LOSS FUNCTION IN ACTURIAL SCIENCE AND ESTIMATION

A Thesis Presented to the Department of Pure and Applied Mathematics, African University of Science and Technology

In Partial Fulfilment of the Requirements for the Degree of

Master of Science

by

Zulaihat Hassan

Abuja, Nigeria

June, 2019.

Loss Function in Acturial Science and Estimation

Certification

This is to certify that the thesis titled "**Loss Function in Acturial Science and Estimation**" submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for the award of the Master's degree is a record of original research carried out by Zulaihat Hassan in the department of Pure and Applied Mathematics.

Approval

LOSS FUNCTION IN ACTURIAL SCIENCE AND ESTIMATION

By

Zulaihat Hassan

A THESIS PRESENTED TO THE DEPARTMENT OF PURE AND APPLIED MATHEMATICS

RECOMMENDED: ==========

Supervisor, Prof. Gane Samb Lo

==========

Head, Department of Pure and Applied Mathematics

APPROVED: ==========

Chief Academic Officer, Prof. C. E. Chidume

==========

Date

Abstract

The non-life insurance pricing consists of establishing a premium or a tariff paid by the insured to the insurance company in exchange for the risk transfer. A key factor in doing that is properly estimating the distribution that the claim and frequency of claim follows. This thesis aim at having a deep knowledge of loss function and their estimation, several concept from Measure Theory, Probability Theory and Statistics were combined in the study of loss function and estimating them is illustrated using insurance data set distributed by the Data Sciences website https://www.kaggle.com. The software R is used to obtained our results.

Key Words. *Loss Fuction, Insurance claim, Premium*

Dedication

Dedicated to my parents, Mr & Mrs Abdulganiyu Hassan and to my late friend Rabiu Musa Baffa. May his soul rest in peace (Amin).

Acknowledgements

I would like to use this section to express my profound gratitude to a quite number of persons that inspired, encouraged, guided and supported me.

I would like to express my untainted appreciation to my supervisor, Professor Gane Samb Lo, for giving me the inspiration to take up this research. The research appeared somewhat impossible at the beginning but with his assistance, support and patience amidst his tight schedules; he successfully guided and supervised me through this thesis work. My sincere appreciation goes to him, for his elderly advice and encouragement throughout the period of my study.

I am extremely thankful to Professor Charles E. Chidume, Vice-President (Academics) and Director of mathematics Institute, African University of Science and Technology (AUST), Abuja for providing me an opportunity and the time necessary to do this thesis. Also, I thank Miss Amaka Udigwe, Administrative assistant to the Vice-President (Academics), AUST, Abuja for her consistent availability and readiness to assist me at all times. To the entire faculty members of the Mathematics Institute, AUST,

I owe you a big thanks.

I am equally grateful to my parents and siblings for their love, encouragements and prayers during the course of this work.

I would like to express my special gratitude to African University of Science and Technology (AUST), Abuja and the African Union (Mwalimu Nyerere) for their financial support. It was their scholarships that supported my study throughout my masters programme.

My appreciation goes to my colleagues and course mates especially, Rahama Sani, Sahura Badamasi, Sani Salisu, Jonathan Areji. I said a big thank you and God bless you.

Finally my utmost appreciation goes to my friends especially, Mubarak Mohammed, Murtadoh Ibrahim, Rabiu Musa Baffa, Nuhu Ibrahim, Shakirat Sa'adudeen. This 18 months journey would have been totally impossible without you guys, I am forever grateful, God bless you all.

Contents

CHAPTER 1

Introduction

Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in insurance, finance and other industries and professions.

Taking the insurance company as a case study, we realize that the fundamental features of these companies is the concept of risk sharing, also known as risk distribution.

Risk distribution is a concept structure in a way that many pay for the expected losses of the few, this is because the risk measure (and, hence, the capital required to support it) for two risks combined is less than that of the risks treated separately. If the number of individuals get large enough, the risk might get nearly to zero.

What insurance company do is to organize such a redistribution for the purpose of making profit. In so doing, the insurance companies develop specific insurance products covering specified risk for client(insured) who contribute their own part of investment through premiums charged by the company.

1. Motivation of Study

In Nigeria, one mandatory insurance product is the third party motor insurance, which indemnifies vehicle driver against third party damages or losses, as specified by the insurance underwriting.

The regulatory framework for insurance companies in Nigeria also places a cap on what amount of premium can be charged. Some vehicle owners take up the comprehensive motor insurance which provides coverage not just for the third party, but for themselves too, as specified in the underwriting. Most often than not, the coverage usually applies to situation such as car theft, accident or fire.

You will no doubt agree that these events have a very low probability of occurrence in motor users. However every motor user frequently struggle with the challenges of routine maintenance cost arising from wear and tear of vehicle through usage. And so a natural question that arises is can we extend motor vehicle insurance product to cover this?

Obvious challenges to answering this question in the affirmative for the insurance companies would be:

(1) What will be the premium? How will it be collected? How will the claims be made?

(2) How will this be designed so that they can avoid losses while maximizing profits and at the same time keeping the premium low? What is even the likelihood of making profit from this product?

A solid foundation that will help in furnishing answers to those questions would be the concept of loss functions.

This Thesis contributes in great details to a thorough exposition and understanding of loss functions as applied in insurance.

What are loss functions and how do they contribute to insurance product development?

Suppose an insurance company has a revenue R and client made claim X the revenue of the company will become $R-X$. So X is a loss to the insurance company and is called the Loss function. Suppose the insurance company has an initial premium u, let us suppose that all premiums are received at a constant rate c so that

$$P_t = u + ct,$$

Suppose at a time t there are $N(t)$ number of claims up to time t, $X_1, X_2, X_3 \cdots X_{N(t)}$, the revenue(surplus) of the company at time t is

$$S_t = u + ct - \sum_{i=1}^{N(t)} X_i.$$

this gives a relationship between premium and claims.

Usually the stochastic process $\{N(t),\ t \geq 0\}$ is the counting process of a Poisson process of intensity $\lambda > 0$. For reader who want to read more on such stochastic processes, we presented it in the Appendix in Chapter 6.

What did we see?

The number of Claims and the Claims are all random.

2. Statement of Problem

A key factor in calculating the premium of an insurance company to avoid ruin and maximize profit is knowing the distribution that the claims and frequency of claims follows. An Actuary is presented with data from the field, a natural question that arise is how can we estimate these distributions from data?

3. Aim

This Thesis aim at estimating loss function from data. This can be achieve through the following objectives

4. Objectives

(1) To have a deep knowledge on real valued random variables and their characterization as well as the related vocabulary in actuarial science.
(2) To have a deep knowledge on Loss Modification

(3) To have deep knowledge on data modeling.

Accordingly to the objectives described above, we organize the body of the dissertation as follows.

A loss function is merely the probability law of a real-valued random variable. Hence, in Chapter 2, we make a round on such laws from a probabilistic approach. We mainly follow the presentations of Lo (2018) and Lo (2018), but the fundamental books of Loève (1997), Chung (1974), Gutt (2005) might be useful. In that chapter, we present the characterization of such laws in a global and comprehensive way, That chapter, of course, goes beyond the scope of the thesis.

In Chapter 3, we present a specific presentation of the probability laws introduced in Chapter 2 in the specialized way they are used in Actuarial Sciences and in their terminology. We

study in it very detailed and particular properties that are very important in the profession. Our main source in that chapter is Klugman et al. (2008).

In Chapter 4, we deal with estimations of the loss functions. It constitutes an initiation of statistical estimation and tests, using data and the software R. We followed the book's project of Lo et al. (2019) which is under development.

In Chapter 5, we applied the knowledge to real-data as a case study of modelling loss data.

Finally, we conclude by final chapter devoted to cocnlusions and perspectives.

5. Basic Definition

(1) **Insurance:** An economic device transferring risk from an individual to a company and reducing the uncertainty of risk through pooling.

(2) **Insured:** Party(ies) covered by an insurance policy

(3) **Insurer:** An insurer or reinsurer authorized to write property and/casualty insurance under the law of any state.

(4) **Premium:** Money charged for the insurance coverage, reflecting expectation of loss

(5) **Actuary:** A business professional who analyzes probabilities of risk and risk management including calculation of premiums, dividends and other applicable insurance policies held.

(6) **Claim:** A request made by the insured for insurer remittance of payment due to loss incured and covered under the policy agreement

(7) **Coinsurance:** A clause contained in most property insurance policies to encourage policy holders to carry a reasonable amount of insurance. If the insured fails to maintain the amount specified in the clause(Usually at least 0.8), the insured shares a higher proportion of the loss.

(8) **Policy:** A written contract ratifying the legality of an insurance agreement

(9) **Ruin Time:** This is the time when the insurance company become bankrupt

(10) **Probability space:** A probability space is a measure space $(\Omega, \mathcal{A}, m)$ where the measure assigns the unity value to the whole

space, that is $m(\Omega) = 1$. Such a measure is called a probability measure. Probability measures are generally denoted in blackboard font $\mathbb{P}$, $\mathbb{Q}$ etc.

CHAPTER 2

Random Variable : A Summary

The state of an Insurance company heavily depends on the level on the claims by the clients. A wrong estimation of the claims, also called losses, surely lead to the ruin of the company.

It is then very important to estimate the level of the losses which are real-valued variables. We already explained in the introduction how these random variable intervene in determining the ruin time and the ruin probability.

Random variable may be studied in general in Probability Theorem. But experts in Actuarial Sciences need to have a deeper insight of the properties of the random variables to be sure to accurately assess the risk faced by the company.

Yet the theoretical study of the random variables may help in the detailed handling of loss function. This chapter is theoretical oriented and relies on the background of Measure Theory Integration. We begin by general considerations.

1. Terminology

Measurable mappings are called random variables. Hence, a mapping $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{E}, \mathcal{B})$ is a random variable, with respect to the σ-algebras $\mathcal{A}$ and $\mathcal{B}$ if and only if it is measurable with respect to the same σ-algebras.

We suppose that all the random variables which are used here are defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

The measure image $\mathbb{P}_X \mathbb{P} X^{-1}$ is called the probability law of X, defined as follows

$$B \ni \mathcal{B} :\to \mathbb{P}_X(B) = \mathbb{P}(X \in B).$$

Determining the probability law of X is the main task in Probability Theory.

There is not a more important phrase in Probability Theory than Probability law. Let me quote Professor Gane Samb Lo

> "The essence of probability Theory is finding probability laws of random phenomena by intellectual means and the essence of Statistical theory is the same but by means of inference from observations or data."

In the present thesis, finding $\mathbb{P}_X$, where X represents the claims, is an essential part of the work of the expert in Actuarial science. That expert uses available data and similar data related

to one product of the company to estimate $\mathbb{P}_X$. We will be dealing with his in Chapter 4 and 5

For now, let us give general considerations. Although the space E is arbitrary, the following cases are usually and commonly studied :

(1) If E is $\mathbb{R}$, endowed with the usual Borel σ-algebra, the random variable is called a real random variables (rrv).

(2) If E is $\mathbb{R}^d, d \geq 1$ endowed with the usual Borel σ-algebra, X is called a d-random vector or a random vector of dimension d, denoted $X = (X_1, X_2, ..., X_d)^t$, where X^t stands for the transpose of X.

(3) E is of the form $\mathbb{R}^T$, where T is an arbitrary non-empty set, then X is simply called a stochastic process.

(4) If E is some metric space (S, d) endowed with the Borel σ-algebra denoted as $B(S)$, the term random variable is simply used although some authors prefer using random element.

Now, let us focus of parameters of random variables of finite dimensions.

2. Parameters of Random vectors

2.1 Parameters of real-valued of rrvs's

1. Mathematical Expectation.

Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a real random variable. Its mathematical expectation with respect to the probability measure $\mathbb{P}$ or its $\mathbb{P}$-mathematical expectation, denoted by $\mathbb{E}_{\mathbb{P}}(X)$. is simply its integral with respect to P whenever it exists and we denote :

$$\mathbb{E}_{\mathbb{P}}(X) = \int_{\Omega} X \ d\mathbb{P}.$$

If there is no confusion, we may drop the subscript $\mathbb{P}$ and have

$$E(X) = \int_{\Omega} X d\mathbb{P}.$$

Also, the parentheses may also be removed and we write EX.

2. Mathematical expectation of a real-valued function of an arbitrary random variable: for any real-valued measurable mapping

$$h : (\mathbb{E}, \mathcal{B}) \to ((\mathbb{R}, \mathcal{B}(\mathbb{R})) \tag{2.1}$$

The composite mapping

$$h(X) = h \diamond X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$$

is a real random variable. We may define the mathematical expectation of $h(X)$ with respect to $\mathbb{P}$ by

$$\mathbb{E}h(X) = \int_{\Omega} h(X) d\mathbb{P} = \int_{\mathbb{R}} h(t) \ \ d\mathbb{P}_X(t)$$

whenever the integral exists. If X is itself a real random variable, its expectation, if it exists, is

$$\mathbb{E}(X) = \int_{\mathbb{R}} t \ \ d\mathbb{P}_X(t).$$

2. Properties of the Mathematical Expectation.

As an integral of real-valued measurable application, the mathematical expectation inherits all the properties of integrals we already had in Measure Theory. Here, we also have that constant real random variables and bounded random variables have finite expectations.

Theorem 2.1. We denote the class of all *rrv* with finite mathematical expectation as $L_1(\Omega, \mathcal{A}, \mathbb{P})$ the mathematical expectation operator has the following properties:

(a) **Linearity** that is for all $(a, b) \in \mathbb{R}^2$, for all $(X, Y) \in L_1(\Omega, \mathcal{A}, \mathbb{P})$,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

(b) It is non-negative, that is for all non-negative X random variable, we have $E(X) \geq 0$.

(c) Satisfies for all non-negative random variable X: E(X) = 0 if and only if X = 0, $\mathbb{P}$-a.e. Besides, we have for all real-valued random variables X and Y defined on $(\Omega, \mathcal{A})$.

$$(|X| \leq Y, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})) \Longrightarrow X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$$

and

$$\left| \int X d\mathbb{P} \right| \leq \int |X| d\mathbb{P} \leq \int Y d\mathbb{P}.$$

3. Moments of covariances of Real Valued Random Variable

3.1. Moments of Real Random Variables.

The moments play a significant role in Probability Theory and in Statistical estimation. Let X and Y are two rrv's. We define the following parameters, whenever the concerned expressions make sense.

(A) **Non centered moments of order k** $k \geq 1$

$$m_k(X) = E(X^k)$$

(B) **Centered Moment of order k** $k \geq 1$

$$\mu_k(X) = E(X - m_1)^k$$

which is defined if $m_1(X) = EX$ exists and is finite.

If EX exists and is finite, the centered moment of second order $\mu_2(X) = E|X - m_1|^2$ is called the variance of X also denoted Var(X) and its square root is called the standard deviation of X.

(C) **Covariance between X and Y.** If EX and EY exist and are finite, we may define the covariance between X and Y by

$$Cov(X,Y) = E((X - EX)(Y - EY))$$

By expanding and using the properties of integral in the definition of variance and covariance we get

$$Var(X) = E(X^2) - (EX)^2 \ \ and \ \ Cov(X,Y) = E(XY) - E(X)E(Y)$$

Hence if X and Y are independent then the covariance is zero.

3.2. Two basic inequalities based on the expectation and the variance.

(i) **Markov's inequality:** For any random variable X, we have for any $\lambda > 0$ that

$$P(|X| > \lambda) \leq \frac{E|X|}{\lambda}$$

(ii) **Tchebychev's inequality:** If $X-E(X)$ is defined a.e., then for any $\lambda > 0$,

$$P(|X - EX| > \lambda) \leq \frac{Var(X}{\lambda^2}$$

3.3. Remarkable properties of Variance and Covariance. Whenever the expressions make sense, we have the following properties.

(P1) Var(X) = 0 if and only if X = E(X) a.e.

(P2) For all $\lambda > 0, Var(\lambda X) = \lambda^2 Var(X)$.

(P3) For any finite sequence of rrv's, $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_k$ we have

$$Var\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2 Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\alpha_i\alpha_j$$

We also have

$$Cov\left(\sum_{i=1}^{n} \alpha_i X_i, \sum_{i=1}^{n} \beta_j Y_j\right) = \sum_{1\leq i\leq n} \sum_{1\leq j\leq k} Cov(X_i, Y_j)\alpha_i\alpha_j$$

(P5) If X and Y are independent, then Cov(X, Y) = 0.

(P6) If $X_1, X_2, ..., X_n$ are pairwise independent, then

$$Var\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2 Var(X_i)$$

(P7) If none of σ_X and σ_Y is null, then the coefficient

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

is called the linear correlation coefficient between X and Y and satisfies $|\rho_{XY}| \leq 1$.

4. Random variables on $\mathbb{R}^d$ or Random Vectors

Random vectors are generalizations of real random variables. A random vector of dimension $d \geq 1$ is a random variable

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \tag{4.1}$$

with values in $\mathbb{R}^d$. From Measure Theory, we know that a random vector,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = (X_1, X_2, ...X_d)^t \tag{4.2}$$

is a random variable if and only if each $X_i, 1 \leq i \leq d$ is a real random variable. If $d = 1$, the random vector becomes a real random variable (rrv).

1. Expectation, Variance and Covariance of random vectors: Let $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^r$ we have

1.1. Expectation: The Mathematical expectation is the column vector

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \mathbb{E}(X_2), ...\mathbb{E}(X_d))^t$$

1.2. Covariance Matrix: the covariance between $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^r$ is the (dxr) matrix obtained as

$$\begin{aligned} Cov(X,Y) &= \Sigma_{XY} = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)^t) \\ &= \mathbb{E}((X_i - \mathbb{E}X_i)(Y_j - \mathbb{E}Y_j)^t)_{1\leq i\leq d, 1\leq j\leq r} \\ &= (Cov(X_i, Y_j))_{1\leq i\leq d, 1\leq j\leq r} \end{aligned}$$

1.3. Variance-Covariance Matrix: The variance of $X \in \mathbb{R}^d$ is a (dxd) matrix obtained as

$$\begin{aligned} Var(X) &= \Sigma_X = \mathbb{E}((X - \mathbb{E}X)(X - \mathbb{E}X)^t) \\ &= \mathbb{E}((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)^t)_{1\leq i\leq d} \\ &= (Cov(X_i, X_j))_{1\leq i\leq d} \end{aligned}$$

1.4. Properties Here are the main properties of the defined parameters.

(a) For any $\lambda \in \mathbb{R}$,

$$E(\lambda X) = \lambda E(X)$$

.

(b) For two random vectors X and Y of the same dimension d, we have

$$E(X+Y) = E(X) + E(Y)$$

.

(c) For any (pxd)-matrix A and any d-random vector X,

$$E(AX) = AE(X) \in \mathbb{R}^p$$

(d) For any d-random vector X and any s-random vector Z,

$$Cov(X,Z) = Cov(Z,X)^t$$

.

(e) For any (pxd)-matrix A, any (q x s)-matrix B, any d-random vector X and any s-random vector Z,

$$Cov(AX,BZ) = ACov(X,Z)B^t$$

which is a (pxq)-matrix.

Random functions are associated to functions which characterize their probability laws. Let us see three of them.

2. Cumulative Distribution Functions (*cdf*).

2.1. Definition. Let

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)),$$

be a d-random vector. The function defined by

$$\mathbb{R}^d \ni x \mapsto F_X(x) = P(X \leq x)$$

where $x^t = (x_1, ..., x_d)$ and

$$F_X(x_1, ..., x_d) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_d \leq x_d) = \mathbb{P}_X \left(\prod_{i=1}^{d}]-\infty, x_i \right)$$

is called the cumulative distribution (cdf) function of X.It has the following important properties.

2.2. Properties of the cdf.

(a) It assigns non-negative volumes to cuboids, that is for all $(a, b) \in (\mathbb{R}^d)^2$ such that $a \leq b, \Delta_{a,b}F \geq 0$. where

$$\Delta_{a,b}F = \sum_{\varepsilon \in \{0,1\}^d} (-1)^{s(\varepsilon)} F(b + \varepsilon(b - a))$$

$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_d), (x, y) * (X, Y) = (x_1X_1, ..., y_dY_d)$ and $s(\varepsilon) = \sum_{i-1}^{d} \varepsilon_i$

(b) It is right-continuous at any point $t \in \mathbb{R}^d$, that is for any decreasing sequence in $\mathbb{R}^d$, $t^{(n)} \downarrow t$ we have

$$F_X(t^{(n)}) \downarrow F_X(t)$$

(c) F_X satisfies the limit conditions :

$$\lim_{\forall i, 1 \leq i \leq k, t_i \to -\infty} F_X(t_1, ..., t_k) = 0$$

and

$$\lim_{\forall i, 1 \leq i \leq k, t_i \to +\infty} F_X(t_1, ..., t_k) = 1$$

Generally, a function $F : \mathbb{R}^d \mapsto [0, 1]$ is cdf if and only if Conditions (1), (2) and (3) above hold. If (1) and (2) only hold, F is called a distribution function.

2.3. Characterization.

There exists a one-to-one correspondence between the class of Probability Lebesgue-Stieljes measures $\mathbb{P}_F$ on $\mathbb{R}^d$ and the class of cdf's F on $\mathbb{R}^d$ according to the relations

$$\forall x \in \mathbb{R}^d, F_P(x) = \mathbb{P}(]-\infty, x])$$

and

$$\forall a, b \in \mathbb{R}^d, a \leq b, \mathbb{P}_F(]a, b]) = \Delta_{a,b} F$$

This implies that two d-random vectors X and Y having the same distribution function have the same probability law.

Remark: Given a cdf F, By Kolmogorov's construction, we can always find a random variable X, such that $F = F_X$

2.4. Joint cdf 's and marginal cdf 's: Let $X^t = (X_1, X_2, ..., X_d)$ be a random vector with cdf $F_X = F_{(X_1, X_2, ..., X_n)}$ defined as in above then F_X is called the joint cdf of $X_1, X_2, ..., X_d$. Their marginal cdf is F_{X_i} defined as

$$F_{X_i}(x) = \mathbb{P}(X_i \leq x)$$

We can obtained the marginal cdf from the joint cdf as

$$F_{X_i}(x_i) = F_{(X_1,X_2,...,X_n)}\left(+\infty, ... + \infty, x_i, +\infty, ... + \infty\right)$$

3. Characteristic function (*cf*).

The characteristics function of a random variable X denoted as ϕ_X is defined as $\phi_X(t) = E(e^{i<X,t>})$ where $i = \sqrt{-1}$ The characteristics function always exist for all distribution since

$$\phi_X(t) = E(e^{i<X,t>}) = E(cos < X, t >) + iE(sin < X, t >)$$

which is defined since the integrands for the real and imaginary parts are bounded.

3. Probability Density Functions (pdf).

Two classes of random variables are frequently used in real cases, although they constitute a small part of the class of all random variable : Discrete probability law and absolutely continuous random variables.

3.1. Definitions.

(Discrete Random Variable) A rrv X is said to be discrete if it takes at most a countable number of values in $\mathbb{R}$ denoted $V_X = \{x^{(j)}, j \in J \subset \mathbb{N}, J \neq \emptyset\}$

We have from Measure Theory that X is measurable if and only if $\forall j \in J, (X = x^{(j)}) \in \mathcal{A}$. Besides for any $B \in \mathcal{B}(\mathbb{R}^d)$, we have

$$\mathbb{P}_X(B) = \sum_{j \in J, x^{(j)} \in B} \mathbb{P}(X = x^{(j)})$$

and

$$\sum_{j \in J} \mathbb{P}(X = x^{(j)}) = 1$$

Now if we define a functionf_X on V_X by

$$V_X \ni x^{(j)} \mapsto f_X(x^{(j)}) = \mathbb{P}(X = x^{(j)}).$$

Consider the counting measure ν on $\mathbb{R}^d$ with support V_X then from the equations above we have

$$\int f_X d\nu = 1$$

For any $B \in \mathcal{B}(\mathbb{R}^d)$ we have

$$\int_B d\mathbb{P}_X = \int_B f_X d\nu$$

We conclude that f_X is the Radon-Nikodym derivative of $\mathbb{P}_X$ with respect to the σ-finite measure ν. If X is discrete, it has a probability density function pdf f_X with respect to the counting measure defined by

$$f_X(x) = P(X = x), x \in \mathbb{R}^d$$

which satisfies

$$f_X(x^{(j)}) = P(X = x(j)) \ \ for \ \ j \in J \ \ and \ \ f_X(x) = 0 \ \ for x \notin V_X$$

DEFINITION 2.2. ***(Absolutely Continuous Probability laws)*** *The probability Law $\mathbb{P}_X$ is said to be absolutely continuous if it is continuous with respect to the Lebesgue measure λ_d on $\mathbb{R}^d$. By extension, we say the random variable itself is said to be absolutely continuous. By Radon-Nikodym's Theorem, there exists a Radon-Nikodym derivative denoted f_X such that for any $B \in \mathbb{R}^d$,*

$$\int_B d\mathbb{P}_X = \int_B f_X d\lambda_d$$

The function f_X satisfies

$$f_X \geq 0 \qquad \textit{and} \qquad \int_\Omega f_X d\lambda_d = 1$$

Such a function is called a probability density function pdf with respect to the Lebesgue measure.

