Notes de cours Estimation/Test :

STID NIORT

R. Priam

L'inférence statistique permet de déduire d'un échantillon des informations sur la population dont il est issu. C'est à dire que partant d'un cas particulier, un échantillon, il devient possible d'obtenir un résultat pour le cas général, l'ensemble de la population, ainsi que pour tout autre échantillon de cette même population.

Le plan de ce cours est organisé comme suit, nous considérons la description d'un échantillon puis suivent des définitions (indicatives) de calcul des probabilités, et des notions de variable aléatoire. Les principales lois de probabilités utiles sont alors communiquées. Les méthodes de calcul sur des intervalles, élément indispensable ici, sont développés. Ensuite, les propriétés de la somme de n variables aléatoires *indépendantes* et ses moments (espérance, variance) sont définies.

Alors, partant d'un échantillon, l'étape de modélisation consiste à faire une hypothèse de loi sur cet échantillon : la variable aléatoire théorique - dont seraient issues nos données empiriques - est posée de façon ad hoc ! Une fois le modèle posé, nous sommes alors en mesure de réaliser les estimations et tests sur les moyennes, variances et proportions, les trois quantités considérées dans la suite.

Les exemples, graphiques, et calculs plus détaillés, sont donnés en cours.

1 Probabilités (de la description à l'inférence)

1.1 Description d'une série univariée

Population
La population est l'ensemble dont provient les données que l'on observe. Chacun de ses éléments est appellé **individu** ou unité statistique.

Échantillon (série univariée observée = les données)
Un échantillon est une partie (sous ensemble) de la population observée. Il s'agit en général des données, c'est à dire, la série de valeurs à disposition ! Dans la suite, il est précisé que l'on rencontre en général trois types de séries, dont la plus classique, la liste de n individus :

$$x_1, x_2, \cdots, x_i, \cdots, x_n$$

Dans le cadre descriptif, en statistique, on ne fait pas référence à la population donc on ne considére que les données à disposition (sans faire d'hypothèse de loi générale dessus). Une hypothèse de loi revient à faire une représentation mathématique de toute la population.

Caractère ou **variable**
Un caractére est la propriété observée dans la population ou l'échantillon considéré, il s'agit de la valeur du x_i ! Par exemple, la taille des élèves dans une classe, ou bien le poids des camions qui passent sur un pont, la circonférence des oranges récoltées sur un oranger. Le caractère peut être :

• soit **qualitatif** donc ne faisant pas l'objet d'une mesure (la région où est né un habitant d'une ville donnée, son sexe, le fait qu'il lise le journal ou non, etc ...). Il prend sa valeur parmi plusieurs modalités possibles.

• soit **quantitatif** donc mesurable (l'âge de l'habitant de la ville, son poids, le montant de ses revenus, etc ...). Il sera sera discret ou continu selon le fait qu'il prenne seulement un nombre fini de valeurs isolées entières ou bien un nombre infini de valeurs réelles.

Classe
Une classe est un sous ensemble de la population, ou de l'échantillon, correspondant à une même valeur ou des valeurs voisines (intervalle) prises par le caractère. En général le choix du nombre de classes n'est pas anodin, et dépend du degrés de compréhension voulue (peu de classes donne peu d'information et trop complique l'interprétation). L'effectif d'une classe est son nombre d'individus. La fréquence est la proportion d'individus de la population (ou de l'échantillon) appartenant à la classe.

Remarque : il est possible de transformer un caractère continu en discret ou inversement en prenant soit d'une part les centres des intervalles (continus) comme valeur discrète soit d'autre part les milieux entre deux valeurs discrètes (entières) pour construire des classes continues. Enfin, il arrive de considérer des valeurs entières comme continues par commodité.

Représentation graphique
Les données sont fournies sous la forme d'un tableau de données (classe, effectif, fréquence). Elles se représentent par un graphique synthétique différent suivant que le caractère est qualitatif (graphique diagramme à bâtons ou camembert) ou quantitatif (graphique histogramme).

Tableau de données, tableau d'effectifs, de fréquences
Graphes diagramme à batons, camembert, histogramme

Caractéristiques de position et dispersion
Pour les caractères quantitatifs, les caractéristiques de position et dispersion permettent de synthétiser davantage les données. Ce sont les moyenne, médiane, et écart-absolu, variance, écart-type, interquartile, qui permettent de construire la boîte à moustaches.

Moyenne
La moyenne arithmétique de n nombres $x_1, x_2, \cdots, x_n$ est

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Les trois cas de séries rencontrées généralement sont les suivantes,
• 1-ier cas, la liste des individus, $x_1, x_2, \cdots, x_n$,
• 2-ième cas, le tableau des effectifs n_i des p classes de valeurs x_i,
• 3-ième cas, le tableau des effectifs n_i des p classes d'intervalles $[a_i; b_i[$ de centre $(a_i + b_i)/2 = c_i$,

Soit le tableau récapitulatif, avec $f_i = n_i/n$ une fréquence :

1-ier cas	$\bar{x}_n = \frac{x_1+x_2+\cdots+x_n}{n}$	$= \frac{1}{n}\sum_{i=1}^{n} x_i$
2-ième cas	$\bar{x}_n = \frac{n_1x_1+n_2x_2+\cdots+n_px_p}{n_1+n_2+\cdots+n_p}$	$= \frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p} n_i x_i$
3-ième cas	$\bar{x}_n = \frac{n_1c_1+n_2c_2+\cdots+n_pc_p}{n_1+n_2+\cdots+n_p}$	$= \frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p} n_i c_i$

Médiane $Q_2 = Me$, quartile Q_1, Q_2, Q_3, quantiles

Graphe pourcentages (des quantiles)
Boîte à moustache (valeurs min, max, aberrantes)

Remarque : la médiane n'est pas sensible aux valeurs aberrantes (extrêmes et isolées) contrairement à la moyenne. Deux séries peuvent être de même moyenne tout en étant évidemment différentes, donc il convient d'étudier la dispersion. Même si les quartiles et l'écart-absolu apportent une information de choix, la variance tient une place particulière du fait de sa facilité de calcul formel, de même que la moyenne vis à vis de la médiane.

Ecart absolu

$$e_n = \frac{|x_1-\bar{x}|+|x_2-\bar{x}|+\cdots+|x_n-\bar{x}|}{n} = \frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$$

Variance, écart-type

1-	$v_n = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$	$= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$
2-	$v_n = \frac{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+\cdots+n_p(x_p-\bar{x})^2}{n_1+n_2+\cdots+n_p}$	$= \frac{1}{n}\sum_{i=1}^{p}n_i(x_i-\bar{x})^2$	$= \frac{1}{n}\sum_{i=1}^{p}n_ix_i^2-\bar{x}^2$
3-	$v_n = \frac{n_1(c_1-\bar{x})^2+n_2(c_2-\bar{x})^2+\cdots+n_p(c_p-\bar{x})^2}{n_1+n_2+\cdots+n_p}$	$= \frac{1}{n}\sum_{i=1}^{p}n_i(c_i-\bar{x})^2$	$= \frac{1}{n}\sum_{i=1}^{p}n_ic_i^2-\bar{x}^2$

avec $n = \sum_{i=1}^{p} n_i$ et l'écart-type $\sigma_n = \sqrt{v_n}$.

Remarque : la variance ici v_n sera notée plus loin s_n^2, la variance non corrigée. On notera $s_n^{2\prime} = \frac{n}{n-1} = s_n^2$ la variance corrigée qui interviendra en estimation pour x_i continu.

Interquartile
$Q_3 - Q_1$

Exemple récapitulatif

Vu en cours et TDs.

Dans la suite, les concepts de probabilité et de variable alétatoire vont permettre de passer du cadre *descriptif* (population ignorée) au cadre *inférentiel* (population prise en compte) pour l'estimation et test : récupérer les informations contenues dans l'échantillon afin de les généraliser à la population entière. Une hypothèse probabiliste sur la génération de l'échantillon sera alors faite à l'étape de la modélisation (+échantillonnage).

1.2 Calculs de probabilités

On introduit le vocabulaire (ensembliste) des événements.

Dans une **expérience aléatoire**, l'**univers** Ω est l'ensemble des résultats possibles.

Un **évenement** est une partie (au sens ensembliste) de l'univers. Un **événement élémentaire** est un événement possédant un seul élément.

Des événements A, B, sont **disjoints**, ou **incompatibles**, si et seulement si, $A \cap B = \emptyset$.

L'événement **contraire** (dit aussi **complémentaire**) d'un événement A est l'ensemble $\bar{A}$ des éléments de Ω n'appartenant pas à A.

On note $\mathcal{A} = \mathcal{P}(\Omega)$ l'ensemble des événements (parties) de l'univers Ω.

Définition

Soit un univers (éventuellement fini) Ω. Une **probabilité** sur Ω est une application P de l'ensemble des événements $\mathcal{A} = \mathcal{P}(\Omega)$ vers l'intervalle $[0, 1]$

$$\begin{array}{rccc} P: & \mathcal{P}(\Omega) & \to & [0;1] \\ & A & \mapsto & P(A) \end{array}$$

telle que :
- $P(\Omega) = 1$.
- Pour tous événements A et B, si $A \cap B = \emptyset$, alors

$$P(A \cup B) = P(A) + P(B)$$

Remarque :

- Pour tout événement A : $P(\bar{A}) = 1 - P(A)$

- Pour tous événements A, B : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- La probabilité d'un événement A est égale à la somme des probabilités des événements élémentaires inclus dans A.

- Ici, on travaille sur $\mathcal{A}$, les parties de Ω, mais plus rigoureusement, il faudrait prendre $\mathcal{A} \subset \mathcal{P}(\Omega)$ en vérifiant que $\mathcal{A}$ satisfait les propriété requises pour que la probabilité P puisse s'exprimer dessus (notamment opérations ensemblistes d'unions et intersections employées par P).

- L'espace $(\Omega, \mathcal{A}, P)$ est appellé espace probabilisé, dans le sens que à

partir de Ω on définit un ensemble particulier $\mathcal{A}$ sur lequel on peut définir une probabilité, puis enfin cette probabilité P.

Définition

L'**équiprobabilité** correspond au cas où tous les événements élémentaires ont même probabilité. Dans ce cas, la probabilité d'un événement A est :

$$P(A) = \frac{\text{Nombre d'éléments de A}}{\text{Nombre d'éléments de } \Omega} = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}$$

Si $\Omega = \{w_1, w_2, \cdots, w_n\}$, et si il y a équiprobabilité pour les événements élementaires, alors $P(\{w_i\}) = \frac{1}{n} = \frac{1}{Card\,\Omega}$

Définition

Soit P une probabilité sur Ω, et soit A un événement de probabilité non nulle. La **probabilité conditionnelle** que A soit réalisé, est l'application P_A qui à tout événement B, associe le nombre :

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

On dit que $P(B/A)$ est la "probabilité de B sachant A" (est réalisé).

Définition

Pour tous événements A et B de probabilités non nulles, on a la formule des probabilités composées :

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

Définition

Les événements A et B, de probabilités non nulles, sont **indépendants** si et seulement si :

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

Ce qui est équivalent, si les probabilités $P(A)$ et $P(B)$ sont non nulles à :

$$P(A|B) = P(A) \text{ ou } P(B|A) = P(B)$$

En remarque, si A et B sont incompatibles alors $P(A \cap B) = P(\emptyset) = 0$. Ainsi deux événements incompatibles de probabilités non nulles ne peuvent être indépendants.

Exemple récapitulatif

Vu en cours et TDs.

1.3 Notion mathématique d'une variable aléatoire réelle

Soit Ω l'ensemble des événements élémentaires observables à l'issue d'une épreuve aléatoire. Soit $\mathcal{A}$ l'ensemble de tous les événements associés à cette épreuve, c'est à dire (pour simplifier, cf. section précédent) on prend les parties $\mathcal{P}(\Omega)$, et on considére un espace probabilisé $(\Omega, \mathcal{A}, P)$.

Définition

Une **variable aléatoire** est une application :

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(w) \end{array}$$

telle que pour tout x réel, $A = \{w | w \in \Omega \ et \ X(w) = x) \in \mathcal{A}$, est un événement, autrement dit, pour tout $x \in \mathbb{R}$, l'ensemble de tous les w ayant x pour image par X est un événement.

On note l'événement A par $\{X = x\}$, $(X = x)$ ou également $X^{-1}(x)$ puisqu'il s'agit de l'image réciproque de x par X.

Remarque

On note $X(\Omega)$ l'ensemble des valeurs prises dans $\mathbb{R}$ par X, sachant que les probabilités se calculent sur des parties de $X(\Omega)$, soit $\mathcal{B} = \mathcal{P}(X(\Omega))$ ou même plus généralement $\mathcal{P}(\mathbb{R})$ (pour simplifier, cf. section précédente).

Le principe fondamental mis en oeuvre ici est que l'on peut définir une loi sur $\mathbb{R}$ à partir de X en calculant l'image d'un événement pris comme une des parties $\mathcal{P}(X(\Omega))$. En effet, si $B \in \mathcal{P}(X(\Omega))$, alors $X^{-1}(B) \in \mathcal{P}(\Omega)$, donc, on utilise la probabilité originale P évaluée sur les parties $\mathcal{P}(\Omega)$.

Définition

La **loi de probabilité ou distribution** de la v.a. X est l'application P_X définie par :

$$\begin{array}{rccc} P_X: & \mathcal{P}(X(\Omega)) & \to & [0;1] \\ & B & \mapsto & P_X(B) = P(X^{-1}(B)) \end{array}$$

Remarque

En pratique, dans la suite, puisque X est une variable aléatoire réelle, $X(\Omega)$ est dans $\mathbb{R}$. Et, il existe deux types de variables aléatoires réelles, suivant que :

$$\begin{cases} X(\Omega) \subset \mathbb{N} & \text{si X v.a. discrète,} \\ X(\Omega) \subset \mathbb{R} & \text{si X v.a. continue.} \end{cases}$$

L'événement B sera[1] :

$$\begin{cases} \text{soit un entier k ou un ensemble de m entiers } \{k_1, k_2, ..., k_m\} & \text{si X v.a. discrète,} \\ \text{soit un intervalle } I_1 \text{ ou une union de deux intervalles } I_1 \text{ et } I_2 & \text{si X v.a. continue.} \end{cases}$$

D'où les événements associés $X^{-1}(B)$, qui correspond à la nature discrète ou continue de la variable aléatoire :

$$\begin{cases} \bullet \text{ soit } \{X = k\} \text{ ou } \{X = k_1\} \cup \{X = k_2\} \cup \cdots \cup \{X = k_m\} & \text{si } \mathbf{X \text{ v.a. discrète}}, \\ \bullet \text{ soit } \{X \in I_1\} \text{ ou } \{X \in I_1\} \cup \{X \in I_2)\} & \text{si } \mathbf{X \text{ v.a. continue}}. \end{cases}$$

Remarque

En résumé, on ne travaillera -dans ce cours- plus avec Ω mais directement des parties de $\mathbb{R}$, donc un entier, un ensemble d'entiers, un intervalle ou bien une union de deux intervalles. On peut représenter ces valeurs sur une droite ! La représentation graphique des probabilités sera donc privilégiée par la suite.

Exemple

Pour une pièce équilibrée parfaite, soit X la variable aléatoire qui donne après son lancer la valeur de la face tournée vers le haut, on sait que

$$P(X = 1) = P(X = 0) = \frac{1}{2}$$

si $(X = 1)$ correspond à l'événement dans Ω du Pile et $(X = 0)$ correspond à l'événement dans Ω du Face qui sont tous les deux équiprobables.
Ici $\Omega = \{Pile, Face\}$, et X ne prend que les valeurs 0 et 1 dans $\mathbb{R}$ donc

$$X(\Omega) = \{0, 1\}.$$

Définition

La **fonction de répartition** associée à la variable aléatoire X est la fonction définie par :

$$\begin{array}{rccc} F_X : & \mathbb{R} & \to & [0;1] \\ & x & \mapsto & F_X(x) = P(X \leq x) \end{array}$$

où on considére[2] donc simplement l'ensemble des réels inférieurs ou égal à x, qui sont dans l'image de X, soit, $X(\Omega) \cap]-\infty; x]$.

Remarques :

• **Si X est discrète**, alors elle prend les valeurs k_i $(i \in I \subset \mathbb{N})$, où I est un ensemble d'indices (m entiers pour simplifier). Donc **pour définir P_X, il suffit d'indiquer les probabilités p_i et les valeurs k_i** :

$$P(X = k_i) = p_i \text{ avec } \sum_{i=1}^{m} p_i = 1.$$

[1] Attention à ne pas confondre les k_i parfois remplacés pas des x_i que sont les valeurs que prennent la v.a. X discrète, et non les individus d'un échantillon.

[2] En revenant à la déf d'une v.a, on a $X^{-1}(B) = \{X \leq x\} = \{w | w \in \Omega \; et \; X(w) \leq x\}$

La fonction de répartition est constante par intervalles (en escaliers). Dans la suite, on considérera uniquement deux cas (Bernoulli et Binomial) pour lesquels les p_i sont connus. Dans le cas dénombrable, m est infini et la somme porte sur $\mathbb{N}$ (pour loi de Poisson par exemple).

• **Si X est continue**, alors elle prend ses valeurs sur des intervalles de $\mathbb{R}$, et la fonction de répartition F_X est continue, croissante, et dérivable. Elle admet une fonction **densité de probabilité** telle que

$$f_X = F'_X \text{ avec } \int_{-\infty}^{+\infty} f_X(t)\, dt = 1.$$

Donc **pour définir P_X il suffit d'indiquer la fonction f_X et son domaine de définition**. En outre, f_X est positive ou nulle, continue sauf en un nombre dénombrable de points (où elle admet une limite finie à gauche et à droite). Par ailleurs, on admet que

$$P_X(\{x\}) = P(\{X = x\}) = P(X^{-1}(x)) = 0 \text{ pour tout } x \in \mathbb{R}.$$

Dans la suite, on considèrera uniquement quatre cas (gaussien, χ^2 et t de Student, F de Fisher) pour lesquels f_X est connue.
Enfin :

$$P(a \leq X \leq b) = P(a < X < b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$

Cette quantité se visualise sur la représentation graphique de f_X comme l'aire sous la courbe f_X pour les abcisses entre a et b.

Exemple :
On calcule la fonction de répartition de la loi exponentielle de paramètre $\lambda = 2$ dont la densité s'écrit sur $\mathbb{R}$:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 2e^{-2x} & \text{si } x \geq 0. \end{cases}$$

Pour tout $x < 0$, $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0.$
Pour tout $x \geq 0$, $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} 2e^{-2t}dt = [-e^{-2t}]_0^x = 1 - e^{-2x}$

qui se représente graphiquement par une courbe concave avec une asymptote $y = 1$ pour x tendant vers l'infini. En remarque, on peut calculer la médiane :

$$F(Me) = 0.5 \Rightarrow Me = ln(2)/2 \text{ soit } 0.35 \text{ environ.}$$

On peut vérifier que l'intégrale de f entre $-\infty$ et $+\infty$ vaut bien 1 puisque la limite de F(x) tend vers 1 lorsque x tend vers $+\infty$!

L'espérance mathématique d'une variable aléatoire est une moyenne arithmétique où les probabilités remplacent les fréquences.

Définition

Lorsque ces quantités existent, on note l'**espérance** de la v.a. X :

$$E(X) = \begin{cases} \sum_{i=1}^{m} k_i \, p_i & \text{si } X \text{ v.a. discréte,} \\ \int_{-\infty}^{+\infty} x f(x) dx & \text{si } X \text{ v.a. continue.} \end{cases}$$

Dans le cas discret, si $X(\Omega)$ est dénombrable (en pratique égal à $\mathbb{N}^+$), alors la somme peut ne pas converger (puisque m tend vers l'infini). De même, dans le cas continu, l'intégrale peut ne pas être finie.

