Université Claude Bernard Lyon 1 L2 UE Statistique pour Informatique
Année 2017-2018
Automne 2017

Résumé du cours

Les notions traduites en anglais sont donnés en *italique* les commandes en R en `tt`.

1 Notions de base

Une **population** est une collection d'individus (ou d'objets) avec certaines caractéristiques.

Un **échantillon** *(sample)* est une partie d'une population.

En statistique, une caractéristique des individus est appellée une **variable**. Si I est l'ensemble des individus, on note x_i la caractéristique de l'individu i.

Une variable peut être

— qualitative : par exemple une couleur
— quantitative (numérique) : un nombre (réel)
— discrète : finie ou dénombrable
— continue : si elle prend ses valeurs dans un intervalle de $\mathbb{R}$ (plus une condition de continuité que j'énoncerai plutard)

Un **paramètre** est un nombre inconnu mais fixe (une variable est variable).

2 Description des données (une variable)

L'échantillon est décrit par une seule variable x. La valeur (caractéristique) de l'individu $i \in I$ est notée x_i.

2.1 Notions

Fréquence d'un échantillon (d'une collection d'individus qui ont une seule caractéristique). La fréquence de la caractéristique c de l'échantillon est le nombre d'individu avec caractéristique c,

$$freq(c) = \#\{i \in I : x_i = c\}.$$

Un **mode** est une caractéristique qui a la fréquence maximale. La **distribution de l'échantillon** est le tableau des fréquences de l'échantillon. La **distribution de la population** est le tableau des fréquences de la population.

La **fréquence relative** de la caractéristique c de l'échantillon est la fréquence divisée par la taille de léchantillon.

$$f(c) = \frac{\#\{i \in I : x_i = c\}}{\text{taille de I}}.$$

Si la variable est numérique on définit la **fréquence relative cumulée** (fonction de répartition empirique) de la caractéristique c de l'échantillon par

$$F(c) = \frac{\#\{i \in I : x_i \leq c\}}{\text{taille de I}}.$$

2.2 Représentation graphique

Histogramme `[hist]` d'un échantillon : On partitionne $\mathbb{R}$ en des intervalles, c.à.d. on choisit des nombres $r_0 < r_1 < \cdots < r_N$ donnant les intervalles $[r_{n-1}, r_n]$ *(break points)*. (On choisit r_0 plus petit et r_N plus grand que les valeurs de la variable). On calcule les fréquences relatives des intervalles $[r_{n-1}, r_n[$,

$$f_n = \frac{\#\{i \in I : r_{n-1} \leq x_i < r_n\}}{\text{taille de I}}.$$

et on trace, au dessus de tout intervalle $[r_{n-1}, r_n]$ un rectangle avec surface égale à f_n (et donc avec une hauteur égale à $h_n = \frac{f_n}{r_n - r_{n-1}}$). L'histogramme dépend de la partition choisie (ce qui peut être trompeur).

2.3 Notions pour les variables numériques

On considère un échantillon de taille L avec une variable réelle x.

La **moyenne de l' échantillon** (moyenne empirique) *(mean)* `[mean]` d'une variable x est

$$\overline{x} = \frac{1}{L} \sum_{i=1}^{L} x_i.$$

La **médiane de l'échantillon** *(median)* `[median]` partage les valeurs de x en deux parties : les petites et les grandes valeurs. Pour ceci il faut ordonner les valeurs. Soit x_k^* la k-ième plus petite valeur de l'échantillon (comptée avec multiplicité [1]). La médiane est alors définie comme

$$mediane(x) = \begin{cases} x^*_{\frac{L+1}{2}} & \text{si } L \text{ est impair} \\ \frac{1}{2}(x^*_{\frac{L}{2}} + x^*_{\frac{L}{2}+1}) & \text{si } L \text{ est pair.} \end{cases}$$

Les **quartiles de l'échantillon** partagent les valeurs de x en quatre parties selon leur grandeur : Si L est divisible par 4, alors les trois quartiles Q_1, Q_2, Q_3 sont à peu près $x^*_{\frac{L}{4}}$, $x^*_{\frac{L}{2}}$, $x^*_{\frac{3L}{4}}$. Dans `R` ils sont plus précisement définis comme

$$Q_j(x) = (1-t)x^*_{n+1} + tx^*_n.$$

où n est la partie entière de $\frac{j(L-1)}{4}$ et t la partie fractionaire de $\frac{j(L-1)}{4}$.

La **variance de l'échantillon** *(variance)* `[var]` d'une variable x est

$$V(x) = \frac{1}{L-1} \sum_{i=1}^{L} (x_i - \overline{x})^2.$$

L'écart type de l'échantillon (déviation standard) *(standard deviation)* `[sd]` d'une variable x est

$$s(x) = \sqrt{V(x)}.$$

3 Description des données (deux variables)

On considère maintenant des échantillons qui sont décrits par deux variables x et y, ou, autrement dit, un couple (x, y). Les valeurs (caractéristiques) de l'individu $i \in I$ sont notées (x_i, y_i).

