

Université Lumière Lyon 2

M2-SISE

Statistique et Informatique pour la Science des données

DIS - Département d'Informatique et Statistique

Support 1 : Inférence Statistique

Rafik Abdesselam

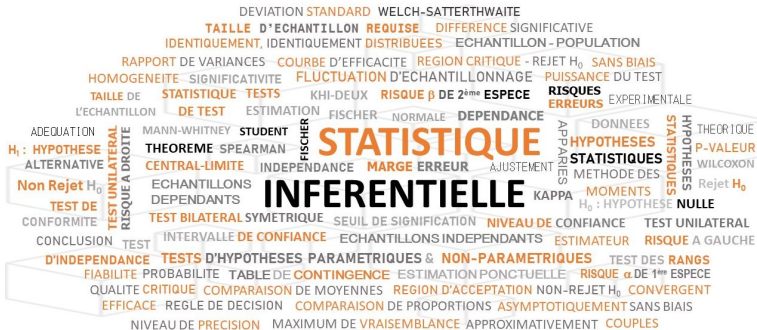
Courriel : rafik.abdesselam@univ-lyon2.fr

Web : <http://perso.univ-lyon2.fr/~rabdesse/fr/>

Support pédagogique : <http://perso.univ-lyon2.fr/~rabdesse/Documents/>

Décembre 2022

Plan du cours



- Introduction
- Estimation ponctuelle & Intervalle de confiance
- Tests d'hypothèses paramétriques

Objectif du cours

- L'objet de ce cours est de présenter des concepts d'**inférence statistique** : présenter des principes qui vont permettre, sur la base de résultats d'échantillon, d'estimer les valeurs des paramètres d'une population avec un niveau de confiance ou encore de vérifier certaines hypothèses statistiques posées sur les valeurs mêmes des paramètres.
- Les problèmes traités sont de deux types : l'**estimation** et les **tests** d'hypothèses de paramètres.
- Une bonne base de statistique et de probabilités est nécessaire pour bâtir une statistique inférentielle solide, qui soit non seulement un ensemble de tests-recettes, effectivement nécessaires, mais aussi l'expression du "**pourquoi**" et du "**comment**" de ces solutions.
- Sur le plan pédagogique, ce cours est accompagné de nombreux exemples réels traités avec les logiciels SAS & SPAD. Aucune connaissance préalable de ces logiciels n'est requise.

Objectif du cours

- **Pré-requis** : Quelques notions de Statistique & Probabilités.
- **Approche pédagogique** : (4) séances de cours magistraux (durée 3h - salle I401) et (4) séances de travaux dirigés (durée 3h - salle Informatique I213).
- Supports informatique : Logiciels SAS V9.4 et SPAD V9.2.
- **Matériel pédagogique** : (1) Polycopié de support de cours, (1) polycopié de travaux dirigés, des fiches aide mémoire des tests, des tables statistiques, (1) polycopié support SAS ainsi que de nombreux fichiers de données réelles (Tables SAS - Bases de données SPAD).
- **Contrôle des connaissances** : Un examen terminal (100%) - (salle Informatique - durée 1h30). Aucun document n'est autorisé.

Objectif du cours

- Quelques références bibliographiques :

[1] Dagnelie, P. Analyse statistique à plusieurs variables. Gembloux, Presses agronomiques, 1986, 362p.

[2] Mervyn, G. Marasinghe, William J. Kennedy, SAS for Data Analysis, Intermediate Statistical Methods. Statistics and Computing, Springer, 2008.

[3] Saporta, G. Probabilités Analyse des données et Statistique. Editions technip, 1990.

Introduction

- La statistique inférentielle ou confirmatoire, passerelle entre la statistique descriptive et la statistique mathématique, établit des relations entre populations et échantillons. On distingue deux types de démarche :
- ① **ECHANTILLONNAGE** (Population $\mapsto$ Echantillon) : démarche **déductive** de la statistique classique "du général au particulier" : **on connaît la population, on s'intéresse à l'échantillon.**
- ② **ESTIMATION** (Echantillon $\mapsto$ Population) : opposée à la démarche précédente, elle vise à étudier, à prédire les paramètres d'une population inconnue à partir des résultats obtenus grâce à des échantillons. C'est une démarche **inductive** "du particulier au général" : **on connaît l'échantillon, on s'intéresse à la population.**

1 - Echantillonnage

● Exemples :

- ① On connaît les professions d'une population cible dans laquelle est prélevé un échantillon. Est-ce que cet échantillon peut-être considéré comme représentatif de la population selon la variable profession ?
- ② Contrôle de la qualité : Est-ce qu'on peut considérer comme constant le diamètre moyen garanti d'une pièce usinée ?

● Méthodes de sélection d'un échantillon :

- ① **Echantillons probabilistes "aléatoires"** : *méthodes basées sur le calcul des probabilités et la théorie de l'estimation (meilleure précision, calcul d'intervalles de confiance).*
 - **aléatoires simples** (table de nombres aléatoires), **systématique** (taux de sondage, liste), en **grappes** (clusters), **strates** (proportionnelles ou non).
- ② **Echantillons non probabilistes "empiriques"** : *méthodes plus connues, faciles à mettre en oeuvre et très utilisées dans les sondages d'opinion (aucun calcul de précision, pas de liste de la population).*
 - **Méthode des quotas** (sondage raisonné), **sondage de convenance** ou de jugement (méthode des individus types).

2 - Estimation

- Exemples :

- 1 Avant des élections, des sondages sont effectués pour 'estimer' les chances des candidats.
- 2 Pour évaluer l'image d'un produit, on réalise un sondage auprès d'un échantillon de clients ; son analyse permet d'estimer l'indice de satisfaction moyen pour ce produit.

- Méthodes d'estimation :

- 1 Ponctuelle et par intervalle de confiance,
- 2 Tests d'hypothèses statistiques.

Estimation de paramètres

- La **théorie de l'estimation** dont l'objectif est d'**estimer** un ou plusieurs paramètres (inconnus) de la population par un **nombre** ou un **intervalle** dit de confiance.
- La **théorie des tests** dont l'objectif est de **confronter une hypothèse** concernant les paramètres théoriques (population) avec la réalité observée (échantillon) puis **décider**.
- Exemples :
 1. On étudie le rendement d'un titre de l'indice CAC40. Le rendement n'est pas constant. La population est l'ensemble des rendements du titre,
 - la **variable aléatoire de l'expérience** : le rendement,
 - les **paramètres d'intérêt** : l'espérance m (rendement moyen) et la variance σ^2 (volatilité) du titre.
 2. On étudie la non-conformité de pièces usinées. La population est l'ensemble des pièces de la production,
 - la **variable aléatoire de l'expérience** : Présence ou Absence de défauts de fabrication,
 - le **paramètre d'intérêt** : la proportion p avec laquelle une pièce est non-conforme.
- Ces méthodes reposent sur le choix d'un modèle statistique : **loi de probabilité** de la **variable aléatoire de l'expérience**.

Théorème Central Limite

- Soit $(X_i)_{i=1,n}$ une suite de v.a. **indépendantes** et de **même loi** (i.i.d.), dont les moments d'ordres un et deux **existent** notés $E(X_i) = m$ et $V(X_i) = \sigma^2$.

Posons la somme $S_n = \sum_{i=1}^n X_i \rightarrow E(S_n) = nm; V(S_n) = n\sigma^2$ alors :

$$\frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - nm}{\sigma\sqrt{n}} \rightarrow_{loi} N(0; 1)$$

La loi de la somme S_n est approximativement normale $N(nm, n\sigma^2)$ lorsque n est assez grand, quelle que soit la loi de probabilité des X_i (connue ou inconnue).

- Le T.C.L. justifie l'importance donnée à la loi normale, loi de probabilité la plus utilisée en statistique.
- Le T.C.L. fournit donc non seulement une méthode simple pour le calcul approximatif de probabilités liées à des sommes de v.a.r., mais il explique également ce fait empirique remarquable que bien des phénomènes naturels admettent une distribution en forme de cloche (fonction densité d'une loi normale).
- Dans beaucoup de situations, le phénomène étudié est la résultante d'un grand nombre de composantes indépendantes et le T.C.L. nous assure alors que la loi normale est tout à fait adéquate.

Moyenne : $\bar{X}_n = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow E(\bar{X}_n) = m; V(\bar{X}_n) = \frac{\sigma^2}{n}$ alors :

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow_{loi} N(0; 1)$$

Estimation ponctuelle

- Obtenir des estimations fiables de certains paramètres (inconnus) de la population échantillonnée.
- La loi, généralement inconnue, de cette population est caractérisée par un voire plusieurs paramètres inconnus notés θ , par exemple :

$$\theta = \begin{cases} \text{moyenne : } m, (m_1 - m_2), \\ \text{proportion : } p, (p_1 - p_2), \\ \text{variance : } \sigma^2, \frac{\sigma_1^2}{\sigma_2^2}. \end{cases}$$

- Le but d'une estimation ponctuelle est de fournir une **valeur approchée de θ** , en utilisant les observations $(x_1, x_2, \dots, x_n)$.
- L'idée de la procédure classique d'estimation consiste à choisir une **statistique** (variable aléatoire) particulière appelée "**estimateur**".
- Cet estimateur, noté $\bar{\theta}_n$, est une fonction des variables aléatoires $(X_1, X_2, \dots, X_n)$ i.i.d.
- L'estimation du paramètre θ de la population est la valeur (numérique) que prend cet estimateur selon les observations $(x_1, x_2, \dots, x_n)$ de l'échantillon.

