Université Claude Bernard Lyon 1
Année 2023-2024

Statistique pour l'Informatique

Notes du cours

version du 29 novembre 2023

Johannes Kellendonk

Chapitre 1

Statistique déscriptive

Les notions traduites en anglais sont donnés en *italique*.

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

Une variable (caractéristique) peut être

- qualitative : par exemple une couleur
- quantitative (ou 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).

L'effectif (ou taille) d'un échantillon I est le nombre de ses individus, c.à.d. le nombre d'éléments dans I, ce qu'on note $\#I$.

1.2 Description des données avec une caractéristique

Dans cette section on considère des échantillons avec individus qui ont une seule caractéristique. On note un tel échantillon $\{x_i : i \in I\}$ ou $\{x_1, \cdots, x_n\}$, où la caractéristique de l'individu $i \in I$ est notée x_i.

1.2.1 Notions

Fréquence absolue (ou l'effectif) de la caractéristique c de l'échantillon est le nombre d'individus avec caractéristique c, noté

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

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** (ou fréquence relative) de la caractéristique c de l'échantillon est la fréquence absolue divisée par la taille de l'échantillon.

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

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

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

1.2.2 Représentation graphique

Histogramme d'un échantillon avec une caractéristique numérique. 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 $[r_{n-1}, r_n[$,

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

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).

1.2.3 Notions pour les variables numériques

On considère un échantillon $x = \{x_1, \cdots, x_n\}$ (de taille n) avec $x_i \in \mathbb{R}$.

La **moyenne de l' échantillon** (moyenne empirique) *(mean)* x est

$$m(x) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

On note aussi $\bar{x} = m(x)$.

La **médiane de l'échantillon** *(median)* est la valeur qui 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

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

Les **quartiles de l'échantillon** partagent les valeurs de x en quatre parties selon leur grandeur, Q_1, Q_2, Q_3 sont les valeur au milieu de ces quatres parties. Comme ceci n'est pas toujours bien défini on prend

$$Q_1 = x^*_{\lfloor \frac{n}{4} \rfloor},\ Q_2 = x^*_{\lfloor \frac{2n}{4} \rfloor},\ Q_3 = x^*_{\lfloor \frac{3n}{4} \rfloor}$$

ou $\lfloor r \rfloor$ est le plus grand entier plus petit ou égal à que r.

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

La **variance de l'échantillon** *(variance)* x est

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

L'écart type de l'échantillon (déviation standard) *(standard deviation)* x est

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

La **variance sans biais de l'échantillon** *(variance)* x est

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

L'écart type sans biais de l'échantillon (déviation standard) *(standard deviation)* x est

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

Il y a juste un facteur $\frac{n}{n-1}$ de difference entre la variance et la variance sans biais, ce qui est negligible si n est grand. Mais la commande en python travaille avec la variance sans biais.

La somme de deux échantillons $x = \{x_1, \cdots, x_n\}$ et $y = \{y_1, \cdots, y_n\}$ est $x + y = \{x_1 + y_1, \cdots, x_n + y_n\}$ et si $r \in \mathbb{R}$ on pose $rx = \{rx_1, \cdots, rx_n\}$.

Si la caractéristique d'un échantillon a une unité, par exemple cm, alors la moyenne, la médiane, les quartiles et l'ècart type ont tous la même unité, pendant que la variance a l'unité au carrée (par exmple cm^2). C'est pour cela que c'est l'écart type qui nous donne l'information de l'ampleur de la déviation des valeurs de la moyenne (statistiquement).

Proposition 1 *Le moyenne est additive $m(x+y) = m(x)+m(y)$ et $m(rx) = rm(x)$ pour $r \in \mathbb{R}$. La variance et l'écart type ne sont pas additives mais $V(rx) = r^2V(x)$ et $\sigma(rx) = |r|\sigma(x)$, pour $r \in \mathbb{R}$.*

1.3 Description des données avec deux caractéristiques

On considère maintenant des échantillons qui ont des individus avec deux caractéristiques. Autrement dit, les individus ont un couple de caractéristiques. On note $(x, y) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$.

