

3. Methods of Generating New Distributions

2024-05-20

Contents

1	Introduction	1
2	Identifying A New Distribution	2
2.1	General Frame Work	2
2.2	Generating Distribution With Hazard Rate Functions	3
2.2.1	Generalized Gompertz Hazard Rate Function	3
2.2.2	Generalized Weibull Hazard Rate Functions	5
3	Transforming A Random Variable	6
3.1	Transforming with No Additional Parameters	7
3.2	Transforming with Additional Parameters	7
4	Transforming A CDF	7
4.1	Family of Exponentiated CDFs	8
4.2	A New Extended Weibull Distribution	8
5	Transforming Involving Two or More Distributions	9
5.1	Convolution	10
5.2	Finite Mixture of Distributions	11
6	Compound Distributions (Optional)	14
6.1	Conditional and Marginal Distributions	14
6.2	n-fold Convolution	15
6.3	Motivation of Compounding Distribution	16
6.4	Compounding Distributions by Examples	16
6.4.1	Compounding Binomial Distribution	17
6.4.2	Compounding Poisson Distribution	17
6.5	Properties of Compound Distributions	18
6.5.1	The Law of Total Expectation and Variance	19
6.5.2	Properties of General Compound Distribution	20
6.6	Summary	20

1 Introduction

The standard 2-parameter Weibull distribution has a monotonic hazard rate. It is common to have diverse hazard rate patterns such as increasing and decreasing hazard rate (IDHR) and Bathtub (upside-down) bathtub hazard rates, or more flexible roller-coaster hazard rates. If none of the existing distributions can capture these hazard patterns, we must identify new distributions to accomplish new tasks.

There are three major approaches to obtaining new distributions:

1. Define a completely new distribution;

2. Transform an existing random variable (a function of a random variable) and use the distribution of the transformed random variable;
3. Expand an existing distribution by adding new parameters.

This note will use several simple examples to illustrate each of these methods and then expand the Weibull family to a new extended Weibull distribution. There are many extensions to the Weibull distribution, this will be a new addition to the existing big family.

2 Identifying A New Distribution

This approach seems to be easy but not straightforward. It is easy because any real-valued (only focusing on univariate random variables) function, say $f(x)$, that satisfies the following conditions could be the distribution of random variable X .

1. $f(x) \geq 0$;
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

However, it is not straightforward because it is hard to identify a desired distribution. For example, if we want to identify a distribution to capture a specific hazard pattern, we need to identify a function that captures the pattern of a hazard rate and use it to define the hazard rate function. Then convert the hazard function to the regular distribution functions.

2.1 General Frame Work

The following is a general framework of this approach.

Theorem: Let $h(x)$ be a non-negative real-valued function that has the desired pattern of the hazard rate associated random variable X . Then the pdf $f(x)$ of random variable X is expressed in terms of $h(x)$ in the following.

$$f(x) = h(x) \exp \left(- \int_0^x h(t) dt \right).$$

Proof: We first notice that

$$f(x) = F'(x) = [1 - S(x)]' = -\frac{dS(x)}{dx} = -S'(x).$$

Since

$$h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)}.$$

We integrate both sides of the above equation over $[0, x]$ and obtain

$$\int_0^x h(t) dt = - \int_0^x \frac{S'(t)}{S(t)} dt = - \int_0^x \frac{1}{S(t)} dS(t) = -\log S(x) + \log S(0).$$

Since $S(0) = 1$, therefore $\log S(0) = 0$. This implies that

$$S(x) = \exp \left(- \int_0^x h(t) dt \right)$$

Therefore,

$$F(x) = 1 - \exp\left(-\int_0^x h(t)dt\right)$$

Taking the derivative of the above equation, we have

$$f(x) = h(x) \exp\left(-\int_0^x h(t)dt\right).$$

Next, we use an example to illustrate the use of the above theorem.

Example. Consider an IDHR pattern associated with the random variable X . It is natural to use a quadratic function $h(x) = \alpha x^2 + \beta x$ to approximate the hazard rate with α and β being the population parameters (which will be estimated from given data in the future). Find the density function with the given hazard function $h(x)$.

Solution. From the above theorem, the density function of X has form

$$\begin{aligned} f(x) &= (\alpha x^2 + \beta x) \exp\left(-\int_0^x (\alpha t^2 + \beta t)dt\right) \\ &= (\alpha x^2 + \beta x) \exp\left[-\left(\frac{\alpha x^3}{3} + \frac{\beta x^2}{2}\right)\right]. \end{aligned}$$

2.2 Generating Distribution With Hazard Rate Functions

This subsection introduces several case studies that modify hazard functions and derive the corresponding extended distributions.

2.2.1 Generalized Gompertz Hazard Rate Function

Recall that the two-parameter Gompertz distribution has the following density function

$$f(x) = b\eta \exp(\eta + bx - \eta e^{bx})$$

where $\eta > 0$ and $b > 0$ are shape and scale parameters. The hazard rate function is given by

$$h(x) = \eta b e^{bx}.$$

The hazard rate function of the Gompertz distribution is increasing. However, if the hazard function is modified in the following