1. par exemple, pour l'échantillon 1 2 4 3 2 on obtient $x_1^* = 1$, $x_2^* = 2$, $x_3^* = 2$, $x_4^* = 3$, $x_5^* = 4$

3.1 Fréquences

Fréquence d'un échantillon (pour un échantillon caractérisé par deux variables). La fréquence de la caractéristique (a, b) de l'échantillon est

$$freq(a, b) = \#\{i \in I : x_i = a, y_i = b\}.$$

La **fréquence relative** de la caractéristique (a, b) de l'échantillon est

$$f(a, b) = \frac{\#\{i \in I : x_i = a, y_i = b\}}{\text{taille de I}}.$$

3.2 Coefficient de corrélation

Soit

$$S_{xy} = \frac{1}{L-1} \sum_{i-1}^{L} (x_i - \overline{x})(y_i - \overline{y}).$$

Donc $S_{yx} = S_{xy}$ et $S_{xx} = V(x)$ est la variance de la variable x.

Le coefficient de corrélation des deux variables x et y est

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}.$$

Si r est proche de 1 ou -1 les données sont proches d'une droite de pente strictement positive ou négative, respectivement. Si r est proche de 0 les données ne sont pas corrélées.

3.3 Régression linéaire

On cherche à déterminer la valeur $a \in \mathbb{R}$ pour la pente et la valeur $b \in \mathbb{R}$ pour l'ordonnée à l'origine d'une droite

$$y = ax + b$$

qui approche le mieux les données (x_i, y_i). Ces valeurs sont données par les formules

$$a = \frac{S_{xy}}{S_{xx}}$$

$$b = \overline{y} - a\overline{x}.$$

4 Probabilité

4.1 Définition empirique

Un **évènement** est le résultat d'une expérience. Un **évènement simple** est un évènement qui ne peut pas être raffiné par l'expérience. L'ensemble de tous les évènements simples est appellé **espace des évènements** et noté Ω. Un évènement est donc une partie de Ω.

Un évènement a eu lieu si un de ses évènement simples a eu lieu. La probabilité $P(A)$ d'un évènement $A \subset \Omega$ est obtenue en répétant l'expérience et comptant

$$P(A) = \lim \frac{\text{nombre de fois } A \text{ a eu lieu}}{\text{nombre des expériences}}$$

dans la limite où l'expérience est répétée infiniment.

4.2 Propriétés élémentaires

Soit $A \subset \Omega$,

1. $0 \leq P(A) \leq P(\Omega)$
2. $P(\emptyset) = 0$ et $P(\Omega) = 1$
3. $P(A^c) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

L'espace des évènements Ω est appellé **discret**, s'il est fini ou dénombrable. Dans ce cas

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

L'espace des évènements Ω est appellé **continu**, si Ω est un intervalle (ou une union d'intervalles) de $\mathbb{R}$. Dans ce cas on peut avoir $P(\{\omega\}) = 0$ pour tous $\omega \in \Omega$. P est alors déterminé par la fonction de répartition

$$F(x) := P(]-\infty, x]).$$

Donc, si $P(\{a\}) = 0$

$$P([a, b]) = F(b) - F(a).$$

4.3 Probabilité uniforme (cas discret)

On considère le cas où tout évènement élementaire a la même probabilité et celle-ci est non nulle. Alors Ω doit être fini et

$$P(A) = \frac{|A|}{|\Omega|}.$$

Ici $|A|$ désigne le nombre des éléments dans A.

Le nombre de possibilités de choisir k parmi n objets (sans tenir compte de leur ordre) est

$$c_n^k := \frac{n!}{k!(n-k)!}.$$

On utilise aussi la notation $c_n^k = \binom{n}{k}$.

4.4 Probabilité conditionelle

Soit $A, B \subset \Omega$. On note $P(A|B)$ la probabilité que A a lieu sachant que B a eu lieu. Alors

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Formule de Bayes : Soient $A_1, A_2, \cdots, A_n$ des évènements incompatibles, c.à.d. $A_i \cap A_j = \emptyset$ pour $i \neq j$, et soit $A = A_1 \cup A_2 \cup \cdots \cup A_n$. Alors, pour $k = 1, \cdots, n$ on a

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Dans le cas particulier où $A_1 = A$, $A_2 = A^c$, le complémentaire de A, on obtient

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

4.5 Évènements indépendants

On dit que A et B sont **indépendants** si $P(A|B) = P(A)$ ou, d'une manière équivalente,

$$P(A \cap B) = P(A)P(B).$$

On dit que $A_1, A_2, \cdots, A_n$ sont indépendants, si

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

5 Variables aléatories et lois de probabilité

5.1 Variables aléatories discrètes

Une **variable aléatoire** (v.a.) est une fonction $X : \Omega \to \mathbb{R}$ d'un espace d'évènements Ω (muni d'une probabilité $\mathbb{P}$) à valeurs réelles. X est discret dans le sens que l'image de X est un ensemble fini ou dénombrable, i.e. $\{x_1, x_2, \cdots\}$. La **loi de probabilité** de X est la donnée

$$P(X = x) = \sum_{\omega : X(\omega) = x} \mathbb{P}(\{\omega\}).$$

Remarques :

1. En pratique, Ω et $\mathbb{P}$ ne jouent pas de rôle et on peut les oublier. **Pour spécifier la loi d'une variable aléatoire X il suffit de donner les probabilités $P(X = x)$ des valeurs x que X peut prendre.**
2. La donnée des $P(X = x)$ permet de calculer la probabilité d'autres évènements comme
$$P(a < X \leq b) = \sum_{a < x \leq b} P(X = x).$$
3. $P(X = x) \in [0, 1]$ et la somme de tous les $P(X = x)$ vaut 1.