Exemple

Pour X prenant la valeur 0 ou 1 (loi de Bernoulli), la loi de X se définit simplement par $P(X = 1) = p$ et $P(X = 0) = 1 - p$. Pour une piéce truquée, $p = 0.6$ par exemple, l'espérance de X s'écrit alors :

$$\begin{aligned} E(X) &= 1 \times P(X = 1) + 0 \times P(X = 0) \\ &= 1 \times p + 0 \times (1 - p) \\ &= p = 0.6 \end{aligned}$$

In fine, la piéce est telle que sur l'ensemble de n lancers effectués, il sera obtenu plus souvent des 1 que des 0. Le joueur connaissant la propriété de la piéce et misant davantage d'argent sur la valeur 1 (Pile) que la valeur 0 (Face) aura un gain supérieur à son adversaire, principe utilisé pour truquer un jeu de hasard simple.

Définition

La **variance** de la v.a. X, lorsqu'elle existe, s'écrit :

$$V(X) = E[(X - E(X))^2] = E(X^2) - E(X)^2$$

L'**écart-type** de la v.a. X s'écrit :

$$\sigma(X) = \sqrt{V(X)} = \sqrt{E(X^2) - E(X)^2}$$

Exemple :

En reprenant le jeu ci-dessus, dans le cas d'une variable aléatoire à valeur 0 ou 1 (loi de Bernoulli), on calcule :

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 \\ &= [1^2 \times P(X = 1) + 0^2 \times P(X = 0)] - E(X)^2 \\ &= p - p^2 \\ &= p(1 - p) = 0.24 \end{aligned}$$

Exemple :

On calcule l'espérance, et la variance de la v.a. de loi uniforme. Cette v.a.

est telle que tout intervalle de taille fixée est de même probabilité. Il s'agit donc en quelque sorte d'équiprobabilité pour une variable aléatoire continue. Elle admet pour densité :

$$f(x) = \begin{cases} 1 & \text{si } x \in [0;1] \\ 0 & \text{sinon.} \end{cases}$$

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 xf(x)dx = [\frac{x^2}{2}]_0^1 = \frac{1}{2} \\ E(X^2) &= \int_{-\infty}^{+\infty} x^2 f(x)dx = \int_0^1 x^2 f(x)dx = [\frac{x^3}{3}]_0^1 = \frac{1}{3} \\ E(X) &= E(X^2) - E(X)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12} \\ \sigma(X) &= \sqrt{V(X)} = \sqrt{\frac{1}{12}} \end{aligned}$$

Représentation graphique de f, F et expression de la fonction F.

En bref :
Le tableau suivant est une synthèse des objets fondamentaux manipulés dans la suite.

v.a. X	discrète , $X(\Omega) \subset \mathbb{N}$	continue , $X(\Omega) \subset \mathbb{R}$
Valeurs prises	entiers k_i $i \in [1;m]$	intervalle $[a;b]$ $a,b \in \mathbb{R}$
Définie par	p_i	f_X
Probabilité	$P(X = k_i) = p_i$	$P(X \in [a;b]) = \int_a^b f(t)dt$
Somme à 1	$\sum_{i=1}^m p_i = 1$	$\int_{-\infty}^{+\infty} f(t)dt = 1$
Espérance	$E(X) = \sum_{i=1}^m k_i\, p_i$	$E(X) = \int_{-\infty}^{+\infty} tf(t)dt$
Variance	$V(X) = \sum_{i=1}^m (k_i - E(X))^2 p_i$	$V(X) = \int_{-\infty}^{+\infty} (t - E(X))^2 f(t)dt$

En outre[3], dans le cas discret, on peut considérer aussi la probabilité d'un ensemble (fini ou dénombrable) d'entiers, et dans le cas continu, d'un intervalle avec bornes ouvertes, fermées ou infinies, voir une union de deux de ces intervalles.

On schématise les lois correspondantes par un diagramme à bâtons théorique pour une v.a. discréte et par la fonction continue de densité[4] théorique pour

[3] On peut construire un tableau similaire pour le cas d'un échantillon, avec l'effectif, la fréquence, la moyenne, la variance, l'écart-type.

[4] En remarque, on peut également définir un histogramme théorique tel que les probabilités remplacent les fréquences dans le calcul des aires des rectangles. Cet histogramme a un sens dans le test d'adéquation qui consiste à vérifier si un échantillon provient d'une loi donnée : on compare alors les histogrammes théorique/empirique avec un test du χ^2.

une v.a. continue. L'échantillon sera représenté par un diagramme à bâtons empirique (ou un camembert) pour le cas discret et un histogramme empirique pour le cas continu. Ces représentations sont vues notamment en TDs.

On rappelle que le terme empirique fait référence à l'expérience donc un échantillon, et le terme théorique fait référence à une loi de probabilité. Ainsi, les moyenne théorique et variance théorique correspondent à l'espérance et la variance d'une variable aléatoire, tandis que les moyenne empirique et variance empirique correspondent à la moyenne et la variance d'un échantillon.

Soit :

Moments	**Notation**	**Calculé sur**
Moyenne théorique ou espérance mathématique ou moyenne de la population, vraie moyenne	μ	une v.a. X ou une population
Variance théorique ou variance mathématique, ou variance de la population, vraie variance	σ^2	une v.a. X ou une population
Moyenne empirique ou moyenne de l'échantillon	$\bar{x}_n$	un échantillon
Variance empirique ou variance de l'échantillon	s_n^2	un échantillon

1.4 Principales lois

Pour rappel, il ne sera plus nécessaire de travailler sur l'espace Ω et on ne s'intéressera qu'aux seules valeurs $X(\Omega) \subset \mathbb{R}$ prises par X. Ainsi, on supposera toujours de façon implicite travailler sur $(\Omega, \mathcal{A}, P)$, un espace probabilisé.

Nom	Param.	$X(\Omega)$	**Loi**	E(X)	V(X)
Loi Bernoulli	p	$\{0,1\}$	$P(X=1)=p$	p	$p(1-p)$
Loi binomiale	n,p	$\{0,1,2,\cdots,n\}$	$P(X=k)=C_n^k p^k(1-p)^{n-k}$	np	$np(1-p)$
Loi Poisson	λ	$\mathbb{N}$	$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ
Loi normale	μ,σ	$\mathbb{R}$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(\frac{x-\mu}{\sigma})^2}$	μ	σ
Loi χ^2					
Loi t-Student					
Loi de Fisher					

Il est également possible de définir de nouvelles variables aléatoires, par somme ou combinaison linéaire notamment, à partir d'existantes.

Propriétés (admises)

Soient a, b, deux réels, et X une v.a. Si on pose la v.a. $Y = aX + b$, alors :

$$\begin{aligned} E(Y) &= E(aX+b) = aE(X)+b \\ V(Y) &= a^2 V(X) \\ \sigma(Y) &= |a|\sigma(X) \end{aligned}$$

On démontre facilement des résultats analogues pour le cas empirique (moyenne, variance et écart-type d'un échantillon, cf. TDs).

1.4.1 Calculs d'intervalles et loi gaussienne

La loi normale est la loi la plus utilisée pour modeliser une mesure entachée d'erreurs ou présentant une certaine variabilité due à la physique ou la biologie. Elle résulte également de la somme de nombreuses causes (variables) aléatoires indépendantes de loi quelconque.

• (Définition) Fonction de densité, de répartition (notée Φ ou Π ou $F_{\mathcal{N}(\mu,\sigma^2)}$), espérance, variance.

• (Théorème) Pour a, b réels, et a non nul, on admet que :

$$X \sim \mathcal{N}(\mu,\sigma^2) \Longleftrightarrow Y = aX+b \sim \mathcal{N}(a\mu+b,(a\sigma)^2)$$

• (Corollaire) Si X est de loi normale $\mathcal{N}(\mu,\sigma^2)$, alors

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

C'est la loi normale centrée réduite qui a été tabulée (cf. table de la loi N(0,1) au format papier). On se raméne toujours à cette loi N(0,1) pour tout calcul d'intervalle $P([a;b])$ ou $P(]-\infty;a])$ ou $P([b;+\infty[)$, pour une loi normale, en effectuant la transformation ci-dessus, puis en lisant la table N(0,1).

• $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
par propriété de l'intégrale ou bien se montre graphiquement.

• $\Phi(-z) = 1 - \Phi(z)$
par changement de variable ou bien se montre graphiquement.

• $P(-z \leq Z \leq z) = \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2 \times \Phi(z) - 1$

Remarque :
Ainsi pour un intervalle $[a;b]$ quelconque de $\mathbb{R}$ et une v.a. de loi $\mathcal{N}(\mu, \sigma^2)$, on calcule $P(X \in [a;b])$ par :

$$\begin{aligned} P(a \leq X \leq b) &= P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) \\ &= P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}) \text{ où } Z \sim \mathcal{N}(0,1) \\ &= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}) \end{aligned}$$

Exemple :
On reprend l'exemple du TD 2 (cf. graphique de la loi), alors, on veut calculer

$$proba = P(X \in [9.98; 10.02])$$

pour $X \sim \mathcal{N}(9.9994, 0.03^2)$, d'où la réduction suivante :

$$\begin{aligned} proba &= \Phi(\frac{10.02 - 9.9994}{0.03}) - \Phi(\frac{9.98 - 9.9994}{0.03}) \\ &\approx \Phi(0.687) - \Phi(-0.646) \\ &\approx \Phi(0.687) - (1 - \Phi(0.646)) \approx 0.49 \end{aligned}$$

De même, pour l'intervalle $]-\infty;0]$ il est évident que, pour f la densité de probabilité de la loi normale centrée réduite, avec $X \sim \mathcal{N}(0, 1^2)$, on a :

$$\Phi(0) = P(X \in]-\infty;0]) = \int_{-\infty}^{0} f(t)dt = 0.5$$

puisque cette même densité est symétrique par rapport à la droite des ordonnées. Et, la médiane coïncide avec la moyenne.

Ensuite, on peut s'intéresser à la loi de la somme de v.a., notamment le cas gaussien qui demeure gaussien, comme considéré dans la section suivante.

1.5 Somme de variables aléatoires indépendantes

Définition

Soient X_1 et X_2 deux variables aléatoires discrètes, prenant les valeurs

$$X_1(\Omega) = \{0, 1, 2, \cdots m_1\} \text{ et } X_2(\Omega) = \{0, 1, 2, \cdots, m_2\}$$

Elle sont dites indépendantes ssi pour tout $k_1 \in X_1(\Omega)$ et $k_2 \in X_2(\Omega)$:

$$P\{(X_1 = k_1) \cap (X_2 = k_2)\} = P(X_1 = k_1) \times P(X_2 = k_2)$$

En remarque, ce résultat s'étend aux variables aléatoires continues en considérant tous les intervalles I_1 et I_2 (au lieu d'entiers) et leur probabilités, soit le produit $P(X_1 \in I_1) \times P(X_2 \in I_2)$ égal à $P\{(X_1 \in I_1) \cap (X_2 = I_2)\}$.
Disposant d'un échantillon, le test du χ^2 vérifie si l'hypothèse d'indépendance doit être rejetée ou non d'après le tableau des effectifs observés.

Propriétés (admises)

Pour a_1 et a_2, deux réels, et X_1 et X_2 deux v.a. indépendantes, alors :

$$\begin{aligned} E(a_1X_1 + a_2X_2) &= a_1E(X_1) + a_2E(X_2) \\ V(a_1X_1 + a_2X_2) &= a_1^2V(X_1) + a_2^2V(X_2) \\ \sigma(a_1X_1 + a_2X_2) &= \sqrt{a_1^2V(X_1) + a_2^2V(X_2)} \end{aligned}$$

On a clairement le même résultat sur une somme de n variables aléatoires indépendantes mutuellement (et non seulement deux à deux).

Définition

Soit une série de n v.a. discrètes $(X_i)_n = (X_1, X_2, \cdots, X_n)$, alors **les X_i sont (mutuellement) indépendantes** ssi le produit défini pour le cas de deux variables indépendantes est généralisé aux n variables à la fois, soit en prenant les n k_i dans les images $X_i(\Omega)$, on a :

$$P\{(X_1 = k_1) \cap (X_2 = k_2) \cap \cdots \cap (X_n = k_n)\} = P(X_1 = k_1) \times \cdots \times P(X_n = k_n)$$

Cette définition s'étend aux v.a. continues en prenant les intervalles au lieu des entiers. Attention : l'indépendance deux à deux n'implique pas l'indépendance mutuelle.

Propriétés (admises)

Pour a_i une suite de réels, et $(X_i)_n$ une série de n v.a indépendantes, alors :

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) \\ V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n) \\ \sigma(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= \sqrt{a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n)} \end{aligned}$$

Définition : **la série $(X_i)_n$ est i.i.d.** si les n v.a. X_i sont indépendantes identiquement distribuées : lorsque les X_i sont mutuellement indépendantes et ont toutes la même loi qu'une v.a. unique X, i.e. $P_{X_i} = P_X$. Cette notion est très utile en estimation/test.

1.5.1 Somme de v.a. de lois de Bernoulli

La loi binomiale donne la probabilité du nombre k de 1 dans une série de n valeurs 0 ou 1, chacune étant tirée aléatoirement et indépendamment suivant $B(p)$, une loi de Bernoulli de même paramétre p.

• Fonction de densité, de répartition (notée $F_{\mathcal{B}(n,p)}$), espérance, variance.

• (Propriété) $Y \sim Bin(n,p) \Longleftrightarrow \left\{ \begin{array}{l} \text{On définit } n \text{ v.a. } X_i \sim B(p) \text{ i.i.d,} \\ \text{et on pose } Y = \sum_{i=1}^{i=n} X_i \end{array} \right.$

Remarque :
Pour le cas de la loi binomiale de paramètres n et p, on a alors :

$$Y = \sum_{i=1}^{i=n} X_i = X_1 + X_2 + \cdots X_n \text{ avec les } X_i \text{ i.i.d. de même loi } B(p)$$

Ainsi, pour chaque X_i, on a $X_i \in \{0,1\}$ et,

$$\begin{array}{rcl} P(X_i = 1) & = & 1 - P(X_i = 0) = p \\ E(X_i) & = & p \\ V(X_i) & = & p(1-p) \end{array}$$

d'où :

$$\begin{array}{rcl} E(Y) & = & E(\sum_{i=1}^{i=n} X_i) \\ & = & E(X_1 + X_2 + \cdots X_n) \\ & = & E(X_1) + E(X_2) + \cdots + E(X_n) \\ & = & p + p + \cdots + p \\ & = & np \end{array}$$

$$\begin{array}{rcl} V(Y) & = & V(\sum_{i=1}^{i=n} X_i) \\ & = & V(X_1 + X_2 + \cdots + X_n) \\ & = & V(X_1) + V(X_2) + \cdots + V(X_n) \\ & = & p(1-p) + p(1-p) + \cdots + p(1-p) \\ & = & np(1-p) \end{array}$$

Cette loi est également tabulée, et pour les exercices on utilise souvent une approximation gaussiennne qui donne une valeur approchée de la fonction de répartition.

Propriétés (admises)

Soit $X \sim \mathcal{B}(n,p)$ une loi binomiale, alors pour n suffisamment grand :

$$X \sim \mathcal{N}(np, np(1-p)) \qquad \text{approximativement}$$

1.5.2 Somme de v.a. de lois gaussiennes

Propriétés (admises)

Pour a_1 et a_2, deux réels, et X_1 et X_2 deux v.a. indépendantes de lois gaussiennes :

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ et } X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Alors :

$$Y = a_1 X_1 + a_2 X_2 \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

avec :

$$\begin{aligned} \mu_Y &= a_1\mu_1 + a_2\mu_2 \\ \sigma_Y &= \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2} \end{aligned}$$

Remarque :

Le résultat sur les moments (espérance et variance) n'est pas nouveau et provient de la propriété précédente plus générale de la somme de v.a. puisque :

$$\begin{array}{rclcrcl} E(X_1) & = & \mu_1 & \text{et} & E(X_2) & = & \mu_2 \\ V(X_1) & = & \sigma_1^2 & \text{et} & V(X_2) & = & \sigma_2^2 \end{array}$$

D'où $E(Y)$ et $V(Y)$ pour $Y = a_1X_1 + a_2X_2$. Par contre, le fait que la combinaison linéaire de deux v.a. gaussiennes reste une v.a. gaussienne est à retenir. Cette propriété se généralise clairement à une combinaison linéaire de n v.a. gaussiennes mutuellement indépendantes, puisque cette combinaison reste également toujours gaussienne.

1.5.3 Somme de v.a. de lois quelconques

Propriétés (admises)

Soit $(X_i)_n$ une série de n v.a. réelles indépendantes de même loi que X,

$$E(X_i) = E(X) = \mu \text{ et } V(X_i) = V(X) = \sigma^2$$

Si on pose

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Alors pour n suffisamment grand,

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad \text{approximativement}$$

2 Estimations et Tests (inférence statistique)

Cf. Suite du Cours.

Table des matières

Cours d'Estimation et Test Paramétriques.

(R. PRIAM)

1 Estimation paramétrique

Ce cours porte sur l'estimation de la moyenne μ d'un caractère dans une population, ou bien la variance σ^2 de ce caractère, ou bien sa probabilité p en tant qu'évènement.

La théorie de l'estimation paramétrique répond à la question suivante.

Question :
On dispose de données - un échantillon de taille n - provenant d'une population. On se demande, à partir de cet échantillon, comment calculer une valeur approchant le mieux possible la moyenne, variance ou proportion d'un caractère de la population totale.

On procèdera à une estimation ponctuelle qui fournit une "unique valeur" puis une estimation par intervalle de confiance qui fournit une "fourchette de valeurs" à partir des données à disposition, un échantillon de n individus :

$$\underline{x}_n = (x_1, x_2, \cdots, x_n) \text{ avec } x_i \in \mathbb{R} \text{ ou } x_i \in \{0, 1\}$$

Cet échantillon est prélevé par tirage avec remise dans une population de taille quelconque ou par tirage sans remise dans une population de taille bien plus grande que la taille de l'échantillon. Dans ces deux cas, on peut considérer les individus comme prélevés de façon indépendante : les variables aléatoires X_i associées à chacune des mesures x_i sont de ce fait mutuellement indépendantes.
On emploie le terme d'échantillon aléatoire pour un échantillon constitué d'individus pris au hasard dans la population.

La théorie de l'échantillonnage consiste à déterminer les propriétés d'un échantillon prélevé dans la population, connaissant les propriétés de cette population totale.
En réalité, on rencontre le plus souvent le problème inverse, celui de l'estimation : on possède des renseignements sur un échantillon, et on cherche à déduire des informations sur la population totale.

En première approche l'estimation ponctuelle fournira une unique valeur mais l'erreur commise comparativement à la vraie valeur estimée (moyenne, variance ou proportion) n'est pas forcément explicite. Au contraire, l'estimation par intervalle de confiance fournit un intervalle dans lequel le vrai paramètre inconnu est fort probablement dedans comme l'explique le cours.

1.1 Vocabulaire

Echantillon aléatoire de X
Soit X une variable aléatoire réelle discrète[1], ou continue[2]. On appelle échantillon aléatoire de loi X un n-uplet de variables aléatoires indépendantes et de même loi que X. Parfois nommé $(X_i)_n$ ou $(X_i)_{1\leq i\leq n}$, on le note dans la suite :

$$\underline{X}_n = (X_1, X_2, \cdots, X_n)$$

pour une taille d'échantillon $n \in \mathbb{N}$.

Echantillon observé de valeurs de X
L'échantillon observé de valeurs de X est le n-uplet :

$$\underline{x}_n = (x_1, x_2, \cdots, x_n)$$

qui est une réalisation de l'échantillon aléatoire $\underline{X}_n$. Il s'agit des données mesurées, celles dont on dispose pour effectuer l'estimation et/ou le test.

Famille paramétrique

On représente une famille paramétrique de lois de probabilité par :

$$\{P_\theta : \theta \in \Theta\} \quad \text{avec } \Theta \subset \mathbb{R}^d$$

On dit que :
- θ est le paramètre,
- Θ est l'ensemble des valeurs admissibles du paramètre,
- d est la dimension du paramètre.

Remarque :
Soit X une v.a. réelle. On modélise le fait que P, la loi de X, n'est pas exactement connue en définissant la loi P comme étant un membre de la famille paramétrique précédente, avec θ inconnu ("à estimer"), $P(X) = P_\theta(X)$.
La question qui se pose alors en statistique paramétrique est à partir d'un échantillon $\underline{x}_n$ de trouver ("estimer") une approximation de ce vrai paramètre θ. Cette approximation est appelée "estimation" et notée $\hat{\theta}$.

