








# MAS-1

## Study Review

Nicholas Langevin

19 septembre 2019

-  Probability Review
-  Stochastic Processes
-  Life Contingencies
-  Simulation
-  Statistics
-  Extended Linear Model
-  Time Series

## Lesson 1 : Probability Review

- > **Bernoulli Shortcut** : If a random variable can only assume two values  $a$  and  $b$  with probability  $q$  and  $1 - q$ , then its variance is  $q(1 - q)(b - a)^2$

## Lesson 2 : Parametric Distributions

- > **Transformations** :
  - Transformed :  $\tau > 0$
  - Inverse :  $\tau = -1$
  - Inverse-Transformed :  $\tau < 0, \tau \neq 1$

## Lesson 4 : Markov Chains

- > **Chapman-Kolmogorov** :
 
$$P_{ij}^{(n+m)} = \sum_{k=0}^{\infty} P_{ik}^{(n)} P_{kj}^{(m)}$$
- > **Gambler's ruin** : Let  $N$  be the target and  $j$  the actual status.
 
$$p_j = \begin{cases} \frac{j}{N} & , r = 1 \\ \frac{r^j - 1}{r^N - 1} & , r \neq 1 \end{cases}$$
 où  $r = \frac{q}{p}$ ,  $p$  : winning prob.
- > **Algorithmic efficiency** : with  $N_j$  = number of steps from  $j^{th}$  solution to best solution.
 
$$E[N_j] = \sum_{i=1}^{j-1} \frac{1}{i}$$

$$\text{Var}(N_j) = \sum_{i=1}^{j-1} \left( \frac{1}{i} \right) \left( 1 - \frac{1}{i} \right)$$
 As  $j \rightarrow \infty$ ,  $E[N_j] \rightarrow \ln j$ ,  $\text{Var}(N_j) \rightarrow \ln j$

## Lesson 5 : Markov Chain Classification

- > An **absorbing** state is one that cannot be exited.
- > State  $j$  is **accessible** ( $i \rightarrow j$ ) from state  $i$  if  $p_{ij}^n > 0$ ,  $\forall n \geq 0$ .
- > Two states **communicate** if  $i \leftrightarrow j$ .
- > A **class** of states is a maximal set of states that communicate with each other.
- > A Markov chain is **irreducible** if it has only one class.
- > A state (class) is **recurrent** if the probability of reentering the state is 1.  $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$
- > A state (class) is **transient** if it is not recurrent.  $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$
- > A finite Markov Chain must have at least one recurrent class. If it is irreducible, then it is recurrent.

## Lesson 6 : Markov Chains Limiting Probability

- > A chain is **positive recurrent** if the expected number of transitions until the state occurs is finite, **null recurrent** otherwise. Null recurrent means that the long-term proportion of time in each state is 0.
- > A chain is **periodic** when states occur every  $n$  periods for  $n > 1$ .
- > A chain is **aperiodic** when the period is 1. In other words,  $P_{ii}^{(1)} > 0, \forall i$
- > A chain is **ergodic** when the chain is aperiodic and positive irreducible recurrent.
- > **Stationary probability** :
 
$$\pi_j = \sum_{i=1}^n P_{ij} \pi_i \quad \sum_{i=1}^n \pi_i = 1$$
- > **Limiting probabilities** : if the chain is ergodic, then

$$P^{(\infty)} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$$

## Lesson 7 : Time in Transient States

- > Tips : **Inverting a matrix**
- >  $S = (I - P_{\text{transient}})^{-1}$ , where  $s_{ij}$  is the time in state  $j$  given that the current state is  $i$ .
- >  $f_{ij} = \frac{s_{ij} - \delta_{i,j}}{s_{jj}} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$ , where  $f_{ij}$  is the probability that state  $i$  ever transitions to state  $j$ .

## Lesson 8 : Branching Processes

- > A branching process is a special type of Markov chain representing the growth or extinction of a population.
- >  $E[X_n] = E[Z]^n$ , where  $E[Z]$  is the expected number of people born in a generation.
- >  $\text{Var}(X_n) = \text{Var}(Z) \cdot E[Z]^{n-1} \sum_{k=1}^n E[Z]^{k-1}$
- > If  $X_0 \neq 1$  mean and variance of  $X_n$  need to be multiplied by  $X_0$ .

- > **Probability of extinction** :

$$\pi_0 = \sum_{j=1}^{\infty} p_j \pi_0^j$$

$$- \mu \leq 1 \Rightarrow \pi_0 \geq 1, \text{ if } X_0 = 1.$$

$$- \mu > 1 \Rightarrow \pi_0 < 1, \text{ if } X_0 = 1.$$

For cubic equation, it is guaranteed to factor ( $\pi_0 - 1$ ). Tips : **Synthetic Division**

## Lesson 9 : Time Reversible

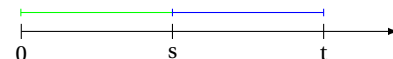
- > If  $Q$  is the reverse-time Markov chain for ergodic  $P$ , then
 
$$\pi_i Q_{ij} = \pi_j P_{ji}$$
 with  $P_{ii} = Q_{ii}$  and if  $p_{ij} = 0 \Leftrightarrow q_{ji} = 0$
- > If  $Q = P$ , then  $P$  is said to be **time-reversible**.

## Lesson 10 : Exponential Distribution

- > **Lack of memory** :
 
$$\Pr(X > k + x | X > k) = \Pr(X > x)$$
- > **Minimum** : if  $X_i \sim \text{Exp}(\lambda_i)$ , then
 
$$\min(X_1, X_2, \dots, X_n) \sim \text{Exp}\left(\sum_{i=1}^n \lambda_i\right)$$
- > The sum of 2 exponential random variables is the sum of the maximum and the minimum, since one must be the min and the other the max.
 
$$X_1 + X_2 = \min(X_1, X_2) + \max(X_1, X_2)$$

## Lesson 11 : Poisson Process

- >  $X(t) \sim \text{Poisson}[m(t)]$ , where  $m(t)$  is **mean value function** representing the mean of the number of events before time  $t$ .
- > Poisson process can't decrease over time.  $N(t) \geq N(s) \quad \text{for } t \geq s$
- >  $N(0) = 0$
- > Increments are **independent** :



$$\Pr[N(t) - N(s) = n | N(s) = k] = \Pr[N(t) - N(s) = n]$$

- > **Non-homogeneous Poisson process** :

$$m(t) = \int_0^t \lambda(u) du$$

where  $\lambda(t)$  is the **intensity function**

- > **Homogeneous Poisson process** : The Poisson process is said to be homogeneous when the intensity function is a constant.

$$m(t) = \int_0^t \lambda du = \lambda t$$

We then say that the process has **stationary increments**.

$$\Pr[N(s - t)] = \Pr[N(t) - N(s)]$$

## Lesson 12 : Poisson Process Time To Next Events

- >  $T_n$  is the time between the  $n^{th}$  event and the  $(n-1)^{th}$  event.
- >  $S_n = \sum_{i=1}^n T_i$ , is the time for the  $n^{th}$  event.
- >  $F_{T_1}(t) = 1 - e^{-\int_0^t \lambda(u) du}$
- > For homogeneous process :
 
$$T_n \sim \text{Exp}(\lambda)$$

$$S_n \sim \text{Gamma}(n, \lambda)$$

## Lesson 13 : Poisson Process Counting Special Type

- > If event of type 1 occur with probability  $\alpha_1(t)$ , then the event follow a Poisson process with intensity  $\lambda(t) \cdot \alpha_1(t)$ .

$$m(t) = \int_0^t \lambda(u) \alpha_1(u) du$$

## Lesson 14 : Poisson Process Other Characteristics

- > Only for homogeneous Poisson processes.
- > The probability of  $k$  event from process 1 is given by :

$$k \sim \text{Binomial}\left(k+l-1, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$$

Then the probability that  $k$  event from process 1 occur before  $l$  from process 2 is :

$$\sum_{i=k}^{k+l-1} \binom{k+l-1}{i} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{k+l-1-i}$$

- > Given that exactly  $N(t) = k$  Poisson events occurred before time  $t$ , the joint distribution of event time is the joint distribution of  $k$  independent uniform random variables on  $(0, t)$ .

$$F_{S_1, \dots, S_n | n(t)}(s_1, \dots, s_n | k) = \frac{k!}{t^k}$$

- > For  $k$  independent uniform random variable on  $(0, t)$ , the expected value of the  $j^{\text{th}}$  order statistics is :  $E[T^{(j)}] = \frac{j \cdot t}{(k+1)}$ .
- > Tips : **Statistic Order**

## Lesson 15 : Poisson Process Sums and Mixtures

- > A **Sums** of independent Poisson random variables is a Poisson random with intensify function  $\lambda(t) = \sum \lambda_i(t)$ . **Warning : Subtraction don't give a Poisson random variable.**
- > A **Mixture** of Poisson processes is not a Poisson processes.

- **Discrete** mixture :

$$F_{X(t)}(t) = \sum_i w_i F_{X_i(t)}(t)$$

where  $w_i > 0, \sum w_i = 1$

- **Continuous** mixture :

$$F_{X(t)}(t) = \int F_{X_u(t)}(t) f(u) du$$

- If  $N(t) | \lambda$  is a Poisson random variable and  $\lambda \sim \text{Gamma}(\alpha, \theta)$ , then  $N(t) \sim \text{NegBin}(r = \alpha, \beta = \theta t)$ .

## Lesson 16 : Compound Poisson Processes

- > A **compound** random variable  $S$  is define by  $S = \sum_{i=1}^N X_i$  where  $N$  is the **primary** distribution and  $X$  the **secondary** distribution.

- > If  $N(t)$  is a Poisson process, then  $S(t)$  is a compound Poisson process with :

$$\begin{aligned} - E[S(t)] &= \lambda t E[X] \\ - \text{Var}(S(t)) &= \lambda t E[X^2] \end{aligned}$$

- > If  $X_i$  is discrete, we can separate the process into a sum of subprocess view in **Lesson 13 : Poisson Process Counting Special Type**.

