Statistique

Suivez le Guide

Brice Franke

Département de Mathématique
Université de Bretagne Occidentale
29200 Brest

1. Introduction au language probabilitste

- tout résultat scientifique doit être reproductible
- ceci nécessite l'existence d'un protocole de l'expérience
- certainnes expériences montrent de la variabilité
- dans ce cas les résultats sont dits aléatoires

Définition (épreuve et événement)

- *On appelle épreuve le protocole d'une expérience dont le résultat est aléatoire.*
- *Les différents résultats possibles d'une épreuve sont appelés les événements élémentaires.*
- *On notera Ω l'ensemble de tous les événements élémentaires d'une épreuve.*
- *Tout sous-ensemble $A \subset \Omega$ est appelé un événement.*
- *$\mathcal{P}(\Omega)$ dénotera l'ensemble de tous les événements de Ω.*

Langage et Notation

- *L'ensemble vide $\{\}$ est appelé l'événement impossible.*
- *L'ensemble Ω est dit événement certain.*
- *Pour deux événements $A, B \subset \Omega$ l'ensemble $A \cup B$ contient tous les événements élémentaires qui sont éléments de A ou de B.*
- *Pour deux événements $A, B \subset \Omega$ l'ensemble $A \cap B$ contient tous les événements élémentaires qui sont éléments de A et de B.*
- *Pour deux événements $A, B \subset \Omega$ l'ensemble $A \backslash B$ contient tous les événements élémentaires qui sont élements de A et pas de B.*
- *Pour un événement $A \in \mathcal{P}(\Omega)$ on pose $\bar{A} := \Omega \backslash A$.*
- *On dit que deux événements A et B sont incompatibles si ils satisfont $A \cap B = \{\}$.*

Un *axiome* est un principe de base non-démontrable.

La théorie des probabilités est basé sur trois axiomes que doit satisfaire l'attribution d'une probabilité $\mathbb{P}(A)$ à tout événement A d'une épreuve.

Définition (mesure de probabilité)

Une mesure de probabilité $\mathbb{P} : \mathcal{P}(\Omega) \to \mathbb{R}$ *doit satisfaire les trois axiomes suivants :*

- *Pour tout* $A \in \mathcal{P}(\Omega)$ *on a* $0 \leq \mathbb{P}(A) \leq 1$*;*
- *Pour tout* $A, B \in \mathcal{P}(\Omega)$ *incompatibles on a* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$*;*
- $\mathbb{P}(\Omega) = 1$.

La *modélisation mathématique* d'une épreuve consitste à choisir une probabilité $\mathbb{P}(A)$ pour chaque événement $A \in \mathcal{P}(\Omega)$ de lépreuve de façon conforme aux axiomes précédents.

Une fois les axiomes fixés on peut en déduire des règles de calculs générales appelés des *théorèmes*.

Théorème

Toute mesure de probabilité $\mathbb{P}$ *satisfait:*

- $\mathbb{P}(\{\}) = 0$;
- *Pour tout* $A \in \mathcal{P}(\Omega)$ *on a :* $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$;
- *Pour tout* $A, B \in \mathcal{P}(\Omega)$ *on a :*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

- *Pour tout* $A, B \in \mathcal{P}(\Omega)$ *avec* $A \subset B$ *on a*

$$\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A).$$

- *Pour tout* $A, B \in \mathcal{P}(\Omega)$ *avec* $A \subset B$ *on a* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Souvent les expériences ont des résultats numériques.

Définition (variable aléatoire)

Le résultat X (encore inconnu) d'une épreuve est appelé une variable aléatoire, s'il s'agit d'un résultat numérique.

- *Une variable aléatoire est dite discrète si le nombre de résultats possible est fini ou dénombrable.*
- *Une variable aléatoire est dite continue si elle prend ses valeurs dans un intervalle de la droite réelle.*

Il faut faire la différence entre l'information disponible sur la variable aléatoire avant et après l'expérience.

Définition (réalisation)

On appelle réalisation d'une variable aléatoire X les valeurs observées après la réalisation de l'expérience.

La modélisation d'une épreuve avec des résultats numériques peut être faite en fixant la loi de la variable aléatoire.

Définition (loi discrète)

La loi d'une variable aléatoire X discrète est l'ensemble des probabilités $\mathbb{P}(X = z_i)$ qui correspondent aux différents résultats possibles $x_1, x_2, x_3, \ldots$ etc..

Le cas particulier avec seulement les deux résultats possibles, zéro et un, a droit à une citation particulière :

Définition (loi de Bernoulli)

On dit qu'une variable aléatoire X suit une loi de Bernoulli de paramètre $p \in [0, 1]$ si on a

$$\mathbb{P}(X = 1) = p \qquad \text{et} \qquad \mathbb{P}(X = 0) = 1 - p.$$

Définition (espérance, variance et écart type)

Soit X une variable aléatoire discrète avec valeurs $x_1, \ldots, x_n$.

- *L'espérance mathématique de X est le nombre réel*

$$\mu_X = \mathbb{E}[X] := x_1 \mathbb{P}(X = x_1) + \ldots + x_n \mathbb{P}(X = x_n).$$

- *La variance de X est le nombre non-négatif*

$$\mathrm{Var}(X) := (x_1 - \mu_X)^2 \mathbb{P}(X = x_1) + \ldots + (x_n - \mu_X)^2 \mathbb{P}(X = x_n).$$

- *L'écart type de X est le nombre $\sigma_X := \sqrt{\mathrm{Var}(X)}$.*

Exemple: Si X suit une loi de Bernoulli de paramètre p alors :

$$\mathbb{E}[X] = p \qquad \text{et} \qquad \mathrm{Var}(X) = p(1-p).$$

Remarque: On peut faire des opérations algébriques sur les variables aléatoires X et Y pour obtenir des nouvelles variables aléatoires : cX, $X + Y$, $X \cdot Y$, X/Y etc.