$$h_1(x) = \eta(2\alpha x + \beta) \exp(\alpha x^2 + \beta x),$$

It will give flexible patterns when α and β take different values. The following are hazard rate curves with different shapes.

```
hfun=function(a, b, d, x){
  d*(2*a*x+b)*exp(a*x^2+b*x)
}
##
par(mfrow=c(2,2), mar=c(2,2,2,1))
x=seq(0,10, length=200)
###
y1 = hfun(a= - 0.1, b=1.5, d = 1, x)
```

```

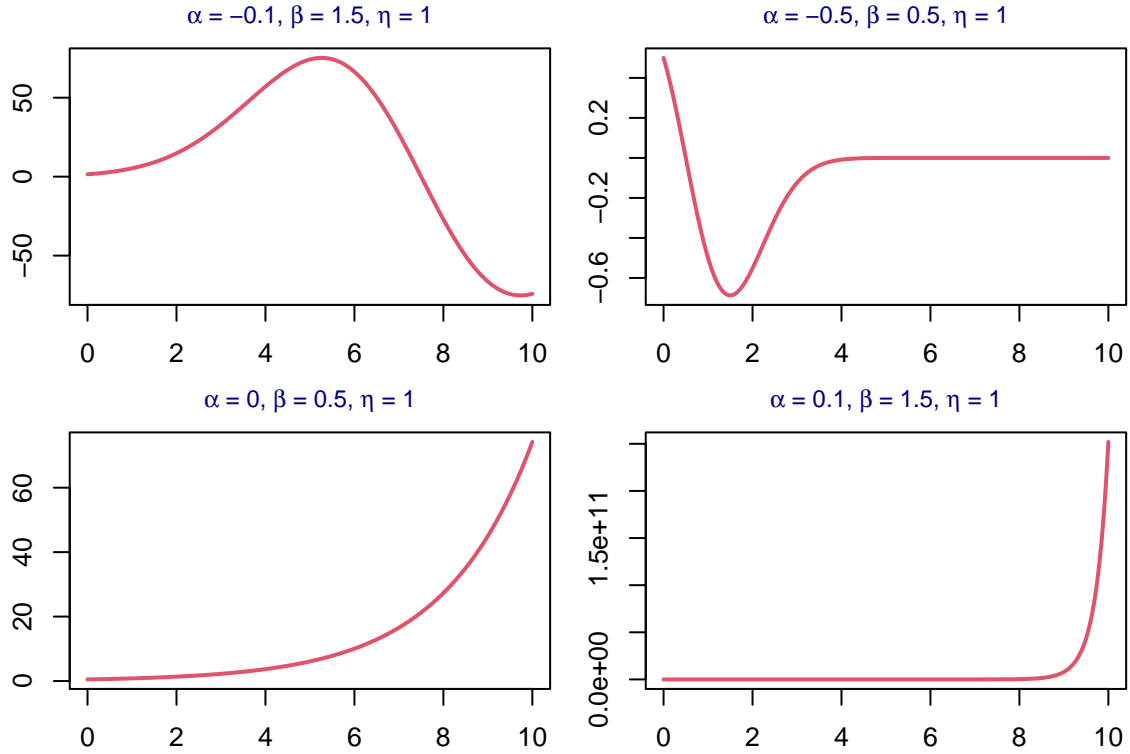
y2 = hfun(a= - 0.5, b=.5, d = 1, x)
y3 = hfun(a= - 0, b=0.5, d = 1, x)
y4 = hfun(a= 0.1, b=1.5, d = 1, x)
###
ymax = max(c(y1, y2, y3, y4))
###
plot(x, y1, type="l", lty=1, col = 2, lwd=2,
      main = expression(paste(alpha, " = -0.1, ", beta, " = 1.5, ", eta, " = 1")),
      cex.main = 0.9,
      col.main = "navy")

plot(x, y2, type="l", lty=1, col = 2, lwd=2,
      main = expression(paste(alpha, " = -0.5, ", beta, " = 0.5, ", eta, " = 1")),
      cex.main = 0.9,
      col.main = "navy")

plot(x, y3, type="l", lty=1, col = 2, lwd=2,
      main = expression(paste(alpha, " = 0, ", beta, " = 0.5, ", eta, " = 1")),
      cex.main = 0.9,
      col.main = "navy")

plot(x, y4, type="l", lty=1, col = 2, lwd=2,
      main = expression(paste(alpha, " = 0.1, ", beta, " = 1.5, ", eta, " = 1")),
      cex.main = 0.9,
      col.main = "navy")

```



Using the theorem, we have

$$f(x) = \eta(2\alpha x + \beta) \exp(\alpha x^2 + \beta x) \exp\left(-\int_0^x \eta(2\alpha t + \beta) \exp(\alpha t^2 + \beta t) dt\right)$$

$$= \eta(2\alpha x + \beta) \exp[\alpha x^2 + \beta x + \eta - \eta \exp(\alpha x^2 + \beta x)]$$

When $\alpha = 0$, the above density function is reduced to the standard Gompertz distribution. **Therefore, this is an extended Gompertz distribution.**

2.2.2 Generalized Weibull Hazard Rate Functions

Recall the two-parameter Weibull distribution with density function has the following form

$$f(x) = b k x^{k-1} e^{-b x^k}.$$

The corresponding hazard function is given by $b k x^{k-1}$ which is a monotonic function.

We can modify the above function to get a more flexible hazard. As an example, consider the modified form

$$h_1(x) = a b k x^{k-1} \exp(-a x^k)$$

where a, b , and k are positive. The following hazard curves shows the more flexible hazard patterns.

```
## write a function to evaluate h(x) for given x and values of parameters
hval=function(a,b, k, x){
  a*b*k*x^(k-1)*exp(-a*x^k)
}

## pre-selected x-values
x = seq(0, 6, length = 200)[-1]
##
## Evaluate y with different combinations of parameters
y1 = hval(a = 0.5, b=.5, k=0.5, x)
y2 = hval(a = 1, b=1, k=1.5, x)
y3 = hval(a = 1.5, b=1.5, k=1.5, x)
y4 = hval(a=0.1, b=.5, k = 3, x)
## range of y
ymax=max(c(y1, y2, y3, y4))
##
plot(x, y1, type="l", ylim=c(0,ymax), xlim=c(0,6),
     xlab = "Modified Weibull score",
     ylab = "hazard rate",
     col= "#8B0000",
     lwd=2,
     main = "shapes of hazard rate curves",
     cex.main = 0.9, # font size of color
     col.main = "navy")
lines(x, y2, col="#46008B", lwd=2,lty=2)
lines(x, y3, col="darkgreen", lwd=2,lty=3)
lines(x, y4, col="coral", lwd=2,lty=4)

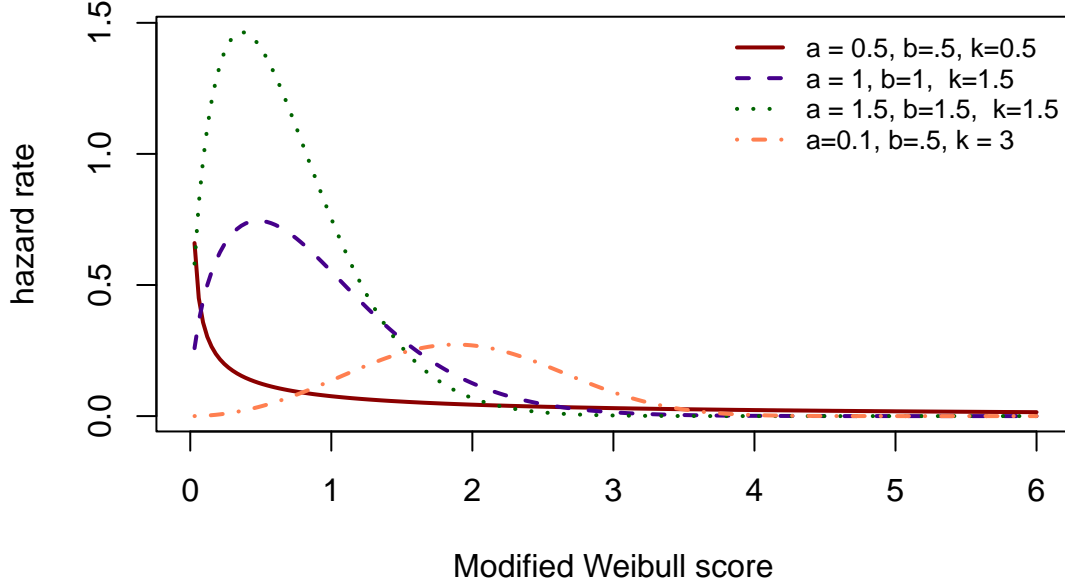
## Add legend to the plot
legend("topright", c("a = 0.5, b=.5, k=0.5", "a = 1, b=1, k=1.5",
```

```

"a = 1.5, b=1.5, k=1.5", "a=0.1, b=.5, k = 3"),
col=c("#8B0000", "#46008B", "darkgreen", "coral"),
lwd=rep(2,4), lty=1:4, cex = 0.8, bty="n")