3.2. Relationship between the cdf and the pdf of absolutely continous rv.

For any $x \in \mathbb{R}^d$

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \int_{]-\infty,x]} d\mathbb{P}_X \\ &= \int_{]-\infty,x]} f_X d\lambda_d \\ &= \int_{-\infty}^{x_1} d\lambda(t_1) \int_{-\infty}^{x_2} d\lambda(t_2) \dots \int_{-\infty}^{x_d} f_X(t_1, t_2, ...t_d) d\lambda(t_d) \end{aligned}$$

The last equation is obtained by using the Fubini's theorem. If f_X is locally bounded and locally Riemann integrable (LL-BRI), we have

$$f_X(x_1, x_2, ..., x_k) = \frac{\partial^k F_X(x_1, x_2, ..., x_k)}{\partial x_1 \partial X_2 \dots \partial X_k)}, \lambda_d, a.e$$

Marginal Probability Density functions.

Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto \mathbb{R}^d$ be a random vector with $X^t = (X_1, .., X_d)$. Suppose that X has a pdf $f_{(X_1,..,X_d)}$ with respect to a σ-finite product measure $m = \otimes_{j=1}^{d} m_j$.
Then each $X_j, 1 \leq j \leq d$, has the marginal pdf 's f_{X_j} with respect to m_j given by: for $x \in \mathbb{R}$,

$$f_{X_j}(x) = \int_{\mathbb{R}^{d-1}} f_{(X_1,..,X_d)}(x_1, x_2, ..., x_k) d\left(\otimes_{1 \leq i \leq d, i \neq j} m_i\right)(x_1, ..., x_{j-1}, x_{j+1}, ...x_d)$$

When d=2, i.e $X = (X_1, X_2)$ the marginal pdf's are

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{(X_1,X_2)}(x_1, x_2) dm_2(x_2), m_1 a.e$$

and

$$f_{X_2}(x_2) = \int_{\mathbb{R}} f_{(X_1,X_2)}(x_1, x_2) dm_1(x_1), m_2 a.e$$

5. Independence

The notion of independence is extremely important in Probability Theory and its applications.

1.Independence of sets.

Let $A_1, A_2, ..., A_n$ be events in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

We have the following definitions.

(a) $A_1, A_2, ..., A_n$ are pairwise independent if and only if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j), forall 1 \leq i \neq j \leq n$$

(b) $A_1, A_2, ..., A_n$ are mutually independent if and only if for any subset $\{i_1, i_2, ...i_k\}$ of $\{1, 2, 3, ...n\}$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

(c) $A_1, A_2, ..., A_n$ fulfills the global factorization formula if and only if

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdots \mathbb{P}(A_n)$$

None of the above definition are equivalent but (b) implies (a) and (c) also independence of set refers to definition (b).

2. Independence of Random variable.

Let $X_1, X_2, ..., X_n$ be n random variables defined on the same probability space i.e

$$X_i : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{E}_i, \mathcal{B}_i)$$

Let $(X_1, X_2, ..., X_n)$ be the n-tuple defined by

$$(X_1, X_2, ..., X_n) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{E}, \mathcal{B})$$

where $E = \prod_{1 \leq i \leq n} \mathbb{E}_i$ is the product space of the $\mathbb{E}_i$'s endowed with the product $\sigma-$ algebra, $B = \prod_{1 \leq i \leq n} \mathcal{B}_i$. On each $(\mathbb{E}_i, \mathcal{B}_i)$, we have the probability law $\mathbb{P}_{X_i}$ of X_i. Each of the $\mathbb{P}_{X_i}$'s is called a marginal probability law of $(X_1, X_2, ..., X_n)$. On $(\mathbb{E}, \mathcal{B})$, we have the following product probability measure define on the semi algebra formed by the set of measurable rectangles and also we have the probability law

$$\mathbb{P}_{(X_1, X_2, ..., X_n)}(B) = \mathbb{P}((X_1, X_2, ..., X_n) \in B)$$

of the n-tuple $(X_1, X_2, ..., X_n)$ on $(\mathbb{E}, \mathcal{B})$, which is the image-measure of $\mathbb{P}$ by $(X_1, X_2, ..., X_n)$ This is called the joint probability.

$$\mathbb{P}_{X_1} \otimes \mathbb{P}_{X_2} \otimes ... \mathbb{P}_{X_n} \left(\prod_{1 \leq i \leq n} A_i \right) = \prod_{1 \leq i \leq n} \mathbb{P}_{X_i}(A_i), A_i \in \mathcal{B}_i$$

$X_1, X_2, ..., X_n$ are linearly independent iff the joint probability equal product of its marginal probability or in another words if is the product measure of its marginal probability laws i.e for any $A_i \in \mathcal{B}_)$,

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, ..., X_n \in A_n) = \prod_{1 \leq i \leq n} \mathbb{P}_{X_i}(A_i).$$

N.B: . The independence is defined for random variables defined on the same probability space. The space in which they take values may differ.

THEOREM 2.3. The random variables $X_1, X_2, ..., X_n$ are independent if and only if, for all non-negative and measurable real-valued functions $h_i : (\mathbb{E}_i, \mathcal{B}_i) \to \mathbb{R}$, we have

$$\mathbb{E}\left(\prod_{1\leq i\leq n} h(X)_i\right) = \prod_{1\leq i\leq n} \mathbb{E}(h_i(X_i)) \tag{5.1}$$

Independence of events is obtained from independence of random variables.Two events $A \in \mathcal{A}$ and $B \in \mathcal{B}$ are independent if and only if the random variables 1_A and 1_B are independent, that is, for all $h_i : \mathbb{R} \to \mathbb{R}$ (i=1,2) non-negative and measurable

$$\mathbb{E}(h_1(1_A)h_2(1_B)) = \mathbb{E}(h_1 1_A)E(h_2 1_B)$$

Now events $A_1, ...A_n$ are independent iff the random variables $1_{A_i}, 1 \leq i \leq n$ are independent iff for any measurable finite mappings $h_i : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbb{E}\left(\prod_{1\leq i\leq n} h_i 1_{A_i}\right) = \prod_{1\leq i\leq n} \mathbb{E}(h_i 1_{A_i}),$$

3. Family of independent random variables.

Consider a family of random variables

$$X_t : (\Omega, \mathcal{A}, \mathbb{P}) \to (E_t, \mathcal{B}_t)), (t \in T).$$

This family $X_t; t \in T$ may be finite, infinite and countable or infinite and non countable. It is said that the random variables of this family are independent iff the random variables in any finite sub-family of the family are independent, that is, for any subfamily $t_1, t_2, ..., t_k \subset T$, the r.v $X_{t_1}, ...X_{t_k}$ are independent.

Theorem 2.4. **Using the cdf to characterize independence:** Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto \mathbb{R}^d$ be a random vector. The margins $X_i, 1 \leq i \leq d$, are independent if and only if the joint cdf of X is factorized in the following way :

$$\forall (x_1, .., x_d) \in \mathbb{R}^d, F_{(X_1, X_2, ..., X_n)}(x_1, .., x_d) = \prod_{i=1}^{d} F_{X_i}(x_i)$$

Theorem 2.5. **Using the pdf to characterize independence:** Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto \mathbb{R}^d$ be a random vector. The margins $X_i, 1 \leq i \leq d$, are independent if and only if the joint cdf of X is factorized in the following way :

$$\forall (x_1, .., x_d) \in \mathbb{R}^d, f_{(X_1, X_2, ..., X_n)}(x_1, .., x_d) = \prod_{i=1}^{d} f_{X_i}(x_i)$$

6. Determining probability laws

There are several methods of determining probability law.

(1) **Using the convolution product to find the probability law of the sum of two independent real-value random variables:** Let X and Y be to real-valued and independent random variables, defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and admitting the probability density functions f_X and f_Y with respect to a σ-finite product measure. Then $Z = X + Y$ has a pdf f_Z which is

the convolution product of f_X and f_Y. If X and Y are absolutely continuous then f_Z has the following expression

$$f_Z(z) = f_X * f_Y = \int_{\mathbb{R}} f_X(z-x) f_Y(x) dx.$$

(2) Using the product of characteristic function to find the probability law of the sum of two independent random variables of equal dimension. the characteristics function is discussed in chapter two.

(3) Finding the cdf of the studied random variable and differentiate it if possible, and try to identify a known probability law. This has been discussed in relationship between the cdf and pdf of absolutely continuous r.v this method is used extensively.

(4) Directly finding the characteristic function of the studied random variable and trying to identify a known probability law. This can be done by using the relationship below.

$$f(x) = \int_{-\infty}^{\infty} e^{-ixu} \phi_X(u) du$$

(5) where ϕ_X is the characteristic function of the r.v X.

(6) **Using the Change of Variable Formula to derive pdf 's if applicable.** Let X be a random variable in $\mathbb{R}^d$ of probability density function f_X with respect to the Lebesgue measure on $\mathbb{R}^d$. Suppose that D is the support of X. Let

$$h : \Delta \mapsto D$$

be a diffeomorphism and $Y = h^{-1}(X)$ be another random vector. Then, the probability density function of Y exists and is given by

$$f_Y(y) = f_X(h(y))|J(h)|1_\Delta(y)$$

This is obtained by simply using definition and then the change of variable.

7. Some Usual Probability Laws and Properties

We introduce some probability laws often used.

(1) **Exponential Random Variable of parameter** $b > 0$ $X \sim \mathcal{E}(b)$ is supported on $\mathbb{R}^+$ has its pdf given by

$$f_X(x) = be^{-bx}1_{(x\geq 0)}$$

(2) **Gamma Random variable with Parameter** $a > 0$ **and** $b > 0$**:** We said X follows the gamma distribution, denoted $X \sim \gamma(a, b)$ if its pdf is define as

$$f_X(x) = \frac{b^a}{\Gamma a}x^{a-1}e^{-bx}1_{(x\geq 0)}$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

(3) **Beta Random variables of parameter** $a > 0$ **and** $b > 0$**.** $X \sim B(a, b)$ is defined on (0, 1) if its pdf is given by

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} 1_{(0,1)}(x),$$

(4) where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

(5) **Pareto Random Variable of parameter** $\alpha > 0, \theta \geq 0$. $X \sim P(\alpha, \theta)$ if its pdf is given by

$$f_X(x) = \frac{\alpha \theta^\alpha}{(x+\theta)^{\alpha+1}} 1_{(x \geq \theta)}(x)$$

(6) **Gaussian Probability Law.** It has support on the whole real line and mean μ, standard deviation σ. If X follows the Gaussian or normal distribution, the pdf is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(7) **The Standard Gaussian Probability law:** It is a special case of the Gaussian probability law when $\mu = 0$ and $\sigma = 1$ then the pdf is

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, x \in \mathbb{R}$$

Now given a standard random variable X with cdf F_X, m a real number and $\sigma > 0$, the random variable

$$Z = \sigma X + m$$

follows the Gaussian or normal probability law of mean m and standard deviation σ with cdf

$$F_Z(x) = \mathbb{P}(Z \leq x) = \mathbb{P}(\sigma X + m \leq x) = \mathbb{P}(X \leq \frac{x-m}{\sigma}) = F_X(\frac{x-m}{\sigma})$$

(8) **Finite linear combination of independent real Gaussian randoms:** Any linear combination of a finite number of independent Normal distribution i.e $X_i \sim \mathcal{N}(m_i, \sigma_i)$ then $Y = \sum_{i=1}^{n} \delta_i X_i$ with $\delta_i \in \mathbb{R}*$, we denote $m^t = (m_1, m_2, ..., m_n), \delta^t = (\delta_1, \delta_2, ..., \delta_n)$ and $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$, Then Y follows a normal distribution with parameter $(m^t\delta, \delta^t\Sigma\delta)$ i.e $Y \sim \mathcal{N}(m^t\delta, \delta^t\Sigma\delta)$

(9) **Chi-square Probability law of parameters** $d \geq 1$. $X \sim X_d^2$ is supported by $V_X = \mathbb{R}^+$.

A Chi-square Probability law d degrees of freedom is simply a Gamma law of parameters $a = \frac{d}{2}$ and $b = \frac{1}{2}$, that is $X_d^2 = \gamma\left(\frac{d}{2}, \frac{1}{2}\right)$

$$f_X(x) = \frac{1}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} x^{\frac{d}{2}-1} e^{-\frac{x}{2}} 1_{(x\geq 0)}$$

Chi-square distributions are generated from Gaussian random variables as follows.

Fact 1. If Z follows a standard Gaussian probability law, Z^2 follows a Chi-square law of one degree of freedom :

$$Z \sim \mathcal{N}(0, 1) \Rightarrow Z^2 \sim X_1^2$$

Fact 2. Let $d \geq 2$. If $X_1, ..., X_d$ are d independent real-valued random variables, defined on the same probability space, identically following a Chi-square law of one degree of freedom, we have that their sum follows a chi-square law of d-degree of freedom i.e

$$\sum_{i=1}^{d} X_i \sim X_d^2$$

We obtained this by using the fact that the characteristic function of d-independent chi-square distribution is their product

(10) **The Student probability law of** $n \geq 1$ **degrees of freedom.**

$X \sim t(n)$ is defined on the whole real line with pdf.

$$f_X(x) = \frac{\Gamma((n+1)/2)}{(n\pi)^{1/2}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

It can be shown using the method of change of variable that a t(n)-random variable has the same law as the ratio of two independent random variables: a standard normal distribution, $\mathcal{N}(0,1)$ random variable by square root of a chi-square random variable divided by its number of freedom degrees. i.e

$$t(n) = \frac{\mathcal{N}(0,1)}{\sqrt{\chi_n/n}}$$

(11) **The Fisher probability law of degrees of of freedom** $n \geq 1$ **and** $m \geq 1$**:**

$X \sim F(n,m)$ is defined on the positive real line with pdf.

$$f_X(x) = \frac{n^{n/2}m^{m/2}\Gamma((n+m)/2)x^{n/2-1}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(n+mx)^{(n+m)/2}}$$

The ratio of two independent random variables: a Chi-square random variable of n degree of freedom divide by n by a Chi-square random variable of m degree of freedom divide by m gives

a fisher random variable of degrees of of freedom $n \geq 1$ and $m \geq 1$ i.e,

$$F(n,m) = \frac{X_2^n/n}{X_2^m/m}$$

(12) **Gauss Probability Law on $\mathbb{R}^d$, Random Vectors** A random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto \mathbb{R}^d$ is said to follow a d-multivariate Gaussian probability law, or in other words, X is a d-Gaussian random vector if and only its mgf is defined by

$$M_X(u) = exp\left(\langle m, u\rangle + \frac{u^t \Sigma u}{2}\right), u \in \mathbb{R}^d$$

where m is a d-vectors or real numbers and Σ is a symmetrical and semipositive d-matrix of real numbers, and we write $X \sim \mathcal{N}_d(m, \Sigma)$, m and Σ are the respective expectation vector and the variance-covariance matrix of X.

Proposition 2.6. *A random vectors with real-valued independent Gaussian components is a Gaussian vector. i.e if $X^t = (X_1, ..., X_n)$ with $X_i \sim \mathcal{N}(m_i, \sigma_i)$ then $X \sim \mathcal{N}_d(m, \Sigma)$.*

We have the following results for Gaussian Random Vectors

Proposition 2.7. *The following assertions hold.*

(a) *Any finite-dimension linear transform of a Gaussian random vector is a Gaussian random vector.*

(b) *Any linear combination of the components of a Gaussian random vector is a real Gaussian random variable.*

(c) *If a random vector $X \sim \mathcal{N}_d(m, \Sigma)$ law and if A is a (kxd)-matrix and B a k-vector, then AX + B follows a $\mathcal{N}_k(Am + B; A\Sigma A^t)$ probability law.*

CHAPTER 3

Loss functions

This chapter focus on how the random variables introduced in the previous chapter are used in actuarial practice.

Consider the relationship between the premium and claims introduce in chapter 1, this motivate us to study deeply how the random variables are being used in actuarial practice. Since we work with real data, our random variables are real valued random variables.

Model: A model in actuarial applications is a simplified mathematical description of a certain Actuarial task. Actuarial models are used by actuaries to form an opinion and recommend a course of action on contingencies relating to uncertain future events.

Survival Function: The survival function of a real valued random variable X denoted as S_X is defined as

$$S_X(x) = P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$$

it satisfies the following

i S_X is non increasing

ii S_X is right continuous

iii $\lim_{x\to-\infty} S_X(x) = 1$ and $\lim_{x\to\infty} S_X(x) = 0$

In mortality terms it is the probability that a person will live beyond the age x. The survival function and the cdf are complement of each other. Historically, when random variable is measuring time the survival function is used while when its measuring dollar we use the distribution function.

Probability Density Function and Probability Mass Function:

(1) If X is discrete we have the probability mass function pmf denoted by $p_X(x)$ define as $p_X(x) = P(X = x)$ it satisfies

i $0 \leq p_X \leq 1$

ii $\sum p_X(x) = 1$

The survival function and the cdf can be obtained from the pmf as follows

$$F_X(x) = \sum_{y\leq x} p_X(y) \ \ and \ \ S_X(x) = \sum_{y\geq x} p_X(y)$$

(2) If X is absolutely continuous, then the probability density function pdf denoted as f_X is the derivative of the cdf or negative derivative of the survival function.

$$f_X(x) = F'_X(x) = -S'_X(x)$$

The cdf and survival functions are obtained from the pdf as

$$S_X(x) = \int_x^\infty f_X(x)dx \ \ and \ \ F_X(x) = \int_{-\infty}^x f_X(x)dx$$

Hazard Rate: This is also known as the force of mortality or failure rate denoted by h_X and defined as

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

In mortality terms it is the annualized probability(not probability) that a person of age x will die in the next instant.

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = \frac{-S'_X(x)}{S_X(x)} = \frac{-d}{dx} \ln S_X(x)$$

$$S_X(x) = e^{-\int_0^x h_X(t)dt}$$

If $h_X(x) = k$ then $S_X(x) = e^{-kx}$. This is the survival function of the exponential distribution with parameter $\frac{1}{k}$ thus the survival function is a constant iff the distribution is the exponential.

1. Basic Distributional Quantities

1.1 Skewness: this is a measure of the symmetry of the pdf. A distribution, or data set, is symmetric if it looks the same to the left and right of the mean. The coefficient of skewness is define by

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

If γ_1 this is close to zero then the distribution is symmetric about its mean e.g the normal distribution with skewness equal to zero.

1.2 Kurtosis: This is a measure of peakedness/flatness of a distribution with respect to the normal distribution. A measure of kurtosis is given by the coefficient of kurtosis define by

$$\gamma_2 = \frac{\mu_4}{\sigma^4}.$$

The coefficient of kurtosis of the normal distribution is 3. The condition $\gamma_2 < 3$ indicates that the distribution is flatter compared to the normal distribution, and the condition $\gamma_2 > 3$ indicates a higher peak (relative to the normal distribution) around the mean value.

Coefficient of Variation (CVR): This is the ratio of standard deviation to mean i.e $CVR = \frac{\sigma}{|\mu|}$

If the CVR $< 30\%$ we say that the data is homogeneous and the mean can represent the data.

DEFINITION 3.1. *For a given d with $P(X > d) > 0$, the excess loss variable is*

$$Y^P = \begin{cases} \text{undefined}, & X \leq d, \\ X - d & X > d \end{cases}$$

Its expected value

$$e_X(d) = e(d) = E(X - d | X > d)$$

is called the mean excess loss function also called the mean residual life function and complete expectation of life. Y^P is also called the left truncated and shifted variable. This is because observations below d are discarded, it is shifted because d is subtracted from the remaining values. The kth moment is

$$\begin{aligned} e_X^k(d) &= \frac{\int_d^\infty (x-d)^k f_X(x)dx}{S_X(d)} \text{ Continous} \\ &= \frac{\sum_{x_j>d}(x_j-d)^k p(x_j)dx}{S_X(d)} \text{ Descrete} \end{aligned}$$

We also have for k = 1

$$e_X(d) = \frac{\int_d^\infty S_X(x)dx}{S_X(d)}$$

Definition 3.2. *left censored and shifted variable is define as:*

$$Y^L = (X-d)_+ = \begin{cases} 0, & X \le d, \\ X-d & X > d \end{cases}$$

This variable is that of the per payment variable. The kth moment is given by:

$$\begin{aligned} E(Y^L)^k = E(X-d)_+^k &= \int_d^\infty (x-d)^k f_X(x)dx \text{ Continous} \\ &= \sum_{x_j>d}(x_j-d)^k P(x_j)dx \text{ Descrete.} \end{aligned}$$

Definition 3.3. *The Limited loss variable is given by:*

$$Y = \min(X,u) = X \wedge u = \begin{cases} X, & X \le u, \\ u, & X > u \end{cases}$$

$E(Y) = E(X \wedge u)$ *is called the limited loss. It can be shown that*

$$E(X \wedge u) = \int_0^u S_X(x)dx,$$

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

An insurance phenomenon that relates to this variable is the existence of policy limit that sets a maximum on the benefit paid. From the above, buying one policy with a deductible of d and another with a policy limit d equal buying a full coverage.

Mode, Percentiles and Quantiles:

a. Percentiles and Quantiles: If the cdf F_X is invertible, the inverse function F_X^{-1} is called the quantile function. $F_X^{-1}(\alpha)$ is called the quantile of order α. Now for any cdf F we may define for $u \in (0,1)$,

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}$$

this is called the generalized inverse function. If $x = F^{-1}(\alpha)$ we have $F(x-0) \leq \alpha \leq F(x)$ and x is called the $\alpha-$ quantile

If $\alpha = 0.5$ then x is the median or second quartile A percentile is the value of a variable below which a certain percent of observations fall. For example, if a score is in the 85th percentile, it is higher than 85 of the other scores. For a random variable X and $0 < p < 1$, we define the 100pth percentile (or the pth quantile) as the number x such that $F(x-0) \leq p \leq F(x)$.
For a continuous random variable, this is the solution to the equation

$$F(x) = p.$$

The 25th percentile is also known as the first quartile, and the 75th percentile as the third quartile.

b. Modes: The mode could be defined as the value that maximizes the probability mass function p_X (discrete case) or the probability density function f_X (continuous case.) In the discrete case, it is the value with the largest probability, while in the continuous case, it is the value for which the pdf is largest, if there are local maxima, these points are considered the mode.

Generating Functions and Sums of Random Variables:

Random variables of the form

$$S_n = X_1 + X_2 + \cdots + X_n$$

appear repeatedly in probability theory and applications. For example, in the insurance context, S_n can represent the total claims paid on all policies where X_i is the ith claim. Thus, it is useful to be able to determine properties of S_n. For the expected value of S_n; we have

$$E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

A similar formula holds for the variance provided that the X_i are independent random variables. In this case,

$$Var(S_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

Moment generating function: For a random variable X, the moment generating function (mgf) is define as $M_X(t) = E(e^{tX})$ for

all t for which the expected value exist.

Remark If the derivative of M exist we have

$$M'_X(t) = \frac{d}{dt}E(e^{tX}) = \sum xe^{tx}p(x) \text{ (descrete case)}$$

The mean is obtained from mgf as

$$M'_X(0) = \mu_1 = E(X)$$

continuing this process we get if the derivatives at zero's exist

$$M^k_X(0) = E(X^k)$$

The moment generating function does not exist for some distribution with heavy tails as will be discuss later.

Probability generating function (pgf): The pgf is define as $P_X(t) = E(t^X)$ for all t for which the expectation exists.

