of a and b.

4.2 The $(a, b, 1)$ Class

4.2.1 One main issue of using distribution in the $(a, b, 0)$ class (Poisson/negative binomial/binomial) to model the number of claims N is that the probability $p_0 = \mathbb{P}(N = 0)$ is usually unreasonably low. In practice, especially when the insurance covers some "rare" loss, the probability p_0 is often quite large.

4.2.2 To deal with this issue, a simple way is to add a flexibility on the choice of p_0 on distribution in the $(a, b, 0)$ class, which gives rise to the $(a, b, 1)$ *class*.

4.2.3 Let N be a nonnegative discrete random variable with pmf $p_k = \mathbb{P}(N = k)$. Then, (the distribution of) N is in the $(a, b, 1)$ **class** if there exist constants a and b such that

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1} \quad \text{for any } k = 2, 3, \dots,$$

while p_0 can take any value in $[0, 1]$.

[Note: The relationship above starts with the pair p_1 and p_2, so the class is called $(a, b, 1)$.]

4.2.4 Like the $(a, b, 0)$ class, all the probabilities $p_0, p_1, \dots$ are fixed after a, b, and p_0 are fixed. Hence, the values of a, b, and p_0 can be used to uniquely characterize a distribution in the $(a, b, 1)$ class.

4.3 The Zero-Modified and Zero-Truncated $(a, b, 1)$ Classes

4.3.1 One systematic method for creating a distribution in the $(a, b, 1)$ class is **zero-modification**:

(1) Pick any distribution of a nonnegative discrete random variable (usually in the $(a, b, 0)$ class).
(2) Modify p_0 to an arbitrary number $p_0^M \in [0, 1]$.
(3) Modify the probabilities p_1, p_2, ... to $p_1^M = cp_1$, $p_2^M = cp_2$,... respectively, for some constant c, such that the resulting probabilities $p_0^M, p_1^M, \dots$ form a valid distribution.

Then, $p_0^M, p_1^M, \dots$ form a distribution in the **zero-modified** $(a, b, 1)$ **class**.

In the case where we start with a distribution in the $(a, b, 0)$ class, for any $k = 2, 3, \dots$, we have

$$p_k^M = cp_k = c\left(a + \frac{b}{k}\right)p_{k-1} = \left(a + \frac{b}{k}\right)p_{k-1}^M.$$

Hence, a distribution in the zero-modified $(a, b, 1)$ class is also in the $(a, b, 1)$ class.

4.3.2 The constant c in [4.3.1] is indeed uniquely determined, due to the constraint that $p_0^M + p_1^M + \cdots = 1$:

$$p_0^M + p_1^M + p_2^M + \cdots = 1 \implies p_0^M + c(p_1 + p_2 + \cdots) = 1 \implies p_0^M + c(1 - p_0) = 1 \implies c = \boxed{\frac{1 - p_0^M}{1 - p_0}}.$$

4.3.3 Let N be a random variable in the $(a, b, 0)$ class, and let N^M be the random variable in the zero-modified $(a, b, 1)$ class obtained by applying zero-modification on the distribution of N. Then, for any $k = 1, 2, \dots$, the kth moment of N^M is

$$\mathbb{E}\left[\left(N^M\right)^k\right] = \sum_{j=1}^{\infty} j^k p_j^M = \sum_{j=1}^{\infty} j^k c p_j = c \sum_{j=1}^{\infty} j^k p_j = \boxed{c\mathbb{E}\left[N^k\right]}$$

where $c = \dfrac{1 - p_0^M}{1 - p_0} = \dfrac{1 - \mathbb{P}\left(N^M = 0\right)}{1 - \mathbb{P}\left(N = 0\right)}$.

[Warning: We do <u>not</u> have $\mathrm{Var}\left(N^M\right) = c\,\mathrm{Var}\left(N\right)$!]

4.3.4 Consider the same setting as [4.3.3]. The pgf of N^M is

$$P_{N^M}(t) = p_0^M + tp_1^M + t^2 p_2^M + \cdots = p_0^M + c(tp_1 + t^2 p_2 + \cdots) = \boxed{p_0^M + c(P_N(t) - p_0)}$$

where $P_N(t)$ is the pgf of N.

[Intuition: The final expression $\boxed{p_0^M + c(P_N(t) - p_0)}$ corresponds to the "actions" taken on the pgf during the zero modification:

(1) $-p_0$: "take away" the original p_0 from the initial pgf
(2) $c(\cdot)$: scale $p_1, p_2, \dots$ by c
(3) $p_0^M +$: "add back" the modified probability p_0^M

]

We can also rewrite the expression as follows:

$$p_0^M + c(P_N(t) - p_0) = \underbrace{p_0^M - cp_0}_{1-c} + cP_N(t) = \boxed{1 - c + cP_N(t)},$$

which only involves c and $P_N(t)$.

4.3.5 If we modify p_0 to $p_0^M = 0$ in the zero modification process, we call the process as **zero-truncation**, and the resulting distribution is in the **zero-truncated $(a, b, 1)$ class**. Furthermore, we usually denote p_k^M by p_k^T instead for any $k = 0, 1, \ldots$.

4.3.6 Since zero-truncated $(a, b, 1)$ class is essentially a special case of zero-modified $(a, b, 1)$ class, previous formulas in [4.3.2] to [4.3.4] also apply, by setting $p_0^M = 0$. Particularly:

- $c = \dfrac{1}{1 - p_0}$, and $p_k^T = cp_k$ for any $k = 1, 2, \ldots$.
- The pgf of N^T (obtained by zero-truncating N) is

$$P_{N^T}(t) = 1 - c + cP_N(t) = \left(1 - \frac{1}{1 - p_0}\right) + \frac{1}{1 - p_0} P_N(t),$$

which only involves p_0 and $P_N(t)$.

4.4 Extended-Truncated Negative Binomial Distribution

4.4.1 In this section we consider a special way to modify a *negative binomial distribution* which involves *extension* and *truncation*. This allows us to create a distribution in the $(a, b, 1)$ class that *cannot* be obtained by zero-modification.

4.4.2 To motivate this way of modification, consider a random variable $N \sim \mathrm{NB}(r, \beta)$. Its pmf is given by

$$p_k = \mathbb{P}(N = k) = \binom{k + r - 1}{k} \left(\frac{1}{1 + \beta}\right)^r \left(\frac{\beta}{1 + \beta}\right)^k$$

for any $k = 0, 1, \ldots$. Recall that the constraints on the parameters are $r > 0$ and $\beta > 0$.

Now, suppose we improperly set $r \in (-1, 0)$, while $\beta > 0$ still. A consequence is that

$$p_0 = (1 + \beta)^{-r} > 1 \quad \text{and} \quad p_k < 0 \text{ for any } k = 1, 2, \ldots,$$

violating the probability axioms.

4.4.3 Nevertheless, in this case, we can observe that $p_0 + p_1 + \cdots = 1$ still holds, and the recursive relation for the $(a, b, 0)$ class is still satisfied:

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1} \quad \text{for any } k = 1, 2, \ldots,$$

where $a = \dfrac{\beta}{1 + \beta}$ and $b = (r - 1)\dfrac{\beta}{1 + \beta}$ (same as [4.1.6]).

So, allowing $r \in (-1, 0)$ is indeed "mostly" appropriate with just "minor" issues. Hence, we are interested in finding a modification on the $\mathrm{NB}(r, \beta)$ distribution to permit $r \in (-1, 0)$ without violating the probability axioms.

4.4.4 This modification contains two elements: *zero-truncation* ("*truncated*") and *extending* the possible range of r ("*extended*"). We perform zero-truncation on $N \sim \mathrm{NB}(r, \beta)$ (where $r \in (-1, 0)$ and $\beta > 0$), which gives

$$p_0^T = 0 \quad \text{and} \quad p_k^T = cp_k \quad \text{for any } k = 1, 2, \ldots$$

where $c = \dfrac{1}{1 - p_0} < 0$ as $p_0 > 1$.

[Note: By construction of zero-truncation, we always have $p_0^T + p_1^T + p_2^T + \cdots + = 1$. Furthermore, p_k^T is nonnegative for any $k = 0, 1, \ldots$, since $p_k < 0$ for any $k = 1, 2, \ldots$ and $c < 0$. Thus, the probability axioms are not violated.]

Then, the probabilities $p_0^T, p_1^T, \ldots$ form the **extended-truncated negative binomial distribution** with parameters $r \in (-1, 0)$ and $\beta > 0$ (denoted by $\mathrm{ETNB}(r, \beta)$).

4.4.5 Since ETNB (r, β) is in the zero-truncated $(a, b, 1)$ class, previous results for this class apply. Particularly, we have the following recursive formula for the $(a, b, 1)$ class:

$$p_k^T = \left(a + \frac{b}{k}\right) p_{k-1}^T \quad \text{for any } k = 2, 3, \ldots$$

where $a = \dfrac{\beta}{1+\beta}$ and $b = (r-1)\dfrac{\beta}{1+\beta}$ (same as [4.1.6]).

4.4.6 As a practical note, to compute probabilities for ETNB (r, β) where $r \in (-1, 0)$ and $\beta > 0$, we carry out the following steps:

(a) Treat it as if it were an ordinary NB (r, β) distribution and use the pmf formula to compute

$$p_k = \binom{r+k-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k$$

for all k needed, depending on the method chosen in the second step below.

(b) The desired ETNB (r, β) probabilities can then be obtained by one of the following methods.

i. Directly use the following:

$$p_0^T = 0 \quad \text{and} \quad p_k^T = c p_k \quad \text{for any } k = 1, 2, \ldots$$

where $c = \dfrac{1}{1 - p_0}$.

ii. First compute $p_1^T = c p_1$, and then use the following recursive formula to compute the rest of the probabilities needed:

$$p_k^T = \left(a + \frac{b}{k}\right) p_{k-1}^T \quad \text{for any } k = 2, 3, \ldots$$

where $a = \dfrac{\beta}{1+\beta}$ and $b = (r-1)\dfrac{\beta}{1+\beta}$ (same as [4.1.6]).

5 Compound Frequency Models

5.1 Compound Frequency Distributions

5.1.1 Apart from mixing discussed in Section 2, another method to create a new discrete distribution is to *compound* ($\neq$ mix!) discrete distributions.