1.3.1 Fréquences

La **fréquence absolue** de la caractéristique (a, b) de l'échantillon (x, y) est $\#\{i \in I : x_i = a, y_i = b\}$.

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

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

On a

$$f(a) = \frac{\#\{i \in I : x_i = a\}}{\#I} = \sum_b \frac{\#\{i \in I : x_i = a, y_i = b\}}{\#I},$$

et

$$f(b) = \frac{\#\{i \in I : y_i = b\}}{\#I} = \sum_a \frac{\#\{i \in I : x_i = a, y_i = b\}}{\#I},$$

La **table de contingence** est la table des fréquences (absolues ou relatives) des couples des caractéristiques.

On peut visualiser un échantillon avec deux caractéristiques numériques dans un diagramme de nuage (scatterplot). C'est l'ensemble des points dans $\mathbb{R}^2$ qui corresponds aux couples (x_i, y_i) de l'échantillon.

1.3.2 Covariance et coefficient de corrélation

La moyenne de (x, y) est

$$m(x, y) = (m(x), m(y))$$

donc c'est un point de $\mathbb{R}^2$. La **covariance** de (x, y) est une matrice 2×2

$$\begin{pmatrix} V(x,x) & V(x,y) \\ V(y,x) & V(y,y) \end{pmatrix}$$

ou

$$V(x, y) = m(x - \bar{x})m(y - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

en particular, $V(x, x)$ est la variance de x et la matrice est symétrique.

Comme pour la variance il y a aussi une version non-biaisée de la covariance, elle est donnée par

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Proposition 2 *La covariance est additive dans la premiére variable et $V(rx, y) = rV(x, y)$ pour $r \in \mathbb{R}$. Par symétrie ceci s'applique aussi à la deuxième variable.*

Le **coefficient de corrélation** des deux variables x et y est

$$cor(x, y) = \frac{V(x, y)}{\sqrt{V(x, x)}\sqrt{V(y, y)}},$$

pourvu que les variances de x et de y ne sont pas nulles. On note que si les deux caractéristiques ont des unités distinctes, par example cm et kg, alors les unités de la matrice de covariances sont mixtes. Par contre le coefficient de corrélation n'a pas d'unité. C'est pour cela que c'est ce coefficient qui nous donne une information sur la corrélation.

Théorème 1 *On suppose que $V(x, x) \neq 0$ et $V(y, y) \neq 0$. On a*

$$-1 \leq cor(x, y) \leq 1$$

avec égalité si et seulement s'il y a des nombres réels $\lambda, \mu \neq 0$ et c t.q. $\forall i \in I : \mu y_i + \lambda x_i = c$. En particulier,

$cor(x, y) = 1$ si et seulement s'il existe $\lambda > 0$, $c \in \mathbb{R}$ t.q. $\forall i \in I : y_i = \lambda x_i + c$
$cor(x, y) = -1$ si et seulement s'il existe $\lambda < 0$, $c \in \mathbb{R}$ t.q. $\forall i \in I : y_i = \lambda x_i + c$

On note que s'il existent avec $\mu, \lambda \neq 0$ t.q. $\forall i \in I : \mu y_i + \lambda x_i = c$ alors y_i est déterminé par x_i et vice versa, donc il y a une correlation parfaite entre les deux characteristiques. Par contre, si $cor(x, y)$ est proche de 0 les données ne sont pas corrélées.