Remark If the derivative of P exists we have

$$P'_X(t) = \frac{d}{dt}E(t^X) = \sum xt^{x-1}p(x) \text{ descrete case}$$

The mean is obtained from pgf as

$$P'_X(1) = \mu_1 = E(X)$$

continuing this process we get

$$P^k_X(1) = E(X(X-1)(X-2)\cdots(X-k+1))$$

and

$$P_X^k(0) = k!p_X(k), \text{ Hence } p_X(k) = \frac{P_X^k(0)}{k!}$$

THEOREM 3.4. We have,

(1) If the pgf exist for a random variable X we have
$P_X(t) = \phi_X(-i\ln t)$ *and* $\phi_X(t) = P_X(e^{it})$
(2) If the pgf and mgf exist we have $M_X(t) = P_X(e^t)$, $P_X(t) = M_X(\ln t)$ and $\phi_X(t) = M_X(it)$

THEOREM 3.5. Let $S_n = X_1 + X_2 + \cdots + X_n$ where the X_i are independent. Then $M_{S_n}(t) = \prod_{j=1}^n M_{X_j}(t)$, $P_{S_n}(t) = \prod_{j=1}^n P_{X_j}(t)$ and $\phi_{S_n}(t) = \prod_{j=1}^n \phi_{X_j}(t)$

Proof: Since the X_i are independent. Then

$$M_{S_n}(t) = E(e^{tS_n}) = E(e^{t(X_1+X_2+\cdots+X_n)}) = E(\prod_{j=1}^n e^{tX_j}) = \prod_{j=1}^n E(e^{tX_j})$$

$$M_{S_n}(t) = \prod_{j=1}^n M_{X_j}(t)$$

Similarly argument can be use for the proof of the pgf and the characteristics function.

2. Classifying and Creating Distribution

A deep knowledge of distributions will be needed in modeling our data to get the loss function, hence knowing how to classify and create new distribution is of utmost important. This will be discuss in this section. **A. Classifying Distributions**

1. Parametric and Scale Distributions: A parametric distribution is one that is completely determined by a set of quantities called parameters. The number of parameters is fixed and finite, these are considered the simplest families of actuarial models.

Models that requires less parameters are considered to be simple models while those with more parameters are said to be complex. Examples of commonly used parametric distributions are listed below.

Name	PDF	Parameters
Exponential	$f(x) = \frac{1}{\theta} e^{\frac{-x}{\theta}}$	$\theta > 0$
Pareto	$f(x) = \frac{\alpha \theta^{\alpha}}{(x+\theta)^{\alpha+1}}$	$\alpha > 0, \theta > 0$
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ, σ
Poisson	$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	$\lambda > 0$

DEFINITION 3.6. *A parametric distribution is a scale distribution if when a random variable from the set of distribution is multiplied by a positive constant, the resulting random variable is also in that set of distribution e.g the exponential and Weibull distributions.*

DEFINITION 3.7. *For random variables with nonnegative support, a scale parameter is a parameter for a scale distribution in which when a member of the scale distribution is multiplied by a constant the scale parameter is multiplied by that constant and all other parameters remain unchanged. e.g* θ *in the gamma distribution with pdf*

$$f(x) = \frac{x^{(\alpha-1)} e^{\frac{-x}{\theta}}}{\theta^{\alpha} \Gamma(\alpha)}$$

It is often possible to recognize a scale parameter from looking at the distribution or density function, the distribution function will always have x appears as x/θ. The larger the scale parameter the more spread out the distribution is.

DEFINITION 3.8. *Let F be a cdf on R, define G as follows*

$$\forall x \in \mathbb{R}, G(x) = F\left(\frac{x-b}{a}\right), a > 0$$

We say that G is centred at b and rescaled with parameter a, b is a location. Location parameter determine the location or shift of the distribution.

DEFINITION 3.9. *A random variable X is a k-point mixture of the random variables $X_1, X_2, \ldots X_k$ if its cdf is given by $F_X(x) = a_1 F_{X_1}(x) + a_2 F_{X_2}(x) + \cdots + a_k F X_k(x)$, where each mixing weight $a_i > 0$ and $a_1 + a_2 + \cdots + a_k = 1$: The mixture X is defined in terms of its pdf or cdf and is not the sum of the random.*

A mixture distribution is the probability distribution of a random variable whose values can be interpreted as being derived from an underlying set of other random variables. For example, a dental claim may be from a check-up, cleaning, filling cavity, a surgical procedure, etc. The mixing weights are discrete probabilities. To see this, let Θ be the discrete random variable with support {1, 2 ... k} and pmf $Pr(\Theta = i) = a_i$. We can think of the distribution of Θ as a conditioning distribution where $X = X_i$ is conditioned on $\Theta = i$ or equivalently,

$$F_{X|\Theta}(x|\Theta = i) = F_{X_i}(x).$$

In this case, X is the unconditional distribution with cdf

$$F_X(x) = a_1 F_{X_1}(x) + a_2 F_{X_2}(x) + \cdots + a_k F_{X_k}(x) = \sum_{i=1}^{k} F_{X|\Theta}(x|\theta = i) P(\Theta = i)$$

In actuarial and insurance terms, discrete mixtures arise in situations where the risk class of a policyholder is uncertain, and the number of possible risk classes is discrete.

DEFINITION 3.10. *A variable component mixture distribution has a distribution function that can be written as*

$$F(x) = \sum_{j=1}^{K} a_j F_j(x), \quad \sum_{j=1}^{K} a_j = 1, \quad a_j > 0, \quad j = 1, 2, 3, ..., K, \quad K = 1, 2, ...$$

DEFINITION 3.11. *A data-dependent distribution is at least as complex as the data or knowledge that produced it, and the number of parameters increases as the number of data points or the amount of knowledge increases.*

We consider two-types of data-dependent distributions:

1. The empirical distribution: The empirical cumulative distribution function ecdf, based on the data $x = c(x_i), i = 1, 2, ..., n$, is defined by

$$Fn(t) = \mathbb{P}(x \leq t) = \frac{1}{n} \#\{i \in [1, n]; x_i \leq t\}$$

2. A kernel smoothed distribution: Given an empirical distribution, we wish to create a continuous distribution whose pdf will be a good estimation of the (discrete) empirical distribution.The density function is given by

$$f_X(x) = \sum_{i=1}^{n} p_n(x) k_i(x)$$

where $p_n(x) = \frac{1}{n}$ and $k_i(x)$ is the kernel smoothed density function which will be discuss in the next chapter on estimation of loss function.

3. Tail Weight

The (right-)tail of a distribution is the portion of the distribution corresponding to large values of the random variable. Alternatively, we can define the tail of a random variable X as the interval (x, ∞). A distribution is said to be a heavy-tailed distribution if it significantly puts more probability on larger values of the random variable e.g the Pareto distribution. In contrast, a distribution that puts less and less probability for larger values of the random variable is said to be light-tailed distribution.There are four ways to show that a distribution is heavy tailed or not.

1. The Existence of Moments:

A distribution $f_X(x)$ is said to be light-tailed if the kth raw moments is finite for all $k > 0$ i.e $|E(X^k)| < +\infty$ for all $k > 0$. The distribution $f_X(x)$ is said to be heavy-tailed if either $E(X^k)$ does not exist for all $k > 0$ or the moments exist only up to a certain value of a positive integer k. For example the Exponential distribution is light tailed since the kth moment exist for all $k > 0$ while the Pareto distribution of parameter (α, θ) is heavy tailed as the moment exist for some k such that $-1 < k < \alpha$

2. Classification Based on the Speed of Decay of the Survival Function:

The survival function $S_X(x) = P(X > x)$ captures the probability of the tail of a distribution. Recall that S_X decreases to 0 as $x \to \infty$, The question is how fast the function decays to zero. If the survival function of a distribution decays slowly to zero (equivalently the cdf goes slowly to one), it is another indication that the distribution is heavy-tailed. In contrast, when the survival function decays to 0 very rapidly then this is indication that the distribution is light-tailed.

Now consider two distribution F_X and F_Y with the same mean, we compare their tail weights by computing the ratio of the tail probabilities or the survival functions which we will refer to as the relative tail weight:

$$\lim_{x\to\infty} \frac{S_X(x)}{S_Y(x)} = \lim_{x\to\infty} \frac{S'_X(x)}{S'_Y(x)} = \lim_{x\to\infty} \frac{f_X(x)}{f_Y(x)} \geq 0$$

If the above limit is 0, we say that the distribution of X has lighter tail than Y. If the limit is finite then we say that the distributions have similar or proportional tails. If the limit diverges to infinity, then more probabilities on large values of X are assigned to the numerator, In this case, we say that the distribution X has heavier tail than Y.

3. Classification based on hazard rate: Distribution with decreasing hazard rate are said to be heavy tailed while those with

increasing hazard rate are light tailed. In comparing two distributions, we think of which decreases faster than the other as heavier tailed than the other.

THEOREM 3.12. If $\frac{f_X(t+x)}{f_X(x)}$ is decreasing then $h_X(x)$ is increasing and vice versa.

Proof:

$$\frac{1}{h_X(x)} = \frac{\int_x^\infty f_X(t)dt}{f_X(x)} = \frac{\int_0^\infty f_X(t+x)dt}{f_X(x)}$$

and the result follows.

4. Classification based on the mean residual loss: A fourth measure of tail weight is the mean excess loss function. For a loss random variable X; if the mean excess loss is increasing the distribution is said to be heavy tailed otherwise it is said to be light tailed. Next, we establish a relationship between mean excess loss and the hazard rate function.

$$e_X(d) = \frac{\int_d^\infty S_X(s)}{S_X(d)} = \int_0^\infty \frac{S_X(y+d)}{S_X(d)}$$

and we have

$$\frac{S_X(y+d)}{S_X(d)} = \frac{exp[-\int_0^{y+d} h_X(x)dx]}{\int_0^d h_X(x)dx} = exp[-\int_d^{y+d} h_X(x)dx] = exp[-\int_0^y h_X(d+t)dt]$$

Therefore if the hazard rate is increasing the mean excess loss will be decreasing and if the hazard rate is decreasing, the mean excess loss will be increasing however the converse is not true for example; take $f(x) = (1+2x^2)e^{-2x}$ has a strictly decreasing mean excess loss while the hazard rate is not strictly increasing. Another relationship between the e(x) and the h(x) is given as

$\lim_{x\to\infty} e(x) = \lim_{x\to\infty} \frac{1}{h(x)}$ as long as these limits exist.

The limiting relationships are useful if the survival function, the hazard rate or the mean excess loss are complicated.

Equilibrium Distributions and tail Weight: Let X be a random variable such that $S_X(0) = 1$ Using definition (1.1) with d = 0 we can write $e(0) = E(X) = \int_0^\infty S_X(x)dx$ or equivalently $1 = \frac{\int_0^\infty S_X(x)}{E(X)}$.

Now we define a random variable X_e with pdf

$$f_e(x) = \frac{S_X(x)}{E(X)}, \quad x \geq 0.$$

We call the distribution of X_e the equilibrium distribution or integrated tail distribution. The corresponding survival function is

$$S_e(x) = \int_x^\infty f_e(x)dx = \frac{\int_x^\infty S_X(x)}{E(x)}, x \geq 0$$

The hazard rate function is given by

$$h_e(x) = \frac{f_e(x)}{S_e(x)} = \frac{S_X(x)}{\int_x^\infty S_X(x)dx} = \frac{1}{e_X(x)}$$

Thus, if $h_e(x)$ is increasing then the distribution X_e (and thus X) is light-tailed and if $h_e(x)$ is decreasing then the distribution X_e (or X) is heavy-tailed. The equilibrium mean is given by

$$E(X_e) = \frac{E(X^2)}{2E(X)}$$

$$\frac{e_X(x)}{e_X(0)} = \frac{S_e(x)}{S_X(x)}$$

If the mean residue function is increasing then

$$e_X(x) \geq e_X(0)$$

and $S_e(x) \geq S_X(x)$ integrating both sides we got

$$\frac{E(X^2)}{2E(X)} \geq E(X)$$

this implies $Var(X) \geq E(X)^2$ and $CVR = \frac{Var(X)}{E(X)^2} \geq 1$

B. Creating New Distribution: There are various ways of creating new distributions.

i) Multiplication by a Constant: This transformation is equivalent to applying inflation uniformly across all loss levels and is known as change of scale. For example if this year's losses are given by a random variable X then uniform inflation of r% indicates that next year's losses can be model with the r.v $Y = (1 + 0.01r)X$

THEOREM 3.13. Let X be a continuous r.v with pdf f_X and cdf F_X. Let $Y = \theta X$ with $\theta > 0$. Then
$F_Y(y) = F_X(\frac{y}{\theta}),\ \ f_Y(y) = \frac{1}{\theta} f_X(\frac{y}{\theta})$
The parameter θ is a scale parameter for the r.v Y

ii) Raising to Power: A new distribution is obtained by raising a random variable to a certain power such as $Y = X^{\frac{1}{\tau}}$ or $Y = X^{\frac{-1}{\tau}}$ where $\tau > 0$. In the first case, Y is called transformed. In the second case, if $\tau = -1$, Y is called inverse and if $\tau < 0$ but $\tau \neq -1$; we call Y the inverse transform of X.

THEOREM 3.14. Let X be a continuous r.v with pdf f_X and cdf F_X, $\tau > 0$ we have

- The transformed: Let $Y = X^{\frac{1}{\tau}}$ then
  $F_Y(y) = F_X(y^\tau)$ and $f_Y(y) = \tau y^{\tau-1} f_X(y^\tau)$
- The inverse transformed: Let $Y = X^{\frac{-1}{\tau}}$ then
  $F_Y(y) = 1 - F_X(y^{-\tau})$ and $f_Y(y) = \tau y^{-\tau-1} f_X(y^{-\tau})$
- The inverse: Let $Y = X^{-1}$ then
  $F_Y(y) = 1 - F_X(y^{-1})$ and $f_Y(y) = y^{-2} f_X(y^{-1})$

Example: the transformed exponential distribution is the Weibull distribution and the inverse transformed exponential distribution is the inverse Weibull distribution.

iii) Exponentiation

THEOREM 3.15. Let X be a continuous random variable with pdf f_X and cdf F_X such that $f_X(x) > 0$ for all $x \in \mathbb{R}$ Let $Y = e^X$: Then, for $y > 0$ we have $F_Y(y) = F_X(\ln y)$ and $f_Y(y) = \frac{1}{y} f_X(\ln y)$

Example: Let X be the normal distribution then the lognormal distribution is obtained as $Y = e^X$

iv) Mixing: We will define the continuous version of mixing where the discrete probabilities are replaced by the pdf of a continuous random variable. In actuarial terms, continuous mixtures arise when a risk parameter from the loss distribution is uncertain and the uncertain parameter is continuous. Suppose that Λ is a continuous random variable with pdf f_Λ Let X

be a continuous random variable that depends on a parameter λ.Then the unconditional pdf of X is

$$f_X(x) = \int f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda$$

The pgf is given by

$$P_X(x) = \int P_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda$$

the corresponding probabilities are given by

$$p_k = \int p_k(\lambda) f_\Lambda(\lambda) d\lambda$$

Theorem 3.16. For the mixture distribution X as defined above, we have

a. $F_X(x) = \int F_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda$
b. $E(X^k) = E(E(X^k|\Lambda))$
c. $Var(X) = E(Var(X|\Lambda)) + Var(E(X|\Lambda))$

v) Frailty (Mixing) Models A continuous mixture model that arises within the context of survival analysis is the frailty model. A frailty model is defined as follows: Let $X|\Lambda$ be a random variable with conditional hazard function given by $h_{X|\Lambda}(x|\lambda) = \lambda a(x)$ where $a(x)$ is a known function of x. The frailty random variable Λ is supposed to have positive support.

Theorem 3.17. The conditional survival function is given by $S_{X|\Lambda}(x|\lambda) = e^{-\lambda A(x)}$ where $A(x) = \int_0^x a(t)dt$ and the unconditional survival function is given by $S_X(x) = M_\Lambda(-A(x))$

vi) Splicing: A k-component spliced distribution has a density functions that can be expressed as follows:

$$f_X(x) = \begin{cases} a_1 f_1(x) & c_o < x < c_1, \\ a_2 f_2(x) & c_2 < x < c_3, \\ \vdots & \\ a_k f_k(x) & c_{k-1} < x < c_k \end{cases}$$

For j = 1, 2, ... k, $a_j > 0$ and f_j must be a legitimate density functions will all probabilities on the interval (c_{j-1}, c_j). Given $f_1, f_2, ...f_k$ distributions we can create a k-component spliced distribution below:

$$f_X(x) = \begin{cases} a_1 g_1(x) & c_o < x < c_1, \\ a_2 g_2(x) & c_2 < x < c_3, \\ \vdots & \\ a_k g_k(x) & c_{k-1} < x < c_k \end{cases}$$

where

$$g_j(x) = \frac{f_j(x)}{\int_{c_{j-1}}^{c_j} f_j(x)dx}$$

vii) Limiting Distributions: New distributions can be obtained as limiting cases of other ones by letting the parameters go to either infinity or zero. Example, for a Pareto distribution with parameter α, θ, letting both α and θ go to infinity with the ratio $\frac{\alpha}{\theta}$ held constant, the result is an exponential distribution.

4. Discrete Distribution

Here we introduce a large class of counting distributions, these are distributions with probabilities only on non-negative integers. In insurance context they describe the number of events such as claims, the understanding of both the number of claims and the size of claims, gives us more information about the claims than knowing only about the total losses. We will adopt the following notation:

If N is the random variable representing the number of events (or claims) then the probability mass function or the probability function $Pr(N = k)$ will be denoted by p_k

A. Poisson Distribution The Poisson random variable is most commonly used to model the number of random occurrences of some phenomenon in a specified unit of space or time. For example, the number of phone calls received by a telephone operator in a 10-minute period or the number of typos per page made by a secretary. The pmf for the Poisson distribution is given by

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!} \quad k = 0, 1, 2...$$

The pgf is given by

$$P_N(z) = E(z^N) = \sum_{k=0}^{\infty} p_k z^k = e^{\lambda(z-1)}, \lambda > 0$$

The mean and variance are

$$E(N) = \lambda \text{ and } Var(N) = \lambda$$

The Poisson distribution has the mean equal its variance, hence when working with counting data having same mean as variance we may guess that the distribution follows Poisson law. Poisson distribution has quiet some important properties

THEOREM 3.18. Let $N_1, N_2, ..., N_n$ be n independent Poisson random variables with parameters $\lambda_1, \lambda_2, ..., \lambda_n$ respectively. Then the random variable $S_n = N_1 + N_2 + \cdots + N_n$ is also a Poisson distribution with parameters $\lambda_1 + \lambda_2 + ... + \lambda_n$

The next property is useful in modeling insurance risk. Suppose the number of claims in a fixed period such as 1 year follows the Poisson distribution, suppose the claims can be classified into m-distinct types then each type also follows the Poisson law.

THEOREM 3.19. Let the total number of events N be a Poisson random variable with mean λ. Suppose that the events can be classified into m types with probabilities $p_1, p_2, ...p_m$ independent of all other events. Then the number of event $N_1, N_2, ..., N_m$ corresponding to event type 1, 2, ...m respectively are mutually independent Poisson r.v with means $\lambda p_1, \lambda p_2, ...\lambda p_m$ respectively.

B. Negative Binomial Distribution (NBD) The Poisson distribution is a one-parameter counting distribution while the negative binomial distribution is a two-parameter counting distribution which makes it more flexible in shape than the Poisson. The pmf of the NGB is:

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

where n = 0, 1, 2, ... and $r > 0$, $\beta > 0$. The pgf is given by

$$P(z) = [1 - \beta(z-1)]^{-r}$$

From this it follows that the mean and variance of the NBD are:

$$E(N) = r\beta \ \ and \ \ Var(N) = r\beta(1+\beta)$$

Hence $Var(N) > E(N)$. This suggests that for a particular set of counting data, if the observed variance is larger than the mean, the NBD might be a better candidate than the Poisson distribution as a model. NBD is a generalization of the Poisson in two different ways; namely as a mixed Poisson distribution with a gamma mixing distribution and as a compound Poisson distribution with a logarithm secondary distribution(this will be discuss later).

There, among other assumptions, the rate at which claims occur is assumed to be constant over time. If the rate is linearly increasing with regard to the number of past claims, then the number of claims in any period will follow NBD.

C. Geometric Distribution: This is a special case of the NBD in which r = 1, it is in some sense the discrete analogue of the exponential distribution(both of them have the memoryless property). Hence it pmf, pgf are same as that of the NBD only with r = 1 in this case.

D. Binomial Distribution: Binomial experiments are problems that consist of a fixed number of trials n; with each trial having exactly two possible outcomes: Success and failure. The probability of a success is denoted by q and that of a failure by $1-q$. Also, we assume that the trials are independent, that is what happens to one trial does not affect the probability of a success in any other trial. It posses some properties different from Poisson and NBD, First it has Mean greater than the variance also it has finite support. This may be useful in modeling the number of individual injured in an automobile accident, e.t.c. The pmf is given by

$$p_k = P(N=k) = \binom{n}{k} q^k (1-q)^{m-k}, \;\; k = 0, 1, ..$$

Its mean and variance are given by

$$E(N) = nq > Var(N) = nq(1-q)$$

the pgf is

$$P(z) = [1 + q(z-1)]^n$$

The Bernoulli distribution is special case of binomial in which n=1

5. The (a, b, 0) Class

The Poisson, negative binomial, geometric, and binomial distributions all satisfy the recursive equation.

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= a + \frac{b}{k} \quad k = 1, 2, 3, ... \\ k\frac{p_k}{p_{k-1}} &= ak + b \end{aligned}$$

for some constants a and b. We will denote the collection of these discrete distributions by C(a, b, 0): The table below list the parameters a and b for each distribution together with the probability function at 0. The second expression is a linear function of k with slope a. Thus, if we graph $k\frac{p_k}{p_{k-1}}$ against k we get a straight line that has a positive slope for the negative binomial or geometric distributions, negative slope for the binomial distribution, and 0 slope for the Poisson distribution.

Distribution	a	b	p_o
Poisson	0	λ	$e^{-\lambda}$
Binomial	$\frac{-q}{1-q}$	$(n+1)\frac{q}{1-q}$	$(1-q)^n$
Negative Binomial	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$(1+\beta)^{-r}$
Geometric	$\frac{\beta}{1+\beta}$	0	$(1+\beta)^{-1}$

Panjer and Willmot shows that these are the only possible distribution satisfying this recursive formular.

6. The (a, b, 1) Class

All the four members of C(a, b, 0) have a fixed positive probability at 0. For some models, an adjustment at 0 is sometimes required. We would like to be able to assign any value in the interval [0, 1) to p_0.

DEFINITION 3.20. *Let p_k be the pmf of a discrete r.v, it is a member of C(a, b, 1) provided there exist constants a, b such that*

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= a + \frac{b}{k} \quad k = 2, 3, ... \\ k\frac{p_k}{p_{k-1}} &= ak + b \end{aligned}$$

The only difference is that the recursion start at k = 2 rather than k=1

It is divided into two classes Zero Truncated and Zero Modified distributions. Let $P(z)$ denote the pgf of a member of the (a, b, 0) class;

(1) Zero Truncated Distributions: Here $p_0 = 0$ and

$$p_k^T = \frac{p_k}{1 - p_0}, \quad k = 1, 2, ...$$

The pgf of the Zero Truncated distribution is given by:

$$P^T(z) = \frac{P(z) - p_0}{1 - p_0}$$

(2) Zero Modified Distributions: Here $p_0 \in (0, 1)$ and

$$p_k^M = \frac{(1 - p_0^M)p_k}{1 - p_0}, \quad k = 1, 2, ...$$

The pgf of the Zero Modified distribution is given by:

$$P^M(z) = (1 - \frac{1 - p_0^M}{1 - p_0}) + \frac{1 - p_0^M}{1 - p_0} P(z)$$

This is a two point mixture of the degenerate distribution and the corresponding (a,b,0) member, we also have

$$p_k^M = (1 - p_o^M)p_k^T, \quad k = 1, 2, ...$$

and

$$P^M(z) = p_o^M + (1 - p_o^M)P^T(z)$$

This is a two point mixture of the degenerate distribution and the corresponding zero truncated (a,b,0) member.

Logarithm Distribution: When $r \to 0$ the limiting case of the extended truncated negative binomial distribution is the logarithm distribution with

$$p_k^T = \frac{[\beta(1+\beta)^{-1}]^k}{k\ln(1+\beta)}, \quad k = 1, 2, ...$$

The pgf is given by

$$P(z) = 1 - \frac{\ln[1 - \beta(z-1)]}{\ln(1+\beta)}$$

Hence the C(a,b,1) consists of the C(a,b,0) class, their Zero Truncated, Their Zero Modified and the logarithm and its Zero Modified logarithm distribution.

7. Compound Frequency Models

A larger class of distribution can be created by the process of compounding any two discrete distribution, i.e the new distribution has pgf $P(z)$ written as

$$P(z) = P_N[P_M(z)]$$

where P_N and P_M are pgf of primary and secondary distribution respectively. It can be easily seen that the pgf of all zero

modified distributions is a compound distribution whose primary distribution is the Bernoulli distribution with $q = \frac{1-p_o^M}{1-p_o}$ and the secondary distribution as its corresponding (a,b,0) member.