Propriétés d'un estimateur

Ayant choisi un estimateur $\bar{\theta}_n$ pour estimer le paramètre θ , on souhaite évidemment que cet estimateur soit de **bonne qualité**, c'est-à-dire :

- ① Estimateur **sans biais** : $E(\bar{\theta}_n) = \theta$.
- ② Estimateur **asymptotiquement sans biais** : $E(\bar{\theta}_n) \rightarrow_{n \rightarrow \infty} \theta$
La précision de l'estimateur doit augmenter avec le nombre n d'observations : taille de l'échantillon.
- ③ Estimateur **convergent** en probabilité :
Sans biais ou **asymptotiquement sans biais** : $Var(\bar{\theta}_n) \rightarrow_{n \rightarrow \infty} 0$.
- ④ Estimateur **efficace** :
soient $\bar{\theta}_n$ et $\bar{\theta}'_n$ deux estimateurs **sans biais** de θ ,
 $\bar{\theta}_n$ est dit plus efficace que $\bar{\theta}'_n$ si $Var(\bar{\theta}_n) \leq Var(\bar{\theta}'_n)$.

Estimation par intervalle de confiance

- Les notions d'intervalle, de niveau de confiance, d'erreur et de risque, liées aux problèmes de distribution d'échantillonnage, se transposent sans difficulté aux problèmes d'estimations.
- L'estimation ponctuelle, bien qu'utile, ne fournit aucune information concernant la **précision de l'estimation** : elle ne tient pas compte de l'erreur possible dans l'estimation ; erreur attribuable aux fluctuations d'échantillonnage.
- On se propose donc d'estimer θ , paramètre inconnu de la population, par intervalle de confiance afin de **le cerner avec une certaine fiabilité**.
- Connaissant l'estimation ponctuelle $\bar{\theta}_n$ du paramètre inconnu θ de la population, il s'agit alors de déterminer un **intervalle** dans lequel il est **vraisemblable** que la vraie valeur de θ s'y trouve.

Estimation par intervalle de confiance

- On obtient cet intervalle en calculant deux limites auxquelles est associée une certaine assurance de contenir la **vraie valeur de θ** (inconnue).
- L'intervalle recherché, dans lequel doit se trouver θ avec une **probabilité $(1 - \alpha)$** , se définit à partir de la distribution de l'estimateur $\bar{\theta}_n$ d'après l'équation suivante :

$$P(\bar{\theta}_n - E \leq \theta \leq \bar{\theta}_n + E) = 1 - \alpha$$

et les limites prendront, après avoir prélevé l'échantillon et calculé l'estimation ponctuelle $\bar{\theta}_n$, la forme suivante :

$$\bar{\theta}_n - E \leq \theta \leq \bar{\theta}_n + E$$

où E : la marge dans l'estimation de θ (**précision**), sera déterminé à l'aide de l'écart-type de la distribution d'échantillonnage de $\bar{\theta}_n$ et du **niveau de confiance $(1 - \alpha)$** ou du **risque d'erreur α** choisi à priori.

Quelques lois de probabilités utiles

1 Loi du Khi-deux : une v.a.r. du khi-deux à n degrés de liberté, notée χ_n^2 , est la somme de n carrés de v.a.r. normales $N(0; 1)$ centrées réduites indépendantes :

$$\chi_n^2 = \sum_{i=1}^n U_i^2 \quad \left\{ \begin{array}{l} \forall (i,j) i \neq j \\ U_i \rightarrow N(0;1) \\ U_i \text{ et } U_j \text{ indépendantes} \end{array} \right.$$

2 Loi de Student : une v.a.r. de Student à n degrés de liberté, notée T_n , est le quotient d'une v.a.r. normale $N(0,1)$ par la racine carrée d'une v.a.r. du χ_n^2 divisée par son nombre de degrés de liberté n :

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}} \quad \left\{ \begin{array}{l} U \rightarrow N(0;1) \\ X \rightarrow \chi_n^2 \\ U \text{ et } X \text{ indépendantes} \end{array} \right.$$

3 Loi de Fisher : une v.a.r. de Fisher à m et n degrés de liberté, notée $F_{(m,n)}$, est le quotient de deux v.a.r. χ_m^2 et χ_n^2 indépendantes, divisées par leur nombre de d.l.l. m et n :

$$F_{(m,n)} = \frac{\frac{X}{m}}{\frac{Y}{n}} \quad \left\{ \begin{array}{l} X \rightarrow \chi_m^2 \\ Y \rightarrow \chi_n^2 \\ X \text{ et } Y \text{ indépendantes} \end{array} \right.$$

Hypothèses et conditions d'application

- **Hypothèses d'application :**

Soient $(x_1, x_2, \dots, x_n)$ n observations réalisées à partir de n v.a.r. $(X_1, X_2, \dots, X_n)$ indépendantes et de même loi de probabilité (i.i.d.) comme X de moyenne m et de variance σ^2 .

- 1 La loi de probabilité de X est normale ($\forall n$ la taille de l'échantillon prélevé).
- 2 Si la loi de probabilité de X est quelconque (connue ou inconnue) et n est grand ($n \geq 30$) : T.C.L. assure la normalité,
- 3 Si la loi de probabilité de X est quelconque (connue ou inconnue) et n est petit ($n < 30$) : ??? Utiliser plutôt des tests non-paramétriques

- **Conclusion :** *On se placera donc toujours dans le cas où l'échantillonnage s'effectue à partir d'une population normale $N(m, \sigma^2)$ ou d'un grand échantillon ($n \geq 30$).*

Estimation de paramètres

- Démarche à suivre pour mener à bien une estimation :
 - 1 Etablir ou construire le ou les estimateurs du paramètre (Méthode du Maximum de Vraisemblance, Méthode des Moments, ...),
 - 2 Choisir le "meilleur" ou le plus efficace (Vérifier les propriétés d'un "bon" estimateur),
 - 3 Déterminer la distribution de probabilité de l'estimateur retenu (Statistique de test),
 - 4 Etablir des intervalles de confiance et/ou des tests d'hypothèses sur le paramètre à estimer (prise de décision).

Exemple de construction d'un estimateur

- Lorsqu'il n'y a pas d'estimateur évident, on est amené à le construire.
- **Exemple** : $(X_1, \dots, X_n)$ i.i.d. comme $X \rightarrow \text{Exp}(\frac{1}{\theta})$: loi exponentielle de paramètre $\frac{1}{\theta}$, ($\theta > 0$), de fonction densité $f(x_i; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ avec $E(X) = \theta$ et $V(X) = \theta^2$
- 1 **Estimateur du Maximum de Vraisemblance (EMV)** : toute fonction $\bar{\theta}_n$ qui vérifie :
 $L(x_1, \dots, x_n; \bar{\theta}_n) = \text{Max}_{\theta} L(x_1, \dots, x_n; \theta) = \text{Max}_{\theta} \prod_{i=1}^n f(x_i; \theta)$
- 2 Vraisemblance : $L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) = \frac{1}{\theta^n} \exp(-\frac{1}{\theta} \sum_{i=1}^n x_i)$
- 3 log-vraisemblance : $\ln L(x_1, \dots, x_n; \theta) = -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^n x_i$
 - $\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{n}{\theta^2} \sum_{i=1}^n x_i = 0$ si $\theta = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$
 - $\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n x_i = \frac{n}{\theta^3} (\theta - 2\bar{x}_n) = -\frac{n}{\bar{x}_n^2} < 0$ pour $\theta = \bar{x}_n$,
- Donc $\bar{\theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
sans biais $E(\bar{\theta}_n) = \theta$ et convergent $V(\bar{\theta}_n) = \frac{\theta^2}{n} \rightarrow 0$.

Estimation ponctuelle de la moyenne : m

- $(X_1, X_2, \dots, X_n)$ i.i.d. $\rightarrow N(m; \sigma^2)$ ou grand échantillon ($n \geq 30$).
 $\forall i = 1, n \ E(X_i) = m; V(X_i) = \sigma^2$.
- Pour estimer la moyenne m de la population, on utilise le plus souvent la distribution d'échantillonnage de la moyenne dont l'estimateur est :

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

- 1 Estimateur **sans biais** :

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n m = \frac{nm}{n} = m$$

- 2 Estimateur **convergent** en probabilité :

$$V(\bar{X}_n) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} \rightarrow 0$$

- *La moyenne $\bar{x}_n$ observée sur l'échantillon est une estimation ponctuelle de la moyenne m de la population.*

Estimation de la moyenne : m (σ^2 connue)

- Lorsque la variance de la population σ^2 est connue, la distribution d'échantillonnage de $\bar{X}_n$ est approximativement normale de moyenne $E(\bar{X}_n) = m$ et de variance connue $Var(\bar{X}_n) = \frac{\sigma^2}{n}$.

- La statistique de test : $\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \hookrightarrow N(0; 1)$

- On peut alors écrire : $P(-u_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \leq u_{\frac{\alpha}{2}}) = 1 - \alpha$

On détermine les fractiles $u_{\frac{\alpha}{2}}$ de la loi $N(0; 1)$:

$$P(-u_{\frac{\alpha}{2}} \leq U \leq u_{\frac{\alpha}{2}}) = P(|U| \leq u_{\frac{\alpha}{2}}) = 1 - \alpha$$

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de m :

$$\bar{x}_n - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x}_n + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- *Marge d'erreur dans l'estimation de m : $E = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$*

- Intervalle "*bilatéral symétrique*" de niveau $1 - \alpha$ de la moyenne m centré en $\bar{x}_n$.

Remarque sur l'estimation de la variance σ^2 (m inconnue)

- Lorsque la variance σ^2 est inconnue on doit d'abord estimer la moyenne m pour estimer σ^2
- Etude de la statistique S_n^2 : Variance empirique d'échantillon

Estimateur : $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

L'estimation : $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ est la variance calculée à partir de l'échantillon de taille n avec $\bar{x}_n$ la moyenne de l'échantillon qui représente l'estimation ponctuelle de m.