```

shapes of hazard rate curves



Using the above theorem, the density function of the modified Weibull is given by

$$\begin{aligned}
 f(x) &= h_1(x) \exp \left(- \int_0^x h(t) dt \right) \\
 &= abkx^{k-1} \exp(-ax^k) \exp \left(- \int_0^t bkt^{k-1} \exp(-at^k) dt \right) \\
 &= abkx^{k-1} \exp \left[-ax^k + b - b \exp(-ax^k) \right].
 \end{aligned}$$

When $a = 0$, the above modified Weibull reduces to the standard two-parameter Weibull distribution.

3 Transforming A Random Variable

The transformation of an existing random variable with a given density will also define a new distribution. For lifetime random variables, the transformation needs to be strictly monotonic so that the old random variable can be re-expressed in terms of the new transformed random variable.

Let X be an existing random variable with CDF $F(x, \theta)$ and $Y = g(X)$ a transformation of X , where $g(x)$ is an inverse function. Then the CDF of Y , denoted by $G(y, \theta)$, is given by

$$G(y, \theta) = P(Y < y) = P[g(X) < y] = P[X < g^{-1}(y)] = F[g^{-1}(x), \theta].$$

The above monotonic transformation could carry additional parameters to make the new distribution more flexible. This method is also called the CDF approach.

3.1 Transforming with No Additional Parameters

Consider the well-known Weibull distribution with density function

$$f(x) = b k x^{k-1} e^{-b x^k}, \quad x > 0.$$

The corresponding CDF is given by

$$F(x) = \int_0^x b k t^{k-1} e^{-b t^k} dt = e^{-b t^k} \Big|_0^x = 1 - e^{-b x^k}.$$

Consider transformation $Y = \sqrt{X}$. What is the density distribution of Y ? Using the CDF approach, we have

$$G(y) = P(Y < y) = P(\sqrt{X} < y) = P(X < y^2) = 1 - e^{-b y^{2k}}.$$

The PDF is given by

$$g(x) = 2b y^{2k-1} e^{-b y^{2k}}.$$

This new distribution still has two parameters. It is also not the 2-parameter Weibull distribution. We can see that the hazard rate function is similar to that of the 2-parameter Weibull distribution.

3.2 Transforming with Additional Parameters

We still use the 2-parameter Weibull as an example. Consider transformation $Y = e^{aX+1}$, The CDF of Y can be found in the following/

$$G(y) = P(Y < y) = P(e^{aX+1} < y) = P(X < \frac{\log y - 1}{a}) = 1 - \exp \left[-b \left(\frac{\log y - 1}{a} \right)^{2k} \right].$$

The transformation brings a new parameter to the transformed distribution. The resulting distribution is a 3-parameter distribution.

4 Transforming A CDF

We have introduced the approach to using the hazard rate function to find distribution in lifetime distribution. This section introduces a general approach to expanding any distribution by adding new parameters based on the CDF such that the new distribution includes the old distribution as a special case.

The general idea is to use the basic properties of CDF $F(x)$.

1. $F(x)$ is non-decreasing;
2. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.

With the above properties in mind, we can transform $F(x)$ to make a new CDF. To be more specific, let $g(x)$ be a function such that

$$G(x) = g[F(x)]$$

meets the following conditions:

1. $G(x)$ is non-decreasing;
2. $\lim_{x \rightarrow -\infty} G(x) = 0$ and $\lim_{x \rightarrow \infty} G(x) = 1$.

Clearly, $G(x)$ is also a legitimate CDF.

4.1 Family of Exponentiated CDFs

As an example, we consider the 2-parameter Weibull distribution with CDF

$$F(x) = 1 - e^{-bx^k}, x > 0$$

Define

$$G(x) = [F(x)]^\theta = \left[1 - e^{-bx^k}\right]^\theta, \quad \theta > 0.$$

It is easy to verify that the above conditions are satisfied. That is, $G(x)$ is an expanded Weibull (when $\theta = 1$, $G(x)$ is reduced to $F(x)$).

The survivor and related reliability functions of the new distribution can be similarly defined.

4.2 A New Extended Weibull Distribution

We first introduce a new extended family of distributions. Let X be a random variable with CDF $F(x)$ and denote $S(x) = 1 - F(x)$. Let θ be a new parameter satisfying $0 \leq \theta \leq 1$. Then

$$G(y) = F(y)[1 - \theta S(y)]$$

is a valid CDF of a new random variable Y . This can be easily verified by checking the conditions required for a CDF.

1. $G(y)$ is increasing: $G'(y) = f(y)[1 - \theta S(y)] + F(y)\theta f(y) = f(y)[1 - \theta] + \theta f(y)F(y) > 0$.
2. Note that $\lim_{y \rightarrow -\infty} F(y) = 0$ and $\lim_{y \rightarrow \infty} F(y) = 1$. Therefore, $\lim_{y \rightarrow -\infty} G(y) = 0(1 - \theta \times 0) = 0$ and $\lim_{y \rightarrow \infty} G(y) = 1(1 - \theta \times 0) = 1$. Hence, $G(y)$ is the valid distribution of random variable Y .