Exemple :
La famille paramétrique de la loi de Bernoulli s'écrit :

$$\{P_\theta(X = x) = \theta^x(1-\theta)^x : \theta \in [0;1]\}$$

[1]Dans ce cours, une v.a. discrète prend seulement un nombre fini de valeurs positives donc on a $X(\Omega) = \{0, 1, \cdots, m\} \subset \mathbb{N}$. La cas positif dénombrable sera également très rarement abordé.

[2]Dans le cas d'une v.a. continue, on rappelle que $X(\Omega) \subset \mathbb{R}$

Ici, le paramètre inconnu θ est la probabilité de succès p, d'ailleurs :

$$\begin{aligned} P_\theta(X=0) &= 1-\theta \\ P_\theta(X=1) &= \theta \end{aligned}$$

Soit un échantillon observé de valeurs de X, des données, tel que le i-ème individu est $x_i \in \{0,1\}$. On peut voir par une simulation (cf. TDs) qu'une bonne valeur approchant le vraie θ est apparemment la moyenne des x_i, soit $\bar{x}_n$ notée f_n dans le cas discret. Ceci est confirmé et justifié par la suite du cours.

Exemple :
La famille paramétrique des lois gaussiennes unidimensionnelles a pour paramètres $\theta = \{\mu, \sigma\} \in \mathbb{R} \times \mathbb{R}_+^*$.

(Une) Statistique

Une statistique T_n est une v.a. (réelle) fonction des v.a. de $\underline{X}_n$, l'échantillon aléatoire de X :

$$T_n(\underline{X}_n) = \tau(X_1, \ldots, X_n) \ .$$

On remarque que la statistique T_n, sur un échantillon observé de valeurs de X, prend ses valeurs dans $\mathbb{R}^d$, soit pour le cas étudié plus loin, $d = 1$:

$$\begin{array}{rccc} T_n : & X(\Omega)^n & \to & \mathbb{R} \\ & \underline{x}_n = (x_1, x_2, \cdots, x_n) & \mapsto & T_n(\underline{x}_n) = \tau(x_1, x_2, \cdots, x_n) \end{array}$$

Ceci est une fonction à plusieurs variables, comme par exemple f(x,y)=x+y qui associe a une abcisse x et une ordonnée y, une hauteur $h = f(x,y)$. Elle associe donc à n valeurs réelles (puisque $X(\Omega) \subset \mathbb{R}$), une valeur réelle. Un exemple classique est la moyenne sur un échantillon.

Exemple :
Par exemple, on pose la v.a. $\bar{X}_n$:

$$T_n(\underline{X}_n) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$\bar{X}_n$ est bien une statistique, et dans ce cas :

$$T_n(\underline{x}_n) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \bar{x}_n$$

On remarque puisque les X_i sont mutuellement indépendants et de même loi que X, que l'on a $E(\bar{X}_n) = E(X)$ et $V(\bar{X}_n) = V(X)/n$.

Remarque :
On suppose disposer d'un échantillon observé (les données !) d'une population. Pour trouver la vraie valeur d'une moyenne, variance, ou probabilité, d'un caractère sur la population totale, on rencontrera deux cas :
- les individus de l'échantillon (et la population) suivent une certaine loi paramétrique de paramètre θ inconnu, gaussienne ou de Bernoulli. Le vrai paramètre inconnu coïncidera avec la valeur recherchée,
- les individus de l'échantillon sont assez nombreux pour appliquer le théorème central limite (pour $\bar{X}_n$), et aucune hypothèse de loi n'est nécessaire.

1.2 Estimation ponctuelle

1.2.1 Définition

Estimateur[3]

On dit que la statistique T_n est un estimateur du paramètre $\theta \in \Theta$ si :

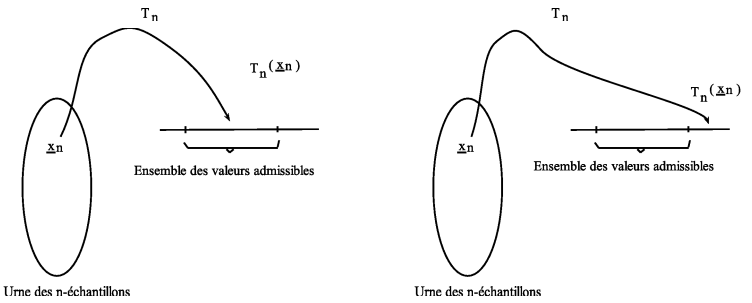
$$T_n(\underline{x}_n) \in \Theta \quad \text{pour tout échantillon } \underline{x}_n \text{ de valeurs de } X$$

Exemple :
Par exemple, pour le cas du modèle de Bernoulli, $\theta = p$ est une probabilité, donc il faut que $T_n(\underline{x}_n) \in [0;1]$ pour tous les échantillons $\underline{x}_n$ (avec $x_i \in X(\Omega)$) afin de pouvoir dire que T_n est une estimateur de p. Par exemple, la statistique $\bar{X}_n$ est un estimateur de p puisque il est clair que l'on a bien :

$$0 \leq \bar{x}_n \leq 1 \text{ si } x_i \in \{0,1\} \text{ pour tout i}$$

Graphiquement :
Sur le graphique de gauche, T_n prend bien sa valeur dans l'ensemble des valeurs admissibles Θ du paramètre θ, ici représenté par un intervalle de $\mathbb{R}$, tandis que sur le graphique de droite la statistique prend sa valeur en dehors de l'intervalle donc ce ne peut être un estimateur de θ. Par exemple, cette situation s'illustre pour Θ l'intervalle $[0;1]$ dans le cas où θ est une probabilité : l'estimateur ne peut prendre la valeur 2 par exemple !

[3]En réalité, avec tolérance aux bords, ce qui signifie que si par exemple $\Theta =]a;b[$ alors on remplace ici Θ par [a ;b], donc on inclut les bords de l'intervalle.
Pour rappel, la statistique T_n est une fonction de l'échantillon aléatoire de X, c'est à dire $\underline{X}_n = (X_1, X_2, \cdots, X_n)$, avec les X_i indépendants mutuellement.



Estimation (ponctuelle)

Soient T_n un estimateur de θ, et $\underline{x}_n$ un échantillon observé de valeurs de X. La valeur réelle $T_n(\underline{x}_n)$,

$$T_n(x_1, \cdots, x_n) \text{ est une estimation de } \theta$$

On notera dans la suite $\hat{\theta}$, une estimation d'un paramètre θ.

Question :
On peut se demander ce qu'est un "bon estimateur" (donc estimation), et quelles sont les propriétés à étudier pour le qualifier ainsi ? Car pour un même θ, il peut exister plusieurs estimateurs, donc comment choisir le meilleur ?

1.2.2 Propriétés

Un bon estimateur doit ainsi être (asymptotiquement) sans biais.

Biais d'un estimateur

On appelle biais de l'estimateur T_n par rapport à θ la quantité :

$$B(T_n, \theta) = E[T_n - \theta] = E[T_n] - \theta \ .$$

Estimateur sans biais

Un estimateur T_n du paramètre θ est dit :

- sans biais si
$$E[T_n] = \theta,$$
- asymptotiquement sans biais si
$$\lim_{n \to +\infty} E[T_n] = \theta.$$

Exemple : dans le cas de la variance et moyenne, pour X gaussien, on a le tableau suivant (expression des estimateurs à la section suivante) :

Paramètre θ	μ	σ^2	σ^2
Estimateur T_n	$\bar{X}_n$	S_n^2	$S_n^{2\prime}$
$E(T_n)$	$E(T_n) = \mu$	$E(S_n^2) = \frac{n-1}{n}\sigma^2$	$E(S_n'^2) = \sigma^2$
Biais ?	sans biais	avec biais	sans biais

Un bon estimateur doit être convergent.

Estimateur convergent

On dit que l'estimateur (T_n) est convergent vers θ si :

$$\lim_{n\to+\infty} E(T_n) = \theta \text{ et } \lim_{n\to+\infty} V(T_n) = 0$$

Remarque :
Cette définition de la convergence est clairement très intuitive et naturelle. La v.a. T_n, estimateur de θ[4] sera dit convergent si on a la propriété suivante : plus la taille des échantillons $\underline{x}_n$ grandit, plus l'estimation $T_n(\underline{x}_n)$ a de fortes chances d'être proche de la valeur à estimer et inconnue θ.

Enfin, entre deux estimateurs convergents, on choisira celui dont la variance est la plus faible (à partir d'un certain entier n), car intuitivement l'estimation résultante sera alors probablement la plus proche de θ dans ce cas.

Remarque :
En effet, dans le cas gaussien, on peut représenter la distribution de deux fonctions de densités de lois gaussiennes avec même espérance et variance différentes. Visuellement, il est clair que dans ce cas il y a "davantage d'individus" proches de la moyenne si la variance est plus petite. Pour le vérifier, on peut calculer la probabilité sur un intervalle symétrique centré en la moyenne.

En résumé, un estimateur est un **bon estimateur de** θ si il :
- prend uniquement ses valeurs parmi celles admissibles par θ ;
- est sans biais ou asymptotiquement sans biais ;
- est convergent vers θ pour assurer que,

lorsque la taille l'échantillon augmente, l'estimation s'améliore ;
- a la plus faible variance possible.

Ces propriétés peuvent se représenter graphiquement.

Dans la suite, les estimations ponctuelles et par intervalles de confiance seront supposées avoir ces bonnes propriétés.

Les estimateurs de la moyenne, variance, et proportion d'une population sont présentés, puis les estimateurs de la différence de deux moyennes, deux proportions et du rapport de deux variances sont développés.

[4](moyenne, variance, probabilité d'une population ou bien le paramètre d'une famille paramétrique)

1.2.3 Tableaux

Cas classiques

Paramètres	Estimateurs sans biais	Estimations ponctuelles	Données
Moyenne $\mu = E[X]$	$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$	$x_i \in \mathbb{R}$
Variance $\sigma^2 = V[X]$	$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$	$s_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$	$x_i \in \mathbb{R}$
Proportion $p = E[X]$	$F_n = \frac{K_n}{n}$ où $K_n = \sum_{i=1}^{n} X_i$	$f_n = \frac{k_n}{n}$ et $k_n = \sum_{i=1}^{n} x_i$	$x_i \in \{0,1\}$

Remarques et propriétés

Moyenne empirique	Variance empirique non corrigée	Proportion empirique
$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$	$F_n = \frac{\sum_{i=1}^{n} X_i}{n}$
$E[\overline{X}_n] = E[X] = \mu$	$E[S_n^2] = \frac{n-1}{n}\sigma^2$	$E[F_n] = p$
$V[\overline{X}_n] = \frac{V[X]}{n} = \frac{\sigma^2}{n}$	$V[S_n^2] \approx \frac{n-1}{n^3}[(n-1)\mu_4 + (3-n)\sigma^4]$	$V[F_n] = \frac{p(1-p)}{n}$

Théorème
La moyenne empirique $\overline{X}_n$ est un estimateur (sans biais) convergent de l'espérance de X.
La variance empirique corrigée S'^2_n (parfois notée $\tilde{S}_n^2$) est un estimateur (sans biais) convergent de la variance de X.
La variance empirique S_n^2 est un estimateur (asymptotiquement sans biais) convergent de la variance de X.

Méthodologie : estimation ponctuelle
Pour estimer ponctuellement en pratique, à partir d'un échantillon $\bar{x}_n$, une moyenne, une variance ou une proportion on évalue respectivement en remplaçant par les valeurs numériques des x_i :

$$\hat{\mu} = \overline{x}_n, \quad \hat{\sigma^2} = \tilde{s}_n^2, \; et \; \hat{p} = f_n$$

On sait d'après la théorie que les valeurs obtenues sont des *estimation ponctuelles* qui pour n grand sont de plus en plus proches de la vraie moyenne, variance ou proportion suivant le cas.

En remarque, la simple valeur de l'estimation ponctuelle ne donne pas d'indication sur l'erreur effectuée, c'est pourquoi, dans la section suivante, l'estimation par intervalles de confiance est présentée.

1.3 Estimation par intervalle de confiance

Un intervalle de confiance au niveau $1-\alpha$ est un intervalle dans lequel le vrai paramètre θ a une probabilité $1-\alpha$ de se trouver. L'estimation par intervalle de confiance se justifie également par le fait que finalement pour deux échantilons différents, l'estimation ponctuelle aboutira a deux valeurs différentes, donc chercher l'ensemble des estimations les plus probables pour un paramètre est pertinent. On note T_n un estimateur de θ.

Intervalle de confiance

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ. Un intervalle de confiance pour le paramètre θ au niveau de confiance $1-\alpha$ $(0<\alpha<1)$, issu d'un échantillon de valeurs de X noté $\underline{x}_n$ est un intervalle de type :

$$I(\underline{x}_n)) = [a(\underline{x}_n); b(\underline{x}_n)]$$

où on a :

$$\begin{aligned} a(\underline{x}_n) &= A_n(x_1, x_2, \cdots, x_n) \\ b(\underline{x}_n) &= B_n(x_1, x_2, \cdots, x_n) \end{aligned}$$

avec A_n et B_n deux variables aléatoires fonction de T_n, telles que :

$$P(A_n \leq \theta \leq B_n) = 1-\alpha$$

Remarque :

• Généralement, pour construire ces intervalles qui sont symétriques, on encadre d'abord la statistique T_n (ou une fonction de T_n) en cherchant deux réels tels que :

$$P(T_n < c) = \frac{\alpha}{2} \quad \text{et} \quad P(T_n > d) = \frac{\alpha}{2}$$

Et on obtient A_n et B_n alors en faisant passer θ du bon coté des inégalités ainsi constituées, puisque on sait que

$$P(c \leq T_n \leq d) = 1-\alpha$$

• Par abus de langage, on dira que pour un intervalle $[a(\underline{x}_n); b(\underline{x}_n)]$ calculé sur un unique échantillon, que l'on a $100(1-\alpha)$ % de chances que le vrai paramètre θ se trouve dedans, alors qu'en réalité, cette probabilité fait référence à l'intervalle constitué des variables aléatoires A_n et B_n.

• Par contre, sur 100 intervalles calculés sur 100 échantillons de valeurs de X, il y aura $100(1-\alpha)$ d'entre eux qui contiendront le vrai paramètre θ.

Exemple : Construction de l'intervalle de confiance pour $X \sim \mathcal{N}(\mu, \sigma)$

On suppose ici σ connu. L'estimation du paramètre inconnu μ par intervalle de confiance de niveau $1-\alpha$ s'écrit :

$$I(\underline{x}_n) = [\bar{x}_n - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x}_n + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

Demonstration :
Soit $(X_1, X_2, \cdots, X_n)$, l'échantillon aléatoire de X de taille n, donc tel que, $E(X_i) = \mu$ et $V(X_i) = \sigma^2$. On pose la statistique :

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Et, directement, en cherchant un intervalle symétrique de centre $\bar{X}_n$, on a avec $r > 0$:

$$\begin{array}{lllll}
 & P(\bar{X}_n - r \leq \mu \leq \bar{X}_n + r) & = & 1-\alpha & \\
\Rightarrow & P(-\frac{r}{\sigma/\sqrt{n}} \leq \frac{\mu - \bar{X}_n}{\sigma/\sqrt{n}} \leq +\frac{r}{\sigma/\sqrt{n}}) & = & 1-\alpha & \\
\Rightarrow & P(-\frac{r}{\sigma/\sqrt{n}} \leq \frac{X_n - \mu}{\sigma/\sqrt{n}} \leq +\frac{r}{\sigma/\sqrt{n}}) & = & 1-\alpha & \\
\Rightarrow & P(-\frac{r}{\sigma/\sqrt{n}} \leq Z \leq +\frac{r}{\sigma/\sqrt{n}}) & = & 1-\alpha & \text{avec } Z \sim \mathcal{N}(0,1) \\
\Rightarrow & P(Z \leq +\frac{r}{\sigma/\sqrt{n}}) - P(Z \leq -\frac{r}{\sigma/\sqrt{n}}) & = & 1-\alpha & \text{avec } \Phi(z) = P(Z \leq z) \\
\Rightarrow & \Phi(\frac{r}{\sigma/\sqrt{n}}) - \Phi(-\frac{r}{\sigma/\sqrt{n}}) & = & 1-\alpha & \text{avec } \Phi(-z) = 1 - \Phi(z) \\
\Rightarrow & 2 \times \Phi(\frac{r}{\sigma/\sqrt{n}}) - 1 & = & 1-\alpha & \\
\Rightarrow & \Phi(\frac{r}{\sigma/\sqrt{n}}) & = & 1-\frac{\alpha}{2} &
\end{array}$$

D'où en prenant $t_{1-\frac{\alpha}{2}}$ tel que $\Phi(t_{1-\frac{\alpha}{2}}) = P(Z \leq t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, il vient :

$$\frac{r}{\sigma/\sqrt{n}} = t_{1-\frac{\alpha}{2}}$$

Soit,

$$r = t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Finalement (et l'intervalle de confiance correspondant) on obtient :

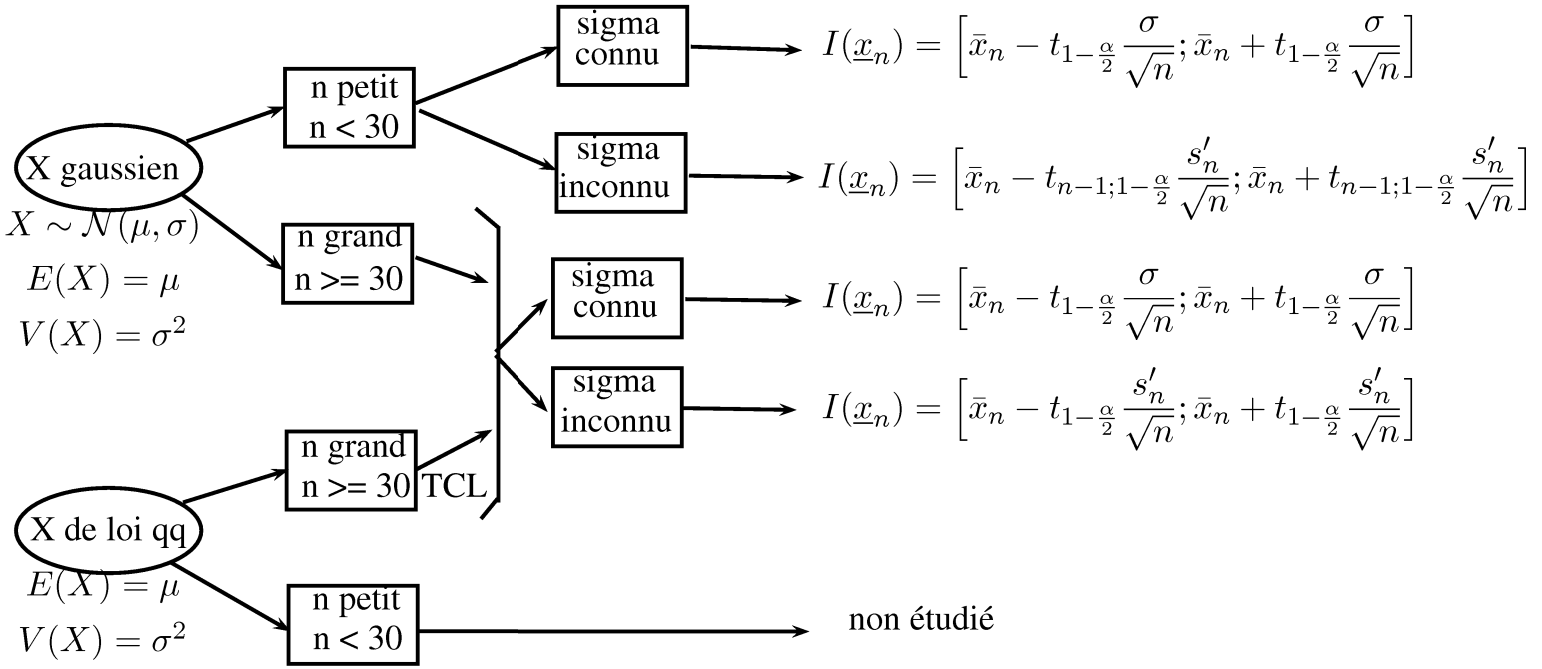
$$P(\bar{X}_n - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Cette démonstration de la construction de l'intervalle est donnée à titre indicative. On se réferrera à l'ouvrage de référence du cours pour les autres intervalles ou une démonstration différente.