5.1.2 We first let:

- N: number of accidents arising in a portfolio of risks
- $M_1, M_2, \ldots$: number of claims from each accident
- S: total number of claims

Also, we assume that $M_1, M_2, \ldots,$ are i.i.d., and are independent of N. When $N = 0$, we set $S = 0$.

Then, S is a **compound random variable**, written as $S = M_1 + \cdots + M_N$. The distribution of S is said to be a **compound distribution**. The distribution of N is called **primary distribution**, and the common distribution of $M_1, M_2, \ldots$ is called **secondary distribution**.

[Note: More generally, even when $M_1, M_2, \ldots$ are *continuous*, S is still said to be a compound random variable. But of course it would not make sense to talk about "pgf" of the secondary distribution anymore. However, the results about mean and variance of S still applies as the proof in [5.1.4] would still work. See Section 7 for more examples about this case.]

Remarks:

- The number of summands is N, a random variable.
- The equation $S = M_1 + \cdots + M_N$ means that given $N = n$, we have $S = M_1 + \cdots + M_n$.

5.1.3 The pgf of a compound random variable takes the following form.

Suppose that $M_1, M_2, \ldots$ are i.i.d. and are independent of N. Let M be a random variable following the secondary distribution. Then, the pgf of S is

$$P_S(t) = P_N(P_M(t))$$

where P_N and P_M denote the pgf's of N and M respectively.

Proof. Note that

$$\begin{aligned}
P_S(t) &= \sum_{k=0}^{\infty} t^k \mathbb{P}(S = k) \\
&= \sum_{k=0}^{\infty} t^k \sum_{n=0}^{\infty} \mathbb{P}(S = k | N = n)\, \mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \sum_{k=0}^{\infty} t^k \mathbb{P}(M_1 + \cdots + M_n = k | N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \sum_{k=0}^{\infty} t^k \mathbb{P}(M_1 + \cdots + M_n = k) && \text{(by independence)} \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n)\, P_{M_1 + \cdots + M_n}(t) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n)\, [P_M(t)]^n && \text{(Theorem 3.1.d, i.i.d. assumption)} \\
&= \mathbb{E}\left[[P_M(t)]^N\right] \\
&= P_N(P_M(t)).
\end{aligned}$$

□

5.1.4 Consider the setting in [5.1.3]. The mean and variance of S are as follows.

- $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[M]$
- $\mathrm{Var}(S) = \mathbb{E}[N]\,\mathrm{Var}(M) + (\mathbb{E}[M])^2\,\mathrm{Var}(N)$

[⚠ Warning: In the variance formula, there is a square for the mean of the random variable following the secondary distribution!]

Proof. Firstly, note that

$$\mathbb{E}[S|N=n] = \mathbb{E}[M_1 + \cdots + M_N|N=n] = \mathbb{E}[M_1 + \cdots + M_n|N=n] = \mathbb{E}[M_1 + \cdots + M_n] = n\mathbb{E}[M].$$

Hence, $\mathbb{E}[S|N] = N\mathbb{E}[M]$. Thus, by law of total expectation,

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]] = \mathbb{E}[N\mathbb{E}[M]] = \mathbb{E}[M]\mathbb{E}[N].$$

Next, for the variance formula, first consider:

$$\begin{aligned}
\mathrm{Var}(S|N=n) &= \mathrm{Var}(M_1 + \cdots + M_N|N=n) \\
&= \mathrm{Var}(M_1 + \cdots + M_n|N=n) \\
&= \mathrm{Var}(M_1 + \cdots + M_n) && \text{(by independence)} \\
&= n\,\mathrm{Var}(M) && \text{(i.i.d. assumption)}.
\end{aligned}$$

Thus, $\mathrm{Var}(S|N) = N\,\mathrm{Var}(M)$, and by law of total variance,

$$\mathrm{Var}(S) = \mathbb{E}[\mathrm{Var}(S|N)] + \mathrm{Var}(\mathbb{E}[S|N]) = \mathbb{E}[N\mathrm{Var}(M)] + \mathrm{Var}(N\mathbb{E}[M]) = \mathrm{Var}(M)\mathbb{E}[N] + (\mathbb{E}[M])^2\,\mathrm{Var}(N).$$

□

5.2 Panjer's Recursion

5.2.1 Consider again the setting in [5.1.3]. Let:

- $g_k = \mathbb{P}(S=k)$
- $p_k = \mathbb{P}(N=k)$
- $f_k = \mathbb{P}(M=k)$

Suppose that p_k and f_k are known for any $k = 0, 1, \ldots$. Then, we are interested in finding the distribution of S, i.e., the value of g_k for any $k = 0, 1, \ldots$. Unfortunately, in general there is no simple way to do this.

5.2.2 However, in the special case where N is in the $(a, b, 0)$ class, we have the following result that allows us to compute g_k recursively.

Theorem 5.2.a (Panjer's recursion ($(a, b, 0)$ class))**.** Suppose that N is in the $(a, b, 0)$ class. Then, for any $k = 1, 2, \ldots$,

$$g_k = \frac{1}{1 - af_0} \sum_{j=1}^{k} \left(a + \frac{bj}{k}\right) f_j g_{k-j}.$$

Proof. Omitted (See, e.g., proof of Theorem 7.1 in Klugman et al. (2019)). □

Remarks:

- The a and b are the constants a and b in the $(a, b, 0)$ class characterization of S.
- To use the Panjer's recursion, we need the initial value g_0. It can be obtained by

$$g_0 = P_S(0) = P_N(P_M(0)) = P_N(f_0).$$

- Some special cases:
  - $(k=1)$ $g_1 = \dfrac{1}{1-af_0}\left(a+\dfrac{b\cdot 1}{1}\right)f_1 g_0$
  - $(k=2)$ $g_2 = \dfrac{1}{1-af_0}\left[\left(a+\dfrac{b\cdot 1}{2}\right)f_1 g_1 + \left(a+\dfrac{b\cdot 2}{2}f_2 g_0\right)\right]$
  - $(k=3)$ $g_3 = \dfrac{1}{1-af_0}\left[\left(a+\dfrac{b\cdot 1}{3}\right)f_1 g_2 + \left(a+\dfrac{b\cdot 2}{3}f_2 g_1\right) + \left(a+\dfrac{b\cdot 3}{3}f_3 g_0\right)\right]$
  - ($a=0$ and $b=\lambda$) $g_k = \dfrac{\lambda}{k}\sum_{j=1}^{k} j f_j g_{k-j}$

5.2.3 When N is in the $(a,b,1)$ class instead, we can slightly modify the Panjer's recursion to obtain the appropriate recursive formula as follows.

Theorem 5.2.b (Panjer's recursion ($(a,b,1)$ class))**.** Suppose that N is in the $(a,b,0)$ class. Then, for any $k=1,2,\ldots$,

$$g_k = \frac{1}{1-af_0}\left\{[p_1-(a+b)p_0]f_k + \sum_{j=1}^{k}\left(a+\frac{bj}{k}\right)f_j g_{k-j}\right\}.$$

Proof. Omitted (See, e.g., proof of Theorem 7.2 in Klugman et al. (2019)). □

5.3 Compound Poisson Frequency Distributions

5.3.1 If S is a compound random variable with the *primary* distribution being $\mathrm{Poi}(\lambda)$, then S follows a **compound Poisson distribution**, and S is called a **compound Poisson random variable**.

5.3.2 Recall from [3.2.2] that the pgf of $N \sim \mathrm{Poi}(\lambda)$ is

$$P_N(s) = \exp\{\lambda(s-1)\}.$$

Thus, the pgf of a compound Poisson random variable S is

$$P_S(t) = P_N(P_M(t)) = \boxed{\exp\{\lambda(P_M(t)-1)\}}$$

where P_M is the pgf for the secondary distribution.

5.3.3 Recall the convolution result for Poisson distribution (Theorem 3.2.a). It turns out that analogous result holds also for *compound* Poisson distribution:

Theorem 5.3.a. Let $S_1,\ldots,S_k$ be k independent compound Poisson random variables with Poisson parameters $\lambda_1,\ldots,\lambda_k$ respectively. Suppose that the pmf of the secondary distribution for S_i is given by $q_i(n)$, for any $i=1,\ldots,k$. Then, the sum $S=S_1+\cdots+S_k$ is also a compound Poisson random variable with Poisson parameter $\lambda=\lambda_1+\cdots+\lambda_k$, and the pmf of the secondary distribution is given by

$$q(n) = \frac{\lambda_1}{\lambda}q_1(n)+\cdots+\frac{\lambda_k}{\lambda}q_k(n).^{5}$

Proof. Let Q_i be the pgf of the secondary distribution for S_i, i.e.,

$$Q_i(t) = \sum_{n=0}^{\infty} t^n q_i(n).$$

[5] $q(n)$ is indeed a legitimate pmf since $q(n)\ge 0$ as $q_1(n),\ldots,q_k(n)\ge 0$ for any $n=0,1,\ldots$ and

$$\sum_{n=0}^{\infty} q(n) = \frac{\lambda_1}{\lambda}\sum_{n=0}^{\infty} q_1(n)+\cdots+\frac{\lambda_k}{\lambda}\sum_{n=0}^{\infty} q_k(n) = \frac{\lambda_1+\cdots+\lambda_k}{\lambda} = 1.$$

Then, for any $i = 1, \ldots, k$, the pgf of S_i can be expressed as

$$P_{S_i}(t) = \exp\{\lambda_i(Q_i(t) - 1)\}.$$

Also, the pgf of the secondary distribution for S is

$$Q(t) = \sum_{n=0}^{\infty} t^n q(n) = \frac{\lambda_1}{\lambda} \sum_{n=0}^{\infty} t^n q_1(n) + \cdots + \frac{\lambda_k}{\lambda} \sum_{n=0}^{\infty} t^n q_k(n) = \frac{\lambda_1}{\lambda} Q_1(t) + \cdots + \frac{\lambda_k}{\lambda} Q_k(t).$$

Since $S_1, \ldots, S_k$ are independent, by Theorem 3.1.d, the pgf of S is

$$P_S(t) = \prod_{i=1}^{k} P_{S_i}(t) = \prod_{i=1}^{k} \exp\{\lambda_i(Q_i(t) - 1)\} = \exp\left\{\sum_{i=1}^{k} \lambda_i Q_i(t) - \sum_{i=1}^{k} \lambda_i\right\} = \exp\{\lambda(Q(t) - 1)\}.$$

Hence, S is a compound Poisson random variable with secondary distribution identified by the pgf $Q(t)$, i.e., the pmf of the secondary distribution is

$$q(n) = \frac{\lambda_1}{\lambda} q_1(n) + \cdots + \frac{\lambda_k}{\lambda} q_k(n).$$

□

5.3.4 Practically, to use Theorem 5.3.a, we may draw a table like below to keep track of the primary and secondary distributions of the sum $S = S_1 + \cdots + S_k$ more conveniently:

	λ	0	1	2	3	
S_1	1	0	0	1	0	row $\times$ 1/6
S_2	2	0	0.5	0	0.5	row $\times$ 2/6
S_3	3	0.3	0.7	0	0	row $\times$ 3/6
S	6	0.15	0.5167	0.1667	0.1667	$\leftarrow$ sum

Basically we are summing rows that are scaled according to the "contributions" of the Poisson parameters. The columns "0", "1", "2", "3" above correspond to the probability masses at those values. Thee columns where all the probability masses are zero have been ignored.