1.3.3 Régression linéaire

Etant donné un échantillon à deux variables numériques $(x, y) = \{(x_1, y_1), \cdots (x_n, y_n)\}$, on cherche à déterminer les paramétres $\lambda \in \mathbb{R}^*, c \in \mathbb{R}$ de la droite

$$y = \lambda x + c$$

qui approche le mieux les données (x_i, y_i). Ceci a un sens seulement si le coefficient de corrélation $cor(x, y)$ est proche de 1 ou de -1 et on peut donc soupconner un corrélation linéaire entre les

deux variables. L'erreur entre la valeur y_i de l'échantillon et la valeur donnée par l'équation de la droite,

$$\mathcal{E}(\lambda, c) := \frac{1}{n}\sum_{i=1}^{n}(y_i - (\lambda x_i + c))^2$$

est minimal si

$$\lambda = \frac{V(x,y)}{V(x,x)}$$
$$c = \overline{y} - a\overline{x}.$$

Chapitre 2

Probabilité

2.1 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.

2.2 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 $P(\{a\}) = 0$

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

2.3 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.

Exemples : Voici trois exemples basés sur le processus de tirer k billes numerotées au hazard d'un sac de n billes. La probabilité de tirer une bille est la même pour chaque bille. On tire plusieurs fois, les tires sont supposés indépendent.

1. On tire une bille après l'autre, chaque fois *avec remise*, et on *prend l'ordre* des résultats *en compte*. Alors l'espace des évènements est le produit Cartesien $\{1, \cdots, n\}^k$ et la probabilité de l'évènement $(i_1, \cdots, i_n)$ est $\frac{1}{n^k}$.
2. On tire une bille après l'autre, *sans remise*, et on *prend l'ordre* des résultats *en compte*. Alors l'espace des évènements consists en des évènements simples $(i_1, \cdots, i_n)$ avec $i_j \neq i_l$ si $j \neq l$ et la probabilité de cette évènement est $\frac{1}{n(n-1)\cdots(n-k+1)} = \frac{(n-k)!}{n!}$.
3. On tire une bille après l'autre, *sans remise*, mais on *ignore l'ordre* des résultats. Alors l'espace des évènements consists en des évènements simples $\{i_1, \cdots, i_n\}$ (notation des ensembles, ou on ne répète pas les éléments) et la probabilité de cette évènement est $\frac{(n-k)!k!}{n!}$. En effet 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 $c_n^k = \binom{n}{k}$.

2.4 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

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

2.5 Évènements indépendants

On dit que deux évènements 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 n évènements $A_1, A_2, \cdots, A_n$ sont indépendants, si, pour tout choix $A_{i_1}, \cdots, A_{i_k}$ de k évènements parmi ces n évènements on a

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

2.6 Variables aléatoires discrètes et leur lois de probabilité

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 spécifier
   (i) **les valeurs que** X **peut prendre**, c.à.d. l'image de X, notée im$X$.
   (ii) **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.

2.6.1 Espérance et variance d'une variable aléatoire discrète

L'espérance d'une variable aléatoire X est

$$\mathbb{E}(X) = \sum_{x \in \text{im}X} xP(X = x).$$

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

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

et **l'écart-type** de X est

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

2.6.2 Fonction d'une variable aléatoire

Soit X une variable aléatoire et $f : \mathbb{R} \to \mathbb{R}$ une fonction. Alors $Y = f(X)$ est une (nouvelle) variable aléatoire dont la loi est donnée par
(i) Y prend les valeurs $f(x)$ ou x est valeur de X, c.à.d. im$Y = \{f(x) : x \in \text{im}X\}$.
(ii)

$$P(Y = k) = \sum_{\substack{x \in \text{im}X \\ f(x) = k}} f(x)$$

L'espérance de Y est alors donné par

$$\mathbb{E}(Y) = \sum_{y \in \text{im}Y} yP(Y = y) = \sum_{x \in \text{im}X} f(x)P(X = x)$$

2.6.3 Les lois discrètes usuelles

Voici les lois discrètes qui sont utilisés le plus souvent, avec leurs espérance et leur variance.

	notation	valeurs de X	$P(X=k)$	espérance	variance
loi de Bernoulli	$\mathcal{B}(p)$	$k \in \{0,1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
loi binomiale	$\mathcal{B}(n,p)$	$k \in \{0,\cdots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
loi géométrique	$\mathcal{G}(p)$	$k \in \mathbb{N}^*$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
loi de Poisson	$\mathcal{P}(\lambda)$	$k \in \mathbb{N}$	$\exp(-\lambda)\frac{\lambda^k}{k!}$	λ	λ
loi uniforme	$\mathcal{U}(n)$	$k \in \{1,\cdots,n\}$	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

La loi Bernoulli correspond à la loi binomiale avec $n = 1$.