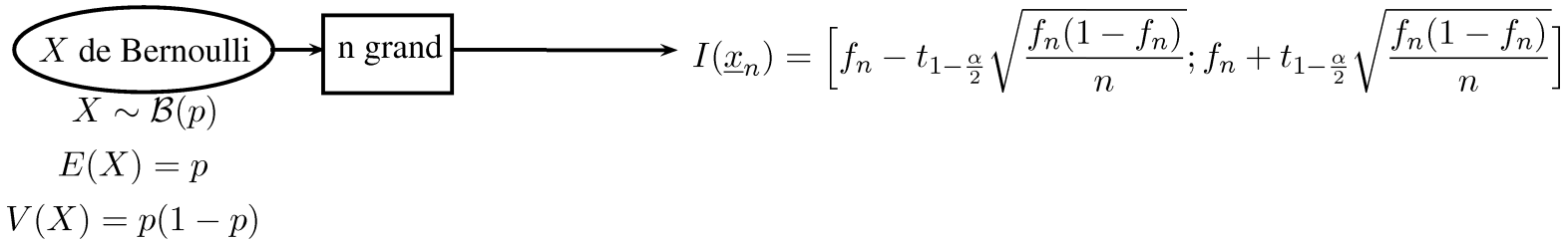
Le formulaire suivant indique les intervalles de confiance utiles en estimation, pour encadrer une moyenne (plusieurs cas possibles), ainsi qu'une proportion.

Soit un échantillon $\underline{x}_n = (x_1, x_2, \cdots, x_n)$.

Estimation de la moyenne μ (d'une population) par intervalle de confiance au niveau $1 - \alpha$



Estimation de la proportion p (d'une population) par intervalle de confiance au niveau $1 - \alpha$



Estimation ponctuelle : $f_n = \bar{x}_n = \dfrac{1}{n} \displaystyle\sum_{i=1}^{i=n} x_i$

Remarque : intervalle de confiance pour une variance

Pour une variance, on ne traite que le cas gaussien, donc $X \sim \mathcal{N}(\mu, \sigma)$, et pour un échantillon $\underline{x}_n = (x_1, x_2, \cdots, x_n)$ de valeurs de X, l'intervalle de confiance au niveau $1-\alpha$ se calcule comme suit :

$$I(\underline{x}_n) = \left[\frac{(n-1)s_n^{2\prime}}{\chi^2_{n-1;1-\frac{\alpha}{2}}}; \frac{(n-1)s_n^{2\prime}}{\chi^2_{n-1;\frac{\alpha}{2}}} \right]$$

avec $\chi^2_{n-1;\beta}$ le quantile d'ordre β de la loi du χ^2 à $n-1$ degrés de liberté,

$$P(Z \leq \chi^2_{n-1;\beta}) = \beta \text{ avec } Z \sim \chi^2_{n-1}.$$

Dans le cas de la variance, la statistique utilisée n'est plus $T_n = \bar{X}_n$ mais $T_n = S'_n$, et sa version "normalisée" $(n-1)S_n^{\prime 2}/\sigma^2$ qui suit une loi du χ^2 à $n-1$ degrés de liberté.

Dans la suite du cours, la concept de test statistique est défini, puis les principaux tests utiles sont développés. Il s'agira notamment
• d'après un seul échantillon de décider si la moyenne μ inconnue d'une population est égale à une valeur μ_0 candidate ou bien,
• d'après deux échantillons, de décider si les moyennes μ_1 et μ_2 de deux populations comparables, et dont sont tirées les échantillons, sont égales.

La notion de risque d'erreur lors de la prise de décision sera également introduite. Enfin, le lien entre estimation par intervalle de confiance de niveau $1-\alpha$ et test statistique paramétrique de *risque* α est abordé.

2 Tests paramétriques

D'une manière générale, en test d'hypothèse, on s'intéresse à préciser comment prendre des décisions concernant l'ensemble de la population, à partir de l'étude d'un ou plusieurs échantillons.
Etant donné qu'on ne dispose pas de renseignements sur l'ensemble de la population, on risque de se tromper en prenant la décision, et il est donc important de contrôler au maximum tout risque d'erreur.

Les définitions suivantes définissent le vocabulaire des tests statistiques. Ensuite, un exemple de test est explicité, juste avant un tableau récapitulatif.

2.1 Vocabulaire

Test statistique paramétrique

Un test statistique est une procédure (décision) basée sur une fonction (statistique) sur un ou deux échantillons, et conduisant à rejeter avec un certain risque d'erreur, une hypothèse nulle notée H_0.

Hypothèses nulle et alternative

L'hypothèse nulle H_0 porte sur la valeur d'un ou plusieurs paramètres de la population mère. ; il s'agit de l'hypothèse que l'on privilégie et que l'on cherche à valider. On parle d'hypothèse nulle car celle-ci peut souvent s'écrire $f(\theta) = 0$ pour θ le paramètre sur lequel porte le test.

L'hypothèse alternative H_1 est l'hypothèse que l'on oppose (confronte) à l'hypothèse nulle.

Erreurs de première et seconde espèce

L'erreur de première espèce est celle commise lorsqu'on rejette l'hypothèse nulle alors que celle-ci est vraie. La probabilité de cette erreur s'appelle le risque de première espèce et se note α.

L'erreur de seconde espèce est celle commise lorsqu'on ne rejette pas l'hypothèse nulle alors que celle-ci est fausse. La probabilité de cette erreur s'appelle le risque de seconde espèce et se note β.

La valeur $1 - \beta$ est appelée *puissance du test.*

Région critique

Soit un test d'hypothèse nulle H_0 et hypothèse alternative H_1.
Soit T_n la statistique associée au test, et soit un échantillon $\underline{x}_n$.
La région critique du test notée $\mathcal{R}$ est un intervalle ou une union de deux intervalles, tels que :

Si $T_n(\underline{x}_n) \in \mathcal{R} \subset \mathbb{R}$, alors, on rejette l'hypothèse H_0 au risque α.

Test de conformité (pour un échantillon)

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ.

• Pour un **test bilatéral symétrique**, soit l'hypothèse nulle, "H_0 le paramètre de la loi de X vaut θ_0", que l'on oppose à, "H_1 le paramètre de la loi de X est différent de θ_0", l'hypothèse alternative. On note alors,

$$H_0 : \ \theta = \theta_0 \text{ contre } H_1 : \ \theta \neq \theta_0$$

• Pour un **test unilatéral**, soit l'hypothèse nulle "le paramètre de la loi de X vaut θ_0" que l'on oppose à l'hypothèse alternative "le paramètre de la loi de X est supérieur à θ_0" ou bien "le paramètre de la loi de X est inférieur à θ_0". On note alors,

$$H_0 : \ \theta = \theta_0 \text{ contre } H_1 : \ \theta > \theta_0 \text{ ou bien } H_1 : \ \theta < \theta_0$$

Exemple :
Pour un test de conformité sur une moyenne, on écrit,

$$H_0 : \ \mu = \mu_0 \text{ contre } H_1 : \ \mu \neq \mu_0$$

Test de comparaison (pour deux échantillon)

Soit X et Y deux variables aléatoires dont les lois respectives dépendent d'un paramètre inconnu θ_1 et θ_2 avec T_{n_1} et T_{n_2} les estimateurs correspondants. Alors on définit un test de comparaison en remplaçant les hypothèses nulles par $\theta_1 - \theta_2 = 0$ pour des moyennes ou $\theta_1/\theta_2 = 1$ pour des variances.

Exemple :
Pour un test de comparaison de deux moyennes, on écrit,

$$H_0 : \ \mu_1 = \mu_2 \text{ contre } H_1 : \ \mu_1 \neq \mu_2$$

2.2 Tests de conformité

Exemple :
Un test de conformité bilatéral symétrique sur la moyenne s'écrit :

$$H_0 : \ \mu = \mu_0 \text{ contre } H_1 : \ \mu \neq \mu_0$$

La régle de décision, de risque α, dans le cas gaussien, avec σ connu, s'écrit :

"Si $\bar{x}_n \in [\mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \mu_0 + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ alors accepter H_0 sinon rejeter H_0".

Démonstration :
Dans le cas gaussien, avec la variance σ^2 connue, on s'intéresse à la v.a. X qui représente notre population totale :

$$X \sim \mathcal{N}(\mu, \sigma)$$

Donc, à l'aide de l'échantillon aléatoire de X de taille n, on pose la statistique :

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

En supposant vraie l'hypothèse H_0, il est clair (cf. rappels proba) que la loi de la statistique est alors :

$$\bar{X}_n \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

Il s'agit donc de la loi des moyennes empirique $\bar{x}_n$ des échantillons $\underline{x}_n = (x_1, x_2, \cdots, x_n)$. Si on détermine un intervalle dans lequel toute moyenne d'un échantillon a une très forte chance de se trouver, on peut poser comme règle de décision de rejeter l'hypothèse H_0 si la moyenne d'un échantillon a disposition ne se trouve pas dans cet intervalle.

Directement, en procédant comme pour la démonstration de l'intervalle de confiance, mais en encadrant non plus μ, mais à la place $\bar{X}_n$, on aboutit à l'intervalle suivant (cf. TDs) :

$$\bar{\mathcal{R}}_{acceptation} = [\mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \mu_0 + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

tel que $P(\bar{X}_n \in \bar{\mathcal{R}}_{acceptation}) = 1-\alpha$. Donc si l'hypothèse H_0 est vraie, alors la probabilité que la moyenne d'un échantillon se trouve dans le complémentaire de $\mathcal{R}_{acceptation}$, i.e. la région critique $\mathcal{R}$ (indicé par "rejet" ici) :

$$\mathcal{R}_{rejet} =]-\infty; \mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}[\ \cup \ ]\mu_0 + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; +\infty[$$

est α, donc faible puisque α est très petit ! Il s'agit de la probabilité de se tromper en rejetant l'hypothèse H_0 alors que celle-ci est vraie.

Remarque :
Dans le cas de la moyenne, le lien entre intervalle de confiance de niveau $1-\alpha$ et test de conformité de risque α est clair, en posant $u_\alpha = t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$:

$$
\begin{array}{ll}
 & \bar{x}_n \in \bar{\mathcal{R}}_{acceptation} \\
 & \bar{x}_n \in [\mu_0 - u_\alpha ; \mu_0 + u_\alpha] \\
\Leftrightarrow & \mu_0 - u_\alpha \leq \bar{x}_n \leq \mu_0 + u_\alpha \\
\Leftrightarrow & \mu_0 - u_\alpha - \bar{x}_n - \mu_0 \leq \bar{x}_n - \bar{x}_n - \mu_0 \leq \mu_0 + u_\alpha - \bar{x}_n - \mu_0 \\
\Leftrightarrow & -u_\alpha - \bar{x}_n \leq -\mu_0 \leq u_\alpha - \bar{x}_n \\
\Leftrightarrow & (-1) \times \Big(- u_\alpha - \bar{x}_n \Big) \geq (-1) \times \Big(- \mu_0 \Big) \geq (-1) \times \Big(u_\alpha - \bar{x}_n \Big) \\
\Leftrightarrow & +u_\alpha + \bar{x}_n \geq +\mu_0 \geq -u_\alpha + \bar{x}_n \\
\Leftrightarrow & \bar{x}_n - u_\alpha \leq \mu_0 \leq \bar{x}_n + u_\alpha \\
\Leftrightarrow & \mu_0 \in [\bar{x}_n - u_\alpha ; \bar{x}_n + u_\alpha] \\
\Leftrightarrow & \mu_0 \in I(\underline{x}_n)
\end{array}
$$

En effet, il y a $1-\alpha$ (par exemple 95 %) chance que le vrai paramètre μ soit dans l'intervalle de confiance $I(\underline{x}_n)$. Alors pour vérifier si une valeur μ_0 candidate est acceptable, il suffit de vérifier si μ_0 est dans l'intervalle de confiance $I(\underline{x}_n)$, et si tel est le cas on accepte l'hypothèse, $\mu = \mu_0$ qui parait vraisemblable. Il y a seulement α (par exemple 5%) chance de se tromper.

Enfin, ce résultat n'est pas vrai dans le cas général. La région critique (et d'acceptation) est alors en effet calculée en supposant H_0 vraie donc en utilisant H_0 (hypothèse que l'on cherche réfuter). Et généralement, l'expression de la région d'acceptation (et de rejet) diffère de celle de l'intervalle de confiance.

Remarque :
On a également le résultat suivant,

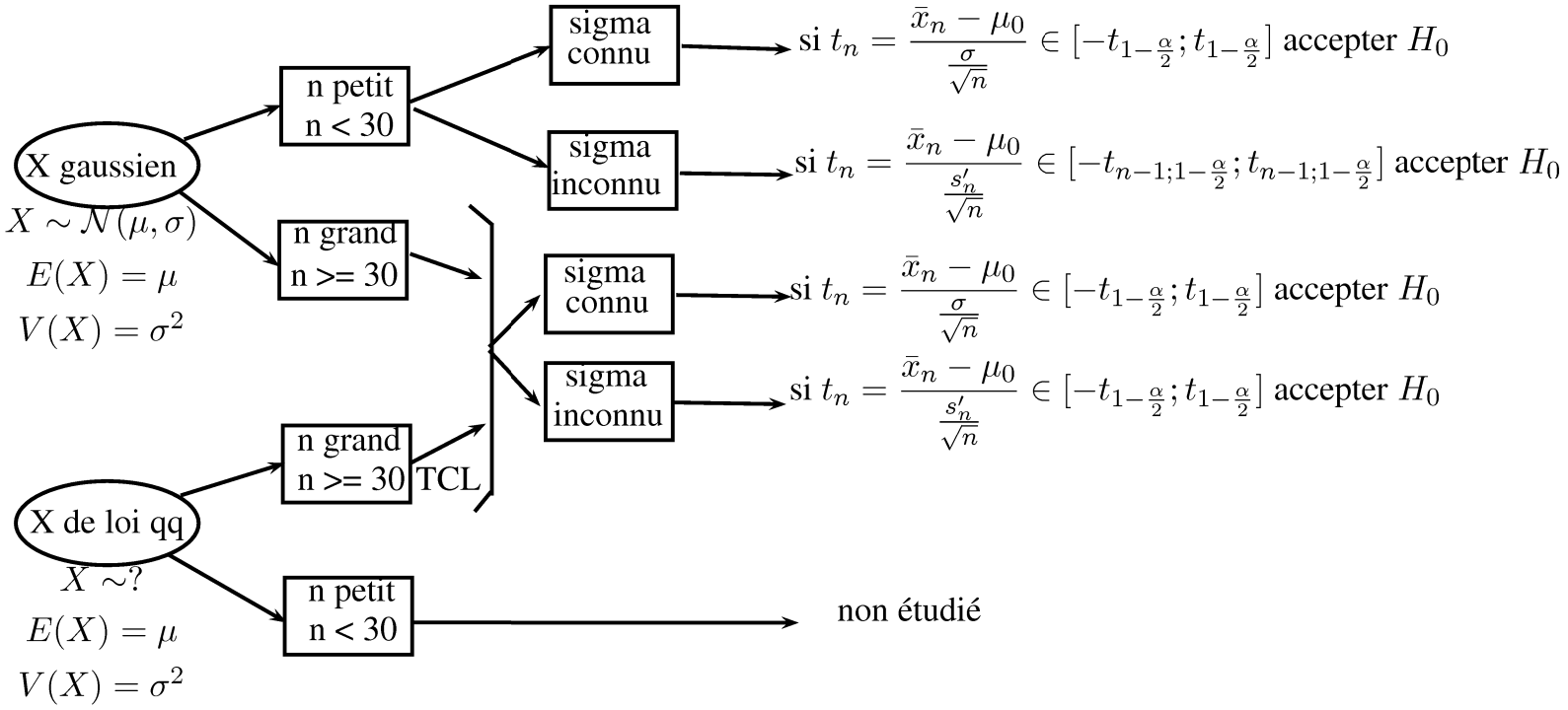
$$
\begin{array}{ll}
 & \bar{x}_n \in [\mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} ; \mu_0 + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \\
\Leftrightarrow & \mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{x}_n \leq \mu_0 + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\
\Leftrightarrow & -t_{1-\frac{\alpha}{2}} \leq \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq +t_{1-\frac{\alpha}{2}} \\
 & t = \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in [-t_{1-\frac{\alpha}{2}} ; +t_{1-\frac{\alpha}{2}}]
\end{array}
$$

Ce qui fait le lien entre la démonstration présentée ci-avant et la règle de décision communiquée dans le tableau récapitulatif.

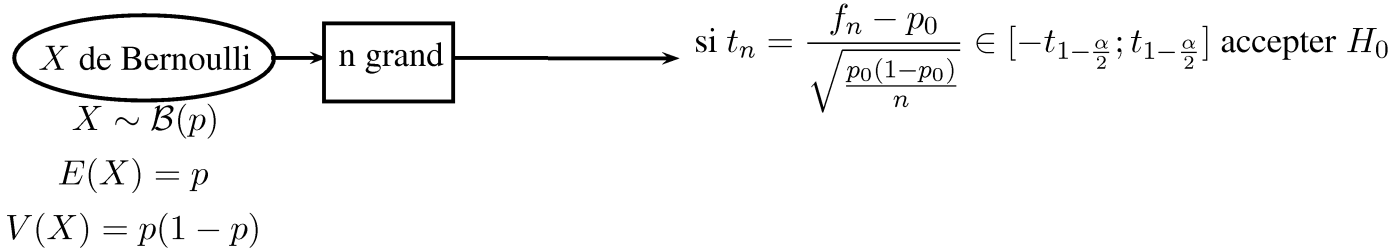
Schéma illustratif revu en TDs.

Soit un échantillon $\underline{x}_n = (x_1, x_2, \cdots, x_n)$.

Test bilatéral symétrique de conformité d'une moyenne H_0 : "$\mu = \mu_0$" contre H_1 : "$\mu \neq \mu_0$" au risque α



Test bilatéral symétrique de conformité d'une proportion H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$" au risque α



Estimation ponctuelle : $f_n = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{i=n} x_i$

2.3 Tests de comparaison

2.3.1 Exemple introductif

Problème :
Dans le même atelier on prélève au hasard et avec remise 40 fiches correspondant à 40 réparations de téléviseurs de la marque B. On obtient les résultats suivants présentés à la suite de ceux déjà obtenus pour 50 réparations de téléviseurs de marque A :

Durée interventions (mn)	[0 ;20[	[20 ;40[	[40 ;60[	[60 ;80[	[80 ;100[	[100 ;120[
Nb d'interventions (cas A)	1	6	11	15	10	7
Nb d'interventions (cas B)	3	7	12	9	6	3

On note les données de type A et de type B, avec $n_1 = 50$, $n_2 = 40$:
$\underline{x}_{n_1} = (x_1, x_2, \cdots, x_{n_1})$ et $\underline{y}_{n_2} = (y_1, y_2, \cdots, , y_{n_2})$

On calcule les moyennes respectives des deux échantillon : $\bar{x}_{n_1} = 69.2$ et $\bar{y}_{n_2} = 58.5$, en supposant que, dans chaque classe, tous les éléments sont au centre. Ensuite, on calcule les estimations ponctuelles des écart-types $s'_{n_1} = 25.86$ et $s'_{n_2} = 27.13$.

On peut se demander si la différence entre les deux moyennes empiriques $\bar{x}$ et $\bar{y}$ provient du fait qu'il s'agit de deux marques différentes donc deux populations de téléviseurs différentes (avec leur moyennes μ_1 et μ_2 différentes) ou bien du fait du prélèvement des échantillons. Ceci fera l'objet d'un test statistique de comparaison de deux moyennes.

Résolution :
Soient $\bar{X}_{n_1}$ (resp. $\bar{Y}_{n_2}$) la variable aléatoire qui a tout échantillon de taille $n_1 = 50$ (resp. $n_2 = 40$) prélevé aléatoirement et avec remise dans la population A (resp. B), associe la moyenne des durées d'intervention sur les téléviseurs de l'échantillon.

On se place dans le cas où ces deux v.a. sont indépendantes et (approximativement) gaussiennes mais de vraie loi inconnue (lois des X_i, Y_j inconnues).