6 Coverage Modifications

6.0.1 In this section, we will discuss different kinds of *coverage modifications*: modifications on the "coverage"/term for an insurance on some loss. Mathematically speaking, with coverage modifications, the amount of actual payment made by the insurer is obtained by modifying the original loss X, which is also known as the **ground-up loss**.

6.0.2 Some of the coverage modifications here have been briefly discussed in Section 1, namely *deductibles* and *policy limit*. Here we will discuss more varieties of coverage modifications in more details.

6.1 Ordinary Deductibles

6.1.1 Consider an insurance policy with a per-loss ordinary deductible d. Then, for every loss claimed:

- If the loss $X \le d$, then there is no payment.
- If the loss $X > d$, then the payment amount is $X - d$.

The first d dollars of the loss is borne by the policyholder. If there is no such deductible and the full loss amount is covered by the insurance, may be incentivized to take "too much" risk as is not responsible for bearing the loss. This issue is known as **moral hazard**, and imposing an ordinary deductible is an useful way to avoid it.

6.1.2 With the presence of a per-loss ordinary deductible d, we can modify the loss X in the following ways:

- loss X → the **per-payment variable** (i.e., excess loss variable):

$$Y^P = (X - d | X > d) = \begin{cases} \text{undefined} & \text{if } X \le d; \\ X - d & \text{if } X > d. \end{cases}$$

  The *per-payment* variable is the payment amount (or cost) *per payment*, which only comes into existence when $X > d$.

- loss X → the **per-loss variable** (i.e., stop loss variable):

$$Y^L = (X - d)_+.$$

  The *per-loss* variable is the payment amount *per loss*. It exists regardless of the amount of loss.

6.1.3 We can also write the per-payment variable as follows:

$$Y^P = \begin{cases} \text{undefined} & \text{if } Y^L \le 0; \\ Y^L & \text{if } Y^L > 0 \end{cases} = \boxed{(Y^L | Y^L > 0)}.$$

This shows that the following three distributions are equal:

(a) distribution of Y^P

(b) conditional distribution of $X - d$ given $X > d$

(c) conditional distribution of Y^L given $Y^L > 0$

6.1.4 To obtain the probabilistic quantities of Y^L and Y^P, we can just use the methods introduced in Section 1.

6.2 Franchise Deductibles

6.2.1 A **franchise deductible** is a modified version of ordinary deductible where the deductible amount is added on top of the payment amount $\$$ *when the payment amount* $\$$ *is positive*. In other words, when there is a franchise deductible d and the loss $X > d$, the payment amount is X instead of $X - d$.

6.2.2 Hence, with the per-loss *franchise* deductible d, we can modify the loss X in the following ways:

- loss X → the **per-payment variable** *in this context*:

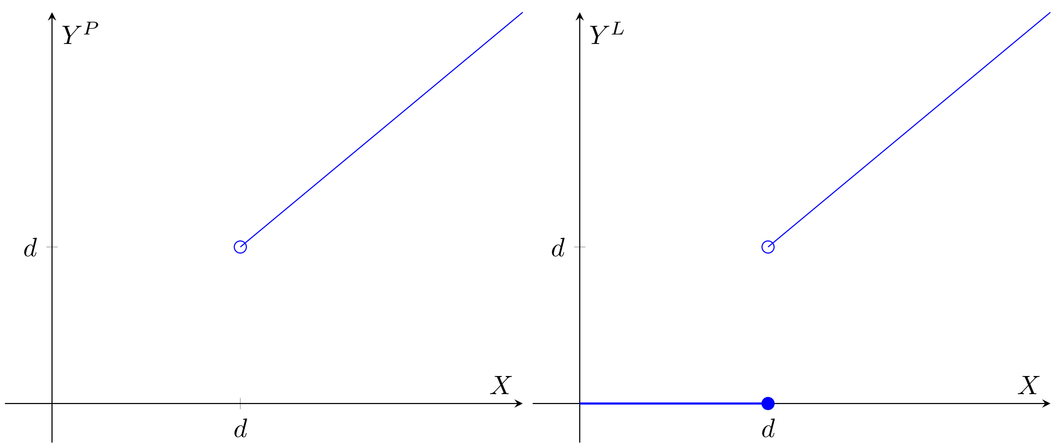
$$Y^P = (X|X > d) = \begin{cases} \text{undefined} & \text{if } X \le d; \\ X & \text{if } X > d. \end{cases}$$

- loss X → the **per-loss variable** *in this context*:

$$Y^L = \begin{cases} 0 & \text{if } X \le d; \\ X & \text{if } X > d. \end{cases}$$

[⚠ Warning: This is neither "X_+" nor "$d + (X - d)_+$"!]

Graphically, the per-payment and per-loss variables look like the following.



These graphs are often useful for obtaining the probabilistic quantities about per-payment and per-loss variables, as the key for getting such quantities is often to relate them with the quantities about X.

6.2.3 Likewise, we can also write the per-payment variable as follows:

$$Y^P = \begin{cases} \text{undefined} & \text{if } Y^L \le 0; \\ Y^L & \text{if } Y^L > 0 \end{cases} = \boxed{(Y^L|Y^L > 0)}.$$

This means that the following three distributions are equal:

(a) distribution of Y^P

(b) conditional distribution of X given $X > d$

(c) conditional distribution of Y^L given $Y^L > 0$

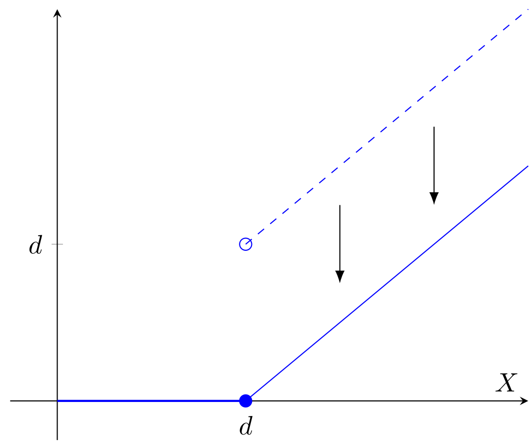
6.2.4 To compute the means of Y^L and Y^P in the case of franchise deductible, consider first the following. Let

$$W = \begin{cases} 0 & \text{if } X \le d; \\ -d & \text{if } X > d. \end{cases}$$

Then,

$$Y^L + W = \begin{cases} 0 & \text{if } X \le d; \\ X - d & \text{if } X > d \end{cases} = (X - d)_+.$$

[Intuition: Basically we are trying to "shift" the line in the graph below downwards by d units, through adding an extra random variable W, to reassemble the graph for the ordinary deductible case.



]

From this relationship, we can readily derive the following formula for $\mathbb{E}\left[Y^L\right]$:

$$\mathbb{E}\left[Y^L\right] = \mathbb{E}\left[(X - d)_+\right] - \underbrace{\mathbb{E}\left[W\right]}_{-d\cdot\mathbb{P}(X>d)} = \boxed{\mathbb{E}\left[(X - d)_+\right] + d \cdot \mathbb{P}\left(X > d\right)}.$$

For the per-payment variable Y^P, the formula takes the same form as the case for ordinary deductible:

$$\mathbb{E}\left[Y^P\right] = \boxed{\frac{\mathbb{E}\left[Y^L\right]}{\mathbb{P}\left(Y^L > 0\right)}} = \boxed{\frac{\mathbb{E}\left[Y^L\right]}{\mathbb{P}\left(X > d\right)}}$$

but of course the meaning of Y^L here is different from that for ordinary deductible.

6.3 Loss Elimination Ratio

6.3.1 The loss elimination ratio quantifies the effect of an *ordinary* deductible in lowering the expected payment made by the insurer per loss, i.e., how much expected loss *for the insurer* (taken out of insurer's pocket) is "eliminated".

6.3.2 More precisely, the **loss elimination ratio** is the ratio of the decrease in the expected payment made by the insurer per loss with an ordinary deductible to the expected payment without the deductible, i.e.,

$$\text{loss elimination ratio} = \frac{\mathbb{E}\left[X\right] - \mathbb{E}\left[(X - d)_+\right]}{\mathbb{E}\left[X\right]} = \frac{\mathbb{E}\left[X \wedge d\right]}{\mathbb{E}\left[X\right]}.$$

6.4 Inflation

6.4.1 In practice, there is often a *delay* between the time at which the loss X is triggered (occurrence of accident) and the time at which the payment is made by the insurer, since it takes time for to "process" a claim.

6.4.2 In case the *inflation rate* is very high, such delay can cause a substantial drop in the *real worth* of the payment received by the policyholder. To protect against this inflation risk, we can modify the terms of the insurance to incorporate also the inflation element.

6.4.3 More specifically, suppose that the inflation rate (over the whole delay period) is r. Here, to incorporate the inflation element, we adjust the ground-up loss X to the inflated loss $X' = X(1+r)$. Then, the insurer makes payment based on the loss after inflation X' instead of X.

[Note: One may argue that it may be more reasonable for the *payment* itself should be multiplied by $1+r$ instead of the ground-up loss. But in our setting here, we consider adjusting the ground-up loss only, by convention.]