Voici des situations ou on peut appliquer ces lois.

Un canal de communication transmet un signal (1 bit) correctement avec une probabilité p.

— J'ai reçu n signaux. Quelle est la probabilité que k de ces signaux soient corrects ? L'événement correspond à la valeur k d'une variable aléatoire X de la loi binomiale $\mathcal{B}(n,p)$. La probabilité est alors

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

— J'ai reçu k signaux. Quelle est la probabilité que les $k-1$ premiers signaux soient faux, alors que le $k$ième signal est correct ? L' évènement correspond à la valeur k d'une variable aléatoire X de loi géométrique $\mathcal{G}(p)$. La probabilité est alors

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

Un canal de communication transmet des signaux. Il fait en moyenne λ erreurs par heures.

— J'ai reçu des signaux pendant une heure. Quelle est la probabilité que k de ces signaux soient faux ? L' évènement correspond à la valeur k d'une variable aléatoire X de la loi de Poisson $\mathcal{P}(\lambda)$. La probabilité est alors

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

— J'ai reçu des signaux pendant **deux** heures. Quelle est la probabilité que k de ces signaux soient faux ? L' évènement correspond à la valeur k d'une variable aléatoire X de la loi de Poisson $\mathcal{P}(2\lambda)$. La probabilité est alors

$$P(X=k) = \exp(-2\lambda)\frac{(2\lambda)^k}{k!}$$

2.6.4 Plusieurs variables aléatoires discrètes

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

$$P(X = k, Y = l) = \sum_{\omega : X(\omega)=k, Y(\omega)=l} \mathbb{P}(\{\omega\})$$

et en pratique la seule chose qui joue un rôle est la donné des probabilités conjointes $P(X = k, Y = l)$ pour les paires de valeurs (k, l) de X et Y.

La loi du couple (X, Y) détermine les lois de X et de Y (les lois marginales), notament

$$P(X = k) = \sum_{l\in \mathrm{im}Y} P(X = k, Y = l), \quad P(Y = l) = \sum_{k\in \mathrm{im}X} P(X = k, Y = l)$$

La **covariance** du couple (X, Y) est

$$\mathbb{V}(X, Y) = \mathbb{E}\big((X - \mathbb{E}(X)(Y - \mathbb{E}(Y))\big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

La variance de X est alors $\mathbb{V}(X) = \mathbb{V}(X, X)$.

2.6.5 Variables aléatoires indépendantes

Les variables X et Y sont **indépendantes** si, pour toute paire de valeurs (k, l)

$$P(X = k, Y = l) = P(X = k)P(Y = l).$$

Les variables $X_1, \cdots, X_n$ sont indépendantes si, pour tout choix $X_{i_1}, \cdots, X_{i_k}$ de k variables parmi ces n variables on a

$$P(X_{i_1} = n_1, \cdots, X_{i_k} = n_k) = P(X_{i_1} = n_1) \cdots P(X_{i_k} = n_k).$$

Si X et Y sont indépendants alors leur covariance est nulle :

$$\mathbb{V}(X, Y) = 0$$

2.6.6 Calcul avec les variables aléatoires

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

$$P(Z = n) = \sum_{k,l:k+l=n} P(X = k, Y = l) = \sum_{k} P(X = k, Y = n - k).$$

De plus

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

et

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathbb{V}(X, Y)$$

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

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

Théorème 2 *Soit X une variable aléatoire de loi $\mathcal{B}(n, p)$ et Y une variable aléatoire de loi $\mathcal{B}(m, p)$. On suppose que les deux variables sont indépendants. Alors $Z = X + Y$ suit la loi $\mathcal{B}(n + m, p)$.*

2.6.7 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 que

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.