En supposant inconnues (ou connues mais valeurs non renseignées) les vraies variances, on écrit qu'approximativement (TCL) puisque n_1, n_2 grands :

$$\bar{X}_{n_1} \sim \mathcal{N}\Big(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\Big) \text{ et } \bar{Y}_{n_2} \sim \mathcal{N}\Big(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\Big)$$

Si bien que la loi (approximative) de $D = \bar{X} - \bar{Y}$ est une loi normale :

$$D \sim \mathcal{N}\Big(\mu_D, \sigma_D\Big)$$

avec :

$$\mu_D = \mu_1 - \mu_2$$

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ensuite, on peut également utiliser les estimations des écart-types à la place des vraies valeurs, car les échantillons sont grands. On a donc approximativement (en remplaçant σ_1 par s'_{n_1} et σ_2 par s'_{n_2}) :

$$\sigma_D \approx \sqrt{\frac{{s'_{n_1}}^2}{n_1} + \frac{{s'_{n_2}}^2}{n_2}} = \sqrt{\frac{25.86^2}{50} + \frac{27.13^2}{40}} = 5.64$$

On s'intéresse alors à savoir si les vraies moyennes inconnues μ_1 et μ_2 sont identiques ou non, soit le test d'hypothèse,

$$H_0 : \ \mu_1 = \mu_2 \text{ contre } H_1 : \ \mu_1 \neq \mu_2$$

Pour décider entre ces deux hypothèses, on suppose H_0 vraie (i.e. en considérant $E(D) = \mu_1 - \mu_2 = 0$), et on cherche un réel positif r tel que :

$$P(-r \leq D \leq r) = 0.95$$

Donc, si H_0 est vraie, on aura 95% de chance que la différence des moyennes $d = \bar{x} - \bar{y}$ de deux échantillons empiriques se trouvent dans l'intervalle $[-r; +r]$, puisque D est la v.a. qui associe à deux échantillons, de taille $n_1 = 50$ et $n_2 = 40$ des type A et B, la différence de leur moyenne respective.

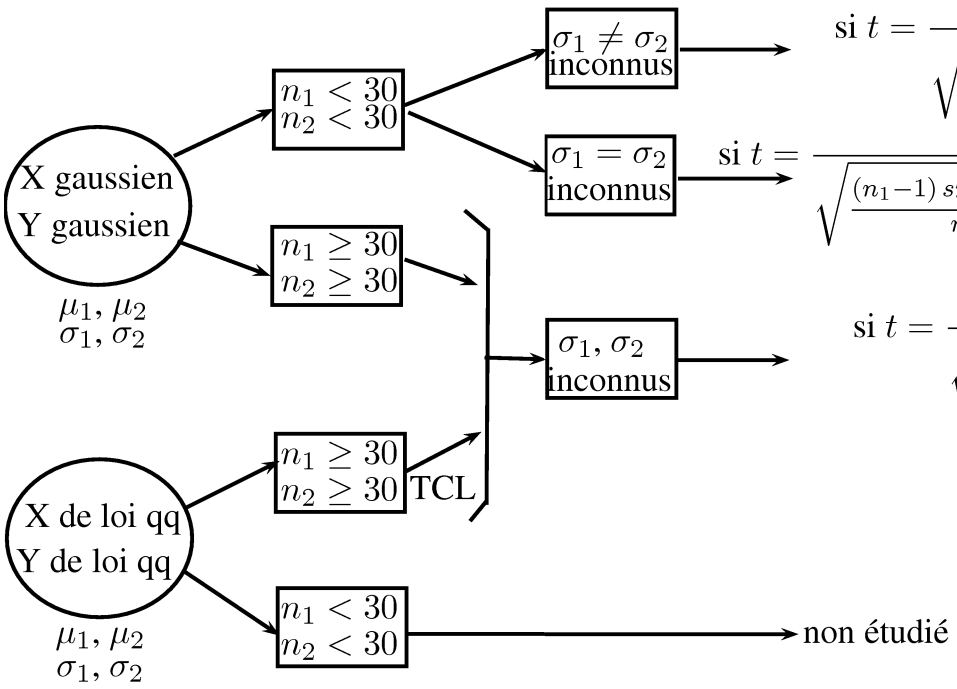
Une fois trouvé r (sous l'hypothèse $\mu_1 - \mu_2 = 0$), on décide donc d'accepter H_0 si la différence de moyennes $d = \bar{x} - \bar{y}$ est dans l'intervalle $[-r; +r]$ sinon on accepte H_1. Il y a toujours un risque de 5% de trouver la différence des moyennes en dehors de l'intervalle (il est construit pour !), on a donc 5% de chance de commettre une erreur et de rejeter l'hypothèse d'égalité des moyennes alors que celle-ci est vraie. C'est l'erreur de première espèce α, définie précédemment. Enfin, on peut représenter graphiquement le test.

Conclusion :
La solution finale (valeur de r et d) permet de prendre la bonne décision. En fait, pour des petits échantillons, une information (σ_1, σ_2 connus, et égaux ou non) sur les écart-types n'est pas anodine, car la loi de la v.a. à considérer est modifiée, il s'agit d'une loi de Student, d'ou le nom du test "t-test". Dans la section suivante, le tableau récapitulatif du test est présenté.

2.3.2 Tableau des tests

Soient deux échantillons $\underline{x}_{n_1} = (x_1, x_2, \cdots, x_{n_1})$ et $\underline{y}_{n_2} = (y_1, y_2, \cdots, , y_{n_2})$.

Test sur l'égalité des moyennes $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ de risque α.

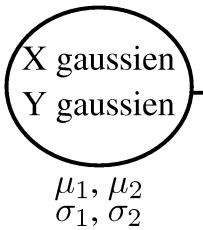


si $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\frac{{s'_{n_1}}^2}{n_1} + \frac{{s'_{n_2}}^2}{n_2}}} \in [\pm t_{\nu^*;1-\frac{\alpha}{2}}]$ accepter H_0

si $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_1-1)\,{s'_{n_1}}^2 + (n_2-1)\,{s'_{n_2}}^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in [\pm t_{n_1+n_2-2;1-\frac{\alpha}{2}}]$ accepter H

si $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\frac{{s'_{n_1}}^2}{n_1} + \frac{{s'_{n_2}}^2}{n_2}}} \in [\pm t_{1-\frac{\alpha}{2}}]$ accepter H_0

Test sur l'égalité des variances $H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$ de risque α.



si $t = max(\dfrac{{s'_{n_1}}^2}{{s'_{n_2}}^2}, \dfrac{{s'_{n_2}}^2}{{s'_{n_1}}^2}) < f_{\nu_a,\nu_b;1-\frac{\alpha}{2}}$ accepter H_0

avec $\nu_a = n_1 - 1$ et $\nu_b = n_2 - 1$ si ${s'_{n_1}}^2 > {s'_{n_2}}^2$ et inversement sinon.

$\nu^* \approx \left(\dfrac{{s'_{n_1}}^2}{n_1} + \dfrac{{s'_{n_2}}^2}{n_2}\right)^2 / \left(\dfrac{{s'_{n_1}}^4}{n_1^2(n_1-1)} + \dfrac{{s'_{n_2}}^4}{n_2^2(n_2-1)}\right)$ (valeur lue sous SAS)

Remarque :
Pour procéder à un test de comparaison de deux moyennes, on doit donc tester l'égalité des variances avant, afin de choisir la bonne statistique de test. En effet, en générale, les variances ne sont pas connues.
Par contre, si les variances sont connues, la seule formule intervenant est celle du cas gaussien classique (vu dans l'exemple introductif).

Test de comparaison de deux proportions

On suppose disposer de deux échantillons $(x_1, x_2, \cdots, x_{n_1})$ et $(y_1, y_2, \cdots, y_{n_2})$, avec les x_i provenant d'une loi de Bernoulli $\mathcal{B}(p_1)$ et les y_j provenant d'une loi de Bernoulli $\mathcal{B}(p_2)$.

Le test de risque α sur l'égalité de deux proportions s'écrit

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 \neq p_2 \end{aligned}$$

Soit

$$f_{n_1} = \frac{x_1 + x_2 + \cdots + x_{n_1}}{n_1} \text{ et } f_{n_2} = \frac{y_1 + y_2 + \cdots + y_{n_2}}{n_2}$$

$$f_{n_1+n_2} = \frac{n_1 f_{n_1} + n_2 f_{n_2}}{n_1 + n_2}$$

On applique alors la règle de décision suivante :

"Si

$$t = \frac{f_{n_1} - f_{n_2}}{\sqrt{f_{n_1+n_2}(1 - f_{n_1+n_2})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \in [-t_{1-\frac{\alpha}{2}}; t_{1-\frac{\alpha}{2}}]$$

alors accepter H_0,
sinon accepter H_1 au risque α."

Remarque :
La règle de test se construit tout comme le cas de la moyenne, en posant la différence des variables aléatoires qui associent à chacun des échantillons les fréquences, puis en utilisant une approximation gaussienne de leur loi (cf TD associé).

2.3.3 Exemple sous SAS

Graphiques illustrant les tests de comparaison.

Procédure du t-test et Interprétation des résultats sous SAS.

Table des matières

Feuille n°1 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1

On a relevé dans une entreprise le nombre de bons de commande enregistrés chaque jour pendant quinze jours de fonctionnement : 8 ; 10 ; 12 ; 10 ; 10 ; 12 ; 9 ; 11 ; 10 ; 9 ; 12 ; 11 ; 12 ; 6 ; 9.

a- Regrouper en classes cette série statistique, puis compléter le tableau d'effectifs et de fréquences correspondant.

b- Construire le diagramme à bâton correspondant.

c- Schématiser le camembert relatif.

d- Calculer les moyenne, variance, écart-type.

Exercice 2

On a relevé le nombre de commandes de copieurs TX 117 chaque jour pendant un mois dans une entreprise spécialisée dans la fourniture de matériels de bureautique : 130 ; 128 ; 128 ; 127 ; 128 ; 127 ; 128 ; 126 ; 127 ; 127 ; 128 ; 128 ; 130 ; 129 ; 128 ; 129 ; 127 ; 128 ; 130 ; 128.

a- Regrouper en classes cette série en remplissant un tableau récapitulatif.

b- Construire un diagramme en bâton correspondant.

c- Interpréter la valeur somme des effectifs pour deux bâtons consécutifs.

d- Calculer les moyenne, variance, écart-type.

Exercice 3

Dans une promotion d'étudiants, la liste des notes obtenues à un devoir de mathématiques par les élèves classés par ordre alphabétique est la suivante : 8 ; 6 ; 9 ; 18 ; 9 ; 11 ; 9 ; 13 ; 7 ; 13 ; 14 ; 7 ; 10 ; 10 ; 10 ; 7 ; 13 ; 14 ; 10 ; 13 ; 15 ; 5 ; 16 ; 13 ; 9 ; 10 ; 7 ; 12 ; 5 ; 12 ; 2 ; 9 ; 9 ; 8 ; 8.

a- Regrouper en classes cette série statistique, puis compléter le tableau d'effectifs et de fréquences correspondant.

b- Représenter graphiquement ces données.

c- Déterminer les valeurs approchées arrondies à 10^{-1} près de $\bar{x}$ et σ.

d- Quel est le pourcentage de notes appartenant à l'intervalle suivant

$$I = [\bar{x} - 2\sigma, \bar{x} + 2\sigma]$$

e- Même question que en d- en remplaçant 2 par 3 dans les bornes de I.

Exercice 4
Dans la production d'une entreprise on prélève 100 rouleaux de papier de tapisserie dont on mesure les longueurs en métres. On obtient les résultats consignés dans le tableau de fréquences ci-dessous.

Longueur (m)	[9,93 ;9,95[	[9,95 ;9,97[	[9,97 ;9,99[	[9,99 ;10,01[	[10,01 ;10,03[	[10,03 ;10,05[	[10,05 ;10,07[
Effectif	5	11	23	25	19	13	4

a- Construire l'histogramme de cette série.
b- Donner le nombre de rouleaux mesurés entre 9,98 m et 10,02 m.
c- Quelle loi théorique pourrait correspondre à la population totale.

Exercice 5
Dans la partie "enseignement général" d'un examen, un candidat a obtenu 11 en français coefficient 2, et 11,5 en anglais coefficient 2. Combien doit il obtenir à l'épreuve de mathématiques, coefficient 3, pour avoir plus de 10 de moyenne sur l'ensemble de ces trois épreuves.

Exercice 6
Montrer que $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$.

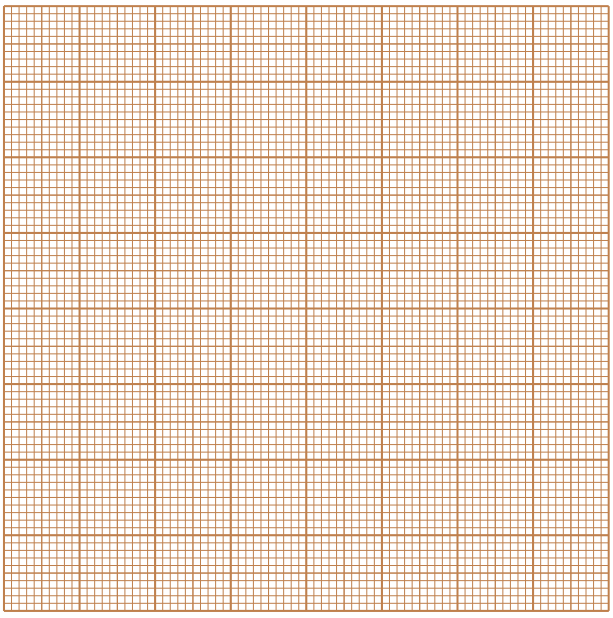
Exercice 7
a- Interpréter la quantité $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ lorsque $x_i \in \{0,1\}$.
b- Ecrire $v = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ en fonction de $\bar{x}$.
c- Donner un exemple de telles données, et la valeur $\bar{x}$ correspondante.

Feuille n°2 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 - Loi normale pour données continues ($x_i \in \mathbb{R}$)
a- Calculer la moyenne $\bar{x}_n$ et la variance corrigée $s_n^{2'}$ (approchées puisque calculées à l'aide des centres des classes) de la série univariée de l'exercice 4 de la feuille précédente n°1.

b- Superposer la courbe de densité de loi normale de paramètres $\mu = \bar{x}_n$ et $\sigma^2 = s_n^{2'}$ sur l'histogramme des fréquences tracé précédemment.



c- Commenter la superposition des deux graphiques obtenue. Interpréter l'aire sous la courbe de densité théorique pour un intervalle [a ;b] donné.

d- Générer sous R un échantillon gaussien de paramétres μ et σ^2 de même taille n=100. Tracer sous R l'histogramme des fréquences en superposant la

courbe de densité de la distribution théorique. Expliquer qualitativement les différences (cf. moyenne empirique, variance empirique, histogramme empirique) entre le nouvel échantillon et celui donné en a-.

Exercice 2 - Loi de Bernoulli pour données binaires ($x_i \in \{0, 1\}$)
On suppose avoir lancé n=30 fois une piéce équilibrée à deux faces 'P' et 'F', et obtenu la série de mesures suivantes (ou échantillon) :

$$FPFFPPFFFFPFFFPFPFFPPPFPPPPFFP$$

a- Les données sont elles équivalentes à un échantillon de valeurs binaires ? Si oui, réécrire les données et indiquer l'intérêt du nouveau codage choisi.

b- Calculer la moyenne $\bar{x}_n$ et la variance non corrigée s_n^2. Vérifier numériquement à l'aide de la calculatrice que $s_n^2 = \bar{x}(1 - \bar{x})$. Rappeller comment interpréter la valeur de $p_n = \bar{x}_n$ dans ce cas.

c- Tracer le diagramme à bâtons empirique correspondant à l'échantillon.

d- Soit X la variable aléatoire ayant générée l'échantillon étudié. Donner les valeurs de $p = P(X = 1)$ et $P(X = 0)$ après avoir défini X.

e- Tracer le diagramme à bâtons théorique correspondant à X.

f- Commenter qualitativement les deux graphiques. Que peut on dire du paramétre p de la loi de X et de la fréquence empirique p_n calculée sur l'échantillon.

g- Si on relance $n = 30$ fois la piéce, la série et sa moyenne empirique vont-elles être modifiées ? Le paramétre théorique p reste-il inchangé et pourquoi ?

Exercice 3 - Loi normale et probabilité d'un intervalle
a- Pour une v.a. X de loi normale centrée-réduite, expliquer graphiquement la valeur de $P(X \in [a; b])$ où $[a; b]$ est un intervalle de la droite réelle. Donner un exemple de tel intervalle et la valeur de la probabilité correspondante.

b- Soit une variable aléatoire X de loi normale $\mathcal{N}(\mu, \sigma^2)$, quelle est la loi de $Z = (X - \mu)/\sigma$? Comment déterminer la valeur de $\Phi(z) = P(Z \leq z)$ pour z réel, par exemple $z = 1.96$. Comment s'appelle la fonction Φ ?

c- Reprenant l'exercice 1, soit X la v.a correspondant à la distribution gaussienne $\mathcal{N}(\mu, \sigma^2)$. Représenter graphiquement puis donner la valeur de la probabilité théorique $P(X \in [9.98; 10.02])$. Comparer avec l'effectif empirique trouvé en 4-b de la feuille n°1.

Exercice 4 - Calculs sur les moments

Soit une série de n valeur continues $x_1, x_2, \cdots, x_n$ de moyenne $\bar{x}$ et variance $v_n(x)$. Soit 2 réels a et b tel que on ait pour tout i, $y_i = ax_i + b$, d'où la nouvelle série de moyenne $\bar{y}$ et variance $v_n(y)$.

a- Montrer que $\bar{y} = a\bar{x} + b$.

b- Montrer que $v_n(y) = a^2 v_n(x)$.

c- A-t-on des résultats comparables lorsqu'on considére des variables aléatoires.

Feuille n°3 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 - Probabilité par le formulaire de la loi N(0,1)
La variable aléatoire X suit une loi normale $\mathcal{N}(20, 5)$. Calculer et représenter graphiquement :
a- $P(X \leq 28)$
b- $P(X \geq 28)$
c- $P(X \geq 12)$
d- $P(X \leq 12)$
e- $P(12 \leq X \leq 28)$

Exercice 2 - Intervalle par le formulaire de la loi N(0,1)
La variable aléatoire X suit la loi normale $\mathcal{N}(20, 5)$. Déterminer à 10^{-2} près le nombre réel a tel que :
a- $P(X \leq a) = 0.99$
b- $P(X \leq a) = 0.01$
c- $P(X \geq a) = 0.05$
d- $P(X \geq a) = 0.90$
e- $P(20 - a \leq X \leq 20 + a) = 0.95$

Exercice 3 - Loi normale
Cet exercice reprend les données de l'exercice 4 de la feuille de n°1 de TD, ainsi que le graphique obtenu. On considére la v.a. X de loi normale $N(10, 0.3)$ et on suppose maintenant que l'échantillon a été généré d'après cette loi.

a- Calculer pour chacune des 7 classes la probabilité de l'intervalle correspondant pour la loi X. Représenter ces probabilités par un histogramme *théorique* superposé à l'histogramme *empirique* (tracé précédemment).

b- Si un nouvel échantillon est tiré d'après la loi N(10,0.3), indiquer comment le graphique est modifié, en considérant que l'on conserve les même classes.

Exercice 4 - Approximation de la loi binomiale
On jette dix fois de suite une pièce de monnaie bien équilibrée en notant chaque fois le résultat, ce qui constitue une partie.

a- On note X la variable aléatoire qui, à chaque partie, associe le nombre de 'Face' obtenu.

i) Justifier que la loi de probabilité suivie par la variable X est une loi binomiale ; on précisera les paramètres de cette loi.

ii) Calculer la probabilité de l'événement E : "Le nombre de 'Face' est compris entre 3 et 6 (bornes incluses)".

b- On décide d'approcher la loi de la variable aléatoire discrète X par la loi normale $\mathcal{N}(\mu, \sigma)$ de paramètres μ et σ.

i) Rappeler la formule de l'approximation de la loi binomiale par la loi normale.

ii) Expliquer pourquoi on prend $\mu = 5$ et $\sigma = \sqrt{2.5}$.

iii) On considère une variable aléatoire Y suivant la loi $\mathcal{N}(5, \sqrt{2.5})$. En utilisant cette approximation calculer la probabilité de l'événement : "Le nombre de 'Face' est compris entre 3 et 6 (bornes comprises)", c'est à dire, $P(2.5 \leq Y \leq 6.5)$ avec la correction de continuité.

iiii) Tracer sous R le diagramme à bâtons de la variable aléatoire modélisant une partie de n=10 lancers, et superposer la fonction de densité de la variable utilisée pour l'approximer. En augmentant n, le nombre de lancers de la pièce, qu'observe-t-on graphiquement ?