Notation: We write for compound distribution primary distribution-secondary distribution. e.g Poisson- ETNB distribution is a compound mixture of a primary Poisson and a secondary ETNB distribution.

THEOREM 3.21. Suppose the pgf P_N satisfies $P_N(z,\theta) = B[\theta(z-1)]$ for some parameter θ and some function B, which is independent of $\theta$and the argument z only appear in the pgf as $\theta(z-1)$, Then $P(z) = P_N[P_M(z);\theta]$ can be written as

$$P(z) = P_N[P_M^T(z);\theta(1-f_o)]$$

where $f_0 = P_M(0)$

This shows that adding, deleting or modifying the probability at 0 in the secondary distribution does not add a new distribution because it is equivalent to modifying the parameter θ of the primary distribution.

The compound Poisson distribution has so many important properties. First the pgf of the compound Poisson distribution may be expressed as

$$P(z) = exp[\lambda(Q(z)-1)]$$

Where Q(z) is the pgf of the secondary distribution,

THEOREM 3.22. Suppose that S_i has a compound Poisson distribution with Poisson parameter λ_i and secondary distribution $\{q_n(i); n = 0, 1, 2, ...\}$ for i = 1, 2, 3, ..., k. Suppose also that $S_1, S_2, ..., S_k$ are independent R.V then $S_n = S_1 + S_2 + \cdots + S_k$ also has a compound Poisson distribution with Poisson parameter $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_k$ and secondary distribution $\{q_n; n = 0, 1, 2, ...\}$ where

$$q_n = \frac{\lambda_1 q_n(1) + \lambda_2 q_n(2) + \cdots + \lambda_k q_n(k)}{\lambda}$$

One main advantage of the theorem above is that if we are interested in computing the sum of independent compound Poisson r.v, then we do not need to compute the distribution of each compound Poisson separately but apply the above theorem.

Another properties of the Poisson distribution is that for the mixed Poisson distribution, the variance is always greater than the mean. There is also an important connection between mixed Poisson and compound Poisson distribution

DEFINITION 3.23. *A distribution is said to be infinitely divisible if for all values of n = 1,2,3,... its characteristic function ϕ can be written as*

$$\phi(z) = [\psi(z)]^n$$

where ψ is the characteristics function of some random variable.

In other words, taking the (1/n)th power of the characteristic function still results in a characteristic function. Examples are the Poisson, NBD, gamma, normal, e.t.c. The binomial distribution is not infinitely divisible.

Effect of Exposure on frequency: Suppose that the current portfolio consist of n entities, each of which could produce claims, Let N_j be the number of claims produced by the jth entity. Then $N = N_1 + N_2 + \cdots N_n$ If we assume that the N_j are independent and identically distributed then
$P_N(z) = [P_{N_1}(z)]^n$

Now suppose the portfolio is expected to expand to g entities with the frequency Ng then

$$P_{Ng}(z) = [P_{N_1}(z)]^g = [P_N(z)]^{g/n}$$

thus if N is infinitely divisible the distribution Ng will have the same form as that of N but with modified parameters. The table below summarize the relationship between the discrete distributions discussed.

Distribution	is a special case of	is a limiting case of
Poisson	ZM Poisson	Negative binomial Poisson–binomial Poisson –inv Gaussian Poisson–geometric(Polya Poisson–Poisson(Neyman–
ZT Poisson	ZM Poisson	ZT negative binomial
ZM Poisson		ZM negative binomial
geometric	negative binomial ZM geometric	geometric – Poisson
ZT geometric	ZT Negative binomial	
ZM geometric	ZM Negative binomial	
Logarithm ZM Logarithm		ZT negative binomial ZM negative binomial
binomial	ZM binomial	
negative binomial	ZM negative binomial Poisson – ETNB	
Poisson – inverse Gaussian Polya–Aeppli Neyman – A	Poisson – ETNB Poisson ETNB	 Poisson ETNB

8. Frequency and Severity with Coverage Modification (Insurance Policies)

We are going to give a brief insight on the various insurance policies and the modifications made to the loss function. Loss are been modified so that the insured share part of the loss hence making the insured to be careful of damages and helping

insurance company not to cover unnecessary loss. For purposes of notation, we shall refer to X as the (ground-up) loss amount random variable before any modifications. We shall denote the modified loss amount to be Y and will be referred to as the claim amount paid by the insurer. We will use the notation Y^L to denote the loss-variable and the notation Y^P to denote the payment-variable.

1. Ordinary Deductibles: For an insurance policy, in order for a claim to be paid, a threshold d must be exceeded. That is, the ground-up loss X must exceed d. In this case, the insurance pays the policyholder the amount $X-d$. For loss amount less than d the insurance pays 0. An ordinary deductibles modifies a r.v into either excess loss or left censored and shifted variable, the difference depends on whether the result of applying the deductibles is to be per payment or per loss. The per payment variable is:

$$Y^P = \begin{cases} \textbf{undefined}, & X \leq d, \\ X - d & X > d \end{cases}$$

While the per-loss variable is:

$$Y^L = \begin{cases} 0, & X \leq d, \\ X - d & X > d \end{cases}$$

The pdf, cdf, Survival function, and hazard rate of the per payment variable are:

$$\begin{aligned} f_{Y^P}(y) &= \frac{f_X(y+d)}{S_X(d)}, \ \ y>0 \\ S_{Y^P}(y) &= \frac{S_X(y+d)}{S_X(d)} \\ F_{Y^P}(y) &= \frac{F_X(y+d)-F_X(d)}{S_X(d)} \\ h_{Y^P}(y) &= \frac{f_X(y+d)}{S_X(y+d)} = h_X(y+d) \end{aligned}$$

As a per - payment variable the excess loss variable places no probability at zero while the left censored and shifted variable has a discrete probability at zero as seen below:

$$F_{Y^L}(0) = P(Y^L \leq 0) = P(X \leq d) = F_X(d)$$

Above zero, the density function is

$$f_{Y^L}(y) = f_X(y+d), \ \ y>0$$

The other key functins are for $y \geq 0$

$$\begin{aligned} S_{Y^L}(y) &= S_X(y+d) \\ F_{Y^L}(y) &= F_X(y+d) \end{aligned}$$

For the continous case we have

$$\begin{aligned} E(Y^L) &= \int_d^\infty (x-d) f_X(x) dx = \int_d^\infty S_X(x) dx \\ E((Y^L)^k) &= \int_d^\infty (x-d)^k f_X(x) dx \end{aligned}$$

An alternative to ordinary deductible is the franchise deductible.

2. Franchise Deductibles A franchise deductible modifies the ordinary deductible by adding the deductible when there is a positive amount paid, that is when the loss exceed the threshold the insured is paid the complete loss. The per payment variable is:

$$Y^P = \begin{cases} \text{undefined}, & X \leq d, \\ X & X > d \end{cases}$$

While the per-loss variable is:

$$Y^L = \begin{cases} 0, & X \leq d, \\ X & X > d \end{cases}$$

other quantities for the per payment variable are $f_{Y^P}(y) = \frac{f_X(y)}{S_X(d)}$

$$F_{Y^P}(y) = \begin{cases} 0, & 0 \leq y \leq d, \\ \frac{F_X(y)-F_X(d)}{S_X(d)}, & y > d \end{cases}$$

$$h_{Y^P}(y) = \begin{cases} 0, & 0 < y < d, \\ h_X(y), & y > d \end{cases}$$

$$S_{Y^P}(y) = \begin{cases} 0, & 0 \leq y \leq d, \\ \frac{S_X(y)}{S_X(d)}, & y > d \end{cases}$$

Other quantities for the per loss variable are

$$f_{Y^L}(y) = \begin{cases} F_X(d), & y = 0, \\ f_X(y), & y > d \end{cases}$$

$$F_{Y^L}(y) = \begin{cases} F_X(d), & 0 \le y \le d, \\ F_X(y), & y > d \end{cases}$$

$$h_{Y^L}(y) = \begin{cases} 0, & 0 < y < d, \\ h_X(y), & y > d \end{cases}$$

$$S_{Y^P}(y) = \begin{cases} S_X(d), & 0 \le y \le d, \\ S_X(y), & y > d \end{cases}$$

THEOREM 3.24. For an ordinary deductible d, we have

- $E(Y^L) = E(X) - E(X \wedge d)$
- $E(Y^P) = \frac{E(Y^L)}{S_X(d)} = \frac{E(X)-E(X\wedge d)}{S_X(d)}$

For the franchise deductible we have

- $E(Y^L) = E(X) - E(X \wedge d) + d(1 - F_X(d))$
- $E(Y^P) = \frac{E(X)-E(X\wedge d)}{S_X(d)} + d$

3. The Loss Elimination Ratio and Inflation Effects for Ordinary Deductibles: The loss elimination ratio is the ratio of the decrease in the expected payment with an ordinary deductible to the expected payment without the deductible. Recall

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

It is a decrease of the overall losses and hence can be considered as savings to the policyholder in the presence of deductibles. The expected savings (due to the deductible) expressed as a percentage of the loss (no deductible at all) is called the Loss Elimination Ratio:

$$LER = \frac{E(X \wedge d)}{E(X)}$$

For a continuous loss amounts X we have

$$LER = \frac{\int_0^d S_X(x)dx}{\int_0^\infty S_X(x)dx}$$

Effect of inflation on ordinary deductibles To see the effect of inflation, consider the following situation: Suppose an event formerly produced a loss of 475. With a deductible of 500, the insurer has no payments to make. Inflation of 12% will increase the loss to 532 and thus results of insurer's payment of 32. A 32% increase in the cost to the insurer.

Theorem 3.25. Let loss amounts be X and let Y be the loss amounts after uniform inflation of r: That is, Y = (1 + r)X: For an ordinary deductible of d, the expected cost per-loss is

$$E(Y^L) = (1+r)[E(X) - E(X \wedge \frac{d}{1+r})]$$

The expected cost per-payment is:

$$E(Y^P) = [1 - F_X(\frac{d}{1+r})]^{-1}(1+r)[E(X) - E(X \wedge \frac{d}{1+r})]$$

4. Policy Limits: If a policy has a limit u; then the insurer will pay the full loss as long as the losses are less than or equal to u otherwise, the insurer pays only the amount u. Thus, the insurer is subject to pay a maximum covered loss of u. Let Y denote the claim amount random variable for policies with limit u. Then

$$Y = \min(X, u) = X \wedge u \begin{cases} X, & X \leq u, \\ u, & X > u \end{cases}$$

We call Y the limited loss random variable. Its cdf and pdf are :

$$F_Y(y) = \begin{cases} F_X(y), & y \leq u, \\ 1, & y \geq u \end{cases}$$

$$f_Y(y) = \begin{cases} f_X(y), & y < u, \\ 1 - F_X(u), & y = u \end{cases}$$

The effect of inflation: Let E(X) be the expected cost before inflation. Suppose that the same policy limit applies after an inflation at rate r, then the after inflation expected cost is given by

$$E((1+r)X \wedge u) = (1+r)E(X \wedge \frac{u}{1+r})$$

5. Combinations of Coinsurance, Deductibles, Limits, and Inflation: Another coverage modification is the use of coinsurance factor. In a policy with a coinsurance factor α, $0 < \alpha < 1$, the insurer portion of the loss is X and the insured portion is $(1-\alpha)X$. For a policy subject to a coinsurance factor α, an ordinary deductible d, policy limit $u*$, and uniform inflation at rate r applied only to losses, we define the maximum covered loss to be $u = u* + d$. Thus, in the absence of a deductible the maximum covered loss is just the policy limit.the claim amount per loss is given by

$$Y^L = \begin{cases} 0, & X < \frac{d}{1+r}, \\ \alpha[(1+r)X - d], & \frac{d}{1+r} \leq X < \frac{u}{1+r} \\ \alpha(u-d), & X \geq \frac{u}{1+r} \end{cases}$$

Theorem 3.26. The expected value of the per loss random variable is

$$E(Y^L) = (1+r)\alpha[E(X \wedge \frac{u}{1+r}) - E(X \wedge \frac{d}{1+r})]$$

The expected value of the per-payment random variable is

$$E(Y^P) = \frac{E(Y^L)}{1 - F_X(\frac{d}{1+r})}$$

for the per loss variable,

$$E[(Y^L)^2] = (1+r)^2\alpha^2\left[E\left(X \wedge u^2*\right) - E\left(X \wedge d^2*\right)\right] - 2d*E(X \wedge u*) + 2d*E(X \wedge d*)$$

Where $u* = \frac{u}{1+r}$ and $d* = \frac{d}{1+r}$. For the second moment of the per payment variable, divide this expression by $1 - F_X(d*)$.

The Impact of Deductibles on the Number of Payments: Deductibles always affect the number of payments. For example, when an imposed deductible is increased the number of payments per period is decreased whereas decreasing deductibles results in an increasing number of payments per period. Let X_j denote the jth ground-up loss and assume that there are no coverage modifications. Let N^L be the total number of losses. Now, suppose that a deductible is imposed. Let $v = Pr(X > d)$ be the probability that a loss will result in a payment. We will let I_j be the indicator random variable whose value is 1 if the jth loss occur (and thus results in a payment) and is 0 otherwise. Then I_j is a Bernoulli random variable such that $Pr(Ij = 1) = v$ and $Pr(Ij = 0) = 1 - v$ The corresponding pgf is $P_{I_j}(z) = vz + 1 - v$. Now, let N^P be the total number of payments. Then $N^P = I_1 +$

$I_2+\cdots+I_{N^L}$. We call N^P a compound frequency model or aggregate claims model. If $I_1; I_2 ... I_j, N^L$, are mutually independent then N^P has a compound distribution with N^L as the primary distribution and a Bernoulli secondary distribution . Thus

$$P_{N^P}(z) = P_{N^L}[1 + v(z-1)]$$

In the important case in which the distribution of N^L depends on parameter θ such that

$$P_{N^L}(z) = P_{N^L}(z;\theta) = B[\theta(z-1)]$$

then $P_{N^P}(z) = B[v\theta(z-1)] = P_{N^L}(z; v\theta)$. This implies that N^L and N^P are both from the same parametric family with only a change in the parameter

CHAPTER 4

Estimation of Loss Function

Insurance claim are raw data hence its important to know the distribution that these data follow, an accurate estimation of the claims of the clients for one product is a sine-qua-none condition to avoid a ruin. In other words, it allows the owner of the company to fix a premium (money paid by the client when contracting the insurance) enabling a non-negative surplus for the company. In this chapter, we discuss how to estimate the loss function from the data, thereby helping us in the calculation of the premium for the insurance company. We would like to fit these data and have a certain degree of confidence in the result obtained.

The general concern in Mathematical Statistics is the following:

Based on the observations, how can we be sure that a random phenomena satisfies certain properties with a high probability of confidence?

1. Mathematical formulation

Suppose that we have a random variable $\mathbf{X} = (X_1, \cdots, X_n)^t$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in some measurable space $(E, \mathcal{B})$.

The random vector X has a probability law $\mathbb{P}_X = \mathbb{P}X^{-1}$. Suppose that we have an observation $\mathbf{x} = (x_1, \cdots, x_n)^t$ on $(E_0, \mathcal{B}_0)$.

The problem becomes : How can we estimate $\mathbb{P}_X$ from the data $\mathbf{x} = (x_1, \cdots, x_n)^t$? How can we define a random variable X which $\mathbf{x}$ follows?
Until know, a big part of Statistical methods are based on the following main hypothesis :

(HG) the data $x_1, \cdots, x_n$ are independent observations of a same random variables of $X*$ of probability law $\mathbb{P}_0$ and taking values in some measurable space $(E_0, \mathbb{B}_0)$.

Now following the hypothesis above, $\mathbf{X} = (X_1, \cdots, X_n)^t$ where $X_i's$ takes values in $(E_0, \mathbb{B}_0)$ and follow same probability law as $X*$, then since the $X_i's$ are independent then

$$\mathbb{P}_X = \mathbb{P}_{(X_1, \cdots, X_n)} = \mathbb{P}_0^{\otimes n}$$

Finally, estimating $\mathbb{P}_{(X_1, \cdots, X_n)}$ reduces to estimating $\mathbb{P}_0$.

The basic idea of the frequentist approach under the general hypothesis is estimate the probability of occurrence of any event

$A \in \mathcal{B}_0$ to its empirical probability law. Given n observations, the empirical probability law is given by

$$\begin{aligned}\mathbb{P}_n(A) &= \frac{1}{n}\#\{i \in [1,n],\ \ x_i \in A\},\ \ A \subset E_0. \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(A) \\ &= \frac{1}{n}\nu(A)\end{aligned}$$

where ν is the counting measure and $\delta_{X_i}(A) = 1_A(x_i)$. Now we use a fundamental theorem in probability theory.

Fundamental Theorem of Frequentist approach of Statistics (Glivenko-Cantelli, 1904).

THEOREM 4.1. If X_1, X_2, $\cdots$ is an infinite sequence of independent and identically distributed(iid) real-valued random variables on common *cdf* F_0 and defined on the same probability space, the sequence of empirical *cdf*,

$$F_n(x) = \frac{1}{n}\#\{i \in [1,n], X_i \leq x\}, = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,\infty)}(x),\ \ x \in \mathbb{R}$$

uniformly converges to

$$F_0(x) = \mathbb{P}(X \leq x),\ \ x \in \mathbb{R}$$

that is

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| \to 0\ as\ n \to +\infty a.s.$$

By strong law of large number we have point wise convergence but Glivenko and Cantelli strengthen this result by proving uniform convergence. Hence by convergence theorem we have that

$$\mathbb{P}_n \rightsquigarrow \mathbb{P}_0.$$

where, $\rightsquigarrow$ stands for weak convergence or convergence in distribution.

Paradigm. Theorem tells us that the exact probability law of the observed random variable X (characterized by F_0) is known if we have an infinite number of observations though the uniform limit of the empirical *cdf*. Based on that remarkable limit of the sequence of empirical probability measures constructed using frequencies of events in the samples $\{X_1, \cdots, X_n\}$, $n \geq 1$, the intuitive recommendation that emerges is the following :

Given a sample $\{x_1, \cdots, x_n\}$ of observations, the probability $\mathbb{P}_n$ **can be taken as an approximation** of $\mathbb{P}_0$ i.e

$$\mathbb{P}_n \approx \mathbb{P}_0.$$

More generally, any parameter of $\mathbb{P}_0$ of the form

$$\theta = \int_{E_0} h(x) \ d\mathbb{P}_0, \ \ with \ \ h : (E_0, \ \mathcal{B}_0) \to \mathbb{R} \in L^1$$

is estimated by replacing $\mathbb{P}_0$ by $\mathbb{P}_n$, that is

$$\theta_n = \int_{E_0} h(x) \ d\mathbb{P}_n = \int_{E_0} \frac{1}{n} h(x) \ d\nu = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

Parameters	Exact Parameters	Empirical Parameters
Mean	$m_X = \mathbb{E}(X) = \int_{\mathbb{R}} x \ \ d\mathbb{P}_X(x)$	$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$
MNC(k)	$m_k(X) = \int_{\mathbb{R}} x^k \ \ d\mathbb{P}_X(x)$	$m_k(x) = \frac{1}{n}\sum_{i=1}^n x_i^k$
MC(k), $k \geq 2$	$\mu_k(X) = \int_{\mathbb{R}} (x - m_X)^k \ \ d\mathbb{P}_X(x)$	$\mu_k(x) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^k$
variance $(n \geq 2)$	$\sigma^2 = \int_{\mathbb{R}} (x - m_X)^2 \ \ d\mathbb{P}_X(x)$	$esd(x)^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$
Quantile	$t(X, \alpha) = F_0^{-1}(\alpha)$	$t(x, \alpha) = (F_n)^{-1}(\alpha)$
cdf	$F_0(x)$	$F_n^(x)$
pdf $\vdots$	f_X $\vdots$	f_n $\vdots$

TABLE 1. Some correspondence between Exact Parameters and Empirical Parameters

The method of replacing $\theta(\mathbb{P}_0)$ by $\theta(\mathbb{P}_n)$ is called a *plug-in* method.

Consequently the parameters of the probability law $\mathbb{P}_X$ of X **should be approximated by** the parameters of the empirical probability $\mathbb{P}_n$. This give an intuitive way for approximating parameters which is explained in the table below:

1.1 Quality of approximation.

The previous developments give an intuitive way for approximating parameters of probability laws : require data and take an empirical analog as estimator, which is called plug-in estimator.

But this does not lead to good approximations automatically. In some cases, we need to have a large sample size to reach a

good approximation. In some other cases, we do not have the convergence of the empirical parameter to the true parameter (if it exists).

The Mathematical Sciences and Probability Theory as well deal with both theoretical convergence questions and optimum questions in those approximation.

The frequentist approach resides in the fact that the n independent and identically observations produce an empirical probability measure $\mathbb{P}_n$ which can be taken as an approximation of the underlying probability measure $\mathbb{P}_0$, for large values of n at least.

2. Intuitive view of Statistical inference

1. General intuitive principle.

From the precedent developments, we have the following fact under certain hypotheses :

(A) A class of properties of a random variables X is approximated more or less accurately by the same properties on a sample of size n from X, meaning that that properties on a sample of size n from X are *similar* or *almost the same* as the same properties for X.

(A)	**Given** a class of properties of X	approximation by $\Longrightarrow$	the corresponding properties on a sample from X
(B)	What Probability law has similar properties	Inference $\Longleftarrow$	**Given** a class of properties on a sample from X

TABLE 2. Approximation and Inference

We may use the inverse sense in the following statement

(B) Suppose that we have a set of observations which we may suppose to be independent and same law observations of a random variable X. Suppose that we observe a class of empirical properties from the data. *We may and do consider* the hypothesis that X is the random variables for which the class of exact properties are near or similar to the empirical ones. By doing so, we say that we are inferring the probability law of X from the data. This is the case with estimation of loss function.

The two assertions (A) and (B) are represented in Table 2

The intuitive statistical inference described above is appealing. But the notions of similarity is still not precised. More importantly, it is obvious that applying principle (B) is more efficient if we can optimize the *similarity* or the *nearness*.

Each inference work is associated to a statistical test which determines the level of confidence of the inference, or simply the quality of the inference.

3. Intuitive view of Statistical tests

a. From Principle (B), we may have an idea of possible probability laws of X. Sometimes, our guess depend on common knowledge. For example, we know that accidental phenomena (and thus having the lack of memory property) may be modeled by exponential laws. As well, life spans of humans can be modeled by Gaussian laws in many cases. Etc.

The most powerful sources of ideas of possible probability laws of X are certainly **descriptive statistics studies or exploratory studies** : The more we explore the data, the deeper knowledge about the probability law of X we extract from the data.

From our knowledge of probability laws and the statistical exploratory studies on the current data, we may have reasons to believe that some fact about the probability law beneath the data should be true. Consequently, we take that fact as an hypothesis **(H)**. In the third step, we have to validate that hypothesis or to reject it.

b. Unfortunately, an hypothesis **(H)** is very difficult to check. For example, let us consider the hypothesis **(H)** : The data come from a $\mathcal{N}(0,1)$ law. We can never check the whole properties from

the data.

But, we usually have one or several consequences **(C)** from the data which are observable from the data and are peculiar to the distribution. We summarize that principle as under a *consistency theorem* :

THEOREM 4.2. ([CT] - Consistency Theorem [exact or asymptotic]) $(H) \to (C)$.

c. Examples.

If we assume that X is Gaussian, that is $\mathbb{P}_X = \mathbb{P}_0$ has the *cdf*

$$F_0(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \ dt, \ \ x \in \mathbb{R}.$$

To test this hypothesis we may use the skewness and the kurtosis parameters, since we know that a standard normal distribution has skewness and kurtosis equal 0 and 3 respectively. We compute the emperical skewness and kurtosis and check if they converge to 0 and 3 respectively.

4. Validating Hypothesis

Now we discuss method of validating hypothesis. Hypothesis can be validated by finding the rejection or critical region, computing the p-value or finding the confidence interval as discuss below.

a. Rejection region or critical region

Frequentist statisticians reason as follows. Since Hypothesis (H) implies the consequence (C), $\neg C$ implies $\neg H$ so if (C) is false, reject (H). Is (C) is true, we do not have any reason to reject (H) : we accept (H) until further tests are preformed.

Let us denote $C = C(X_1, \cdots, X_n)$ since (C) is observable through the observation $X_j = x_j,\ \ j \in [1, n]$.