2.7 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 $\rho_X(x) = F'_X(x)$.

2.8 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

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

2.8.1 Les lois continues usuelles

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

La loi normale de paramètres $\mu \in \mathbb{R}$, $\sigma > 0$, notée $\mathcal{N}(\mu, \sigma)$ (ou $\mathcal{N}(\mu, \sigma^2)$) a la densité

$$\rho_{\mathcal{N}(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.1}$$

Son espérance est μ et son écart type σ. On note que $\rho_{\mathcal{N}(\mu,\sigma)}(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$ est la valeur maximale de la densité et $\rho_{\mathcal{N}(\mu,\sigma)}(\mu \pm \sigma) = \frac{1}{\sqrt{e}}\rho_{\mathcal{N}(\mu,\sigma)}(\mu) = 0.607\rho_{\mathcal{N}(\mu,\sigma)}(\mu)$. Donc à l'écart de σ de sa moyenne la densité est 60.7% de la densité maximale.

La 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

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

La fonction de repartition $F_{\mathcal{N}(0,1)}$ de la loi $\mathcal{N}(0, 1)$ est strictement croissante et donc admit une fonction reciproque. Etant donné $\alpha \in [0, 1]$ il existe alors un unique réel z_α t.q.

$$F_{\mathcal{N}(0,1)}(z_\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} e^{-\frac{x^2}{2}} dx = 1 - \alpha.$$

La loi exponentielle de paramètre $\lambda > 0$

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 $\mathcal{P}(\lambda)$:
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$
2. La variable aléatoire Y, qui compte le temps d'arrivée du premier évènement. Elle suit la loi exponentielle $\mathcal{E}(\lambda)$:
$$P(Y > t) = e^{-\lambda t}.$$
C'est une loi continue qui a comme densité
$$\rho_{\mathcal{E}}(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

La loi uniform de paramètres $a < b$

Il s'agit de la loi, notée $\mathcal{U}[a,b]$ avec la densité

$$\rho_{\mathcal{U}[a,b]}(t) = \begin{cases} \frac{1}{b-a} & \text{si } t \in [a,b] \\ 0 & \text{sinon.} \end{cases}$$

Théorème 3 *Soit X une variable aléatoire de loi $\mathcal{N}(\mu,\sigma)$ et Y une variable aléatoire de loi $\mathcal{N}(\nu,\tau)$. Si X et Y sont indépendents alors $Z = X + Y$ suit la loi $\mathcal{N}(\mu+\nu, \sqrt{\sigma^2+\tau^2})$.*

2.8.2 Limites des variables aléatoires

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

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

Théorème 4 (Loi des grands nombres) *Soit $X_1, \cdots, X_n$ une suite des variables aléatoires indépendantes de même loi. On suppose que la moyenne de la loi (l'esperence de X_1) et μ et que sa variance soit finie. Alors, pour tout $a > 0$ on a*

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

On peut alors dire que dans la limite où n tend vers $+\infty$ la moyenne empirique perd son caractère aléatoire.

Théorème 5 (Théorème central limite) *Soit $X_1, \cdots, X_n$ une suite des variables aléatoires indépendantes de même loi. On suppose que la moyenne de la loi (l'esperence de X_1) et μ et que sa variance est σ^2. Soit $\overline{Y}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$, la moyenne empirique centrée réduite associée à $\bar{X}_n$. 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 variable aléatoire 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 la somme de n variable aléatoire de Bernoulli, $n\overline{X}_n = X_1 + \cdots + X_n$, 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 $\mathcal{N}(np, \sqrt{np(1-p)})$, c.à.d. si Y suit la $\mathcal{B}(n,p)$ pour grand n alors

$$P(a \leq Y \leq b) \cong \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(y-np)^2}{2np(1-p)}} dx.$$

Chapitre 3

Inférence statistique

Dans le premier chapitre on a vu les notions de la statistique déscriptive, comme les caracteristiques, un échantillon, sa moyenne et sa variance. Dans le deuxième chapitre on a vu les notions de la théorie des probabilités, qui servent a modéliser les caractéristiques par des variables aléatoires (dans le cas ou les caractéristiques sont numériques). On peut résumer ceci dans le tableau suivant.