- Décomposition de S_n^2 : $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

Partant de l'expression $(X_i - m) = (X_i - \bar{X}_n) + (\bar{X}_n - m)$,

on a alors, $\sum_{i=1}^n (X_i - m)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^n (\bar{X}_n - m)^2 + 2(\bar{X}_n - m) \sum_{i=1}^n (X_i - \bar{X}_n)$

comme, $\sum_{i=1}^n (X_i - \bar{X}_n) = 0$

on a, $\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + (\bar{X}_n - m)^2$

$$\Rightarrow S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X}_n - m)^2$$

Remarque sur l'estimation de la variance σ^2 (m inconnue)

- Biais de S_n^2 :

$$\begin{aligned} E(S_n^2) &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right] = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X}_n - m)^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i - m)^2 - E(\bar{X}_n - m)^2 \\ &= \frac{1}{n} \sum_{i=1}^n V(X_i) - V(\bar{X}_n) \\ &= \frac{1}{n} \sum_{i=1}^n \sigma^2 - \frac{\sigma^2}{n} = \frac{(n-1)}{n} \sigma^2 \end{aligned}$$

Cela montre que $E(S_n^2) \neq \sigma^2$.

S_n^2 est une statistique biaisée pour σ^2 . Le biais vaut $\frac{\sigma^2}{n}$

- Variance de S_n^2 :

On montre sans difficulté que :

$$\begin{aligned} V(S_n^2) &= V\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X}_n - m)^2\right] \\ &= \frac{n-1}{n^3} [(n-1)\mu_4 - (n-3)\sigma^4] \rightarrow \frac{\mu_4 - \sigma^4}{n} \rightarrow 0 \text{ si } n \rightarrow +\infty. \end{aligned}$$

$\mu_4 = E[(X_i - m)^4]$ le moment centré d'ordre 4.

Estimation de la moyenne : m (σ^2 inconnue)

- C'est généralement le cas. Lorsque la variance σ^2 est inconnue on doit d'abord estimer la moyenne m pour estimer σ^2 :

$$\text{Estimateur sans biais : } S_n^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

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

Dans ce cas, la distribution d'échantillonnage de $\bar{X}_n$ a pour moyenne $E(\bar{X}_n) = m$ et de variance estimée $\text{Var}(\bar{X}_n) = \frac{s_n^{*2}}{n}$.

- La statistique de test : $\frac{\bar{X}_n - m}{s_n^*/\sqrt{n}} \rightarrow T_{n-1} \text{ d.d.l.}$
- On peut alors écrire : $P(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - m}{s_n^*/\sqrt{n}} \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$

Les fractiles $t_{\frac{\alpha}{2}}$ de la loi de Student à n d.d.l. (cf. table) :

$$P(-t_{\frac{\alpha}{2}} \leq \bar{T}_n \leq t_{\frac{\alpha}{2}}) = P(|T_n| \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$$

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de m :

$$\bar{x}_n - t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}} \leq m \leq \bar{x}_n + t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$$

- Marge d'erreur dans l'estimation de m : $E = t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$

Statistique de test de la moyenne m (σ^2 inconnue)

- La statistique de test : $\frac{\bar{X}_n - m}{\frac{s_n^*}{\sqrt{n}}} \rightarrow T_{n-1} \text{ d.d.l.}$

- Preuve :

$$(1) \quad U = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0 ; 1)$$

$$(2) \quad S_n^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sigma} \right)^2$$

$$Y = \frac{n-1}{\sigma^2} S_n^{*2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sigma} \right)^2 = \sum_{i=1}^n U_i^2 \rightarrow \chi_{n-1}^2$$

$$\text{Définition : } T_{n-1} = \frac{U}{\sqrt{\frac{Y}{n-1}}} \quad \left\{ \begin{array}{l} U \rightarrow N(0 ; 1) \\ Y \rightarrow \chi_{n-1}^2 \\ U \text{ et } Y \text{ indépendantes} \end{array} \right.$$

On a alors,

$$\frac{(1)}{\sqrt{\frac{(2)}{n-1}}} = \frac{\frac{\sqrt{n}(\bar{X}_n - m)}{\sigma}}{\sqrt{\frac{(n-1)S_n^{*2}}{\sigma^2}}} = \frac{\bar{X}_n - m}{\frac{s_n^*}{\sqrt{n}}} \rightarrow T_{n-1}$$

Exemple d'application 1

- **Remarque - Approximation** : si la taille de l'échantillon est grande ($n \geq 30$) alors on peut remplacer la valeur du fractile $t_{\frac{\alpha}{2}}$ de Student à $(n-1)$ d.d.l. par celle du fractile $u_{\frac{\alpha}{2}}$ de la loi normale centrée-réduite $N(0, 1)$.
- Le chiffre d'affaires (C.A.) mensuel d'une entreprise est supposé normalement distribué.
Sur les 16 derniers mois, on a observé un C.A. mensuel moyen de 250 K.€. avec un écart-type de 52 K.€..
Etablir un intervalle de confiance niveau $1 - \alpha = 98\%$ du C.A. mensuel moyen.

Estimation ponctuelle d'une proportion : p

- Soient $A_1, \dots, A_i, \dots, A_n$ n événements indépendants de probabilité p. Pour estimer la proportion p de la population, on utilise la proportion de réalisation des événements A_i :

Estimateur sans biais : $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n 1_{A_i}$

Estimation : $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n P(A_i)$

- L'estimateur ainsi définit,

1 Estimateur **sans biais** :

$$E(\hat{P}_n) = E\left(\frac{1}{n} \sum_{i=1}^n 1_{A_i}\right) = \frac{1}{n} \sum_{i=1}^n E(1_{A_i}) = \frac{1}{n} \sum_{i=1}^n P(A_i) = \frac{np}{n} = p$$

2 Estimateur **convergent** en probabilité :

$$\begin{aligned} V(\hat{P}_n) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n 1_{A_i}\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(1_{A_i}) \\ &= \frac{1}{n^2} \sum_{i=1}^n P(A_i)(1 - P(A_i)) = \frac{1}{n^2} \sum_{i=1}^n p(1 - p) = \frac{npq}{n^2} \rightarrow 0 \end{aligned}$$

- La proportion $\hat{p}_n$ observée sur l'échantillon est une estimation ponctuelle de la proportion p de la population.*

Estimation ponctuelle d'une proportion p

- Remarque : on est ramené au cas précédent : estimation de la moyenne d'une loi de Bernouilli. En effet,

$$\left\{ \begin{array}{l} (X_1, \dots, X_i, \dots, X_n) \text{ i.i.d.} \\ X_i \rightarrow B(p) \text{ Bernoulli} \\ \forall i E(X_i) = p \text{ et } V(X_i) = pq \end{array} \right. \Rightarrow \left\{ \begin{array}{l} Y = \sum_i^n X_i \rightarrow B(n, p) \text{ Binomiale} \\ E(Y) = np \\ V(Y) = npq \text{ avec } q = 1 - p \end{array} \right.$$

↓

$$\left\{ \begin{array}{l} \hat{P}_n = Y/n = 1/n \sum_{i=1}^n X_i \\ E(\hat{P}_n) = \frac{1}{n} E(Y) = p \\ V(\hat{P}_n) = \frac{1}{n^2} V(Y) = \frac{pq}{n} \end{array} \right. \rightarrow (T.C.L. : n \geq 30) \left\{ \begin{array}{l} \hat{P}_n \rightarrow N(p; (\sqrt{\frac{pq}{n}})^2) \\ E(\hat{P}_n) = p \text{ et } V(\hat{P}) = \frac{pq}{n} \\ \frac{(\hat{P}_n - p)}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0; 1) \text{ Approximativement Normale} \end{array} \right.$$

Intervalle de confiance de p (n grand : $n \geq 30$)

- Dès lors que la taille de l'échantillon prélevé est grande ($n \geq 30$), l'estimation par intervalle de confiance de p (inconnue) de la population se déduit de la distribution d'échantillonnage de la proportion :

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (X_1, \dots, X_i, \dots, X_n) \text{ i.i.d. } X_i \rightarrow B(p)$$

- La statistique de test : $\frac{\hat{P}_n - p}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0; 1)$
- La distribution d'échantillonnage de $\hat{P}_n$ est **approximativement normale** de moyenne $E(\hat{P}_n) = p$ et de variance en fonction de p (inconnue) $Var(\hat{P}_n) = \frac{pq}{n} \leftarrow \frac{\hat{p}_n \hat{q}_n}{n}$ (estimation de la variance).

On peut alors écrire : $P(-u_{\frac{\alpha}{2}} \leq \frac{\hat{P}_n - p}{\sqrt{\frac{\hat{p}_n \hat{q}_n}{n}}} \leq u_{\frac{\alpha}{2}}) = 1 - \alpha$

Intervalle de confiance de p (n grand : $n \geq 30$)

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de p :

$$\hat{p}_n - u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n \hat{q}_n}{n}} \leq p \leq \hat{p}_n + u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n \hat{q}_n}{n}}$$

- Marge d'erreur dans l'estimation de p : $E = u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n \hat{q}_n}{n}}$.
- Intervalle "*bilatéral symétrique*" de niveau $1 - \alpha$ de la proportion p centré en $\hat{p}_n$.

Exemple d'application 2

- Un sondage effectué auprès d'un échantillon aléatoire de 50 employés d'une entreprise donne la répartition selon les catégories salariales suivantes :

Cat. salariale/mois	Nombre de salariés
Moins de 2 M.€.	18
[2 M.€. - 4 M.€.[	20
4 M.€. et plus	12
Total	50

- Donner une estimation ponctuelle de la proportion vraie d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€..
- Donner un intervalle de confiance pour p , la proportion vraie d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€., avec un risque d'erreur de 10%, de 5% et de 1%.