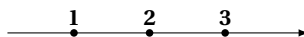
- > **Sums of compound** homogeneous Poisson process is also a Poisson process with :

$$\begin{aligned} - N(t) &\sim \text{Pois}(\sum \lambda_i) \\ - F_X(x) &= \sum_i w_i F_{X_i(t)}(t), \quad w_i = \frac{\lambda_i}{\sum \lambda_i} \end{aligned}$$

## Lesson 17 : Reliability Structure Functions

- >  $\phi(\mathbf{x})$  is the **structure** function for a system. It equal 1 if the system work, 0 otherwise.

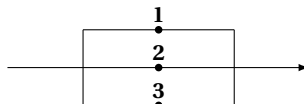
- > A **series** system is define as a **minimal path set**. The system is working if all components are working.



The serie structure function is define as

$$\phi(\mathbf{x}) = \prod_{i=1}^n x_i$$

- > A **parallel** system is define as a **minimal cut set**. The system is working if at least 1 components is working.



The parallel structure function is define as

$$\phi(\mathbf{x}) = 1 - \prod_{i=1}^n (1 - x_i)$$

- > Tips : Minimal path set is all way for the system to work, and the minimal cut set is all the way for the system to not work.

- > Tips : If set is  $\{1, 2, 3\}$  and  $\{1, 2\}$ , the **minimal** mean we only take  $\{1, 2\}$ .

- > Tips : **Minimal cut** is a serie of parallel structure and **minimal path** is a parallel of serie structure.

## Lesson 18 : Reliability Probabilities

- >  $r(\mathbf{p})$  is the same polynomial as  $\phi(\mathbf{x})$ .

- > Inclusion/exclusion bounds using minimal path :

$$\begin{aligned} r(\mathbf{p}) &\leq \sum A_i \\ r(\mathbf{p}) &\geq \sum A_i - \sum A_i \cup A_j \end{aligned}$$

$$r(\mathbf{p}) \leq \sum A_i - \sum A_i \cup A_j + \sum A_i \cup A_j \cup A_k$$

where  $A_i = \sum p_i$  is the probability of the  $i^{\text{e}}$  minimal path set work.

- > Inclusion/exclusion bounds using minimal cut :

$$1 - r(\mathbf{p}) \leq \sum A_i$$

$$1 - r(\mathbf{p}) \geq \sum A_i - \sum A_i \cup A_j$$

$$1 - r(\mathbf{p}) \leq \sum A_i - \sum A_i \cup A_j + \sum A_i \cup A_j \cup A_k$$

where  $A_i = \sum (1 - p_i)$  is the probability of the  $i^{\text{e}}$  minimal cut set work.

- > Bounds using intersections :

$$\prod \phi(\mathbf{X})^{\text{min. cut}} \leq r(\mathbf{p}) \leq \prod \phi(\mathbf{X})^{\text{min. path}}$$

- > **Random graph** :

$$1 - P_n = \sum_{k=1}^{n-1} \binom{n-1}{k-1} q^{k(n-k)} P_k$$

$$1 - P_n \leq (n+1) q^{n-1}$$

$$P_1 = 1$$

## Lesson 19 : Reliability Time to Failure

- > Expected amount of time to failure :

$$E[\text{system life}] = \int_0^\infty r(\bar{F}(t)) dt$$

where,

- For serie system :

$$r(\bar{F}(t)) = \prod_{i=1}^n \bar{F}_i(t)$$

- For parallel system :

$$r(\bar{F}(t)) = 1 - \prod_{i=1}^n F_i(t)$$

- > **Shortcut** :  $k$  out of  $n$  system with exponentials( $\theta$ ) :  $E[T] = \theta \sum_{i=k}^n \frac{1}{i}$

- > **Hazard rate function** (failure rate function) :

$$h(t) = \frac{f(t)}{\bar{F}(t)}$$

and we say that the distribution

- is an increasing failure rate if  $h(t)$  is non-decreasing function of  $t$ .

- is an decreasing failure rate if  $h(t)$  is non-increasing function of  $t$ .

- > **Cumulative hazard function** :

$$H(t) = \int_0^t h(u) du = -\ln \bar{F}(t)$$

with  $\frac{H(t)}{t}$  the average of the hazard rate.

## Lesson 20 : Survival Models

$${}_t p_x = \frac{\ell_{x+t}}{\ell_x}, \quad {}_t q_x = \frac{\ell_x - \ell_{x+t}}{\ell_x}$$

$${}_t |u q_x = \frac{\ell_{x+t} - \ell_{x+t+u}}{\ell_x}$$

$${}_t + u p_x = {}_t p_x \cdot {}_t p_{x+u}$$

$${}_t |u q_x = {}_t + u q_x - {}_t q_x = {}_t p_x \cdot {}_t + u q_{x+t}$$

- > Let be  $N_x$  the number of life surviving to age  $x$ , then

$$(N_{x+t} | N_x = n) \sim \text{Bin}(n, {}_t p_x)$$

- > **Force of mortality** :

$$\mu_{x+t} = \frac{f_{T_x}(t)}{{}_t p_x} = -\frac{d}{dt} \ln {}_t p_x$$

### Linear interpolation(D.U.D) :

$$\ell_{x+t} = (1-t)\ell_x + t\ell_{x+1}$$

Shortcut :  $\forall t \in (0, 1), \forall x \in \mathbb{N}, x < x+t < x+1 :$

$$\rightarrow tq_x = t \cdot q_x$$

$$\rightarrow \mu_{x+t} = \frac{q_x}{1-t \cdot q_x}$$

- > **Expected life time** : Let  $k_x = \lfloor T_x \rfloor$ , the full years until death. Then  $e_x$  is the **curtate life expectancy** and  $\bar{e}_x$  the **complete life expectancy**.  $\omega$  is the age where  $\ell_\omega = 0$  and  $\omega = \infty$  by convention is nothing is said.

$$e_x = E[K_x] = \sum_{k=1}^{\omega-x-1} k p_x$$

$$\bar{e}_x = E[T_x] = \int_0^{\omega-x} t p_x dt \stackrel{\text{D.U.D}}{=} e_x + 0.5$$

## Lesson 21 : Contingent Payments

The contract here are define with  $K_x$  to pay at the end of death year. All same contract can be define with  $T_x$  to pay at the moment of death. Then we use integral instead of sum and use

$$\Pr(K = k) = k p_x q_{x+k} \Rightarrow f_{T_x}(t) = t p_x \mu_{x+t}$$

### Life Insurance :

- Whole Life insurance :

$$A_x = \sum_{k=0}^{\infty} v^{k+1} k p_x q_{x+k}$$

- Term Life insurance :

$$A_{x:\overline{n}|} = \sum_{k=0}^n v^{k+1} k p_x q_{x+k}$$

- Deferred insurance :

$${}_m|A_x = \sum_{k=m}^{\infty} v^{k+1} k p_x q_{x+k}$$

- Endowment insurance :

$$A_{x:\overline{n}|} = A_{x:\overline{n}|}^1 + n E_x$$

- Pure Endowment :

$${}_n E_x = v^n n p_x$$

### Life Annuities :

- Whole Life annuity

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k k p_x$$

- Temporary Life annuity

$$\ddot{a}_{x:\overline{n}|} = \sum_{k=0}^n v^k k p_x$$

- Deferred annuity

$${}_m|\ddot{a}_x = \sum_{k=m}^{\infty} v^k k p_x$$

- Certain and life annuity

$$\ddot{a}_{x:\overline{n}|} = \ddot{a}_{\overline{n}|} + {}_m|\ddot{a}_x$$

### Illustrative Life Table :

- $A_x = v q_x + v p_x A_{x+1}$
- $\ddot{a}_x = 1 + v p_x \ddot{a}_{x+1}$
- $A_{x:\overline{n}|} = A_x - n E_x A_{x+n}$
- $\ddot{a}_{x:\overline{n}|} = \ddot{a}_x - n E_x \ddot{a}_{x+n}$
- ${}_m|A_x = m E_x A_{x+m}$
- ${}_m|\ddot{a}_x = m E_x \ddot{a}_{x+m}$
- $\ddot{a}_x = 1 + a_x$
- $A_x = 1 - d \ddot{a}_x$

- > **Joint life annuity**( $\ddot{a}_{xy}$ ) make payments until the earliest death pf two lives.

- > **Last survivor annuity**( $\ddot{a}_{\overline{xy}|}$ ) make payments until the last death of two lives.

$$\ddot{a}_x + \ddot{a}_y = \ddot{a}_{xy} + \ddot{a}_{\overline{xy}|}$$

- > **Premiums :**

$$M \cdot A_x = P \ddot{a}_x$$

$$P = \frac{M \cdot A_x}{\ddot{a}_x} = \frac{M}{\ddot{a}_x} - M \cdot d$$

## Lesson 22 : Simulation Inverse Method

- > **Linear congruential generators :**

$$x_k = (a x_{k-1} + c) \bmod m$$

$$x_k = b - \left\lfloor \frac{b}{m} \right\rfloor m$$

where  $b = (a x_{i-k} + c)$  and  $x_0 \equiv \text{seed}$

- > **Inverse transformation method :**

$$\Pr(F^{-1}(u) \leq x) = \Pr(u \leq F(x)) = F(x)$$

then  $x = F^{-1}(u)$  where  $U \sim \text{Unif}(0, 1)$

- Normal Case :  $x = \mu + \sigma z$

- Log-Normal Case :  $x = e^{\mu + \sigma z}$

where  $z = \Phi^{-1}(u)$ , with linear interpolation.