Statistique dèscriptive sondage	**Probabilité** modèle théorique
caractéristique numerique	variable aléatoire X
échantillon $x_1, \cdots, x_n$	échantillon aléatoire $X_1, \cdots, X_n$
moyenne de l'échantillon	moyenne empirique $\overline{X}_n$
variance de l'échantillon	variance de l'échantillon aléatoire S^2

Ici 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 aléatoire. On donnera la définition de S^2 plus bas.

Le but de l'inférence statistique est de déduire des observations, c.à.d. des sondages d'une caractéristique des individus d'un échantillon, la distribution de la caractéristique dans la population. Autrement dit, on veut déduire de l'échantillon (résultat d'un sondage) la loi mère de la variable aléatoire qui décrit la caractéristique.

Or, il y a deux problèmes :

1. Pour déterminer une loi de probabilité, il faut une quantité infinie d'informations. Or, un échantillon ne contient qu'un nombre fini de caractéristiques.
   Solution : On fait des hypothèses à priori sur la loi mère. Par exemple qu'il s'agit d'une loi normale $\mathcal{N}(\mu, \sigma)$. On se contente alors de déterminer ses paramètres μ et σ.
2. Tout est soumis au hasard. On ne pourrait donc pas prédire une loi avec une certitude parfaite.
   Solution : On fournit chaque résultat avec une marge d'erreur α. On dit aussi que le résultat est valable avec un niveau de confiance $1 - \alpha$.

3.1 Estimateurs

Un **estimateur** est une fonction $f(X_1, \cdots, X_n)$ de l'échantillon aléatoire $X_1, \cdots, X_n$, qui sert à estimer un paramètre θ de la loi mère. Ceci veut dire que l'ésperence de l'estimateur tend

vers le paramétre quand n tend vers $+\infty$

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

Mieux encore sont les **estimateurs sans biais**. Un estimateur sans biais satisfait

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

déjà pour n fini.

L'idée est alors que, pour estimer θ, on effectue un sondage (une réalisation $x_1, \cdots, x_n$ de l'échantillon aléatoire) et puis 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 $f(x_1, \cdots, x_n)$ donnera la valeur précise dans la limite où n tend vers $+\infty$.

3.1.1 L'estimateur pour la moyenne

Soit $X_1, \cdots, X_n$ un l'échantillon aléatoire, c.à.d. une suit de variable aléatoire indépendente de même loi. Soit μ la moyenne de cette loi mère. Soit

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

la moyenne empirique. On a vu que

$$\mathbb{E}(\overline{X}) = \mu.$$

La moyenne empirique $\overline{X}$ est alors un estimateur sans biais pour la moyenne de la loi mère. Pour estimer μ, on effectue un sondage $x_1, \cdots, x_n$ et puis calcule la moyenne de l'échantillon

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

$\bar{x}$ est une estimation pour μ.

3.1.2 L'estimateur pour la variance

Soit $X_1, \cdots, X_n$ un l'échantillon aléatoire, c.à.d. une suit de variable aléatoire indépendente de même loi. Soit σ^2 la variance de cette loi mère. On appelle

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

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

$$\mathbb{E}(S^2) = \sigma^2.$$

S^2 est donc un estimateur sans biais pour la variance de la loi mère. Pour estimer σ^2, on effectue un sondage $x_1, \cdots, x_n$ et puis calcule la variance **sans biais** de l'échantillon

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

$var(x)$ est une estimation pour σ^2. Donc $s := \sqrt{var(x)}$ est une estimation pour l'écart type σ.

