MAS-1
Study Review

Nicholas Langevin

2 avril 2019

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

Lesson 1 : Probability Review

> **Bernouilli Shortcut :** If a random variable can only assume two values a and b with probability q and $1 - q$, then is 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 :**
$$p_j = \begin{cases} \frac{j}{N} & , r = 1 \\ \frac{r^j - 1}{r^N - 1} & , r \neq 1 \end{cases}$$
où $r = \dfrac{q}{p}$, p : winning prob.

> **Algorithmic efficency :** with N_j = number of steps from j^{th} solution to best solution.
$$E[N_j] = \sum_{i=1}^{j-1} \frac{1}{i}$$
$$Var(N_j) = \sum_{i=1}^{j-1} \left(\frac{1}{i}\right)\left(1 - \frac{1}{i}\right)$$
As $j \to \infty$, $E[N_j] \to \ln j$, $Var(N_j) \to \ln j$

Lesson 5 : Markov Chain Classification

> An **absorbing** state is one that cannot be exited.

> State j is **accessible**$(i \to j)$ 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 state that communicate with each other.

> A Markov chain is **irreductible** 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) si **transcient** 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 irreductible, then it is recurrent.

Lesson 6 : Markov Chains Limiting Probability

> A chain is **positive recurrent** is the expected number of transitions until the state occur is finite, **null recurrent** otherwise. Null recurrent mean 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 world, $P_{ii}^{(1)} > 0$, $\forall i$

> A chain is **ergodic** when the chain is aperiodic and positive irreductible recurrent.

> **Stationary probability :**
$$\pi_j = \sum_{i=1}^{n} P_{ij}\pi_i \qquad \sum_{i=1}^{n} \pi_i = 1$$

> **Limiting probabilities :** if the chain is ergodic, then
$$\mathbf{P}^{(\infty)} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$$

Lesson 7 : Time in Transient States

> Tips : Inverting a matrix

> $\mathbf{S} = (\mathbf{I} - \mathbf{P}_{\text{transient}})^{-1}$, 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.

> $Var(X_n) = Var(Z) \cdot E[Z]^{n-1} \sum_{k=1}^{n} E[Z]^{k-1}$

> If $X_0 \neq 1$ mean and variance of X_n need to be multiplicated by X_0.

> **Probability of extinction :**
$$\pi_0 = \sum_{j=1}^{\infty} p_j \pi_0^j$$
>   – $\mu \leq 1 \Rightarrow \pi_0 \geq 1$,if $X_0 = 1$.
>   – $\mu > 1 \Rightarrow \pi_0 < 1$,if $X_0 = 1$.

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

Lesson 9 : Time Reversible

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

> If $\mathbf{Q} = \mathbf{P}$, then $\mathbf{P}$ 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
$$\min(X_1, X_2, ..., X_n) \sim \text{Exp}\left(\sum_{i=1}^{n} \lambda_i\right)$$

> The sum of 2 Exponentials randoms 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 events before time t.

> Poisson process can't decrease over time. $N(t) \geq N(s)$

> $N(0) = 0$

> Increament are **independent** :
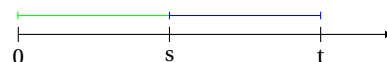
$$\Pr[N(t) - N(s) = n | N(s) = k] = \Pr[N(t) - N(s) = n]$$

> **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 have **stationary increments**.
$$\Pr[N(s)] = \Pr[N(t) - N(s)]$$

Lesson 12 : Poisson Process Time To Next Events

> T_n is the time between the nth event and the $(n-1)^{th}$ event.

> $S_n = \sum_{i=1}^{n} T_i$, is the time for the n^e 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 $\lambda(t) \cdot \alpha_1(t)$.
$$m(t) = \int_0^t \lambda(u)\alpha_1(u)\,\mathrm{d}u$$

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 occured 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 jth order statistics is : $E[T^{(j)}] = \frac{jt}{(k+1)}$.

