« Cours Statistique et logiciel R »

Rémy Drouilhet [1], Adeline Leclercq-Samson [1],
Frédérique Letué [1], Laurence Viry [2]

[1] Laboratoire Jean Kuntzmann, Dép. Probabilites et Statistique,

[2] Laboratoire Jean Kuntzmann, Dép. Modele et Algorithmes Déterministes

du 24 février au 7 avril 205

Plan de la présentation

Introduction à la démarche statistique

A notre disposition :

- mise en place d'une expérience,
- n résultats (appelés **données**) obtenus sur un **échantillon** de la population,
- calcul d'indicateurs sur ces n données,
- élaboration d'hypothèses à partir de ces calculs.

Objectif de la statistique inférentielle ?

- étendre les propriétés trouvées sur les n données à la population totale,
- vérifier la validité de ces propriétés,
- calculer l'erreur commise en étendant les propriétés à la population entière.

Approche par modélisation

Modélisation : Utilisation de l'outil mathématique afin de résoudre un problème concret (mise en équation, mise en forme à l'aide de fonctions, etc...). Un **modèle** est le choix de particularités ou d'hypothèses dans la modélisation.

Exemples :

- Equations de Boltzman,
- modèle de propagation d'une épidémie,
- modèle de profil de langue (phonétique), ...

Si le **modèle** *est mal choisi, il engendrera des résultats erronés.*

Modélisation statistique

Le statisticien cherche à modéliser...

l'aléatoire

L'aléa (ou **hasard**) se trouve dans le choix de l'échantillon parmi la population. **Modéliser** ce hasard permet d'étudier le problème considéré et/ou d'en extraire un protocole expérimental adapté.

Le hasard est appréhendé via des **variables aléatoires** : chaque **donnée** x_i (**après** échantillonnage) est considérée comme **une réalisation** $X_i(\omega)$ d'une variable aléatoire X_i (**avant** échantillonnage, où le choix de l'échantillon est modélisé par ω) ayant pour loi de probabilité $\mathbb{P}$.

$\mathbb{P}$ a priori **inconnue** $\Rightarrow$ **modèle** = **forme** ou **type** choisi(e) pour $\mathbb{P}$ (binômiale, Poisson, normale, exponentielle, etc...).

Etapes de la démarche statistique

1. **Choix du modèle de probabilité**
   - Dépend de la nature du phénomène étudié
   - variables qualitatives, quantitatives
   - le choix n'est pas unique en général, plusieurs modèles peuvent être comparés

2. **Estimation des quantités inconnues**
   - A partir des données disponibles, extraire de l'information pour estimer/apprendre le modèle de probabilité. Ce problème est appelé **problème d'ajustement**.
   - **modèle paramétrique** : on se restreint à une famille de lois de probabilités et on estime les paramètres inconnus de la famille de lois
   - **modèle semi- ou non-paramétrique** : on ne met pas de contrainte de formes sur une partie du modèle

Rappels de probabilité

Rappels de probabilité

On distingue les lois de probabilité de variables qualitatives ou quantitatives discrètes, des lois de variables quantitatives continues

Variables qualitatives ou quantitative discrète

Une variable discrète X prend un nombre fini (ex $\{1, 2, \ldots, 6\}$) ou dénombrable de valeurs $\mathbb{N}$.

On note $x_k, k = 1, 2, \ldots$ les valeurs prises par X. Définir la loi de probabilité de X, c'est définir les quantités

$$\mathbb{P}(X = x_k) = p_k, \quad \text{pour } k = 1, 2, \ldots$$

La somme des probabilités de toutes les éventualités vaut 1 : $\sum_k p_k = 1$.

Lois discrètes usuelles (1)

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

- X modélise la survenue ou non d'un événement de probabilité p.
- C'est une loi binaire, X prend deux valeurs, 0 ou 1 et

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

- exemples : on s'intéresse à l'événement
  - A="le patient a la grippe"
  - A = "l'enfant tend le doigt vers l'objet qu'on lui montre"
  - A = "l'électeur a voté pour le parti P", etc.

  Alors $X=1$ si l'événement A est réalisé, $X=0$ sinon.
- Espérance $\mathbb{E}(X)=p$; Variance $Var(X)=p(1-p)$.
- Simulation de N réalisations sous R : `rbinom(N,1,p)`

Lois discrètes usuelles (2)

Loi Binomiale $\mathcal{B}(n,p)$, avec n un entier et $p \in [0,1]$

- X modélise le nombre de fois où l'évènement A s'est produit parmi n expériences.
- X prend ses valeurs entre 0 et n.

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

- exemples :
  - nombre d'enfants dans la classe de 30 enfants qui ont attrapé la grippe
  - nombre d'objets vers lequel l'enfant a tendu le doigt parmi les 10 présentés
  - nombre de fois où l'électeur a voté pour le parti P lors des 5 derniers scrutins
- Espérance $\mathbb{E}(X) = np$; Variance $Var(X) = np(1-p)$.
- Simulation de N réalisations sous R : `rbinom(N,n,p)`

Lois discrètes usuelles (3)

Loi géométrique $\mathcal{G}(p)$, avec $p \in [0,1]$

- X modélise le premier instant où l'évènement A s'est produit dans une suite d'expériences successives.
- X prend ses valeurs dans $\mathbb{N}$.

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

- exemples :
  - première semaine où un enfant a eu la grippe
  - premier objet que l'enfant a attrapé dans une série d'objets présentés
  - première fois où un électeur a voté pour le parti P depuis qu'il vote
- Espérance $\mathbb{E}(X) = 1/p$; Variance $Var(X) = (1-p)/p^2$.
- Simulation de N réalisations sous R : `rgeom(N,p)`

Lois discrètes usuelles (4)

Loi de Poisson $\mathcal{P}(\lambda)$, avec $\lambda > 0$

- X modélise le nombre d'objets possédant un caractère rare, dans une population
- X prend ses valeurs dans $\mathbb{N}$.

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

- exemples :
  - nombre d'accidents de la route dans la vie d'un individu
  - nombre de mutations sur un gène
- Espérance $\mathbb{E}(X) = \lambda$; Variance $Var(X) = \lambda$.
- Simulation de N réalisations sous R : `rpois(N,lambda)`

Variable quantitative

Variables quantitatives continues

Une variable continue X prend un nombre indénombrable de valeurs (ex $\mathbb{R}$, $\mathbb{R}^+$).

On ne peut pas noter les valeurs prises par X les unes après les autres et on ne peut pas définir la probabilité de chaque valeur comme pour une variable discrète.
On introduit la notion de densité de probabilité, qui est une fonction f, positive ($f(x) \geq 0$) telle que, pour tout intervalle $[a, b]$

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

La "somme" des probabilités de toutes les éventualités vaut 1 :
$\int_{\mathbb{R}} f(x)dx = 1$.