The pdf is given by

$$g(y) = f(y)[1 - \theta] + \theta f(y)F(y) = f(y)[1 - \theta S(y)].$$

The survivor function is given by

$$S(y) = 1 - G(y) = 1 - F(y)[1 - \theta S(y)] = [1 + \theta F(y)]S(y)$$

Therefore, the hazard rate function is defined by

$$\begin{aligned} h_G(y) &= g(y)/S(y) = \frac{f(y)[1 - \theta S(y)]}{[1 + \theta F(y)]S(y)} \\ &= h_F(y) \left[1 - \frac{\theta}{1 + \theta F(y)} \right]. \end{aligned}$$

Therefore, adding the new parameter changes the hazard rate patterns of the original distribution.

A New Extended Weibull Distribution flexible Hazard Rates

Let $F(x)$ be the 2-parameter Weibull CDF and θ be a parameter between 0 and 1. Then

$$G(y) = F(y)[1 - \theta S(y)] = (1 - bky^{k-1}e^{-by^k})[1 - \theta e^{-by^k}].$$

This defines a new distribution that contains the two-parameter Weibull as a special case.

Since no work has ever been done about this distribution (or model), we can study the following properties (not limited to...)

- **Mathematical Property Analyses**
 - Density shape analysis
 - Survival and hazard rate analyses
 - Turning points of hazard rate
 - Entropy
 - Other Risk Measures (such as actuarial measures like value at risk - VaR, etc.)
- **Inferential Analyses** (*most of the following will be addressed using examples in the subsequent notes*)
 - MLE of parameters
 - Asymptotic properties of the MLE (estimated standard error of MLE)
 - Goodness of fit and significant tests
 - Confidence intervals of hazard rate turning points.
- **Numerical Experiments** (*most of the following will be addressed using examples in the subsequent notes*)
 - Simulation studies (coverage probabilities)
 - Identifying turning points of hazard rates
 - Likelihood ratio and significance test of the new parameter $H_0 : \theta = 0$ vs $H_a : \theta \neq 0$.

5 Transforming Involving Two or More Distributions

There are cases in which a newly generated distribution may involve two or more random variables. A simple example in elementary mathematical statistics is the mean of a set of IID (**independently and identically distributed**) normal random variables $\{X_1, X_2, \dots, X_n\}$ from $N(\mu, \sigma^2)$. We consider the new random variable defined to be of

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

We can easily (?) show that \bar{X} is also a normal random variable. This section will outline the transformations with two or more random variables (try this proof using the MGF method.). For ease of illustration, we focus on the two special cases that involve two **non-IID** () random variables: convolution and compounding.

5.1 Convolution

The mathematical term **convolution** is an operation of producing a new function $h(x)$ based on two existing functions, say $f(\cdot)$ and $g(\cdot)$ in the following integral transformation

$$h(z) = (f \otimes g)(z) = \int_D f(x)g(z-x)dx.$$

D is the domain of $f(\cdot)$ and $g(\cdot)$. \otimes is called the convolution operator.

In probability, we can convolution operation to define a new variable based on existing ones. To derive the distribution, we need the following more general result.

Theorem (Section 6.6 of the textbook): Suppose that Y_1 and Y_2 are continuous random variables with **joint density function** $f_{Y_1, Y_2}(y_1, y_2)$. Define bivariate transformation

$$u_1 = h_1(y_1, y_2) \quad \text{and} \quad u_2 = h_2(y_1, y_2)$$

to be one-to-one from (y_1, y_2) to (u_1, u_2) and assume further that the inverse transformation is

$$y_1 = h^{-1}(u_1, u_2) \quad \text{and} \quad y_2 = h^{-1}(u_1, u_2).$$

If both inverse transformations have partial derivatives with respect to u_1 and u_2 and the **Jacobian** matrix is not singular, that is,

$$J = \det \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \frac{\partial h_1}{\partial u_2} \\ \frac{\partial h_2}{\partial u_1} & \frac{\partial h_2}{\partial u_2} \end{vmatrix} \neq 0,$$

then the joint density of U_1 and U_2 is

$$f_{u_1, u_2}(u_1, u_2) = f_{Y_1, Y_2}[h^{-1}(u_1, u_2), h^{-1}(u_1, u_2)] \times |J|,$$

$|J|$ is the absolute value of the Jacobian.

Remark: The above theorem can be generalized to the multivariate case.

To be more specific,

let X be a random variable with density function $f(x)$ and Y be a random variable with density $g(y)$. We define a new random variable $Z = X + Y$. The CDF of Z is given by

$$F(z) = P(Z < z) = P(X + Y < z) = P(Y < z - X) = \int_{-\infty}^{z-X} g(t)dt$$

Convolution Distribution: Let X and Y be two independent random variables with density functions $f(x)$ and $g(y)$, respectively. Define $Z = X + Y$, the density function of Z is given by

$$h_Z(z) = \int_{-\infty}^{\infty} f(z-y)g(y)dy.$$

To use the above theorem, we define a bivariate transformation

$$Z = X + Y \quad \text{and} \quad Y = Y.$$

The inverse transformation is

$$X = Z - Y \quad \text{and} \quad Y = Y$$

The Jacobian matrix is

$$J = \det \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

Therefore, the **joint density** of U and Y is given by

$$h(z, y) = f(z - y)g(y)|J|$$

The distribution of Z is given by

$$h_Z(z) = \int_{-\infty}^{\infty} f(z - y)g(y)dy$$

Example: Let X and Y be Weibull and Gamma distribution respectively with density functions

$$f(x) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} y^{\alpha_1-1} e^{-\beta_1 y}. \quad \text{and} \quad g(y) = \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} y^{\alpha_2-1} e^{-\beta_2 y}.$$

Then the distribution of $Z = X + Y$ has density function

$$\begin{aligned} h(z) &= \int_{-\infty}^{\infty} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} (z - y)^{\alpha_1-1} e^{-\beta_1(z-y)} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} y^{\alpha_2-1} e^{-\beta_2 y} dy \\ &= \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1) \Gamma(\alpha_2)} \int_{-\infty}^{\infty} (z - y)^{\alpha_1-1} y^{\alpha_2-1} e^{-\beta_1(z-y) - \beta_2 y} dy \end{aligned}$$

The convolution of two density functions gives a new distribution. Statistically, it is a very special bivariate transformation of two independent random variables. As stated earlier, one can generate the bivariate transformation to multivariate transformation to combine more distributions including dependent random variables.