Théorème

Soient X et Y deux variables aléatoires et c un nombre réel.

- $\mathbb{E}[cX] = c\mathbb{E}[X]$;
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$;
- $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$;
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

Remarque: On dit que l'espérance est linéaire alors que la variance est quadratique.

Attention: La formule $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ n'est pas valable en toutes circonstances. Il faut de l'indépendance.

- Dans certainnes situations deux épreuves n'ont aucune influence l'une sur l'autre. On parle alors d'indépendance.

Définition (indépendance d'événements)

Deux événements A et B sont dit indépendants si on a

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

- On peut prolonger la défintion sur les variables aléatoires

Définition (indépendance de variables aléatoires)

Deux variables aléatoires X et Y sont dites indépendants si pour tout choix de deux ensembles A et B on a

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

Théorème

Si X et Y sont indépendants alors on a :

- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$;
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Attention: Les deux opérations suivantes sont interdites:

- $\mathbb{E}[X/Y] = \mathbb{E}[X]/\mathbb{E}[Y]$;
- $\mathbb{E}[1/X] = 1/\mathbb{E}[X]$.

Théorème

Soit X une variable aléatoire avec $\mathbb{E}[X] = \mu_X$ et $\mathrm{Var}(X) = \sigma_X^2$. Alors pour tout choix de deux nombres réels a, b la variable aléatoire $Y = aX + b$ satisfait

$$\mathbb{E}[Y] = a\mu_X + b \qquad \text{et} \qquad \mathrm{Var}(Y) = a^2\sigma_X^2.$$

Le nombre de possibilités de sélectionner k sujets parmi n est

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

avec $0! := 1$ et $n! := n(n-1) \cdot .... \cdot 3 \cdot 2 \cdot 1$ pour $n \geq 1$.

Définition (loi binimiale)

Une variable aléatoire X suit une loi binomiale $\mathcal{B}(n,p)$ de paramètres n et p si elle prend les valeurs $0, 1, ..., n$ avec les probabilités

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \textit{pour} \quad 0 \leq k \leq n.$$

- la loi de Bernoulli de paramètre p est le cas particulier d'une loi binomiale de paramètres 1 et p.

Théorème

Soient $Y_k; 1 \leq k \leq n$ des variables aléatoires indépendantes qui suivent la même loi de Bernoulli de paramètre p. Alors leur somme $X := Y_1 + ... + Y_n$ suit une loi binomiale $\mathcal{B}(n, p)$.

Ce résultat nous donne la possibilité de calculer l'espérance et la variance d'une loi binomiale:

$$\mathbb{E}[X] = \mathbb{E}[Y_1 + ... + Y_n] = \mathbb{E}[Y_1] + ... + \mathbb{E}[Y_n] = np$$

$$\mathrm{Var}(X) = \mathrm{Var}(Y_1 + ... + Y_n) = \mathrm{Var}(Y_1) + ... + \mathrm{Var}(Y_n) = np(1-p)$$

Théorème

Soit X une variable aléatoire avec une loi binomiale $\mathcal{B}(n, p)$. Alors on a $\mathbb{E}[X] = np$ et $Var(X) = np(1-p)$.

Théorème (Siméon Denis Poisson 1781-1840)

Pour toute suite X_n de variables aléatoires avec lois $\mathcal{B}(n, p_n)$ de sorte que $\lim_{n\to\infty} np_n = \lambda$ on a

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \textit{pour tout } k \in \mathbb{N}.$$

- La loi limite du théorème porte le nom de son inventeur.

Définition (lois de Poisson)

On dit qu'une variable aléatoire X suit une loi de Poisson $\mathcal{P}(\lambda)$ de paramètre λ si elle prend ses valeurs dans les nombres naturels $\mathbb{N}$ avec probabilités :

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \textit{pour tout } k \in \mathbb{N}.$$

- soit X le nombre de fois qu'un certain événement de probabilité p est apparu dans n répétitions d'une épreuve;
- la loi $\mathcal{B}(n, p)$ modélise le nombre X si les épreuves sont indépendantes;
- sous certaines conditions la loi $\mathcal{P}(\lambda)$ est un bon modèle pour le nombre total X.

Théorème

Si X suit une loi de Poisson de paramètre λ, alors

$$\mathbb{E}[X] = \lambda \qquad \textit{et} \qquad \mathrm{Var}(X) = \lambda$$

- pour obtenir une bonne approximation de $\mathcal{B}(n, p)$ par $\mathcal{P}(\lambda)$ il est nécessaire que np et $np(1-p)$ soient proches de λ;
- ceci est le cas si p est petit et n grand.

Application en prévision des catastrophes:

- chaque année il y a un grand nombre de tremblements de terre en France (souvent très faibles)
- sur une année la probabilité d'enregistrer un tremblement de terre d'une puissance supérieure à 6 en France est faible
- en moyenne il y a un tremblement de terre de puissance supérieure à 6 en France tous les dix ans;
- on peut donc dire que $\mathbb{E}[X] = 1/10$.
- le nombre X des tremblement de terre supérieur à 6 en France suit une loi de Poisson de paramètre $\lambda = 10$.
- la probabilité d'enregistrer plus de deux tremblements de terre de puissance supérieure à 6 en 2014 est donc

$$\begin{aligned} \mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\ &= 1 - e^{-10} - 10e^{-10} \end{aligned}$$

Définition (loi uniforme discrète)

On dit qu'une variable aléatoire suit une loi uniforme sur l'ensemble $\{x_1, ..., x_n\}$ *si pour tout* $k \in \{1, ..., n\}$ *on a* $\mathbb{P}(X = x_k) = 1/n$.

- la loi uniforme modélise des épreuves où tous les résultats ont la même probabilité.