- > **Tips : Discrete Cumulative Function**

- > **Tips :** if  $\uparrow U \equiv \downarrow X$  then  $(1 - u_i) \Rightarrow u_i$

## Lesson 23 : Simulation Application

- >  $\Pr(X \leq x) \approx \frac{1}{m} \sum_{j=1}^m \mathbb{1}_{\{x^{(j)} \leq x\}}$

- >  $E[X^k] \approx \frac{1}{m} \sum_{j=1}^m [x^{(j)}]^k$

- >  $\text{VaR}_k(X) \approx X^{[j_0]}$

- >  $\text{TVaR}_k(X) \approx \frac{1}{m(1-k)} \sum_{j=j_0+1}^m X^{(j)} \mathbb{1}_{\{X^{(j)} > X^{[j_0]}\}}$   
 $\approx \frac{1}{m-j_0} \sum_{j=j_0+1}^m X^{[j]}$

where

- $j_0 = \lfloor m \cdot k \rfloor$
- $m$  is the number of simulations.
- $X^{(j)}$  is the  $j^{\text{th}}$  simulations.
- $X^{[j]}$  is the  $j^{\text{th}}$  simulations in order statistics.

## Lesson 24 : Simulation Rejection Method

- > **General method** : Let  $f(x)$  be the density function of variable to simulate, and let  $g(x)$  be the **base distribution**, a random density function that is easy-to-simulate with nonzero wherever  $f(x) \neq 0$ .

$$c = \max \frac{f(x)}{g(x)}$$

Generate two uniform number  $u_1, u_2$ . Let

$x_1 = G^{-1}(u_1)$ . Accept  $x_1$  only if

$$u_2 \leq \frac{f(x_1)}{c \cdot g(x_1)}$$

- > **Simulating gamma distribution** : Use  $\text{Exp}(\alpha \cdot \theta)$  as the base distribution and  $x = \alpha \cdot \theta$  that maximize  $c$ .

- > **Simulating standard normal distribution** : Generate 3 uniform  $u_1, u_2, u_3$ . Let  $y_1 = -\ln u_2$  and  $y_2 = -\ln u_3$ . Accept  $y_1$  if

$$y_2 \geq \frac{(y_1 - 1)^2}{2}$$

and add  $(-)$  if  $u_3 \geq 0.5$

- > The **Number of iteration** is a Ross-geometric distribution with mean  $c$ . Let be  $\beta$  the mean of a geometric distribution given in the exam appendix :

$$E[N] = 1 + \beta = c$$

$$\text{Var}(N) = \beta(1 + \beta)$$

## Lesson 25 : Estimator Quality

- > **Bias** : This quality measures if, on average, the estimator is on the expected value of the parameter.

$$E[\hat{\theta}] = \theta + \text{bias}_{\hat{\theta}}(\theta)$$

- If  $\text{bias}_{\hat{\theta}}(\theta) = 0$ , then  $\hat{\theta}$  is **unbiased**.

- If  $\lim_{n \rightarrow \infty} \text{bias}_{\hat{\theta}}(\theta) = 0$ , then  $\hat{\theta}$  is **asymptotically unbiased**.

- If  $\text{bias}_{\hat{\theta}}(\theta) \neq 0$ , then  $\hat{\theta}$  is **biased**.

- > **Consistency** : This quality measures if the probability that the estimator is different from the parameter by more than  $\varepsilon$  goes to 0 as  $n$  goes to infinity.

$$\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| > \varepsilon) \rightarrow 0, \forall \varepsilon > 0$$

In other word, as  $n \rightarrow \infty$ ,  $E[\hat{\theta}] \rightarrow \theta$ ,  $\text{Var}(\hat{\theta}) \rightarrow 0$

- > **Efficiency** : This quality measures the variance of the estimator. Lower the variance is, more efficient is the estimator.

$$\text{Efficiency of } \hat{\theta} = \frac{\text{Var}(\hat{\theta})^{\text{rao}}}{\text{Var}(\hat{\theta})}$$

$$\text{Relative efficiency of } \hat{\theta}_1 \text{ to } \hat{\theta}_2 = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

See the **rao-cramer lower bound**.

- > **Mean Square Error** : This quality measures the expected value of the square difference between the estimator and the parameter.

$$\text{MSE}_{\hat{\theta}}(\theta) = E[(\hat{\theta} - \theta)^2] = (\text{bias}_{\hat{\theta}}(\theta))^2 + \text{Var}(\hat{\theta})$$

- > An estimator is called a **uniformly minimum variance unbiased estimator(UMVUE)** if it's unbiased and if there is no other unbiased estimator with a smaller variance for any true value  $\theta$ .

- > Some estimator :

- $\bar{x} = \frac{1}{n} \sum x_i$  is a unbiased estimator of the mean  $\mu$ .  $\text{Var}(\bar{x}) = \frac{1}{n} \text{Var}(x)$

- $s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$  is a unbiased estimator of the variance  $\sigma^2$ .
- $\hat{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{n}$  is an asymptotically unbiased of the variance  $\sigma^2$ .
- $\hat{\mu}'_k = \frac{1}{n} \sum x_i^k$ , where  $\hat{\mu}'_1 = \bar{x}$  and  $\hat{\mu}_k = \frac{1}{n} \sum (x_i - \bar{x})^k$ , where  $\hat{\mu}_1 = 0$  and  $\hat{\mu}_2 = \hat{\sigma}^2$ .

## Lesson 26 : Kernel Density Estimation

- > **Empirical distribution** : All data is assigning a probability of  $\frac{1}{n}$ . This is the same method used for simulation, see **Lesson 23 : Simulation Application**.

$$F_e(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}}$$

$$f_e(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i = x\}} \\ = F_e(x) - F_e(x_{i-1})$$

- > **Kernel Density** is a empirical distribution smoothed with a base function. Let define the scaling factor  $b$  called **bandwidth**.

- The kernel-density estimate of the density function is :  $\hat{f}(x) = \frac{1}{n} \sum k\left(\frac{x-x_i}{b}\right) \Leftrightarrow \sum f_e(x) k\left(\frac{x-x_i}{b}\right)$
- The kernel-density estimate of the distribution is :  $\hat{F}(x) = \frac{1}{n} \sum K\left(\frac{x-x_i}{b}\right)$

- > **Rectangular(uniform, box) kernel** :

$$k(x) = \begin{cases} \frac{1}{2b}, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$K(x) = \begin{cases} 0, & x < -1 \\ 0.5(x+1), & -1 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

$$\hat{f}(x) = \frac{F_e(x+b) - F_e(x-b)}{2b}$$

- > **Triangular kernel** :

$$k(x) = \begin{cases} \frac{1-|x|}{b}, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$K(x) = \begin{cases} 0, & x < -1 \\ \frac{(1+x)^2}{2}, & -1 \leq x \leq 0 \\ 1 - \frac{(1-x)^2}{2}, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

- > **Gaussian kernel** : The distribution become normal with  $\mu = x_i$  and  $\sigma = b$ .

$$k(x) = \frac{e^{-x^2/2}}{b\sqrt{2\pi}}$$

$$K(x) = \Phi(x)$$

- > Other kernel :  $k(x) = \beta(x)$  and  $K(x) = B(x)$

- > **kernel moments** : Let  $X$  be the kernel density estimate and  $x_i$  the empirical estimate.

We then condition on  $x_i$ .

$$E[X] = E[E[X|x_i]] = E[x_i]$$

$$\text{Var}(X_R) = \text{Var}(x_i) + \frac{b^2}{3}$$

$$\text{Var}(X_T) = \text{Var}(x_i) + \frac{b^2}{6}$$

$$\text{Var}(X_G) = \text{Var}(x_i) + b^2$$

- > Tips : For rectangular kernel,  $E[x|x_i]$  is a uniform( $x_i - b, x_i + b$ ).

## Lesson 27 : Method of Moments

- > **Types of data** :

- Complete data : Data is complete if we are given the exact value of each observation.
- Grouped data : Set of interval and we know how many observation are in each.
- Censored data : Value that are in a interval, but we don't know the exact value. Like limits ( $\min(X, u)$ ).
- Truncated data : We have data only when it in certain range, otherwise we don't know. Like deductible ( $X|X > d$ ).

- > **Method of Moments** : We match  $\hat{\mu}'_k = E[X^k]$  and find the parameters. If data is Censored or Truncated, we need to match the censored or truncated moment :  $\hat{\mu}'_k = E[\min(X, u)^k]$  or  $\hat{\mu}'_k = E[X^k | X > d]$ .

- > For pareto distribution, if  $\hat{\mu}'_2 = \hat{\sigma}^2 + \bar{x}^2 \leq 2\bar{x}^2$ , the method of moment is unstable and can't be used.

## Lesson 28 : Percentile Matching

- > **Percentile Matching** : We match  $F_e(\hat{\pi}_p) = p$  and find the parameters.

- For censored data, we need select percentiles within the range of the uncensored portion of the data.
- For truncated data, we need to match the percentiles of the conditional distribution :

$$F(x|X > d) = \frac{\Pr(d < X \leq x)}{\Pr(X > d)} = \frac{F(x) - F(d)}{1 - F(d)}$$

$$S(x|X > d) = \frac{S(x)}{S(d)}$$

- > **Smoothed empirical percentile** :

$$\hat{\pi}_p = (1-h)X^{[j]} + hX^{[j+1]}$$

where

- $j = \lfloor (n+1)p \rfloor$
- $h = (n+1)p - j$
- $X^{[j]}$  is the  $j^{\text{th}}$  order statistics.