Lois continues usuelles (1)

Loi uniforme $\mathcal{U}([a,b])$, avec $a < b$

- X prend ses valeurs dans $[a,b]$ de façon équiprobable.

$$f(x) = \frac{1}{b-a} 1_{x \in [a,b]}$$

- exemples :
  - temps d'attente aux urgences en supposant que cette attente ne peut pas dépasser 6h.
  - temps d'attente à un arrêt de tramway sachant que le tram passe tous les 5 min.
- Espérance $\mathbb{E}(X) = (a+b)/2$; Variance $Var(X) = (b-a)^2/12$.
- Simulation de N réalisations sous R : `runif(N,a,b)`

Lois continues usuelles (2)

Loi normale $\mathcal{N}(\mu, \sigma^2)$

- X prend ses valeurs dans $\mathbb{R}$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- exemples :
  - poids d'un individu
  - température à Grenoble
- Espérance $\mathbb{E}(X) = \mu$; Variance $Var(X) = \sigma^2$.
- Simulation de N réalisations sous R : `rnorm(N,mu, sigma)`

Lois continues usuelles (3)

Loi exponentielle $\mathcal{E}(\lambda)$

- X prend ses valeurs dans $\mathbb{R}^+$

$$f(x) = \lambda e^{-\lambda x} 1_{x>0}$$

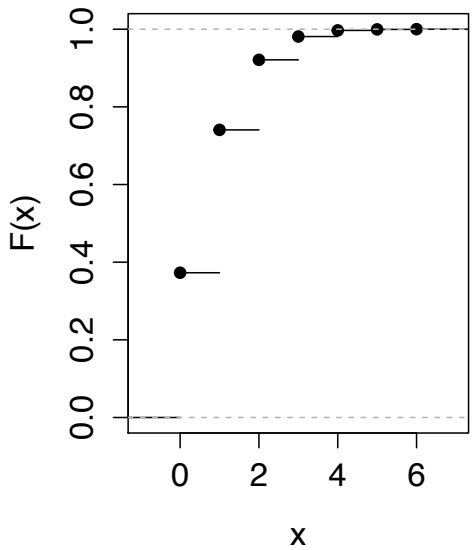
- exemples :
  - durée avant que l'enfant atteigne l'objet
  - durée avant la première avalanche
- Espérance $\mathbb{E}(X) = 1/\lambda$; Variance $Var(X) = 1/\lambda^2$.
- Simulation de N réalisations sous R : `rexp(N,lambda)`

Fonction de répartition

Pour tout x,

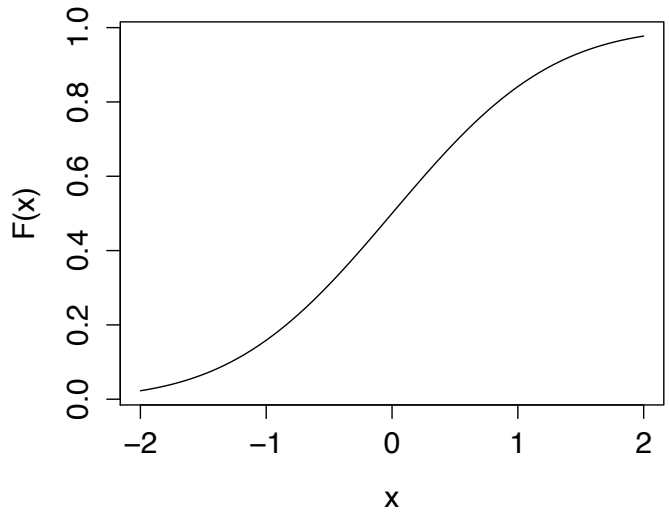
$$F(x) = \mathbb{P}(X \leq x)$$

Variables discrètes




```
plot(ecdf(X))
```

Variables continues

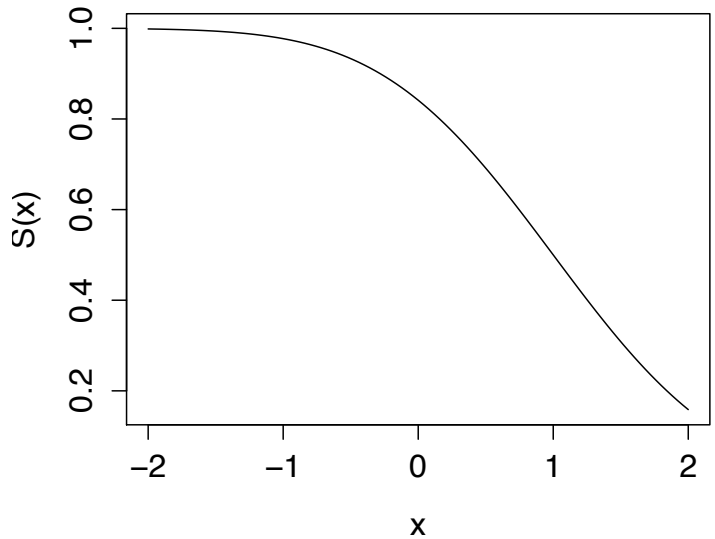



```
curve(pnorm(x,0,1), -2, 2)
```

Fonction de survie

Pour tout x

$$S(x) = \mathbb{P}(X > x) = 1 - F(x)$$




```
curve(1-pnorm(x,0,1),-2, 2)
```

Fonction quantile

Pour $u \in]0;1[$, on définit

$$Q(u) = \inf(x : F(x) \geq u)$$

- $u = 50\%$, $Q(u)$ est la valeur qui séparer la population (la variable) en deux : **médiane**
- $u = 25\%$, 25% des valeurs de la variables sont inférieures à $Q(u)$, 75% sont au dessus : **premier quartile**
- $u = 75\%$, 75% des valeurs de la variables sont inférieures à $Q(u)$, 25% sont au dessus : **troisième quartile**
- $u = 10, 20, \ldots, 90\%$: **déciles**

Théorèmes limites

Il existe deux grands théorèmes de convergence en probabilité

- **Loi de grands nombres**
- **Théorème centrale limite**

Ces théorèmes font appel à des notions de convergence de variables aléatoires. Il en existe de plusieurs sortes.

- Convergence en loi = convergence des fonctions de répartition
- Convergence presque sure = toute trajectoire converge avec probabilité 1

Lois des grands nombres

Loi des grands nombres

Soit $(X_i)_{1\leq i\leq n}$ une suite de v.a. iid, de même espérance μ. Pour tout $n \geq 1$, soit $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors, **quelle que soit la loi des** X_i, quand $n \to \infty$,

$$\overline{X}_n \xrightarrow{\mathbb{P},ps} \mu.$$

En d'autres termes, quelle que soit la loi des X_i, si n est grand, $\overline{X}_n$ est proche de μ.

Théorème central limite (TCL)

Théorème central limite

Soit $(X_i)_{1 \leq i \leq n}$ une suite de v.a. iid, de même espérance μ et de même variance σ^2. Pour tout $n \geq 1$, soit $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Alors, **quelle que soit la loi des** X_i, la v.a.

$$T_n = \sqrt{n} \, \frac{\overline{X}_n - \mu}{\sigma}$$

converge en loi vers une v.a. de loi $\mathcal{N}(0, 1)$.

En d'autres termes, quelle que soit la loi des X_i, si n est grand, $\overline{X}_n$ suit **approximativement** la loi $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Echantillon – Estimateur

Population et Echantillon

Population $\mathcal{P}$: Ensemble total des objets ou individus sur lequel on étudie un caractère statistique X.
Supposons $\mathrm{Card}(\mathcal{P}) = N$.