6.4.4 For an insurance with ordinary deductible d which incorporates inflation, the expected payment made by per loss is

$$\mathbb{E}\left[Y^L\right] = \mathbb{E}\left[(X'-d)_+\right] = \mathbb{E}\left[((1+r)X-d)_+\right] = \boxed{(1+r)\mathbb{E}\left[\left(X-\frac{d}{1+r}\right)_+\right]}.$$

6.4.5 For an insurance with ordinary deductible d which incorporates inflation, the expected payment made by per payment is

$$\mathbb{E}\left[Y^P\right] = \frac{\mathbb{E}\left[Y^L\right]}{\mathbb{P}\left(X'>d\right)} = \boxed{\frac{\mathbb{E}\left[Y^L\right]}{\mathbb{P}\left(X>\frac{d}{1+r}\right)}}.$$

6.5 Policy Limits

6.5.1 Consider an insurance policy with a policy limit u. Then, recall that:

- If the loss $X \le u$, then the insurer pays the full amount u to the policyholder.
- If the loss $X > u$, then only pays u dollars to.

The policy limit u is the maximum amount of payment made by.

6.5.2 With the policy limit u, the loss X is modified to $Y = X \wedge u$, which is the payment made by. As Y is just the same as the limited loss variable discussed in Section 1, we can again use the methods there to get the probabilistic quantities of Y. [Note: A thought question: Why don't we distinguish "per-payment" and "per-loss" variables here?]

6.6 Coinsurance

6.6.1 The final coverage modification introduced here is the *coinsurance*. The idea is that both parties — policyholder and insurer — contribute to the insurance coverage together (hence "co").

6.6.2 For an insurance policy with **coinsurance** element added, the insurer only pays a certain fixed proportion α of the original payment amount, where $\alpha \in [0,1]$[6]. Hence, assuming no other coverage modifications, the payment made by is $X' = \alpha X$.

6.7 General Policy

6.7.1 Here, a **general policy** refers generally to any insurance with possibly multiple kinds of coverage modifications.

6.7.2 We consider a general policy with the following coverage modifications:

[6]The minimum contribution is to have no contribution for both parties, so no "negative" contribution is allowed! A negative contribution effectively allows the party to *benefit from loss*, which can lead to some moral issues.

(a) inflation (rate: r)

(b) policy limit u

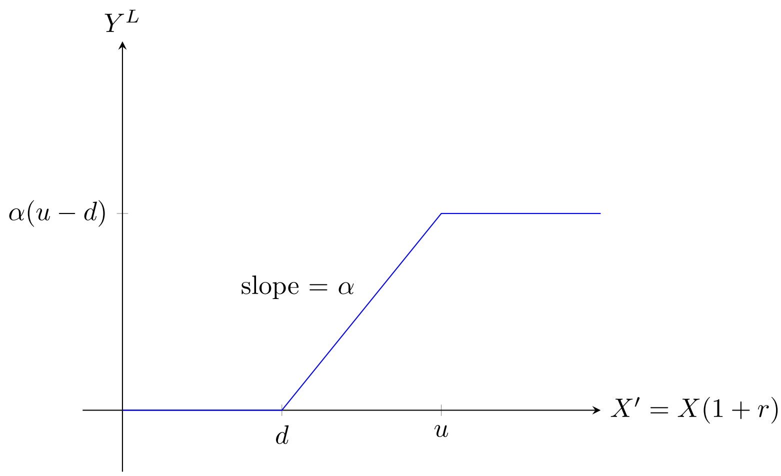
(c) ordinary deductible d

(d) coinsurance (proportion: α)

We shall assume that $u > d$ and that the modifications are applied *in the order above*:

$$X \overset{\text{inflation}}{\rightarrow} X(1+r) \overset{\text{policy limit}}{\rightarrow} X(1+r) \wedge u \overset{\text{deductible}}{\rightarrow} [X(1+r) \wedge u - d]_+ \overset{\text{coinsurance}}{\rightarrow} \alpha[X(1+r) \wedge u - d]_+.$$

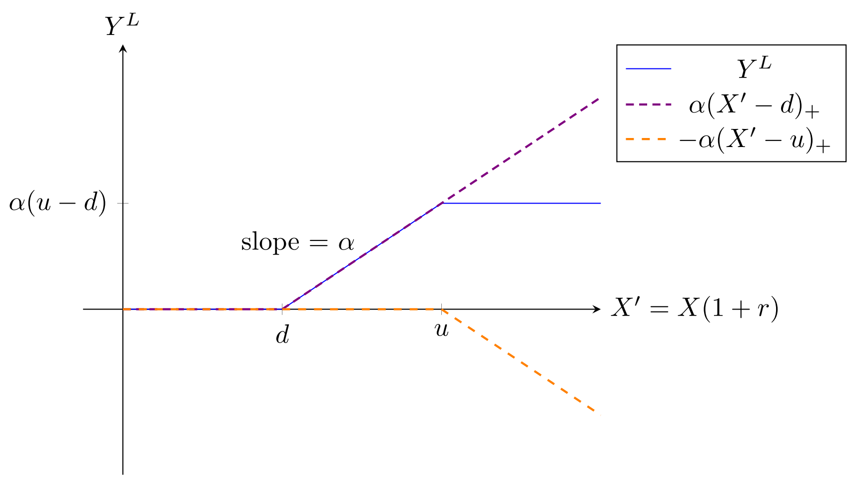
6.7.3 We can graphically show the final payment amount per loss $Y^L = \alpha[X(1+r) \wedge u - d]_+$ as follows:



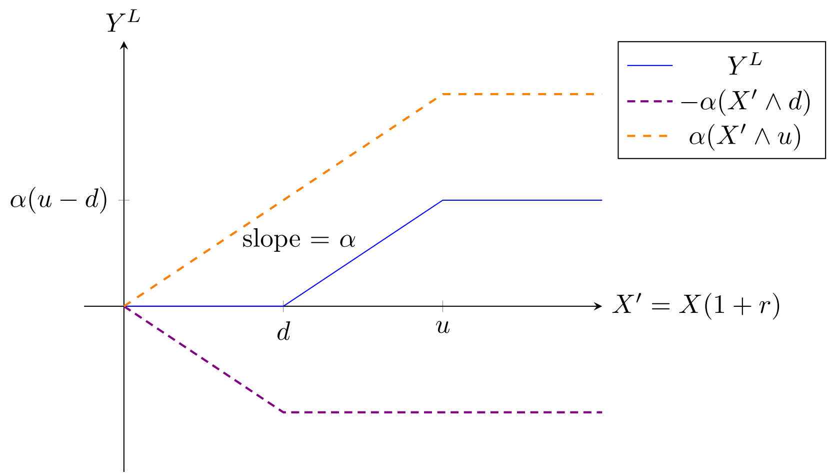
[Note: It has a similar shape as the payoff graph of *bull spread* in STAT3905.]

6.7.4 Based on the graph in [6.7.3], we can deduce the following formulas for Y^L:

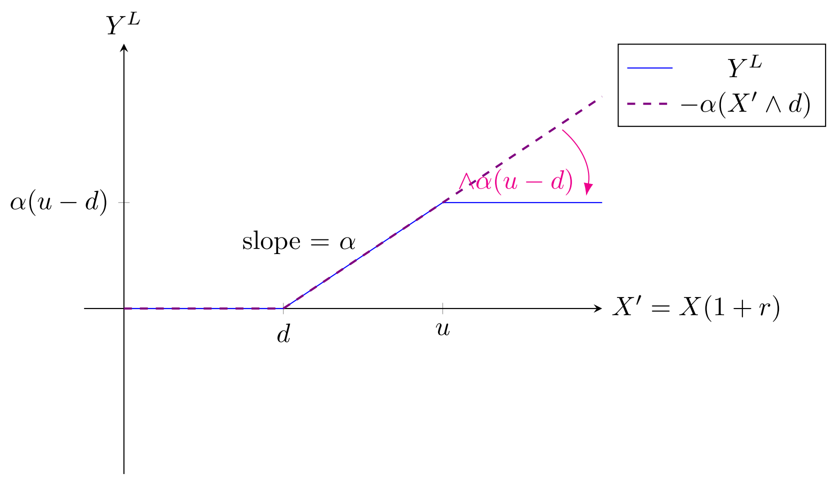
(a) $Y^L = \boxed{\alpha[(X'-d)_+ - (X'-u)_+]}$



(b) $Y^L = \boxed{\alpha[(X' \wedge u) - (X' \wedge d)]}$

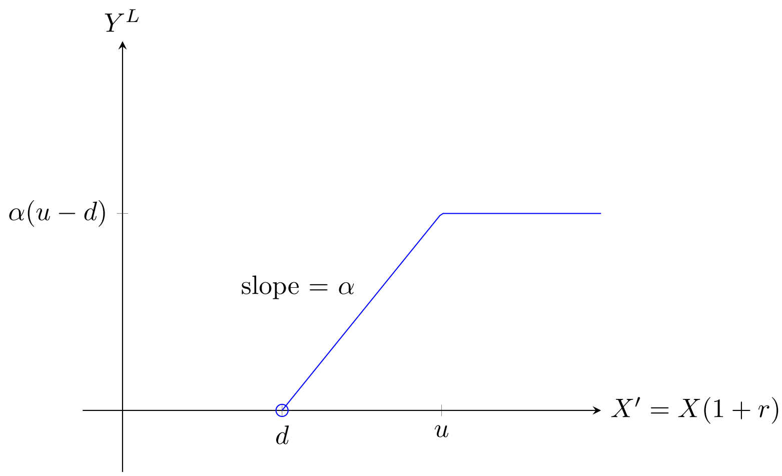


(c) $Y^L = \boxed{\alpha[(X'-d)_+ \wedge (u-d)]}$



6.7.5 Also, from the graph in [6.7.3], we see that despite the "nominal" policy limit applied is u, the *true* policy limit for the resulting insurance, i.e., the maximum amount of payment made by the insurer is $\alpha(u-d)$. Since no additional payment is made by the insurer when the loss after inflation X' exceeds u, we call u the **maximum covered loss** in this case (which may not equal the policy limit $\alpha(u-d)$!).

6.7.6 In a similar manner, we can graphically show the final payment amount *per payment* $Y^P = (Y^L | Y^L > 0)$ as follows:



6.7.7 The key to obtain the probabilistic quantities of Y^L and Y^P is often to utilize the expressions in [6.7.4]b.

6.8 Number of Payments and Losses With the Presence of Deductibles

6.8.1 With the presence of deductibles, some payments that would be made if there were not deductibles may vanish. Hence, the number of payments made by the insurer would become lower. The number of payments made may no longer be the same as the number of losses. So, here we will investigate the relationship between these two numbers with the presence of deductibles.