Feuille n°4 de TD du cours Estimation/Test

STID NIORT 2007/2008

Distribution d'échantillonnage de la moyenne

On considére une variable aléatoire réelle X de loi normale :

$$X \sim \mathcal{N}(5, 2^2)$$

1) Indiquer pour la v.a. X, les valeurs de l'espérance $\mu = E(X)$ et de l'écart-type $\sigma = \sqrt{V(X)}$, d'après les paramètres de sa loi.

2) Simuler sous R, un premier échantillon de 30 valeurs indépendamment générées d'après la distribution de X. Calculer sa moyenne $\bar{x}^1$.

3) Recommencer pour 9 autres échantillons, chacun ayant également 30 individus, dont les moyennes respectives sont notées $\bar{x}^2$, $\bar{x}^3$, $\cdots$, $\bar{x}^{10}$.

4) Calculer la moyenne $\bar{x}^\bullet$ et l'écart-type (corrigé) $s^{\bullet\prime}$ de la série statistique univariée :

$$\bar{x}^1, \bar{x}^2, \cdots, \bar{x}^{10}.$$

5) Comparer (numériquement, sans justifier) d'une part μ et $\bar{x}^\bullet$, puis d'autre part $s^{\bullet\prime}$ et $\frac{\sigma}{\sqrt{30}}$. Leurs valeurs numériques sont elles proches ?

6) Déterminer les bornes de l'intervalle :

$$I = \left[\mu - 1.96\frac{\sigma}{\sqrt{30}}; \mu + 1.96\frac{\sigma}{\sqrt{30}}\right]$$

Placer sur le même graphique, l'intervalle I, et les 10 moyennes de l'échantillon $\bar{x}^1, \bar{x}^2, \cdots, \bar{x}^{10}$ obtenues aux questions 2 et 3. Quel pourcentage de ces nombres est situé dans l'intervalle I ?

7) Pour chaque i-ème échantillon, déterminer les bornes de l'intervalle :

$$I_i = \left[\bar{x}^i - 1.96\frac{\sigma}{\sqrt{30}}; \bar{x}^i + 1.96\frac{\sigma}{\sqrt{30}}\right]$$

où i=1,2, ..., 10. Faire un tableau récapitulatif des intervalles ainsi établis.

Placer ces dix intervalles sur dix axes parallèles en alignant verticalement les abcisses communes (origines sur la même droite verticale et même unité). Quel pourcentage de ces intervalles contient l'espérance ?

8) Soit la v.a. normale

$$Y \sim \mathcal{N}\left(5, \sqrt{4/30}^2\right)$$

Calculer la probabilité

$$P(Y \in I)$$

et comparer (numériquement, sans justifier) au pourcentage obtenu à la question 6. Leurs valeurs numériques sont elles proches ?

Question bonus : Justifier la définition de Y à partir de la théorie de l'échantillonnage. Interpréter les deux pourcentages calculés en question 6) et 7).

<u>Évaluation :</u>

Un rapport récapitulatif est à rendre, en début de scéance du TD prochain, contenant :

- *les réponses rédigées manuscrites, de 1 à 8,*

- *Le code en langage R imprimé en annexe.*

Feuille n°9 de TD du cours Estimation/Test

STID NIORT 2007/2008

Ce TD porte sur le test d'une moyenne/proportion d'un caractère dans une population de laquelle on dispose un unique échantillon de n individus : $\underline{x}_n = (x_1, x_2, \cdots, x_n)$; $x_i \in \mathbb{R}$.

Exercice 1 - construction du test bilatéral sur la moyenne

Une grande surface vend du matériel photographique. On note X la variable aléatoire qui, à chaque ticket de caisse prélevé au hasard dans le stock de tickets d'un mois associe son montants en euros. On admet que X suit la loi normale de moyenne $\mu = 550$ et d'écart-type $\sigma = 195$.

A la suite d'une campagne promotionnelle d'un concurrent, le responsable de la grande surface redoute que le montant moyen des tickets (donc des ventes) soit modifié. Afin de contrôler que la moyenne μ des achats reste égale à 550 euros il se propose de construire un *test d'hypothèse bilatérale*.

On désigne par $\bar{X}$ la variable aléatoire qui, à chaque échantillon observé de valeurs de X, c'est à dire 50 tickets de caisse, associe la moyenne des montants en euros de ces tickets (le stock de tickets est assez important pour qu'on puisse assimiler des prélèvements à des tirages de 50 tickets avec remise, d'où l'hypothèse d'indépendance entre tirages).

• L'hypothèse nulle est H_0 : "$\mu = 550$".
• L'hypothèse alternative est H_1 : "$\mu \neq 550$".
• Le seuil de signification du test est fixé à 0,05.

a- Sous l'hypothèse H_0, quelle est la loi et les paramètres de $\bar{X}$?

b- Déterminer à 1 près, le nombre réel positif h tel que :

$$P(550 - h \leq \bar{X} \leq 550 + h) = 0.95$$

c- Déduire du b) la règle du test qui permet de décider si un échantillon provient bien d'une population de tickets de moyenne $\mu = 550$

d- La moyenne des montants d'un échantillons de 50 tickets prélevés après la campagne du concurrent est $\bar{x} = 597$. Utiliser le test du c) sur cet échantillon et conclure.

Exercice 2 - construction du test bilatéral sur une proportion

Les statistiques ont permis d'établir qu'en période de compétition la probabilité, pour un sportif pris au hasard, d'être déclaré positif au contrôle anti-dopage est égale à 0.02. On décide de construire un test qui, à la suite des contrôles sur un échantillon de 50 sportifs prélevés au hasard, permette de décider si au seuil de signification de 10%, le pourcentage de sportifs contrôlés positifs est p=0.02.

a- construction du test bilatéral.

Dans ce qui suit, tous les résultats approchés seront arrondis à 10^{-3}.

Soit F la variable aléatoire qui, à tout échantillon observé de valeurs de X (tirées avec remise) de 50 sportifs contrôlés, associe le pourcentage de sportifs contrôlés positivement. On suppose que F suit la loi normale :

$$\mathcal{N}\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big) \text{ où } n = 50$$

i) Enoncer une hypothèse nulle H_0 et une hypothèse alternative H_1 pour ce test bilatéral ;

ii) Déterminer, sous l'hypothèse H_0, le réel positif a tel que

$$P(p - a \leq F \leq p + a) = 0.9$$

iii) Enoncer la règle de décision du test.

b- Utilisation du test.

Dans un échantillon $\underline{x}_{50}$, deux contrôles anti-dopages ont été déclarés positifs. En appliquant la règle de décision du test à cet échantillon assimilé à un échantillon dont les individus sont choisis indépendamment et avec remise, peut-on conclure au risque 10 %, que l'échantillon observé est bien représentatif de l'ensemble de la population sportive ?

Feuille n°10 de TD du cours Estimation/Test

STID NIORT 2007/2008

Ce TD porte sur le test de conformité d'une moyenne ou d'une proportion d'un caractère dans une population de laquelle on dispose d'un unique échantillon de n individus :

$$\underline{x}_n = (x_1, x_2, \cdots, x_n)\,;\ \text{avec } x_i \in \mathbb{R} \text{ ou } x_i \in \{0, 1\}.$$

Exercice 1 - test bilatéral sur une moyenne
Dans une expérience sur l'acuité visuelle, un chercheur a demandé à 15 individus d'évaluer la distance d'un objet placé à 20 cm. Il a obtenu les résultats suivants en centimètres :

17	20	21	14	18	19	19	16	24	21	16	23	15	21	20

Il se demande s'il peut affirmer que les individus ont de la difficulté à évaluer correctement la distance ? On considère un risque $\alpha = 0.01$ et on suppose que l'évaluation de la distance par un individu suit une loi normale.
a- Poser le test d'hypothèse à employer en pareille situation.
b- Calculer la statistique de test et énoncer la décision résultante.

Exercice 2 - test bilatéral sur une proportion
Sur 1000 candidats au baccalauréat, 675 ont réussi. Testez au niveau de 10 % l'hypothèse selon laquelle la probabilité de réussite est 0.7.

Exercice 3 - test bilatéral sur une proportion
Dans plusieurs pays, les prévisions météorologiques sont données sous la forme de probabilités. La prévision "la probabilité de pluie pour demain est 0.4" a été faite 25 fois au cours de l'année et il a plu 13 fois. Tester l'exactitude de la prévision au niveau de 5%.

Exercice 4 - test bilatéral sur une moyenne
Un boulanger souhaite effectuer une vérification de la charte qualité de ses petits pains de lait en mesurant le poids moyen (en grammes) d'un échantillon de 7 pains produits dans la semaine ; il obtient le tableau suivant :

141	140	145	135	147	141	154

Le poids moyen standard est fixé à $\mu_0 = 146$ g. En supposant une loi normale pour le poids des petits pains, et en utilisant un risque $\alpha = 10\%$, peut on considérer que la qualité de fabrication est bonne ?

Exercice 5 - test bilatéral sur une moyenne
Des contenants d'un litre sont remplis de lait mécaniquement par une machine de telle sorte qu'ils contiennent normalement en moyenne 1,01 L avec un écart-type égal à 0,01 L. De plus, on a une loi normale pour le contenu.

a- Un échantillon de taille 25 permet d'obtenir un contenu moyen de 1,005 L. Doit on faire un ajustement de la machine si on se donne 5 % de risque de ne pas faire un ajustement requis ?

b- Quelle est la probabilité qu'un ajustement requis ne soit pas fait si le contenu réel moyen est de 1 L, et qu'on fait un ajustement si la moyenne échantillonnale de 25 observations est inférieure à 1,005 L ?

Exercice 6 - Risque de première espèce et seconde espèce
La quantité d'acide nitrique (en microgrammes) dans un mélange chimique doit être égale à 10. Cependant, des erreurs de manipulation font en sorte que cette quantité suit une loi normale de moyenne μ et de variance 0.09. On décide de tester $H_0 : \mu = 10$ contre $H_1 : \mu \neq 10$, à l'aide de résultats d'observation de 20 mélanges prélevés au hasard, et de rejeter H_0 si $\bar{x}_{20} < 9.80$ ou si $\bar{x}_{20} > 10.20$, où $\bar{x}_{20}$ est la quantité moyenne d'acide nitrique dans les 20 mélanges. Calculez, en posant $\bar{X}_{20}$ la v.a. de la moyenne de 20 mesures :

a- La probabilité d'erreur de première espèce, i.e. en reprenant la définition du cours, il s'agit de

$$\alpha = 1 - P(9.80 \leq \bar{X}_{20} \leq 10.20),$$

la probabilité de rejeter l'hypothèse H_0, alors que celle-ci est vraie, donc en supposant $\mu = \mu_0$ d'où $\bar{X}_{20} \sim \mathcal{N}\Big(\mu_0, \frac{\sigma}{\sqrt{20}}\Big)$, avec $\mu_0 = 20$.

b- La probabilité d'erreur de deuxième espèce dans le cas où $\mu = 9.90$.
En reprenant la définition du cours, il s'agit de

$$\beta = P(9.80 \leq \bar{X}_{20} \leq 10.20),$$

la probabilité d'accepter l'hypothèse H_0, alors que celle-ci est fausse, donc en supposant $\mu \neq \mu_0$, et ici, $\mu = 9.90$, d'où $\bar{X}_{20} \sim \mathcal{N}\Big(\mu, \frac{\sigma}{\sqrt{20}}\Big)$.

c- La valeur de la (fonction de) puissance lorsque $\mu = 9.90$, c'est à dire $1 - \beta$.

d- Représenter graphiquement les probabilités précédentes. Que faut-il remarquer ?

Exercice 7 - Test de conformité d'une moyenne sous SAS

Effectuer la résolution de l'exercice 1 à l'aide du logiciel SAS par les commandes suivantes :

```
DATA donnees;
   INPUT distance_evaluee @@;
   DATALINES;
  17 20 21 14 18 19 19 16 24 21 16 23 15 21 20
;
RUN;

PROC UNIVARIATE DATA=donnees CIBASIC (TYPE=TWOSIDED ALPHA=0.01) MU0=20;
VAR distance_evaluee;
OUTPUT OUT=stats MEAN=m N=nbobs;
RUN;
```

Il est important de remarquer que SAS ici traite un unique cas : le cas gaussien à variance inconnue. Et ne faisant pas d'approximation, la statistique utilisée suit un loi de Student pour toute taille d'échantillon. En remarque, pour rappel, pour un taille d'échantillon suffisamment grande, la loi de Student est proche d'une loi normale centrée réduite.

Vérifier que vous obtenez bien la sortie suivante :

Retrouver à la calculatrice l'ensemble des valeurs calculées par le logiciel pour l'intervalle de confiance, et le test de Student (nom du test de conformité mis en oeuvre).

Feuille n°11 de TD du cours Estimation/Test

STID NIORT 2007/2008

Ce TD porte sur le test de comparaison de deux moyennes ou proportions d'un caractère. On dispose de deux échantillons de n_1 et n_2 individus :

$$\underline{x}_{n_1} = (x_1, x_2, \cdots, x_{n_1}) \text{ et } \underline{y}_{n_2} = (y_1, y_2, \cdots, y_{n_2}).$$

Un des principaux objectifs d'un test de comparaison consiste à décider si ces deux échantillons proviennent d'une seule population ou bien de deux populations différentes.

Exercice 1 - construction d'un test de comparaison (proportions)

Une école de voile sur une île depuis deux ans connait une bonne activité en période estivale. Elle souhaite proposer une régate en bord de mer pour faire connaitre son activité à davantage d'habitants. A la question "Souhaitez vous assister à cet événement", sur les 400 habitants interrogés dans la région (N), 63 % ont répondu oui d'après une étude publiée dans le journal local. Lors d'une enquête similaire sur 500 habitants de la région voisine (S), 67 % réponses sont positives.

On note F_S la variable aléatoire qui, à tout échantillon de 400 habitants pris au hasard et avec remise dans la région S, associe la proportion de personnes de cet échantillon qui sont intéressées par un tel événement.
On note F_N la variable aléatoire qui, à tout échantillon de 500 personnes pris au hasard et avec remise dans la région N, associe la proportion de personnes de cet échantillon qui sont intéressées par un tel événement.
On suppose que la loi de la variable $D = F_S - F_N$ est approximativement une loi normale de moyenne $p_S - p_N$ inconnue et d'écart-type 0.032 (p_S et p_N étant les pourcentages de préférence dans la population S et N).

On se propose de construire un test bilatéral permettant de décider s'il y a une différence significative, au seuil 5 %, entre les pourcentages issus des deux échantillons de ces enquêtes.

a- Construction du test bilatéral.

i) L'hypothèse H_0 est donnée par $p_S = p_N$. Enoncer l'hypothèse alternative H_1.

ii) Déterminer l'intervalle $[-a; a]$, tel que, sous l'hypothèse H_0,

$$P(-a \leq D \leq a) = 0.95$$

b- Utiliser ce test avec les deux échantillons de l'énoncé et conclure.

c- Comment est calculée la variance de D ?

Exercice 2 - comparaison de deux proportions

On dispose de deux sirops pour la toux, notés A et B. On a obtenu en 2 jours d'utilisation, 49 guérisons sur 80 cas pour le sirop A et 75 guérisons sur 105 cas pour le sirop B. Y-a-t-il une différence entre les deux sirops ?

Exercice 3 - comparaison de deux moyennes

Un test est donné à 2 groupes de personnes, le premier de taille 75 et le second de taille 50. La moyenne des résultats dans le premier groupe est 80 avec un écart-type de 10, alors que la moyenne dans le second est 70 avec un écart-type de 12. Avec un niveau de 20 %, testez l'hypothèse que les personnes des 2 groupes sont des individus pris au hasard dans une même population.

Exercice 4 - comparaison de deux moyennes

Des biologistes veulent étudier l'effet de la domestication sur la croissance d'une variété de mollusques. Pour cela, un échantillon de $n_T = 32$ mollusques a été prélevé parmi des mollusques cultivés dans un bassin artificiel. La longueur des mollusques calculée à partir de cet échantillon a donné une moyenne égale à $\bar{x}_T = 3.0$ cm. Un autre échantillon de taille $n_R = 35$ prélevé parmi les mollusques vivant en milieu naturel a donné une moyenne égale à $\bar{x}_R = 3.5$ cm. Des études antérieures suggèrent que la longueur des mollusques domestiques devrait suivre une loi $N(\mu_T, 1)$, et, celle des mollusques marins une loi $N(\mu_R, 1)$. On se propose de tester l'hypothèse suivant laquelle l'élevage en milieu domestique n'a pas d'effet sur la longueur totale des mollusques. On utilisera une niveau de risque $\alpha = 0.05$.

a- Indiquer les deux hypothèses à tester.

b- Calculer la statistique de test et conclure.

Exercice 5 - comparaison de deux moyennes

On veut comparer les durées de vie de deux types de disques durs A et B. Pour cela, on prélève aléatoirement deux échantillons, l'un pour les disques durs de type A et l'autre pour ceux de type B. Les durée de vie, en centaines d'heures, sont consignées dans le tableau suivant ; on suppose qu'elles sont normalement distribuées.

type A	232, 228, 237, 225, 214, 213, 205, 233, 219, 236 ;
type B	222, 234, 244, 235, 229, 260, 232, 224.

Y-a-t-il une différence de durée de vie moyenne au seuil de 5 % ?

Exercice 6 - Test de comparaison de moyenne sous SAS

Répondre à l'exercice 5 à l'aide du logiciel SAS et la procédure TTEST.

Retrouver avec l'aide de la calculatrice l'ensemble des valeurs obtenues sous SAS pour les statistiques de test sur les moyennes ainsi que sur les variances.

Feuille n°12 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1
Les diamètres de 20 vis produites par une machine sont les suivantes :

1.05	1.04	1.06	1.02	1.03	1.04	1.07	1.09	1.02	1.03
1.05	1.03	1.09	1.07	1.03	1.05	1.07	1.04	1.02	1.01

a- Donner avec l'aide du logiciel SAS une valeur estimée :
- du diamètre moyen des vis produites par cette machine,
- de la variance du diamètre des vis produites par cette machine.

b- Donner avec l'aide du logiciel SAS une estimation de la moyenne par intervalle de confiance au niveau 5%.

c- Vérifie par le calcul les questions a et b.

Exercice 2
Un échantillon de 40 cigarettes d'une certaine marque a donné les teneurs en goudron (mg) suivantes :

12.9	11.9	12.4	12.8	14.5	13.1	12.9	14.5	14.7	12.3	13.4	14.7	13.4	13.9	12.9
14.5	16.5	12.7	14.8	11.8	14.3	14.4	13.5	11.9	12.8	13.5	15.6	12.8	11.8	14.7
14.4	15.0	15.2	11.8	12.9	13.6	14.6	12.9	11.8	14.2					

a- La norme en vigueur recommande une teneur en goudron d'au plus 13 mg par cigarette. Donner une valeur estimée avec l'aide du logiciel SAS :

- de la proportion des cigarettes de cette marque qui respectent la norme de la teneur en goudron ;

- de la teneur en goudron moyenne des cigarettes de cette marque ;

- de l'écart-type de la teneur en goudron des cigarettes de cette marque.

b- Donner avec l'aide du logiciel SAS une estimation par intervalle de confiance de la moyenne et de l'écart-type au niveau de confiance 99%.