5.2 Finite Mixture of Distributions

Mixture modeling is another way to combine two or more distributions for cases in which data may come from multiple populations with different distributional patterns of certain numeric characteristics of interest.

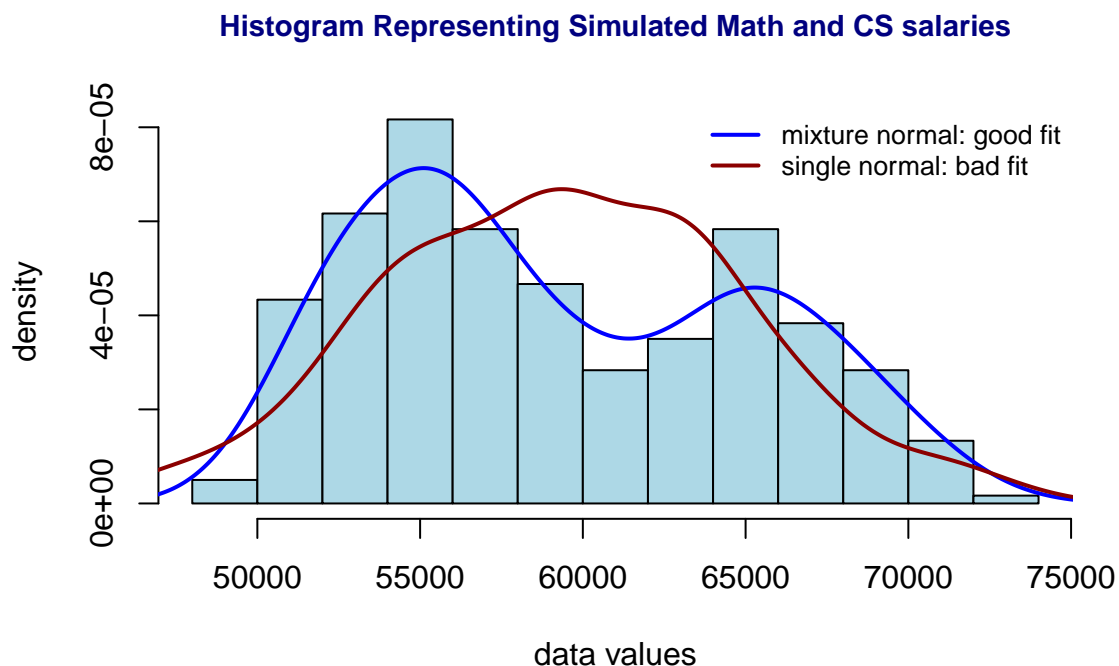
For example, consider studying the distribution of starting salaries of recent graduates in the Department of Mathematics and Computer Science at a university. Assume that the department has only two majors, mathematics and computer science. The distribution of salaries is commonly assumed to have a normal distribution in a homogeneous population. If we assume all starting salaries of recent graduates in the department follow the sample normal distribution, we only need to know (estimate) the mean (μ) and variance (σ^2) to uniquely determine the distribution. However, due to the nature of the two majors, the starting salaries of CS majors are higher than that of math majors, so fitting a single normal distribution to the data from a heterogeneous population may lead to a poor fit. See the following figure.

```

# This function generates data from two normal distributions that allow users
# to choose the values of different population parameters and mixing proportions.
mix2Norm <- function(n, mu1, sig1, mu2, sig2, mixp){
  # n = sample size;
  # mixp = mixing proportion
  # mu1, sig1: mean and standard deviation of population 1
  # mu2, sig2: mean and standard deviation of population 2
  set.seed(666)
  y0 <- rnorm(n, mean=mu1, sd = sig1)
  y1 <- rnorm(n, mean=mu2, sd = sig2)
  #
  flag <- rbinom(n, size=1, prob=mixp)
  y <- y0*(1 - flag) + y1*flag
  #
  y
}

## Plot distributions of correct and specified distributions fit to the data
dataset= mix2Norm(300,55000, 2500, 65000, 3500, .45)
hist(dataset, probability = TRUE,
      col = "lightblue",
      main = "Histogram Representing Simulated Math and CS salaries",
      cex.main = 0.9,
      col.main = "navy",
      xlab = "data values",
      ylab = "density")
lines(density(dataset), lwd=2, col = "blue")
lines(density(rnorm(300, mean(dataset), sd(dataset))), lwd=2, col = "darkred")
legend("topright", c("mixture normal: good fit", "single normal: bad fit"),
      lwd=rep(2,2), col=c("blue", "darkred"), bty="n", cex= 0.8)

```



The above figure is based on a simulated data set assuming that 45% are CS majors with a mean salary \$55,000 and a standard deviation of \$3500 and 55% are math majors with a mean salary \$55,000 and a standard deviation \$2500. We can see that a single normal distribution does not fit well.

The question is how to find a distribution that captures the heterogeneity of the salaries due to **two** significantly different groups in the population. A natural approach is to combine the **two distributions** by taking the weighted average. If we know the proportion (α) of math and CS majors, the weight can be defined as the proportion. *Note that the sum of all weights must be 1.*

To proceed with mathematical and statistical analysis, we need a **crystal analytic formulation** of the vague problem of **weighted average of two distributions**. That is, the weighted average of two random variables? density functions? or something else?

- Case I: $Z = \alpha X + (1 - \alpha)Y$ is not an choice. This requires a bivariate transformation of random variables.
- Case II: $h(x) = \alpha f(x) + (1 - \alpha)g(x)$ - the weighted average of the two density functions. This is a natural and interpretable approach. We can easily show that $h(x)$ is a valid density function ($0 < \alpha < 1$).
- CASE III: $H(x) = \alpha F(x) + (1 - \alpha)G(x)$ - This is also valid and equivalent to Case II.

Definition: Let $f(x)$ and $g(x)$ be two density functions representing the distribution of random variables X_1 and X_2 . Then $h(x) = \alpha f(x) + (1 - \alpha)g(x)$ is a valid density and called the **mixture of $f(x)$ and $g(x)$** . $0 < \alpha < 1$ is called the **mixing parameter**.