Définition (loi géométrique)

On dit qu'une variable aléatoire X *suit une loi géométrique* $\mathcal{G}(p)$ *de paramètre* p *si elle prend ces valeurs dans* $\mathbb{N}$ *si on a* $\mathbb{P}(X = k) = p(1-p)^k$ *pour tout* $k \in \mathbb{N}$.

- $\mathcal{G}(p)$ modélise le nombre de répétitions indépendantes d"épreuves de Bernoulli au paramètre p à réaliser pour voir apparaître le premier succès.

- la loi d'une variable aléatoire continue est déterminée par un poids attribué à chaque résultat possible de l'épreuve

Définition (densité et fonction de répartition)

La densité d'une variable aléatoire continue X est une fonction positive $f_X(x)$ de sorte que pour tout choix de $a < b$ on a

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

On appelle fonction de répartition de X la fonction

$$F_X(y) := \mathbb{P}(X \leq y) = \int_{-\infty}^{x} f_X(y)dy.$$

- on a $\lim_{y\to\infty} F_X(y) = 1$ et $\lim_{y\to-\infty} F_X(y) = 0$;
- la fonction de répartition est croissante; i.e.: $F_X(y) \leq F_X(z)$ pour tout choix de $y \leq z$;
- on a la formule : $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$.

Définition (espérance et variance)

Soit X une variable aléatoire continue avec densité $f(x)$.

- *L'espérance mathématique de X est le nombre réel*

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

- *La variance de X est le nombre réel non-négatif*

$$\mathrm{Var}(X) := \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx.$$

- *L'écart type de X est le nombre $\sigma_X := \sqrt{\mathrm{Var}(X)}$.*

- L'espérance et la variance de variables aléatoires continues satisfont les mêmes règles de calculs que l'on a déjà vues dans le cas de variables aléatoires discrètes.

Définition (loi normale)

On dit qu'une variable aléatoire X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ de paramètres μ et σ^2 si elle a la densité

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

- La loi normale est la loi la plus utilisée en statistique.

Théorème

Si X suit une loi normale de paramètre μ et σ^2 alors on a

$$\mathbb{E}[X] = \mu \qquad \textit{et} \qquad \mathrm{Var}(X) = \sigma^2.$$

- Il est impossible de donner une formule explicite pour la fonction de répartition F_X de la loi normale.
- Il y a des tables dans lesquelles on peut trouver les valeurs de la fonction de répartition d'une loi normale $\mathcal{N}(0, 1)$.

Langage et Notation:

- On appelle la loi $\mathcal{N}(0,1)$ la loi normale centrée réduite.
- $\Phi(x)$ denote la fonction de répartition de $\mathcal{N}(0,1)$
- $X \sim \mathcal{L}$ est une notation pour dire que X suit la loi $\mathcal{L}$

Théorème

- *Si* $X \sim \mathcal{N}(\mu, \sigma^2)$ *alors* $Y := \frac{1}{\sigma}(X - \mu) \sim \mathcal{N}(0,1)$.
- *Si* $X \sim \mathcal{N}(0,1)$ *alors* $Y := \sigma(X + \mu) \sim \mathcal{N}(\mu, \sigma^2)$.
- *si* $X \sim \mathcal{N}(0,1)$ *alors* $Y := -X \sim \mathcal{N}(0,1)$.

- la fonction de répartition d'un X avec loi $\mathcal{N}(\mu, \sigma^2)$ est

$$\Phi_{\mu,\sigma^2}(x) = \mathbb{P}(X \leq x) = \mathbb{P}(Z \leq (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma)$$

- $\Phi(-x) = \mathbb{P}(Z \leq -x) = \mathbb{P}(-Z > x) = \mathbb{P}(Z > x) = 1 - \Phi(x)$

Théorème

Si $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *et* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ *sont indépendants alors* $Z := X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

- L'importance de la loi normale en statistique est basée sur le théorème central limite suivant:

Théorème (Pierre-Simon Laplace 1749-1827)

Soit Y_n *une suite de variables aléatoires de lois binomiales* $\mathcal{B}(n, p)$ *alors on a si* $n \to \infty$

$$\mathbb{P}\left(a \leq \frac{1}{\sqrt{np(1-p)}}(Y_n - np) \leq b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx.$$

- ce résultat permet de calculer des probabilités sous des lois binomiales lorsque n est grand

- sous la loi normale on voit aussi des résultats négatifs;
- on trouve dans la nature des phénomènes où les seuls résultats sont positifs (p.e.: durée de vie d'un individu).

Définition (loi exponentielle)

On dit qu'une variable aléatoire X suit une loi exponentielle $\mathcal{E}(\lambda)$ de paramètre λ si elle a la densité

$$f_X(x) = \begin{cases} 0 & \text{pour } x \leq 0 \\ \lambda e^{-\lambda x} & \text{pour } x > 0. \end{cases}$$

Théorème

Si X suit une loi exponentielle $\mathcal{E}(\lambda)$ alors on a

$$\mathbb{E}[X] = 1/\lambda \qquad \text{et} \qquad \mathrm{Var}(X) = 1/\lambda^2.$$

- dans certaines situations les résulttats d'une épreuve se trouvent dans un intervalle $[A, B]$

Définition (loi uniforme continue)

On dit qu'une variable aléatoire X suit une loi uniforme $\mathcal{U}[A, B]$ sur un intervalle $[A, B]$ si elle a une densité de la forme

$$f_X(x) = \begin{cases} 0 & \text{si } x < A \text{ ou } x > B \\ \frac{1}{B-A} & \text{si } x \in [A, B]. \end{cases}$$

- la loi $\mathcal{U}[A, B]$ donne le même poids à chaque point de $[A, B]$

Théorème

Si X suit une loi uniforme $\mathcal{U}[A, B]$ alors on a

$$\mathbb{E}[X] = \frac{B + A}{2} \qquad \text{et} \qquad \mathrm{Var}(X) = \frac{(B - A)^2}{12}.$$

2. L'estimation statistique

- La première étape d'une modélisation consiste à trouver une loi pour décrire la répartition des observations.
- Souvent ces lois contiennent des paramètres.
- Ces paramètres doivent être mis en accord avec une série d'observations faites sur l'expérience.
- On effectue alors une estimation des paramètres.