> 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 : Substraction 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 :
$$F_{X(t)}(t) = \int F_{\{X_u(t)\}}(t)f(u)\,\mathrm{d}u$$

  – 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 :
  – $E[S(t)] = \lambda t E[X]$
  – $\text{Var}(S(t)) = \lambda t E[X^2]$

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

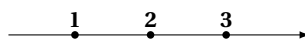> **Sums of compound** homogeneous Poisson process is also a Poisson process with :
  – $N(t) \sim \text{Pois}(\sum \lambda_i)$
  – $F_X(x) = \sum_i w_i F_{X_i(t)}(t), \quad w_i = \frac{\lambda_i}{\sum \lambda_i}$

Lesson 17 : Reliability Structure Functions

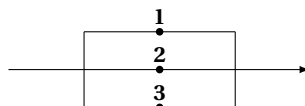> $\phi(\mathbf{x})$ is the **structure** function for a systeme. It equal 1 if the systeme function, 0 otherwise.

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

$$\underset{\bullet}{1} \quad \underset{\bullet}{2} \quad \underset{\bullet}{3}$$

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 systeme is working if at least 1 components is working.

$$\boxed{\begin{array}{c} 1 \\ 2 \\ 3 \end{array}}$$

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 :
$$r(\mathbf{p}) \leq \sum A_i$$
$$r(\mathbf{p}) \geq \sum A_i - \sum A_i \cup A_j$$
$$r(\mathbf{p}) \leq \sum A_i - \sum A_i \cup A_j + \sum A_i \cup A_j \cup A_k$$
where $A_i = \sum p_i$ is the probability of the i^{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^{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 amound of time to failure :
$$E[\textbf{system life}] = \int_0^\infty r(\bar{\mathbf{F}}(\mathbf{t}))\,\mathrm{d}t$$
where,

  – For serie system :
$$r(\bar{\mathbf{F}}(\mathbf{t})) = \prod_{i=1}^n \bar{F}_i(t)$$

  – For parallel system :
$$r(\bar{\mathbf{F}}(\mathbf{t})) = 1 - \prod_{i=1}^n F_i(t)$$

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

> **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-deacreasing function of t.

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

> **Cumulatice hazard function** :
$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\ln \bar{F}(t)$$
with $\frac{H(t)}{t}$ the average of the hazard rate.

Lesson 20 : Survival Models

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

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

> $_{t+u}p_x = {}_u p_x \cdot {}_t p_{x+u}$

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

> 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** :
$$\mu_{x+t} = \frac{f_{T_x}(t)}{{}_t p_x} = -\frac{\mathrm{d}}{\mathrm{d}t}\ln {}_t p_x$$

> **Linear interpolation(D.U.D)** :
$$\ell_{x+t} = (1-t)\,\ell_x + t\,\ell_{x+1}$$
Shortcut : $\forall\, t \in (0,1),\ \forall\, x \in \mathbb{N},\ x < x+t < x+1$:

� ➡ $_t q_x = t \cdot q_x$

➡ $\mu_{x+t} = \dfrac{q_x}{1 - t \cdot q_x}$

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

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

$$\mathring{e}_x = \mathrm{E}[T_x] = \int_0^{\omega - x} {}_t p_x \,\mathrm{d}t \overset{\mathrm{D.U.D}}{=} e_x + 0.5$$

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 :
$$A_{x:\overline{n}|}^1 = \sum_{k=0}^{n} v^{k+1}\, {}_k p_x\, q_{x+k}$$
  – 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
$$\ddot{a}_{\overline{x:\overline{n}|}} = \ddot{a}_{\overline{n}|} + {}_{m|}\ddot{a}_x$$

> **Illustrative Life Table :**
  – $A_x = v^n\, q_x + p_x\, A_{x+1}$
  – $\ddot{a}_x = 1 + v\, p_x\, \ddot{a}_{x+1}$
  – $A_{x:\overline{n}|}^1 = A_x - {}_n E_x\, A_{x+n}$
  – $\ddot{a}_{x:\overline{n}|} = \ddot{a}_x - {}_n E_x\, \ddot{a}_{x+n}$
  – ${}_{m|}A_x = {}_m E_x\, A_{x+m}$
  – ${}_{m|}\ddot{a}_x = {}_m E_x\, \ddot{a}_{x+m}$
  – $\ddot{a}_x = 1 + a_x$
  – $A_x = 1 - d\, \ddot{a}_x$

> **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 $x_0 \equiv$ seed

> **Inverse transformation method :**
$$\Pr\left(F^{-1}(u) \le x\right) = \Pr\left(u \le F(x)\right) = F(x)$$
then $x = F^{-1}(u)$ where $U \sim \mathrm{Unif}(0,1)$
  – 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