Exercice 3
Voici les longueurs (en m) de baleines bleues :

22.74	17.25	16.95	18.97	18.00	19.40	23.40	18.45	17.60	21.30
19.80	21.65	24.50	22.10	22.05	21.80	18.45	25.83	17.50	20.15
21.50	17.80	21.70	24.40	21.70	19.64	19.45	23.50	20.30	18.60
18.70	20.00	19.20	22.45	24.80	19.60	24.90	17.53	19.00	25.20
18.60	19.20	18.20	18.00	21.60	17.90	18.10	19.00	20.35	19.90
21.50	18.60	16.90	18.95	21.90	18.55	24.40	22.55	20.30	18.90
19.80	25.45	18.05	18.40	18.80	23.80	17.85	20.50	19.10	18.50
19.85	22.90	21.40	24.55	24.40	22.60	24.50	18.70	19.10	19.00
25.50	19.35	18.95	17.75	22.70	18.80	17.80	17.50	20.55	21.70
24.90	26.30	18.35	24.45	19.30	18.07	22.70	18.30	19.40	23.30

Déterminer avec l'aide du logiciel SAS un intervalle de confiance au niveau 95% de leur longueur moyenne. Vérifier par le calcul l'intervalle obtenu.
Indiquer l'interprétation que l'on peut faire de cet intervalle de confiance pour la population totale des baleines (bleues).

Exercice 4
Dans une population le temps d'écoute de la télévision en une semaine (en heures) est une variable de la loi normale dont la moyenne μ fluctue d'une semaine à l'autre selon l'intérêt que suscitent les émissions. Un sondage auprès de 16 individus de cette population a donné les temps d'écoute suivant en une semaine :

12.0	23.5	15.0	11.5	23.0	29.5	34.0	19.5
20.0	27.0	13.0	22.5	06.0	25.0	18.0	20.5

Déterminer un intervalle de confiance à 90% pour μ cette semaine là.

Exercice 5
Dans lequel des cas suivants la variance de la moyenne (échantillonnale) $\bar{X}_n$ de n observations indépendantes d'une variable X d'écart-type σ est-elle la plus grande ?

a- $\sigma = 10$, $n = 100$.
b- $\sigma = 20$, $n = 200$.
c- $\sigma = 10$, $n = 200$.
d- $\sigma = 20$, $n = 100$.

Feuille n°13 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 : estimation/test pour une proportion avec <u>EXCEL</u>

On veut connaitre la probabilité p d'obtenir le résultat pile avec une pièce de monnaie. On procède donc à n lancers de la pièce. Le résultat de l'expérience correspond à : 0 1 1 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 1 0 0 1 1 1 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0.
Une valeur "1" indique que le lancer correspondant a donné pour résultat le coté pile.

1.1 Estimation

a) Écrire le programme SAS -en utilisant la PROC FREQ- pour calculer les estimations ponctuelle et par intervalle de confiance de p.

b)En déduire par lecture de la SORTIE SAS :
- n, la taille de l'échantillon,
- f, l'estimation ponctuelle de p,
- I l'estimation par intervalle de confiance au niveau 95% de p.

Effectuer les estimations précédentes à l'aide des formules du cours.

c) Utilisation de Excel pour effectuer ce calcul à partir de n, f, et α.

On renseigne tout d'abord :
- la valeur de n, soit 100 en cellule B6,
- la valeur de α, soit 0.05 en cellule C6
- la valeur de f, soit 0.36 en cellule D6.

Pour obtenir sous *Excel* les bornes de l'intervalles de confiance de la proportion inconnue, on utilise les formules suivantes :

```
D6-LOI.NORMALE.STANDARD.INVERSE(1-C6/2)*RACINE(D6*(1-D6)/B6)
D6+LOI.NORMALE.STANDARD.INVERSE(1-C6/2)*RACINE(D6*(1-D6)/B6)
```

1.2 Test de conformité bilatéral

Disposant des données précédentes sur la pièce, on veut décider si la pièce est équilibrée.

C'est à dire, on veut savoir si on peut accepter l'hypothèse que la proportion inconnue p d'obtenir pile est égale à $p_0 = \frac{1}{2}$.

H_0 : "$p = p_0$" contre H_1 : "$p \neq p_0$"

a) Ajouter dans PROC FREQ, la déclaration de l'option *p=0.5* dans les parenthèses après l'option BINOMIAL, en conservant un risque 5%.

L'option $p =$ indique la valeur du p_0 dans le test d'hypothèse, c'est à dire $p_0 = 0.5$ ici. Si on n'ajoute pas l'option, 0.5 est la valeur par défaut.

b) Quelle décision prend on d'après la sortie SAS obtenue ?

On accepte $H_0 : p = p_0$ ssi *P-value="Pr > |Z| bilatéral"* $\geq \alpha$, sinon, on accepte $H_1 : p \neq p_0$ au risque α.

c) Comparer le résultat avec celui de la règle de décision du cours.

d) Ajouter à la feuille Excel précédente le calcul de la statistique de test, pour un p_0 candidat, ainsi que l'intervalle qui permet de décider entre H_0 et H_1. Éventuellement, calculer la P-value.

Placer la valeur de p_0, en B20. Alors, pour calculer la statistique de test (c'est à dire la valeur de t), on place dans une nouvelle cellule, la formule :

```
(D6-B20)/RACINE(B20*(1-B20)/B6)
```

1.3 Utilisation de la feuille Excel obtenue

a) Utiliser la feuille Excel précédente pour le calcul de la statistique de test avec la nouvelle valeur candidate $p_0 = 0.4$ pour la proportion p du nombre de lancer coté pile.

b) Recommencer l'estimation par intervalle de confiance et le test pour un risque 1% et 10%.

c) Représenter graphiquement les intervalles de confiance. Commenter.
Représenter graphiquement les intervalles du test. Commenter.

Exercice 2 : estimation/test pour une moyenne avec <u>EXCEL</u>

Nous nous plaçons dans le cadre le plus fréquent en pratique, la moyenne et l'écart-type de la population sont inconnus.

Reprendre les données du TD 12, exercice 2. Un échantillon de 40 cigarettes d'une certaine marque a donné les teneurs en goudron (mg) suivantes :

12.9	11.9	12.4	12.8	14.5	13.1	12.9	14.5	14.7	12.3	13.4	14.7	13.4	13.9	12.9
14.5	16.5	12.7	14.8	11.8	14.3	14.4	13.5	11.9	12.8	13.5	15.6	12.8	11.8	14.7
14.4	15.0	15.2	11.8	12.9	13.6	14.6	12.9	11.8	14.2					

a- Écrire un programme Excel pour :
→ le calcul de l'intervalle de confiance moyenne μ,
→ le calcul de l'intervalle de confiance de l'écart-type σ,
en considérant un niveau de confiance $1-\alpha$ fixé.

On a noté :
→ μ, la teneur moyenne inconnue de goudron sur l'ensemble de la population des cigarettes produites.
→ σ, l'écart-type inconnu de la teneur en goudron sur l'ensemble de la population des cigarettes produites.

Pour la moyenne, proposer le calcul exact et celui approché pour n grand, à l'aide des fonctions suivantes[1] :

```
LOI.NORMALE.STANDARD.INVERSE(1-C6/2)
LOI.STUDENT.INVERSE(C6;B6-1)
KHIDEUX.INVERSE(1-C6/2;B6-1)
```

qui correspondent respectivement à $t_{1-\alpha/2}$, $t_{1-\alpha/2;n-1}$ et $\chi^2_{\alpha/2;n-1}$.

b- Vérifier que les résultats correspondent à ceux de SAS
(cf. PROC UNIVARIATE ou éventuellement PROC TTEST sans 'CLASS').

c- Compléter la feuille de calcul Excel pour effectuer un test de conformité (bilatéral) pour la moyenne. On testera les hypothèses suivantes

H_0 : "$\mu = 13$" contre H_1 : "$\mu \neq 13$"

[1] (cf. Exercice 1 pour le contenu des cellules C6 et B6)

ANNEXE

Estimation de la proportion de lancers donnant pile.

On note $x_i \in \{0,1\}$, le résultat du i-ème lancer pour $i = 1, \cdots, n$.
L'estimation ponctuelle de la proportion inconnue p s'écrit pour $n = 100$

$$f = \sum_{i=1}^{i=n} x_i = \frac{0+1+1+1+1+\cdots+0+0+1+0++1+0}{100} = \frac{36}{100} = 0.36$$

puis avec $1-\alpha = 95\%$ donc $\alpha = 0.05$.
L'estimation par intervalle de confiance de la proportion inconnue p s'écrit :

$$\begin{aligned} I &= \left[f - t_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f \times (1-f)}{n}} \; ; \; f + t_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f \times (1-f)}{n}} \right] \\ &= \left[0.36 - 1.96\sqrt{\frac{0.36 \times (1-0.36)}{100}} \; ; \; 0.36 + 1.96\sqrt{\frac{0.36 \times (1-0.36)}{100}} \right] \\ &= [0.26592 \; ; \; 0.45408]. \end{aligned}$$

Test de conformité de la proportion de lancers donnant pile.

On note $p_0 = 0.5$, la probabilité candidate pour le test. La statistique de test s'écrit :

$$\begin{aligned} t &= \frac{f - p_0}{\sqrt{\frac{p_0 \times (1-p_0)}{n}}} \\ &= \frac{0.36 - 0.5}{\sqrt{\frac{0.5 \times (1-0.5)}{100}}} \\ &\approx -2.8 \end{aligned}$$

Pour un risque $\alpha = 0.05$, on obtient $t_{1-\frac{\alpha}{2}} = t_{0.975} = 1.96$ d'après lecture inverse de la table N(0,1). La règle de décision consiste à vérifier si t est dans $[-t_{1-\frac{\alpha}{2}}; t_{1-\frac{\alpha}{2}}]$, or

$$-2.8 \notin [-1.96; 1.96],$$

donc on rejette H_0 au risque 5%.

Feuille n°14 de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 : test de comparaison et logiciel SAS

Les données étudiées correspondent à des mesures sur des crabes australiens de deux espèces (bleu et orange). Une entreprise alimentaire se demandent si les crabes ont un poids différents suivant leur sexe ou leur espèce ?

Sur un échantillon de n=200 crabes, les mesures effectuées sur le poids, sur chacun des crabes, en tenant compte de son espèce et de son poids, donnent les résultats indiqués dans les tableaux ci dessous. Ici l'unité de mesure est le gramme, et les écart-types sont corrigés.

Espèce	**Orange**	**Bleue**
Effectif	100	100
Moyenne	109,7202	135,1975
Écart-type	48,17162	50,93990

Espèce	**Mâle**	**Femelle**
Effectif	100	100
Moyenne	127,6742	117,2435
Écart-type	54,86659	46,66808

On fait l'hypothèse que le poids d'un crabe est l'observation d'une variable aléatoire de loi normale, de paramètres inconnus.

a- Quel est le moyenne du poids sur l'ensemble des 200 crabes ?

b- Quel est l'écart-type du poids sur l'ensemble des 200 crables ?

c- Peut on accepter l'hypothèse d'égalité des variances ?

d- Peut on accepter l'hypothèse d'égalité des moyennes ?

e- Vérifier vos résultats à l'aide de SAS en interprétant les *P-value* correspondants.

Exercice 2 : test de conformité et logiciel SAS

a- Tester pour chaque sous échantillon (mâle ou espèce) l'hypothèse que le poids moyen de la sous population correspondante est égal au poids moyen total de l'échantillon arrondi au gramme.

b- Comparer vos résultats à ceux obtenus par la feuille *Excel* du TD précédent.

c- Vérifier vos résultats à l'aide de SAS en interprétant les *P-value* correspondants.

Exercice 3 : intervalle de confiance et logiciel SAS

a- Encadrer pour chaque sous échantillon (mâle ou espèce) le poids moyen de la sous population correspondante.

b- Comparer vos résultats à ceux obtenus par la feuille *Excel* du TD précédent.

c- Vérifier vos résultats à l'aide de SAS.

Exercice 4 : calcul de probabilité

Quelle est la probabilité qu'un crabe ait un poids inférieur à 100 grammes ?

Solutions de le feuille n°10
de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 - test bilatéral sur une moyenne

a- H_0 : "$\mu = 20$", H_1 : "$\mu \neq 20$"

b- $t = \frac{\bar{x}_n - \mu_0}{s'_n/\sqrt{n}} = \frac{18.93-20}{2.91/\sqrt{15}} \approx -1.42$ et $t_{n-1;1-\frac{\alpha}{2}} = t_{14;0.995} \approx 2.977$

$-1.42 \in [-2.977; +2.977]$ donc on accepte H_0 au risque 1%.

Exercice 2 - test bilatéral sur une proportion

H_0 : "$p = 0.7$", H_1 : "$p \neq 0.7$"

$t = \frac{f_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.675-0.7}{\sqrt{\frac{0.7(1-0.7)}{1000}}} \approx -1.72$ et $t_{1-\frac{\alpha}{2}} = t_{0.95} \approx 1.645$

$-1.72 \notin [-1.645; +1.645]$ donc on rejette H_0 au risque 10%.

Exercice 3 - test bilatéral sur une proportion

H_0 : "$p = 0.4$", H_1 : "$p \neq 0.4$"

$t = \frac{f_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{13/25-0.4}{\sqrt{\frac{0.4(1-0.4)}{25}}} \approx 1.22$ et $t_{1-\frac{\alpha}{2}} = t_{0.975} \approx 1.96$

$1.22 \notin [-1.96; +1.96]$ donc on accepte H_0 au risque 5%.

Les prévisions météo sont fidèles à la vraie météo.

Exercice 4 - test bilatéral sur une moyenne

H_0 : "$\mu = 146$", H_1 : "$\mu \neq 146$"

$t = \frac{\bar{x}_n - \mu_0}{s'_n/\sqrt{n}} = \frac{143.29-146}{6.07/\sqrt{7}} \approx -1.18$ et $t_{n-1;1-\frac{\alpha}{2}} = t_{6;0.95} \approx 1.943$

$-1.18 \in [-1.943; +1.943]$ donc on accepte H_0 au risque 10%.

Exercice 5 - test bilatéral sur une moyenne

a- test classique.

H_0 : "$\mu = 1.01$", H_1 : "$\mu \neq 1.01$" (et $\sigma = 0.01$ connu)

$t = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{1.005-1.01}{0.01/\sqrt{25}} \approx -2.5$ et $t_{1-\frac{\alpha}{2}} = t_{0.975} \approx 1.96$

$-2.5 \notin [-1.96; +1.96]$ donc on rejette H_0 au risque 5%.

On doit donc effectuer un ajustement de la machine au risque 5%.

b- calcul d'une probabilité.
On pose X, la variable aléatoire qui associé à un contenant son contenu. D'aprés l'énoncé, X est de loi normale $N(1.0, 0.01)$. On doit alors calculer :

Probabilité("moyenne empirique de 25 observations est supérieure à 1.005")

donc on utilise la variable aléatoire $\bar{X}_n$ qui associe à un échantillon de taille 25 contenants, leur contenu moyen. D'après le cours, on pose directement $\bar{X}_n \sim N(1.0, \frac{0.01}{\sqrt{25}})$, puis on traduit la phrase de l'énoncé sous la forme :

$$P(\bar{X}_n > 1.005) = P(Z > 2.5) = 1 - P(Z < 2.5) = 0.0062.$$

Exercice 6 - Risque de première espèce et seconde espèce

a- On a :

$$\begin{aligned} \alpha &= \text{Probabilité(Rejeter } H_0 \text{ alors que } H_0 \text{ est vrai)} \\ &= \text{Probabilité(Décider } \mu \neq 10 \text{ alors que } \mu = 10) \\ &= 1 - P(9.80 \leq \bar{X}_{20} \leq 10.20) \text{ avec } \bar{X}_{20} \sim \mathcal{N}\Big(10, \frac{0.01}{\sqrt{20}}\Big) \\ &\approx 0.0028 \end{aligned}$$

b- On a :

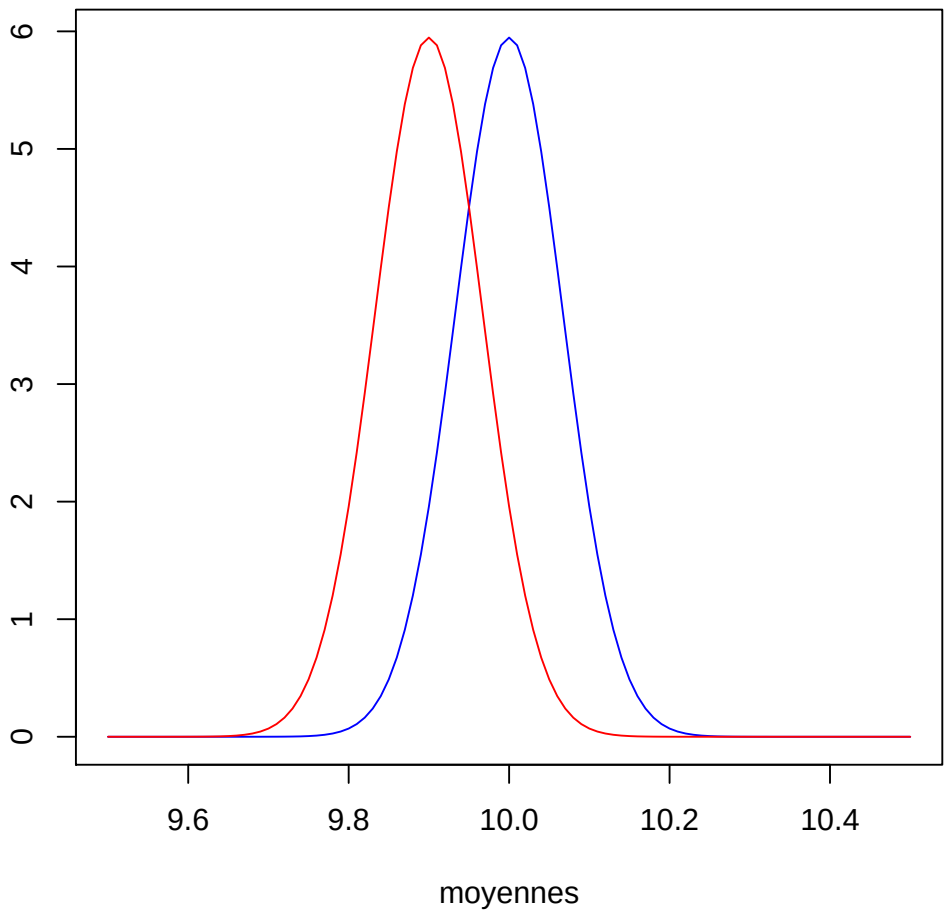
$$\begin{aligned} \beta_{9.9} &= \text{Probabilité(Accepter } H_0 \text{ alors que } H_1 \text{ est vrai)} \\ &= \text{Probabilité(Décider } \mu = 10 \text{ alors que } \mu = 9.9) \\ &= P(9.80 \leq \bar{X}_{20} \leq 10.20) \text{ avec } \bar{X}_{20} \sim \mathcal{N}\Big(9.9, \frac{0.01}{\sqrt{20}}\Big) \\ &\approx 0.9319 \qquad\qquad \text{dans le cas particulier } \mu = 9.9 \end{aligned}$$

b- On a :

$$1 - \beta = 0.0681$$

Il faut remarque que β est calculé pour une vraie moyenne μ donnée. Donc si la vraie moyenne μ est différente β est différente. Ici, la valeur est grande car la valeur proposée μ_0 choisie pour valeur de μ peut très bien être remplacée par une autre valeur candidate !

Remarque, on peut représenter graphiquement la loi de $\bar{X}_n$ dans le cas $\mu = 10$ et $\mu = 9.90$.



Exercice 7 - Test de conformité d'une moyenne sous SAS

On calcule pour la moyenne :

- $\bar{x}_n = 18.93$ et $s'_n = 2.91$
- $\bar{x}_n - t_{n-1;1-\frac{\alpha}{2}} \frac{s'_n}{\sqrt{n}} = 18.93 - \underbrace{2.977}_{t_{14;0.995}} \times \frac{2.91}{\sqrt{15}} = 16.69$
- $\bar{x}_n + t_{n-1;1-\frac{\alpha}{2}} \frac{s'_n}{\sqrt{n}} = 18.93 + 2.977 \times \frac{2.91}{\sqrt{15}} = 21.17$

Ensuite, pour la variance,

- $\frac{(n-1){s_n}'^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}} == \frac{14\times 2.91^2}{\chi^2_{14;0.995}} = \frac{14\times 2.91^2}{31.32} \approx 3.79$
- $\frac{(n-1){s_n}'^2}{\chi^2_{n-1;\frac{\alpha}{2}}} = \frac{14\times 2.91^2}{\chi^2_{14;0.005}} = \frac{14\times 2.91^2}{4.075} \approx 29.10$

Les quantiles de la loi du χ^2 se lisent soit sur une table de loi, soit avec le logiciel R, par le calcul $qchisq(0.005, 14)$ et $qchisq(0.995, 14)$.