5.2 Espérance et variance d'une v.a. discrète

L'espérance d'une v.a. X est

$$\mathbb{E}(X) = \sum_x x P(X = x)$$

(la somme est sur toutes les valeurs que X peut prendre).

La **variance** d'une v.a. X est

$$\mathbb{V}(X) = \sum_x (x - \mathbb{E}(X))^2 P(X = x) = \left(\sum_x x^2 P(X = x) \right) - E(X)^2$$

et **l'écart-type** de X est

$$\sigma(X) = \sqrt{\mathbb{V}(X)}.$$

5.3 Plusieurs variables aléatoires discrètes

Deux variable aléatoires discrètes (donc deux fonctions $X, Y : \Omega \to \mathbb{R}$) forment un couple (X, Y). La loi du couple est la donnée des probabilités que X prenne la valeur x et qu'en même temps Y prenne la valeur y,

$$P(X = x, Y = y) = \sum_{\omega : X(\omega) = x, Y(\omega) = y} \mathbb{P}(\{\omega\})$$

et en pratique la seule chose qui joue un rôle est les données $P(X = x, Y = y)$ pour les paires de valeurs de X et Y.

Deux variables sont **indépendantes** si, pour toute paire de valeurs (x, y)

$$P(X = x, Y = y) = P(X = x) P(Y = y)$$

et $X_1, \cdots, X_n$ sont indépendantes si $P(X_1 = x_1, \cdots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$.

5.4 Calcul avec les variables aléatoires

Soit X une variable aléatoire et $f : \mathbb{R} \to \mathbb{R}$ une fonction. Alors $Y = f(X)$ est une variable aléatoire de loi

$$P(Y = y) = \sum_{x:f(x)=y} P(X = x).$$

Soient X et Y deux variables aléatoires. Alors $Z = X + Y$ est une variable aléatoire de loi

$$P(Z = z) = \sum_{x,y:x+y=z} P(X = x, Y = y) = \sum_{x} P(X = x, Y = z - x).$$

De plus

$$\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y)$$

et, si X et Y sont indépendantes

$$\mathbb{V}(Z) = \mathbb{V}(X) + \mathbb{V}(Y).$$

Si $X_1, \cdots, X_n$ sont n variables aléatoires de même loi, alors

$$\mathbb{E}(X_1 + \cdots + X_n) = n\mathbb{E}(X_1)$$

et si en plus les variables sont indépendantes, alors

$$\mathbb{V}(X_1 + \cdots + X_n) = n\mathbb{V}(X_1)$$

et donc

$$\sigma(X_1 + \cdots + X_n) = \sqrt{n}\sigma(X_1).$$

5.5 Les lois discrètes usuelles

5.5.1 Loi de Bernoulli de paramètre $p \in [0, 1]$, $\mathcal{B}(1, p)$

X prend les valeurs $0, 1$ avec probabilités

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

$\mathbb{E}(X) = p$ et $\mathbb{V}(X) = p(1 - p)$.

5.5.2 Loi binomiale de paramètres $n \in \mathbb{N}_*$, $p \in [0, 1]$, $\mathcal{B}(n, p)$, `binom`

X prend les valeurs $0, 1, \cdots, n$ avec probabilités

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$\mathbb{E}(X) = np$ et $\mathbb{V}(X) = np(1 - p)$.

5.5.3 Loi de Poisson de paramètre $\lambda > 0$, $\mathcal{P}(\lambda)$, `pois`

X prend les valeurs $0, 1, \cdots$ avec probabilités

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$\mathbb{E}(X) = \lambda$ et $\mathbb{V}(X) = \lambda$.

5.5.4 Loi géométrique de paramètre $p \in]0,1[$, $\mathcal{G}eo(p)$, `geom`

X prend les valeurs $0, 1, \cdots$ avec probabilités

$$P(X = k) = (1-p)^k p$$

$\mathbb{E}(X) = \frac{1-p}{p}$ et $\mathbb{V}(X) = \frac{1-p}{p^2}$. *Attention, dans la litterature americaine, et aussi dans R, la loi géométrique correspond à la variable $X + 1$.*

5.6 Comparaison de variables aléatoires

Pour comparer deux variables aléatoires on les met sous forme centrée réduite. Si X est une variable aléatoire alors

$$Y = \frac{X - \mathbb{E}(X)}{\sigma(X)}$$

est une variable aléatoire qui est **centrée**,

$$\mathbb{E}(Y) = 0$$

et **réduite**,

$$\mathbb{V}(Y) = 1.$$

On compare maintenant les lois des variables centrées réduites de loi binomiale avec paramètre n qui tend vers $+\infty$ et $p = \frac{1}{2}$: Si X est de loi $\mathcal{B}(n, \frac{1}{2})$ alors $\mathbb{E}(X) = \frac{n}{2}$ et $\sigma(X) = \frac{\sqrt{n}}{2}$. Donc $Y = \frac{X - \mathbb{E}(X)}{\sigma(X)}$ prend les valeurs $y_k = -\sqrt{n} + \frac{2k}{\sqrt{n}}$, $k = 0, \cdots, n$ avec $P(Y = y_k) = \binom{n}{k} \frac{1}{2^n}$. On observe (voir aussi le fichier R C1.R) :

1. $P(Y = y_k)$ tend vers 0 si $n \to +\infty$.
2. La distance entre deux valeurs consécutives de Y est $y_{k+1} - y_k = \frac{1}{\sigma(X)}$. Elle tend vers 0 si $n \to +\infty$.
3. L'écart $y_n - y_0 = 2\sigma(X)$ tend vers $+\infty$ quand $n \to +\infty$.
4. $\sigma(X) P(Y = y_k)$ tend vers $\rho_{\mathcal{N}}(y_k)$ quand $n \to +\infty$.