3.1.3 Les lois associées aux estimateurs

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

$$\rho_{t(n)}(x) = c\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{3.1}$$

   où $c^{-1} = \int_{-\infty}^{+\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$. Pour grand n ($n \geq 30$) la loi de Student peut être approximée par la loi $\mathcal{N}(0,1)$.
3. Si on ne connaît pas la loi mère 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 ($n \geq 30$).

3.2 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

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

où $\rho_{\mathcal{N}(0,1)}$ est la densité de la loi normale (2.1). Voici quelques valeurs pour $z_{\frac{\alpha}{2}}$ en fonction de $1-\alpha$

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

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

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

où $\rho_{t(n)}$ est la densité de la loi de Student (3.1). Voici quelques valeurs pour $t(n)_{0.025}$ ($1-\alpha = 95\%$) en fonction de n

n	8	9	10	30	100
$t(n)_{0.025}$	2.31	2.26	2.23	2.04	1.98

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 en avance**. 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 en avance**. 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 sans biais, 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 ($n \geq 30$) 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 sans biais de l'échantillon.

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

$$p \in \left[\overline{x} - \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{x} + \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}} z_{\frac{\alpha}{2}}\right].$$

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

3.2.1 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 sur la base des données d'un sondage (un échantillon tiré au hasard).

On appelle H_0 (l'hypothèse nulle) l'hypothèse qu'on veut rejeter ("nullifier") à la base des observations. Typiquement H_0 est l'hypothèse qu'on a toujours acceptée, mais maintenant, sur la base des observations qu'on a fait, on y croit plus. On appelle H_1 l'hypothèse qu'on veut affirmer. H_1 est souvent l'opposé de H_0, mais ce n'est pas necessaire, il suffit que H_1 implique que H_0 soit fausse. Par exemple, H_0 pourrait être l'hypothèse que la temperature moyenne dans l'année est 15^o, une moyenne établie par des observations il y a 30 ans, mais on observe une augmentation. H_1 serait alors l'hypothèse que la moyenne de la temperature dans l'année est plus que 15^o.

Vu qu'une statistique ne donne pas des résultats avec certitude, il y a un risque d'erreur de rejeter H_0 à tort ou d'affirmer H_0 à tort. Comme H_0 est plutot l'hypothèse conservative (celle dont on a tourjours cru qu'elle est vraie) on préfère de minimiser le risque de rejeter H_0 à tort sans s'occuper du risque d'affirmer H_0 à tort. Concrètement cela veut dire que, pour une petite marge d'erreur $\alpha \in [0,1]$ (typiquement $\alpha = 5\%$) on voudrait que

$$\mathbb{P}(\text{rejeter } H_0 \text{ à tort}) \leq \alpha. \tag{3.2}$$

Modèle du test

On modélise la situation avec un échantillon aléatoire $X_1, \cdots, X_n$. L'hypothèse concerne une propriété de la loi mère de l'échantillon aléatoire, et un sondage (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 pour le paramètre, notamment une fonction

$f = f(X_1, \cdots, X_n)$ de l'échantillon aléatoire. Après un sondage on obtient alors un nombre $f(x_1, \cdots, x_n)$ qui donne une estimation pour le paramètre.

Pour décider si on affirme H_1, et donc rejette H_0, on désigne une région $C \subset \mathbb{R}$ telle que le critère d'affirmation de H_1 soit $f(x_1, \cdots, x_n) \in C$. C doit être choisi d'une telle manière que la probabilité que f appartient à C bienque H_0 soit vrai est au plus α,

$$\mathbb{P}(f \in C | H_0 \text{ est vrai}) \leq \alpha \tag{3.3}$$

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

Si l'hypothèse H_0 spécifie la loi mère, alors $\mathbb{P}(f \in C | H_0 \text{ est vrai})$ est la probabilité que $f(X_1, \cdots, X_n) \in C$ sous cette loi mère. (Si H_0 spécifie une famille de lois mères il faut calculer la probabilité que $f \in C$ avec une loi de la famille qui maximise cette probabilité.)