Définition

Un échantillon de taille n d'une variable aléatoire X est composé de n répétitions indépendantes de l'épreuve qui mène au résultat décrit par X. L'échantillon est donc un n-uplet $(X_1, ..., X_n)$ *de variables aléatoires indépendantes dont les composantes on toutes la même loi que X.*

Remarques:

- la loi de la variable aléatoire X qui est à la base de l'échantillon $(X_1, ..., X_n)$ est applelée la loi parente
- on différencie entre le n-uplet de variables aléatoires et le n-uplet des observations après réalisation de l'expérience

Définition (échantillon de données)

On appelle échantillon des données le n-uplet $(x_1, ..., x_n)$ des valeurs observées après réalisation de n répétitions de l'épreuve qui est décrite par la variable aléatoire X.

Remarques:

- l'échanillon est un objet imaginaire qui dépend de notre choix de la modélisation de l'expérience
- l'échantillon des données est un objet réel obtenu par la réalisation d'une expérience

Définition (moyenne arithmétique)

- *Soit* $(X_1, ..., X_n)$ *un échantillon de variables aléatoires. On appelle la variable aléatoire* $M_n := \frac{1}{n}(X_1 + ... + X_n)$ *la moyenne arithmétique de l'échantillon.*
- *Soit* $(x_1, ..., x_n)$ *un échantillon de données. On appelle le nombre réel* $m_n := \frac{1}{n}(x_1 + ... + x_n)$ *la moyenne empirique des données.*

Théorème

Soit M_n *la moyenne arithmétique d'un échantillon* $(X_1, ..., X_n)$ *de taille n d'une variable aléatoire X. Alors on a*

$$\mathbb{E}[M_n] = \mathbb{E}[X] \qquad \textit{et} \qquad \mathrm{Var}(M_n) = \frac{1}{n}\mathrm{Var}(X)$$

- On voit que la variance décroît avec la taille de l'échantillon

Définition (estimateur de paramètre)

Soit $(X_1, ..., X_n)$ un échantillon d'une variable aléatoire dont la loi contient un paramètre θ. On appelle estimateur de θ toute formule basée sur $(X_1, ..., X_n)$ qui est supposée faire une prévision sur θ.

Exemple:

- Soit $(X_1, ..., X_n)$ un échantillon avec loi parente de type Bernoulli $\mathcal{B}(p)$. Le paramètre $\theta = p$ est inconnu;
- la moyenne arithmétique $M_n = \frac{1}{n}(X_1 + ... + X_n)$ est alors un estimateur de p;
- le minimum de l'échantillon $X_{\text{inf}} := \min\{X_1, ..., X_n\}$ est aussi un estimateur de p.
- X_{inf} est un estimateur moins bon que M_n pour estimer p.

- On doit trouver un moyen de caractériser les estimateurs performants.
- Pour toute variable aléatoire X dont la loi dépend d'un paramètre θ nous utilisons la notation $\mathbb{E}_\theta[X]$ pour indiquer que l'espérance de X dépend aussi de θ.

Définition (biais d'un estimateur)

Soit $T(X_1, ..., X_n)$ un estimateur d'un paramètre θ.

- *On appelle biais de l'estimateur le nombre réel $b(T) := \mathbb{E}_\theta[T(X_1, ..., X_n)] - \theta$.*
- *On dit que l'estimateur $T(X_1, ..., X_n)$ est sans biais si $\mathbb{E}_\theta[T(X_1, ..., X_n)] = \theta$.*

- Le biais mesure l'écart entre une estimation moyenne exprimée par $\mathbb{E}_\theta[T(X_1, ..., X_n)]$ et la vraie valeur θ.

Cas particulier:

- Soit M_n la moyenne arithmétique d'un échantillon de variables aléatoires du type Bernoulli de paramètre p
- On a

$$\mathbb{E}[M_n] = \frac{1}{n}(\mathbb{E}[X_1] + ... + \mathbb{E}[X_n]) = \frac{1}{n}(p + ... + p) = p.$$

- Donc M_n est un estimateur sans biais du paramètre p.
- De plus on a pour l'estimateur $X_{\text{inf}} = \min\{X_1, ..., X_n\}$

$$\begin{aligned} \mathbb{E}[X_{\text{inf}}] &= 1\mathbb{P}(X_{\text{inf}} = 1) + 0\mathbb{P}(X_{\text{inf}} = 0) \\ &= \mathbb{P}(X_1 = 1, ..., X_n = 1) = p^n. \end{aligned}$$

- Si $n > 1$ alors $\mathbb{E}[X_{\text{inf}}] = p^n < p$
- Alors X_{inf} est un estimateur du paramètre p avec biais.

Définition (estimateur convergent)

Soit $T(X_1, ..., X_n)$ un estimateur d'un paramètre θ On dit que l'estimateur est convergent si l'écart quadratique moyen $\mathbb{E}_\theta \left[T(X_1, ..., X_n) - \theta)^2\right]$ tend vers zéro lorsque n tend vers infini.

- Si $T(X_1, ..., X_n)$ est un estimateur sans biais, alors on a $\theta = \mathbb{E}_\theta[T(X_1, ..., X_n)]$ et l'écart quadratique moyen est égal à la variance

$$\begin{aligned} & \mathbb{E}_\theta \left[T(X_1, ..., X_n) - \theta)^2\right] \\ = \quad & \mathbb{E}_\theta \left[T(X_1, ..., X_n) - \mathbb{E}_\theta[T(X_1, ..., X_n)])^2\right] \\ = \quad & \mathrm{Var}_\theta(T(X_1, ..., X_n)) \end{aligned}$$

- La variance est souvent plus facile à calculer.