- 1 **La marge d'erreur ou niveau de précision** recherché dans l'estimation par intervalle de confiance, lorsqu'on utilise l'estimation $\bar{\theta}_n$ de l'échantillon pour estimer la vraie valeur θ de la population, est l'écart (en valeur absolue), noté $E = |\bar{\theta}_n - \theta|$.
- 2 En pratique, on peut fixer la marge d'erreur qu'on ne veut pas excéder et déterminer **la taille minimale de l'échantillon requise**.
- 3 On peut déduire **le risque d'erreur** ou le niveau de confiance attribué à une estimation par intervalle.

Paramètre	Marge d'erreur	Taille d'échantillon	Risque d'erreur
Moyenne m (σ^2 connue)	$E = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$n = (u_{\frac{\alpha}{2}} \frac{\sigma}{E})^2$	$u_{\frac{\alpha}{2}} = \frac{\sqrt{n}}{\sigma} E$
Moyenne m (σ^2 inconnue)	$E = t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$	$n = (t_{\frac{\alpha}{2}} \frac{s_n^*}{E})^2$	$t_{\frac{\alpha}{2}} = \frac{\sqrt{n}}{s_n^*} E$
Proportion p	$E = u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$n = (\frac{u_{\frac{\alpha}{2}}}{E})^2 \hat{p}\hat{q}$	$u_{\frac{\alpha}{2}} = \sqrt{\frac{n}{\hat{p}\hat{q}}} E$

Exemples d'application 3

- **Exemple 3.1 :** Le chiffre d'affaires (C.A.) mensuel d'une entreprise est supposé normalement distribué. Sur les 16 derniers mois, on a observé un C.A. mensuel moyen de 250 K.€. avec un écart-type de 52 K.€..

Déterminer la taille de l'échantillon requise de sorte que la marge d'erreur n'excède pas 15 K.€. avec un risque d'erreur $\alpha = 2\%$

- **Exemple 3.2 :** Le sondage effectué auprès d'un échantillon aléatoire de 50 employés d'une entreprise a montré que 32 employés ont un salaire supérieur ou égal à 2 M.€..

Déterminer la taille de l'échantillon requise "nombre d'employés à interroger" de sorte que la marge d'erreur, dans l'estimation de la proportion d'employés dont le salaire est supérieur ou égal à 2 M.€., n'excède pas 4% avec un risque d'erreur $\alpha = 5\%$

- **Exemple 3.3 :** On effectue une enquête afin de déterminer la proportion de personnes en âge de voter qui exerceront leur droit de vote lors de la prochaine élection que l'on peut estimer à 50%. On exige une estimation à 95% de la proportion des votants avec une marge d'erreur dans l'estimation qui n'excède pas 3%.

- 1 Quel doit-être le nombre minimal de personnes à interroger pour respecter les conditions imposées ?
- 2 Quel serait ce nombre si on vous demandait une précision de 1.5% ?
- 3 Quel est le niveau de confiance $1 - \alpha$ que l'on peut attribuer à cet intervalle bilatéral symétrique : $p \in [48.71\% ; 51.29\%]$ de la proportion des votants sachant que l'on a interrogé 10 000 personnes ?

Estimation de la variance σ^2 (m connue)

- $x_1, x_2, \dots, x_n$ n observations indépendantes de même loi de moyenne m et de variance σ^2 . Pour estimer σ^2 , si **la moyenne m est connue**, on peut construire l'estimateur :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 \quad \leftarrow \text{(variance mathématique).}$$

- 1 Estimateur **sans biais** :

$$\begin{aligned} E(S_n^2) &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - m)^2\right) = \frac{1}{n} \sum_{i=1}^n E[(X_i - m)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n} = \sigma^2 \end{aligned}$$

- 2 Estimateur **convergent** :

$$\begin{aligned} V(S_n^2) &= V\left(\frac{1}{n} \sum_{i=1}^n (X_i - m)^2\right) = \frac{1}{n^2} \sum_{i=1}^n V[(X_i - m)^2] \\ &= \frac{1}{n^2} \sum_{i=1}^n (E[(X_i - m)^4] - [E(X_i - m)^2]^2) \\ &= \frac{1}{n^2} \sum_{i=1}^n (E[(X_i - m)^4] - \sigma^4) = \frac{1}{n^2} \sum_{i=1}^n \text{cste} \rightarrow 0 \end{aligned}$$

- *La variance $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$ observée sur l'échantillon est une estimation ponctuelle de la variance σ^2 de la population échantillonnée lorsque la moyenne m de la population est connue.*

Estimation de la variance σ^2 (m connue)

- Lorsque la moyenne m est connue, on peut montrer que :

$$\sum_{i=1}^n \frac{(X_i - m)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 = \sum_{i=1}^n U_i^2 \rightarrow \chi_{n \text{ d.d.l.}}^2 \text{ avec } U_i \rightarrow N(0, 1).$$

(cf. définition d'une variable aléatoire du khi-deux comme somme de carrés de variables aléatoires normales centrées réduites indépendantes).

- La statistique de test : $\sum_{i=1}^n \frac{(X_i - m)^2}{\sigma^2} = n \frac{S_n^2}{\sigma^2} \rightarrow \chi_{n \text{ d.d.l.}}^2$
- On peut alors écrire : $P(k_1 \leq n \frac{S_n^2}{\sigma^2} \leq k_2) = 1 - \alpha$

où, $k_1 = \chi_{\frac{\alpha}{2}}^2$ et $k_2 = \chi_{1 - \frac{\alpha}{2}}^2$ sont les fractiles de la du khi-deux à n degrés de liberté (cf. table du khi-deux).

c'est-à-dire : $P(\chi_n^2 \leq k_1) = \frac{\alpha}{2}$ et $P(\chi_n^2 \leq k_2) = 1 - \frac{\alpha}{2}$

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de σ^2 :

$$n \frac{S_n^2}{k_2} \leq \sigma^2 \leq n \frac{S_n^2}{k_1}$$

Estimation de la variance σ^2 (m inconnue)

- Lorsque la moyenne m est inconnue (cas le plus fréquent), pour estimer σ^2 , on pourrait utiliser naturellement l'estimateur :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \leftarrow \text{après avoir estimé } m.$$

Cependant, l'estimateur S_n^2 est biaisé : $E(S_n^2) = \frac{n-1}{n} \sigma^2$, on préfère alors utiliser l'estimateur : $S_n^{*2} = \frac{n}{n-1} S_n^2$ appelé : **carré de la déviation standard empirique**.

- 1 Estimateur **sans biais** :

$$E(S_n^{*2}) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} E(S_n^2) = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2$$

- 2 Estimateur **convergent** :

$$V(S_n^{*2}) = V\left(\frac{n}{n-1} S_n^2\right) = \frac{n^2}{(n-1)^2} V(S_n^2) \approx V(S_n^2) \rightarrow 0$$

- *La variance $s_n^{*2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n}{n-1} s_n^2$ observée sur l'échantillon est une estimation ponctuelle de la variance σ^2 de la population échantillonnée lorsque la moyenne m de la population*

Estimation de la variance σ^2 (m inconnue)

- Lorsque la moyenne m est inconnue, on peut également montrer que :

- la statistique de test :

$$\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} = (n-1) \frac{S_n^{*2}}{\sigma^2} \rightarrow \chi_{(n-1)}^2 \text{ d.d.l.}$$

- On peut alors écrire : $P(k_1 \leq (n-1) \frac{S_n^{*2}}{\sigma^2} \leq k_2) = 1 - \alpha$

où, $k_1 = \chi_{\frac{\alpha}{2}}^2$ et $k_2 = \chi_{1-\frac{\alpha}{2}}^2$ sont les fractiles de la du khi-deux à $(n-1)$ degrés de liberté (cf. table du khi-deux).

c'est-à-dire : $P(\chi_{(n-1)}^2 \leq k_1) = \frac{\alpha}{2}$ et $P(\chi_{(n-1)}^2 \leq k_2) = 1 - \frac{\alpha}{2}$

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de σ^2 :

$$(n-1) \frac{S_n^{*2}}{k_2} \leq \sigma^2 \leq (n-1) \frac{S_n^{*2}}{k_1}$$

- Ou encore pour l'écart-type σ :

$$\sqrt{(n-1) \frac{S_n^{*2}}{k_2}} \leq \sigma \leq \sqrt{(n-1) \frac{S_n^{*2}}{k_1}}$$

Exemple d'application 4

- Le chiffre d'affaires (C.A.) mensuel d'une entreprise est supposé normalement distribué. Sur les 16 derniers mois, on a observé un C.A. mensuel moyen de 250 K.€. avec un écart-type de 52 K.€..
 - 1 Donner une estimation ponctuelle de l'écart-type σ du chiffre d'affaires mensuel de cette entreprise.
 - 2 Etablir un intervalle de confiance de niveau $1 - \alpha = 95\%$ de l'écart-type σ .
 - 3 Peut-on conclure avec un risque d'erreur $\alpha = 5\%$, que l'écart-type σ du C.A. mensuel de cette entreprise est significativement différent de 80 K.€ ?