## Lesson 29 : Maximum Likelihood Estimators

- > **Maximum Likelihood Estimators** : We maximize the probability of observing the data.

$$L(\theta) = \prod g(x_i; \theta)$$

$$l(\theta) = \sum \ln g(x_i; \theta)$$

- Individual data :  $g(x_i; \theta) = f(x_i)$
- Grouped data :  $g(x_i; \theta) = F(x_i) - F(x_{i-1})$
- Censored data :  $g(x_i; \theta) = S(x_i)$
- Truncated data :  $g(x_i; \theta) = \frac{f(x)}{s(x)}$

## Lesson 30 : MLE Special Techniques

- > Case MLE equals MME

- For Exponential,  $\hat{\theta}^{\text{MLE}} = \bar{x}$
- For Gamma with fixed  $\alpha$ ,  $\hat{\theta}^{\text{MLE}} = \hat{\theta}^{\text{MME}}$
- For Normal,  $\hat{\mu}^{\text{MLE}} = \bar{x}$  and  $(\hat{\sigma}^2)^{\text{MLE}} = \frac{1}{n} \sum (x_i - \bar{x})^2$
- For Binomial,  $mq = \bar{x}$  then given  $m$ ,  $\hat{q}^{\text{MLE}} = \frac{\bar{x}}{m}$
- For Poisson,  $\hat{\lambda}^{\text{MLE}} = \hat{\lambda}^{\text{MME}}$
- For Binomial Negative, given  $r$  or  $\beta$ ,  $(r\beta)^{\text{MLE}} = \bar{x}$

- > Parametrization and Shifting :

- Parametrization : MLE's are independent of parametrization  $\lambda = \frac{1}{\theta} \Leftrightarrow \hat{\lambda}^{\text{MLE}} = \frac{1}{\hat{\theta}^{\text{MLE}}}$

- Shifting the distribution is equivalent of shifting the data.

- > Transformations : MLE's are invariant under one-to-one transformation. Then if we have a transformed variable, we can untransform the data and find the parameter of the untransform distribution.

Tips : **Transformations of distribution**

- > Weibull distribution : If the data is censored(u) or truncated(d), then

$$(\hat{\theta}^{\text{MLE}})^{\tau} = \frac{\sum x_i^{\tau} - \sum d_i^{\tau}}{\sum \mathbb{1}_{\{x_i \leq u\}}}$$

if  $\tau = 1$ , then the distribution is Exponential.

- > Pareto distribution with fixed  $\theta : \hat{\alpha} = -\frac{n}{K}$

$$K = \sum_{i=1}^{n+c} \ln(\theta + d_i) - \sum_{i=1}^{n+c} \ln(\theta + x_i)$$

where  $n \equiv$  number of non-censored(c) data.

- > Single-parameter Pareto :  $\hat{\alpha} = -\frac{n}{K}$

$$K = \sum_{i=1}^{n+c} \ln \max(\theta, d_i) - \sum_{i=1}^{n+c} \ln x_i$$

where  $n \equiv$  number of non-censored(c) data.

- > Uniform(0,  $\theta$ ) : We take the smallest  $\theta$  possible,  $\hat{\theta}^{\text{MLE}} = \max(x_1, \dots, x_n)$

- Censored(u) :  $\hat{\theta}^{\text{MLE}} = \frac{nd}{\sum \mathbb{1}_{\{x_i < d\}}}$



- Grouped : We take the highest interval (L, U).  $\hat{\theta}^{\text{MLE}} = \min(U, \hat{\theta}_{\text{Censored}(L)}^{\text{MLE}})$
- > Bernoulli : Let have a random variable that can take 2 values,  $n$  and  $m$ . Then  $\hat{p} = \frac{n}{n+m}$
- > Tips : If  $L(\theta)$  look like a density distribution,  $\hat{\theta}^{\text{MLE}} \equiv \text{mode of this distribution}$ .
- > **Ground-up loss** is define as  $(x|x>d)$ .

## Lesson 31 : Variance of MLE

- > **Fisher information matrix** :

$$I(\theta) = nE \left[ \left( \frac{d \ln f(x; \theta)}{d\theta} \right)^2 \right]$$

$$= -nE \left[ \frac{d^2 \ln f(x; \theta)}{d\theta^2} \right]$$

using the loglikelihood function

$$I(\theta) = E \left[ \left( \frac{d l(x_1, \dots, x_n; \theta)}{d\theta} \right)^2 \right]$$

$$= -E \left[ \frac{d^2 l(x_1, \dots, x_n; \theta)}{d\theta^2} \right]$$

- > **Rao-Cramer lower bound** is the lowest possible variance for a unbiased estimator  $\hat{\theta}$  of  $\theta$ . Then  $\hat{\theta} \sim \text{Normal}(0, \text{Var}(\hat{\theta})^{\text{rao}})$

$$\text{Var}(\hat{\theta})^{\text{rao}} \geq \frac{1}{I(\theta)}$$

under certain regularity conditions

- The seconde derivative of the loglikelihood exist.
- The support of the density function is not function of  $\theta$ .

## Lesson 32 : Sufficient Statistics

- > A **sufficient statistics** are statistics that capture all the information about the parameter we are estimating that the sample as to offer.
- > A statistics is sufficient when the distribution of a sample given a statistics does not depend on the parameter.  $Y$  is a sufficient statistics for a parameter  $\theta$  if and only if

$$L(x_1, \dots, x_n; \theta | Y) = h(x_1, \dots, x_n)$$

$$L(x_1, \dots, x_n; \theta) = g(y)h(x_1, \dots, x_n)$$

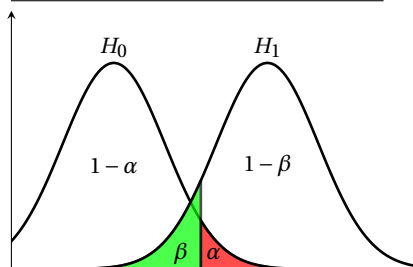
where  $h(x_1, \dots, x_n)$  is a function that does not involve  $\theta$ .

- > **Rao-Blackwell Theorem** : For any unbiased estimator  $\hat{\theta}$  and sufficient statistic  $Y$ , the estimator  $E[\hat{\theta}|Y]$  is unbiased and has variance less than or equal to  $\text{Var}(\hat{\theta})$ .
- > The maximum likelihood estimator is a function of a sufficient statistic.

## Lesson 33 : Hypothesis Testing

- > Let be  $H_0$  the **null hypothesis** and  $H_1$  the **alternative hypothesis**.

	Accept $H_0$	Reject $H_0$
$H_0$ True	$1 - \alpha$	$\alpha$
$H_1$ True	$\beta$	$1 - \beta$



- > The  $\alpha$  value is usually name :

- Probability of Type I error
- Size of critical region
- significance level

The  $\beta$  value is usually name :

- Probability of Type II error

The  $(1 - \beta)$  value is usually name :

- The power of test.

- > We will reject  $H_0$  in favor of  $H_1$  if a certain condition occurred ( $X > \gamma$ ), named the **critical region**. Then the probability of rejecting  $H_0$  is giving by

$$\Pr(X > \gamma | H_0 \equiv \text{true}) = \alpha$$

- > Lowering the probability of type I error came at the cost of raising the probability of type II error. One way to lower both is to increase sample size.

- > The **p-value** is the probability of being greater or equal to the observation if  $H_0$  is true.  $H_0$  is rejected if and only if the p-value is less than the significance level.

$$P_{\text{value}} < \alpha$$

## Lesson 34 : Confidence Interval and Sample Size

- > Let be  $c$  the **confidence coefficient**. Then we can say that we're 100c% confident that the parameter is between  $(a, b)$ , called the **confidence interval**.  $\alpha = 1 - c$

$$\theta \in \hat{\theta} \pm z_{1+\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\theta})}$$

- > We can found the probability that the half-width of the interval is less than  $k$ .

$$\Pr(|\hat{\theta} - \theta| \leq k) \geq \frac{1+c}{2}$$

$$\Phi \left( \frac{k}{\sqrt{\sigma^2/n}} \right) \geq \frac{1+c}{2}$$

- > To find the sample size needed to have a certain  $(\alpha)$  and  $(1 - \beta)$ , we resolve

$$\Pr(\bar{x} > k | H_0) = 1 - \Phi \left( \frac{k - \mu_0}{\sqrt{\sigma^2/n}} \right) = \alpha$$

$$\Pr(\bar{x} > k | H_1) = 1 - \Phi \left( \frac{k - \mu_1}{\sqrt{\sigma^2/n}} \right) = 1 - \beta$$

## Lesson 35 : Confidence Intervals for Means

- > The **chi-square** is a one-parameter family distribution. In definition, it's a gamma with  $\alpha = \frac{n}{2}$  and  $\theta = 2$ . The only parameter  $n$  is called **degree of freedom**.

- Let  $X_i, i = 1, \dots, n$  be normal random variable with mean  $\mu$  and variance  $\sigma^2$ .

$$Y = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_{(n)}^2$$

- Let  $x_i, i = 1, \dots, n, n \geq 2$  be random sample from normal distribution with variance  $\sigma^2$ .

$$W = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi_{(n-1)}^2$$

- Tips :  $\chi_{(2)}^2 \sim \text{Exp}(\theta = 2)$

- > The **student** is a one-parameter family distribution. We define it as

$$T_{(n)} = \frac{Z}{\sqrt{W/n}}$$

where  $Z \sim N(0, 1)$  and  $W \sim \chi_{(n)}^2$ .

Note that as  $n \rightarrow \infty$ ,  $T_{(n)} \rightarrow N(0, 1)$

- > When the variance is unknown, we need to estimate it with the unbiased estimator  $S^2$ .

$$T_{(n-1)} = \frac{\bar{x} - \mu}{\sqrt{S^2/n}}$$

- > Testing the difference of means from two population.

$$x_1, \dots, x_n \sim N(\mu_x, \sigma_x^2)$$

$$y_1, \dots, y_m \sim N(\mu_y, \sigma_y^2)$$

$$T_{(n+m-2)} = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$\text{where } S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{m+n-2}$$

- > Testing for mean of Bernoulli population. Let  $p_0$  the probability on  $H_0$ .

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

## Lesson 36 : Kolmogorov-Smirnov Tests

- > The **Kolmogorov-Smirnov test** is one method for determining how well a parametric model fits its data. This test is only appropriate for continuous distribution.

$$D = \max |F_e(x) - F^*(x; \hat{\theta})|$$

where  $d \leq x \leq u$  and  $F^*(x) = \frac{F(x) - F(u)}{S(d)}$ .