Ici $\rho_{\mathcal{N}} : \mathbb{R} \to \mathbb{R}$ est la **fonction gaussienne**

$$\rho_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Il s'en suit :

$$\sum_{a < y \leq b} P(Y = y) \overset{n \to +\infty}{\longrightarrow} \int_a^b \rho_{\mathcal{N}}(y) dy$$

Il doit y avoir une loi qui est limite des lois binomiales centrées réduites.

5.7 Variables aléatoires continues

Une **variable aléatoire** X est **continue** si sa loi est donnée par une **densité de probabilité** $\rho_X : \mathbb{R} \to \mathbb{R}^+$, c.a.d. si les probabilités sont données par

$$P(a < X \leq b) = \int_a^b \rho_X(x) dx$$

pour tout $a < b$. La densité ρ_X est une fonction positive, continue par morceaux, qui satisfait

$$\int_{-\infty}^{+\infty} \rho_X(x) dx = 1.$$

La probabilité que X prenne précisement la valeur x est nulle : $P(X = x) = 0$.

La **fonction de répartition** de la variable aléatoire est $F_X(x) := P(X \leq x)$, c.à.d.

$$F_X(x) = \int_{-\infty}^{x} \rho_X(x')dx'.$$

On a donc $\rho_X(x) = F'_X(x)$.

5.8 Espérance et variance d'une variable aléatoire continue

L'espérance d'une variable aléatoire continue X est

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x\rho_X(x)dx.$$

La **variance** d'une variable aléatoire continue X est

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2\rho_X(x)dx = \int_{-\infty}^{+\infty} x^2\rho_X(x)dx - \mathbb{E}(X)^2$$

et **l'écart-type** de X est

$$\sigma(X) = \sqrt{\mathbb{V}(X)}.$$

5.9 Les lois continues usuelles

5.9.1 La loi normale de paramètres $\mu \in \mathbb{R}$, $\sigma > 0$

La loi normale de paramètres $\mu \in \mathbb{R}$, $\sigma > 0$, notée $\mathcal{N}(\mu, \sigma)$ (`dnorm`) a la densité

$$\rho_{\mathcal{N}(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Son espérance est μ et son écart type σ. Sa version centrée réduite est la loi normale centrée réduite $\mathcal{N}(0, 1)$.

Si X est une variable de loi $\mathcal{N}(\mu, \sigma)$, alors

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-\frac{y^2}{2}} dy.$$

5.9.2 La loi exponentielle

Le processus de Poisson décrit des évènements (de même nature) qui arrivent aléatoirement d'une manière indépendante dans un intervalle de temps $\Delta t = t_1 - t_0$. Soit λ la fréquence moyenne des évènements qui arrivent en temps Δt. Associé au processus de Poisson sont deux lois, une discrète et une continue.

1. La variable aléatoire X, qui compte le nombre des évènements en temps Δt. Elle suit la loi de Poisson $\mathcal{P}(\lambda)$ (`dpois`) :

   $$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

2. La variable aléatoire Y, qui compte le temps d'arrivée du premier évènement. Elle suit la loi exponentielle $\mathcal{E}(\lambda)$ (`dexp`) :

   $$P(Y > t) = e^{-\lambda t}.$$

   C'est une loi continue qui a comme densité

   $$\rho_{\mathcal{E}}(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

5.10 La moyenne empirique

Soit X une variable aléatoire avec espérance $\mathbb{E}(X) = \mu$ et écart-type $\sigma(X) = \sigma$. Soient $X_1, \cdots, X_n$ des variables aléatoires indépendantes de même loi que X. On appelle

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

la **moyenne empirique** de X. On trouve

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

5.10.1 Loi des grands nombres

Dans la limite où n tend vers $+\infty$ la moyenne empirique perd son caractère aléatoire : Pour tout $a > 0$ on a

$$P(\mu - a \leq \overline{X}_n < \mu + a) \overset{n\to+\infty}{\longrightarrow} 1.$$

5.10.2 Théorème central limite

$\overline{Y}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ est la moyenne empirique centrée réduite. La loi de $\overline{Y}_n$ tend vers la loi normale centrée réduite :

$$P(a \leq \overline{Y}_n \leq b) \overset{n\to+\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx.$$

Pour grand n un peut alors approcher $\overline{X}_n$ avec une variable aléatoire de loi $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

$$P(a \leq \overline{X}_n \leq b) \cong \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{n(x-\mu)^2}{2\sigma^2}} dx.$$

Pour une v.a. X de loi de Bernoulli de paramètre $p \in [0, 1]$ on trouve $\overline{Y}_n = \frac{n\overline{X}_n - np}{\sqrt{np(1-p)}}$. Comme $n\overline{X}_n = X_1 + \cdots + X_n$ suit la loi binomiale de paramètre n, p on déduit que pour grand n la loi $\mathcal{B}(n, p)$ peut être approchée par la loi $\mathcal{N}(np, \sqrt{np(1-p)})$.