> $\Pr(X \le x) \simeq \dfrac{1}{m} \sum\limits_{j=1}^{m} \mathbb{1}_{\{x^{(j)} \le x\}}$

> $\mathrm{E}[X^k] \simeq \dfrac{1}{m} \sum\limits_{j=1}^{m} [x^{(j)}]^k$

> $\mathrm{VaR}_k(X) \simeq X^{[j_0]}$

> $\mathrm{TVaR}_k(X) \simeq \dfrac{1}{m(1-k)} \sum\limits_{j=j_0+1}^{m} X^{(j)} \mathbb{1}_{\{X^{(j)} > X^{[j_0]}\}}$
$$\simeq \frac{1}{m - j_0} \sum_{j=j_0+1}^{m} X^{[j]}$$

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 $f(x) \ne 0$.
$$c = \max \frac{f(x)}{g(x)}$$

Generate two uniform number u_1, u_2. Let $x = G^{-1}(u_i)$. Accept x_1 only if
$$u_2 \le \frac{f(x_1)}{c \cdot g(x_1)}$$

> **Simulating gamma distribution :** Use $\mathrm{Exp}(\alpha \cdot \theta)$ 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 $y_2 = -\ln u_2$. Accept y_1 if
$$y_2 \ge \frac{(y_1 - 1)^2}{2}$$
and add $(-)$ if $u_3 \ge 0.5$

> The **Number of iteration** is a Ross-geometric distribution with mean c. Let be β the mean of a geometric distribution given in the exam appendix :
$$\mathrm{E}[N] = 1 + \beta = c$$
$$\mathrm{Var}(N) = \beta(1 + \beta)$$

Lesson 25 : Estimator Quality

> **Bias :** This quality measures if, on average, the estimator is on the expected value of the parameter.
$$\mathrm{E}[\hat{\theta}] = \theta + \mathrm{bias}_{\hat{\theta}}(\theta)$$
  – If $\mathrm{bias}_{\hat{\theta}}(\theta) = 0$, then $\hat{\theta}$ is **unbiased**.
  – If $\lim\limits_{n \to \infty} \mathrm{bias}_{\hat{\theta}}(\theta) = 0$, then $\hat{\theta}$ is **asymptotically unbiased**.
  – If $\mathrm{bias}_{\hat{\theta}}(\theta) \ne 0$, then $\hat{\theta}$ is **biased**.

> **Consistency :** This quality measures if the probability that the estimator is different from the parameter by more than ε goes to 0 as n goes to infinity.
$$\lim_{n \to \infty} \Pr\left(|\hat{\theta} - \theta| > \varepsilon\right) \to 0,\ \forall \varepsilon > 0$$
In other word, as $n \to \infty$, $\mathrm{E}[\hat{\theta}] \to \theta$, $\mathrm{Var}(\hat{\theta}) \to 0$

> **Efficiency :** This quality measures the variance of the estimator. Lower the variance is, more efficient is the estimator.
$$\text{Efficiency of } \hat{\theta} = \frac{\mathrm{Var}(\hat{\theta})^{\mathrm{rao}}}{\mathrm{Var}(\hat{\theta})}$$
$$\text{Relative efficiency of } \hat{\theta}_1 \text{ to } \hat{\theta}_2 = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}$$
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.
$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] = \left(\mathrm{bias}_{\hat{\theta}}(\theta)\right)^2 + \mathrm{Var}(\hat{\theta})$$

> 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 θ.

> Some estimator :
  – $\bar{x} = \frac{1}{n} \sum x_i$ is a unbiased estimator of the mean μ. $\mathrm{Var}(\bar{x}) = \frac{1}{n} \mathrm{Var}(x)$

- $s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$ is a unbiased estimator of the variance σ^2.

- $\hat{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{n}$ is an asymptotically unbiased of the variance σ^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.

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

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

$$= F_e(x) - F_e(x_{i-1})$$

> **Kernel Density** is a empirical distribution smoothed with a base fonction. Let define the scaling factor b called **bandwith**.