Echantillon $\mathcal{E}$: Sous-ensemble de $\mathcal{P}$.
$\mathrm{Card}(\mathcal{E}) = n$.

Observations : Tirages successifs de n individus dans $\mathcal{P}$
$\Rightarrow$ n v.a. $(X_1, \ldots, X_n)$ **avant** *échantillonnage* (plusieurs choix d'échantillons).
Obervations $(x_1, \ldots, x_n)$ **après** *échantillonnage* (**un tirage = une réalisation** de $(X_1, \ldots, X_n)$).

Propriété recherchée : Chaque variable avant échantillonnage suit **la même loi** que celle de X dans la population.

Quantités inconnues

On se place dans le **cadre paramétrique**, c'est-à-dire qu'on restreint l'étude de la loi de X à une famille de loi (binomiale, Poisson, normale, exponentielle, etc). Cette famille de loi est fixée, choisie en fonction du contexte de modélisation, mais son ou ses paramètres sont généralement inconnus.
On note θ **les paramètres inconnus** de la loi de X.
Exemples

- Loi binomiale : $\theta = p$
- Loi de Poisson : $\theta = \lambda$
- Loi normale : $\theta = (\mu, \sigma^2)$
- Loi exponentielle : $\theta = \lambda$

On cherche à approcher au mieux θ grâce aux n valeurs de l'échantillon : c'est le **problème d'ajustement**.

Exemple de la moyenne et la variance

Moyenne μ et **variance** σ^2 du caractère X **dans la population** $\mathcal{P}$: quantités **déterministes** ($\mathcal{P}$ fixe) liées aux valeurs de X sur $\mathcal{P}$:

$$\mu = \mathbb{E}(X), \qquad \sigma^2 = \mathrm{Var}(X),$$

μ et σ^2 **inconnues** : **on n'a pas accès** à toutes les valeurs de X sur $\mathcal{P}$.

Echantillonnage : approcher au mieux μ et σ^2 grâce aux n valeurs $(x_1, \ldots, x_n)$ tirées aléatoirement dans l'ensemble des N valeurs de X sur $\mathcal{P}$.

- **Estimateurs** de μ et σ^2 : **variables aléatoires** calculées à partir de $(X_1, \ldots, X_n)$ (avant échantillonnage)
- **Estimations** de μ et σ^2 : **quantités déterministes**, réalisations des estimateurs sur l'échantillon (après échantillonnage).

Estimateur : Avant Echantillonnage

Avant échantillonage les estimateurs de μ et σ^2 sont des **variables aléatoires** (aléa de l'échantillonnage) :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{estimateur de } \mu,$$

$$S_I^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{estimateur de } \sigma^2.$$

Remarque : Un autre estimateur de σ^2 est souvent utilisé

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

Estimation : Après Echantillonnage

On tire **un** échantillon $\mathcal{E}$ de taille n dans $\mathcal{P}$.
$\Rightarrow (x_1, \ldots, x_n)$ n valeurs du caractère sur $\mathcal{E}$.
$(x_1, \ldots, x_n) =$ **une** réalisation de $(X_1, \ldots, X_n)$.

$\Rightarrow$ valeurs numériques correspondant à la réalisation des estimateurs **sur l'échantillon** :

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{réalisation de } \overline{X}_n,$$

$$s_l^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \quad \text{réalisation de } S_l^2,$$

$$s^2 = \frac{n}{n-1}s_l^2.$$

Après échantillonnage, $\overline{x}_n$, s_l^2 et s^2 sont des valeurs **déterministes**
= réalisation des estimateurs $\overline{X}_n$, S_l^2 et S^2 sur l'échantillon en question.

Résumé

Exemple : Estimer la moyenne μ du caractère X dans la population.

Ensemble d'individus	**Observations/Paramètres**
Population $\mathcal{P}$	μ moyenne des valeurs de X sur $\mathcal{P}$. $\mu = \mathbb{E}(X)$. μ est **déterministe** (valeur inchangée).
Echantillon $\mathcal{E}$ (choisi aléatoirement dans $\mathcal{P}$)	$X_1, \ldots, X_n$ sur l'échantillon aléatoire. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ avant échantillonnage. $\overline{X}_n$ est un **estimateur** de μ. $\overline{X}_n$ est **aléatoire** (dépend de $\mathcal{E}$). $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ moyenne sur $\mathcal{E}$. $\overline{x}_n$ est **déterministe** une fois $\mathcal{E}$ choisi.

Problème Statistique

Inconnues : distribution d'un caractère X dans une population $\mathcal{P}$.

Hypothèse : la loi de X dans $\mathcal{P}$ (généralement de forme connue) dépend d'un paramètre θ **inconnu** (espérance, variance,...), éventuellement multi-dimensionnel.

A disposition : n observations $x_1, x_2, \ldots, x_n$, mesures du caractère faites sur un échantillon $\mathcal{E}$ de taille n.

$\Longrightarrow$ $(x_1, x_2, \ldots, x_n)$ réalisation d'un n-uplet de variables aléatoires $(X_1, X_2, \ldots, X_n)$ **indépendantes et identiquement distribuées (i.i.d.)**, de même loi que X.
On dit que $(X_1, X_2, \ldots, X_n)$ est un **échantillon aléatoire simple**, ou un **n-échantillon**, de la loi de X.

Problème : Comment **estimer** θ à partir des n-observations $(x_1, x_2, \ldots, x_n)$?

Estimateurs et Estimation

Estimateur

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon d'une loi P_θ dépendant d'un paramètre réel inconnu θ (déterministe, déduit de $\mathcal{P}$ ou du modèle). On appelle **estimateur de θ** une **variable aléatoire**

$$T_n = f(X_1, X_2, \ldots, X_n).$$

Remarque : Un estimateur est en fait une suite de v.a. $(T_n)_{n\geq 1}$. On assimile souvent l'estimateur avec le terme général de la suite T_n.

Estimation

Soit T_n un estimateur de θ. On appelle **estimation de θ** la réalisation t_n de la v.a. T_n, obtenue à partir de la réalisation **déterministe** $(x_1, x_2, \ldots, x_n)$ du n-échantillon $(X_1, X_2, \ldots, X_n)$ sur $\mathcal{E}$.

$$t_n = f(x_1, x_2, \ldots, x_n).$$

Exemple de la proportion

Problème : Estimer la proportion p de "Pile" obtenue au cours d'une infinité de lancers.

Réalisation : n résultats binaires $(x_1, \ldots, x_n)$ sur n lancers indépendants.

Modèle : X binaire (1 si succès, 0 sinon) de loi $\mathcal{B}(1, p)$
$\Longrightarrow (X_1, \ldots, X_n)$ iid de loi $\mathcal{B}(1, p)$

Paramètre : θ est la proportion inconnue p.

Estimateur : $T_n = \hat{p} = \overline{X}$ proportion **aléatoire** (dépend des n lancers).

Estimation : $t_n = \overline{x}$ proportion obtenue **après** les n lancers.

Exemple : Sur 30 lancers, on obtient 17 "Pile".
Une estimation de p sera alors $17/30 = 0,57$.