Remarks

1. α does not have to be given just like in the salary example. It could also be an unknown parameter to be estimated from the data. This is why it is called *mixing parameter*.
2. The above definition can be generalized to a fixed number of heterogeneous distributions with densities $f_i(x)$ and corresponding mixing parameters α_i for $i = 1, 2, \dots, k$ to obtain a new distribution with density function

$$h(x) = \sum_{i=1}^k \alpha_i f_i(x).$$

where $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^k \alpha_i = 1$.

• Mixture Models are Machine Learning Algorithms

- If the number of heterogeneous distributions is **known**, the objective is to determine each individual distribution. This method (also referred to as an algorithm) is called **supervised machine learning algorithm**.
- If the number of heterogeneous distributions is **unknown** (but countable), the objective is to determine the number of distributions and each individual distribution. This method is called **unsupervised machine learning algorithm**.

Example: The above starting salary example using the following two-component normal mixture model with density function

$$f(x) = 0.55 \times \frac{1}{\sqrt{2\pi} \times 2500} \exp \left[-\frac{(x - 55000)^2}{2 \times 2500^2} \right] + 0.45 \times \frac{1}{\sqrt{2\pi} \times 3500} \exp \left[-\frac{(x - 65000)^2}{2 \times 3500^2} \right].$$

6 Compound Distributions (Optional)

In the convolution method (multi-variate transformation) and mixture distributions, the number of the involved distributions is either known or unknown but fixed. That is, $\{X_1, X_2, \dots, X_N\}$ is the set of random variables, we have outlined the distribution of

$$Y = X_1 + X_2 + \dots + X_N,$$

where N is a fixed number (could be unknown). However, if the number of the involved distributions is uncertain (i.e., random), for example, the number of distributions follows a Poisson distribution. How to determine the distribution of Y ?

To answer this question, we need another random variable to characterize the random number of distributions to pose a condition on the number of distributions in $Y = X_1 + X_2 + \dots + X_N$.

The process of making conditional distributions is illustrated in the following (thinking about the aggregated claims $Y = \sum X_i$ of an insurance policy);

Case 1 if $N = 0$, then $Y = 0$. As a convention, set $F(x) = 1$.

Case 2 if $N = 1$, then $P(Y < y) = P(Y < y|N = 1) \times P(N = 1) = P(X_1 < y) \times P(N = 1)$

Case 3 if $N = 2$, then $P(Y < y) = P(Y < y|N = 2) \times P(N = 2) = P(X_1 + X_2 < y) \times P(N = 2)$

Case 4 if $N = 3$, then $P(Y < y) = P(Y < y|N = 3) \times P(N = 3) = P(X_1 + X_2 + X_3 < y) \times P(N = 3)$

.....

Case k if $N = k$, then $P(Y < y) = P(Y < y|N = k) \times P(N = k) = P(X_1 + X_2 + \dots + X_k < y) \times P(N = k)$

.....

Combining the above cases, we have the CDF of the distribution of $Y = X_1 + X_2 + \dots + X_N$ as

$$F(y) = \sum_{k=0}^{\infty} P(Y < y|N = k) \times P(N = k) = \sum_{k=0}^{\infty} P\left(\sum_{i=1}^k X_i < y\right) P(N = k),$$

where $\{X_1, X_2, \dots, X_N\}$ are IID random variables such exponential random variables and N follows discrete distribution such as Poisson or binomial distributions. To better understand the above general definition of compound distribution, we outline the concepts of marginal and conditional distributions and k-fold convolution in the subsequent subsections.

6.1 Conditional and Marginal Distributions

When working with two or more distributions, it is inevitable to have a mathematical understanding of concepts such as independence, conditional, and marginal distributions. Next, we use random variables X and Y to explain these concepts. Assume that X and Y have joint density functions $f(x, y)$ (i.e., X and Y are assumed to be not independent). **The textbook covers these concepts in Chapter 5.**

Marginal Distributions of X and Y : The two marginal distributions of X and Y are defined respectively by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Independence of X and Y If X and Y are independent **if and only if** the joint density function can be decomposed as the product of the two marginal density functions. That is,

$$f(x, y) = f_X(x) \times f_Y(y)$$

Conditional Distributions If X and Y are jointly continuous random variables with joint density function $f(x, y)$, then the conditional distribution function of X given $Y = y$ is

$$F(x|y) = P(Y \leq x|Y = y).$$

To find the unconditional CDF, we integrate y out from the above expression and obtain

$$F(x) = \int_{-\infty}^{\infty} F(x|y) f_Y(y) dy.$$

On the other hand, $f_X(x)$ is the marginal density of $f(x, y)$. That is

$$\begin{aligned} F(x) &= \int_{-\infty}^x f_X(t) dt = \int_{-\infty}^x \int_{-\infty}^{\infty} f(t, y) dy dt \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^x f(t, y) dt dy. \end{aligned}$$

That is,

$$F(x|y) f_Y(y) = \int_{-\infty}^x f(t, y) dt.$$

Taking derivative on both sides of the above expression, we have

$$f(x|y) f_Y(y) = f(x, y)$$

Therefore, the conditional density function of (X, Y) conditioning on Y is given by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

6.2 n-fold Convolution

The example in the opening paragraph of this section involves the conditional distribution $P(Y < y|N = n) = P(X_1 + X_2 + \dots + X_n < y)$. The density of $X_1 + X_2 + \dots + X_n$ requires the concept and notation of n -fold convolution.