$\bowtie$ Performing a statistical hypothesis (H) **consists in choosing a reject region also named as critical region**, usually based on a consequence entirely known with the observations, in the form.

$$R = (\neg C) = \{\omega \in \Omega, C(X_1(\omega), \cdots, X_n(\omega)) is false\}.$$

Outside R, we cannot reject and thus we accept, unless we have supplementary information and we proceed to a new test.

(Errors of test). We usually test an hypothesis (H) against its negation $(\neg H)$, called the alternative hypothesis (Ha). In some cases, (Ha) may not be the negation of (H). By alternative hypothesis, we mean that if (H) is rejected, then (Ha) is accepted.

By choosing a rejection region R for the test of (H) against (Ha), we risk to commit two kinds of errors, all of them being conditional probabilities.

Error of the first kind or Type I error : rejecting (H) while (H) is true

$$\alpha = \mathbb{P}(R|H)$$

Error of the second kind or Type II error : accepting (H) while (Ha) is true

$$1 - \beta = \mathbb{P}(\neg R|Ha)$$

We have the following facts from the course of Mathematical Statistics.

(1) It is not possible lower both errors simultaneously. So we have to lower the most damaging error. Usually the test is designed such that α is the error we search to have low. So we say that we perform a test of level α. In most of the cases, we take $\alpha = 5\%$.

(2) If we fix $\alpha = 5\%$ and require that

$$\mathbb{P}(R|H) \leq \alpha,$$

and any test for which this inequality is true is called a test of level α and we put the subscript α to R to have

$$\mathbb{P}(R_\alpha|H) \leq \alpha \qquad (RGA1)$$

Remark. In may cases, the set $\mathbb{P}(R_\alpha|H)$ is non-increasing with α and, if possible, we determine the critical region form the equation

$$\mathbb{P}(R_\alpha|H) = \alpha.(RGA2)$$

We say that R_α is the rejection region at level α. Once α is fixed and R_α determined, we observe the data, check if the consequence (C) is observed and say whether or not we accept (H). The maximum probability of making type I error given that the null hypothesis is true is called **the level of confidence.**

But, finding R_α from Equation (RGA2) requires knowing the probability law to be used in computing $\mathbb{P}(R_\alpha|H)$. So we need a result of probability law or of a convergence in law.

Summary. Performing a statistical test is based on two steps :

Step 1. A consistency result from which we draw an observable event (C) and form the rejection region as $R = (\neg C)$.

Step 2. A probability law or limit law result which allows to precise the rejection region R_α at level α.

b. Notion of p-value of a test.

A great deal of statistical test are based on critical regions of the form

$$R_c = (T > c)$$

where T is some statistic. Let us denote its *cdf* by G. We suppose that G is strictly increasing. So the critical region at level α is given by

$$\alpha = \mathbb{P}(T > c) = 1 - G(c) \Leftrightarrow c = c_\alpha = G^{-1}(1-\alpha)$$

Hence, the is

$$R_{c_\alpha} = (T > G^{-1}(1-\alpha)).$$

where G^{-1} is the quantile function Now suppose that we have the

data and observe T as t_{obs}. We compute

$$p = \mathbb{P}(T > t_{obs}) = 1 - G(t_{obs}).$$

Since $1 - G$ is strictly decreasing, we have :

(1) $p \geq \alpha$ implies that $t_{obs} \leq c$ and thus, the hypothesis is accepted at level α.

(2) $p < \alpha$ implies that $t_{obs} > c$ and thus, the hypothesis is rejected at level α.

Conclusion. If the critical region of a statistical test is of the form $(T > c)$ where the *cdf* G of T is strictly increasing, the number $p = \mathbb{P}(T > t_{obs})$ is called the p-value of the statistical test. We accept the hypothesis (H) at level α if and only if $p \geq \alpha$.

c. Confidence Domain.

Another way to proceed to statistical tests is using confidence sets or domains. Let us suppose that we have prior information (based of our experience or the experience of others, or our descriptive statistics study) which led us to test the hypothesis that the data come from a parametric model $\mathbb{P}_\theta$, where the parameter $\theta = (\theta_1, \cdots, \theta_k)$ is a vector of dimension $k \geq 1$. Knowing the law requires knowing a specific value of the θ. So the problem is to estimate θ by a statistic $\hat{\theta}_n$ based on the data.

In many cases, it is possible that for a fixed level $\alpha \in]0, 1[$, we get a domain $D(\alpha) = D(\hat{\theta}_n, \alpha)$ of $\mathbb{R}^k$ which is based on $\hat{\theta}_n$ and is such that

$$\mathbb{P}(\theta \in D(\hat{\theta}_n, \alpha)) \geq 1 - \alpha.$$

The so-called $(1-\alpha)$-Confidence Domain $D(\hat{\theta}_n, \alpha)$ is interesting only for small values of α (usually $\alpha = 5\%$) and when it has small diameter, such that $\|\theta' - \theta\|$ is small for any pair $(\theta', \theta) \in D(\alpha)^2$ i.e,

$$\left\|\hat{\theta}' - \theta\right\| \leq c_n$$

and c_n is small enough.

In such situations, we may do two things.

(1) We may take any value θ_0 in $D(\alpha)$ as an estimator of θ.

(2) If we want to test the hypothesis $(H) : \theta = \theta_0$ with $\theta_0 \in D(\alpha)$, and we take the rejection region as

$$(RC3) \quad : \quad (\theta \notin D(\alpha)).$$

It is clear that this test is of level α since

$$\begin{aligned} \mathbb{P}((\theta \notin D(\alpha))|H) &= \mathbb{P}((\theta_0 \notin D(\alpha)) \\ &\leq \alpha, \end{aligned}$$

(3) In many cases, $D(\alpha)$ can be taken as a closed ball $B_n(\theta_0, b_n, \alpha)$ of radius $b_n/2$. $(RC3)$ becomes

$$(RC3) \quad : \quad (|\theta_n - \theta| > c_n). \qquad (CI1)$$

Hence, the p-value of the tests is

$$p = \mathbb{P}(|\theta_n - \theta| > t_0). \quad (CI2)$$

where t_0 is the observed value of $|\theta_n - \theta|$.

In the specific case of $\mathbb{R}$, we work with confidence interval (CI) so that the results *CI1* and *CI2* hold.

The above final remarks do link confidence intervals, p-values and critical region, which are three way of validating an hypothesis.

d. Simulations.

Once a statistical test is completely designed, it can be applied to available data. But it is **strongly** advised to *test* the statistical test itself, meaning checking if the statistical test can detect cases for which we already know that (H) is valid. Especially when the rejection region is obtained through an approximation.

The most famous way to do this is to proceed to a Monte-Carlo Study. Here, we use generated data such that (H) is valid and we apply the statistical test and check if it works.

But we should not forget that the rejection region is determined with a possible error of α, say $\alpha = 5\%$. So, if the test fails while H is true, we cannot conclude that the test is inaccurate since of one the 5% of erroneous cases may happened.

So we have to iterate the application of the test. We make B individuals applications and take the proportion p_B of successful tests. If p_B is less than α, we conclude that the *statistical test* is accurate and recommend it to users in similar environments.

Simulating (RC1). This simple simulation works as follows.

0. Fix α
1. Fix B large enough to ensure stability of the conclusion
2. Generate data for which (H) is valid.
3. Apply the statistical test.
4. Repeat Steps 2 an 3 B times
5. Report the proportion p_B of successes
6. Conclude that the test is accurate if $1-p_B$ is less than α.

5. Empirical probability density functions

1. Understanding the curve of the histogram.

Considering histogram of a continuous data over a partition

$$x_{(0)} - \varepsilon = c_0 < c_1 < \cdots < c_m = x_{(m)} + \varepsilon.$$

The number $\varepsilon > 0$ is a negligible and is used to ensure that all the classes $C_j =]c_{j-1},\ c_j]$, $j \in [2, m]$ and $C_0 = [c_o, c_1]$, catch all the observation x_i, $i \in [1, n]$.

$$f_j = \frac{1}{n}\#\{i \in [1,n],\ \ x_i \in C_j, j \in [1,m]\} \text{= area of rectangle of histogram}$$

Let x_j^* be the midpoint of c_j and c_{j-1} hence $c_j = x_j^* + h_j/2$ and $c_{j-1} = x_j^* - h_j/2$ where $h_j = (c_j - c_{j-1})$ the amplitude of the class C_j and we have

$$\begin{aligned} f_j &= \frac{1}{n}\#\{i \in [1,n],\ \ x_i \in]x_j^* - h_j/2,\ \ x_j^* + h_j/2]\} \\ &= \sum_{i=1}^{n} \frac{1}{n} 1_{]x_j^* - h_j/2,\ \ x_j^* + h_j/2]}(x_i) \\ &= \sum_{i=1}^{n} \frac{1}{n} 1_{]-1/2,\ \ 1/2]}\left(\frac{x_i - x_j^*}{h_j}\right) \end{aligned}$$

The height of that rectangle is $\ell_j = f_j/h_j$.

So the area of all the rectangles in the the histogram is the area under the following curve :

$$\begin{aligned} \hat{f}_n(x) &= \ell_j \ \ for \ \ x \in C_j,\ \ j \in [1,m] \\ &= \sum_{j=1}^{m} \ell_j 1_{C_j} = \sum_{j=1}^{m} \frac{f_j}{h_j} 1_{C_j} \end{aligned}$$

Let us apply 5.1 to the middles of classes $x = x_j^*$, It is clear that for a fixed j and have

$$\hat{f}_n(x_j^*) = \sum_{j=1}^{j=m} \Big(\sum_{i=1}^{n} \frac{1}{nh_j} 1_{]-1/2,1/2]} \left(\frac{x_i - x_j^*}{h_j} \right) \Big) 1_{C_j} \quad (DA).$$

The function

$$f_n(x) = \sum_{j=1}^{j=m} \Big(\sum_{i=1}^{n} \frac{1}{nh_j} 1_{]-1/2,1/2]} \left(\frac{X_i - x}{h_j} \right) \Big) 1_{C_j} \quad (DG)$$

coincides with $\hat{f}_n$ on the set of the midpoints of the classes $\mathcal{M} = \{x_i^*, \cdots, x_m^*\}$.

Now let us suppose that the subdivision is uniform : $h_j = h$ for all $j \in [1, m]$, we get

$$f_n(x) = \sum_{i=1}^{n} \frac{1}{nh} 1_{]-1/2,1/2]} \left(\frac{X_i - x}{h} \right) (DU)$$

The above function is usually considered as an estimator of the *pdf*. The number $h = h_n$ is called the bandwidth of the estimator. The following value is recommended (see **?**)

$$h_n = 1.06 \hat{\sigma} n^{-1/5}, \hat{\sigma} = min(sd(x), Q/1.34)$$

where $Q = q_3(x) - q_1(x)$ is the inter-quartile range.

In general, the indicator function

$$K(x) = 1_{]-1/2,1/2]}.$$

Boxcar Kernel	$K(x) = 1_{[-1/2,1/2]}$
Gaussian Kernel	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Epanechnikov	$K(x) = \frac{3}{4}(1-x^2)1_{\|x\|\leq 1}$
Tricube Kernel	$K(x) = \frac{70}{81}(1-\|x\|^3)^3 1_{\|x\|\leq 1}$

TABLE 3. Four kernel function of frequent use

may be replaced by any function $K : \mathbb{R} \mapsto \mathbb{R}_+$ called a Kernel satisfying the following condition :

$$\int_{\mathbb{R}} K(x)\ dx = 1,\ \int_{\mathbb{R}} xK(x)\ dx = 0,\ \sigma_K = \int_{\mathbb{R}} x^2 K(x)\ dx = 1 x \in]0, +\infty[.$$

Here are four kernel of frequent use in Table 3.

The boxcar Kernel is related to the histogram as explained above. But the curve we get is not smooth, in the sense it is not differentiable. The others are more interesting as smooth functions. In that sense, K is called a smoother.

(A) The function define in (DU) is named as the Parzen estimator of f.

(B) The curve corresponding to area of the histogram $\hat{f}_n$ coincides with the estimator f_n of the *pdf* on $\mathcal{M} = \{x_j^*, \cdots, x_m^*\}$, the set of middle points of the classes.

(C) The linear interpolation of the points $(x_j^*, \hat{f}_n(x_j^*) = (x_j^*, f_n(x_j^*)$ gives a curve p_n that we name as the polygon of frequencies. Under regularity condition, the polygon of frequencies is an efficient estimator of the *pdf*.

So the histogram, when completed with the polygon of frequencies, helps in finding the true distribution of the data.

Let us do some remarks of the empirical *cdf* and how to relate to the description above.

2. Variation of the *ecdf*.

The empirical distribution function may written as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,x]}(X_i)$$

We have the following growth rate

$$\Delta(F_n, h) = \frac{F_n(x + h/2) - F_n(x - h/2)}{h}$$

is the empirical analog of the exact growth rate

$$\Delta(F, h) = \frac{F(x + h/2) - F(x - h/2)}{h}$$

If the *cdf* F related to the observations is differentiable, that is F has a probability density function f, we have

$$\Delta(F,h) = \frac{F(x+h/2) - F(x-h/2)}{h} \to f(x) \ \ as \ \ h \to 0.$$

Also $\Delta(F_n, h) = f_n(x)$.

If we do not use a uniform subdivision, the function f_n may be obtained from

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} F_{ni}(x), \quad F_{ni}(x) = 1_{]-\infty,x]}(X_i)$$

by taking the growth rate of each F_{ni} over an amplitude of variation of h_j

$$\hat{f}_n(x_j^*) = \sum_{j=1}^{j=m}\left(\sum_{i=1}^{n}\frac{1}{nh_j}\Delta(F_{ni}, h_j)\right)1_{C_j} \ \ (DG)$$

and the linearly interpolated function may still be considered as estimator of the *pdf*.

3. Conclusion. The *pdf* is usually estimated by the Parzen estimator :

$$f_n(x) = \sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{X_i - x}{h}\right)(DU)$$

6. Continous Data Modeling

Here, we restrict ourselves to continuous data. Indeed count data (non-negative integer data) requires specific methods. However, some methods presented here apply to discrete and continuous data. First we give outlines about parameter estimations.

1. Parameter estimation

Suppose we have reasons to believe that the data are parametric, with parameter $\theta = (\theta_1, \cdots, \theta_r)$ *iid* having a *cdf*

$$\{F_\theta,\ \theta \in \Theta\}$$

Let us consider the hypothesis

$$(H)\ x_1, \cdots, x_n\ come\ from\ F_\theta,\ \theta \in \Theta.$$

Suppose that (H) is true. So θ should take a value θ_n such that the empirical properties of the data are similar to those of F_{θ_n}. And this remark leads to two sources of estimation.

A. First source : the moment method. If (H) is true, the empirical moments should be approximately equal to the exact moments. Let us suppose that the observed random variable is X has r parameters. We equate the first r emperical moment to the exact non centred moment (since they should be approximately

equal) and solve the r-system of equation to find the r parameters

$$\left\{\begin{array}{lcl} \overline{X}_n & = & m_1 = \mathbb{E}(X) \\ s_n^2(x) & = & \mu_2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ \mu_3(x) & = & \mu_3 = \mathbb{E}(X - \mathbb{E}(X))^3 \\ \vdots & \vdots & \vdots \\ \mu_r(x) & = & \mu_r = \mathbb{E}(X - \mathbb{E}(X))^r \end{array}\right. . \tag{6.1}$$

The estimators obtained by this method are called Moment Estimators (MOE). We can use centred or non centred moments except the first moment which is used as non-centered.

B. Second source : the maximum likelihood method. This method works for parametric families having *pdf*'s with respect to a σ-finite measure ν. For now, we focus on absolutely continuous *cdf*'s on $\mathbb{R}$, having *pdf*'s f_θ with respect to Lebesgue measure. Let us denote $X^{(\theta)}$ a random variable wixh *cdf* F_θ.

We recall that, as $\Delta x \to 0$,

$$\frac{F_\theta(x + \Delta x/2) - F_\theta(x - \Delta x/2)}{\Delta x} = \frac{\mathbb{P}(X^{(\theta)} \in [x - \Delta x/2,\ \ x + \Delta x/2])}{\Delta x} \to f_\theta(x)$$

Let us denote $I(x, \Delta) = [x - \Delta x/2,\ \ x + \Delta x/2]$. For Δx small, we may denoted the event

$$(X^{(\theta)} \approx) = (X^{(\theta)} \in I(x,\ \Delta)).$$

If (H) holds and θ_0 is the true value of θ, we have, for small values of Δ,

$$\mathbb{P}(X \in I(x, \Delta)) \approx f_{\theta_0}(x)\Delta x$$

Given the observation x of X :

(a) We have that the probability of $X_{\theta^0} \approx x$ is proportional to $f_{\theta_0}(x)$.

Is clear that :

(b) if $\theta \neq \theta_0$, $X^{(\theta)}$ will certainly distance itself from x, so that the event $X_\theta \approx x$ is less likely and finally $f_\theta(x)$ [which is proportional to $f_\theta(x)$], is less than $f_{\theta_0}(x)$.

By combining (a) and (b), we understand that the true value should be the value of θ which maximizes f_{θ_0}. So we take as estimator, if it exists,

(6.2)
$$\hat{\theta}_n = Arg \max_{\theta \in \Theta}\ f_\theta(x).$$

Given the data x, the function

$$\theta \to f(\theta|x) = f_\theta(x),$$

is called the likelihood function (in θ) given the observation x. As a consequence, the estimator in (6.2) is the Maximum Likelihood Estimator (MLE) of θ.

Given n-data of $iid \sim f_\theta$, the pdf is

$$f_{\theta,n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

To maximize $f_{\theta,n}(x_1, \cdots, x_n)$, a frequent method consists in maximizing (in θ) the following log-likelihood function

$$\log f_{\theta,n}(x_1, \cdots, x_n) = \sum_{i=1}^{n} \log f_\theta(x_i),$$

given the data $(x_1, \cdots, x_n)$. Under regularity conditions (usually continuous differentiability), we determine the *MLE* by solving the log-likelihood equations

$$\frac{\partial \log f_{\theta,n}(x_1, \cdots, x_n)}{\partial \theta_j} = 0, \quad j \in [1, r]. \quad (LLE).$$

we then show that solutions of Equations (LLE) are global maximum or local maximum. For $r = 1$, it will be enough to check that

$$\frac{\partial^2 \log f_{\theta,n}(x_1, \cdots, x_n)}{\partial \theta^2} < 0.$$

For $r > 1$, we use the partial derivatives of second order and check that their matrix generates a definite-negative quadratic form.

7. Selecting models

In this section we discuss the steps taken in estimation of the loss function or selecting a model for the data.

1.1. Cleaning the data.

An outlier data is a data which is not similar to most of the other data. We search the outliers by the Wilk's test, we compute the following

$$d_1(x) = me(x) - 1.5Q \ \ and \ \ d_2(x) = me(x) + 1.5Q$$

Where me is the median and Q is the interquatile range.

Empirical Wilk's rule. Any observation outside the interval $[d_1(x),\ d_2(x)]$ can be considered as an outlier. Usually, an outlier has some particular pattern not shared by the other data. For example, if we consider the age of people in a class (students and the professor). If the professor is old when compared to the students, his age is detected by the Wilk's test as an outlier. This is the first step in any statistical analysis.

We then remove the outliers and make our statistical analysis without them. Graphically, the box-plot tells us whether or not the data has outliers as the bounds of the box-plot are $d_1(x)$ and $d_2(x)$. The points represented before $d_1(x)$ and after $d_2(x)$ are signaled outliers. Finally, we look at the homogeneity of data by computing the Relative Coefficient of variations(rcv), if the rcv is small say less than 30% we say our data is homogenous and proceed.

2. Visualizing the interpolated empirical distribution functions and the Kernel density probability function.

2.1. Visualizing the interpolated empirical distribution functions (iecdf).

Let us sort the data $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
It is clear that, for each $n \geq 1$, the empirical distribution function satisfies

$$F_n(X_j) = \frac{j}{n}, \;\; j \in [1, n].$$

The *iecdf* is simply the linear interpolation of the joints $(X_{j,n}, j/n)$.

We may also use the Kernel estimator of the empirical distribution function *kedf* :

$$F_{n,h_n} = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)$$

where h_n is already given in 5

By using the Gaussian Kernel for example, we will get a smoother curve for the estimator.

2.2. Visualizing the Parzen estimator of the empirical *pdf* [epdf].

We consider the Parzen estimator introduced in the previous as follows

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{X_i - x}{h_n}\right) (DU)$$

We draw the functions described above: the *iecdf* and the Parzen estimator. If we are lucky, we may recognize the *cdf* or the *pdf* of some probability law.

2.3 the $Q-Q$ plot

We have for all $1 \leq j \leq n$.

$$F_n(X_j) = j/n$$

and thus by inverting

$$X_{j,n} = Q_{1,n}(j/n), \;\; 1 \leq j \leq n. \quad (QQ1)$$

where $Q_{1,n}$ is the qnantile function related to the current data $(x_1, \cdots, x_n)$. For testing the hypothesis (H) that the data are $iid \sim F_0$, we can generate a sample $(Z_1, ..., Z_n)$ form F_0 in R, if $Q_{2,n}$ is the quantile function based on $(z_1, \cdots, z_n)$, we also have

$$Z_{j,n} = Q_{2,n}(j/n), \ \ 1 \leq j \leq n. \quad (QQ2)$$

If the hypothesis (H) is true, both samples would come from F_0 and thus $Q_{1,n}$ and $Q_{2,n}$ would be almost equal, As a consequence, we would have $X_{j,n} \approx Z_{j,n}$, $j \in [1, n]$, meaning that :

[CT5] The points $(X_{j,n}, \ Z_{j,n})_{1 \leq j \leq n}$ should be approximately on the bi-sector line.

Following the negative reasoning of statisticians, we make the following rule :

For testing the hypothesis that the data $(x_1, \cdots, x_n)$ of size $n \geq 1$ are from a *cdf* F_0, we generate a sample $(z_1, \cdots, z_n)$ from F_0 of size n. We draw the scatter plot of the points $(x_{j,n}, \ z_{j,n})_{1 \leq j \leq n}$ and we reject the hypothesis if we do not obtain an almost strait line.

Warning. Generally, the convergence of the empirical quantile function Q_n to the exact quantile function Q is not fast in neighborhoods of 0 and 1. So the approximation $x_{j,n} \approx z_{j,n}$ is not so

good for j low (near 1) and for j large (near n), So by checking that we get a straight line, we should focus on the central part of the graphics. It is normal, even if the hypothesis is true, that the curve deviates from the bi-sector at the beginning and at the end.

Some times, we can fail to fit some data to available probability laws. Such kinds of situations, mixture distributions can be tried. Search of new probability laws is also a possibility as a matter of facts, the field of finding new interesting new laws is very active. For example, the class of non-symmetrical normal laws is proposed to model *pdf*'s like the Gaussian law but presenting non-symmetry trends.

8. A General statistical tests for fitting distribution

We use two general tools for testing our statical test.

1.0 Chi-square tests Let us suppose that the range of F_0 is $[a, b]$, with $a = lep(F_0)$ (low endpoint) and $b = uep(F_0)$ (upper endpoint) so that $a < b$. So the data $x_1, \cdots, x_n$ will be necessarily between a and b. So as a first rule, we cannot expect that F_0 can fit the data if there is a significant number of data outside the range $[a, b]$. This is an important reason for cleaning the data by putting aside the outliers.

We proceed as follows. First, we partition the range into m classes $a = c_0 < c_1 < \cdots < c_m = b$ so that the relative frequencies of classes

$$f_1 = \frac{1}{n}\#\{i, \in [1,n],\ \ x_i \in [c_0, c_1]\}\ \ and\ \ f_j = \frac{1}{n}\#\{i, \in [1,n],\ \ x_i \in]c_{j-1}, c_j]\},\ \ j \in [2,m].$$

is between 10% and 90%.

Now if the hypothesis that the data are *iid* observations from F_0, then the *iid* observations fall in the m classes one after another independently. So the vector $N = (N_1, \cdots, N_m)^t$ of the scores, where

$$N_1 = \#\{i \in [1,n],\ \ X_i \in [c_0, c_1]\}\ \ and\ \ N_j = \#\{i \in [1,n],\ \ X_i \in]c_{j-1}, c_j]\},\ \ j \in [2,m],$$

follows a multinomial law of dimension n and of vector parameter $p = (p_1, ..., p_m)$ where

$$p_j = F_0(c_j) - F_0(c_{j-1}),\ \ j \in [2,m].$$

For j=1, we would have $p_1 = F_0(c_1) - F_0(c_0 - 0)$ [$F_0(c_0 - 0)$ being the left limit of F_0 at c_0].