- 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) kermel :**

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

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

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

> **Triangular kernel :**

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

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

> **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 $\hat{\mu}'_2 = \hat{\sigma}^2 + \bar{x}^2 \le 2\bar{x}^2$, 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 :

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

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

> **Smoothed empirical percentile :**
$$\hat{\pi}_p = (1 - h) X^{[j]} + h X^{[j+1]}$$
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 α, $\hat{\theta}^{\text{MLE}} = \hat{\theta}^{\text{MME}}$

- For Normal, $\hat{\mu}^{\text{MLE}} = \bar{x}$ and $(\hat{\sigma}^2)^{\text{MLE}} = \frac{1}{n} \sum (x_i - \hat{\mu})^2$

- 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 β, $(r\beta)^{\text{MLE}} = \bar{x}$

> 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
$$\left(\hat{\theta}^{\text{MLE}}\right)^\tau = \frac{\sum (x_i - d_i)^\tau}{\sum \mathbb{1}_{\{x_i \le u\}}}$$
if $\tau = 1$, then the distribution is Exponential.

> Pareto distribution with fixed θ : $\hat{\alpha} = \frac{n}{K}$
$$K = \sum_{i=1}^{n+c} \ln(\theta + d_i) - \sum_{i=1}^{n+c} \ln(\theta + x_i)$$
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 smalest θ possible, $\hat{\theta}^{\text{MLE}} = \max(x_1, ..., x_n)$

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

 – Grouped : We take the heighest interval(L, U). $\hat{\theta}^{\text{MLE}} = \min(U, \hat{\theta}^{\text{MLE}}_{\text{Censored(L)}})$

› Bernouilli : 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, $\hat{\theta}^{\text{MLE}} \equiv$ mode of this distribution.

Lesson 31 : Variance of MLE

› **Fisher information matrix** :
$$I(\theta) = n\mathrm{E}\left[\left(\frac{\mathrm{d}\ln f(x;\theta)}{\mathrm{d}\theta}\right)^2\right]$$
$$= -n\mathrm{E}\left[\frac{\mathrm{d}^2\ln f(x;\theta)}{\mathrm{d}\theta^2}\right]$$
using the loglikehood function
$$I(\theta) = \mathrm{E}\left[\left(\frac{\mathrm{d}l(x_1,...,x_n;\theta)}{\mathrm{d}\theta}\right)^2\right]$$
$$= -\mathrm{E}\left[\frac{\mathrm{d}^2 l(x_1,...,x_n;\theta)}{\mathrm{d}\theta^2}\right]$$

› **Rao-Cramer lower bound** is the lowest possible variance for a unbiased estimator $\hat{\theta}$ of θ. 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 certains regularity conditions

 – The seconde derivative of the loglikelihood exist.

 – The support of the density function is not function of θ.

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 θ if and only if
$$L(x_1,...,x_n;\theta|Y) = h(x_1,...,x_n)$$
$$L(x_1,...,x_n;\theta) = g(y)h(x_1,...,x_n)$$
where $h(x_1,...,x_n)$ is a function that does not involve θ.

› **Rao-Blackwell** Theorem : For any unbiased estimator $\hat{\theta}$ and sufficient statistic Y, the estimator $\mathrm{E}[\hat{\theta}|Y]$ 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	$1-\alpha$	α
H_1 True	β	$1-\beta$

› The α value is usuly name :
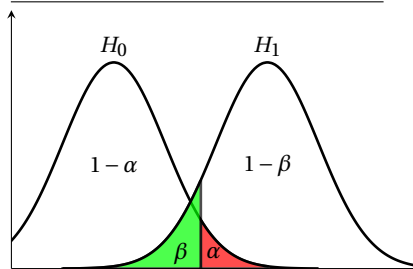
 – Probability of Type I error

 – Size of critical region

 – signifiance level

The β value is usuly name :

 – Probability of Type II error

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

 – The power of test.

› We will reject H_0 in favor of H_1 if a certain condition occured $(X > \gamma)$, 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 then the signifiance level.
$$P_{\text{value}} < \alpha$$

Lesson 34 : Confidence Interval and Sample Size

› Let be c the **confidence coefficient**. Then we can say the we're $100c\%$ confident that the parameter is between (a, b), called the **confidence interval**. $\alpha = 1 - c$
$$\theta \in \hat{\theta} \pm z_{\frac{1+c}{2}}\sqrt{\text{Var}(\hat{\theta})}$$

› We can found the probability that the half-width of the interval is less then 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 (α) 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-sqare** is a one-parameter family distribution. In definition, it a gamma with $\alpha = \frac{n}{2}$ and $\theta = 2$. The only parameter n is called **degree of freedom**.