$x_i$	$F^*(x_i)$	$F_e(x_i)$	$F_e(x_i)$	max
$x_1$	0.5	0.2	0.6	0.3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

## Lesson 37 : Chi Square Test

- The **Chi Square** look for equality of means between  $k$  group. Let  $O_i$  be the observation and  $E_i = np_i$  the expected on each group.

$$H_0 : \mu_1 = \dots = \mu_k$$

$$Q = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \left( \frac{O_i^2}{E_i} \right) - n \sim \chi_{(k-1-\theta')}^2$$

Note : This test can be use to test the fit of as parametric model.  $\theta'$  is the number of parameter fitted with the same data as the test.

- Two-dimensional chi-square :**

$$Q = \sum_{i=1}^k \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(k-1)(c-1)}^2$$

## Lesson 38 : Confidence Interval for Variances

- To find a confidence interval for the variance, we need the following statistic.

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2$$

- Warning :  $W$  is on the denominator, so for upper one-sided interval, we take the lower percentile  $\alpha$  and  $1 - \alpha$  for lower one-sided interval.

- $\left( 0, \frac{(n-1)S^2}{w_\alpha} \right)$
- $\left( \frac{(n-1)S^2}{w_{1-\alpha}}, \infty \right)$
- $\left( \frac{(n-1)S^2}{w_{1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{w_{\frac{\alpha}{2}}} \right)$

- The **Fisher** distribution is define as

$$F(r_1, r_2) = \frac{W_1/r_1}{W_2/r_2}$$

where  $r_1$  and  $r_2$  are the degree of freedom.

- If  $T \sim$  Student, then  $T^2 \sim$  Fisher.

- To find a confidence interval for variance ratio, we need the following statistic.

$$F(n_x-1, n_y-1) = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$$

## Lesson 39 : Uniformly Most Powerful critical Regions

- The **Neyman-Pearson lemma** states that for tests of one *simple* hypothesis against another, the best critical region for any ( $\alpha$ ) is to select all that minimize the likelihood ratio.

$$h(x) = \frac{L(x_1, \dots, x_n; \theta | H_0)}{L(x_1, \dots, x_n; \theta | H_1)} < c$$

- If  $h(x)$  is increasing,  $F(k|H_0) < \alpha$ .
- If  $h(x)$  is decreasing,  $S(k|H_0) < \alpha$ .

- If the alternative hypothesis is *composite*, then we can find the **uniformly most powerful critical region** with the same likelihood ratio. This region only exist for one-sided test.

## Lesson 40 : Likelihood Ratio Tests

- This test is usefull when there is no uniformly most powerful critical region.

$$h(x) = \frac{g(x_1, \dots, x_n; \theta | H_0)}{g(x_1, \dots, x_n; \theta | H_1)}$$

where  $g(x_1, \dots, x_n; \theta)$  is the maximum likelihood.

- For large sample, we can use the asymptotic distribution of the likelihood.

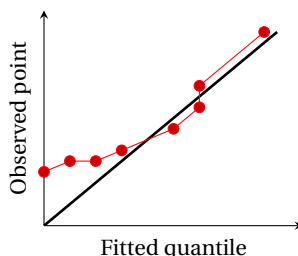
$$-2[l(\theta|H_0) - l(\theta|H_1)] \sim \chi_{(k-l)}^2$$

where  $k$  is the number of parameters specifies by  $H_0$  and  $l$  is the combinaison of numbers of parameters specifies by  $H_0$  and  $H_1$ .

- The last test can also be use to decide if it worth to add parameter to a distribution fit.

## Lesson 41 : q-q Plots

- This plot compare quantile of two distribution. It consiste of a plot of coordinate pairs :  $(x_i, F^{-1}(p_i))$  where  $p_i$  is the **empirical percentile** of  $x_i$ . Then the fit is good if the point are close to a 45° line.



## Lesson 42 : Introduction to Extended Linear Models

There are two purposes in building a extended linear model.

- Prediction :** We want to predic the valu of the *response* variable given specific values of the *explanatory* variables.
- Inference :** We want to understand which *explanatory* variables explain the *response* variable and how much their explain it.

To evaluate the accuracy of a model, we estimate it mean square error.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## Lesson 43 : How a Generalized Linear Model Works

- Linear Model :**

$$Y = \eta + \varepsilon = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

where

$$\varepsilon \sim N(0, \sigma^2)$$

$$Y \sim N(\eta, \sigma^2)$$

**Hypothesis :**

$$(H_1) \quad E[\varepsilon] = 0 \quad (\text{Linearity})$$

$$(H_2) \quad \text{Var}(\varepsilon) = \sigma^2 \quad (\text{Homoscedasticity})$$

$$(H_3) \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (\text{Independence})$$

- The **Box-Cox transformation** is a general set of transformation. When the variance of the error terms is not constant ( $H_2$ ), we need to transforme  $Y$ .

$$Y^* = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln Y & \lambda = 0 \end{cases}$$

where  $\lambda$  is chosen to best stabilize the variance of the error terms.

- The *feature* must be linearly independent. That mean their can't be a function of another. Ex :  $X_3 = 1 - X_2$ .

- We need to encode categorials variables with  $k$  levels into  $(k - 1)$  indicators variables (called *dummy* variables) to avoid *feature* to be dependent. For interaction with 2 categorials variables,  $(k - 1)(l - 1)$  dummy variables are needed.

- GLM :**

$$g(E[Y]) = \beta_0 + \sum_{i=1}^n \beta_i x_i$$

where  $g(\cdot)$  is the link function.

- Exponential Family :**

$$f(y; \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}$$

with

$$E[a(y)] = -\frac{c'(\theta)}{b'(\theta)}$$

$$\text{Var}(a(y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

- Tweedie** distribution :

$$\text{Var}(Y) = aE[Y]^p$$

- link function :** The GLM estimate is unbiased when the canonical link is used.

Distribution	Canonical link
Normal	$g(y) = y$
Binomial	$g(y) = \ln \frac{y}{1-y}$
Poisson	$g(y) = \ln y$
Gamma	$g(y) = \frac{1}{y}$

- Offset :** We add  $\ln n_i$  for cell with  $n_i$  exposure.

- Rate ratio :**

$$RR = \frac{E[Y_i | x_j = 1]}{E[Y_i | x_j = 0]}$$

## Lesson 44 : Categorical Response

### Binomial Response

- Let  $\pi_i \in (0, 1)$  be the response variable. We then need to have link that map  $\eta$  into  $(0, 1)$ .

- logit** :  $\ln\left(\frac{\pi}{1-\pi}\right) = \eta$
- Probit** :  $\Phi^{-1}(\pi) = \eta$
- Log-log** :  $\ln(-\ln(1-\pi)) = \eta$

- Odds Ratio** :  $o = \frac{\pi}{1-\pi}$

### Nominal Response

- Suppose the response can be  $J$  values. Then we create a model of relative odds.

$$\ln \frac{\pi_j}{\pi_1} = \eta_j \Leftrightarrow \pi_j = \pi_1 e^{\eta_j}$$

- $\pi_i = \frac{1}{1 + \sum_{j=2}^J e^{\eta_j}}$
- $\pi_j = \frac{e^{\eta_j}}{1 + \sum_{j=2}^J e^{\eta_j}}$

- If  $x_i$  is a binary feature, then the odds ratio of this **variable** in the category  $j$  to the base categorie is  $e^{\beta_{ij}}$ .

### Ordinal Response

Ordinal response variables have several categories in logical order.

- Cumulative logit and proportional odds models** :

$$\ln o_j = \ln \frac{\sum_{k=1}^j \pi_k}{1 - \sum_{k=1}^j \pi_k} = \eta_j$$

Tips : The model is cumulative, so to find  $\pi_2$ , we need to find  $\pi_1$  and  $\pi_1 + \pi_2$ .

This model is proportional so if we **fix** the categorie but consider two set of feature  $x_{i1}$  and  $x_{i2}$ , the relative odds are

$$\frac{(o_j | x_i = x_{i1})}{(o_j | x_i = x_{i2})} = e^{\sum \beta_i (x_{i1} - x_{i2})}$$

- Adjacent categorie logit model** :

$$\ln \frac{\pi_j}{\pi_{j+1}} = \eta_j$$

$$\sum_{j=1}^J \pi_j = 1$$

- Continuation ratio logit model** :

$$\ln \frac{\pi_j}{\sum_{k=j+1}^J \pi_k} = \ln \frac{\pi_j}{1 - \sum_{k=1}^j \pi_k} = \eta_j$$

Tips : Resolve for  $\pi_1$  then for  $\pi_2$  and so on ...

## Lesson 45 : Estimating Parameters

- Let  $\mathbf{X}$  be the **design matrix**, the  $p \times n$  features matrix.