From a theorem on asymptotic law of multinomial distribution we have

$$\sum \frac{(OAF - ENO)^2}{ENO} \sim \chi^2_{nb-1},$$

where nb is the number of class, OAF (the N_j's) stands for *Observed Absolute Frequency* for classes, *EO* (the np_j's) for *Expected Number of Observations*. Here, we have

$$Q_n = \sum_{j=1}^{j=m} \frac{(N_j - np_j)^2}{np_j} \chi^2_{m-1}.$$

Rule 1. If F_0 is entirely known, Q_n is observable and since Q_n is bounded in probability, we use the p-value test where the p-value of the test is given by

$$p = \mathbb{P}(\chi^2_{m-1} > q_n).$$

where q_n is the observed value of Q_n.

Rule 2. In some cases, F_0 depends on r unknown parameters, $r < m$ and has a *cdf* F_0. then we use the MLE of parameters and plug-in them into F_0 to obtain $\hat{F}_0$. The computations of the probability become

$$\hat{p}_j = \hat{F}_0(c_j) - \hat{F}_0(c_{j-1}), \;\; j \in [1, m].$$

By plug-in the estimators in the model, the chi-square law loses r degrees of freedom and Q_n becomes $\hat{Q}_n$ with

$$\hat{Q}_n = \sum_{j=1}^{j=m} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j} \chi^2_{m-r-1}.$$

The p-value of the test is

$$p = \mathbb{P}(\chi^2_{m-r-1} > \hat{q}_n).$$

where where $\hat{q}_n$ is the observed value of $\hat{Q}_n$.

Remark. This test depends on the choice of the number of subdivisions of the range and the partition itself by the user. Two users backing of different partitions do not necessarily have the same p-value. As as long as the frequencies of classes are not too strong (greater that 90%) or too weak (less that 10%) for both partitions, the corresponding p-values should not be too far one from another. In general, this test is never used alone. We have to validate it by another test.

1.1 Kolmogorov-smirnov tests

By Glivenko-Cantelli theorem we have

$$n^{-1/2} D_n = \|F_n - F_0\|_{+\infty} = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \to 0 \ \ as \ \ n \to +\infty,$$

under our current working hypothesis that the data are generated from F_0. From that consistency result the following critical region is very intuitive

$$(RCX) \quad : \quad R_c = (n^{-1/2} D_n > c), \quad c > 0.$$

So we need the limiting distribution of D_n given by the Kolmogorov-Smirnov Theorem.

THEOREM 4.3. (Kolmogorov-Smirnov) Under the hypothesis (H), we have

$$\lim_{n \to +\infty} \mathbb{P}(\sqrt{n} \left\| F_n - F_0 \right\|_{+\infty} \leq x) = 1 - 2 \sum_{j=1}^{+\infty} (-1)^{j+1} \exp(-2j^2 x^2).$$

$$Let KS(x) = 1 - 2 \sum_{j=1}^{+\infty} (-1)^{j} \exp(-2j^2 x^2), \quad x > 0.$$

Instead of D_n which requires to take all the values of range of F_0, we simply take the values of the functions at the ordered observations $X_{i,n}$, $i \in [1.n]$ and this gives

$$D_n^* = \sqrt{n} \max_{1 \leq i \leq n} \left| \frac{i}{n} - F_0 \left(X_{i,n} \right) \right\|$$

and the same limiting distribution holds

$$\lim_{n \to +\infty} \mathbb{P}(D_n^* \leq x) = 1 - 2 \sum_{j=1}^{+\infty} (-1)^{j+1} \exp(-2j^2 x^2), \quad x > 0.$$

By consider the statistical test of critical region $(D_n^* > c)$, $c > 0$, the p-value of the test becomes

$$p = 1 - KS(d_n^*),$$

where d_n^* is the observed value of D_n^*.

Let us show how we an easily compute the function KS. Let fix a level of error $0 < \varepsilon < 1$. For $x > 0$, we have for $J \geq 2$

$$KS(x) = 1 - 2\sum_{j=1}^{J-1}(-1)^{j+1}\exp(-2j^2x^2) + R_J(x),$$

where

$$|R_J(x)| \leq 2\sum_{j\geq J}\exp(-2j^2x^2) \leq 2\sum_{j\geq J}\exp(-2jx^2) = \frac{2\exp(-2Jx^2)}{1-\exp(-2x^2)}$$

By allowing an error not greater than ε, that is

$$\frac{1\exp(-2Jx^2)}{1-\exp(-2x^2)} \leq \epsilon \Leftrightarrow J \geq \left(-\frac{\log\left((\epsilon/2)\left(1-e^{-2x^2}\right)\right)}{2x^2}\right) = J(x)$$

$KS(x)$ can be approximated as

$$KS(x) = 1 - 2\sum_{j\geq J(x)}\exp(-j^2x^2/2) \pm \varepsilon.$$

CHAPTER 5

A Case Study

We will illustrate estimation of distribution with an insurance data set distributed by the Data Sciences website https://www.kaggle.c It has 39 variables and 10211 rows. The name of the corresponding external file is *insurance_claims_2019.csv*, in particular to the following variables: injury_claim and vehicle_claim. We recall that the data concerns clients of an insurance company. The object of the insurance is the vehicle. The clients may contract insurance on the vehicle itself (to be repaired in case of accident) and/or the injury of the people in the vehicle. The software R is used to obtained our results.

Based on our knowledge on distribution that accident have the memory-loss property, we guess that these data follow a gamma distribution

(1) Modeling data for injury claim under the rear collision.

(A) We noticed that the data contains zero values. We interpret them as claims X under the deductible level d and the non-zero values $X-d$. This may explain the presence of null values. Insurance company are usually interested in large claims,

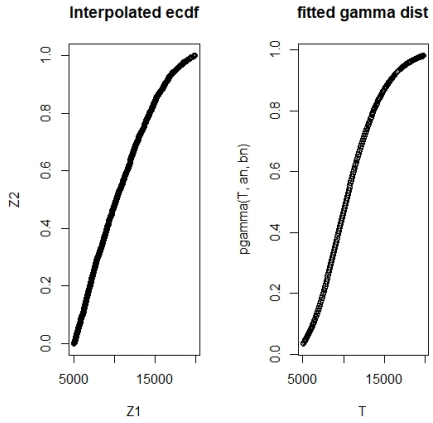
let us focus on the big claims, exceeding some threshold say TSH(5000). We then clean the data and obtained the relative coefficient of variation as 35.24%.

(B) We obtained obtained the parameter of the distribution using the MOE

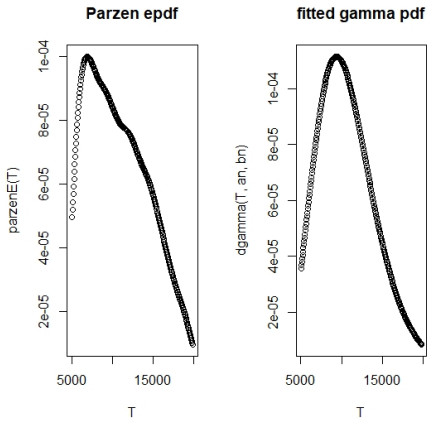
(C) We plot the emperical cdf and that of the fitted gamma distribution and obtained the image below.

(D) We plot the Parzen estimator of the emperical pdf and that of the fitted gamma distribution as shown below.

Compairing the ecdf and epdf with that of the fitted gamma distribution, we are tempted to accept the result. But plotting the q-q plot suggest we reject the result, we further test our result using the chi-square and the KS test. the p-values obtained are 0 and $1.535339e^{-7}$ respectively which also supported the result obtained from the Q-Q plot, hence we reject our hypothesis.



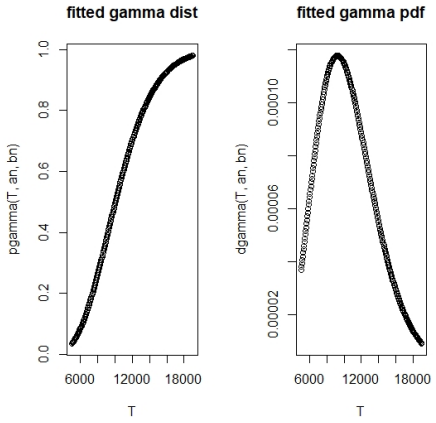
(a) Comparison of cdf's



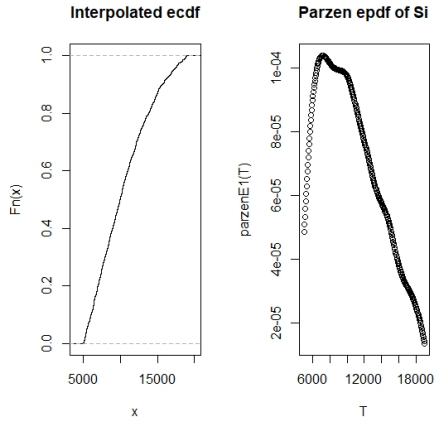
(b) comparison on pdf's

(2) Modeling data for injury claim under the side collision.

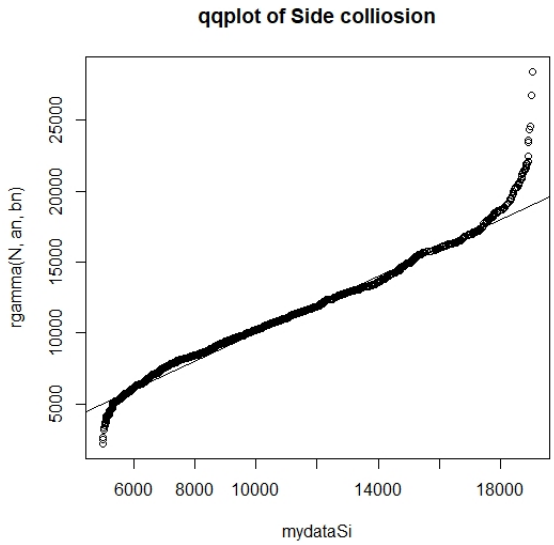
We also set a threshold of 5000 for same reason as in the rear collision type, we clean the data removing 96 outliers and obtained rcv as 34.15% which is reasonable then we proceed in plotting the epdf and ecdf. We obtained the parameter of the distribution using the MOE and compare the epdf and ecdf with that of gamma distribution. We further plot the q-q plot as shown below which is not so good, then further our test using the chisquare and ks test and obtained very low p values of $5.69392e^{-10}$ and 0.000198832 respectively which made us reject our hypothesis that the data follows a gamma distribution.



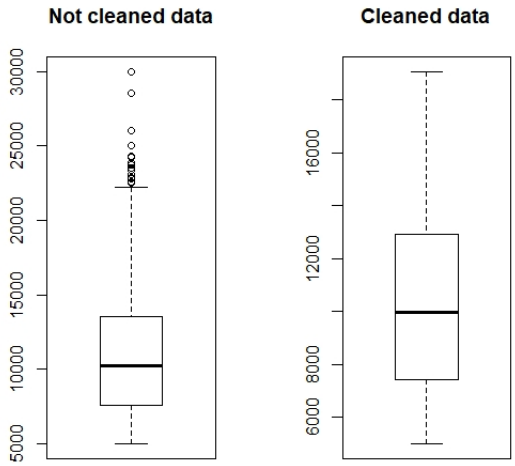
(a) fitted distribution



(b) ecdf and epdf of IC Side



(c) QQ-Plot



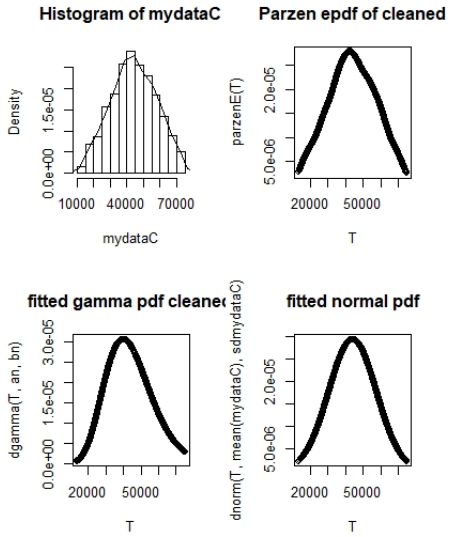
(d) cleaning the data

(3) Modeling data for vehicle claim rear collision

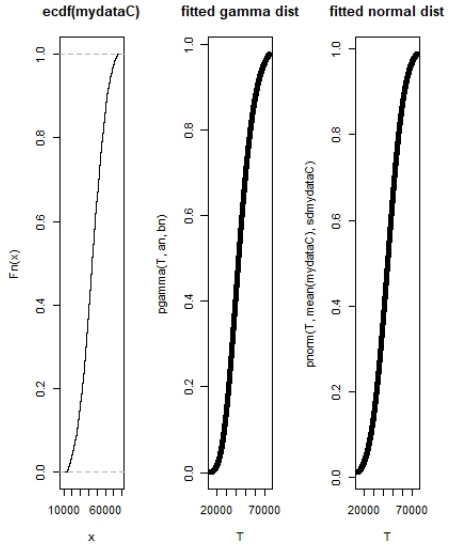
Here we didn't set a threshold, we cleaned the data by removing 181 outliers and obtained an rcv of 30.69% which is reasonable so we proceed. The epdf and ecdf looks like that of normal distribution, so we test for normality and gamma distributions.

(A) Normal Distribution: The q-qnorm and the chisquare (p=0.0006125113) suggest we reject the result but the ks test gives a p-value of 0.08405986 which suggest we accept the result, in this case we further do other test of normality. The Shapiro test for normality also gives a very low p value of $1.621e^{-}12$ which made us reject the hypothesis that the data follows a normal distribution.

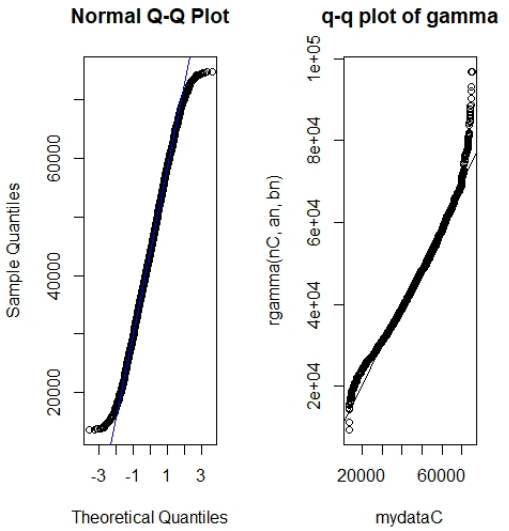
(B) Gamma Distribution: The q-qnorm, chisquare (p=0) and ks test (p=$8.041712e^{-6}$) all suggest we reject the result. Hence we conclude that the data is not following a gamma distribution.



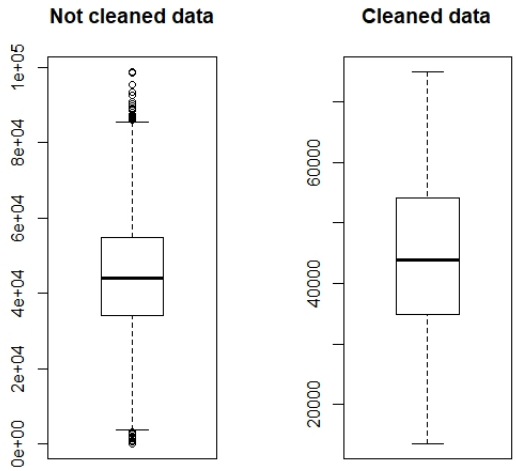
(a) epdf, fitted gamma and fitted normal distribution



(b) ecdf, normal and gamma cdf of vehicle claim for rear



(c) qqnorm and qqplot



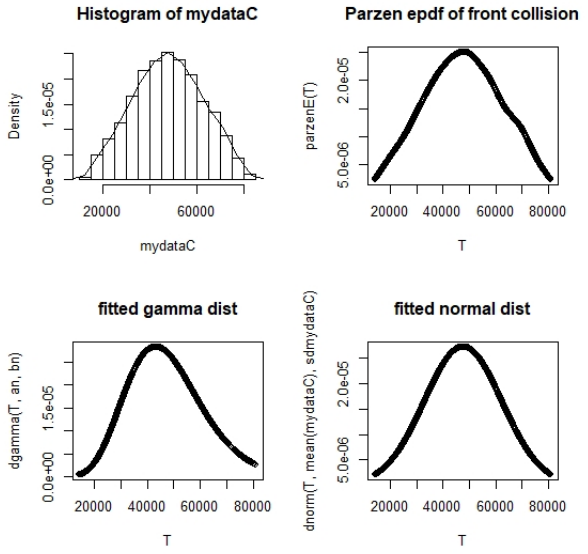
(d) cleaning the data

(4) Modeling data for vehicle claim front collision

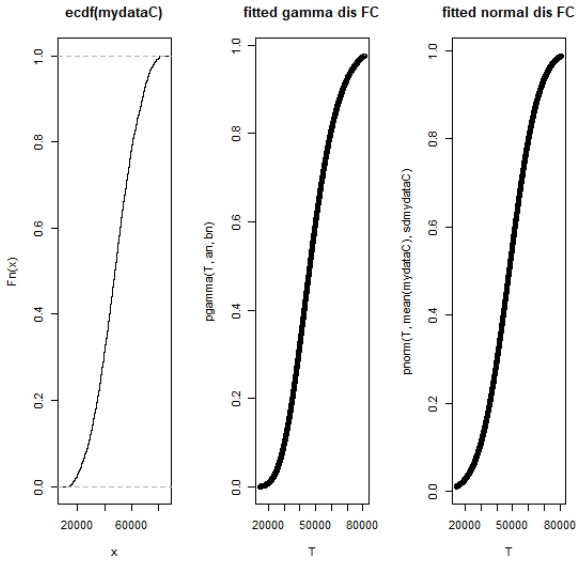
Here we didn't set a threshold also, we cleaned the data by removing 118 outliers and obtained an rcv of 30.74% which is also reasonable so we proceed. The epdf and ecdf looks like that of normal distribution than gamma distribution, qqnorm is better than the qqplot for gamma.

(A) Normal Distribution: The chisquare (p=0.0008535301) suggest we reject the result but the ks test gives a p-value of 0.2144677 which suggest we accept the result, in this case we further do other test of normality. The Shapiro test for normality also gives a very low p value of $5.282e^{-}11$ We suggest we do other test for normality.

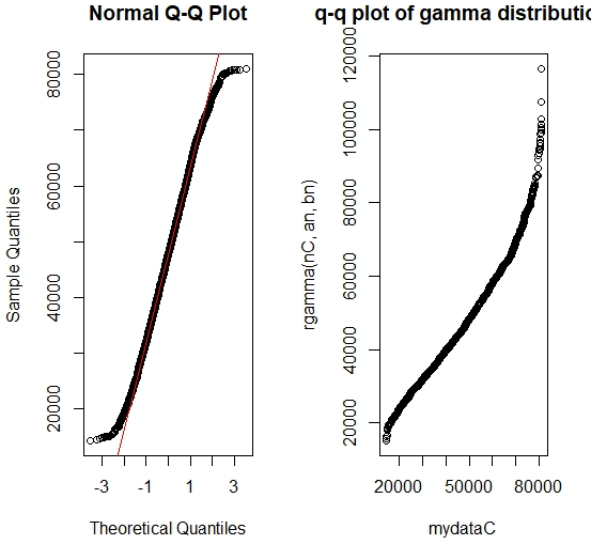
(B) Gamma Distribution: The chisquare (p=0) and ks test (p=0.00012730 all suggest we reject the result. Hence we conclude that the data is not following a gamma distribution.



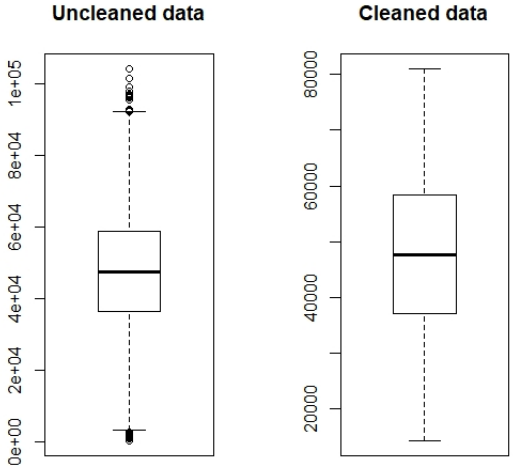
(a) epdf, fitted gamma and fitted normal distribution



(b) ecdf, normal and gamma cdf of vehicle claim for front



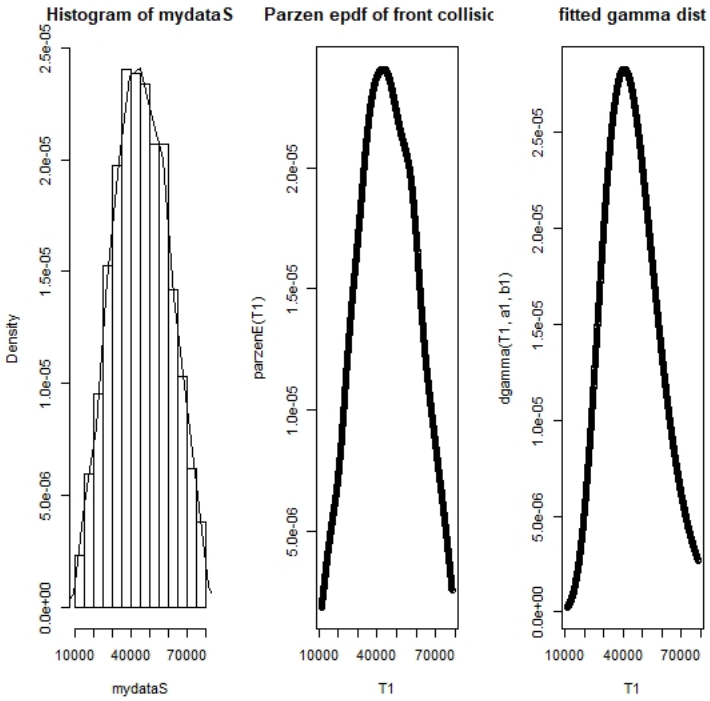
(c) qqnorm and qqplot



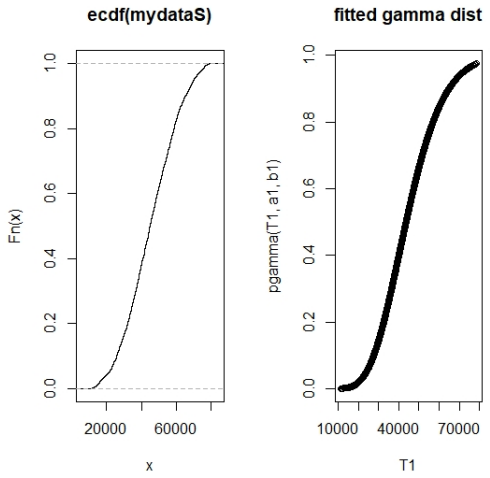
(d) cleaning the data

(5) Modeling data for vehicle claim under the side collision.

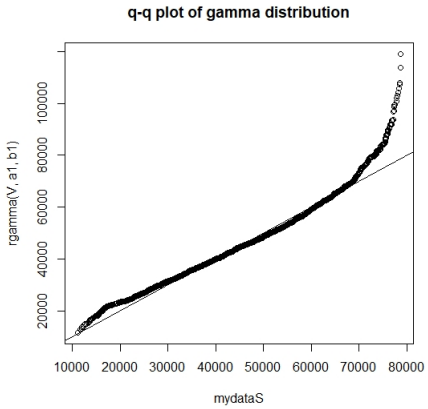
We clean the data removing 114 outliers and obtained rcv as 32.59% which is reasonable then we proceed in plotting the epdf and ecdf. We obtained the parameter of the distribution using the MOE and compare the epdf and ecdf with that of gamma distribution. We further plot the q-q plot as shown below which suggest we reject the result, then further our test using the chisquare and ks test and obtained very low p values of 0 and $1.14114e^{-5}$ respectively which made us reject our hypothesis that the data follows a gamma distribution.