Question : On aurait pu prendre $T_n = 0$ (fonction constante), ou T_n la proportion aléatoire de "Face" obtenus au cours des n lancers. Ce sont également des fonctions des $(X_1, \ldots, X_n)$.
$\longrightarrow$ Quelles propriétés essentielles devra avoir T_n pour être considéré comme convenable ?

Propriétés Recherchées

Deux types de propriétés :

- **Propriétés à temps fixe**, comme par exemple sur $\mathbb{E}(T_n)$, $Var(T_n)$, ...
- **Propriétés asymptotiques**, lorsque n tend vers l'infini : par exemple, un estimateur raisonnable se concentrera de plus en plus autour de sa cible (i.e. le paramètre θ) lorsque n croît.

Propriétés étudiées de manière conjointe afin de définir un estimateur raisonnable.

Estimateurs sans biais

En moyenne sur tous les échantillons possibles, l'estimateur "vise bien" :

Biais

Soit T_n un estimateur de θ.

Biais de T_n : $B(T_n) = \mathbb{E}(T_n) - \theta$.

Si $B(T_n) = 0$ (i.e. $\mathbb{E}(T_n) = \theta$), T_n est dit **estimateur sans biais de θ**.
Si $B(T_n) \neq 0$ (i.e. $\mathbb{E}(T_n) \neq \theta$), T_n est dit **biaisé**.

Une bonne propriété est donc que l'estimateur soit sans biais.

Si T_n est biaisé, on espère que le biais tende vers 0 lorsque $n \to +\infty$:

Soit T_n un estimateur de θ. Si $B(T_n)$ tend vers 0 lorsque n tend vers l'infini, T_n est dit **estimateur asymptotiquement sans biais de θ**.

Exemple d'Estimateur Biaisé

Dans le cas d'une proportion, X suit la loi $\mathcal{B}(1, p)$. On considère 3 estimateurs pour p :

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T'_n = \frac{1}{n}\sum_{i=1}^{n}(1 - X_i), \quad T''_n = 0.$$

Alors

- $\mathbb{E}(T_n) = p$, T_n est sans biais,
- $\mathbb{E}(T'_n) = 1 - p$, T'_n est biaisé, avec un biais égal à $1 - 2p$,
- $\mathbb{E}(T''_n) = 0$, T''_n est biaisé, avec un biais égal à $-p$.

Remarque : T'_n est un estimateur sans biais de $1 - p$ (proportion de "Face").

Consistance

Autre propriété recherchée : **consistance**= plus n est grand, et plus T_n est "probablement proche" de θ.

Consistance

L'estimateur T_n de θ est **consistant** de θ s'il **converge en probabilité vers** θ quand n tend vers l'infini :

$$\forall \varepsilon > 0,\ \mathbb{P}(|T_n - \theta| \geq \varepsilon) \underset{\mathrm{n}\to\infty}{\longrightarrow} 0.$$

C'est la loi (faible) des grands nombres pour l'estimateur $\overline{X}_n$:

Loi (faible) des grands nombres

Soient n variables aléatoires $X_1, X_2, \cdots, X_n$ de même loi qu'une variable aléatoire X telle que $\mathbb{E}(X) = \mu$ et $\mathrm{Var}(X) = \sigma^2$ sont finis. Alors,

$$\overline{X}_n \underset{\mathrm{n}\to\infty}{\overset{\mathrm{P}}{\longrightarrow}} \mu$$

Propriétés Recherchées : Résumé

Propriétés recherchées pour T_n estimateur de θ :

- $B(T_n) = 0$ ou $B(T_n) \underset{n\to\infty}{\longrightarrow} 0$,
- $\mathrm{Var}(T_n)$ minimale si $B(T_n) = 0$ ou. $B(T_n) \underset{n\to\infty}{\longrightarrow} 0$,
- T_n consistant.

Exemple d'estimateur vérifiant toutes ces propriétés ?

$\overline{X}_n$ **estimateur de** μ**.**
$\hat{p}$ **estimateur de** p**.**

Cas de la Variance

Deux estimateurs de σ^2, variance du caractère X dans la population :

1. Cas où la moyenne μ dans la population est **connue** :
   - $S_I^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ (non-biaisé),
   - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$ (biaisé).

2. Cas où la moyenne μ dans la population est **inconnue** :
   - $S_I^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (biaisé),
   - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (non-biaisé).

Cas de la Variance (2)

On suppose que $\mu_4 = \mathbb{E}((X-\mu)^4)$ existe. Alors

- $\mathrm{Var}(S_I^2) \simeq \frac{1}{n}(\mu_4 - \sigma^4)$,
- $\mathrm{Var}(S^2) \simeq \frac{n}{(n-1)^2}(\mu_4 - \sigma^4)$,

avec égalité lorsque μ est connue.

Comme l'un est sans biais et l'autre est asymptotiquement sans biais, S_I^2 **et** S^2 **sont tous les deux consistants**, que μ soit connue ou non.

Estimation par intervalle de confiance

Introduction

$(X_1, \ldots, X_n)$ n-échantillon d'une loi dépendant d'un paramètre θ. On sait construire un estimateur T_n de θ et en déduire une estimation (donc une valeur) de θ à partir d'une réalisation $(x_1, \ldots, x_n)$.
Exemple : si $\theta = \mu$ moyenne sur la population, $T_n = \bar{X}_n$ et $t_n = \bar{x}_n$ estimation de θ sur l'échantillon.

Problème : quelle **confiance** accorder à cette estimation ?
Solution : trouver un **encadrement** de θ du type $[t_n - \varepsilon; t_n + \varepsilon]$ avec un **niveau de confiance** donné. Ce niveau de confiance sera atteint via la **loi** de l'estimateur T_n.

Exemple : pour $\theta = \mu$ et à variance connue, à un niveau de confiance donné de 95%, trouver ε de telle sorte que

$$\mathbb{P}(\theta \in [\bar{X}_n - \varepsilon; \bar{X}_n + \varepsilon]) \geq 0.95.$$

Comme il s'agit d'une **probabilité**, ε dépend évidemment de la loi de $\bar{X}_n$.

Intervalle de Confiance

But : encadrer θ avec une probabilité α (par exemple $\alpha = 5\%, 1\%, \ldots$) de se tromper.

$\Longrightarrow$ Intervalle **aléatoire** $I_{\theta,\alpha} = [A, B]$, où A et B sont deux **variables aléatoires** telles que

$$\mathbb{P}(\theta \in [A; B]) = \mathbb{P}(\theta \in I_{\theta,\alpha}) = 1 - \alpha.$$

Intervalle de Confiance (IC)

Soit α un réel tel que $0 < \alpha < 1$. On appelle **intervalle de confiance du paramètre θ de niveau de confiance $1 - \alpha$** (ou de risque α) un intervalle **aléatoire** I_α tel que

$$\mathbb{P}(\theta \in I_\alpha) = 1 - \alpha.$$

I_α construit à partir de la loi de l'estimateur T_n de θ. Deux cas : les échantillons gaussiens et les échantillons de loi quelconque de taille suffisamment grande (pour pouvoir appliquer le TCL).

Echantillon Gaussien

$(X_1, \ldots, X_n)$ n-échantillon de loi gaussienne $\mathcal{N}(\mu, \sigma)$.