Ensuite, pour l'écart-type,

- $2.91^2 = 8.47$
- $\sqrt{3.79} \approx 1.95$
- $\sqrt{29.10} \approx 5.39$

Et le test de Student,

- $t = \frac{\bar{x}_n - \mu_0}{s'_n/\sqrt{n}} = \frac{18.93-20}{2.91/\sqrt{15}} \approx -1.42$
- Pr $>$ |t| sous SAS, est égal à p-value$= 2 \times P(T > |t|)$ où $T \sim \mathcal{T}_{n-1}$, une loi de Student à n-1=14 ddl (degrés de liberté).

On lit sur un formulaire que pour $t = 1.42$, on a $P(-1.42 < T < 1.42) \approx 0.20$, ce qui est cohérent avec la valeur lue sous SAS. Plus exactement sous R, on calcule $(1 - pt(1.42, 14)) * 2$ qui donne le résultat 0.177.

Solutions de la feuille n°11
de TD du cours Estimation/Test

STID NIORT 2007/2008

Exercice 1 - construction d'un test de comparaison (proportions)

a- Construction du test bilatéral.

Attention : on prend $n_S = 500$, $f_S = 67\%$, $n_N = 400$ et $f_N = 63\%$ (cf. séance).

i) H_0 : "$p_S = p_N$", H_1 : "$p_S \neq p_N$"

ii) sous H_0, $p_S - p_N = 0$ donc $D \sim \mathcal{N}(0, 0.032)$,
puis en se ramenant à une loi N(0,1) par centrage/réduction,

$$\begin{aligned} & P(-a < D < a) = 0.95 \\ \Rightarrow \quad & P(\tfrac{-a}{0.032} < \tfrac{D}{0.032} < \tfrac{a}{0.032}) = 0.95 \\ \Rightarrow \quad & 2 \times \Phi(\tfrac{a}{0.032}) - 1 = 0.95 \\ \Rightarrow \quad & \Phi(\tfrac{a}{0.032}) = 0.975 \\ \Rightarrow \quad & a = 1.96 * 0.032 \\ \Rightarrow \quad & a \approx 0.063 \end{aligned}$$

b- La règle du test est donc
" si $d = f_S - f_N \in [-6.3\%; 6.3\%]$ alors accepter H_0 sinon H_1 au risque 5%".

en effet, on vient de montrer par le calcul de probabilité, que si H_0 est vraie, alors il y a 95% de chance que la différence d se trouve dans l'intervalle $[-a; a] = [-6.3\%; 6.3\%]$!

Ici, $d = 67\% - 63\% = +4\%$ donc on accepte H_0, c'est à dire, les proportions sur les deux populations ne sont pas significativement différentes.

c- On calcule l'écart-type vu dans le formulaire du cours

$$f_{SN} = \frac{n_S f_S + n_N f_N}{n_S + n_T} = \frac{500 \times 67\% + 400 \times 63\%}{500 + 400} \approx 65\%$$

Et, ensuite,

$$s_D = \sqrt{f_{SN}(1-f_{SN})(\frac{1}{n_S}+\frac{1}{n_N})} = \sqrt{0.65 \times 0.35 \times (\frac{1}{400}+\frac{1}{500})} \approx 0.032$$

Donc l'énoncé utilise bien la même formulation du test que le cours !

Exercice 2 - comparaison de deux proportions

H_0 : "$p_A = p_B$", H_1 : "$p_A \neq p_B$"

$$f_{AB} = \frac{n_A f_A + n_B f_B}{n_A + n_B} = \frac{49+75}{80+105} \approx 67\%$$

$$t = \frac{f_A - f_B}{\sqrt{f_{AB}(1-f_{AB})(\frac{1}{n_A}+\frac{1}{n_B})}} = \frac{\frac{49}{80}-\frac{75}{105}}{\sqrt{0.67(1-0.67)(\frac{1}{80}+\frac{1}{105})}} \approx -1.46$$

or $t_{1-\frac{\alpha}{2}} = t_{0.975} \approx 1.96$ (on choisit α =5%)

donc $-1.46 \in [-1.96; 1.96]$ et on accepte H_0 au risque 5%.

Exercice 3 - comparaison de deux moyennes

H_0 : "$\mu_1 = \mu_2$", H_1 : "$\mu_1 \neq \mu_2$"

$n_1 = 75, n_2 = 50$ grands, et σ_1, σ_2 inconnus.
$\bar{x}_{n_1} = 80$, $\bar{y}_{n_2} = 70$, $s_{n_1} = 10$ $s_{n_2} = 12$ (ecart-types non corrigés !)

$$t = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\frac{s'^2_{n_1}}{n_1}+\frac{s'^2_{n_2}}{n_2}}} = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\frac{s^2_{n_1}}{n_1-1}+\frac{s^2_{n_2}}{n_2-1}}} = \frac{80-70}{\sqrt{\frac{10^2}{74}+\frac{12^2}{49}}} \approx 4.83$$

car la variance corrigée s'écrit $s'^2_n = \frac{n}{n-1}s^2_n$

or $t_{1-\frac{\alpha}{2}} = t_{0.90} \approx 1.28$

donc $4.83 \notin [-1.28; 1.28]$ et on rejette H_0 au risque 20%.

Exercice 4 - comparaison de deux moyennes

a- H_0 : "$\mu_T = \mu_R$", H_1 : "$\mu_T \neq \mu_R$"

b- nous sommes dans le cas (rare) des écart-types connus

$n_T = 32, n_R = 35$ (grands), et $\sigma_T = 1, \sigma_R = 1$ connus.
$\bar{x}_{n_T} = 3.0$, $\bar{y}_{n_R} = 3.5$.

$$t = \frac{\bar{x}_{n_T} - \bar{y}_{n_R}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} = \frac{3.0 - 3.5}{\sqrt{\frac{1^2}{32} + \frac{1^2}{35}}} \approx -2.04$$

or $t_{1-\frac{\alpha}{2}} = t_{0.975} \approx 1.96$

donc $-2.04 \notin [-1.96; 1.96]$ et on rejette H_0 au risque 5%.

on conclut que l'élevage en milieu domestique a un effet sur la croissance du mollusque de cette variété.

Exercice 5 - comparaison de deux moyennes

$n_1 = 10, n_2 = 8$ petits, et σ_1, σ_2 inconnus.
$\bar{x}_{n_1} = 224.2$, $\bar{y}_{n_2} = 235.0$, $s'_{n_1} = 10.96$ $s'_{n_2} = 12.20$

i) on compare les deux vrais écart-types avant de pouvoir comparer les moyennes ! en effet, à la lecture de l'énoncé, on ne sait pas si les écart-types sont égaux, donc il faut effectuer un test pour trancher !

<u>test d'hypothèses</u> H_0 : "$\sigma_1 = \sigma_2$", H_1 : "$\sigma_1 \neq \sigma_2$"

pour ce faire, on calcule, $max(\frac{s'^2_{n_1}}{s'^2_{n_2}}, \frac{s'^2_{n_2}}{s'^2_{n_1}}) = \frac{s'^2_{n_2}}{s'^2_{n_1}} = \frac{12.20^2}{10.96^2} \approx 1.24$

or, $f_{n_2-1;n_1-1;1-\frac{\alpha}{2}} = f_{7;9;0.975} \approx 4.2$

d'où, 1.24<4.2 et on accepte l'hypothèse d'égalité des variances, et dans la suite, on considère <u>$\sigma_1 = \sigma_2$</u>.

i) on compare les deux moyennes après avoir décidé si les deux vrais écart-types étaient égaux ou non !

<u>test d'hypothèses</u> H_0 : "$\mu_1 = \mu_2$", H_1 : "$\mu_1 \neq \mu_2$"

comme $\sigma_1 = \sigma_2$, X et Y de loi normale, et n_1, n_2 petits, alors on utilise la statistique suivante.

$$t = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\frac{(n_1-1)\times s'^2_{n_1} + (n_2-1)\times s'^2_{n_2}}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{224.2-235.0}{\sqrt{\frac{(10-1)\times 10.96^2 + (8-1)\times 12.20^2}{10+8-2}\left(\frac{1}{10}+\frac{1}{8}\right)}} \approx -1.98$$

or $t_{n_1+n_2-2;1-\frac{\alpha}{2}} = t_{16;0.975} \approx 2.12$

donc $-1.98 \in [-2.12; 2.12]$ et on peut accepter H_0 au risque 5%, et conclure à l'égalité des vraies moyennes, soit $\underline{\mu_1 = \mu_2}$.

Exercice 6 - Test de comparaison de moyenne sous SAS

On écrit sous SAS le programme suivant.

```
/* données : durée de vie pour deux types de disques durs (A et B) */
DATA disques;
   INPUT duree type @@;
   DATALINES;
    232 1 228 1 237 1 225 1 214 1 213 1 205 1 233 1 219 1 236 1
    222 2 234 2 244 2 235 2 229 2 260 2 232 2 224 2
;
RUN;

PROC MEANS DATA=disques;
  VAR duree;
  BY type;
RUN;

PROC TTEST DATA=disques;
CLASS type;
VAR duree;
TITLE 'Test d''egalite de moyenne pour deux types de disques durs';
RUN;
```

Il faut remarquer que SAS ne traite que le cas gaussien, et n'utilise pas de formule approchée pour n grand, donc les deux formules du polycopié du cas n_1, n_2 petits s'appliquent ici quelque soit la valeur de n_1 et n_2.
Évidement, plus les tailles des échantillons deviennent importantes et plus les deux statistiques de test de la moyenne (cf Méthode *Pooled* et *Satterthwaite*) s'approchent du cas limite donné en cours $t = (\bar{x}_{n_1} - \bar{y}_{n_2})/\sqrt{\frac{s'^2_{n_1}}{n_1} + \frac{s'^2_{n_2}}{n_2}}$.

Pour l'interprétation, il faut procéder ainsi (par défaut $\alpha = 5\%$) :

1) On teste l'égalité des variances.
$\Rightarrow$ "Pr $>$ F"=0.7473 $\geq \underbrace{0.05}_{\alpha}$ donc on accepte l'égalité des variances.

2) On teste l'égalité des moyennes (cas "*Pooled - Variances Equal*").
$\Rightarrow$ "Pr $>$ |t|"=0.0656 $\geq \underbrace{0.05}_{\alpha}$ donc on accepte l'égalité des moyennes.

En conclusion, les deux disques durs de type A et B ont des durées de fonctionnement identiques.

STID S3 14/11/2007

Exercice.

On s'intéresse au test de conformité bilatéral de la moyenne μ pour les données suivantes (n=7) :

141 140 145 135 147 141 154

La moyenne candidate μ_0 est égale à 146.

Il s'agit ici de reprendre l'exercice 7 de la feuille de TD n°10 d'estimation et tests en traitant les données de l'exercice 4, afin de retrouver les résultats vus en séance.

1) Indiquer le code SAS utilisé pour calculer les intervalles de confiance et le test de conformité ;

2) Communiquer la sortie SAS obtenue pour les intervalles de confiance de la moyenne, variance et écart-type, ainsi que le test de t de Student ;

3) Retrouver -cf. correction de l'exercice 7- par l'application directe des formules du cours l'ensemble des résultats chiffrés de la sortie SAS. Il doit être fait uniquement usage de la calculatrice, les tables de loi, et le logiciel R ;

4) Rappeler le test d'hypothèse mis en œuvre dans la procédure SAS. Indiquer si l'hypothèse H_0 peut être acceptée au risque 10% en interprétant la *P-value* obtenue.

Cet exercice est à rendre en début de séance au prochain TD (20/11/2007).
Pour rappel, un devoir sur les tests de conformité est prévu ce même jour.

Solution au Contrôle d'Estimation et Tests Paramétriques

STID S3 Niort - 21 décembre 2007

EXERCICE 1

Partie A (4 pts)

a- La population tout comme l'échantillon a pour individus des mots.
L'échantillon est constitué des 200 mots sur chacune des 10 pages, soit un total de n=2000 mots. Il observé sur chaque mot, l'évènement : "bien retranscrit" ou "mal retranscrit",

soit pour i=1,...,2000, on pose $x_i = \begin{cases} 0 \text{ si le ième mot est mal retranscrit} \\ 1 \text{ si le ième mot est bien retranscrit} \end{cases}$

La taille de la population est inconnue, et on suppose qu'elle est importante.

On pose : p = "proportion de mots mal retranscrits par le logiciel d'ocr".

b- On calcule une estimation du pourcentage d'erreur inconnu p, par l'estimation ponctuelle

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{29}{2000} = 1.45\%$$

puis étant donné que n=2000 est grand l'estimation par intervalle de confiance s'écrit :

$$\begin{aligned} I &= \left[\hat{p} - t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \\ &= \left[\frac{29}{2000} - 1.96\sqrt{\frac{\frac{29}{2000}(1-\frac{29}{2000})}{2000}}; \frac{29}{2000} + 1.96\sqrt{\frac{\frac{29}{2000}(1-\frac{29}{2000})}{2000}}\right] \\ &\approx [0.93\%; 1.97\%] \end{aligned}$$

au niveau de $1 - \alpha = 95\%$.

c- On veut résoudre le test de conformité suivant :

$$\begin{aligned} H_0 &: p = 1\% \\ H_1 &: p \neq 1\% \end{aligned}$$

Soit, avec $p_0 = 0.01$, on calcule la statistique :

$$t = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{29}{2000} - \frac{1}{100}}{\sqrt{\frac{\frac{1}{100}(1-\frac{1}{100})}{2000}}} \approx 2.02$$

Donc, $2.02 \notin [-1.96; +1.96]$ soit $t \notin [-t_{1-\frac{\alpha}{2}}; t_{1-\frac{\alpha}{2}}]$,
on rejette au risque 5% l'hypothèse H_0.

Ainsi, étant données les informations disponibles, on conclut que le logiciel ZOCR ne vérifie pas l'hypothèse d'un taux d'erreur égal à 1% (au risque 5%).

Attention : 2.02 est 'très proche' de 1.96, donc modérer la décision (cf. P-value) !

Partie B (3 pts)

a- On pose :
p_1 = "proportion de mots mal retranscrits par le logiciel ZOCR".
p_2 = "proportion de mots mal retranscrits par le logiciel LOCR".

D'où le test de comparaison permettant de comparer les performances des logiciels :

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 \neq p_2 \end{aligned}$$

b- On calcule, au risque 5% :

$$t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{12}(1-\hat{p}_{12})(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{\frac{29}{2000} - \frac{26}{2000}}{\sqrt{\frac{55}{4000}(1-\frac{55}{4000})(\frac{1}{2000}+\frac{1}{2000})}} \approx 0.40$$

avec $n_1 = 2000$, $n_2 = 2000$, et

$$\hat{p}_{12} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{2000 \times \frac{29}{2000} + 2000 \times \frac{26}{2000}}{2000+2000} = \frac{55}{4000} = 13.75\%$$

Donc $0.40 \in [-1.96; +1.96]$ soit $t \in [-t - 1_{\frac{\alpha}{2}}; t_{1-\frac{\alpha}{2}}]$,
on accepte au risque 5% l'hypothèse H_0.

c- En conclusion, on peut choisir le logiciel LOCR dont la proportion d'erreur sur l'échantillon est plus faible, même si la différence est non significative. Un test unilatéral de comparaison de deux proportions serait intéressant à mettre en oeuvre ici.

EXERCICE 2

Partie A (6 pts)

a- La population est l'ensemble des grains sur la plage, et sa taille est donc pratiquement incommensurable, c'est à dire d'une taille très très grande. L'échantillon résulte d'un tirage au hasard de n=15 grains, parmi tous les grains de la plage. Après mesure, il en résulte les tailles x_i en mm des grains pour i allant de 1 à 15.
Le caractère mesuré sur chaque individu de la population et de l'échantillon est donc la taille de l'individu ou grain, c'est à dire son diamètre.

b- On évalue la moyenne, la variance non corrigée et celle corrigée :

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n} = 4.647$$

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n} = 0.324$$

$$s_n'^{\,2} = \frac{n}{n-1} s_n^2 = 0.347$$

à la calculatrice. En remarque, la valeur de s_n' apparait dans les sorties SAS !

c- On calcule, avec $t_{1-\frac{\alpha}{2};n-1} = t_{0.995;14} = 2.977$, l'estimation par intervalle de confiance de la moyenne au niveau 1% :

$$\begin{aligned} I &= \left[\bar{x}_n - t_{1-\frac{\alpha}{2};n-1}\frac{s_n'}{\sqrt{n}}; \bar{x}_n + t_{1-\frac{\alpha}{2};n-1}\frac{s_n'}{\sqrt{n}}\right] \\ &= \left[4.647 - 2.977\sqrt{\frac{0.347}{15}}; 4.647 + 2.977\sqrt{\frac{0.347}{15}}\right] \\ &= [4.194; 5.099] \end{aligned}$$

et on lit sur la sortie SAS communiquée, l'estimation par intervalle de confiance de l'écart-type au niveau 1% :

$$I' = [0.394; 1.092]$$

d- On pose :

$$\begin{aligned} H_0 &: \mu = 5 \text{ mm} \\ H_1 &: \mu \neq 5 \text{ mm} \end{aligned}$$

D'où :

$$t = \frac{\bar{x}_n - \mu_0}{\frac{s_n'}{\sqrt{n}}} = \frac{4.647 - 5.000}{\sqrt{\frac{0.347}{15}}} \approx -2.32$$

Donc, $-2.32 \in [-2.977; +2.977]$ soit $t \in [-t_{1-\frac{\alpha}{2};n-1}; t_{1-\frac{\alpha}{2};n-1}]$,
on accepte au risque 1% l'hypothèse H_0.

e- La plage est (principalement) constituée de sable d'après l'intervalle de confiance.

Partie B (7 pts)

a- Le test de comparaison des deux variances a pour P-value = "Pr $>$ F", la valeur 0.4285 donc au risque 1% on accepte l'égalité des variances.

Le test de comparaison des deux moyennes a pour P-value = "Pr $>$ |t|", la valeur 0.0891, donc on accepte l'égalité des deux tailles moyennes de grain au risque 1%.

b- On calcule pour comparer les variances :

$$t = max\left\{\left(\frac{0.589}{0.4859}\right)^2, \left(\frac{0.4859}{0.589}\right)^2\right\} = \left(\frac{0.589}{0.4859}\right)^2 \approx 1.47$$

et pour comparer les moyennes :

$$t' = \frac{4.647 - 4.965}{\sqrt{\frac{(15-1)0.589^2 + (20-1)0.4859^2}{15+20-2}\left(\frac{1}{15} + \frac{1}{20}\right)}} \approx -1.75$$

c- Au risque 1%, on calcule :

$$t = \frac{\bar{x}_n - \mu_0}{\frac{s'_n}{\sqrt{n}}} = \frac{4.965 - 5.200}{\frac{0.4859}{\sqrt{20}}} \approx -2.16$$

Donc la statistique est dans l'intervalle d'acceptation de la règle de décision puisque $t_{1-\frac{\alpha}{2};n-1} = t_{0.995;19} = 2.861$ donc on accepte l'hypothèse H_0, c'est à dire que la moyenne des tailles de grains de la seconde place vaut 5.2 mm.

d- Soit μ_1 la taille de tous les grains de la première plage, μ_2 celle de la seconde plage.

Non, les décisions ne concordent pas puisque $5.0 \neq 5.2$.
Il faut considérer ici l'erreur de seconde espèce, pour l'expliquer.

e- Puisque n est grand, pour les deux plages à la fois, l'étendue de intervalle de confiance de la moyenne, de valeur $\Delta 0.02$, s'écrit, au risque 1% :

$$\Delta = 2 \times t_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Soit, avec $t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.57$ et $\sigma = 0.5$:

$$n \geq \left(\frac{2 \times 2.57 \times 0.5}{0.02}\right)^2 \approx 128.5^2$$

Soit un nombre minimal de grains égal à :

$$n = 16513$$