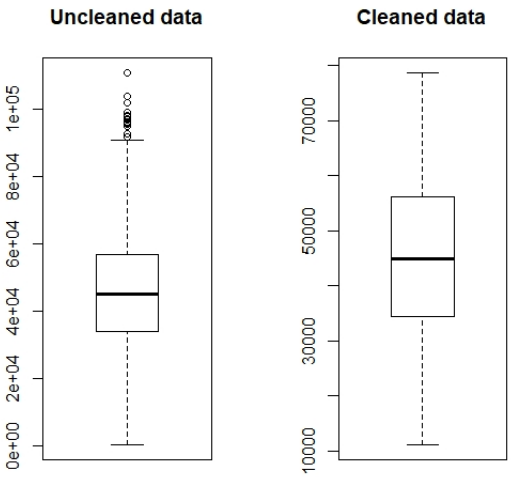
(a) histogram, epdf, fitted pdf of VC Side collision



(b) ecdf fitted cdf of VC Side collision



(c) QQ-Plot



(d) cleaning the data

CHAPTER 6

Poisson Stochastic Processes

In this chapter, we are going to study the Poisson stochastic from three points of view. This stochastic process is mainly used to described occurrence times of some random events over a continuous time. Here are some general examples :

(a) Arrival times to a desk in some bank.

(b) Occurrence times of failure of machines in a company.

(c) Arrival times of tasks to the central unity of a computer.

(d) etc.

Let us introduce the stochastic process from several points of view.

1. Description by exponential inter-arrival

Let us consider a bank desk which opens at 08H00 for example. Assuming that at that time there is no clients and we denote $Z = 0$ at that initial time taken as $t = 0$. In this model we

want simple, clients arrive one after another at random times [Later, models in which several clients might come at the same times will be studied],

$$0 < Z_1 < Z_2 < .... < Z_n < ...$$

so that the inter-arrival times

$$X_1 = Z_1 - Z_0, X_2 = Z_2 - Z_1, X_3 = Z_3 - Z_2, ..$$

are independent and follow an exponential of parameters $\lambda > 0$, denoted

$$X_1, X_2, ...iid \ \sim \ \mathcal{E}(\lambda). \tag{1.1}$$

Let us begin by giving the finite-distributions of the sequences $(X_n)_{n\geq 1}$ and $(Z_n)_{n\geq 1}$. This will allow to derive the characteristics of this important stochastic process.

1.1. Probability law of arrival times and that of the inter-arrival times.

We have for all $n \geq 1$

$$f_{(X_1,...,X_n)}(x_1, ..., x_n) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i) \ \ 1_{(x_1\geq 0,...,x_n\geq 0)}. \tag{1.2}$$

From there, we can derive the *pdf* of $Z^{(n)} = {}^t(Z_1, ..., Z_n)$. Indeed we have :

Proposition 6.1. *The finite-distributions of arrival times $Z_1 < Z_2 < .... < Z_n < ...$ are given as follows. For all $n \geq 1$,*

$$f_{(Z_1,...,Z_n)}(z_1, ..., z_n) = \lambda^n \exp(-\lambda z_n) 1_{\Gamma_n}(z) \tag{1.3}$$

with

$$\Gamma_n = \{z = (z_1, ..., z_n) \in \mathbb{R}^n; 0 < z_1 < z_2 < ... < z_n\}.$$

Each $i \geq 1, Z_i$ follows a Gamma of parameters (i, λ) : $Z_i \sim \gamma(i, \lambda)$.

Proof. Let $n \geq 1$, each of the two $Z^{(n)} = Z^t(Z_1, ..., Z_n)$ and $X^{(n)} =^t (X_1, ..., X_n)$ are linear transformations of the others:

$$\begin{cases} Z_1 = & X_1 \\ Z_2 = & X_1 + X_2 \\ ... & ... \\ Z_n = & X_1 + X_2 + ... + X_n \end{cases} \iff \begin{cases} X_1 = & Z_1 - Z_2 \\ X_2 = & Z_2 - Z_1 \\ ... & ... \\ X_n = & Z_n - Z_{n-1} \end{cases}.$$

Let us denote the matrices in the linear transformations above by A and B and we write

$$Z^{(n)} = AX^{(n)} \Leftrightarrow X^{(n)} = BZ^{(n)}.$$

The support of $(X_1, ..., X_n)$ is $D_n = \mathbb{R}_+^n$ and that of $(Z_1, ..., Z_n)$ is

$$\Gamma_n = \{z = (z_1, ..., z_n) \in \mathbb{R}^n; 0 < z_1 < z_2 < ... < z_n\}.$$

The mapping $X^{(n)} = BZ^{(n)}$ is a diffeomorphism between D_n and Γ_n and the jacobian determinant satisfies $|\det(B)| = 1$, The change of variables (see Lo (2018), Chapter 3) entails

$$f_{(Z_1,...,Z_n)}(z_1, ..., z_n) = f_{(X_1,...,X_n)}(x_1, ..., x_n) \ 1_{\Gamma_n}(z_1, ..., z_n) \tag{1.4}$$

with

$$(x_1, ..., x_n) = B(z_1, ..., z_n) = (z_1 - z_0, z_2 - z_1, ..., z_n - z_{n-1}). \tag{1.5}$$

By plugging (1.5) in (1.2), (1.4) becomes

$$f_{(Z_1,...,Z_n)}(z_1, ..., z_n) = \lambda^n \exp(-\lambda z_n) \ 1_{\Gamma_n}(z_1, ..., z_n).$$

The first approach is finished. □

Let us go the second approach.

2. Counting function

Let use the independent and λ-exponentially distributed inter-arrival times $X_j,\ j \geq 1$ corresponding to the arrival times $Z_0 = 0 < Z_2 < .... < Z_n < ....$ We define the counting stochastic processes $N(t)$ as the number of arrivals up to time $t \geq 1$:

$$\forall t \geq 0,\ \ N_t = N(]0,t]) = \#\{i \geq 1, Z_i \leq t\}.$$

We easily can express events of the form $(N_t = k)$ in function the arrival times Z_i. Indeed, we have

$$\forall t \geq 0,\ \ (N_t = k) = (Z_k \leq t < Z_{k+1})$$

We also can write

$$N_t = \sum_{k \geq 1} 1_{]0,t]}(Z_i).$$

The counting function itself is called the Poisson process of intensity $\lambda > 0$.

We are going to give the characteristic properties of the counting function. To do short, we propose the following formula [which will be used in the computations later] as an exercise (see solution in page **??**)

Exercise. For all $k \geq 1$, show that

$$\int_{(0<z_1<z_2<...<z_k\leq t)} dz_1...dz_k = \frac{t^k}{k!}.$$

Here are interesting properties of the counting process.

2.1. One-dimensional margins of the counting process.

For $t=0$, we have $N_0=0$. For $t>0$, we have for $k\geq 0$,

$$\begin{aligned}\mathbb{P}(N_t &= k) = \mathbb{P}(Z_k \leq t < Z_{k+1}) \\ &= \mathbb{P}(Z_1 \leq Z_2 \leq ... \leq Z_k \leq t < Z_{k+1}).\end{aligned}$$

From there, we use the *pdf* of $(Z_1, Z_2, ..., Z_k, Z_{k+1})^t$ given in Formula (1.3) and the Fubini's formula to get

$$\begin{aligned}\mathbb{P}(N_t &= k) = \lambda^{k+1} \int_{z_1\leq z_2\leq ...\leq z_k\leq t<z_{k+1}} \exp(-\lambda z_{k+1}) dz_1....dz_k dz_{k+1} \\ &= \lambda^{k+1} \int_{z_1\leq z_2\leq ...\leq z_k\leq t} dz_1....dz_k \int_{\leq t<z_{k+1}} \exp(-\lambda z_{k+1}) dz_{k+1} \\ &= \lambda^{k+1} \int_{z_1\leq z_2\leq ...\leq z_k\leq t} dz_1....dz_k \left[-\frac{\exp(-\lambda z_{k+1})}{\lambda}\right]_{z_{k+1}=t}^{z_{k+1}=+\infty} \\ &= \lambda^k \exp(-\lambda t) \times \int_{z_1\leq z_2\leq ...\leq z_k\leq t} dz_1....dz_k.\end{aligned}$$

By using the exercise above, we arrive at

$$\mathbb{P}(N_t = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!},$$

which proves that $N(t) \sim \mathcal{P}(\lambda t)$. $\square$

This means that each margin N_t follows a Poisson Law of parameter λt : $N_t \sim \mathcal{P}(\lambda t)$.

The name *Poisson* given to that stochastic process certainly derives from that fact.

2.2. finite-distribution laws.

Proposition 6.2. *For all $k \geq 1$, for all $t_0 = 0 < t_1 < ... < t_k$, for all $(n_1, ..., n_k) \in \mathbb{N}^k$, we have (with the convention that $n_0 = 0$),*

(2.1)

$$\mathbb{P}(N_{t_1} = n_1, ..., N_{t_k} = n_k) = \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp(-\lambda(t_j - t_{j-1})) 1_{(0 \leq n_1 \leq n_2 \leq ... \leq n_k)}.$$

or, in other words,

(2.2)

$$\mathbb{P}(N_{t_1} = n_1, ..., N_{t_k} = n_k) = \exp(-\lambda t_k) \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} 1_{(0 \leq n_1 \leq n_2 \leq ... \leq n_k)}.$$

Before we give the proof, we should remark this important fact about the memory loss property.

Remark (R) Based on the independence of the *iidness* inter-arrival times, we surely have that given $N_t = k$ at time t, the $N^*(u) =$

$N(t+u)-k$, $u \geq 0$ in again a counting process based on the interval times $Z_0^* = 0$, $Z_i^* = Z_{i+k}^* - k$, $i \geq 1$, with inter-arrival-times $Z_{i+1}^* - Z_i^* = Z_{k+i+1}^* - Z_{k+i}^*$ $iid \sim \mathcal{P}(\lambda)$, independent to events based on $(Z_1, ..., Z_k)$. So, we have

$$N_u \stackrel{d}{=} N_{t+u} - N_t.$$

As well, $N_t = k$, $N_{t+u} - N_t$ does not depend on k. We can generalize that for any $k \geq 1$, for any $t_0 = 0 < t_1 < ... < t_k$, we have

$$(N_{t_1} = n_1, ..., N_{t_k} = n_k) = (N_{t_1} - N_{t_0} = n_1 - n_0, N_{t_2} - N_{t_1} = n_2 - n_1 ..., N_{t_k} - N_{t_{k-1}} = n_{k-1} - n_k),$$

From there, we may get the following facts :

(a) Each $N_{t_j} - N_{t_{j-1}}$, $1 \leq j \leq k$, has the same law as $N_{t_j - t_{j-1}}$, that is $\mathcal{P}(\lambda(t_j - t_{j-1}))$.

(b) The $N_{t_j} - N_{t_{j-1}}$ are independent.

Since N_t non-decreasing in $t \geq 1$, the event $(N_{t_1} = n_1, ..., N_{t_k} = n_k)$ is possible if only $n_0 = 0 \leq n_1 \leq n_2 \leq ... \leq n_k$, So we get

$$\mathbb{P}(N_{t_1} = n_1, ..., N_{t_k} = n_k) = \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp(-\lambda(t_j - t_{j-1})) 1_{(0 \leq n_1 \leq n_2 \leq ... \leq n_k)},$$

which entails

$$\mathbb{P}(N_{t_1} = n_1, ..., N_{t_k} = n_k) = \exp(-\lambda t_k) \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} 1_{(0 \leq n_1 \leq n_2 \leq ... \leq n_k)}.$$

From these facts, let us characterize the counting stochastic process.

2.3. Characterization of the counting process.

We already proved this.

Proposition 6.3. *The counting stochastic process satisfies the following properties.*

(IC)[Initial condition] $N_0 = 0$ *a.s.*

(PM) [Poisson margins] The margins N_t, $t \geq 1$, *follow Poisson laws* $\mathcal{P}(\lambda t)$.

(SI) [Stationarity of increments] N_t *has strong stationary increments, that is, for any* $k \geq 2$, *for any* $t_0 = 0 < t_1 < ... < t_k$,

$$(N_{t_j} - N_{t_{j-1}},\ 1 \leq j \leq k-1)^t =_d (N_{t_j - t_{j-1}},\ 1 \leq j \leq k-1)^t.$$

(I2) [Independent increments] N_t *has independent increments : for any* $k \geq 2$*, for any* $t_0 = 0 < t_1 < ... < t_k$*, the random variables* $N_{t_1} - N_{t_0}$*,* $N_{t_2} - N_{t_1}$*,* $\cdots$*,* $N_{t_k} - N_{t_{k-1}}$ *are independent.*

As explained in **?***, Chapter 2 [Definitions of types of stationary], with condition (I2) and the fact that* $N(0) = 0$*, Property (SI) is equivalent to*

$$\forall t \geq 0,\ u \geq 1,\ \ N(t+u) - N(t) =_d N(u).$$

(RC) [Paths right-continuity] The paths $t \hookrightarrow N_t(\omega)$ *are right continuous.*

Le point (RC) comes of the fact that for any $k \geq 0$ and for any $t \geq 0$,

$$(N_t = k) = (Z_k \leq t < Z_{k+1}).$$

entails that for any $\omega \in \Omega$ and for $s_0 > 0$ so small that $t + s_0 < Z_{k+1}(\omega)$, we still have for any $0 \leq s < s_0$

$$(Z_k(\omega) \leq t < t + s < Z_{k+1}(\omega)),$$

and $N_{\omega,t+s} = N(\omega, t)$ for $0 \leq s < s_0$. $\square$

In terms of Probability laws, we will have the Properties (IC), (PM), (SI) and (I2) do exactly characterize the probability law of $N(t)$, A stochastic process $N(t)$ with time $t \geq 0$ and sate space $\mathbb{N}$ will be called a Poisson process of intensity $\lambda > 0$ if Conditions (IC), (PM), (SI) and (I2) hold. If, on top of them,

(RC) holds *a.s.*, we say that a **standard** Poisson process of intensity $\lambda > 0$.

Let us use the Kolmogorov Existence Theorem ***(KET)*** to give birth the Poisson process from another point of view. But before we proceed to that, let us make the following remarks.

(A) The finite-distributions in Formula (2.1) hold if Conditions (PM), (SI) and (I2) hold.

(B) If the finite-distribution of a stochastic process process (N_t) are given by Formula (2.1), we easily prove the assertions (IM), (SI), $(I2)$ by the very simple definition of the independence (seeLo (2018), Chapter 2).

Let use the Kolmogorov approach.

3. Approach of the Kolmogorov Existence Theorem

From a probabilistic point of view, the best approach begins with the Kolmogorov's theorem with the characterization of finite distributions. However, this approach does not always guarantee properties we want to have for the trajectories. The approach of Kolmogorov only guarantees the probability laws. For

example, the right-continuity of the paths of $(N_t)_{t\geq 0}$ must be established beyond the Kolmogorov's theorem to build a right-continuous version. We already did this for the Brownian movement for which we created a continuous version.

3.1. Construction of the Counting of the Poisson Process by the *KET*.

Theorem 6.4. Consider the family of discrete finite-distributions determined by their discrete *pdf*'s : for $k \geq 1$, for $t_0 = 0 < t_1 < ... < t_k$, par

(3.1)
$$f_{(t_1,...,t_k)}(n_1, ..., n_k) = \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp(-\lambda(t_j - t_{j-1}))1_{(0\leq n_1\leq n_2\leq ...\leq n_k)}.$$

Then this family is consistent and there exist a stochastic process

$$(\Omega, A, \mathbb{P}, (N_t)_{t\geq 0}, \mathbb{R}_+, \mathbb{N})$$

of finite-distributions defined by Formula (3.1) and that Properties (IC), (PM), (SI) and (I2) holds.

Proof. Let us establish the consistency condition (CHSD2) (see page **??**, Chapter **??**). We have to prove that for any $k \geq 1$, for any $t_0 = 0 < t_1 < ... < t_k < t_{k+1}$

$$\int_{\mathbb{N}} f_{(t_1,...,t_k,t_{k+1})}(n_1, ..., n_k, n_{k+1}) d\nu(n_{k+1}) = f_{(t_1,...,t_k,t_{k+1})}(n_1, ..., n_k),$$

that is

$$\sum_{n_{k+1}\geq 0} f_{(t_1,...,t_k,t_{k+1})}(n_1,...,n_k,n_{k+1}) = f_{(t_1,...,t_k,t_{k+1})}(n_1,...,n_k).$$

Let us treat the nontrivial case: $0 \leq n_1 \leq n_2 \leq ... \leq n_k$, otherwise the equation is simply $0 = 0$. Furthermore, the terms $f_{(t_1,...,t_k,t_{k+1})}(n_1,...,n_k,n_{k+1})$ are null for $n_{k+1} < n_k$. So, $\sum_{n_{k+1}\geq 0} f_{(t_1,...,t_k,t_{k+1})}(n_1,...,n_k,n_{k+1})$ is equal to :

$$\begin{aligned}
&= \sum_{n_{k+1}\geq n_k} f_{(t_1,...,t_k,t_{k+1})}(n_1,...,n_k,n_{k+1}) \\
&= \sum_{n_{k+1}\geq n_k} \prod_{j=1}^{k+1} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp(-\lambda(t_j - t_{j-1})) \\
&= \prod_{j=1}^{k} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp(-\lambda(t_j - t_{j-1})) \\
&\times \sum_{n_{k+1}\geq n_k} \frac{(\lambda(t_{k+1} - t_k))^{n_{k+1} - n_k}}{(n_j - n_{j-1})!} \exp(-\lambda(t_{k+1} - t_k)).
\end{aligned}$$

By the change of variables $s = n_{k+1} - n_k$, we have

$$\begin{aligned}
&\sum_{n_{k+1}\geq n_k} \frac{(\lambda(t_{k+1} - t_k))^{n_{k+1} - n_k}}{(n_j - n_{j-1})!} \exp(-\lambda(t_{k+1} - t_k)) \\
&= \exp(-\lambda(t_{k+1} - t_k)) \sum_{s\geq 0} \frac{(\lambda(t_{k+1} - t_k))^s}{s!} \\
&+ \exp(-\lambda(t_{k+1} - t_k)) \times \exp(\lambda(t_{k+1} - t_k)) \\
&= 1.
\end{aligned}$$

So, we have the desired result. To finish, we apply the *KET* and Remark (R) above. □

By Remark (R), (page 131), Conditions (IC), (PM), (SI) and (I2) together determine the finite-distributions of $N(t)$ as in Formula (3.1) which in turns allows a unique extension to a Poisson Counting process N^*, which is equal to N in law. So, we have

THEOREM 6.5. The counting stochastic process of a Poisson process is entirely characterized by Properties (IC), (PM), (SI) and (I2).

3.2. Right-Continuous version.

Now, we have created a Poisson counting stochastic process from the finite-distributions in (3.1), can we transform it into an *a.s* right-continuous Poisson counting stochastic process? It is possible according to **?**.

4. More properties for the Standard Poisson Process

It is time to get more deeper properties of the standard Poisson process (from the three points view of *iid* exponential inter-arrival times, its characterization by properties (IC), (PM), (SI) and (I2) and its its characterization by the finite-distributions in (3.1)).

4.1. Conditional Laws.

PROPOSITION 6.6. *Conditionally on the event* $(N_t = n)$*, the vector* $(Z_1, ..., Z_n)$ *of arrival time has the following pdf*

$$f_{(Z_1,...,Z_n)}(z|N_t = n) = \frac{n!}{t^n} 1_{(0<z_1<z_2<...<z_n\leq t)}. \tag{4.1}$$

Proof. For all Borel subset of $\mathbb{R}^n$, we have

$$\begin{aligned}\mathbb{P}((Z_1, ..., Z_n) \in A|N_t = n) &= \frac{\mathbb{P}(Z_1, ..., Z_n) \in A, N_t = n)}{P(N_t = n)} \\ &= \frac{\mathbb{P}(Z_1, ..., Z_n) \in A, Z_n \leq t < Z_{n+1})}{P(N_t = n)}.\end{aligned}$$

Let us use the unconditional *pdf* of $(Z_1, ..., Z_{n+1})$ and The Funibi's theorem to get

$$\begin{aligned}\mathbb{P}((Z_1, ..., Z_n) \in A|N_t = n) &= \frac{\lambda^{n+1}}{\mathbb{P}(N_t = n)} \\ &\quad \times \int_{(z_1,...,z_n)\in A, 0\leq z_1\leq...\leq z_n\leq t<z_{n+1}} \exp(-\lambda z_{n+1}) dz_1....dz_n dz_{n+1} \\ &= \frac{\lambda^{n+1}}{\mathbb{P}(N_t = n)} \int_{(z_1,...,z_n)\in A, 0\leq z_1\leq...\leq z_n\leq t} dz_1....dz_n \left(\int_{t<z_{n+1}} \exp(-\lambda z_{n+1}) dz_{n+1} \right) \\ &= \frac{\lambda^{n+1}}{\mathbb{P}(N_t = n)} \int_{(z_1,...,z_n)\in A, 0\leq z_1\leq...\leq z_n\leq t} dz_1....dz_n \left(\frac{\exp(-\lambda t)}{\lambda} \right) \\ &= \frac{\lambda^{n+1}}{(\lambda t)^n \exp(-\lambda t)/n!} \int_{(z_1,...,z_n)\in A, 0\leq z_1\leq...\leq z_n\leq t} dz_1....dz_n \left(\frac{\exp(-\lambda t)}{\lambda} \right) \\ &= \int_A \frac{n!}{t^n} 1_{\Gamma_n}(z_1, ..., z_n) \ dz_1....dz_n.\end{aligned}$$

By definition of the *pdf* with respect to the Lebesgue measure, we have for all $n \geq 1$,

$$d\mathbb{P}_{(Z_1,...,Z_n)}(\circ|N_t = \frac{n!}{t^n}1_{\Gamma_n}.$$

4.2. Law of arrival times when the order of arrival is lost. A very interesting property of Poisson process is the following. Let us suppose that clients arrive to some bank desk and each time one clients arrives, his name is put on a sheet and his arrival time in an ordered list and the arrival time on a card. Once n clients has already arrived up to time t, they sat in a waiting room without any order. The only way to have the different arrival times is to use the ordered list. Let us suppose that the list which determines the the correspondence between the cards and their owners is lost. The bank officer decides then to take a random order of the clients and to shuffle the cards and to give a card to each client in the current order that was set by the bank officer. In that situation, we will proof that the arrival times are independent and are uniformly distributed on $[0,\ t]$.