In this subsection, we consider the case that $\{X_1, X_2, \dots, X_N\}$ is a set of IID random variables with common CDF $F(x)$. Recall that the convolution distribution of $Y = X_1 + X_2$

$$(f \otimes f)(y) \equiv f^{(2\otimes)}(x) = \int_{-\infty}^{\infty} f(y - t) f(t) dt.$$

$f^{(2\otimes)}(x)$ is called 2-fold convolution of $f(x)$. In general, we define **k-fold convolution of $f(x)$** recursively as

$$f^{[n\otimes]}(x) = \left(f^{[(n-1)\otimes]} \otimes f \right)(x), \quad f^{[1\otimes]}(x) = f(x), \quad \text{and} \quad f^{[0\otimes]}(x) = 0.$$

We can also re-express the convolution density function into the CDF form as

$$\begin{aligned}
F^{(2\otimes)}(z) &\equiv P[X + Y < z] = \iint_{x+y < z} f(x)f(y)dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x)f(y)dy dx. \\
&= \int_{-\infty}^{\infty} f(y) \left(\int_{-\infty}^{z-y} f(x)dx \right) dy = \int_{-\infty}^{\infty} F(z-y)f(y)dy.
\end{aligned}$$

By using integral by parts, we have

$$F^{(2\otimes)}(z) = \int_{-\infty}^{\infty} F(z-y)f(y)dy = \int_{-\infty}^{\infty} F(z-y)dF(y) = \int_{-\infty}^{\infty} F(y)f(z-y)dy.$$

We denote $F^{0\otimes} = 1$ and $F^{1\otimes} = F(x)$. we can then define k-fold convolution in terms of CDF in the following

$$F^{[n\otimes]}(z) = \left(F^{[(n-1)\otimes]} \otimes f \right) (z) = \int_{-\infty}^{\infty} F^{[(n-1)\otimes]}(z)f(z-y)dy,$$

where

$$F^{[1\otimes]}(z) = F(z), \quad \text{and} \quad F^{[0\otimes]}(z) = 1.$$

6.3 Motivation of Compounding Distribution

The compounding method is used to define the distribution explicitly in the following.

Definition: The random variable Y is said to have a compound distribution if Y is of the following form

$$Y = X_1 + X_2 + \cdots + X_N,$$

where

- 1). The number of terms N is uncertain,
- 2). The random variables X_i are independent and identically distributed (with common distribution X) and
- 3). each X_i is independent of N.

Real-world Connections: Why compounding distribution and how it is used in practice? Consider the aggregated claim of an insurance policy. Let $\{X_1, X_2, \dots, X_N\}$ be the amounts of claims within the policy period (e.g., one year). Assume the amount of the claims has the same distribution. $Y = X_1 + X_2 + \cdots + X_N$ is the aggregated amount of N claims.

- 1). N is random: think about auto insurance, the number of accidents within the policy period (usually 1 year) is random!
- 2). X_1, X_2, \dots, X_N are assumed to have identical distribution. Using the same auto insurance example, the claims amount of one claim is apparently independent of the other.
- 3). It is also obvious that the claim amount X_i is independent of how many accidents have occurred and how many will happen.

6.4 Compounding Distributions by Examples

The following two illustrative examples illustrate the steps for finding new distributions through compounding.

6.4.1 Compounding Binomial Distribution

Consider an auto insurance policy that is modeled by a binomial distribution $N \sim \text{Bin}(2, p)$ and claim amounts follow an exponential distribution $\exp(\lambda)$ with CDF $X_i \sim F(x) = 1 - e^{-\lambda x}$.

The objective is to find the distribution of $Y = X_1 + \dots + X_N$.

Since $N \sim \text{Bin}(2, p)$,

- $P(N = 0) = (1 - p)^2$,
- $P(N = 1) = \frac{2!}{1!(2-1)!}p(1 - p) = 2p(1 - p)$, and
- $P(N = 2) = \frac{2!}{2!(2-2)!}p^2 = p^2$.

Note that

- $F^{[0\otimes]}(y) = 1$,
- $F^{[1\otimes]}(y) = F(y) = 1 - e^{-\lambda y}$, and
- The 2-fold convolution of F and F is given by

$$\begin{aligned} F^{[2\otimes]}(y) &= \int_0^y (1 - e^{-\lambda(y-x)})\lambda e^{-\lambda x} dx \\ &= \int_0^y \lambda e^{-\lambda x} dx - \int_0^y \lambda e^{-\lambda y} dx = -(e^{-\lambda y} - 1) - \lambda y e^{-\lambda y} \\ &= 1 - e^{-\lambda y} - \lambda y e^{-\lambda y}. \end{aligned}$$

Therefore, the CDF of compound binomial $\text{Bin}(2, p)$ with IID exponential distributions is the given by

$$\begin{aligned} F_Y(y) &= F^{[0\otimes]}(y)P(N = 0) + F^{[1\otimes]}(y)P(N = 1) + F^{[2\otimes]}(y)P(N = 2) \\ &= 1 \times (1 - p)^2 + (1 - e^{-\lambda y}) \times 2p(1 - p) + (1 - e^{-\lambda y} - \lambda y e^{-\lambda y}) \times p^2. \end{aligned}$$

The corresponding pdf of Y is given by

$$f_Y(y) = 2p(1 - p)\lambda e^{-\lambda y} + p^2\lambda^2 y e^{-\lambda y}.$$

6.4.2 Compounding Poisson Distribution

Before introducing compound Poisson distribution, we first present a result on **n-fold convolution** of exponential distribution with density function

$$f(x) = \lambda e^{-\lambda x}.$$

Lemma Let $\{X_1, X_2, \dots, X_n\}$ be IID exponential random variables with density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Then the n-fold convolution is given by

$$f^{[n\otimes]}(y) = \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!}$$

Proof (by induction). We have shown in the previous section that

$$f^{[n\otimes]}(y) = \lambda^2 y e^{-\lambda y}.$$

That is, the Lemma is true when $n = 2$. Assume that the lemma is true for $n = k$, i.e.,

$$f^{[k\otimes]}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-2)!}.$$

Then,

$$\begin{aligned} f^{[(k+1)\otimes]}(y) &= [f^{[k\otimes]} \otimes f](y) = \int_{-\infty}^{\infty} f^{[k\otimes]}(y-u) f(u) du \\ &= \int_{-\infty}^{\infty} f^{[k\otimes]}(u) f(y-u) du = \int_0^y \frac{\lambda^k u^{k-2} e^{-\lambda u}}{(k-2)!} \times \lambda e^{-\lambda(y-u)} du \\ &= \frac{\lambda^{k+1} e^{-\lambda y}}{(k-2)!} \int_0^y u^{k-2} du = \frac{\lambda^{k+1} e^{-\lambda y}}{(k-2)!} \times \frac{y^{k-1}}{k-1} = \frac{\lambda^{k+1} y^{(k+1)-2} e^{-\lambda y}}{[(k+1)-2]!}. \end{aligned}$$

This means that the lemma is true when $n = k + 1$. By the principle of induction, the lemma is true for all n .