Cas d'un tirage sans remise

- Si la population échantillonnée a un nombre fini d'individus de taille N , on conçoit que la loi de la population change après chaque tirage et que les tirages ne soient pas indépendants.
- Cependant, pour l'estimation de la moyenne $E(\bar{X}_n) = m$:

$$V(\bar{X}_n) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \quad (\text{variance connue})$$

$$V(\bar{X}_n) = \frac{s^{*2}}{n} \frac{N-n}{N-1} \quad (\text{variance inconnue})$$

de même, pour l'estimation d'une proportion $E(\hat{P}_n) = p$:

$$V(\hat{P}_n) = \frac{\hat{p}\hat{q}}{n} \frac{N-n}{N-1}$$

- Donc, lorsque l'échantillonnage s'effectue sans remise (tirage exhaustif) à partir d'une population finie de taille N , on doit apporter un **facteur de correction** : $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$ à la variance de l'estimateur, si le **taux de sondage** $t = \frac{n}{N} > 10\%$.

Toutefois ce facteur de correction peut être ignoré ($\frac{N-n}{N-1} \approx 1$) si le **taux de sondage** $t = \frac{n}{N} \leq 10\%$.

Exemple d'application 5

- Supposons que cette entreprise compte 200 employés et que l'échantillon de 50 employés a été prélevé au hasard parmi les deux cents.

Cat. salariale/mois	Nombre de salariés
Moins de 2 M.€.	18
[2 M.€. - 4 M.€.]	20
4 M.€. et plus	12
Total	50

- Donner une estimation de la proportion de l'ensemble des employés dont le salaire mensuel est de 2 M.€ et plus.
- Quel est le taux de sondage ?
- Etablir un intervalle de confiance de niveau $1 - \alpha = 95\%$ de la proportion d'employés de cette entreprise dont le salaire est supérieur ou égal à 2 M.€.
- Déterminer la probabilité qu'au moins 30 employés de cet échantillon possèdent un salaire mensuel de 2 M.€ et plus, lorsque la population échantillonnée en contient 64%.

Comparaisons

- Il existe de nombreuses applications qui consistent, par exemple, à **comparer deux groupes d'individus** en regard d'un caractère particulier (poids, taille, rendement,...), ou comparer deux procédés de fabrication selon une caractéristique (résistance, diamètre, longueur,...), ou encore comparer les proportions d'apparition d'un caractère de deux populations (proportion de défectueux, proportion de gens favorisant un parti politique,...).
- Les distributions d'échantillonnage qui sont alors utilisées pour effectuer ces comparaisons 'Tests d'hypothèses' ou 'calcul d'intervalles de confiance' sont celles correspondant aux fluctuations d'échantillonnage de la **différence** de **2 moyennes**, de **2 proportions** ou encore le **rapport** de **2 variances** observées .

Estimation de la différence de 2 moyennes

- On prélève des échantillons $x_1, x_2, \dots, x_n$ et $y_1, y_2, \dots, y_p$ dans deux populations distinctes. On considère que ces échantillons sont des réalisations de v.a.r. indépendantes $X_1, X_2, \dots, X_n$ et $Y_1, Y_2, \dots, Y_p$ les premières de loi de probabilité L_x , les secondes de loi de probabilité L_y telles que :

2 Populations, 2 échantillons **indépendants**

$$\begin{cases} x_1, x_2, \dots, x_n \leftrightarrow \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \\ y_1, y_2, \dots, y_p \leftrightarrow \bar{Y}_p = \frac{1}{p} \sum_{j=1}^p Y_j \end{cases}$$

- On suppose la **normalité** de la variable dans les 2 populations ou **grands échantillons**, avec respectivement des moyennes m_x et m_y , et des variances σ_x^2 et σ_y^2 qui **existent** (connues ou inconnues).

$$\begin{cases} \forall i = 1, n \\ E(X_i) = m_x \text{ et } V(X_i) = \sigma_x^2 \end{cases} \quad \begin{cases} \forall j = 1, p \\ E(Y_j) = m_y \text{ et } V(Y_j) = \sigma_y^2 \end{cases}$$

Estimation de la différence de 2 moyennes

- ① Il est évidemment intéressant d'estimer la différence $(m_x - m_y)$ et on le fait naturellement par la différence des distributions d'échantillonnages des moyennes : $\bar{X}_n - \bar{Y}_p$

Estimateur **sans biais** :

$$E(\bar{X}_n - \bar{Y}_p) = E(\bar{X}_n) - E(\bar{Y}_p) = m_x - m_y$$

Estimateur **convergent** :

$$V(\bar{X}_n - \bar{Y}_p) = V(\bar{X}_n) + V(\bar{Y}_p) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}.$$

La différence des moyennes $(\bar{x}_n - \bar{y}_p)$ observée sur les échantillons est une estimation ponctuelle de la différence $(m_x - m_y)$ des moyennes des populations.

- ② **Remarque** : De même pour l'estimation ponctuelle de la différence de 2 proportions, la différence $(\hat{p}_x - \hat{p}_y)$ observée sur les échantillons est une estimation ponctuelle de la différence des proportions $(p_x - p_y)$ des populations.

Estimation de la différence de 2 moyennes

Cas où les variances σ_x^2 et σ_y^2 sont connues

- Moyennes : $E(\bar{X}_n) = m_x$ et $E(\bar{Y}_p) = m_y \Rightarrow E(\bar{X}_n - \bar{Y}_p) = m_x - m_y$.

$$\text{Variances : } V(\bar{X}_n) = \frac{\sigma_x^2}{n} \text{ et } V(\bar{Y}_p) = \frac{\sigma_y^2}{p} \Rightarrow V(\bar{X}_n - \bar{Y}_p) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}$$

- La distribution d'échantillonnage de la différence $(\bar{X}_n - \bar{Y}_p)$:

$$(\bar{X}_n - \bar{Y}_p) \rightarrow N(m_x - m_y ; (\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}})^2)$$

- La statistique de test :
$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}} \rightarrow N(0; 1)$$

- I.C. $_{1-\alpha}$ de la différence de moyennes $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) - u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}} \leq m_x - m_y \leq (\bar{x}_n - \bar{y}_p) + u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}$$

- Marge d'erreur dans l'estimation de $(m_x - m_y)$:

$$E = u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}$$

Exemple d'application 6

- Le temps mis par une machine pour fabriquer une pièce est supposé suivre une loi normale de paramètres m et σ^2 . Dans un atelier, deux machines A et B fabriquent la même pièce. Pour un échantillon de 9 pièces fabriquées, on a obtenu les résultats suivants :

	Machine A	Machine B
Nombre de pièces fabriquées	9	9
Temps moyen observé (mn)	50	45
Variances des populations	25	36

- Déterminer un intervalle de confiance, de niveau $(1 - \alpha) = 95\%$, de la différence des temps moyens des deux machines $m_a - m_b$.
- Question : La machine B est-elle aussi performante que la machine A ?

Estimation de la différence de 2 moyennes

Cas où les variances σ_x^2 et σ_y^2 sont inconnues

Grands échantillons : $n \geq 30$ et $p \geq 30$

- Le cas précédant est évidemment peu courant en pratique ; voyons à présent, dans les mêmes conditions que ci-dessus, les cas les plus fréquents.
- Si les échantillons prélevés dans chaque population (quelconque, par forcément normale) sont de grandes tailles alors on peut remplacer les variances inconnues σ_x^2 et σ_y^2 par leur estimation respective s_x^{*2} et s_y^{*2} .
- La distribution d'échantillonnage de la différence $(\bar{X}_n - \bar{Y}_p)$ est approximativement normale.
- La statistique de test :
$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}}} \rightarrow N(0; 1)$$
- Ce qui fournit aisément un intervalle de confiance de niveau $(1 - \alpha)$ pour la différence $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) - u_{\frac{\alpha}{2}} \sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}} \leq m_x - m_y \leq (\bar{x}_n - \bar{y}_p) + u_{\frac{\alpha}{2}} \sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}}$$

Exemple d'application 7

- On fait subir à des cadres intermédiaires de deux grandes entreprises (une oeuvrant dans la fabrication d'équipement de transport et l'autre dans la fabrication de produits électriques) un test d'appréciation et d'évaluation. La compilation des résultats pour chaque groupe à l'issue de cette évaluation s'établit comme suit :

	1 Equipement	2 Produits Electriques
Nombre de cadres	34	32
Appréciation globale moyenne	184	178
Somme des Carrés des Ecart	15774	9858

- 1 Déterminer un intervalle de confiance qui a 95 chances sur 100 de contenir la valeur vraie de la différence des moyennes ($m_1 - m_2$) des deux groupes de cadres.
- 2 Question : Selon cet intervalle, que peut-on conclure quant à la performance des cadres de ces deux secteurs au test d'évaluation ? Est-ce qu'en moyenne, la performance est vraisemblablement identique ou semble-t-il une différence significative entre ces deux groupes ?

Intervalle de confiance de la différence de 2 moyennes

1. Cas où les variances sont inconnues et égales $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Petits échantillons n (et/ou) $p < 30$: Populations normales

- Dans ce cas, on ne peut pas remplacer les variances inconnues σ_x^2 et σ_y^2 par leur estimation s_x^{*2} et s_y^{*2} calculées sur chacun des échantillons (elles seront peu précises).
- Puisqu'on les suppose égales à une valeur inconnue σ^2 , on se servira de l'information des deux échantillons pour obtenir une estimation unique s^{*2} , de la variance $\sigma^2 = \sigma_x^2 = \sigma_y^2$.