- Soit $(X_1, ..., X_n)$ un échantillon avec loi parentale du type Bernoulli $\mathcal{B}(p)$. Le paramètre p est inconnu.
- Dans ce cas la moyenne arithmétique devient alors la proportion des uns dans l'échantillon

$$M_n = \frac{1}{n}\text{Card}\{i \in \{1, ..., n\} : X_i = 1\}.$$

- M_n est un estimateur sans biais pour l'estimation de p

$$\begin{aligned}
\mathbb{E}_\theta[(M_n - \theta)^2] &= \text{Var}(M_n) \\
&= \text{Var}\left(\frac{1}{n}(X_1 + ... + X_n)\right) \\
&= \frac{1}{n^2}\left(\text{Var}(X_1) + ... + \text{Var}(X_n)\right) \\
&= \frac{1}{n}p(1-p) \longrightarrow 0 \quad \text{si } n \to \infty.
\end{aligned}$$

.

- Donc M_n est un estimateur sans biais convergent du paramètre p d'une loi de Bernoulli.
- Nous voulons estimer la proportion p des individus dans la population globale qui sont porteurs d'une caractéristique.
- n individus sont tirés au sort pour vérification.
- Dans l'échantillon chaque 1 représente un individu porteur et chaque 0 représente un individu sans la caractéristique.
- Pour un échantillion de données $\{x_1, ..., x_n\}$, composé de 1 et de 0 uniquement, on introduit la proportion empirique : $\hat{p} := \frac{1}{n}\text{Card}\{i \in \{1, ..., n\} : X_i = 1\}$.
- Chaque tirage peut être modélisé par une loi Bernoulli $\mathcal{B}(p)$
- Sous cette modélisation la proportion $\hat{p}$ des individus porteurs dans l'échantillon suit la même loi que M_n.
- $\hat{p}$ est donc une bonne estimation de la proportion p dans la population globale.

- Souvent on veut estimer l'espérance et la variance d'une variable aléatoire X.

Théorème

Soit $(X_1, ..., X_n)$ un échantillon dont la loi parentale est égale à celle de X.

- $M_n = \frac{1}{n}(X_1 + ... + X_n)$ *est un estimateur convergent sans biais de l'espérance de X.*
- $S_n^2 = \frac{1}{n-1}((X_1 - M_n)^2 + ... + (X_n - M_n)^2)$ *est un estimateur convergent sans biais de la variance de X.*

- En pratique on utilise donc la moyenne empririque m_n d'un échantillon de données $(x_1, ..., x_n)$ pour estimer $\mathbb{E}[X]$
- Pour estimer $\mathrm{Var}(X)$ on utilise la variance empirique de l'échantillon de données

$$s_n^2 := \frac{1}{n-1}((x_1 - m_n)^2 + ... + (x_n - m_n)^2)$$

- Dans certains cas au lieu de donner une valeur précise pour θ on préfère donner une fourchette de valeurs.
- Une loi a été fixée pour modéliser l'épreuve qui sera répétée.
- Cette loi contient toujours un paramètre θ inconnu.
- Un échantillon de données $(x_1, ..., x_n)$ a été observé sur n répétitions de l'épreuve.

Définition (intervalle de confiance)

L'intervalle de confiance au degré de confiance γ d'un paramètre θ est un intervalle $[\beta_0, \beta_1]$ de sorte que si la valeur de θ se trouvait dans cet intervalle alors la probabilité d'observer l'échantillon des données $(x_1, ..., x_n)$ serait supérieure à γ.

- Souvent en pratique on fixe le degré de confiance à 95%.

Langage et Notation

- *Pour $y \in [0,1]$ notons u_y le nombre réel de sorte que $\Phi(u_y) = y$. On utilise la table de $\mathcal{N}(0,1)$ pour trouver u_y.*

- Soit $(X_1, ..., X_n)$ un échantillon avec loi parentale du type Bernoulli $\mathcal{B}(p)$.
- La variable $Y_n := nM_n = X_1 + ... + X_n$ suit une loi $\mathcal{B}(n,p)$ donc le théorème de Laplace donne :

$$\begin{aligned} & \mathbb{P}\left(p - u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq M_n \leq p + u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right) \\ = & \mathbb{P}\left(-u_{\frac{\gamma+1}{2}} \leq \frac{1}{\sqrt{np(1-p)}}(Y_n - p) \leq u_{\frac{\gamma+1}{2}} \right) \\ \simeq & \Phi\left(u_{\frac{\gamma+1}{2}}\right) - \Phi\left(-u_{\frac{\gamma+1}{2}}\right) = 2\Phi\left(u_{\frac{\gamma+1}{2}}\right) - 1 = \gamma \end{aligned}$$

- On vient de voir que M_n tombe avec probabilité γ dans l'intervalle de pari $\left[p - u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.
- Donc on peut supposer que $\hat{p}$ tombe avec probabilité γ dans l'intervalle $\left[p - u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
- On a que: $\hat{p} \in \left[p - u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est équivalent à $p \in \left[\hat{p} - u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}, \hat{p} + u_{\frac{\gamma+1}{2}} \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
- Finalement on remplace p par $\hat{p}$ pour obtenir l'intervalle de confiance au degré de confiance γ pour l'estimation de p :

$$\mathcal{I}_\gamma = \left[\hat{p} - u_{\frac{\gamma+1}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + u_{\frac{\gamma+1}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$$

- On calcule l'intervalle de confiance pour l'estimation de l'espérance $\mathbb{E}[X]$ dans différentes situations.