 – Let $X_i, i = i,...,n$ be normal random variable with mean μ and varianve σ^2.
$$Y = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$$

 – Let $x_i, i = i,...,n, n \geq 2$ be random sample from normal distribution with variance σ^2.
$$W = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

 – Tips : $\chi^2_{(2)} \sim \text{Exp}(\theta = 2)$

› The **strudent** is a one-parameter family distribution. We define it as
$$T_{(n)} = \frac{Z}{\sqrt{W/n}}$$
where $Z \sim \mathrm{N}(0,1)$ and $W \sim \chi^2_{(n)}$.
Note that as $n \to \infty$, $T_{(n)} \to N(0,1)$

› When the variance is unknow, 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.
$$x_1,...,x_n \sim N(\mu_x, \sigma_x^2)$$
$$y_1,...,y_m \sim N(\mu_y, \sigma_y^2)$$
$$T_{(n+m-2)} = \frac{(\bar{x}+\bar{y})-(\mu_x-\mu_y)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$$
where $S^2 = \frac{(n-1)S_x^2+(m-1)S_y^2}{m+n-2}$.

› Testing for mean of bernouilli 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 methode 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
⋮	⋮	⋮	⋮	⋮

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.
$$H_0 : \mu_1 = ... = \mu_k$$

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

Note : This test can be use to test the fit of as parametric model. θ' is the number of parameter fited with the same data as the test.

› **Two-dimensional chi-square :**
$$Q = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(k-1)(c-1)}$$

Lesson 38 : Confience Interval for Variances

› To find a confidence interval for the variance, we need the following statistic.
$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

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

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

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

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

› 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 $T \sim$ Strudent, 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 (α) is to select all that minimize the likelihood ratio.
$$h(x) = \frac{L(x_1,...,x_n;\theta|H_0)}{L(x_1,...,x_n;\theta|H_1)} < c$$

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

   – If $h(x)$ is deacreasing, $S(k|H_0) < \alpha$.

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

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

› For large sample, we can use the asymptotic distribution of the likehood.
$$-2[l(\theta|H_0) - l(\theta|H_1)] \sim \chi^2_{(k-l)}$$

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 dicide 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 : $\left(\mathbf{x_i}, \mathbf{F^{-1}(p_i)} \right)$ where p_i is the empirical percentile of x_i. Then the fit is good if the point are close to a 45° line.
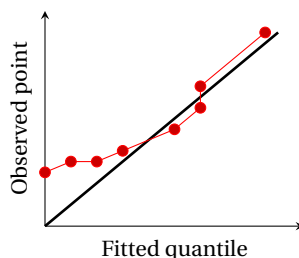
Lesson 42 : Introduction to Extended Linear Models

There are two purposes in building a extended linear model.

1. **Prediction :** We want to predic the valu of the *response* variable given specific values of the *explanatory* variables.

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

Lesson 43 : How a Generalized Linear Model Works

› **Linear Model :**
$$Y = \eta + \varepsilon = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon$$

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

Hypothesis :

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

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

(H_3) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ (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 λ is chossen 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 :**
$$g(\mathrm{E}[Y]) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

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

› **Exponential Family :**
$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}$$

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

› **Tweedie** distribution :
$$\mathrm{Var}(Y) = a\mathrm{E}\big[y\big]^p$$

› **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 :**
$$RR = \frac{\mathrm{E}\big[Y_i | x_j = 1\big]}{\mathrm{E}\big[Y_i | x_j = 0\big]}$$

Lesson 44 : Categorial Response

Binomial Response

> Let $\pi_i \in (0,1)$ be the response variable. We than need to have link that map η 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 :**
$$o_j = \ln\frac{\sum_{k=1}^{j}\pi_k}{1-\sum_{k=1}^{j}\pi_k} = \eta_j$$
Tips : The model is cumulative, so to find π_2, we need to find π_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 π_1 then for π_2 and so on ...