Intervalle de confiance de la différence de 2 moyennes

1. *Cas où les variances sont inconnues et égales* $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Petits échantillons n (et/ou) $p < 30$: Populations normales

- On montre que : $S^{*2} = \frac{nS_x^2 + pS_y^2}{n+p-2}$ est un bon estimateur de σ^2 .

$$\text{Moyenne : } E(\bar{X}_n - \bar{Y}_p) = m_x - m_y$$

$$\text{Variance : } V(\bar{X}_n - \bar{Y}_p) = \frac{s^{*2}}{n} + \frac{s^{*2}}{p} = s^{*2} \left(\frac{1}{n} + \frac{1}{p} \right)$$

- La statistique de test :

$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{s^* \sqrt{\frac{1}{n} + \frac{1}{p}}} \rightarrow T_{(n+p-2)} d.d.l.$$

- D'où l'intervalle de confiance de niveau $(1 - \alpha)$ de $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) - t_{\frac{\alpha}{2}} s^* \sqrt{\frac{1}{n} + \frac{1}{p}} \leq m_x - m_y \leq (\bar{x}_n - \bar{y}_p) + t_{\frac{\alpha}{2}} s^* \sqrt{\frac{1}{n} + \frac{1}{p}}$$

Intervalle de confiance de la différence de 2 moyennes

1. Cas où les variances sont inconnues et égales $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Petits échantillons n (et/ou) $p < 30$: Populations normales

- On montre que : $S^{*2} = \frac{nS_x^2 + pS_y^2}{n+p-2}$ est un bon estimateur de σ^2 .

$$\text{Moyenne : } E(\bar{X}_n - \bar{Y}_p) = m_x - m_y$$

$$\text{Variance : } V(\bar{X}_n - \bar{Y}_p) = \frac{s^{*2}}{n} + \frac{s^{*2}}{p} = s^{*2} \left(\frac{1}{n} + \frac{1}{p} \right)$$

- La statistique de test :

$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{s^* \sqrt{\frac{1}{n} + \frac{1}{p}}} \rightarrow T_{(n+p-2) d.d.l.}$$

- D'où l'intervalle de confiance de niveau $(1 - \alpha)$ de $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) - t_{\frac{\alpha}{2}} s^* \sqrt{\frac{1}{n} + \frac{1}{p}} \leq m_x - m_y \leq (\bar{x}_n - \bar{y}_p) + t_{\frac{\alpha}{2}} s^* \sqrt{\frac{1}{n} + \frac{1}{p}}$$

Cas particulier

Variances inconnues et égales $\sigma_x^2 = \sigma_y^2 = \sigma^2$

- Si $n = p$ (échantillons indépendants de même taille), on a plus simplement : $s^{*2} = \frac{n(s_x^2 + s_y^2)}{2(n-1)}$
- La statistique de test :

$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{s^* \sqrt{\frac{2}{n}}} \rightarrow T_{2(n-1) d.d.l.}$$

- Les Limites de l'intervalle de confiance de $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) \pm t_{\frac{\alpha}{2}} s^* \sqrt{\frac{2}{n}}$$

!!! A NE PAS CONFONDRE AVEC LE CAS DE
2 ECHANTILLONS DEPENDANTS - APPARIES!!!

Preuve - Statistique de test de Student à $n+p-2$ d.d.l. :

s^{*2} : estimation ponctuelle de la variance commune : $\sigma^2 = \sigma_x^2 = \sigma_y^2$.

par définition,
$$T_{n+p-2} = \frac{U}{\sqrt{\frac{Y}{n+p-2}}} \quad \left\{ \begin{array}{l} U \rightarrow N(0 ; 1) \\ Y \rightarrow \chi_{n+p-2}^2 \\ U \text{ et } Y \text{ indépendantes} \end{array} \right.$$

On sait :
$$\left\{ \begin{array}{ll} \text{(1)} \bar{X}_n \rightarrow N(m_x ; \frac{\sigma_x^2}{n} = \frac{\sigma^2}{n}) & \text{et } \bar{Y}_p \rightarrow N(m_y ; \frac{\sigma_y^2}{p} = \frac{\sigma^2}{p}) \\ \text{(2)} \frac{(n-1)S_x^{*2}}{\sigma^2} \rightarrow \chi_{n-1}^2 & \text{et } \frac{(p-1)S_y^{*2}}{\sigma^2} \rightarrow \chi_{p-1}^2 \end{array} \right.$$

On a,
$$\text{(1)} \left\{ \begin{array}{l} (\bar{X}_n - \bar{Y}_p) \rightarrow N(m_x - m_y ; \frac{\sigma^2}{n} + \frac{\sigma^2}{p} = \left(\sigma \sqrt{\frac{1}{n} + \frac{1}{p}} \right)^2) \\ \frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{p}}} \rightarrow N(0 ; 1) \end{array} \right.$$

Différence de 2 v.a.r. normales indépendantes.

Propriété,
$$\left. \begin{array}{l} X \rightarrow \chi_m^2 \\ Y \rightarrow \chi_n^2 \\ X \text{ et } Y \text{ indépendantes} \end{array} \right\} \Rightarrow X + Y \rightarrow \chi_{m+n}^2$$

Preuve - Statistique de test de Student à $n+p-2$ d.d.l. :

On a, (2) $\left\{ \begin{array}{l} \frac{(n-1)S_x^{*2}}{\sigma^2} \rightarrow \chi_{n-1}^2 \\ \frac{(p-1)S_y^{*2}}{\sigma^2} \rightarrow \chi_{p-1}^2 \end{array} \right. \Rightarrow \frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{\sigma^2} \rightarrow \chi_{n+p-2}^2$

Somme de 2 v.a.r. du Khi-deux χ_{n-1}^2 et χ_{p-1}^2 indépendantes.

$$\left. \begin{array}{l} (1) \frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{p}}} \rightarrow N(0; 1) \\ (2) \frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{\sigma^2} \rightarrow \chi_{n+p-2}^2 \end{array} \right\} \Rightarrow \frac{(1)}{\sqrt{\frac{(2)}{n+p-2}}} \rightarrow T_{n+p-2}$$

on a alors, $\frac{(1)}{\sqrt{\frac{(2)}{n+p-2}}} = \frac{\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{p}}}}{\sqrt{\frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{\sigma^2} \cdot \frac{1}{n+p-2}}} = \frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sqrt{\frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{n+p-2}} \sqrt{\frac{1}{n} + \frac{1}{p}}} = \frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{s^* \sqrt{\frac{1}{n} + \frac{1}{p}}}$

avec, $s^* = \sqrt{\frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{n+p-2}}$

L'estimation de la variance commune $\sigma^2 = \sigma_x^2 = \sigma_y^2$:

$$s^{*2} = \frac{(n-1)S_x^{*2} + (p-1)S_y^{*2}}{n+p-2} = \frac{nS_x^2 + pS_y^2}{n+p-2} = \frac{SCE_x + SCE_y}{n+p-2}$$

Exemple d'application 8

- Un laboratoire indépendant a effectué, pour le compte d'une revue sur la protection du consommateur, un essai de durée de vie sur un type d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur de produits d'éclairage. Les durées de vie des ampoules des deux fabricants sont supposées normalement distribuées et dont **les variances sont supposées égales**. Les essais effectués dans les mêmes conditions, sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants :

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Ecart	2400	2800

- Déterminer un intervalle de confiance de niveau 95% de la différence des durées de vie moyennes des ampoules de ces deux fabricants.
- Question : Est-ce que la revue peut affirmer, qu'en moyenne, les durées de vie des ampoules des deux fabricants sont identiques (ou différentes) ? En d'autres termes, est-ce que la différence observée lors des essais est significative ?

Intervalle de confiance de la différence de 2 moyennes

2. Cas où les variances sont inconnues et différentes $\sigma_x^2 \neq \sigma_y^2$

Petits échantillons n (et/ou) $p < 30$: Populations normales

- Dans ce cas, les variances inconnues σ_x^2 et σ_y^2 sont remplacés par leur estimation s_x^{*2} et s_y^{*2} , mais le degré de liberté $\mathbf{v}$ de la loi de Student suivie par la statistique n'est pas connu :

$$\frac{(\bar{X}_n - \bar{Y}_p) - (m_x - m_y)}{\sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}}} \rightarrow T_{\mathbf{v}}$$

- Il faut déterminer $\mathbf{v}$ pour pouvoir lire le fractile $t_{(\frac{\alpha}{2}; \mathbf{v})}$ dans la table de la loi de Student.
- Le calcul de $\mathbf{v}$ est approché par l'équation de **Welch-Satterthwaite** :

$$\mathbf{v} = \frac{(\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p})^2}{(\frac{s_x^{*2}}{n})^2 + (\frac{s_y^{*2}}{p})^2} = \frac{(\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p})^2}{\frac{s_x^{*4}}{n^2(n-1)} + \frac{s_y^{*4}}{p^2(p-1)}}$$

Intervalle de confiance de la différence de 2 moyennes

2. Cas où les variances sont inconnues et différentes $\sigma_x^2 \neq \sigma_y^2$

Petits échantillons n (et/ou) $p < 30$: Populations normales

- Le degré de liberté ν calculé est généralement un **nombre réel**. Pour déterminer le fractile $t_{(\frac{\alpha}{2}, \nu)}$, on peut appliquer une interpolation linéaire :
 $t_{(\frac{\alpha}{2}, [\nu]+1)} < t_{(\frac{\alpha}{2}, \nu)} < t_{(\frac{\alpha}{2}, [\nu])}$: $[\nu]$ désigne la partie entière de ν .
- Le degré de liberté de ce test t de Welch, $\nu < n + p - 2$, (cas précédent où les variances étaient égales).
- Lorsque $n \approx p$, alors les degrés de liberté sont approximativement les mêmes dans les deux cas (variances égales ou variances inégales),
 $\nu \simeq n + p - 2$.
- D'où l'intervalle de confiance de niveau $(1 - \alpha)$ de $(m_x - m_y)$:

$$(\bar{x}_n - \bar{y}_p) - t_{(\frac{\alpha}{2}, \nu)} \sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}} \leq m_x - m_y \leq (\bar{x}_n - \bar{y}_p) + t_{(\frac{\alpha}{2}, \nu)} \sqrt{\frac{s_x^{*2}}{n} + \frac{s_y^{*2}}{p}}$$

Exemple d'application 9

- Un organisme financier a comparé les rendements annuels de deux types de banques, l'une privée, l'autre publique.