Let S_n be the set of permutations $\{1, 2, ..., n\}$ endowed with the σ-algebra $\mathcal{P}(\{1, 2, ..., n\}) = \mathcal{P}_n$ and the discrete $\mathbb{U}$ on $(S_n, \mathcal{P}_n)$

$$\mathbb{U}(\{\sigma\}) = \frac{1}{n!},\ \ \sigma \in S_n.$$

Since the ordered arrival times is lost, a random ordering is denoted by

$$(Z_{\sigma(1)}, ..., Z_{\sigma(n)})(\omega) = \sigma(Z_1, ..., Z_n)(\omega)$$

with two elementary events : $\omega \in \Omega$ and $\sigma \in S_n$. The study takes place in the probability space

$$(S_n \times \Omega, \ \mathcal{P}_n \otimes \mathcal{A}, \ \mathbb{L}),$$

where

$$\mathbb{L} = \mathbb{U} \otimes \mathbb{P}(\cdot | N_t = n).$$

We have :

PROPOSITION 6.7. *Let us suppose that conditionally on* $(N_t = n)$ *and that we have lost the order of arrival times. Then the vector* $(Z_1, ..., Z_n)$ *has has margins independent and uniformly distributed on* $[0, t]$*, that is, pour tout* $B \in \mathcal{B}(\mathbb{R}^n)$*,*

$$\mathbb{L}(\sigma(Z_1, ..., Z_n) \in B) = \frac{1}{t^n} \int_B \prod_{i=1}^{n} 1_{[0,t]}(u_i) du_1 ... du_n.$$

Proof. Before we go further, we recall that the hyperplans of $\mathbb{R}^k$ are null sets with respect to the Lebesgue measure. Let us denote by λ_n by the Lebesgue measure on $\mathbb{R}^n$. Then we have

$$\lambda_n(\{(z_1, ..., z_n), \ \exists (1 \leq i \neq j \leq n), \ z_i = z_j\}) = 0. \tag{4.2}$$

Now we have

$$\begin{aligned}\mathbb{L}(\sigma(Z_1,...,Z_n)\in B) &= \sum_{\sigma_0\in S_n}\mathbb{L}(\sigma(Z_1,...,Z_n)\in B,\sigma=\sigma_0)\\ &= \sum_{\sigma_0\in S_n}\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma=\sigma_0)\times\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma(Z_1,...,Z_n)\in B|(\sigma=\sigma_0))\\ &= \sum_{\sigma_0\in S_n}\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma=\sigma_0)\times\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma_0(Z_1,...,Z_n)\in B|(\sigma=\sigma_0)).\end{aligned}$$

But $\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma=\sigma_0)$ does not depend on σ. Hence we use the marginal probability on $\mathbb{U}$ and we have

$$\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma=\sigma_0)=\frac{1}{n!}.$$

Also, we have

$$\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma_0(Z_1,...,Z_n)\in B|(\sigma=\sigma_0))$$

depends on ω (since σ fixed as σ_0), we utilize that the the marginal probability $\mathbb{P}(\cdot|N_t=n)$ and we have

$$\mathbb{U}\otimes\mathbb{P}(\cdot|N_t=n)(\sigma_0(Z_1,...,Z_n)\in B)=\mathbb{P}(\cdot|N_t=n)((Z_1,...,Z_n)\in\sigma_0^{-1}(B)).$$

Moreover, we have

$$\begin{aligned}\mathbb{L}(\sigma(Z_1,...,Z_n) \in B) &= \frac{1}{n!}\sum_{\sigma_0\in S_n}\mathbb{P}(\cdot|N_t=n)((Z_1,...,Z_n)\in\sigma_0^{-1}(B)) \\ &= \frac{1}{n!}\sum_{\sigma_0\in S_n}\mathbb{P}(\cdot|N_t=n)((Z_1,...,Z_n)\in\sigma_0^{-1}(B)),\end{aligned}$$

since σ_0^{-1} runs over S_n if σ_0 does. By applying (4.1), we have

$$\begin{aligned}\mathbb{L}(\sigma(Z_1,...,Z_n) \in B) &= \frac{1}{n!}\int\sum_{\sigma_0\in S_n}\mathbb{P}(\cdot|N_t=n)((Z_1,...,Z_n)\in\sigma_0(B)) \\ &= \frac{1}{t^n}\sum_{\sigma_0\in S_n}\int_{\sigma_0(B)}1_{(0<z_1<z_2<...<z_n\le t)}dz_1....dz_n.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{L}(\sigma(Z_1,...,Z_n) \in B) &= \frac{1}{t^n}\sum_{\sigma_0\in S_n}\int 1_{\sigma_0(B)}(z_1...,z_n)1_{(0<z_1<z_2<...<z_n\le t)}dz_1....dz_n \\ &= \frac{1}{t^n}\sum_{\sigma_0\in S_n}\int 1_B(\sigma_0^{-1}(z_1...,z_n))1_{(0<z_1<z_2<...<z_n\le t)}dz_1....dz_n.\end{aligned}$$

In each integral, we make the change of variable $\sigma_0^{-1}(z_1...,z_n) = (u_1,...u_n)$. σ_0 is diffeomorphism on $\mathbb{R}^n$, is linear and the Jacobian of the diffeomorphism the $|J|=1$. Thus,

$$\begin{aligned}\mathbb{L}(\sigma(Z_1,...,Z_n) \in B) &= \frac{1}{t^n}\sum_{\sigma_0\in S_n}\int_B 1_{(0<u_{\sigma_0(1)}<u_{\sigma_0(2)}<...<u_{\sigma_0(n)}\le t)}du_1....du_n \\ &= \frac{1}{t^n}\int_B\left\{\sum_{\sigma_0\in S_n}1_{(0<u_{\sigma_0(1)}<u_{\sigma_0(2)}<...<u_{\sigma_0(n)}\le t)}\right\}du_1....du_n.\end{aligned}$$

But $]0,t]^n$ in the sum of

$$C_t = \{(u,...,u_n) \in]0,t]^n, \forall (1 \leq i \neq j \leq n), \ \ u_i \neq u_j\}$$

and of sets included in hyperplans. We have

$$\begin{aligned} C_t &= \sum_{\sigma_0 \in S_n} \{(u_1,...,u_n) \in [0,t]^n, \ \ u_{\sigma_0(1)} < u_{\sigma_0(2)} < ... < u_{\sigma_0(n)}\} \\ &= \sum_{\sigma_0 \in S_n} \{(u_1,...,u_n) \in \mathbb{R}^n, \ \ 0 \leq u_{\sigma_0(1)} < u_{\sigma_0(2)} < ... < u_{\sigma_0(n)} \leq t\} \end{aligned}$$

Hence,

$$1_{C_t} = \left\{ \sum_{\sigma_0 \in S_n} 1_{(0<u_{\sigma_0(1)}<u_{\sigma_0(2)}<...<u_{\sigma_0(n)}\leq t)} \right\}$$

and thus

$$\mathbb{L}(\sigma(Z_1,...,Z_n) \in B) = \frac{1}{t^n} \int_B 1_{C_t}(u,...,u_n) \ \ du_1....du_n.$$

Finally, by (4.2), the set

$$[0,t]^n \backslash C_t = \{(z_1,...,z_n) \in [0,t]^n, \quad \exists (1 \leq i \neq j \leq n), \ \ z_i = z_j\}$$

is null and that leads to

$$\begin{aligned} \mathbb{L}(\sigma(Z_1,...,Z_n) \ \ &\in \ \ B) = \frac{1}{t^n} \int_B 1_{[0,t]^n}(u,...,u_n) \ \ du_1....du_n \\ &= \int_B \left\{ \prod_{i=1}^n \left(\frac{1}{t} 1_{[0,t]}(u_i) \right) \right\} du_1...du_n. \end{aligned}$$

Here, we identify the *pdf* of a vector of dimension n with *iid* components uniformly distributed on $[0,t]$. □

4.3. Superposition of Poisson Processes.

Proposition 6.8. *Let us assume that we have two independent Poisson processes with respective intensities λ_1 and λ_2 and with respective counting functions $N_t^{(1)}$ and $N_t^{(2)}$, taking place at the same time. By superposing the arrival times of the two processes, nous obtenons a Poisson process of intensity $\lambda = \lambda_1 + \lambda_2$.*

Proof. By superposing the arrival times of the two processes, the total number of arrival times N_t is on the form of $N_t^{(1)}$ and $N_t^{(2)}$:

$$N_t = N_t^{(1)} + N_t^{(2)}.$$

Now, each $N_t^{(i)}$, $i \in \{1,2\}$, only depends on the arrivals times. Since the two stochastic processes $N_t^{(i)}$, $i \in \{1,2\}$'s are independent. So the law of their sum is the convolution product of each element od the sum. So,

$$N_t^{(1)} + N_t^{(2)} \sim \mathcal{P}((\lambda_1 + \lambda_2)t).$$

for all $t \geq 0$. Form there we easily get (IC), (PM), (SI) and (I2) listed in Subsection 2.3 (page 133). Finally, N_t is a counting process of intensity Poisson $\lambda = \lambda_1 + \lambda_2$, which right-continuous if both processes $N_t^{(i)}$, $i \in \{1, 2\}$, are. □

4.4. Decomposition of Stochastic processes.

Proposition 6.9. *We consider a Poisson stochastic of arrival times of intensity $\lambda > 0$. Let us associate to the stochastic processes a Bernoulli trial of probability $p \in]0, 1[$, which is independent of the arrival times. To each arrival, we perform the Bernoulli experience independently from the former performances. If we have a success, the client is directed to a desk A, and to a desk B otherwise. Let $N_t^{(1)}$ the counting functions at the desk A and by $N_t^{(2)}$ that of the desk B. We have :*

(a) *$N_t^{(1)}$ and $N_t^{(2)}$ are independent.*

(b) $N_t = N_t^{(1)} + N_t^{(2)}$.

(c) *$N_t^{(1)}$ is a Poisson stochastic process of intensity $p\lambda$ and $N_t^{(2)}$ Poisson stochastic process of intensity $(1-p)\lambda$.*

Proof. Let us denote the event $(N_t^{(1)} = k | N_t = n)$: after n arrival times, we have k successes in the Bernoulli trial. Hence,

if $0 \leq k \leq n$, we have

$$\mathbb{P}(N_t^{(1)} = k | N_t = n) = \binom{n}{k} \ p^k (1-p)^{n-k}.$$

Hence

$$\begin{aligned}
\mathbb{P}(N_t^{(1)} &= k) = \sum_{n=0}^{\infty} \mathbb{P}(N_t = n) \times \mathbb{P}(N_t^{(1)} = k | N_t = n) \\
&= \sum_{n \geq k}^{\infty} \frac{n!}{p!(n-p)!} p^k (1-p)^{n-k} \times \frac{(\lambda t)^n}{n!} \exp(-\lambda t) \\
&= \frac{(\lambda t p)^k}{k!} \exp(-\lambda t) \sum_{n \geq k}^{\infty} \frac{(\lambda (1-p) t)^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p t)^k}{k!} \exp(-\lambda t) \exp(\lambda (1-p) t) \\
&= \frac{(\lambda p t)^k}{k!} \exp(-\lambda p t),
\end{aligned}$$

leading to

$$N_t^{(1)} \sim \mathcal{P}(\lambda p t).$$

At the same time, we have

$$N_t^{(2)} \sim \mathcal{P}(\lambda (1-p) t).$$

Let us show that the two margins are independent. We notice that $(N_t^{(1)} = k, N_t^{(2)} = \ell)$ is included in $(N_t = k + \ell)$. We get

$$\begin{aligned}\mathbb{P}(N_t^{(1)} &= k, N_t^{(2)} = \ell)) = P(N_t^{(1)} = k, N_t^{(2)} = n-k, N_t = k+\ell) \\ &= \mathbb{P}(N_t = k+\ell) P(N_t^{(1)} = k, N_t^{(2)} = n-k | N_t = k+\ell) \\ &= \frac{(\lambda t)^{k+\ell}}{(k+\ell)!} \times \binom{k+\ell}{k} p^k (1-p)^\ell = \frac{(\lambda p t)^k}{k!} \times \frac{(\lambda(1-p)t)^\ell}{\ell!} \\ &= \mathbb{P}(N_t^{(1)} = k) \times \mathbb{P}(N_t^{(2)} = \ell). \ \square\end{aligned}$$

4.5. The Bus paradox.

Let us suppose that the Poisson arrival times are those of buses at a given station S_0 and that $\lambda > 0$ is the intensity of the stochastic process. Let us take two passengers arriving between two arrival times of buses, in a random inter-arrival time interval. The bus paradox says that : what ever be the arrival times of these two passengers in a random inter-arrival time interval, the mathematical expectations of their waiting times before the next bus are the same and are equal to $1/\lambda$.

Indeed, if a is the arrival time of the passenger between Z_ν and $Z_{\nu+1}$, where ν is the random variable defined by

$$\nu = \max\{k \geq 0, Z_k \leq a\}.$$

The next bus arrives at the time $Z_{\nu+1}$ and the waiting time

$$U = Z_{\nu+1} - a.$$

Let us compute the expectation of U. We may use a drawing to better understand the event $Z_\nu \leq a < Z_{\nu+1}$. Let us see the relation between

$$V = a - Z_{\nu+1} \ \ and \ \ U = Z_{\nu+1} - a.$$

We notice that for $u \geq 0$, $(U > u) = (Z_{v+_1} > a+u)$. This event exactly means that there is no arrivals between a and $a+u$, Hence,

$$(U > u) = (N_{a+u} - N_a = 0)$$

As well, for all $v \geq 0$, we have $(V > u) = (Z_v < a - v)$. If $a - v \leq 0$, the event $(V > u) = (Z_v < a - v)$ is impossible since $Z_v > 0$ $a.s.$ and if $a - v > 0$, that is $Z_\nu \leq a - v \leq a$, it is exactly $N_a - N_{a-v} = 0$. Hence for $0 < v \leq a$, we get

$$(V > v) = (N_a - N_{a-v} = 0).$$

Hence,

$$\begin{aligned}\mathbb{P}(U &> u, V > v) = \mathbb{P}(N_{a+u} - N_a = 0, N_a - N_{a-v} = 0)\\ &= \mathbb{P}(N_{a+u} - N_a = 0) \times \mathbb{P}(N_a - N_{a-v} = 0)\\ &= e^{-\lambda u} \times e^{-\lambda v}\\ &= \mathbb{P}(U > u)\mathbb{P}(V > v).\end{aligned}$$

For $v \geq a$,

$$\begin{aligned}\mathbb{P}(U &> u, V > v) = \mathbb{P}(N_{a+u} - N_a = 0, N_a - N_{a-v} = 0)\\ &= \mathbb{P}(N_{a+u} - N_a = 0, \emptyset)\\ &= 0\\ &= \mathbb{P}(U > u)\mathbb{P}(V > v), \quad since \quad (, \quad since \quad (\mathbb{P}(V > v) = 0).\end{aligned}$$

with

$$\mathbb{P}(V > v) = \begin{cases} 0 & si \quad v \geq a \\ e^{-\lambda v} & si \quad 0 < v < a \end{cases}.$$

and for all $u > 0$,

$$\mathbb{P}(U > u) = e^{-\lambda u}.$$

So, for all $u > 0, \ v > 0$,

$$\mathbb{P}(U > u, V > v) = \mathbb{P}(U > u)\mathbb{P}(V > v).$$

It follows that U and V are independent with the concerned laws, We deduce that

$$\begin{aligned}\mathbb{E}(U) &= \int_0^{+\infty} \mathbb{P}(U > u)du = \int_0^{+\infty} e^{-\lambda v}dv \\ &= \lambda^{-1}.\end{aligned}$$

Hence the expectation of waiting times for two passengers arriving at the station in a random inter-arrival time interval, are the same.

Concerning V and the inter-arrival time $Z_{v+1} - Z_v$, it follows that

$$\begin{aligned}\mathbb{E}(V) &= \int_0^{+\infty} \mathbb{P}(V > v)dv = \int_0^{a} e^{-\lambda v}dv \\ &= \lambda^{-1}(1 - e^{-\lambda a}) \to \lambda^{-1},\end{aligned}$$

as $a \to +\infty$ and

$$\begin{aligned}\mathbb{E}(Z_{v+1} - Z_v) &= \mathbb{E}(Z_{v+1} - a + a - Z_v) \\ &= \mathbb{E}(U) + \mathbb{E}(V) \to \lambda^{-1}\end{aligned}$$

as $a \to +\infty$.

5. Kolmogorov equations

One of the most important and useful characterization of counting functions of Poisson processes resides in the expansions of the function

$$0 \leq t \to p_k(t) = \mathbb{P}(N_t = k), \;\; t \geq 0, \;\; k \geq 0$$

in neighborhoods of zero, in particular differential equations satisfied by those expansions. The following proposition gives interesting properties of $p(t), \;\; t \geq 0$.

Proposition 6.10. *For any counting function of a Poisson stochastic processes* $\{N(t), \; t \geq 0\}$ *and by denoting*

$$p_k(t) = \mathbb{P}(N_t = k), \; t \geq 0, \; k \geq 0$$

we have, as $h \downarrow 0$,

(D1) $p_0(h) = 1 - \lambda h + o(h)$.

(D2) $p_1(h) = \lambda h + o(h)$.

(D3) $p_k(h) = o(h)$, *uniformly for* $k \geq 2$, *meaning*

$$\sup_{k \geq 2} p_k(h) = o(h).$$

Comments. Property (D1) means that the probability of having more than one arrival in a small interval is λ-proportional to the length of that interval as it goes to zero :

$$\mathbb{P}(N_h > 1) = 1 - P(N_h = 0) = h(\lambda + o(1)),$$

that is : the more the interval is small, the less the probability of finding more than one arrival in the interval, with a constant ratio λ.

Property (D2) says that, for all $t \geq 0$, for all $k \geq 0$,

$$\frac{\mathbb{P}(N_{t+h} = k+1) - \mathbb{P}(N_t = k)}{h} = \frac{\mathbb{P}(N_h = 1)}{h} \to \lambda,$$

meaning that the probability to have exactly one arrival in an interval of length $h > 0$ is proportional to h with ratio λ, meaning also that the probability of having one arrival at each instant is λ which become an instant probability (exactly as a *pdf*). This explains the name of intensity of λ.

Finally, (D3) means that the probability of having more than two arrivals in a small interval is infinitely small (of order 1) with respect to the length of that interval.

Proof. The formulas we want to prove derive from expansions of the exponential function at zero. Let us prove each property.

(D1) For $k = 0$,

$$p_0(h) = \frac{(\lambda h)^0}{0!} \exp(-\lambda h) = 1 - \lambda h + o(h).$$

(D2) For $k = 1$,

$$\begin{aligned} p_1(h) &= \frac{(\lambda h)^1}{1!} \exp(-\lambda h) = \lambda h(1 - \lambda h + o(h)) = \lambda h - \lambda^2 h^2 + o(\lambda h^2) \\ &= \lambda h + o(h). \end{aligned}$$

(D3) We have

$$\mathbb{P}(N_h \geq 2) = 1 - \mathbb{P}(N_h = 0) - \mathbb{P}(N_h = 1).$$

By using the two previous properties, we get

$$\begin{aligned} \mathbb{P}(N_h &\geq 2) = 1 - (1 - \lambda h + o(h)) - (\lambda h + o(h)) \\ &= o(h). \end{aligned}$$

By decreasingness of the events $(N_h \geq k)$ in k, it follows that

$$\forall k \geq 2, \ \ (N_h \geq k) \subset (N_h \geq 2)$$

and hence

$$\sup_{k\geq 2}\mathbb{P}(N_h \geq k) \leq \mathbb{P}(N_h \geq 2) = o(h).\square$$

Now, we are going to do the reverse way.

PROPOSITION 6.11. *A stochastic process $(N_t)_{t\geq 0}$ with non-negative integer values and satisfying Properties (D1), (D2), (D3),*

(IC) $N_0 = 0$ a.s.,

(SI) $(N_t)_{t\geq 0}$ has stationarity increments

and

(I2) $(N_t)_{t\geq 0}$ has independent increments.

Then $(N_t)_{t\geq 0}$ is a counting function of a stochastic process, that is we have, on top of (IC), (SI) and (I2),
(PM) for all $t \geq 0$, $N_t \sim \mathcal{P}(\lambda t)$.

Proof. Let as assume that the properties (D1), (D2),(D3), (IC), (SI) and (I2) hold. We fix $t \geq 0$ and $k \geq 0$. So on only have to prove (PM) to complete the proof. We are going to determine the marginal laws *through* the Kolmogorov equations. We have :

$$\begin{aligned} p_0(t+h) &= \mathbb{P}(N_{t+h}=0) \\ &= \sum_{m\geq 0}\mathbb{P}(N_{t+h}=0, N_h=m) \\ &= \mathbb{P}(N_{t+h}=0, N_h=0) = \mathbb{P}(N_{t+h}-N_h=0, N_h=0) \\ &= \mathbb{P}(N_{t+h}-N_h=0)\times\mathbb{P}(N_t=0) \\ &= p_0(t)p_0(h) = p_0(t)-\lambda h p_0(t)+o(h), \end{aligned}$$

It follows by (D1) that

$$\frac{p_0(t+h)-p_0(t)}{h} = -\lambda p_0(t)+o(1) \to p'(t) = -\lambda p_0(t).$$

Hence for $k\geq 1$,

$$\begin{aligned} p_k(t+h) &= \mathbb{P}(N_{t+h}=k) \\ &= \sum_{m\geq 0}\mathbb{P}(N_{t+h}=k, N_h=m) \\ &= \sum_{m\geq 0}\mathbb{P}(N_{t+h}-N_h=k-m, N_h=m) \\ &= \mathbb{P}(N_{t+h}-N_h=k, N_h=0)+\mathbb{P}(N_{t+h}-N_h=k-1, N_h=1)+A, \end{aligned} \tag{5.1}$$

with

$$A = \sum_{m\geq 2}\mathbb{P}(N_{t+h}-N_h=k-m, N_h=m).$$

In the two last lines, we have decomposed the expression according to the different values of m : $m=0$, $m=1$ and $m\geq 2$. It follows that

$$A = \sum_{m\geq 2} \mathbb{P}(N_{t+h} - N_h = k-m, N_h = m) \tag{5.2}$$
$$= \sum_{m\geq 2} \mathbb{P}(N_{t+h} - N_h = k-m,) \times \mathbb{P}(N_h = m)$$
$$= \sum_{m\geq 2} \mathbb{P}(N_{t+h} - N_h = k-m \times \mathbb{P}(N_h = m).$$

Since $\mathbb{P}(N_h = m) = p_m(h) = o(h)$ uniformly in m, it comes that

$$A = o(h) \sum_{m\geq 2} \mathbb{P}(N_{t+h} - N_h = k-m) \tag{5.3}$$
$$= o(h)\mathbb{P}(N_{t+h} - N_h \geq k-2) = o(h).$$

Furthermore,

$$\mathbb{P}(N_{t+h} - N_h = k, N_h = 0) = \mathbb{P}(N_{t+h} - N_h = k) \times \mathbb{P}(N_h = 0) \tag{5.4}$$
$$= p_k(t)p_0(h) = p_k(t) - \lambda h p_k(t) + o(h)$$

and, finally,

$$\mathbb{P}(N_{t+h} - N_h = k-1, N_h = 1) \tag{5.5}$$
$$= \mathbb{P}(N_{t+h} - N_h = k-1) \times \mathbb{P}(N_h = 1)$$
$$= p_{k-1}(t)p_1(h) = \lambda h p_{k-1}(t) + o(h).$$

By putting Formulas (5.1), (5.3), (5.4) and (5.5) together, we get

$$\frac{p_k(t+h)-p_k(t)}{h} = -\lambda(p_k(t)+p_{k-1}(t)) + o(1).$$

Thus, $p_k(t)$ is differentiable and

$$p_k'(t) = \lambda(p_{k-1}(t) - p_k(t)).$$

So we have obtained the following equations, called Kolmogorov Equations,

$$\forall t \geq 0, \forall k \geq 0, \quad \frac{1}{\lambda}p_k'(t) = p_{k-1}(t) - p_k(t)),$$

with the convention that $p_{k-1}(t) \equiv 0$.

In the second part of this proof, we are going to solve the Kolmogorov Equations by determining the moment function of each N_t, By definition,

$$f(t,z) = \mathbb{E}(z^{N_t}) = \sum_{k\geq 0} z^k \mathbb{P}(N_t = k) = \sum_{k\geq 0} z^k p_k(t), z \geq 0.$$

This function is uniformly summable on the domain $|z| \leq 1$. It is differentiable on it term by term. By the Kolmogorov equations, we have

$$\begin{aligned} f'(t,z) &= \sum_{k\geq 0} z^k p_k'(t) \\ &= \lambda \sum_{k\geq 0} z^k (p_{k-1}(t) - p_k(t)) \\ &= \lambda \left\{ \sum_{k\geq 0} z^k p_{k-1}(t) - \sum_{k\geq 0} z^k p_k(t)) \right\} \\ &= \lambda \left\{ \sum_{k\geq 1} z^k p_{k-1}(t) - \sum_{k\geq 0} z^k p_k(t)) \right\} \\ &= \lambda \left\{ z \sum_{k\geq 0} z^{k-1} p_{k-1}(t) - \sum_{k\geq 0} z^k p_k(t)) \right\} \\ &= \lambda(z-1) f(t,z). \end{aligned}$$

It follows

$$\frac{f'(t,z)}{f(t,z)} = \lambda(z-1).$$

The general solution of that equation is

$$f(t,z) = K(t) \exp(\lambda t(z-1)), \tag{5.6}$$

where $K(t)$ is the integration constant. For $t \geq 0$, for $z = 1$,

$$f(t,1) = \mathbb{E}(1^{N_t}) = \sum_{k\geq 0} p_k(t) = 1$$

and Equation (5.6) gives $z = 1$

$$K(t) = 1.$$

Hence

$$f(t, z) = \exp(\lambda(z - 1)), z \geq 0 \tag{5.7}$$

and since this is the moment function of a Poisson law of parameter λt, we get the marginal law of N_t. We conclude by using Proposition 6.3. ■

Bibliography

Kai Lai Chung (1974). *A Course in Probability Theory*. Academic Press. New-York.

(2005). Allan Gutt (2005). *Probability Theory : a graduate course*. Springer.

Michel Loève (1997). *Probability Theory I*. Springer Verlag. Fourth Edition.

Lo, G.S.(2016). A Course on Elementary Probability Theory. SPAS Editions. Saint-Louis, Calgary, Abuja. Doi : 10.16929/sbs/2016.0003.

Lo, G.S.(2018). *Measure Theory and Integration by and for the learner*. SPAS Editions. Saint-Louis, Calgary, Abuja. Doi : http://dx.doi.org/10.16929/sbs/2016.0005, ISBN : 978-2-9559183-5-7.

Lo, G.S.(2018). *Mathematical Foundation of Probability Theory*. SPAS Books Series. Saint-Louis, Senegal - Calgary, Canada. Doi : http://dx.doi.org/10.16929/sbs/2016.0008. Arxiv : arxiv.org/pdf/1808.01713

Lo, G.S.(2019). *Introduction to Statistical Modelling*. SPAS Books Series.

Klugman S., Panjer H., Willmot G.E.(2004). *Loss Models from Data to Decision*. Wiley. New-York