6.8.2 Suppose that there are N^L independent losses, and let X_j be the amount/severity of jth loss. For a given insurance policy with possibly coverage modification, let v be the probability that a loss results in a payment. For example, $v = \mathbb{P}(X > d)$ for an insurance with just an ordinary deductible of d.

Define the indicator random variable I_j for the jth loss by

$$I_j = \begin{cases} 1 & \text{if } j\text{th loss results in payment;} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we can express the total number of payments N^P as a *compound random variable*:

$$N^P = I_1 + I_2 + \cdots + I_{N^L}.$$

where $I_1, I_2, \ldots$ are i.i.d. $\text{Ber}(v)$ random variables that are independent of N^L.

6.8.3 Since $\text{Ber}(v) \equiv \text{Bin}(1, v)$, the pgf of the secondary distribution is $P_I(t) = 1 - v + vt$. Hence, by [5.1.3], the pgf of the compound random variable N^P is

$$P_{N^P}(t) = P_{N^L}(P_I(t)) = \boxed{P_{N^L}(1 - v + vt)}.$$

6.8.4 To obtain the mean and variance of the number of payments N^P, we can just utilize [5.1.4] as it is just a compound random variable.

6.8.5 The following result suggests a relationship between N^L and N^P under the condition that N^L is in the $(a, b, 0)$ or zero-modified $(a, b, 1)$ class.

Theorem 6.8.a. Suppose that N^L is in the $(a, b, 0)$ or zero-modified $(a, b, 1)$ class. Then, N^P and N^L belong to the same parametric family of distributions (with possibly different parameters)[7].

Proof. It can be proved by checking the cases one by one, but we will omit the proof here. □

[7] In other words, their pgf take the same form, with possibly different parameters.

7 Aggregate Loss Model

7.0.1 Previously, when multiple losses are involved, we consider them in isolation. Here, we study the behaviour of their *sum*, called **aggregate loss**.

A practical setting where such sum arises is when an insurer sells many insurance policies and the insurer is considering the *total* loss from those policies.

7.1 Collective Risk Model

7.1.1 In the **collective risk model**, the aggregate loss is a compound random variable with primary (secondary) distribution being claim frequency distribution (severity distribution):

$$S = X_1 + \cdots + X_N.$$

Remarks:

- The random variable N is called **claim count** (or **claim frequency**).
- Each random variable X_j is called **severity**/**single loss**/**individual loss**.
- Do not forget the assumptions imposed in the definition of compound distribution in Section 5!

We work in the collective risk model henceforth in this section.

7.1.2 By [5.1.4], the mean and variance of S are:

- $\mathbb{E}[S] = \boxed{\mathbb{E}[N]\,\mathbb{E}[X]}$.
- $\mathrm{Var}(S) = \boxed{\mathbb{E}[N]\,\mathrm{Var}(X) + (\mathbb{E}[X])^2\,\mathrm{Var}(N)}$.

7.1.3 Consider the collective risk model equipped with the setting in [6.8.2]. Suppose that the given insurance policy has an ordinary/franchise deducible of d applied to individual losses. Then, the probability that a loss results in a payment is $v = \mathbb{P}(X > d)$. Recall that the total number of payments is

$$N^P = I_1 + I_2 + \cdots + I_{N^L}$$

where N^L is the total number of losses.

Now, for any $i = 1, 2, \ldots$, let the per-loss and per-payment variables be $Y_i^L = (X_i - d)_+$ and $Y_i^P = (X_i - d | X_i > d)$ respectively.

In this setting, the **aggregate payment** (i.e., the sum of all payments) can be calculated in the following ways:

- per-loss perspective: aggregate payment $= \boxed{Y_1^L + \cdots + Y_{N^L}^L}$ [Intuition: This sums "payments" associated to all losses (N^L of them), including those with zero value.]
- per-payment perspective: aggregate payment $= \boxed{Y_1^P + \cdots + Y_{N^P}^P}$ [Intuition: This sums only positive payments (N^P of them).]

Both methods would result in the same aggregate payment, and the choice of the method used depends on what information is given.

7.2 Convolution

7.2.1 To obtain more distributional quantities for the aggregate loss S in the collective risk model, we would like to deduce the actual probability distribution of S. Of course in the special case where $X_1, X_2, \ldots$ are all discrete and N belongs to the $(a, b, 0)$ or $(a, b, 1)$ family, we can use Panjer's recursion.

But in general, we would need to resort to finding the cdf of S. By Proposition 2.2.b, we can write

$$
\begin{aligned}
F_S(s) &= \mathbb{P}(S \le s) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(S \le s | N = n)\, \mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X_1 + \cdots + X_n \le s | N = n)\, \mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X_1 + \cdots + X_n \le s)\, \mathbb{P}(N = n).
\end{aligned}
$$

[Note: Here "$X_1 + \cdots + X_n$" is supposed to be 0 when $n = 0$.] We can then observe that the cdf's of *sums* of random variables $\mathbb{P}(X_1 + \cdots + X_n \le s)$ are needed for finding the cdf of S. The concept of *convolution* provides a systematic way to deduce such cdf's.

7.2.2 Let $X_1, X_2, \ldots$ be i.i.d. continuous random variables with the same cdf F_X and pdf f_X. Define $S_n = X_1 + \cdots + X_n$ which is a sum of n i.i.d. random variables. Denote its cdf by F_X^{*n}:

$$
F_X^{*n}(x) = \mathbb{P}(X_1 + \cdots + X_n \le x).
$$

7.2.3 To find the cdf F_X^{*n}, we can perform the following recursive process (**convolution**):

(1) Define $F_X^{*1} = F_X$ (the common cdf of $X_1, X_2, \ldots$).

(2) Use the following recursive formula to find $F_X^{*2}, F_X^{*3}, \ldots$, up to F_X^{*n}:

$$
\begin{aligned}
F_X^{*k}(x) &= \mathbb{P}(X_1 + \cdots + X_k \le x) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X_1 + \cdots + X_k \le x | X_k = y)\, f_X(y)\,\mathrm{d}y \quad \text{(law of total probability, continuous case)} \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X_1 + \cdots + X_{k-1} \le x - y)\, f_X(y)\,\mathrm{d}y \\
&= \boxed{\int_{-\infty}^{\infty} F_X^{*(k-1)}(x-y) f_X(y)\,\mathrm{d}y}
\end{aligned}
$$

for any $k = 2, 3, \ldots$.

[⚠ Warning: Be careful about when the cdf $F_X^{*(k-1)}(x-y)$ is zero and when the pdf $f_X(y)$ is zero!] When $X_1, X_2, \ldots$ are nonnegative, the integral above can be written as

$$
\int_0^x F_X^{*(k-1)}(x-y) f_X(y)\,\mathrm{d}y
$$

since $F_X^{*(k-1)}(x-y) = 0$ when $y > x$ and $f_X(y) = 0$ when $y < 0$.

7.2.4 Generally speaking, it is extremely challenging to carry out the convolution approach as the complexity of the expressions would generally skyrocket as the recursive process above goes on, unless the distributions involved have some "nice" properties, e.g., normal distribution has the remarkable property that sum of independent normal random variables is still normal.

Hence, in practice, one often just *approximates* the distribution of S instead of finding its exact form. There are two main methods:

(a) method of rounding (to be introduced in Section 7.3)

(b) normal approximation

Carrying out normal approximation is quite simple, we just suppose that S is "approximately normally distributed" and approximate the cdf of S by

$$\mathbb{P}(S \leq s) \approx \Phi\left(\frac{S - \mathbb{E}[S]}{\sqrt{\operatorname{Var}(S)}}\right)$$

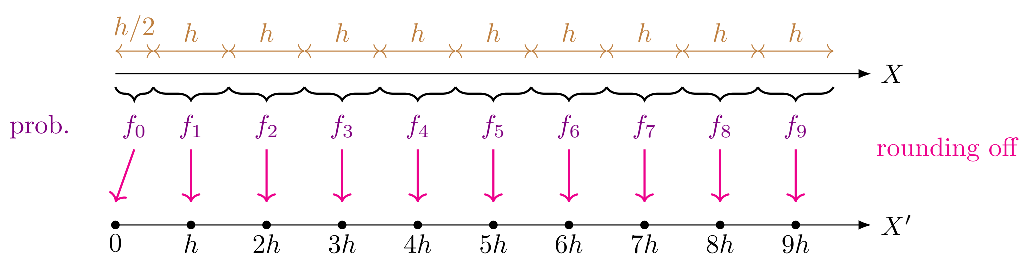
where Φ is the standard normal cdf. Computing $\mathbb{E}[S]$ and $\operatorname{Var}(S)$ should be a manageable task since we have the formulas from [7.1.2].

7.3 Method of Rounding

7.3.1 Even when the claim count N is in the $(a, b, 0)$ class or $(a, b, 1)$ class, Panjer's recursion would not be applicable if the severity is *continuous*. However, by *approximating* the continuous severity distribution by a discrete distribution (known as **discretization**), we may then apply Panjer's recursion to *approximate* the distribution of S. Here, we introduce a way of discretization: *method of rounding*.

7.3.2 The **method of rounding** is as follows. Let X be the original continuous severity with cdf F_X. Then, we approximate X by a discrete random variable X' defined by:

$$X' = \begin{cases} 0 & \text{w.p. } f_0 = F_X(h/2); \\ h & \text{w.p. } f_1 = F_X(h + h/2) - F_X(h - h/2); \\ 2h & \text{w.p. } f_2 = F_X(2h + h/2) - F_X(2h - h/2); \\ \vdots & \vdots \\ jh & \text{w.p. } f_j = F_X(jh + h/2) - F_X(jh - h/2); \\ \vdots & \vdots \end{cases}$$



[Note: The value h is said to be the **span**.]

7.3.3 After that, we can approximate the aggregate loss S by the discretized version: $S' = X'_1 + \cdots + X'_N$, where the secondary distribution now becomes the distribution of X' above. Then, we can apply Panjer's recursion to S' as long as N is in the $(a, b, 0)$ class or $(a, b, 1)$ class.