	1 : Privée	2 : Publique
Nombre d'essais	$n_1 = 8$	$n_2 = 10$
rendement moyen observé (%)	$\bar{x}_1 = 9.8$	$\bar{x}_2 = 6.9$
Somme des Carrés des Ecart	$SCE_1 = 250$	$SCE_2 = 61$

Les rendements annuels sont normalement distribués et dont **les variances inconnues sont supposées différentes**.

- 1 Etablir l'intervalle de confiance de niveau $1 - \alpha = 95\%$ de la différence des rendements annuels moyens. Peut-on conclure avec un risque d'erreur $\alpha = 5\%$, que les rendements annuels moyens de ces deux banques sont différents ?
- 2 Question : Peut-on affirmer que l'écart observé des rendements annuels moyens de 2.9% est significatif ?

Exemple - Solution 9

- Remarques : petits échantillons $n_1 = 8$ et $n_2 = 10$ indépendants. Les rendements annuels sont normalement distribués avec des variances σ_1^2 et σ_2^2 inconnues supposées différentes $\sigma_1^2 \neq \sigma_2^2$.

- Statistique de test :
$$\frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}} \mapsto T_{v \text{ d.d.l.}}$$

- Les données : $n_1 = 8$; $s_1^{*2} = \frac{SCE_1}{(n_1 - 1)} = 35.71$; $n_2 = 10$; $s_2^{*2} = \frac{SCE_2}{(n_2 - 1)} = 6.78$.

Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.

Calcul du nombre de degrés de liberté v à partir de l'expression de Welch-Satterthwaite :

$$v = \frac{(\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2})^2}{\frac{\frac{s_1^{*4}}{n_1^2} + \frac{s_2^{*4}}{n_2^2}}{(n_1 - 1) + (n_2 - 1)}} = \frac{(\frac{35.71}{8} + \frac{6.78}{10})^2}{\frac{35.71^4}{8^2 \times 7} + \frac{6.78^4}{10^2 \times 9}} = 9.1233$$

Interpolation linéaire du fractile de la loi de Student : cf. Table à $v = 9$ d.d.l. et à $v = 10$ d.d.l.

$$t_{(2.5\% ; v=10 \text{ d.d.l.})} = 2.2281 \text{ et } t_{(2.5\% ; v=9 \text{ d.d.l.})} = 2.2622 \Rightarrow t_{(\frac{\alpha}{2}=2.5\% ; v=9.12 \text{ d.d.l.})} = \mp 2.2580$$

$$\text{Marge d'erreur : } E = t_{(\frac{\alpha}{2} ; v=9.12)} \sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}} = 2.2580 \times \sqrt{\frac{35.71}{8} + \frac{6.78}{10}} = 5.12\%.$$

Estimation ponctuelle de la différence ($m_1 - m_2$) : $\bar{x}_1 - \bar{x}_2 = 9.8 - 6.9 = 2.9\%$.

- Intervalle de confiance de niveau 95% de ($m_1 - m_2$) :

$$2.9 - 5.12 = -2.22 \leq (m_1 - m_2) \leq 8.02 = 2.9 + 5.12$$

$$I.C._{95\%} : (m_1 - m_2) \in [-2.22\% , 8.02\%]$$

- Conclusion : 0 appartient à $I.C._{95\%}$, l'écart observé de 2.9% sur les rendements annuels moyens n'est pas significatif (avec un risque d'erreur de 5%). Cet écart est donc attribuable aux fluctuations d'échantillonnage. L'organisme financier peut conclure avec un risque d'erreur de 5%, que les rendements annuels moyens de ces deux banques sont identiques : $m_1 \approx m_2$.

Différence de 2 moyennes : Echantillons appariés

Echantillons dépendants - Données associées par paires

- **Exemple 1** : On compare 2 méthodes de mesures en soumettant à ces méthodes les mêmes individus. Les 2 échantillons sont issus de deux lois différentes, mais ne sont pas indépendants (en général !).
- **Exemple 2** : Lorsque nous avons, pour chaque élément de l'échantillon, deux valeurs obtenues à des périodes différentes (avant / après) ou selon des traitements différents.
- Dans ce cas les deux séries de mesures **ne sont pas indépendantes** l'une de l'autre, elles sont **dépendantes**.
- Ce cas est **différent** du test de comparaison de moyennes d'échantillons indépendants.

Différence de 2 moyennes : Echantillons appariés

Echantillons dépendants - Données associées par paires

La méthode pour comparer deux échantillons appariés - Méthode des couples consiste à établir :

- pour chaque paire, la différence $d_i = x_i - y_i$, on se ramène ainsi à **un seul échantillon** différence $(d_1, d_2, \dots, d_n)$ de taille n ,
- $D_1 = (X_1 - Y_1), \dots, D_n = (X_n - Y_n)$ n variables aléatoires **i.i.d.**,
- $\bar{D}_n = \frac{1}{n} \sum_{i=1}^n D_i = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)$ est l'estimateur sans biais et convergent de la moyenne différence m_d ,
- la distribution d'échantillonnage $\bar{D}_n$ de la différence est **supposée normalement distribuée** de moyenne $E(\bar{D}_n) = m_d = (m_x - m_y)$ et de variance $V(\bar{D}_n) = \sigma_d^2 = \sigma_{x-y}^2$ inconnues,
- $\bar{d}_n = \bar{x}_n - \bar{y}_n = \frac{1}{n} \sum_{i=1}^n d_i$, la moyenne calculée sur l'échantillon différence $(d_1, d_2, \dots, d_n)$ est une **estimation ponctuelle** de la moyenne différence m_d ,

Différence de 2 moyennes : Echantillons appariés

Echantillons dépendants - Données associées par paires

- $s_d^{*2} = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$ est une **estimation ponctuelle** de la variance σ_d^2 lorsque la moyenne m_d est inconnue,
- On établit cf. paragraphe "Estimation de la moyenne : m (**σ^2 inconnue**)", la statistique de test :

$$\frac{\bar{D}_n - m_d}{s_n^*/\sqrt{n}} \rightarrow T_{n-1} \text{ d.d.l.}$$

- On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de $m_d = (m_x - m_y)$:

$$\bar{d}_n - t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}} \leq m_d \leq \bar{z}_n + t_{\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$$

Exemple d'application 10

- On mesure 12 pièces avec des méthodes différentes. On a obtenu les résultats suivants :

$$\bar{x}_n = 1, \bar{y}_n = 2.08, SCE_x = 106.16, SCE_y = 118.19 \text{ et } SCE_{x-y} = 14.58.$$

Déterminer un intervalle de confiance de niveau 95% de la différence des deux méthodes de mesures.

Rapport de 2 variances

(comparaison de 2 variances)

- La comparaison de 2 populations normales peut porter non seulement sur leur valeur centrale (moyenne), mais également sur leur dispersion. La caractéristique de dispersion la plus utilisée est la variance.
- Rappelons qu'une des conditions d'application de la loi de Student dans le cas de comparaison de moyennes est que les échantillons proviennent de 2 populations normales de variances identiques : $\sigma_1^2 = \sigma_2^2$. Cette hypothèse peut être maintenant vérifiée à l'aide de l'intervalle de confiance du rapport des 2 variances : **Test d'égalité de 2 variances**.
- On suppose que l'on a prélevé deux échantillons indépendants de tailles n_1 et n_2 de deux populations normales $N(m_1 ; \sigma_1^2)$ et $N(m_2 ; \sigma_2^2)$ de paramètres inconnus.

Rapport de 2 variances

(comparaison de 2 variances)

- On sait déjà que :

$$\sum_{i=1}^{n_1} \frac{(X_i - \bar{X}_1)^2}{\sigma_1^2} = (n_1 - 1) \frac{S_1^{*2}}{\sigma_1^2} \rightarrow \chi_{(n_1-1)}^2 \text{ d.d.l.}$$

$$\sum_{i=1}^{n_2} \frac{(X_i - \bar{X}_2)^2}{\sigma_2^2} = (n_2 - 1) \frac{S_2^{*2}}{\sigma_2^2} \rightarrow \chi_{(n_2-1)}^2 \text{ d.d.l.}$$

- On peut alors montrer que la statistique de test :

$$\frac{\sigma_2^2}{\sigma_1^2} \frac{S_1^{*2}}{S_2^{*2}} \rightarrow F_{(n_1-1), (n_2-1)} \text{ d.d.l.}$$

- On en déduit, au niveau $(1 - \alpha)$, un intervalle de confiance pour le rapport $\frac{\sigma_2^2}{\sigma_1^2}$: $f_1 \frac{S_2^{*2}}{S_1^{*2}} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2 \frac{S_2^{*2}}{S_1^{*2}}$

où, $f_1 = f_{1-\frac{\alpha}{2}}$; $P(F_{(n_1-1, n_2-1)} > f_1) = 1 - \frac{\alpha}{2}$

et $f_2 = f_{\frac{\alpha}{2}}$; $P(F_{(n_1-1, n_2-1)} > f_2) = \frac{\alpha}{2}$

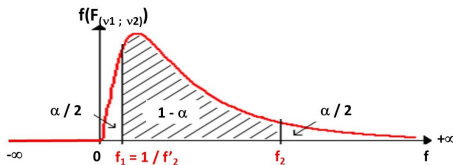
sont les fractiles de la de Fisher-Snédecour à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté (cf. table).