6 Inférence statistique

Le but de l'inférence statistique est de déduire des observations, c.à.d. des mesures d'une caractéristique des individus d'un échantillon, la distribution de la caractéristique dans la population.

6.1 Échantillonage aléatoire

Un **échantillon aléatoire** de taille n est une suite de n variables aléatoires indépendantes $X_1, \cdots, X_n$ de même loi. La loi commune est appelée **loi mère** de l'échantillon. On notera μ sa moyenne et σ son écart type.

L'idée est que la distribution d'une caractéristique dans la population correspond à une variable aléatoire dont la loi est la loi mère. Mesurer la caractéristique des individus d'un échantillon correspond à réaliser les variables aléatoires $X_1, \cdots, X_n$, c.à.d. à obtenir des valeurs $x_1, \cdots, x_n$ pour les variables.

Une **statistique** est une variable aléatoire qui est fonction de l'échantillon. Les moments empiriques sont des statistiques. Le k-ième moment est la variable aléatoire

$$\overline{X^k} := \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Le moment d'ordre 1, $\overline{X}$ est donc la moyenne empirique, qu'on a aussi notée $\overline{X}_n$. Relié au moment d'ordre 2 est l'expression

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

qu'on appelle la **variance de l'échantillon**. Donc $S^2 = \frac{n}{n-1}(\overline{X^2} - \overline{X}^2)$. On a

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2.$$

6.2 Les lois associées aux statistiques

1 Si la loi mère est la loi normale $\mathcal{N}(\mu, \sigma)$, alors $\overline{X}$ suit la loi $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. D'une manière équivalente, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ suit la loi $\mathcal{N}(0,1)$.

2 Si la loi mère est la loi normale $\mathcal{N}(0,1)$, alors $\overline{X^2}$ suit la loi $\chi^2(n)$, dite la loi khi de n degrés de liberté. C'est une loi continue qui a comme densité

$$\rho_{\chi^2(n)}(x) = cx^{\frac{n}{2}-1}e^{-\frac{x}{2}} \text{ si } x \geq 0 \quad \text{et 0 si } x < 0$$

où $c^{-1} = \int_0^{+\infty} x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx$.

3 Si la loi mère est la loi normale $\mathcal{N}(\mu, \sigma)$, alors $\frac{n-1}{\sigma^2}S^2$ suit la loi $\chi^2(n-1)$.

4 Si on ne connaît pas la loi mère (mais qu'elle est numérique) alors les lois de $\overline{X}$ et S^2 peuvent être approximées par les lois données en 1 et 3, quand n est grand ($n \geq 30$ pour $\overline{X}$ et $n \geq 100$ pour S^2).

5 Si la loi mère est la loi normale $\mathcal{N}(\mu, \sigma)$, alors $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit la loi $t(n-1)$. $t(n)$ est appellée la loi de Student à n degrés de liberté. C'est une loi continue qui a comme densité

$$\rho_{t(n)}(x) = c\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

où $c^{-1} = \int_{-\infty}^{+\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$. Pour grand n ($n \geq 100$) la loi de Student peut être approximée par la loi $\mathcal{N}(0,1)$.

6 Si on ne connaît pas la loi mère (mais qu'elle est numérique) alors la loi de $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ peut être approximée par la loi $\mathcal{N}(0,1)$, quand n est grand ($n \geq 100$).

6.3 Estimateurs

Un **estimateur** est une fonction de l'échantillon aléatoire $f(X_1, \cdots, X_n)$ (donc une statistique), qui sert à estimer un paramètre θ de la loi mère. Les estimateurs que nous allons utiliser sont les statistiques dont nous avons discuté plus haut, en effet, principalement $\overline{X}$, qui sert à estimer la moyenne de la loi mère.

Le **biais** d'un estimateur $f(X_1, \cdots, X_n)$ pour le paramètre θ est la différence

$$\mathbb{E}(f(X_1, \cdots, X_n)) - \theta.$$

$\overline{X}$ est un estimateur sans biais pour la moyenne μ de la loi mère. S^2 est un estimateur sans biais pour la variance σ^2 de la loi mère.

Un bon estimateur devait au moins être aymptotiquement sans biais, c.à.d.

$$\mathbb{E}(f(X_1, \cdots, X_n)) - \theta \overset{n\to+\infty}{\longrightarrow} 0.$$

L'idée est alors que, pour estimer θ, on effectue une réalisation $x_1, \cdots, x_n$ de l'échantillon aléatoire et on calcule la valeur $f(x_1, \cdots, x_n)$. Cette valeur donne une estimation pour θ, qui sera de plus en plus précise, quand la taille de l'échantillon augmente. Par la loi des grands nombres l'estimation devient déterministe dans la limite où n tend vers $+\infty$.

6.4 Intervalle de confiance

L'idée de base des intervalles de confiance est de donner une information supplémentaire sur l'erreur de l'estimation d'un paramètre. On se fixe un **niveau de confiance** $1-\alpha$, $\alpha \in [0,1]$, et on détermine un intervalle autour de la valeur estimée pour θ, dans lequel le vrai θ devait se situer avec confiance $1-\alpha$.

On note $z_{\frac{\alpha}{2}}$ la valeur réelle positive telle que

$$\int_{-z_{\frac{\alpha}{2}}}^{z_{\frac{\alpha}{2}}} \rho_{\mathcal{N}(0,1)}(x)dx = 1-\alpha$$

($\rho_{\mathcal{N}(0,1)}$ la densité de la loi normale).