Théorème

Si la variance σ_X^2 de X est connue alors l'intervalle de confiance au degré de confiance γ pour l'estimation de l'espérance μ_X de X est $\mathcal{I}_\gamma := \left[m_n - u_{\frac{\gamma+1}{2}} \sqrt{\frac{\sigma_X^2}{n}}, m_n + u_{\frac{\gamma+1}{2}} \sqrt{\frac{\sigma_X^2}{n}} \right].$

- Si σ_X^2 est inconnu alors on la remplace par la variance empririque de l'échantillon : $s_n^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - m_n)^2$.

Théorème

Si la variance de X est inconnue alors l'intervalle de confiance au degré de confiance γ pour l'estimation de l'espérance μ_X de X est $\mathcal{I}_\gamma := \left[m_n - u_{\frac{\gamma+1}{2}} \sqrt{\frac{s_n^2}{n}}, m_n + u_{\frac{\gamma+1}{2}} \sqrt{\frac{s_n^2}{n}} \right].$

- Si la taille de l'échantillon est trop faible ($n \leq 30$), alors l'approximation de σ^2 par s_n^2 n'est pas bonne et l'on doit modifier la constante devant $\sqrt{\frac{s_n^2}{n}}$.

Définition (loi de Student-Fisher)

Une variable aléatoire X suit une loi de Student-Fisher avec n degrés de liberté $\mathcal{T}_n$ si elle a la densité

$$f_n(x) := \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

- Lorsque n tend vers l'infini la densité de Student Fisher converge vers la densité de la loi normale centrée réduite :

$$\lim_{N\to\infty} f_n(x) = \varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{pour tout } x \in \mathbb{R}.$$

Théorème

Soit $(X_1, ..., X_n)$ *un échantillon dont la loi parentale est* $\mathcal{N}(0, 1)$. *Alors*

$$Y_n := \frac{M_n}{\sqrt{\frac{S_n}{n}}} \sim \mathcal{T}_{n-1}.$$

- Supposons $X \sim \mathcal{T}_n$. Il existe un nombre unique $t_n(\gamma)$ de sorte que $\mathbb{P}(|X| > t_n(\gamma)) = \gamma$.

Théorème

Si la variance de X est inconnue et la taille de l'échantillon est inférieure à 30, *alors l'intervalle de confiance au degré de confiance* γ *pour l'estimation de l'espérance* μ_X *de X est*

$$\mathcal{I}_\gamma := \left[m_n - t_{n-1}(1-\gamma)\sqrt{\frac{s_n^2}{n}}, m_n + t_{n-1}(1-\gamma)\sqrt{\frac{s_n^2}{n}} \right].$$

3. La notion du test statistique

- Le test statistique vise à donner une règle permettant de décider si on peut rejeter une hypothèse sur la base de données relevées sur un échantillon.

Définition (hypothèse)

- *L'hypothèse nulle $\mathcal{H}_0$ est l'hypothèse dont on cherche à savoir si elle peut être rejetée grâce aux observations dont on dispose.*
- *L'hypothèse alternative $\mathcal{H}_1$ est l'hypothèse en concurrence avec l'hypothèse nulle.*

- Le principe du test de Neyman Pearson est de définir avant l'expérience une zone de rejet de l'hypothèse $\mathcal{H}_0$.
- La probabilité de tomber dans cette zone de rejet doit être faible lorsque l'hypothèse $\mathcal{H}_0$ est satisfaite.

Définition (zone de rejet)

La zone de rejet est l'ensemble des valeurs expérimentales pour lesquelles on décide à l'avance de rejeter l'hypothèse $\mathcal{H}_0$.

- Il est important de controller le risque de se tromper en rejetant l'hypothèse $\mathcal{H}_0$.

Définition (risque de première espèce)

On appelle risque de première espèce la probabilité que l'on a de rejeter l'hypothèse $\mathcal{H}_0$ avec le test statistique employé, quand l'hypothèse $\mathcal{H}_0$ est vraie.

- Au lieu de risque de première espèce on dit souvent le *seuil du test*.
- Souvent on fixe le seuil à 5% et ensuite on cherche une zone de rejet de sorte que le risque de première espèce est 5%.

Exemple: Tester la proportion dans une population

- On veut tester l'hypothèse nulle $\mathcal{H}_0$ qu'une proportion $p = p_0$ d'une population a une caractéristique A.
- L'hypothèse alternative $\mathcal{H}_1$ est $p \neq p_0$.
- Dans un échantillon de n individus de la population on détermine la proportion emprique $\hat{p}$ des individus qui montrent la caractéristique A.
- Le test au seuil α consiste à rejeter l'hypothèse si

$$|\hat{p} - p_0| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Rappellons que u_y est le nombre de sorte que $\phi(u_y) = y$.
- Souvent on utilise $\alpha = 0.05$ et dans ce cas on a

$$u_{1-\frac{\alpha}{2}} = u_{1-0.025} = u_{0.975} = 1.96.$$

- Lorsque le test tombe dans la zone de rejet on dit que le *test est significatif*.
- Dans ce cas les données mènent à rejeter l'hypthèse $\mathcal{H}_0$.
- Sinon on dit que le *test n'est pas significatif*.
- Le fait que le test est non-significatif n'est pas une garantie pour l'exactitude de l'hypothèse $\mathcal{H}_0$.

Définition (degré de signifcation)

Dans le cas d'un test significatif (on a rejeté $\mathcal{H}_0$) on appelle degré de signification le plus petiti seuil α_0 qu'aurait pu avoir le test en permettant toujours de rejeter $\mathcal{H}_0$.

- Dans notre test $\mathcal{H}_0 : p = p_0$ contre $\mathcal{H}_1 : p \neq p_0$ le niveau de signification est

$$\delta = 2 - 2\Phi\left(\frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)}}\right).$$

Définition (puissance)

La puissance d'un test est la probabilité avec laquelle on rejete l'hypothèse $\mathcal{H}_0$ alors que celle-ci n'est pas valide.