Intervalle de confiance de μ à σ connue

$\bar{X}_n$ estimateur sans biais et consistant de μ.
$\Longrightarrow$ on l'utilise pour construire un intervalle de confiance de μ de niveau de confiance $1-\alpha$.

$(X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\mu, \sigma)$ donc

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

On cherche donc c_α tel que $\mathbb{P}\left(-c_\alpha \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq c_\alpha\right) = 1-\alpha$.

c_α : quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$.

$$\Longrightarrow I_\alpha = \left[\bar{X}_n - \frac{c_\alpha \sigma}{\sqrt{n}}; \bar{X}_n + \frac{c_\alpha \sigma}{\sqrt{n}}\right].$$

Echantillon Gaussien (2)

σ connue (suite)

Exemple : Pour $\alpha = 5\%$, $c_\alpha = 1,96 \Longrightarrow \mu \in I_\alpha$ avec une **probabilité fixe** égale à 95% : si on répétait l'expérience aléatoire de tirage d'un échantillon de taille n un grand nombre de fois, pour 95% d'entre eux (950 sur 1000 par exemple) l'intervalle contiendrait la valeur μ.

Pour un **échantillon donné** de taille n, l'intervalle correspondant est donc

$$\left[\overline{x}_n - \frac{c_\alpha \sigma}{\sqrt{n}}; \overline{x}_n + \frac{c_\alpha \sigma}{\sqrt{n}}\right].$$

Cet intervalle sera alors **l'estimation par intervalle de μ**, qui **contiendra** (**échantillon typique**) ou **ne contiendra pas** (**échantillon atypique**) la vraie valeur de μ.

Remarque : on peut aussi construire des intervalles unilatéraux tels que $\mathbb{P}(\mu \leq B_\alpha) = 0.95$ ou $\mathbb{P}(\mu \geq A_\alpha) = 0.95$.

Simulation d'intervalle de confiance de μ dans le cas gaussien

Simulations

```
niveau<-0.95
mu<-5
sigma<-1
N<-100
n<-50
X <- matrix(rnorm(N*n, mu, sigma),N,n)
```

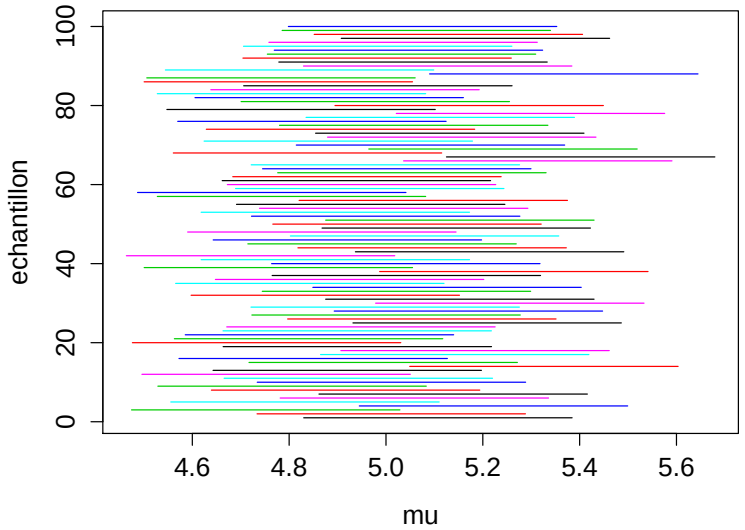
Calculs des intervalles de confiance

```
Xbar <- rowMeans(X)
ual2 <- (1+niveau)/2
ca <- qnorm(ual2)
amp <- ca*sigma/sqrt(n)
bi <- Xbar - amp
bs <- Xbar + amp
```

Simulation d'intervalle de confiance de μ dans le cas gaussien (2)

Graphique des intervalles de confiance

```
matplot(rbind(bi,bs),rbind(1:N,1:N), type = 'l', lty=1)
abline(v=mu)
```



Simulation d'intervalle de confiance de μ dans le cas gaussien (3)

Proportion d'intervalles contenant la vraie valeur

```
sa<-(length(which((bi<mu)&(bs>mu))))/N
```

Faire varier N et n

Echantillon Gaussien (3)

Cas μ et σ^2 inconnus

$\bar{X}_n$ estimateur sans biais et consistant de μ.
$\Longrightarrow$ on l'utilise pour construire un intervalle de confiance de μ de niveau de confiance $1-\alpha$.

S^2 estimateur sans biais et consistant de σ^2.
$\Longrightarrow$ on l'utilise pour construire l'intervalle de confiance de niveau de confiance $1-\alpha$ de σ^2.

Echantillon Gaussien (4)

Cas μ et σ^2 inconnus (suite)

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma)$. Soient

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Alors

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sqrt{S^2}} \sim \mathcal{T}(n-1) \qquad \text{(loi de Student à } n-1 \text{ degrés de liberté)}$$

Echantillon Gaussien (6)

Cas μ et σ^2 inconnus (suite)

Donc

- on cherche c_α tel que $\mathbb{P}\left(-c_\alpha \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sqrt{S^2}} \leq c_\alpha\right) = 1-\alpha$ à l'aide de la loi de Student,
- c_α est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de libertés

On obtient l'intervalle de confiance de μ de niveau de confiance $1-\alpha$,

$$I_\alpha = \left[\bar{X}_n - \frac{c_\alpha\sqrt{S^2}}{\sqrt{n}}; \bar{X}_n + \frac{c_\alpha\sqrt{S^2}}{\sqrt{n}}\right]$$

Simulation d'intervalle de confiance de μ dans le cas gaussien

Simulations

```
niveau<-0.95
mu<-5; sigma<-1
N<-100
n<-50
X <- matrix(rnorm(N*n, mu, sigma),N,n)
```

Calculs des intervalles de confiance

```
Xbar <- rowMeans(X)
S <- apply(X,1,sd)
ual2 <- (1+niveau)/2
ca <- qt(ual2, n-1)
amp <- ca*S/sqrt(n)
bi <- Xbar - amp
bs <- Xbar + amp
```

Simulation d'intervalle de confiance de μ dans le cas gaussien (2)

Graphique des intervalles de confiance

```
matplot(rbind(bi,bs),rbind(1:N,1:N), type = 'l', lty=1)
abline(v=mu)
```

Proportion d'intervalles contenant la vraie valeur

```
sa<-(length(which((bi<mu)&(bs>mu))))/N
```

Faire varier N et n

Echantillon Gaussien (3)

Intervalle de confiance de σ^2 à μ inconnue

S^2 estimateur sans biais et consistant de σ^2.
$\Longrightarrow$ on l'utilise pour construire l'intervalle de confiance de niveau de confiance $1-\alpha$ de σ^2.

$(X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\mu, \sigma)$ donc

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n).$$

On cherche donc c_α et d_α tels que $\mathbb{P}\left(c_\alpha \leq (n-1)\frac{S^2}{\sigma^2} \leq d_\alpha\right) = 1-\alpha$.

$$\Longrightarrow I_\alpha = \left[(n-1)\frac{S^2}{d_\alpha}; (n-1)\frac{S^2}{c_\alpha}\right].$$

Remarque : Pour un intervalle unilatéral, on cherche par exemple d_α tel que $\mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} \leq d_\alpha\right) = 1-\alpha$.