Lesson 45 : Estimating Parameters

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

> **Linear Regression :**
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\hat{x}$$
$$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X}^{-1})\mathbf{X}^\mathsf{T}\mathbf{y}$$

> The **score** function is define as the derivative of the loglikehood
$$\mathbf{U}(\beta) = \ell'(\beta)$$

> **Newton-Raphson** algorithm :
$$\beta^{(k+1)} = \beta^{(k)} - \frac{\mathbf{U}(\beta^{(k)})}{\mathbf{U}'(\beta^{(k)})}$$

> **Fisher Scoring** algorithm :
$$\beta^{(k+1)} = \beta^{(k)} - \frac{\mathbf{U}(\beta^{(k)})}{\mathrm{E}[\mathbf{U}'(\beta^{(k)})]}$$

> The score vector has components
$$U_j = \sum_{i=1}^{n}\frac{y_i-\mu_i}{\mathrm{Var}(y_i)}x_{ij}\left(\frac{\mathrm{dg}(\mu_i)}{\mathrm{d}\mu_i}\right)$$

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

> Let $\mathbf{W}$ be the diagonal matrix with entries
$$w_{ii} = \left(\frac{\mathrm{dg}(\mu_i)}{\mathrm{d}\mu_i}\mathrm{Var}(y_i)\right)^{-1}$$

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

Lesson 46 : Measures of Fit

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

> The **deviance** statistic test compare a model to the satured 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.
$$2(\tilde{\ell}_{p+q} - \hat{\ell}_p) \sim \chi^2_{(q)}$$
$$\hat{D} - \tilde{D} \sim \chi^2_{(1)}$$

> **Wald test :** To test wheter a single parameter $\beta_j = r$.
$$W = \frac{(\hat{\beta}_j - r)^2}{\mathrm{Var}(\hat{\beta}_j)} \sim \chi^2_{(1)}$$

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

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

> **Score text :** $\mathbf{U}^\mathsf{T} I(\theta)^{-1}\mathbf{U} \sim \chi^2_{(q)}$

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 Strudent Statistic

$$SST = SSE + SSR$$

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

> **Error sum of square :** $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$$SSE = \varepsilon^\mathsf{T}\varepsilon = \mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{b}^\mathsf{T}\mathbf{x}^\mathsf{T}\mathbf{y}$$

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

> **Strudent test :** To test $\beta_i = \beta^*$
$$t_{n-p'} = \frac{\beta_i - \beta^*}{S_{\beta_i}}$$

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

> Simple linear regression :

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

  – $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

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

Lesson 48 : Fisher Statistic and VIF

> The **Fisher** statistic test the signifiance od 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 signifiance.
$$F_{\Delta_{df}, n-p'} = \frac{SSE^{(0)} - SSE^{(1)} / \Delta_{df}}{SSE^{(1)} / (n-p')}$$

> The **Variance Inflation Factor** test the collinearity of the features. To mesure it, we take the x_j feature and take it as the response. Let $R^2_{(j)}$ be the R^2 of this regression.
$$\text{VIF} = \frac{1}{1 - r^2_{(j)}}$$

We want the lowest VIF.

> **Coeficient 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 $R^2_{(y)} = r^2$.

Lesson 49 : Validation

> The **Hat matrix** put a hat on y since $\hat{\mathbf{y}} = \mathbf{Hy}$.
$$\mathbf{H} = \mathbf{X(X^\top X)^{-1} x^\top}$$

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

> For simple linear regression :
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

> The **studentized residuals** are define as
$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{S^2(1 - h_{ii})}}$$

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

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

> Two mesure for influence point.

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

  – **Cook :** $D_i = r_i^2 \frac{h_{ii}}{p'(1 - h_{ii})}$

    $D_i > 1$ 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

$$\text{SST} = \text{SSE} + \text{SSTR}$$

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

> **Within sum of square**
$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\Sigma})^2$$

> **Between sum of square**
$$\text{SSTR} = \sum_{i=1}^{k} n_i (\bar{y}_{i\Sigma} - \bar{y}_{\Sigma\Sigma})^2 = \sum_{i=1}^{k} \left(\frac{y_{i\Sigma}^2}{n_i} \right) - n \bar{y}_{\Sigma\Sigma}^2$$

> **Total sum of square**
$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\Sigma\Sigma})^2 = \sum_{j=1}^{n_i} (y_{ij}^2) - n \bar{y}_{\Sigma\Sigma}^2$$

> **Fishier test**
$$F_{(k-1, n-k)} = \frac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)} = \frac{(D_M - D_A)/(k-1)}{D_A/(n-k)}$$
where D_M is the *scale deviance* of the minimal model.