- Une fois le risque de première espèce fixé la puissance sert à mesurer la qualité d'un test.
- Normalement la puissance du test grandit lorsque l'on augmente la taille de l'échantillon

Définition (risque de seconde espèce)

On appelle risque de seconde espèce la probabilité de ne pas rejeter l'hypothèse $\mathcal{H}_0$ grace au test statistique alors que l'hypothèse $\mathcal{H}_0$ n'est pas valide.

- Si l'on utilise les notations usuelles pour la puissance P et le risque de seconde espèce β alors on a $P = 1 - \beta$.

Exemple: Tester l'égalité de deux proportions

- Soient U et V deux populations.
- On veut tester si la caractéristique A est présente avec la même proportion dans les population U et V.
- On a à notre disposition deux échantillons de tailles n_U et n_V issues des populations U et V.
- Soient p_U et p_V les proportions dans les deux populations.
- Soient $\hat{p}_U$ et $\hat{p}_V$ les proportions dans les deux échantillons.
- On veut tester $\mathcal{H}_0 : p_U = p_V$ contre $\mathcal{H}_1 : p_U \neq p_V$.
- Nous rejetons l'hypothèse au seuil α si

$$\left|\hat{p}_U - \hat{p}_V\right| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_U(1-\hat{p}_U)}{n_U} + \frac{\hat{p}_V(1-\hat{p}_V)}{n_V}}.$$

- Dans beaucoup de situations nous avons à faire à des variables aléatoires qui suivent des lois normales $\mathcal{N}(\mu, \sigma^2)$.
- Nous voulons faire des tests sur la moyenne μ.
- Soient m_n la moyenne arithmétique et s_n^2 la variance empirique de l'échantillon $(x_1, ..., x_n)$ des données.
- Si σ^2 est connu nous rejetons l'hypothèse $\mathcal{H}_0 : \mu = \mu_0$ contre l'alternative $\mathcal{H}_1 : \mu \neq \mu_0$ au seuil α si

$$|m_X - \mu_0| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}.$$

- Si σ^2 n'est pas connu nous rejetons $\mathcal{H}_0 : \mu = \mu_0$ contre l'alternative $\mathcal{H}_1 : \mu \neq \mu_0$ au seuil α si

$$|m_X - \mu_0| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{s_n^2}{n}}.$$

- Si la taille de l'échantillon est trop petite, il faut utiliser la loi de Student Fisher $\mathcal{T}_{n-1}$ avec $n-1$ degrés de liberté.
- Si σ^2 est inconnu et $n \leq 30$ nous rejetons l'hypothèse $\mathcal{H}_0 : \mu = \mu_0$ contre l'alternative $\mathcal{H}_1 : \mu \neq \mu_0$ au seuil α si

$$|m_X - \mu_0| > t_{n-1}(\alpha)\sqrt{\frac{s_n^2}{n}}.$$

- Dans certainnes situations nous voulons tester une hypothèse *non-symétrique*.
- Nous rejetons au seuil α l'hypothèse $\mathcal{H}_0 : p \leq p_0$ contre l'alternative $\mathcal{H}_1 : p > p_0$ si

$$\hat{p} - p_0 > u_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

- Si σ^2 est connu nous rejetons l'hypothèse $\mathcal{H}_0 : \mu \leq \mu_0$ contre l'alternative $\mathcal{H}_1 : \mu > \mu_0$ au seuil α si

$$m_X - \mu_0 > u_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}.$$

- Si σ^2 est inconnu nous rejetons l'hypothèse $\mathcal{H}_0 : \mu \leq \mu_0$ contre l'alternative $\mathcal{H}_1 : \mu > \mu_0$ au seuil α si

$$m_X - \mu_0 > u_{1-\alpha}\sqrt{\frac{s_n^2}{n}}.$$

- Il y a aussi un test utilisant la loi de Student Fisher pour l'hypothèse non-symétrique lorsque σ^2 est inconnu et $n \leq 30$.

- Nous voulons tester si des données $(x_1, ..., x_n)$ suivent une loi préscrite.

Définition (loi du chi-deux)

Soit $(Z_1, ..., Z_n)$ un échantillon avec loi parente $\mathcal{N}(0, 1)$ et soit $K_n := Z_1^2 + ... + Z_n^2$. La loi χ_n^2 de K_n est appelée loi du chi-deux à n degrés de liberté.

- *Notons $\chi_n^2(y)$ le nombre qui satisfait $\mathbb{P}(K_n > \chi_n(y)) = y$.*

- Il existe des tables pour la fonction de répartition de χ_n^2.
- Supposons une population avec k catégories d'individus que l'on rencontre avec proportions $p_1, ..., p_k$.
- On a : $p_1 + ... + p_k = 1$.
- Soient $n_1, ..., n_k$ les nombres d'individus issus des catégories $1, ..., k$ dans un échantillon de taille n.

- On a $n_1 + ... + n_k = n$.
- Avant l'expérience les nombres de sujets $N_1, ..., N_k$ issus des diffférentes catégories suivent des lois binomiales avec paramètres respectifs $(n_1, p_1), .., (n_k, p_k)$.

Théorème

Si les nombres $np_1, ..., np_k$ sont grands, alors la variable aléatoire $Q := \sum_{i=1}^{n} \frac{(N_i - np_i)^2}{np_i}$ suit approximativement χ^2_{k-1}.

- Nous rejetons l'hypothèse $\mathcal{H}_0 : p_1 = p_1^0, ..., p_k = p_k^0$ contre l'alternative $\mathcal{H}_1$: il existe un i avec $p_i \neq p_i^0$ au seuil α si

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} > \chi^2_{k-1}(\alpha).$$

4. Étude de la dépendance

Définition (loi jointe)

Soient X et Y deux variables aléatoires discrètes avec valeurs respectives dans $\{x_1, ..., x_n\}$ et $\{y_1, ..., y_m\}$. On appelle la famille de probabilités

$$p_{ij} = \mathbb{P}(X = x_i, Y = y_j); \quad i = 1, ..., n, \ j = 1, ..., m$$

la loi jointe du couple X et Y.