Simulation d'intervalle de confiance de σ^2 dans le cas gaussien

Simulations

```
niveau<-0.95
mu<-5
sigma2<-1
N<-100
n<-50
X <- matrix(rnorm(N*n, mu, sqrt(sigma2)),N,n)
```

Calculs des intervalles de confiance

```
S2 <- apply(X,1,var)
ca <- qchisq((1-niveau)/2, n-1)
da <- qchisq(1-(1-niveau)/2, n-1)
bi <- (n-1)*S2/da
bs <- (n-1)*S2/ca
```

Simulation d'intervalle de confiance de σ^2 dans le cas gaussien (2)

Graphique des intervalles de confiance

```
matplot(rbind(bi,bs),rbind(1:N,1:N))
abline(v=sigma2)
```

Proportion d'intervalles contenant la vraie valeur

```
sa<-(length(which((bi<sigma2)&(bs>sigma2))))/N
```

Faire varier N et n

Echantillon de Grande Taille

$(X_1, \ldots, X_n)$ n-échantillon d'une loi inconnue d'espérance $\mu = \mathbb{E}(X_1)$ et de variance $\sigma^2 = \mathrm{Var}(X_1)$. **On suppose** $n \geq 30$.

Problème : construire un intervalle de confiance de μ de niveau de confiance $1 - \alpha$.
Solution : Utiliser le théorème central limite (TCL) pour approcher la loi de $\bar{X}_n$ par une loi connue.

Comme $n \geq 30$, on peut approcher la loi de $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma}$ par une loi normale centrée et réduite.

On peut aussi montrer que on peut approcher la loi de $\sqrt{n}\dfrac{\bar{X}_n - \mu}{S}$ par une loi normale centrée et réduite.

Echantillon de Grande Taille (2)

Cas où σ est connue

Intervalle de confiance de μ de niveau de confiance $1-\alpha$

$$I_\alpha = \left[\bar{X}_n - \frac{c_\alpha \sigma}{\sqrt{n}}; \bar{X}_n + \frac{c_\alpha \sigma}{\sqrt{n}}\right],$$

où c_α obtenu via la table de la loi normale centrée réduite.

Cas où σ est inconnue

Intervalle de confiance de μ de niveau de confiance $1-\alpha$

$$I_\alpha = \left[\bar{X}_n - \frac{c_\alpha \sqrt{T_n}}{\sqrt{n}}; \bar{X}_n + \frac{c_\alpha \sqrt{T_n}}{\sqrt{n}}\right],$$

où c_α obtenu via la table de la loi normale centrée réduite.

Simulation d'intervalle de confiance de μ dans le cas non gaussien

Simulation d'un échantillon de Bernoulli de parametre p

```
niveau<-0.95
p<-0.6
N<-100
n<-50
X <- matrix(rbinom(N*n, 1, p),N,n)
```

Calculs des intervalles de confiance

```
Xbar <- rowMeans(X)
S <- apply(X,1,sd)
ual2 <- (1+niveau)/2
ca <- qt(ual2, n-1)
amp <- ca*S/sqrt(n)
bi <- Xbar - amp
bs <- Xbar + amp
```

Simulation d'intervalle de confiance de μ dans le cas gaussien (2)

Graphique des intervalles de confiance

```
matplot(rbind(bi,bs),rbind(1:N,1:N), type = 'l', lty=1)
abline(v=mu)
```

Proportion d'intervalles contenant la vraie valeur

```
sa<-(length(which((bi<p)&(bs>p))))/N
```

Calcul d'intervalle de confiance avec RStudio

Jeu de données

```
cystifibr  install.packages("ISwR")
library(ISwR)
help(cystfibr)
summary(cystfibr)
attach(cystfibr)
```

Calcul de l'IC par la fonction `t.test`

```
t.test(fev1, conf.level=0.95)$conf.int
t.test(fev1, alternative = "less")$conf.int
t.test(fev1, alternative = "greater")$conf.int
```

Différences entre les sexes

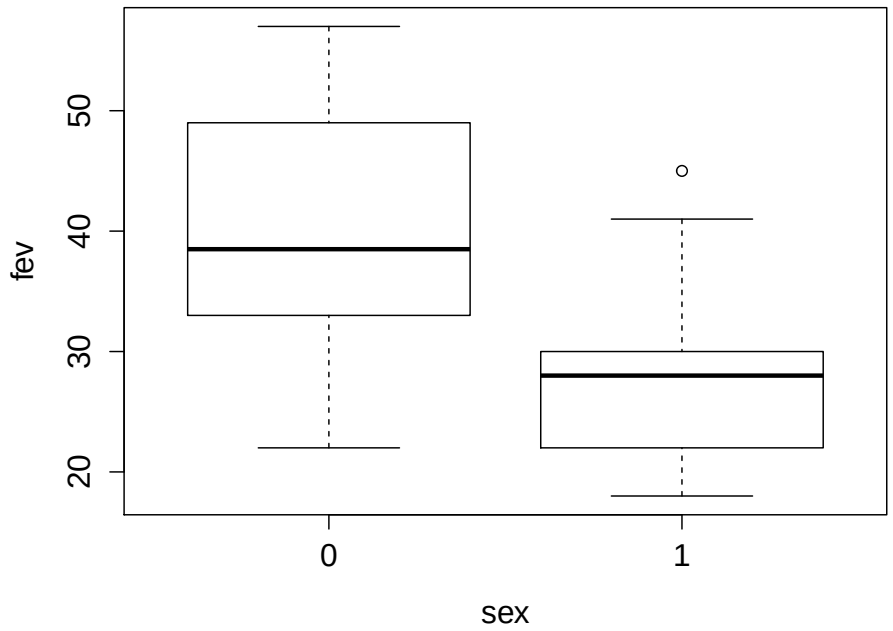
```
tapply(fev1,sex,summary)
t.test(fev1[sex==0], conf.level=0.95)$conf.int
t.test(fev1[sex==1], conf.level=0.95)$conf.int
```

Tests

Problématique

Exemple de la comparaison de deux échantillons

```
boxplot(fev1 sex)
```



On observe une différence entre les deux groupes. Est-elle significative ?

Pour cela on effectue un test statistique.

Démarche générale

On dispose de n données $x_1, \cdots, x_n$ réalisations de n variables aléatoires $X_1, \cdots, X_n$. On va chercher à tester des hypothèses portant sur la loi de probabilité du n-uplet $(X_1, \cdots, X_n)$ modélisant les observations.

On se placera essentiellement dans le cadre où les X_i sont **indépendantes et identiquement distribuées (iid).**

On effectue un test de H_0 contre H_1, H_0 **et** H_1 **étant deux hypothèses portant sur la loi de** $X_1, \cdots X_n$.

Démarche générale (2)

La construction d'un test va consister à établir une règle de décision permettant de faire un choix entre les deux hypothèses H_0 et H_1 au vu d'un échantillon de même loi que X.

En faisant ce test nous allons faire deux types d'erreurs

- La **première erreur** consiste à rejetter l'hypothèse H_0 alors qu'elle est vraie.
- La **deuxième erreur** consiste à garder l'hypothèse H_0 alors qu'elle est fausse.