Two-factor ANOVA without replication

$$\text{SST} = \text{SSE} + \text{SSTR} + \text{SSB}$$

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

> 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 satured model, I for additive model.

> ANCOVA

Measures of Fit II

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

> Deviance : $D = 2 \sum O_i \ln \frac{O_i}{E_i}$

> Pearson : $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

> Likelihood ratio chi-square : $C = 2[\ell - \ell_{\min}]$

> **Pseudo** R^2 : 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 signe of $y_k - \hat{y}_i$

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 $x^3 - 12x^2 - 81$

	1	-12	0	-81
3		3	-27	-81
	1	-9	-27	0

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

> $Y_1 = \min(X_1, ..., X_n)$
$$f_{Y_1}(y) = n f(y)[S(y)]^{n-1}$$
$$S_{Y_1}(y) = \prod_{i=1}^{n} \Pr(X_i > x)$$

> $Y_n = \max(X_1, ..., X_n)$
$$f_{Y_n}(y) = n f(y)[F(y)]^{n-1}$$
$$F_{Y_n}(y) = \prod_{i=1}^{n} \Pr(X_i \le x)$$

> $Y_k \in (Y_1, ..., Y_k, ..., Y_n)$
$$f_{Y_k}(y) = \frac{n! \cdot f(y)[F(y)]^{k-1}[S(y)]^{n-k}}{(k-1)!(n-k)!}$$
$$F_{Y_k}(y) = \Pr(\text{at least k of n } X_i \text{ are} \le y)$$
$$= \sum_{i=k}^{n} \binom{n}{i}[F(y)]^i [S(y)]^{n-j}$$

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

Mode : Most likely probability

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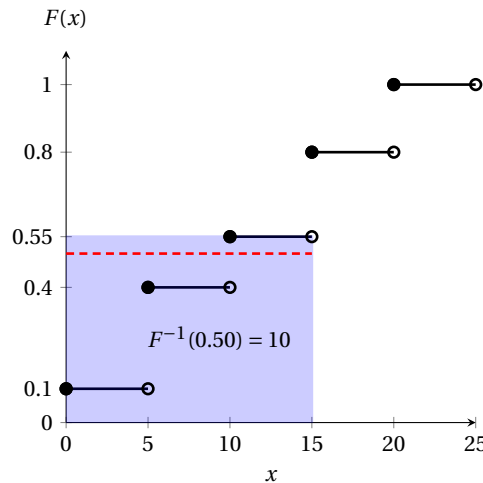> **Mode** is the x that respects : $g'(x) = 0$

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

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

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

Contract

> **Deductible(d)**
> **Maximum(u)**
> **Inflation(r)**
> **Coinsurance(α)**

$$Y = \begin{cases} 0 & x \le \frac{d}{1+r} \\ \alpha[(1+r)x - d] & \frac{d}{1+r} < x < \frac{u}{1+r} \\ \alpha[u - d] & x \ge \frac{u}{1+r} \end{cases}$$
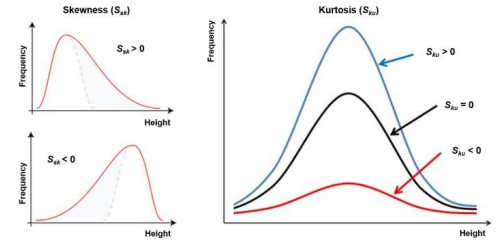
Warning : The maximal don't include the deductible.

Moments

> k^e moment about the origin. $\mu'_k = E\left[X^k\right]$
> k^e moment about the mean. $\mu_k = E\left[(X - \mu)^k\right]$

> The **Skewness** moment give infomation 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 infomation 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.
$$\text{CV} = \frac{\sigma}{E[X]}$$

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
$$Y \sim \text{Weibull}(\sqrt[\tau]{\theta})$$
$$X \sim \text{Exponential}(\theta)$$

Parameter interpretation

> **Scale parameter** (θ, β, σ) **:** Affect the spread of the distribution.
> **Rate parameter** (λ) **:** Affect the rate of data at mean. (1/scale)
> **Shape parameter** (α, τ, γ) **:** Affect the shape rather then simply shift the distribution.

Produit de convolution

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

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