Test pour la moyenne μ

On veut tester une hypothèse sur la **moyenne** d'une caractéristique numérique de la population. 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 $f = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 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 H_0 est que la moyenne est μ_0. On distingue trois cas :

$$\begin{aligned} &\text{l'hypothèse } H_1 \text{ est que } \mu < \mu_0 \\ &\text{l'hypothèse } H_1 \text{ est que } \mu > \mu_0 \\ &\text{l'hypothèse } H_1 \text{ est que } \mu \neq \mu_0 \end{aligned}$$

Dans les deux premiers cas on parle d'une hypothèse unilatérale, dans le troisième d'une hypothèse bilatérale.

Considerons le dernier cas, c.à.d. l'hypothèse H_1 est que $\mu \neq \mu_0$. Vu que hypothèse est bilatérale on cherche une region C qui est symétrique autour de μ_0. Soit I l'intervalle de confiance de niveau $1 - \alpha$ pour la moyenne. Alors

$$\mathbb{P}(\overline{X} \in I | H_0 \text{ est vrai}) = 1 - \alpha$$

car H_0 est vrai veut dire que $E(\overline{X}) = \mu_0$. D'ou

$$\mathbb{P}(\bar{X} \in I^c | H_0 \text{ est vrai}) \leq \alpha.$$

On prend alors pour C le complémentaire de l'intervalle de confiance de niveau $1-\alpha$. Autrement dit, on affirme $\mu \neq \mu_0$ avec confiance $1-\alpha$ si le sondage donne une moyenne $\bar{x}$ qui est en dehors de l'intervalle de confiance de niveau $1 - \alpha$.

Dans les cas unilateraux on désigne pour C une région de la forme $]-\infty, \mu_0 - \delta]$ ou $[\mu_0 + \delta, +\infty[$, ou 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 les critères d'affirmation de H_1 (rejet de H_0) en terme de la moyenne $\bar{x}$ du sondage suivants :

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

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 sans biais de l'échantillon.

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 $\mathcal{T}(n-1)$. On obtient comme critère de rejet d'affirmation de H_1 (rejet de H_0) :

$$\begin{aligned}
&\text{On affirme } \mu < \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \leq \mu_0 - \frac{s}{\sqrt{n}} t(n-1)_\alpha \\
&\text{On affirme } \mu > \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad \overline{x} \geq \mu_0 + \frac{s}{\sqrt{n}} t(n-1)_\alpha \\
&\text{On affirme } \mu \neq \mu_0 \text{ avec confiance } 1-\alpha \quad \text{si} \quad |\overline{x} - \mu_0| \geq \frac{s}{\sqrt{n}} t(n-1)_{\frac{\alpha}{2}}
\end{aligned}$$

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

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

$$\begin{aligned}
&\text{si } H_1 \text{ est } \mu < \mu_0 \text{ alors} \quad && \text{p-val}(x_1, \cdots, x_n) = P(\overline{X} < \overline{x}) = \mathcal{F}_{\mathcal{N}(0,1)}(\bar{y}) \\
&\text{si } H_1 \text{ est } \mu > \mu_0 \text{ alors} \quad && \text{p-val}(x_1, \cdots, x_n) = P(\overline{X} > \overline{x}) = 1 - \mathcal{F}_{\mathcal{N}(0,1)}(\bar{y}) \\
&\text{si } H_1 \text{ est } \mu \neq \mu_0 \text{ alors} \quad && \text{p-val}(x_1, \cdots, x_n) = P(|\overline{X} - \mu_0| > |\overline{x} - \mu_0|) = 2\mathcal{F}_{\mathcal{N}(0,1)}(\bar{y})
\end{aligned}$$

où $\bar{y} = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$. 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 $\sigma = \sqrt{p_0(1-p_0)}$.

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 $\bar{z} = \frac{\overline{x}-\mu}{s/\sqrt{n}}$ (s est l'écart-type sans biais de l'échantillon) on obtient la formule

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

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