7.4 Stop-Loss Premium

7.4.1 Practically, it is common for an insurance to not apply deductible *individually* on each loss claimed, but apply it on the losses claimed for a certain period *in aggregate*.

7.4.2 Consider an aggregate loss S, which may be interpreted as sum of individual losses in a certain period. An insurance on the aggregate loss S subject to an aggregate deductible d is called **stop-loss insurance**. The expected payment per loss of this insurance is called the **(net) stop-loss premium**: $\mathbb{E}[(S - d)_+]$.

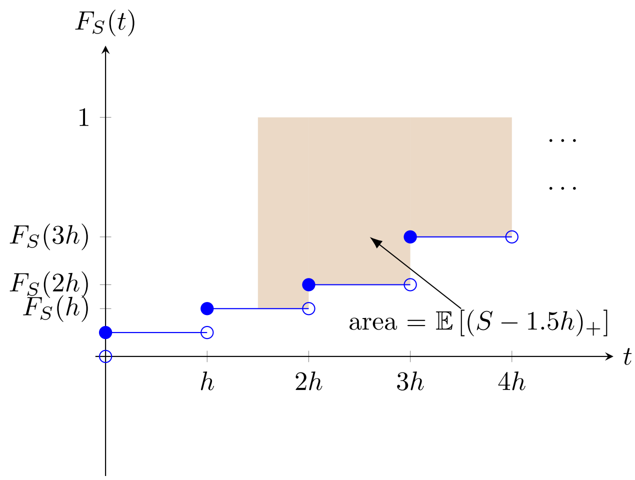
Remarks:

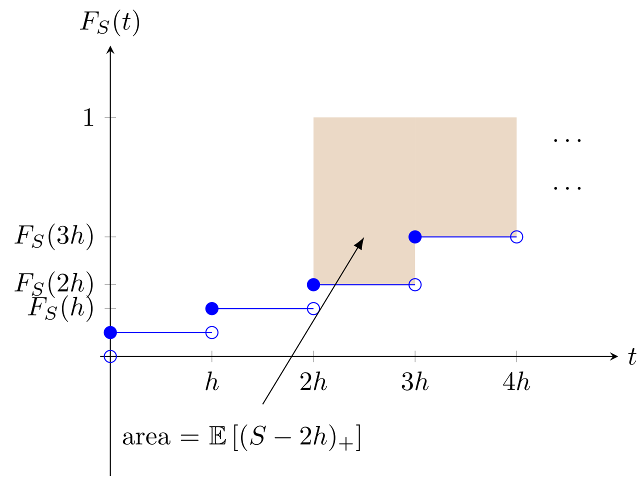
- Here we are not required to work in the collective risk model.

- The insurance "stops" the aggregate loss suffered by the policyholder (once d dollar is reached), hence "stop-loss insurance".
- If one uses the *equivalence principle* (without considering expenses) to price the insurance policy, the net premium obtained would be $\mathbb{E}\left[(S-d)_+\right]$ (the expected amount of benefit), hence hence "net stop-loss premium".

7.4.3 The key to compute the stop-loss premium is to utilize the methods introduced for computing the means of stop loss variables in Section 1. Particularly, the *geometrical approach* (finding area) is usually very helpful. For example, if the severity is discrete, then finding the stop-loss premium may look like the following, where the values of cdf's may be determined by Panjer's recursion.

[Note: This kind of method is also very helpful for computing various *risk measures* to be introduced in Section 8.]





8 Risk Measures

8.1 Introduction

8.1.1 For a loss, we use a nonnegative loss random variable X to measure it. Now, consider a *risk*, which is an exposure to possibility of loss. The main goal in this section is to study how to *measure* a *risk* (quantifying exposure/risk "level"). It turns out that, unlike a loss, there are multiple ways to measure a risk.

8.1.2 A **risk measure** is a function $\rho : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the set of all possible nonnegative loss random variables. For any loss random variable X, we assign a number $\rho(X)$ to it, which serves as a measure for the risk corresponding to that loss.

The value $\rho(X)$ is to be interpreted as the amount of *capital* needed for the insurer to *protect* against the loss X. Then, larger value of $\rho(X)$ indicates higher risk.

[Note: This interpretation is more natural if the function ρ satisfies certain properties, which will be discussed in Section 8.8. However, these properties are not *required* in the definition.]

8.2 Premium Principles

8.2.1 The first use of risk measures in actuarial science was the development of *premium principles* (e.g., equivalence principle in STAT3901). The premium principles are applied to the distribution of loss X for determining a suitable premium to charge for insuring the loss X.

8.2.2 Some examples of premium principles expressed in the language of risk measures are:

- **expected value premium principle**: $\rho(X) = (1 + k)\mathbb{E}[X]$ for some $k \geq 0$
- **standard deviation premium principle**: $\rho(X) = \mathbb{E}[X] + k\sqrt{\operatorname{Var}(X)}$ for some $k \geq 0$
- **variance premium principle**: $\rho(X) = \mathbb{E}[X] + k \operatorname{Var}(X)$ for some $k \geq 0$

[Note: To follow one of these premium principle, the premium to be charged for insuring the loss X is $\rho(X)$.
]

8.2.3 Each of these premium principles gives a premium that is *at least* the expected loss $\mathbb{E}[X]$. The excess amount serves as a "cushion" against adverse experience. Such excess amount is known as **premium loading**:

$$\text{premium loading} = \text{premium} - \mathbb{E}[X].$$

In the standard deviation and variance principles, the premium loading is related to the variance of the loss X:

- standard deviation premium principle: premium loading $= \alpha\sqrt{\operatorname{Var}(X)}$
- variance premium principle: premium loading $= \alpha \operatorname{Var}(X)$

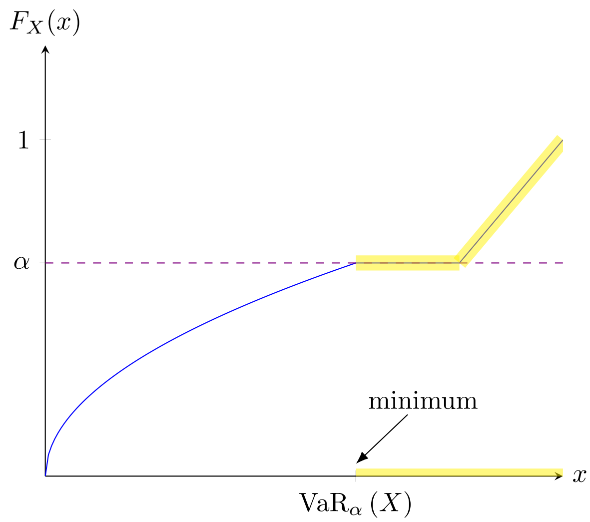
8.3 Value-at-Risk

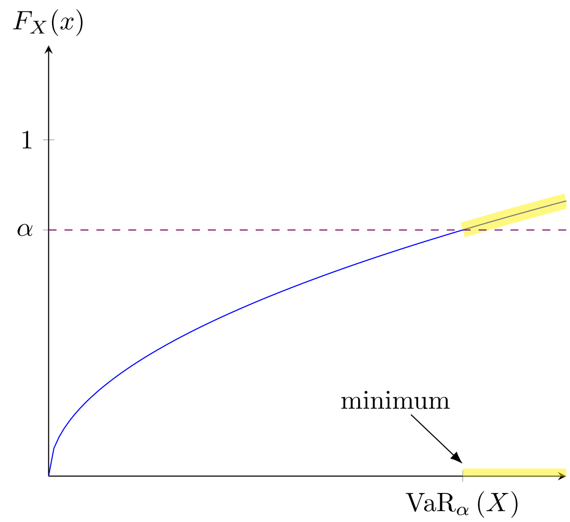
8.3.1 A popular risk measure is the *value-at-risk* (VaR). Let $\alpha \in (0, 1)$. [Warning: Here 0 and 1 are *excluded*!] The **value-at-risk at confidence level** α (or α**-VaR**) of a loss X is the αth quantile (or $100\alpha\%$ percentile) of X:

$$\operatorname{VaR}_\alpha(X) = \min\{x \geq 0 : F_X(x) \geq \alpha\}$$

where F_X is the cdf of X.

The following plots illustrate how to find the value-at-risk graphically in two cases: (i) the line $y = \alpha$ intersects with the cdf at infinitely many points and (ii) the line $y = \alpha$ intersects with the cdf at exactly one point.

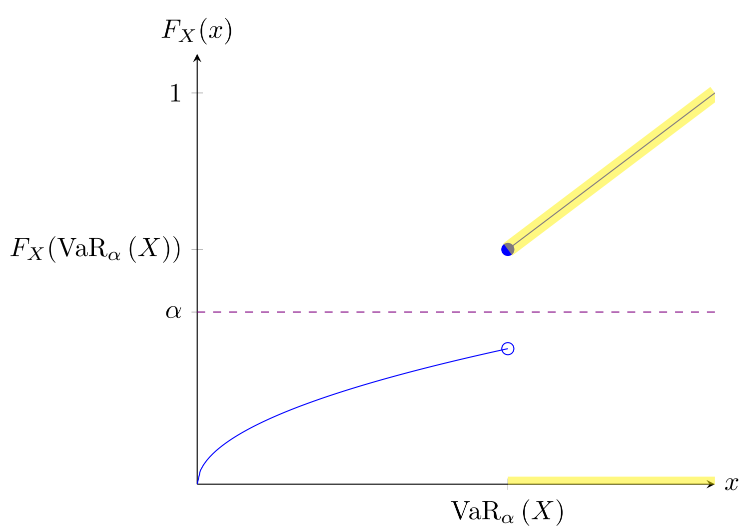




<u>Remarks</u>:

- In many (but not all!) cases, α-VaR is the value that is not exceeded by the loss X with probability α, i.e., $\mathbb{P}(X \le \mathrm{VaR}_\alpha(X)) = \alpha$. (See [8.3.2] for an example where this is not the case.)
- It can be interpreted as the amount of capital required to ensure that the loss can be "absorbed" by the insurer 🏢 (so that 🏢 would not bankrupt) with a high degree of certainty (when α is large).
- If F_X is continuous and strictly increasing, then the inverse function F_X^{-1} exists and we have $\mathrm{VaR}_\alpha(X) = F_X^{-1}(\alpha)$.