With the above lemma, the compound Poisson distribution $Y = X_1 + X_2 + \cdots + X_N$ is given by

$$f(x) = \sum_{n=0}^{\infty} f^{[n\otimes]}(y) P(N = n) = \sum_{n=0}^{\infty} \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-2)!} \times \frac{\gamma^n e^{-\gamma}}{n!} = e^{-(\lambda y + \gamma)} \sum_{n=0}^{\infty} \frac{(n-1)(\lambda \gamma)^n y^{n-1}}{n!}$$

One can easily show that the function of y is a density function.

6.5 Properties of Compound Distributions

We briefly discuss the properties of compound distributions based on moments in this subsection. Recall that the compound distribution of $Y = X_1 + X_2 + \cdots + X_N$, (X_1, X_2, \dots, X_N) are IID and follow some distributions. N has a distribution that independent on X_i for $i = 1, 2, \dots, X_N$. The compound distribution Y is given by

$$f_Y(y) = \sum_{n=0}^N f^{[n\otimes]}(y) \times P(N = n).$$

Not that

$$f^{[n\otimes]}(y) = \frac{dP(X_1 + X_2 + \cdots + X_n \leq y | N = n)}{dy}$$

is a conditional density. For convenience, we use the expression of conditional density in the compound density in the following

$$f_Y(y) = \sum_{n=0}^{\infty} f^{[n\otimes]}(y | N = n) \times P(N = n).$$

6.5.1 The Law of Total Expectation and Variance

Since the derivations of moment-based properties involve conditional events, we next introduce the *law of total expectation and variance*.

- **About The Notation of Expectation E**

Every expectation is associated with a random variable. When working with more than one variable (particularly conditional distribution), we should explicitly specify the distribution associated with the expectation by using a subscript. For example $E_X(aX + b) = \int_D (aX + b)f(x)dx$.

- **The Law of Total Expectation:** Let X and Y be two random variables. the expected value of the conditional expectation of X given Y is the same as the expected value of X :

$$E_X(X) = E_Y[E_X(X|Y)].$$

To illustrate, we assume X and Y have a joint density $f(x, y)$. By definition,

$$\begin{aligned} E_Y[E_X(X|Y)] &= E_Y \left(\int_{D_X} xf(x|Y)dx \right) \\ &= \int_{D_Y} \int_{D_X} xf(x|y)f_Y(y)dx dy = \int_{D_Y} \int_{D_X} \frac{xf(x, y)}{f_Y(y)} f_Y(y)dx dy \\ &= \int_{D_X} x \int_{D_Y} f(x, y)dy = \int_{D_X} xf_X(x)dx = E_X(X). \end{aligned}$$

That is,

$$E_Y[E_X(X|Y)] = E_X[X].$$

The domain of X and Y are denoted respectively by D_X and D_Y . Similarly, we can derive the law of total variance.

$$\text{var}_X(X) = E_Y[\text{var}_X(X|Y)] + \text{var}_Y[E(X|Y)].$$

We will use the above law of total expectation to derive the law of total variance in the following two steps

$$E_Y[\text{var}(X|Y)] = E_Y [E_X(X^2|Y) - [E(X|Y)]^2] = E_X(X^2) - E_Y [E_X(X|Y)]^2.$$

and

$$\text{var}_Y[E_X(X|Y)] = E_Y [E_X(X|Y)]^2 - \{E_Y [E_X(X|Y)]\}^2 = E_Y [E_X(X|Y)]^2 - [E_X(X)]^2$$

Adding up the above two equations, we have the law of total variance

$$\text{var}_X(X) = E_Y[\text{var}_X(X|Y)] + \text{var}_Y[E(X|Y)].$$

6.5.2 Properties of General Compound Distribution

- **Expectation**

Since N and X_i (for $X_i = 1, 2, \dots, X_N$) are independent and X_i 's are continuous with common CDF F , then

$$\begin{aligned} E[Y] &= E_N[E_Y(Y|N)] = E_N[E_X(X_1 + X_2 + \dots + X_N|N)] \\ &= E_N[NE(X_1)] = E[X_1]E[N]. \end{aligned}$$

Example For compound Poisson distribution discussed in the above subsection, The expectation of i.i.d X_i is

$$E[X_1] = \frac{1}{\lambda} \quad \text{and} \quad E[N] = \gamma$$

Therefore, $E[Y] = E[X_1]E[N] = \gamma/\lambda$.

- **Variance of Compound Distributions**

We use the definition of variance to derive the variance of compound distribution

$$\begin{aligned} \text{var}_Y[Y] &= E_N[\text{var}_Y(Y|N)] + \text{var}_N[E(Y|N)] \\ &= E_N[\text{var}_Y(X_1 + X_2 + \dots + X_N|N)] + \text{var}_N[E(X_1 + X_2 + \dots + X_N|N)] \\ &= E_N[N \times \text{var}_Y(X_1)] + \text{var}_N[N \times E(X_1)] = E_N[N]\text{var}_Y(X_1) + \text{var}_N[N][E(X_1)]^2. \end{aligned}$$

- **Moment Generating Functions**

By the moment generating function, we have

$$\begin{aligned} M_Y(t) &= E_Y[e^{tY}] = E_N \left\{ E[e^{t(X_1 + \dots + X_N)}|N] \right\} \\ &= E_N \left\{ E[e^{t(X_1)}]^N \right\} = E_N \left\{ e^{N \ln E_{X_1}[e^{tX_1}]} \right\} \\ &= E_N \left\{ e^{N \ln E_{X_1}[e^{tX_1}]} \right\} = E_N \left\{ e^{N \ln M_{X_1}(t)} \right\} = M_N[\ln M_{X_1}(t)], \end{aligned}$$

were $M_{X_1}(t)$ is the common moment generating function (MGF) of IID random variables X_1, \dots, X_n and $M_N(\cdot)$ is the moment generating function of a random variable N .

6.6 Summary

We have discussed three different approaches to introducing new distributions based on two or more existing distributions in this and previous sections.

- **Convolution Method**; finds the distribution (pdf or CDF) of the sum of (IID or non-IID) random variables. n-fold convolution of identical distributions is commonly used in practice partly because some of the distributions could have a closed form of analytic expression. For example, the n-fold exponential distribution has a nice form (which is an **Erlang distribution**).
- **Mixture Distribution Method**: defines a new distribution by taking the weighted average of existing distributions. The weights could be unknown and can be estimated from the given data.
- **Compounding Method**: defines a distribution of the sum of a set of **random numbers** of random variables.