On note $t(n)_{\frac{\alpha}{2}}$ la valeur réelle positive telle que

$$\int_{-t(n)_{\frac{\alpha}{2}}}^{t(n)_{\frac{\alpha}{2}}} \rho_{t(n)}(x)dx = 1-\alpha$$

($\rho_{t(n)}$ la densité de la loi de Student).

Voici quelques valeurs pour $z_{\frac{\alpha}{2}}$.

$1-\alpha$	0.80	0.85	0.90	0.95	0.99
$z_{\frac{\alpha}{2}}$	1.28	1.44	1.645	1.96	2.58

6.4.1 Intervalle de confiance pour la moyenne

On construit un intervalle de confiance pour la moyenne μ d'une loi mère dont la variable est numérique.

1. On suppose que la loi mère est $\mathcal{N}(\mu, \sigma)$ et que σ **est connu**. On prend $\overline{X}$ comme estimateur. On a alors
$$\mathbb{P}\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$
Si $x_1, \cdots, x_n$ est une réalisation de l'échantillon aléatoire et $\overline{x}$ sa moyenne, alors l'intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est
$$\mu \in \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right).$$

2. On suppose que la loi mère est $\mathcal{N}(\mu, \sigma)$ et que σ **n'est pas connu**. On prend $\overline{X}$ comme estimateur pour μ et S^2 comme estimateur pour σ^2. On a alors
$$\mathbb{P}\left(\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right| \leq t(n-1)_{\frac{\alpha}{2}}\right) = 1-\alpha.$$
Si $x_1, \cdots, x_n$ est une réalisation de l'échantillon aléatoire, $\overline{x}$ sa moyenne et s^2 sa variance, alors l'intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est
$$\mu \in \left(\overline{x} - \frac{s}{\sqrt{n}} t(n-1)_{\frac{\alpha}{2}}, \overline{x} + \frac{s}{\sqrt{n}} t(n-1)_{\frac{\alpha}{2}}\right).$$
Pour grand n grand ($n \geq 100$) on peut remplacer $t(n-1)_{\frac{\alpha}{2}}$ par $z_{\frac{\alpha}{2}}$ dans cette expression.
3. Si la loi mère n'est pas connue, on peut quand même travailler avec l'intervalle de confiance $\mu \in \left(\overline{X} - \frac{s}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{X} + \frac{s}{\sqrt{n}} z_{\frac{\alpha}{2}}\right)$ si n est grand. Ici, si σ est connu, on prend $s = \sigma$, sinon s^2 est la variance de l'échantillon.

6.4.2 Intervalle de confiance pour une proportion

On considère le cas où la loi mère est celle d'une variable aléatoire qualitative, le choix entre deux alternatives (vote d'une population entre deux candidats). Ceci se décrit avec la loi de Bernoulli de paramètre p : la variable X vaut 1 pour l'alternative "oui" et 0 pour l'alternative "non". p serait alors la proportion des individus dans la population qui votent "oui". On construit un intervalle de confiance pour la probabilité p de la loi Bernoulli.

L'estimateur est $\overline{X}$. $n\overline{X} = X_1 + \cdots + X_n$ suit alors une loi binomiale $\mathcal{B}(n,p)$. Pour n grand on approche cette loi par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$. Donc $\frac{\overline{X}-p}{\sqrt{\overline{X}(1-\overline{X})}/\sqrt{n}}$ suit à peu près la loi $\mathcal{N}(0,1)$ et
$$\mathbb{P}\left(\left|\frac{\overline{X}-p}{\sqrt{\overline{X}(1-\overline{X})}/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Si $x_1, \cdots, x_n$ est une réalisation de l'échantillon aléatoire et $\overline{x}$ sa moyenne, alors l'intervalle de confiance de niveau $1-\alpha$ pour la proportion p est
$$p \in \left(\overline{x} - \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{x} + \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}} z_{\frac{\alpha}{2}}\right).$$

6.4.3 Taille d'un échantillon

La largeur d'un intervalle de confiance dépend de la taille de l'échantillon. On peut se demander quelle taille on doit prendre pour obtenir une largeur qui ne dépasse pas une valeur maximale 2δ donnée.

1. Cas d'un intervalle pour la moyenne μ d'une loi mère numérique. On suppose que la loi mère est une loi normale ou que n est suffisamment grand. Dans ce cas l'intervalle de confiance de niveau $1-\alpha$ a une largeur au plus 2δ si

$$n \geq \left(\frac{\sigma}{\delta} z_{\frac{\alpha}{2}}\right)^2.$$

   Ceci ne dépend pas de la valeur de μ mais de la variance σ^2, qui doit donc être connue en avance.

2. Cas d'un intervalle pour la proportion p d'une loi mère de Bernoulli. Dans ce cas la variance σ^2 dépend de p et pour obtenir une formule qui ne dépend pas d'un choix apriori pour σ, on prend la valeur maximale de $\sigma^2 = p(1-p)$ qui est $\frac{1}{4}$. Avec l'approximation de la loi binomiale par une loi normale, on trouve que l'intervalle de confiance de niveau $1-\alpha$ a une largeur au plus 2δ si

$$n \geq \left(\frac{1}{2\delta} z_{\frac{\alpha}{2}}\right)^2.$$

6.5 Test d'hypothèses

Un test statistique a comme but d'affirmer ou de rejeter une hypothèse sur la base d'une statistique. Ici une hypothèse concerne les caractéristiques (des individus) d'une population. La décision d'affirmer ou de rejeter une hypothèse se fait après observation d'un échantillon tiré au hasard, sur la base des données de ses caractéristiques.