8.3.2 If F_X is continuous, then we always have $F_X(\mathrm{VaR}_\alpha(X)) = \alpha$. Thus, α-VaR is the value that is not exceeded by the loss X with probability α. However, this is not necessarily the case if F_X is not continuous:

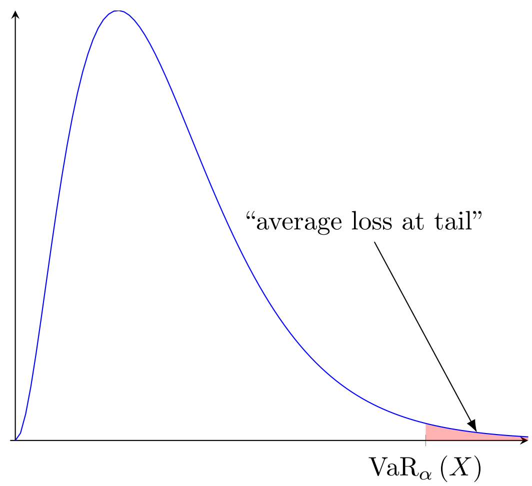


This is the third (and last) possible case: the line $y = \alpha$ does not intersect with the cdf. In this case, we have $F_X(\mathrm{VaR}_\alpha(X)) > \alpha$.

8.4 Conditional Tail Expectation

8.4.1 Let $\alpha \in (0,1)$. The **conditional tail expectation at confidence level α** (or **α-CTE**) is the expected loss given that the loss exceeds its α-VaR:

$$\mathrm{CTE}_\alpha(X) = \mathbb{E}[X | X > \mathrm{VaR}_\alpha(X)].$$



[Note: From [8.3.2], the probability $\mathbb{P}(X > \mathrm{VaR}_\alpha(X))$ is $1-\alpha$ if F_X is continuous, and it is not necessarily $1-\alpha$.]

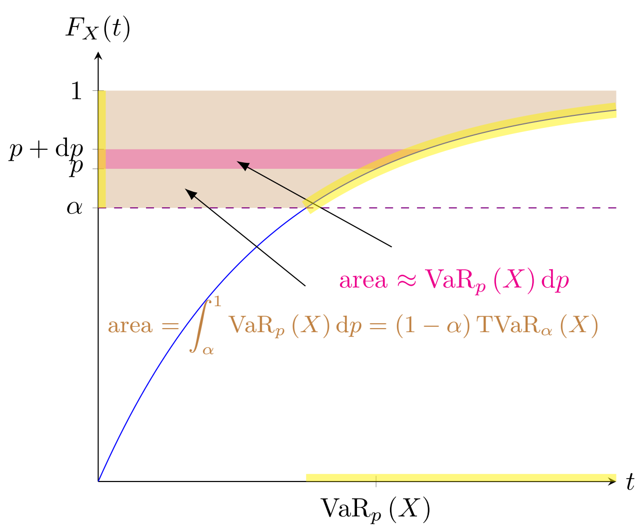
8.4.2 To compute conditional tail expectation, a typical approach is to utilize the relationship between CTE and *tail value at risk* (to be introduced in Section 8.5) as suggested in [8.7.1]b. Alternatively, one may directly compute the expectation if the distribution involved is "nice".

8.5 Tail Value-at-Risk

8.5.1 Let $\alpha \in (0,1)$. The **tail value-at-risk at confidence level α** (or **α-TVaR**) of the loss X is the "average" of the VaR at confidence level not less than α:

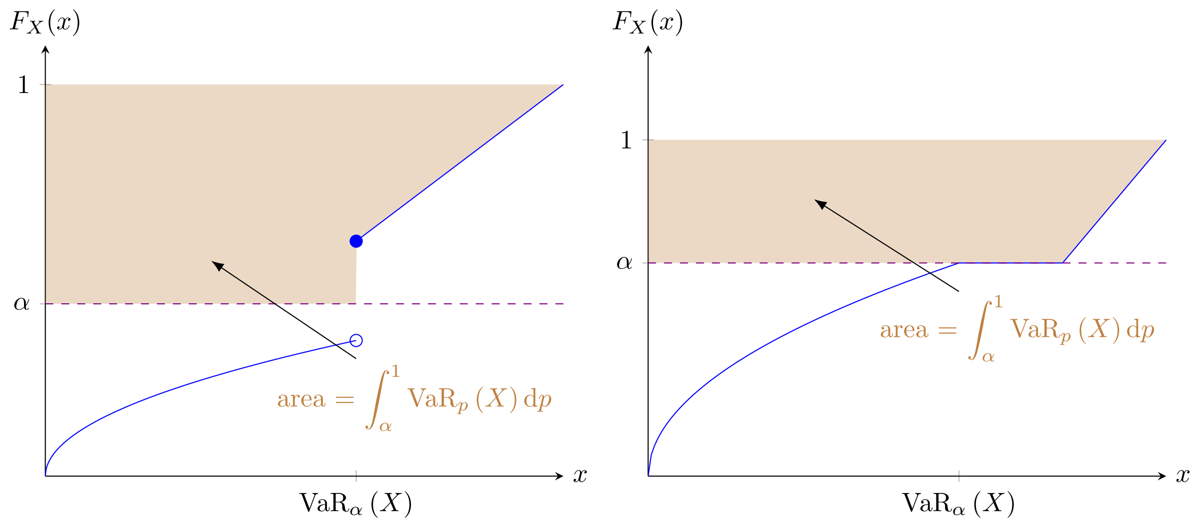
$$\mathrm{TVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_p(X)\,\mathrm{d}p\,.$$

Tail value-at-risk can be graphically illustrated as follows.



[Note: We consider VaR at confidence level in the "right tail".]

8.5.2 For the cases where the line $y=\alpha$ does not intersect with the cdf or intersects with cdf at infinitely many points, the areas look like the following.



8.5.3 Like finding stop-loss premium, the key to compute the α-TVaR is often to use the geometric approach, namely finding the area of the appropriate region, *and then divide the area by* $1-\alpha$ [Warning: Do not forget this step!].

8.6 Expected Shortfall

8.6.1 Before considering *expected shortfall*, let us introduce the concept of *shortfall* first. Here we assume that the capital reserve is calculated according to VaR at a certain confidence level α. Then the shortfall is given by

$$\text{shortfall} = \begin{cases} X - \mathrm{VaR}_\alpha(X) & \text{if } X > \mathrm{VaR}_\alpha(X), \\ 0 & \text{if } X \le \mathrm{VaR}_\alpha(X) \end{cases} = (X - \mathrm{VaR}_\alpha(X))_+$$

where X is the loss faced by the company. In words, the shortfall is the excess amount of loss over the α-VaR (capital reserved based on VaR) if the loss exceeds the α-VaR, and is zero otherwise (in this case the capital reserve is sufficient for covering the loss completely, so there is no "shortfall").

8.6.2 Then, as anticipated, the expected shortfall is just the expected value of shortfall introduced above. Let $\alpha \in (0,1)$. The **expected shortfall at confidence level α** (or **α-ESF**) of the loss X is the expected value of shortfall (with respect to α-VaR):

$$\mathrm{ESF}_\alpha(X) = \mathbb{E}\left[(X - \mathrm{VaR}_\alpha(X))_+\right].$$

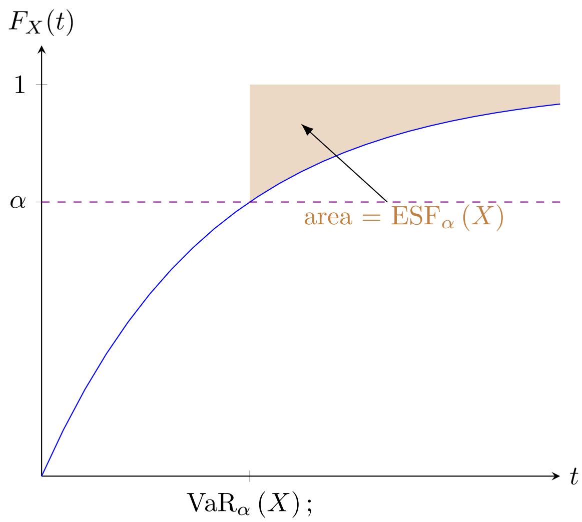
[Warning: The concepts of conditional tail expectation, tail value at risk, and expected shortfall can be defined very differently in other places. For example, the "expected shortfall" in some other places may mean what we call "tail value at risk", etc.]

Although expected shortfall is a function assigning a number to each possible nonnegative loss random variable (hence is a risk measure mathematically), it is not an "interpretable" one in the sense that the number should not be regarded as the amount of capital needed to protect against the loss X. It merely tells us how much our reserve (based on VaR) falls short when facing such loss, on average!

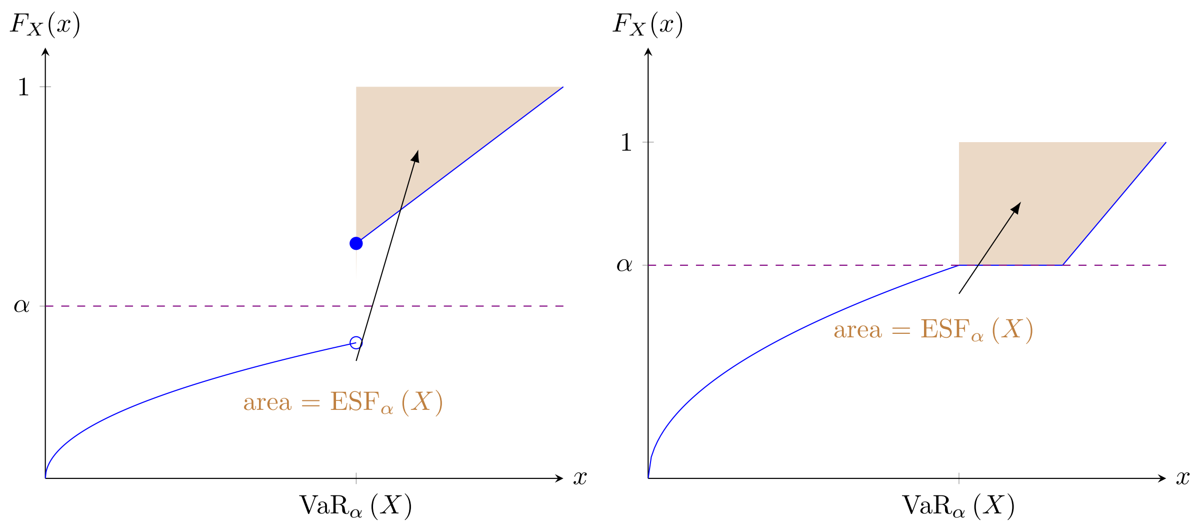
8.6.3 By Proposition 1.2.a, we can express $\mathrm{ESF}_\alpha(X)$ as:

$$\mathrm{ESF}_\alpha(X) = \int_{\mathrm{VaR}_\alpha(X)}^\infty [1 - F_X(x)]\,\mathrm{d}x\,.$$