Choix des hypothèses

H_0 est l'hypothèse privilégiée

L'hypothèse H_0 appelée *hypothèse nulle* est celle que l'on garde si le résultat de l'expérience n'est pas clair. On **conserve H_0 sauf si les données conduisent à la rejeter**.

Quand on ne rejette pas H_0, **on ne prouve pas qu'elle est vraie** ; on accepte de conserver H_0 car on n'a pas pu accumuler suffisamment d'éléments matériels contre elle ; les données ne sont pas incompatibles avec H_0, et l'on n'a pas de raison suffisante de lui préférer H_1 compte-tenu des résultats de l'échantillon.

L'hypothèse H_1 contre laquelle on *teste* H_0 est appelée *contre hypothèse ou hypothèse alternative*.

Niveau du test

En pratique, on va imposer pour le test que **la probabilité de rejetter H_0 à tort (alors qu'elle est vraie) soit petite**, inférieure à α, que l'on appelera le **niveau du test**.

Lorsque le test conduira à rejeter H_0 pour H_1, on dira que **le test de niveau α est significatif**. Dans ce cas, on aura démontrer H_1 avec un risque α de se tromper.

Erreurs de première et de seconde espèces

Lors de la prise de la décision de rejetter ou non l'hypothèse H_0, on peut commettre **deux erreurs** :

- **L'erreur de première espèce** est l'erreur que l'on commet lorsqu'on rejette H_0 à tort, ie lorsqu'on choisit H_1 alors que H_0 est vraie.
  La probabilité de commettre cette erreur, que l'on appelle **le risque de première espèce**, est notée
  $\mathbb{P}(\textit{rejeter } H_0 \textit{ à tort}) = \mathbb{P}_{H_0}(\textit{rejeter } H_0)$. Pour **un niveau de test** α, on impose que
  $$\mathbb{P}_{H_0}(\textit{rejeter } H_0) \leq \alpha.$$
- **L'erreur de deuxième espèce** est l'erreur que l'on commet lorsqu'**on accepte H_0 à tort**, ie lorsqu'on ne rejette pas H_0 alors qu'elle est fausse ; la probabilité de commettre cette erreur, que l'on appelle **le risque de deuxième espèce**, est notée β avec
  $$\beta = \mathbb{P}(\textit{accepter } H_0 \textit{ à tort}) = \mathbb{P}_{H_1}(\textit{accepter } H_0)$$

Erreurs de première et de seconde espèces (2)

Décision / Réalité	H_0	H_1
H_0	Décision correcte Probabilité : $1-\alpha$	**Erreur de première espèce** Probabilité α
H_1	**Erreur de seconde espèce** Probabilité β	Décision correcte Probabilité $1-\beta$

Stratégie

La stratégie consiste à fixer **le niveau du test** α.

Si $\alpha = 5\%$, il y a 5 chances sur 100 que, si H_0 est vraie, l'échantillon ne donne pas une valeur de l'observation comprise dans la zone d'acceptation de H_0.
On est donc **prêt à rejeter H_0 si le résultat fait partie d'une éventualité improbable n'ayant que 5% de chances de se produire**.

L'hypothèse H_0 est privilégiée : on veut avant tout contrôler le risque de rejeter H_0 à tort.

Une fois le risque de première espèce controlé par α, on cherchera alors à **minimiser le risque de deuxième espèce** β : une fois qu'on a controlé le risque de rejetter H_0 à tort, on va chercher à minimiser le risque de la garder à tort.

Choix de H_0

Selon les cas, le choix de l'hypothèse nulle est réservé

- **l'hypothèse qu'il est le plus grave de rejeter à tort**. On choisit H_0 et H_1 de telle sorte que le scénario catastrophique soit d'accepter H_1 alors que H_0 est vraie.
  Ce scénario "le pire" a ainsi une petite probabilité de se réaliser α fixée (le scénario catastrophique dépend souvent du point de vue considéré).
- **l'hypothèse dont on a admis jusqu'à présent la validité** et H_1 représente la contre hypothèse suggérée par une nouvelle théorie ou une expérience récente. Si le test conduit à rejeter H_0, on aura donc démontrer H_1 au risque α de se tromper.
- **l'hypothèse qui permet de faire le test** (seule hypothèse facile à formuler, permettant calcul de la loi d'une variable aléatoire sur laquelle on peut fonder le test).

Statistique de test- zone de rejet de H_0

Le test est basé sur une **statistique de test**, T_n, fonction des variables aléatoires $X_1, \cdots, X_n$.
La statistique de test T_n est une variable aléatoire dont la loi sous H_0 va déterminer la zone de rejet de H_0. On rejette H_0 quand T_n est "loin" de H_0, vers H_1.
Cette **zone de rejet** ZR_α est déterminée de telle sorte que

$$\alpha \geq P(\textit{rejeter } H_0 \textit{ à tort}) = \mathbb{P}_{H_0}(\textit{rejeter } H_0) = \mathbb{P}_{H_0}(T_n \in ZR_\alpha).$$

Une fois cette zone de rejet de H_0 déterminée, le risque de deuxième espèce est donné par est

$$\beta(\alpha) = P(\textit{accepter } H_0 \textit{ à tort}) = P_{H_1}(\textit{accepter } H_0) = P_{H_1}(T_n \notin ZR_\alpha).$$

Résumé sur les étapes d'un test statistique

- 2 hypothèses : H_0 (nulle) / H_1 (alternative)
- 1 statistique de test : T dont on connait exactement la loi sous H_0 et la tendance sous H_1.
- seuil de significativité α
- On détermine une région de rejet ZR_α dans laquelle se trouve T avec proba α sous H_0.
- Si la statistique observée sur les données t^{obs} se trouve dans la zone de rejet, on rejette l'hypothèse H_0 au profit de H_1. Si non, on ne peut rejeter H_0, on ne peut pas prouver que H_0 est fausse.

Test d'adéquation d'un échantillon à une valeur

Soit X un caractère quantitatif d'une population $\mathcal{P}$ dont on veut étudier l'espérance $E(X) = \mu$. On note $Var(X) = \sigma^2$.

Soit $(X_1, \cdots, X_n)$ un échantillon i.i.d. de même loi que X ; et $(x_1, \cdots, x_n)$ une série statistique, réalisation de l'échantillon.

Soit μ_0 est une valeur de référence connue et fixée. On veut tester

$$H_0 : \mu = \mu_0$$

contre une des trois alternatives suivantes

- $H_1 : \mu \neq \mu_0$
- $H_1 : \mu > \mu_0$
- $H_1 : \mu < \mu_0$

Construction du test

Il est naturel de rejeter H_0 si l'écart $\bar{X} - \mu_0$ est "trop grand".

Il faut donc connaître la loi de $\bar{X} - \mu_0$ sous H_0.

On distingue plusieurs cas :

- la variance σ^2 de X est-elle connue ou inconnue ?
- A t'on un grand échantillon (de loi quelconque) pour utiliser le TCL, ou un échantillon de loi normale

Cas d'un échantillon gaussien, variance inconnue

On suppose que l'on a un n-échantillon $(X_1, \cdots, X_n)$ gaussien, $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
On estime μ par $\bar{X}_n$, et σ^2 par S^2. On sait alors que

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}} \sim_{H_0} St(n-1)$$

(loi de Student à $(n-1)$ degrés de liberté).

Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$

Statistique de test : sous H_0, $T_n = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}} \sim_{H_0} St(n-1)$.

Règle de décision :
Si $\mid T_n \mid > s_\alpha$, on rejette H_0
Si $\mid T_n \mid \leq s_\alpha$, on ne rejette pas H_0
où s_α tq $P_{H_0}(\mid T_n \mid > s_\alpha) = \alpha$ (quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student).

Autre point de vue

La p-valeur

Définition

- La p-valeur est le risque minimal pour accepter H_1. C'est le plus petit niveau α pour lequel le test est significatif.
- C'est aussi la probabilité que la statistique de test dépasse la valeur observée sur les données, sous H_0.
- La $p - valeur$ ne peut être calculée qu'une fois que les observations ont été faites (t_n calculé).

C'est la valeur habituellement donnée par les logiciels.
Plus la p-valeur est faible, moins on prend de risque en rejetant H_0.

$p - valeur$	significativité du test
$0,01 < p - valeur \leq 0,05$	significatif
$0,001 < p - valeur \leq 0,01$	très significatif
$p - valeur \leq 0,001$	hautement significatif

p-valeur du test d'adéquation d'un échantillon

Pour le test d'adéquation d'un échantillon à une valeur μ_0

$p - valeur = P_{H_0}(\mid T_n \mid > \mid t_n \mid) = 2(1 - P_{H_0}(T_n \leq \mid t_n \mid)) = 2(1 - P(St(n-1) \leq \mid t_n \mid))$

car $T_n \sim_{H_0} St(n-1)$.

Fonction t.test

Exemples des données cystfibr avec $\mu_0 = 35$

- $H_0 : \mu = 35$ contre $H_1 : \mu \neq 35$
  `t.test(fev1, mu=35)`
- $H_0 : \mu = 35$ contre $H_1 : \mu > 35$
  `t.test(fev1, mu=35, alternative = "greater")`
- $H_0 : \mu = 35$ contre $H_1 : \mu < 35$
  `t.test(fev1, mu=35, alternative = "less")`

Réessayer avec $\mu_0 = 30$.

Erreur de première espèce

Pour $\alpha = 5\%$, on rejette H_0 avec une probabilité de se tromper en prenant cette décision égale à 5%. C'est **l'erreur de première espèce**.

Si on répétait l'expérience aléatoire de tirage d'un échantillon de taille n dans la loi $\mathcal{N}(\mu_0, \sigma^2)$ un grand nombre de fois, pour 95% d'entre eux (950 sur 1000 par exemple) on ne rejetterait pas l'hypothèse $H_0 : \mu = \mu_0$.

```
mu0<-5
sigma<-1
N<-1000
n<-100
X <- matrix(rnorm(N*n, mu0, sigma),N,n)
pvalue=rep(0,N)
for (i in 1:N){
pvalue[i]=t.test(X[i,], mu=mu0)$p.value
}
hist(pvalue)
length(which(pvalue<0.05))/N
```

Cas d'un échantillon non gaussien mais de grande taille, variance inconnue

On suppose que l'on a un n-échantillon $(X_1, \cdots, X_n)$ gaussien, $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
On estime μ par $\bar{X}_n$, et σ^2 par S^2. On sait alors que (TCL)

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}} \approx_{H_0} \mathcal{N}(0,1)$$

Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$

Statistique de test : sous H_0, $T_n = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}} \approx_{H_0} \mathcal{N}(0,1)$.

Règle de décision :
Si $\mid T_n \mid > s_\alpha$, on rejette H_0
Si $\mid T_n \mid \leq s_\alpha$, on ne rejette pas H_0
où s_α tq $P_{H_0}(\mid T_n \mid > s_\alpha) = \alpha$ (quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$).

Cas grand échantillon sous R

Même si la loi asymptotique est une loi $\mathcal{N}(0,1)$, **on utilise la fonction** `t.test`, qui suppose que la loi asymptotique est une loi de Student.

Exemple : jeu de données `faithful` sur les temps d'irruption du geyser Old Faithful dans le parc de Yellowstone.
On veut tester si l'espérance de la variable `éruptions` est égale à 3 :

$$H_0 : \mu = 3, \quad \text{contre} \quad H_1 : \mu \neq 3$$

```
hist(faithful$eruptions
length(faithful$eruptions)
t.test(faithful$eruptions, mu=3)
tvalue <- t.test(faithful$eruptions, mu=3)$statistic
# calcul de la p-valeur avec la loi normale
(1-pnorm(tvalue))*2
(1-pt(tvalue, 271))*2
```

Test sur la valeur d'une probabilité à un échantillon

On s'intéresse à **la probabilité p inconnue** d'avoir un caractère A dans une population $\mathcal{P}$.
On modélise ce caractère par une variable $X \sim \mathcal{B}(p)$. On a $E(X) = p(= \mu)$ et $Var(X) = p(1-p)(= \sigma^2)$.

Test de $H_0 : p = p_0$ contre $H_1 : p \neq p_0$
p_0 est une valeur connue fixée. Le test va être construit à partir de la différence $\bar{X}_n - p_0$. Si n suffisamment grand ($n > 30$), on peut appliquer le TCL.
Statistique de test

$$T_n = \frac{\sqrt{n}(\bar{X}_n - p_0)}{\sqrt{p_0(1-p_0)}} \quad \text{et} \quad \text{sous } H_0,\ T_n \approx_{H_0} \mathcal{N}(0,1)$$

Règle de décision : si $\mid T_n \mid > s_\alpha$, on rejette H_0
Si $\mid T \mid \leq s_\alpha$, on ne rejette pas H_0
où le seuil s_α tq $P_{H_0}(\mid T_n \mid > s_\alpha) = \alpha$.

Programmation sous R

Exemple avec la base cycstfibr

On s'intéresse à la masse corporelle et en particulier aux individus qui ont une masse inférieure à 80% de la masse normale. On souhaite tester si

$$H_0 : p = p_0 = 0.5,$$

c'est à dire si 50% des individus ont une masse inférieure à 80% de la masse normale.

```
attach(cystfibr)
bmpbinaire <- bmp<80
table(bmpbinaire)
t.test(bmpbinaire, mu=0.5)
```

Test de normalité

On dispose d'un n-échantillon $X_1, \ldots, X_n$ i.i.d. de même loi que X. On veut tester

$$H_0 : X \text{ est gaussien} \quad H_1 : X \text{ n'est pas gaussien}$$

- **Test de Kolmogorov-Smirnov** : On suppose connus (non estimés)μ et σ^2. On teste :

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2) \quad H_1 : X \text{ non } \mathcal{N}(\mu, \sigma^2)$$

- **Test de Shapiro-Wilks**

$$H_0 : X \text{ est gaussien} \quad H_1 : X \text{ n'est pas gaussien}$$

Exemples

```
shapiro.test(faithful$eruptions)
ks.test(faithful$eruptions, 'pnorm')
shapiro.test(fev1)
ks.test(fev1, 'pnorm')
```