Vu qu'une statistique ne donne pas des résultats avec certitude, il y a un risque d'erreur. Celui se decompose en deux types :

`Type I` On affirme l'hypothèse bien qu'elle est fausse.

`Type II` On rejete l'hypothèse bien qu'elle est juste.

Logiquement, affirmer une hypothèse est la même chose que rejeter son opposé. Mais en pratique, affirmer une hypothèse bien qu'elle soit fausse peut être plus grave que la rejeter bien qu'elle soit juste. L'exemple qui illustre bien cela est l'hypothèse qu'un accusé jugé par un tribunal soit coupable. Affirmer cette hypothèse bien qu'elle soit fausse veut dire condamner un innocent, ce qui est bien plus grave que de laisser aller un coupable impuni. C'est pour cela qu'en statistique, il y a une asymétrie entre l'hypothèse et son opposé. En particulier, on veille à ce que le risque de type `I` soit minimal et on ne s'occupe pas trop du risque de type `II`.

On note H_1 l'hypothèse dont on veut que le risque de son affirmation à tort soit minimal. On note H_0 son opposé. Affirmer H_1 veut donc dire rejeter ("nullifier") H_0 (l'hypothèse nulle). Limiter le risque de type `I` veut donc dire que, pour un $\alpha \in [0,1]$ petit (typiquement $\alpha = 5\%$) on a

$$\mathbb{P}(\text{rejeter } H_0 \text{ à tort}) \leq \alpha. \tag{1}$$

6.5.1 Modèle du test

On modélise la situation avec un échantillon aléatoire $X_1, \cdots, X_n$. L'hypothèse concerne donc une propriété de la loi mère de l'échantillon aléatoire, et l'observation d'un échantillon tiré au hasard correspond à une réalisation $x_1, \cdots, x_n$ des variables aléatoires.

Nous considérons ici le cas où l'hypothèse concerne un paramètre de la loi mère. L'hypothèse est formulée à l'aide d'un estimateur T pour le paramètre. T est une fonction de l'échantillon aléatoire, $T = T(X_1, \cdots, X_n)$. Après observation on obtient alors un nombre $T(x_1, \cdots, x_n)$.

Pour décider si on rejette H_0 (et donc affirmer H_1), on désigne une région $C \subset \mathbb{R}$ telle que le critère de rejet de H_0 soit $T(x_1, \cdots, x_n) \in C$. C doit être choisi d'une telle manière que (1) soit satisfait, c.à.d., la probabilité que T appartient à C bienque H_0 soit vrai est au plus α,

$$P(T \in C | H_0 \text{ est vrai}) \leq \alpha \tag{2}$$

En particulier, C dépend du risque α (confiance $1 - \alpha$).

Si l'hypothèse H_0 spécifie la loi mère, alors $P(T \in C | H_0 \text{ est vrai})$ est la probabilité que $T \in C$ sous cette loi mère. (Si H_0 spécifie une famille de lois mères il faut calculer la probabilité que $T \in C$ avec une loi de la famille qui maximise cette probabilité.)

6.5.2 Test pour la moyenne μ

On veut tester une hypothèse sur la **moyenne** d'une caractéristique **numérique** de la population. Avec `R` ce test se fait avec la commande `t.test`.

La caractéristique est modélisée par une variable aléatoire dont la loi est la loi mère de notre échantillon aléatoire. L'estimateur pour la moyenne μ de la loi mère est $T = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, la moyenne empirique. Mais si la variance n'est pas connue autrement, on aura aussi besoin de l'estimateur pour la variance de l'échantillon $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

L'hypothèse se formule à l'aide d'une valeur μ_0 donnée. On distingue trois cas :

l'hypothèse H_1 est que $\mu < \mu_0$
l'hypothèse H_1 est que $\mu > \mu_0$
l'hypothèse H_1 est que $\mu \neq \mu_0$

Dans les deux premiers cas on parle d'une hypothèse unilatérale, dans le troisième d'une hypothèse bilatérale. On désigne pour C une région de la forme $]-\infty, \mu_0 - \delta]$, $[\mu_0 + \delta, +\infty[$, $]-\infty, \mu_0 - \delta] \cup [\mu_0 + \delta, +\infty[$, respectivement. Le δ dépend de α et de la loi de l'estimateur.

La taille de l'échantillon est grande ($n \geq 30$) Dans ce cas on peut approcher la loi de $\overline{X}$ par la loi normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. On obtient comme critère de rejet de H_0 (affirmation de H_1) après observation de l'échantillon :

$$\text{On affirme } \mu < \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \leq \mu_0 - \frac{s}{\sqrt{n}} z_\alpha$$
$$\text{On affirme } \mu > \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \geq \mu_0 + \frac{s}{\sqrt{n}} z_\alpha$$
$$\text{On affirme } \mu \neq \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad |\overline{x} - \mu_0| \geq \frac{s}{\sqrt{n}} z_{\frac{\alpha}{2}}$$

Ici, $\overline{x}$ est la moyenne de l'échantillon et la valeur de s dépend de la situation.

1. Si la valeur de σ est connue on prend $s = \sigma$.
2. Sinon, on prend $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ l'écart-type de l'échantillon.