- on retrouve les lois marginales dans la loi jointe du couple

Théorème

Soit (X, Y)un couple de variables aléatoires avec loi jointe p_{ij}; $i = 1, ..., n, \ j = 1, ..., m$. Alors on a

- $\mathbb{P}(X = x_i) = p_{i1} + ... + p_{im}$ *et* $\mathbb{P}(Y = y_j) = p_{1j} + ... + p_{nj}$.

Définition

Soient X et Y deux variables aléatoires discrètes avec loi jointe p_{ij}; $i = 1, ..., n,\ j = 1, ..., m$.

- $\sigma_{XY} = \mathrm{Cov}(X, Y) := \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - \mathbb{E}[X])(y_j - \mathbb{E}[Y])$ *est appelée la covariance de X et Y.*
- $\rho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ *est appelée la corrélation de X et Y.*

- la corrélation est une mesure de dépendance

Théorème

Pour deux variables aléatoires discrètes X et Y on a :

- $-1 \leq \rho_{XY} \leq 1$;
- *si X et Y sont indépendants alors $\rho_{XY} = 0$*

- Dans le cas d'un couple de variables aléatoires continues, les sommes deviennent des intégrales.

Définition (densité du couple)

Soient X et Y deux variables aléatoires continues.

- *$F_{X,Y}(x,y) := \mathbb{P}(X \leq x, Y \leq y)$ est appelée la fonction de répartition jointe du couple*
- *Une fonction non-négative $f(x,y)$ est appelée densité du couple si elle satisfait $F_{X,Y}(u,v) = \int_{-\infty}^{u}\int_{-\infty}^{v} f(x,y)dxdy$.*
- *La covariance de X et Y est* $\mathrm{Cov}(X,Y) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathbb{E}[X])(y-\mathbb{E}[Y])f(x,y)dxdy$.
- *La corrélation de X et Y est* $\rho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$

- Covariance et corrélation ont les mêmes propriétés que dans le cas d'un couple de variables discrètes.

Définition (loi normale bivariée)

On dit que le couple de variables aléatoires suit une loi normale bivariée si la densité du couple est

$$f(x,y) := \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}+2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}.$$

On utilise alors la notation : $(X,Y) \sim \mathcal{N}\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}\right)$.

- La loi normale bivariée définie des variables aléatoires normales avec de la dépendance.

Théorème

Si on a $(X,Y)) \sim \mathcal{N}\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}\right)$, *alors :*

- $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *et* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$;
- *la corrélation de* X *et* Y *est* ρ_{XY}.

Théorème

Si $(X, Y)) \sim \mathcal{N}\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}\right)$, *alors :* $\rho_{XY} = 0$ *implique que X et Y sont indépendants.*

- Soit $((x_1, y_1), ..., (x_n, y_n))$ un échantillon de données;
- $m_X := \frac{1}{n}(x_1 + ... + x_n)$ et $m_Y := \frac{1}{n}(y_1 + ... + y_n)$;
- $s_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)^2$ et $s_Y^2 := \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_Y)^2$;
- on définit la covariance emprique

$$s_{XY} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y);$$

- on définit la corrélation empirique

$$r_{XY} := \frac{\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_X)^2 \sum_{i=1}^{n}(y_i - m_Y)^2}}.$$

Théorème

Soit $((X_1, Y_1), ..., (X_n, Y_n))$ *un échantillon de variables aléatoires bivariées.*

- $\sum_{i=1}^{n}(X_i - \mu_X)(Y_j - M_Y)$ *est un estimateur sans biais de la covariance de* X *et* Y*.*
- *La quantité*

$$R := \frac{\sum_{i=1}^{n}(X_i - M_X)(Y_i - M_Y)}{\sqrt{\sum_{i=1}^{n}(X_i - M_X)^2 \sum_{i=1}^{n}(Y_i - M_Y)^2}}.$$

  est un éstimateur sans biais de la corrélation ρ_{XY}
- *Si la loi parente est* $\mathcal{N}\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}\right)$ *alors*

$$T := R\sqrt{\frac{n-2}{1-R^2}} \sim \mathcal{T}_{n-2}.$$

Tester l'indépendance

- Nous voulons tester si deux variables aléatoires X et Y sont indépendantes.
- Si nous supposons que le couple (X, Y) a une loi normale bivariée alors il suffit de tester l'hypothèse $\rho_{XY} = 0$.
- Nous rejetons l'hypothèse de l'indépendance au seuil α si

$$|T| = \left| R\sqrt{\frac{n-2}{1-R^2}} \right| > t_{n-2}(\alpha).$$

- Dans le cas $T > t_{n-2}(\alpha)$ on conclut qu'il existe une liaison positive entre les deux variables.
- Dans le cas $T < -t_{n-2}(\alpha)$ on conclut qu'il existe une liaison négative entre les deux variables.

- Soient X et Y deux variables aléatoires dépendantes.
- Il est envisageable que Y dépend de X de façon linéaire : $Y = aX + b + \epsilon$ avec $a, b \in \mathbb{R}$ et ϵ une erreur aléatoire.
- nous voulons trouver a et b à partir d'un échantillon de données bivariées $((x_1, y_1), ..., (x_n, y_n))$.

Définition

La droite de régression basée sur l'échantillon de données $((x_1, y_1), ..., (x_n, y_n))$ *est la droite linéaire définie par* $y(x) := a\dot{x} + b$ *avec* $a = \frac{s_{XY}}{\sigma_X^2}$ *et* $b := m_Y - a\dot{m}_X$.

- La droite de régression est l'unique droite $y(x)$ de sorte que la somme des carrés $\sum_{i=1}^{n}(y_i - y(x_i))^2$ est minimale.
- On appelle cette approche la méthode des moindres carrés (inventée par Gauß 1777-1855) .

Merci pour votre attention!