- Linear Regression** :

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- The **score** function is define as the derivative of the loglikelihood

$$\mathbf{U}(\beta) = \ell'(\beta)$$

- Newton-Raphson** algorithm :

$$\beta^{(k+1)} = \beta^{(k)} - \frac{\mathbf{U}(\beta^{(k)})}{\mathbf{U}'(\beta^{(k)})}$$

- Fisher Scoring** algorithm :

$$\beta^{(k+1)} = \beta^{(k)} - \frac{\mathbf{U}(\beta^{(k)})}{\mathbf{E}[\mathbf{U}'(\beta^{(k)})]}$$

- The score vector has components

$$U_j = \sum_{i=1}^n \frac{y_i - \mu_i}{\text{Var}(y_i)} x_{ij} \left( \frac{d g(\mu_i)}{d \mu_i} \right)$$

- The information matrix :  $I(\theta) = \mathbf{X}^T \mathbf{W} \mathbf{X}$

- Let  $\mathbf{W}$  be the diagonal matrix with entries

$$w_{ii} = \left( \left( \frac{d g(\mu_i)}{d \mu_i} \right)^2 \text{Var}(y_i) \right)^{-1}$$

- Let  $\mathbf{G}$  be the diagonal matrix with entries

$$G_{ii} = \frac{g(\mu_i)}{\mu_i}$$

- The regression variable for one iteration

$$\mathbf{z}^{(k-1)} = \mathbf{X} \mathbf{b}^{(k-1)} + \mathbf{G}^{(k-1)} (\mathbf{y} - \mu^{(k-1)})$$

- The **Weighted Least Square** :

$$\mathbf{b}^{(k)} = (\mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{z}^{(k-1)}$$

## Lesson 46 : Measures of Fit

- The **saturated** model is when we have as much feature as parameters ( $p = n$ ).  $g^{-1}(\mathbf{X}^T \mathbf{b}) = \mathbf{y}$

- The **deviance** statistic test compare a model to the saturated model.

$$D = 2[\ell(\mathbf{b}_{max}) - \ell(\mathbf{b})] \approx n - p'$$

where  $p' = p + 1$  and  $p$  the number of feature.

- Binomial :

$$D = 2 \sum_{i=1}^n \left( y_i \ln \frac{y_i}{\hat{y}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{y}_i} \right)$$

- Normal (*scaled deviance*) :

$$\sigma^2 D = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Poisson :

$$D = 2 \sum_{i=1}^n \left( y_i \ln \frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i) \right)$$

- Gamma :

$$D = 2\alpha \sum_{i=1}^n \left( -\ln \frac{y_i}{\hat{y}_i} + \frac{y_i - \hat{y}_i}{\hat{y}_i} \right)$$

## Significance of Feature

- Loglikelihood ratio test** : These tests compare a **unconstrained** modele with  $p + q$  parameters versus another **constrained** model with  $p$  parameters.

$$2(\bar{\ell}_{p+q} - \bar{\ell}_p) \sim \chi_{(q)}^2$$

$$\hat{D} - \bar{D} \sim \chi_{(1)}^2$$

- Wald test** : To test wheter a single parameter  $\beta_j = r$ .

$$W = \frac{(\hat{\beta}_j - r)^2}{\text{Var}(\hat{\beta}_j)} \sim \chi_{(1)}^2$$

$\sqrt{W} \sim N(0, 1)$ , is usefull for confidence interval.

$I(\theta)^{-1} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$  is the covariance matrix.

- Score test** :  $\mathbf{U}^T I(\theta)^{-1} \mathbf{U} \sim \chi_{(q)}^2$

If  $q = 1$ ,  $\frac{U}{\sqrt{I(\theta)}} \sim N(0, 1)$ .

- We want the lowest AIC and BIC.

## Lesson 47 : Standard Error, $R^2$ , and Student Statistic

$$SST = SSE + SSR$$

- Total sum of square** :  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

- Error sum of square** :  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

$$SSE = \varepsilon^T \varepsilon = \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{x}^T \mathbf{y}$$

- Regression sum of square** :  $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

ANOVA			
SS	df	MS	F
SSR	p	MSR = SSR/df	$\frac{MSR}{MSE}$
SSE	n-p'	MSE = SSE/df	
SST	n-1	MST = SST/df	

- The standort error of the regression is

$$s = \sqrt{MSE}$$

- The **coefficient of determination** is the proportion explain by the regression.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Student test** : To test  $\beta_i = \beta^*$

$$t_{n-p'} = \frac{\hat{\beta}_i - \beta^*}{s_{\beta_i}}$$

Matrice variance-covariance :  $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$

- Simple linear regression :

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{s_{xx}}$$



## Lesson 48 : Fisher Statistic and VIF

- The **Fisher** statistic test the significance of the entire regression, in other word if all  $\beta_i = 0$ . For simple linear regression  $F = T^2$ . Tips : Divide numerator and denominator of  $F$  by SST to find  $R^2$ .

- For simple linear regression, since  $p = 1$ , then  $T(n) = \sqrt{F_{1,n}}$ .

- **Partial Fisher test** : To test is  $q$  added variables have significance.

$$F_{\Delta df, n-p'} = \frac{SSE^{(0)} - SSE^{(1)} / \Delta df}{SSE^{(1)} / (n-p')}$$

- The **Variance Inflation Factor** test the collinearity of the features. To measure it, we take the  $x_j$  feature and take it as the response. Let  $R_{(j)}^2$  be the  $R^2$  of this regression.

$$VIF = \frac{1}{1 - R_{(j)}^2}$$

We want the lowest VIF.

- **Coefficient of correlation** :

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- For two-feature model  $R_{(y)}^2 = r^2$ .

## Lesson 49 : Validation

- The **Hat matrix** put a hat on  $y$  since  $\hat{y} = H y$ .  
 $H = X(X^T X)^{-1} X^T$

- It follow that  $\text{Var}(\hat{e}) = (I - H) \sigma^2$

- For simple linear regression :

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

- The **studentized residuals** are define as

$$r_i = \frac{\hat{e}_i}{\sqrt{S^2(1 - h_{ii})}}$$

where  $h_{ii}$  is the **leverage**. Average leverage should be at  $\frac{p'}{n}$ .  $\sum h_{ii} = p'$

- A **influence point** is a observation that influence a lot  $y$ . A **outliers** is a observation that have  $|r_i| > 3$ .

- Two measure for influence point.

$$- \text{DFITS}_i = r_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

$$- \text{Cook} : D_i = \frac{\sum (\hat{y}_j - y_{j(i)})^2}{p' S^2} = r_i^2 \frac{h_{ii}}{p'(1 - h_{ii})}$$

$D_i > 1$  is too high.

## Lesson 50 : Prediction

- A **confidence interval** for predicted values.

$$y^* \in \hat{y}^* \pm t_{(n-2)} \sqrt{S^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

- A **prediction interval** for predicted values.

$$y^* \in \hat{y}^* \pm t_{(n-2)} \sqrt{S^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

## Lesson 51 : ANOVA

### One-factor ANOVA

$$SST = SSE + SSTR$$

Model	Sum of square	Deviance
$Y = \mu + \varepsilon_{ij}$	SST	$D_M$
$Y = \mu_i + \varepsilon_{ij}$	SSE	$D_A$

- **Within sum of square**

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

- **Between sum of square**

$$SSTR = \sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^k \left( \frac{y_{i\cdot}^2}{n_i} \right) - n \bar{y}_{\cdot\cdot}^2$$

- **Total sum of square**

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{j=1}^{n_i} (y_{ij}^2) - n \bar{y}_{\cdot\cdot}^2$$

- **Fishier test**

$$F_{(k-1, n-k)} = \frac{SSTR / (k-1)}{SSE / (n-k)} = \frac{(D_M - D_A) / (k-1)}{D_A / (n-k)}$$

where  $D_M$  is the *scale deviance* of the minimal model.

### Two-factor ANOVA without replication

$$SST = SSE + SSTR + SSB$$

Model	Sum of square (DF)
$Y = \mu + \varepsilon_{ij}$	$SST(bk - 1)$
$Y = \mu + \alpha_i + \varepsilon_{ij}$	$SSTR(k - 1)$
$Y = \mu + \beta_j + \varepsilon_{ij}$	$SSB(b - 1)$
$Y = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$	$SSE(k - 1)(b - 1)$

- The formula are the same but  $n_i$  is  $k$  for SSTR and  $b$  for SSB.

### Two-factor ANOVA with replication

- To test interaction :

$$F_{(I-1)(J-1), IJ(K-1)} = \frac{(D_I - D_S) / (I-1)(J-1)}{D_S / IJ(k-1)}$$

- To test factor A :

$$F_{(I-1), IJ(K-1)} = \frac{(D_B - D_I) / (I-1)}{D_S / IJ(k-1)}$$

- To test factor B :

$$F_{(J-1), IJ(K-1)} = \frac{(D_M - D_B) / (J-1)}{D_S / IJ(k-1)}$$

where  $D_S$  is the saturated model, I for additive model.

- ANCOVA

## Lesson 52 : Measures of Fit II

For **contingencies table** with binomial or poisson distribution.

- Pearson :  $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-p')}$

- Likelihood ratio chi-square :

$$C = 2[\ell - \ell_{\min}] \sim \chi^2_{(p'-1)}$$

- **Pseudo  $R^2$**  :  $\text{pseudo} R^2 = \frac{\ell_{\min} - \ell}{\ell_{\min}}$

### Residus

- Pearson residual :  $X_k = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{\mu}_i)}}$

- Deviance residual :  $d_k = s_k \sqrt{\text{deviance}}$   
where  $s_k$  is the sign of  $y_k - \hat{y}_i$

- To standartize them, divide by  $\sqrt{1 - h_{ii}}$

## Lesson 53 : Resampling Methods

- **Cross-Validation** :

$$CV(K) = \frac{1}{k} \sum_{i=1}^k MSE_i$$

As  $k \rightarrow n$  bias  $\downarrow$   $\text{Var}(CV_k) \uparrow$

If  $k = n$  then is the LOOCV statistic.

- LOOCV for least-square regression :

$$CV(K) = \frac{1}{n} \sum_{i=1}^n \left( \frac{\varepsilon_i}{1 - h_{ii}} \right)^2$$

- **Bootstap** :

$$SE_B(\alpha) = \sqrt{\frac{1}{1-B} \sum_{i=1}^B (\hat{\alpha} - \bar{\alpha})^2}$$

## Lesson 54 : Subset Selection

Using a lot of feature will result of lower standard error on training data, but poor prediction. We need to keep only the feature that truly impact the response.

- **Subset selection** For  $k$  possible feature,  $2^k$  different model are possible.

- For 2 model with same number of feature, we take the one with lowest SSE.
- Otherwise, we compare with : Mallows'  $C_p$ , AIC, BIC and adjusted  $R^2$ .