On note que dans le cas bilatéral, le critère d'affirmation de H_0 (donc de $\mu = \mu_0$) est exactement que μ_0 appartienne a l'intervalle de confiance déterminé par $\overline{x}$.

La taille de l'échantillon est petite et la loi mère une loi normale Dans ce cas $\overline{X}$ suit la loi normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

Si σ est connu, on peut procéder comme en haut (grande taille) avec $s = \sigma$.

Si σ n'est pas connu on doit remplacer la loi $\mathcal{N}(0,1)$ par la loi de Student. En effet $Y = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit la loi $t(n-1)$. On obtient comme critère de rejet de H_0 (affirmation de H_1) après observation de l'échantillon :

$$\text{On affirme } \mu < \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \leq \mu_0 - \frac{s}{\sqrt{n}} t(n-1)_\alpha$$

$$\text{On affirme } \mu > \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \geq \mu_0 + \frac{s}{\sqrt{n}} t(n-1)_\alpha$$

$$\text{On affirme } \mu \neq \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad |\overline{x} - \mu_0| \geq \frac{s}{\sqrt{n}} t(n-1)_{\frac{\alpha}{2}}$$

où $\overline{x}$ est la moyenne et s l'écart-type de l'échantillon.

6.5.3 Test pour la proportion p

On veut tester une hypothèse sur la **proportion** de la population qui est caractérisée par une **alternative** (la proportion des gens qui vote "oui"). Avec `R` ce test se fait avec la commande `prop.test`.

La situation est modélisée par un échantillon aléatoire dont la loi mère est la loi de Bernoulli, le paramètre p correspondant à la proportion. Vu que la moyenne de la loi de Bernoulli est $\mu = p$, la situation est comme pour les tests sur la moyenne μ. De plus, l'écart type σ de la loi de Bernoulli est $\sigma = \sqrt{p(1-p)}$ et est donc connu pour la loi qui correspond à la proportion de comparaison p_0.

$$\text{On affirme } p < p_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \leq p_0 - \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} z_\alpha$$

$$\text{On affirme } p > p_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \geq p_0 + \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} z_\alpha$$

$$\text{On affirme } p \neq p_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad |\overline{x} - p_0| \geq \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} z_{\frac{\alpha}{2}}$$

Si la taille de l'échantillon est petite on peut se baser sur la loi binomiale $\mathcal{B}(n,p)$, qui est la loi exacte pour $n\overline{X}$, mais comme $\mathcal{B}(n,p)$ est une loi discrète, les formules sont plus compliquées.

6.5.4 La p-valeur d'une observation

La p-valeur d'une observation (seuil d'importance observé) est un nombre entre 0 et 1 qui quantifie la confiance (ou le risque) de la décision d'affirmer une hypothèse sur la base d'une observation $x_1, \cdots, x_n$ d'un échantillon. La p-valeur de $x_1, \cdots, x_n$ est définie comme le plus petit α pour lequel on affirme H_1 sur la base de l'observation. Autrement dit, si $\alpha = \text{p-val}(x_1, \cdots, x_n)$ est la p-valeur de l'observation, la confiance maximale avec laquelle on peut affirmer H_1 est $1-\alpha$. Plus la p-valeur est petite, plus on peut avoir confiance dans l'affirmation de H_1.

La formule pour la p-valeur dépend du test. Si σ est connu, et la loi mère est normale ou la

taille est grande, la p-valeur pour un test de la moyenne est donnée par

$$\begin{aligned}
&H_1 \text{ est } \mu < \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = P(\overline{X} < \overline{x}) = \mathcal{F}_{\mathcal{N}(0,1)}\left(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)\\
&H_1 \text{ est } \mu > \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = P(\overline{X} > \overline{x}) = 1 - \mathcal{F}_{\mathcal{N}(0,1)}\left(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)\\
&H_1 \text{ est } \mu \neq \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = P(|\overline{X} - \mu_0| > |\overline{x} - \mu_0|) = 2\mathcal{F}_{\mathcal{N}(0,1)}\left(-\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Ici P est la probabilité calculée avec la loi $\mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$ pour $\overline{X}$ et $\mathcal{F}_{\mathcal{N}(0,1)}$ la fonction de répartition pour la loi $\mathcal{N}(0,1)$. Si la taille est grande, la p-valeur pour un test de la proportion est donnée par la même formule que celle pour la moyenne, si on prend $\mu = p$ et $\mu_0 = p_0$ et $\sigma = \sqrt{p_0(1-p_0)}$.

Finalement, dans le cas d'une loi mère qui est normale, mais sans connaissance de σ, il faut se ramener à la variable aléatoire $Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$. Avec $z = \frac{\overline{x}-\mu}{s/\sqrt{n}}$ (s est l'écart-type de l'échantillon) on obtient la formule

$$\begin{aligned}
&H_1 \text{ est } \mu < \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = \mathcal{F}_{t(n-1)}(z)\\
&H_1 \text{ est } \mu > \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = 1 - \mathcal{F}_{t(n-1)}(z)\\
&H_1 \text{ est } \mu \neq \mu_0 \qquad \text{p-val}(x_1, \cdots, x_n) = 2\mathcal{F}_{t(n-1)}(-|z|)
\end{aligned}$$

Ici $\mathcal{F}_{t(n-1)}$ est la fonction de répartition pour la loi de Student à $n-1$ degrés de liberté.