We can represent the expected shortfall $\mathrm{ESF}_\alpha(X)$ geometrically as follows:



8.6.4 For completeness, we also consider the cases where the line $y = \alpha$ does not intersect with the cdf or intersects with cdf at infinitely many points here:

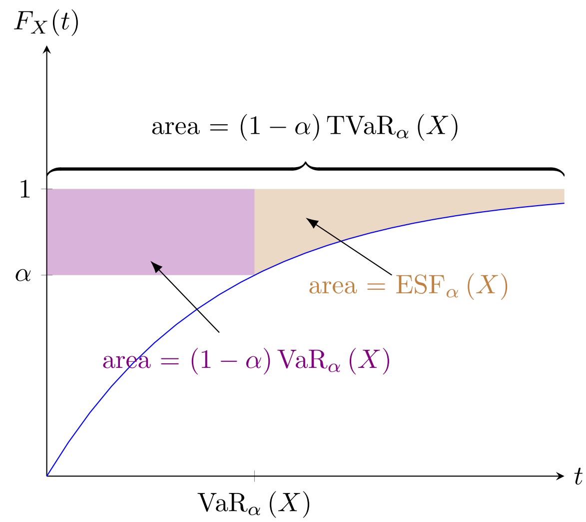


8.7 Relationships Between VaR, CTE, TVaR and ESF

8.7.1 For any $\alpha \in (0, 1)$, we have:

(a) $\mathrm{TVaR}_\alpha(X) = \boxed{\mathrm{VaR}_\alpha(X) + \frac{1}{1-\alpha} \cdot \mathrm{ESF}_\alpha(X)}$

Proof.



From this, we see that

$$(1-\alpha)\,\mathrm{TVaR}_\alpha(X) = (1-\alpha)\,\mathrm{VaR}_\alpha(X) + \mathrm{ESF}_\alpha(X),$$

which implies that

$$\mathrm{TVaR}_\alpha(X) = \mathrm{VaR}_\alpha(X) + \frac{1}{1-\alpha}\cdot \mathrm{ESF}_\alpha(X).$$

(The argument is similar when the cdf has other shape.) □

(b) $\mathrm{CTE}_\alpha(X) = \boxed{\mathrm{TVaR}_{F_X(\mathrm{VaR}_\alpha(X))}(X)}$ (when $0 < F_X(\mathrm{VaR}_\alpha(X)) < 1$).

Proof. Omitted. □

[Note: If $F_X(\mathrm{VaR}_\alpha(X)) = \alpha$ (which is the case when F_X is continuous), then we can simplify the formula in [8.7.1]b to:

$$\mathrm{CTE}_\alpha(X) = \boxed{\mathrm{TVaR}_\alpha(X)}.$$

This shows CTE and TVaR are equivalent in this case!]

Since [8.7.1]a should be obvious based on the geometric approaches for finding the risk measures, the more important relationship here is [8.7.1]b, which provides a "bridge" for applying the geometric approach to compute the CTE indirectly. Knowing this, we can see that the key 🔑 for computing risk measures is the geometric approach (finding area).

8.8 Coherent Risk Measures

8.8.1 When the risk measure ρ satisfies certain properties, the interpretation of $\rho(X)$ as amount of capital needed for the insurer to protect against the loss X is more natural.

8.8.2 The properties are as follows. [Note: The letters X and Y denote arbitrary nonnegative loss random variables (in the set $\mathcal{X}$) in the following.]

- **translation invariance** (TI):

$$\rho(X+c) = \rho(X) + c \quad \text{for any constant } c \in \mathbb{R}.$$

[Note: This means adding/subtracting an amount of loss implies addition/subtraction of the same amount to the required capital for protecting that loss.]

- **positive homogeneity** (PH):

$$\rho(\lambda X) = \lambda\rho(X) \quad \text{for any constant } \lambda > 0.$$

[Note: This means changing the unit of loss only leads to the corresponding unit change for the amount of capital required (but not change in the "actual" amount).[8]]

- **monotonicity** (M):

$$\mathbb{P}(X \le Y) = 1 \implies \rho(X) \le \rho(Y).$$

[Note: This means when a loss Y is not less than another loss X (with probability 1), the required capital for protecting the loss Y should also be not less than that for the loss X.]

- **subadditivity** (S):

$$\rho(X + Y) \le \rho(X) + \rho(Y).$$

[Note: This means diversification (adding/combining different losses together) cannot possibly make the resulting *total* amount of capital required greater than what we would have if there were no diversification.]

A risk measure satisfying TI, PH, M, and S is called a **coherent risk measure**.

8.8.3 The following summarizes the coherence of some common risk measures.

Risk measure	Coherent?
$\mathbb{E}[\cdot]$	yes[9]
expected value premium principle	no (fails TI) (when $k > 0$)
standard deviation premium principle	no (fails M) (when $k > 0$)
variance premium principle	no (fails M) (when $k > 0$)
VaR	no (fails S)
CTE	no (fails S)
ESF	no (fails S)
TVaR	yes[10]

[8]For example, suppose that the exchange rate is 1 USD = 160 JPY. Let X be the loss in USD. Suppose the amount of capital required in USD is $\rho(X) = 100$ (USD). Changing the currency unit to JPY, the loss becomes $160X$ (in JPY), and the capital required in JPY should be

$$160\rho(X) = 16000 \text{ (JPY)}$$

in order to have no change in the "actual" amount. Hence, we should have

$$\rho(160X) = 160\rho(X).$$

[9]TI, PH, and S follow from linearity of expectation. M follows from monotonicity of expectation.

[10]It turns out to be highly non-trivial to prove this.

References

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). *Loss models: From data to decisions* (5th ed.). John Wiley & Sons.

Sundt, B., & Jewell, W. S. (1981). Further results on recursive evaluation of compound distributions. *ASTIN Bulletin*, *12*(1), 27–39.

Concepts and Terminologies

Results

Section 1

- Proposition 1.1.a: formula for $\mathbb{E}[X]$ based on survival function
- Corollary 1.1.b: formula for $\mathbb{E}[X]$ based on survival function where X is discrete
- Proposition 1.2.a: formula for $\mathbb{E}[(X-d)_+]$ based on survival function
- [1.3.4]: relationship between MRL $e_X(d)$ and $\mathbb{E}[(X-d)_+]$
- [1.4.3]: relationship between stop loss and limited loss variables
- Proposition 1.5.a: existence of kth (positive) moment implies existence of all smaller positive moments
- [1.5.7]: equality of limit of ratio of survival functions and limit of ratio of pdfs
- Proposition 1.5.b: X has DFR (IFR) implies X has IMRL (DMRL)

Section 2

- [2.1.2]: expression of probability function from mixing finitely many random variables
- [2.1.4]: expression of probability function from mixing countably infinitely many random variables
- [2.1.5]: expression of probability function from mixing uncountably infinitely many random variables
- Proposition 2.2.b: expression of cdf from mixing random variables
- Theorem 2.2.a: law of total expectation
- Proposition 2.2.c: law of total variance

Section 3

- Proposition 3.1.a: probability generating property of pgf
- Corollary 3.1.b: same pgf implies same distribution
- Proposition 3.1.c: moment generating property of pgf
- Theorem 3.1.d: pgf of a sum of independent nonnegative discrete random variables
- [3.2.2]: pgf of $N \sim \text{Poi}(\lambda)$
- [3.2.3]: mean and variance of $N \sim \text{Poi}(\lambda)$
- Theorem 3.2.a: convolution for Poisson distribution
- Theorem 3.2.b: thinning for Poisson distribution
- [3.3.3]: pgf of $N \sim \text{NB}(r, \beta)$
- [3.3.4]: mean and variance of $N \sim \text{NB}(r, \beta)$
- Proposition 3.4.a: memoryless property of geometric distribution
- [3.5.2]: pgf of $N \sim \text{Bin}(m, q)$
- [3.5.3]: mean and variance of $N \sim \text{Bin}(m, q)$

Section 4

- Theorem 4.1.a: the only types of distributions in the $(a, b, 0)$ class
- [4.1.5]: characterization of $\text{Poi}(\lambda)$ using the $(a, b, 0)$ class
- [4.1.6]: characterization of $\text{NB}(r, \beta)$ using the $(a, b, 0)$ class
- [4.1.7]: characterization of $\text{Bin}(m, q)$ using the $(a, b, 0)$ class
- [4.3.2]: formula for constant c in the zero-modified $(a, b, 1)$ class
- [4.3.3]: formula for kth moment of a random variable in the zero-modified $(a, b, 1)$ class
- [4.3.4]: formula for pgf of a random variable in the zero-modified $(a, b, 1)$ class
- [4.3.6]: formulas for constant c and pgf of a random variable in the zero-truncated $(a, b, 1)$ class
- [4.4.6]: steps for computing $\text{ETNB}(r, \beta)$ probabilities

Section 5

- [5.1.3]: pgf of a compound random variable
- [5.1.4]: mean and variance of compound distribution
- Theorem 5.2.a: Panjer's recursion ($(a, b, 0)$ class)
- Theorem 5.2.b: Panjer's recursion ($(a, b, 1)$ class)
- Theorem 5.3.a: convolution for compound Poisson distribution

Section 6

- [6.2.4]: means of Y^L and Y^P in the case of franchise deductible
- Theorem 6.8.a: same family for N^P and N^L when N^L is in the $(a, b, 0)$ or zero-modified $(a, b, 1)$ class

Section 7

- [7.1.2]: mean and variance of S in the collective risk model
- [7.1.3]: formulas for aggregate payment in the collective risk model

Section 8

- [8.7.1]a: formula of TVaR in terms of VaR and ESF
- [8.7.1]b: formula of CTE in terms of TVaR
- [8.8.3]: coherence of some common risk measures