- **Forward stepwise selection** consist of starting with the null, then fit  $k$  models with one variable and select the best base on SSE, then fit  $k - 1$  variables and so on. We obtain  $k + 1$  model, the best for each number of predictor, and select the final one with cross-validation or the 4 statistics. For categorial variables, each categorie is added independently.

- **Total fitted model** :

$$- \text{Forward} : 1 + \sum_{i=0}^{\min(p, n)} (\min(p, n) - i)$$

$$- \text{Backward} : 1 + \sum_{i=1}^{\min(p, n)} i$$

### Choosing the best model

- **Cross-validation** is the more accurate.

- **Mallows'  $C_p$**  :  $C_p = \frac{1}{n} (SSE + 2p\hat{\sigma}^2)$

IF  $\hat{\sigma}^2$  is unbiased then  $C_p$  is unbiased.

- **Adjusted  $R^2$**  :  $R_a^2 = 1 - \frac{MSE}{MST}$

- We want the lowest Mallows'  $C_p$ , AIC, BIC and the heighest  $R_a^2$ .

## Lesson 55 : Shrinkage and Dimension Reduction

> **Ridge Regression** : Minimize

$$\left( \sum_{i=1}^n y_i - \beta_0 - \sum_{j=1}^{p'-1} \beta_j x_{ij} \right) + \lambda \sum_{j=1}^{p'-1} \beta_j^2$$

> **Lasso Regression** : Minimize

$$\left( \sum_{i=1}^n y_i - \beta_0 - \sum_{j=1}^{p'-1} \beta_j x_{ij} \right) + \lambda \sum_{j=1}^{p'-1} |\beta_j|$$

> **Standard Predictors**  $\tilde{x} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum (x_{ij} - \bar{x})^2}}$

$$\lambda \rightarrow \infty \Leftrightarrow \beta_j \rightarrow 0$$

$$\lambda \rightarrow 0 \Leftrightarrow \beta_j \rightarrow \hat{\beta}_j^{\text{normal}}$$

PCA	Partial Least Square
unsupervised	supervised
variables are linear combination of the original	
Higher bias	Lower bias
Lower variance	Higher variance

## Lesson 56 : Extension to the Linear Model

> **Extension** : These type can be treated as same as GLM.  $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \varepsilon_i$

For **polynomial regression**,  $b_j(x) = x^j$ .

For **piecewise constant regression**

$$b_j(x) = \mathbb{1}_{\{a \leq x < b\}}(x)$$

> **Generalized Additive Model** :

$$y_i = \beta_0 + \sum_{j=1}^p f_j(x_{ij}) + \varepsilon_i$$

- Allows nonlinear fits for each explanatory variable.
- Effect on each explanatory is separate, so easily identifiable.
- Does not allow for effect of interaction among variable.

## Lesson 57 : Trend and Seasonality

> **Trend** measures the amount by which the series increase from period to period.

> **Seasonal** variation measure cycle within a year.

> **Decomposition models**

- Additive Model :  $x_t = m_t + s_t + z_t$
- Multiplicative Seasonality :  $x_t s_t + z_t$
- Multiplicative Model :  $x_t = m_t s_t z_t$

> **Centered moving average** :

$$\hat{m}_t = \frac{0.5m_{t-k} + m_{t-k+1} + \dots + m_t + \dots + m_{t+k-1} + 0.5m_{t+k}}{2k}$$

> **Seasonal variation factor** :

- Additive Seasonality :  $\hat{s}_t = x_t - \hat{m}_t$   
Adjusted so that  $\sum (s_t + c) = 0$ .
- Multiplicative Seasonality  $\hat{s}_t = \frac{x_t}{\hat{m}_t}$   
Adjusted so that  $\sum \frac{(\hat{s}_t + c)}{n} = 1$ .

## Lesson 58 : correlation

> if  $\mu(t)$  and  $\sigma^2(t)$  does not vary with  $t$  then the time series is **second order stationary**

> **Variance** :  $\sigma^2(t) = E[(x_t - \mu(t))^2]$

**Stationary Time series**

> **sample variance**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_t - \bar{x})^2$

> **Covariance** at lag  $k$  :

$$\text{Cov}(x_t, x_{t+k}) = \gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)] \quad (\text{acvf})$$

$$c_k = \frac{1}{n} \sum_{i=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \quad (\text{sample acvf})$$

> **Auto-correlation**

$$\rho_k = \frac{\text{Cov}(x_t, x_{t+k})}{\sigma^2} \quad (\text{acf})$$

$$r_k = \frac{c_k}{c_0} \quad (\text{sample acf})$$

**Relationships of different time series**

> A **leading variable** is one that impacts another.

> **Cross-covariance**

$$\gamma_k(x, y) = E[(x_{t+k} - \mu_x)(y_t - \mu_y)] \quad (\text{ccvf})$$

$$c_k = \frac{1}{n} \sum_{i=1}^{n-k} (x_{t+k} - \bar{x})(y_t - \bar{y}) \quad (\text{sample ccvf})$$

> **Cross-correlation**

$$\rho_k(x, y) = \frac{\gamma_k(x, y)}{\sigma_x \sigma_y} \quad (\text{ccf})$$

$$r_k = \frac{c_k(x, y)}{\sqrt{c_0(x, x) c_0(y, y)}} \quad (\text{sample ccf})$$

> Notice

$$\gamma_k(x, y) = \gamma_{-k}(y, x)$$

$$\rho_k(x, y) = \rho_{-k}(y, x)$$

$$c_0(x, x) = c_0$$

## Lesson 59 : White Noise and Random Walks

> **White noise** each term are independent and variance  $\sigma^2$ . The correlogram has autocorrelations all close to 0 except for  $r_0$ .

$$w \sim N(0, \sigma^2)$$

> A **Random Walks** is a nonstationary time series which is the accumulation of white noise. The correlogram will decrease slowly from 1 to 0.

$$x_1 = w_1$$

$$x_t = x_{t-1} + w_t$$

with

$$\mu(t) = 0$$

$$\sigma^2(t) = t\sigma^2_w$$

$$\gamma_k(t) = t\sigma^2_w$$

$$\rho_k(t) = \frac{1}{\sqrt{1 + \frac{k}{t}}}$$

> A **Walk with drift** drift the mean  $\mu(t) = t\delta$  by don't affect variance and autocorrelations.

$$x_t = x_{t-1} + \delta + w_t$$

## Lesson 60 : Autoregressive Models

> An **autoregressive** model of order ( $p$ ), or AR( $p$ ) is a time series where term may be expressed in term of previous terms plus white noise.

$$x_t - \mu = \alpha_1(x_{t-1} - \mu) + \alpha_2(x_{t-2} - \mu) + \dots + \alpha_p(x_{t-p} - \mu) + w_t$$

> An AR(1) process is stationary if  $|a| < 1$ . correlogram is decreasing exponentially. For a stationary AR(1) process

$$\mu_k = 0$$

$$\gamma_k = \frac{\alpha^k \sigma_w^2}{1 - \alpha^2}$$

$$\rho = \alpha^k$$

> Notation :  $\mathbf{B}^k x_t = x_{t-k}$

$$w_t = x_t - \alpha_1 x_{t-1} - \alpha_2 x_{t-2}$$

$$= (\alpha_2 \mathbf{B}^2 - \alpha_1 \mathbf{B} + 1)x_t$$

$$= \theta_p(\mathbf{B})x_t$$

where  $\theta_p(\mathbf{B})$  is the **characteristic equation**.

> Testing stationarity : Root  $< 1$

$$(\text{given}) \quad x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + w_t$$

$$(\text{solve}) \quad \theta_p(\mathbf{B}) = 0$$

(answer) If  $|\mathbf{B}| > 1$ , the process is stationary.

> tips : For 2 param :

$$\alpha_2 - \alpha_1 < 1$$

$$\alpha_2 + \alpha_1 < 1$$

$$|\alpha_2| < 1$$

> Forecast  $\hat{x}_{n+1|n}$  is the same equation omitting  $w_t$ .

## Lesson 61 : Regression

> Variance of sample mean with correlation is given by

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n} \left( 1 + 2 \sum_{k=1}^{n-1} \left( 1 - \frac{k}{n} \right) \rho_k \right)$$

> **Harmonic Seasonal model**

$$x_t = m_t + \sum_{i=1}^{\lfloor S/2 \rfloor} s_i \sin(2\pi i t / s) + c_i \cos(2\pi i t / s) + z_t$$

> Forecast correction

- Lognormal :  $e^{\sigma^2/2}$

- Empirical :  $\frac{\sum e^{z_t}}{n}$

## Lesson 62 : Moving Average Models

- > A **moving average** time serie (MA(q)) is always stationary. It define as

$$x_t = \mu + w_t + \beta_1 w_{t-1} + \dots + \beta_q w_{t-q}$$

$$= \mu + \phi(\mathbf{B}) w_t$$

with

$$\mu(t) = 0$$

$$\gamma_k = \sigma_w^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \quad \beta_0 = 1$$

and  $\gamma_k = 0$  for  $k > q$  so MA(q) may be good fit is we observe  $\gamma_q = 0$  in correlogram.

- >  $q$  beta +  $\mu + \sigma_w^2 = q + 2$  parameters fits.
- > A MA(q) is **Inversible** if all the root of  $\phi(\mathbf{B})$  are  $|\mathbf{B}| > 1$
- > Express MA(q) in form of AR( $\infty$ ). If  $\phi(\mathbf{B})$  is reversible :  

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \dots$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + x^4 + \dots$$
- > ARIMA with  $p = d = 0$  is a MA(q) model.
- > conditional sum of squared residuals :  $\sum w_t^2$

## Lesson 63 : ARMA Models

- > The ARMA(p,q) models :  

$$x_t = \alpha_1 x_{t-1} + \dots + \alpha_p x_{t-p} + \beta_1 w_{t-1} + \dots + \beta_q w_{t-q} + w_t$$

$$\theta_p(\mathbf{B})x_t = \phi_q(\mathbf{B})w_t$$
- > The process is stationary if all roots of  $\theta(x) > 1$  and the process is invertible if all roots of  $\phi(x) > 1$

$$\gamma_0 = \sigma_w^2 \left( \frac{1 + 2\alpha\beta + \beta^2}{1 - \alpha^2} \right)$$

$$\gamma_k = \sigma_w^2 (\alpha + \beta) \alpha^{k-1} \left( \frac{1 + \alpha\beta}{1 - \alpha^2} \right)$$

$$\rho_k = \alpha \rho_{k-1} \text{ for } k \geq 2.$$

- > If the process is stationary,  $E[x_t] = E[x_{t-1}]$ .

## Lesson 64 : ARIMA and SARIMA models

- >  $\nabla x_t + x_t - x_{t-1} = (1 - \mathbf{B})x_t$
- > An **ARIMA** model is a nonstationary process. If  $x_t$  is an ARIMA model, then  $y_t = \nabla^d x_t$  is an ARMA(p,q). Then the ARIMA(p,d,q) is  

$$\theta(\mathbf{B})(1 - \mathbf{B})^d x_t = \phi(\mathbf{B})w_t$$
  - With no MA(q), this is ARI(p,d)
  - With no AR(p), this is IMA(d,q)
- > An **SARIMA** model is a ARIMA with seasonal effect.
- > To forecast, we take the difference and then forecast ARMA(p,q) model.

## Appendix

### Inverting a matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Ajouter pour une matrice 3x3

### Synthetic Division

**Exemple :** Factorize  $x^3 - 12x^2 - 81$

$$\begin{array}{r|rrrr} & 1 & -12 & 0 & -81 \\ 3 & & 3 & -27 & -81 \\ \hline & 1 & -9 & -27 & 0 \end{array}$$

then,  $x^3 - 12x^2 - 81 = (x - 3)(x^2 - 9x - 27)$

### Deductible and Limite

$$X = \min(X; d) + \max(0; X - d)$$

$$E[X] = E[\min(X; d)] + E[\max(0; X - d)]$$

$$= E[(X \wedge d)] + E[(x - d)_+]$$

$$= E[(X \wedge d)] + e_x(d) \cdot S_x(d)$$

### Statistic Order

$$Y_1 = \min(X_1, \dots, X_n)$$

$$f_{Y_1}(y) = n f(y) [S(y)]^{n-1}$$

$$S_{Y_1}(y) = \prod_{i=1}^n \Pr(X_i > y)$$

$$Y_n = \max(X_1, \dots, X_n)$$

$$f_{Y_n}(y) = n f(y) [F(y)]^{n-1}$$

$$F_{Y_n}(y) = \prod_{i=1}^n \Pr(X_i \leq y)$$

$$Y_k \in (Y_1, \dots, Y_k, \dots, Y_n)$$

$$f_{Y_k}(y) = \frac{n! \cdot f(y) [F(y)]^{k-1} [S(y)]^{n-k}}{(k-1)!(n-k)!}$$

$$F_{Y_k}(y) = \Pr(\text{at least } k \text{ of } n X_i \text{ are } \leq y)$$

$$= \sum_{i=k}^n \binom{n}{i} [F(y)]^i [S(y)]^{n-i}$$

- >  $x + y = \min(x, y) + \max(x, y)$ , since one is for sure the max and the other the min.

### Mode : Most likely probability

- >  $g(x) = f(x)$  or some time  $g(x) = \ln f(x)$
- > **Mode** is the  $x$  that respects :  $g'(x) = 0$

### Normal Approximation

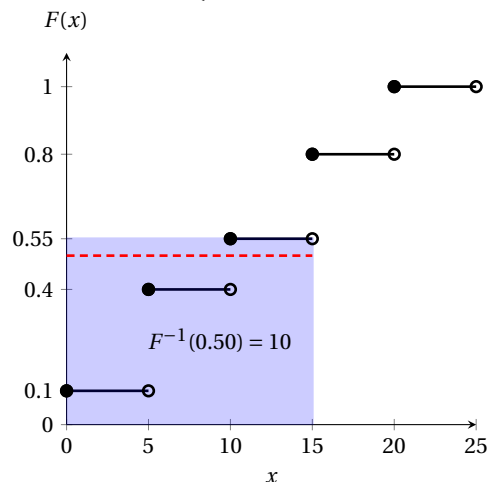
$$F_X(x) = \Phi\left(\frac{x - E[X]}{\sqrt{\text{Var}(X)}}\right)$$

- > **Continuity correction** is necessary when  $X$  is discrete.  $F_X(x) = \Phi\left(\frac{(X \pm k) - E[X]}{\sqrt{\text{Var}(X)}}\right)$  where  $k$  is the mid-point of the discrete value.

### Discrete Cumulative Function

$$\Pr(X = x) = \begin{cases} 0.10, & x = 0 \\ 0.30, & x = 5 \\ 0.15, & x = 10 \\ 0.25, & x = 15 \\ 0.20, & x = 20 \end{cases}$$

$$\Pr(X \leq x) = \begin{cases} 0.10, & 0 \leq x < 5 \\ 0.40, & 5 \leq x < 10 \\ 0.55, & 10 \leq x < 15 \\ 0.80, & 15 \leq x < 20 \\ 1, & x \geq 20 \end{cases}$$



### Contract

- > **Deductible(d)**
- > **Maximum(u)**
- > **Inflation(r)**
- > **Coinurance( $\alpha$ )**

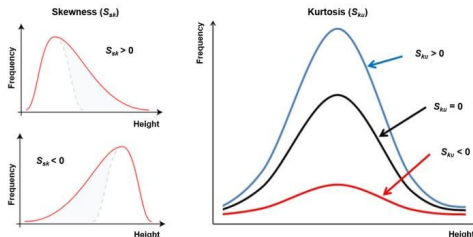
$$Y = \begin{cases} 0 & x \leq \frac{d}{1+r} \\ \alpha[(1+r)x - d] & \frac{d}{1+r} < x < \frac{u}{1+r} \\ \alpha[u - d] & x \geq \frac{u}{1+r} \end{cases}$$

**Warning :** The maximal don't include the deductible.

## Moments

- ›  $k^e$  moment about the origin.  $\mu'_k = E[X^k]$
- ›  $k^e$  moment about the mean.  $\mu_k = E[(X - \mu)^k]$
- › The **Skewness** moment give information about the asymmetry of the distribution. If  $S_{sk} = 0$ , the distribution is normal.

$$S_{sk} = E\left[\left(\frac{X - \mu}{\sigma^2}\right)^3\right]$$



- › The **kurtosis** moment give information about the flattening of the distribution. If  $S_{ku} = 0$ , the distribution is normal.

$$S_{ku} = E\left[\left(\frac{X - \mu}{\sigma^2}\right)^4\right]$$

- › The **coefficient of variation** give information about the dispersion of the distribution.

$$CV = \frac{\sigma}{E[X]}$$

## Transformations of distribution

- › Lognormal :  $Y = e^X$ , where  
 $Y \sim \text{Lognormal}(\mu, \sigma)$   
 $X \sim \text{Normal}(\mu, \sigma)$
- › Inverse Exponential :  $Y = \frac{1}{X}$ , where  
 $Y \sim \text{Inverse Exponential}(1/\theta)$   
 $X \sim \text{Exponential}(\theta)$
- › Weibull :  $Y = X^{1/\tau}$ , where  
 $Y \sim \text{Weibull}(\tau/\theta)$   
 $X \sim \text{Exponential}(\theta)$

## Parameter interpretation

- › **Scale parameter** ( $\theta, \beta, \sigma$ ) : Affect the spread of the distribution.
- › **Rate parameter** ( $\lambda$ ) : Affect the rate of data at mean. (1/scale)
- › **Shape parameter** ( $\alpha, \tau, \gamma$ ) : Affect the shape rather than simply shift the distribution.

## Produit de convolution

The convolution of 2 random variable is define as the sum of the two.

$$f_{X_1+X_2}(x) = \int_{-\infty}^{\infty} f_{X_1}(x-s)f_{X_2}(s)ds$$

$$F_{X_1+X_2}(x) = \int_{-\infty}^x F_{X_1}(x-s)f_{X_2}(s)ds$$

## Shifting Exponential

$$f(x; \theta; d) = \frac{1}{\theta} e^{-(x-d)/\theta}$$

$$E[x] = \theta + d$$

$$\text{Var}(x) = \theta^2$$

## Norme d'un vecteur

$$\ell_1 = x_1 + x_2 + \dots + x_n$$

$$\ell_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

## Greedy algorithm

A group of  $n$  peoples is to be assigned to  $k$  job, one to each job. The cost of job  $j$  is  $c_{ij}$  if person  $i$  is assigned. Select the assignment to have the minimal cost. We fixe the job or the people such as the  $\min(n, k)$  and take the **minimum** cost for this personne(job).

$$c_{ij} \sim \text{Exp}(\theta)$$

$$E[\text{Total cost}] = \sum_{i=1}^{\min(n,k)} \frac{\theta}{\max(n,k)}$$

## Potential outliers

Let  $Q_1$  and  $Q_2$  be the 25<sup>e</sup> and the 75<sup>75</sup> quantile. Then  $h = 1.5(Q_3 - Q_1)$ . Observation are potential outliers if their are not in  
 $(Q_1 - h, Q_3 + h)$

## Test III

Type III tests are tests on variable or set of variable that assume that all other variables are present in the model. The **Wald** test is type III.

## Overdispersion

If the variance of the observation is **higher** than indicated by the binomial or poisson model. **High diviance** may indicate Overdispersion. We can solve it with **quasi-likelihood** :  $\phi \text{Var}(y)$ . This don't affect the fit but many statistic like Pearson. For Poisson, we can use a Binomial Negative model.