Loi de Fisher-Snedecor

- Propriété :

$$X = \frac{\frac{\chi_{v_1}^2}{v_1}}{\frac{\chi_{v_2}^2}{v_2}} \rightarrow F_{(v_1, v_2)} \quad ; \quad \frac{1}{X} = \frac{\frac{\chi_{v_2}^2}{v_2}}{\frac{\chi_{v_1}^2}{v_1}} \rightarrow F_{(v_2, v_1)}$$

- Lecture des fractiles f_1 et f_2 de la table de Fisher-Snedecor



- $F(f_2) = P(F_{(v_1, v_2)} \leq f_2) = 1 - \frac{\alpha}{2}$: f_2 se lit sur la table $F_{(v_1, v_2)}$.

- $F(f_1) = P(F_{(v_1, v_2)} \leq f_1) = \frac{\alpha}{2}$

$$\Rightarrow P\left(\frac{1}{F_{(v_1, v_2)}} \leq \frac{1}{f_1}\right) = 1 - \frac{\alpha}{2}$$

$$\Rightarrow P(F_{(v_2, v_1)} \leq \frac{1}{f_1} = f'_2) = 1 - \frac{\alpha}{2} \quad : \quad f'_2 = \frac{1}{f_1} \text{ se lit sur la table } F_{(v_2, v_1)},$$

on en déduit $f_1 = \frac{1}{f'_2}$.

Exemple d'application 11

- Reprenons l'exemple de la durée de vie moyenne de 2 types d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur de produits d'éclairage. Les essais effectués dans les mêmes conditions, sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants : La durée de vie d'une ampoule est supposée normalement distribuée. **On ne dispose d'aucune information sur les variances des deux populations.**

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Ecart	2400	2800

- Déterminer un intervalle de confiance de niveau 95% du rapport des variances des populations d'ampoules de ces deux fabricants.
- Question : Peut-on considérer l'égalité des variances $\sigma_2^2 = \sigma_1^2$?

Exemple d'application 12

- Reprenons l'exemple sur la rentabilité des banques : Un organisme financier a comparé les rendements annuels de deux types de banques, l'une privée, l'autre publique.

	1 : Privée	2 : Publique
Nombre d'essais	$n_1 = 8$	$n_2 = 10$
rendement moyen observé (%)	$\bar{x}_1 = 9.8$	$\bar{x}_2 = 6.9$
Somme des Carrés des Ecart	$SCE_1 = 250$	$SCE_2 = 61$

- Déterminer un intervalle de confiance de niveau 95% du rapport des variances des rendements annuels de ces deux banques.
- Question : Peut-on considérer l'égalité des variances $\sigma_2^2 = \sigma_1^2$?

Différence de 2 proportions

(Grands échantillons : $n_1 \geq 30$ et $n_2 \geq 30$)

- Il y a de nombreuses applications où nous devons décider si l'écart observé entre deux proportions échantillonnales est significatif ou s'il est plutôt attribuable au hasard de l'échantillonnage.
- Comme dans le cas de la comparaison de deux moyennes, on doit connaître la distribution d'échantillonnage de la différence $(P_1 - P_2)$ des deux proportions pour estimer, par intervalle de confiance, cette différence.
- On traite uniquement le cas où nous sommes en présence de **grands échantillons** prélevés au hasard et indépendamment de deux populations. Dans ce cas,
- La statistique de test :
$$\frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \hookrightarrow N(0; 1)$$
- D'où l'intervalle de confiance de niveau $(1 - \alpha)$ de $(p_1 - p_2)$:

$$(\hat{p}_1 - \hat{p}_2) - u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Différence de 2 proportions

(Grands échantillons : $n_1 \geq 30$ et $n_2 \geq 30$)

- On peut également supposer l'hypothèse d'égalité des proportions inconnues p_1 et p_2 à une **valeur commune p** ($p_1 = p_2 = p$) que l'on estime par $\hat{p}$ en combinant les proportions observées dans chaque échantillon comme suit :

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

- On peut donc aussi utiliser la statistique de test :

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}} \hookrightarrow N(0; 1)$$

- D'où l'intervalle de confiance de niveau $(1 - \alpha)$ de $(p_1 - p_2)$:

$$(\hat{p}_1 - \hat{p}_2) - u_{\frac{\alpha}{2}} \sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + u_{\frac{\alpha}{2}} \sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$$

Exemple d'application 13

- Dans deux municipalités avoisinantes, on a effectué un sondage pour connaître l'opinion des contribuables sur un projet d'aménagement d'un site. Les résultats de l'enquête se résument comme suit :

	Municipalité 1	Municipalité 2
Nombre de personnes interrogées	250	250
En faveur du projet	110	118

- 1 Quelle est l'estimation ponctuelle de la différence de proportions des contribuables de chaque municipalité favorisant l'aménagement du site ?
- 2 Déterminer l'intervalle de confiance de niveau $(1 - \alpha) = 95\%$ de contenir la valeur vraie de la différence des proportions, $(p_1 - p_2)$?
- 3 Question : Avec l'intervalle calculé en 2), est-ce que l'on rejeterait, au seuil de signification $\alpha) = 5\%$, l'hypothèse selon laquelle les contribuables des deux municipalités favorisent dans la même proportion l'aménagement du site sur leur territoire ?

Tests d'hypothèses paramétriques

- Les tests d'hypothèses constituent un aspect important de la statistique inférentielle dont le principe général peut s'énoncer comme suit :
- ❶ Soit un caractère d'une population dont la valeur du paramètre (moyenne, proportion, variance, etc.) est inconnue. Une hypothèse dite "nulle" ou H_0 est alors formulée sur ce paramètre inconnu, elle résulte de considérations théoriques, pratiques ou encore plus simplement basée sur un pressentiment.
- ❷ On veut porter un jugement sur cette hypothèse, sur la base des résultats d'un échantillon prélevé de cette population.
- ❸ Pour décider si l'hypothèse "nulle" formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue à partir de l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable aux fluctuations d'échantillonnages.

Tests d'hypothèses paramétriques

- La construction d'un test d'hypothèse consiste alors à déterminer entre quelles valeurs peut varier la statistique en **supposant vraie l'hypothèse nulle H_0** formulée.
- Les distributions d'échantillonnage des paramètres - **Statistiques de test** que nous avons traitées dans le chapitre précédent vont être particulièrement utiles dans l'élaboration d'un test statistique.
- Concepts importants dans l'élaboration d'un test statistique :
 - 1 **Hypothèse statistique** : est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une population.
 - 2 **Test d'hypothèse (ou test statistique)** : est une démarche qui a pour but de fournir une **règle de décision** permettant, sur la base de résultats d'échantillon, de faire un choix : **accepter ou refuser l'hypothèse nulle H_0** .

Tests d'hypothèses paramétriques

- Hypothèses statistiques :

- 1 Hypothèse nulle H_0 : hypothèse selon laquelle on fixe *a priori* un paramètre inconnu de la population à une valeur particulière.
- 2 Hypothèse alternative H_1 (ou contre-hypothèse) : n'importe quelle autre hypothèse qui diffère de l'hypothèse nulle H_0 .

- Remarque : c'est toujours l'hypothèse nulle qui est soumise au test et toute la démarche du test s'effectue en considérant ou supposant cette hypothèse H_0 comme vraie.

- Conclusion : Rejet ou Non-Rejet de l'hypothèse nulle H_0 , en aucun cas l'Acceptée. Si H_0 est rejetée, on accepte l'hypothèse alternative H_1 .

- Seuil de signification d'un test statistique : c'est le risque d'erreur α , consenti à l'avance de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie (favoriser alors l'hypothèse H_1) :

$$\alpha = P(\text{Rejeter } H_0 / H_0 \text{ vraie}) = P(\text{Choisir } H_1 / H_0 \text{ vraie})$$

Tests d'hypothèses paramétriques

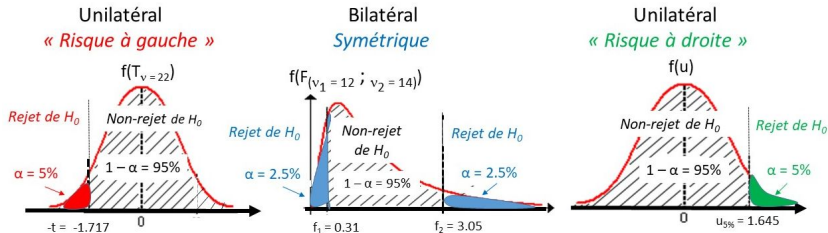
- Au seuil de signification correspond sur la distribution d'échantillonnage de la statistique une **région critique** de **Rejet H_0** : son aire correspond à la probabilité α .
- Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite **région d'acceptation** ou encore région de **Non-rejet de H_0** de probabilité $1 - \alpha$.
- La valeur observée de la statistique, **calculée toujours sous H_0** déduite des résultats de l'échantillon appartient, soit à la région de Rejet H_0 (on favorisera alors l'hypothèse H_1), soit à la région de Non-rejet de H_0 (on favorisera alors l'hypothèse H_0).
- Formulation des hypothèses - Types de tests : à partir d'exemples, nous allons résumer les différents types de test **"Bilatéral"** ou **"Unilatéral"** qui peuvent se présenter et schématiser les régions de rejet et d'acceptation de l'hypothèse nulle H_0 .

Remarques importantes

- 1 Pour un test bilatéral, les **2 valeurs critiques** (tables statistiques) sont des limites de la statistique qui conduisent au rejet de H_0 , selon le seuil de signification α choisi.
- 2 Un test unilatéral "risque à droite" ou "risque à gauche" ne comporte qu'**une seule valeur critique**.
- 3 Quelle que soit le type de test, l'hypothèse nulle H_0 comporte toujours **le signe égal** ($=, \leq, \geq$) et **spécifie la valeur du paramètre**.
- 4 l'hypothèse alternative H_1 est formulée en choisissant l'une ou l'autre des trois formes ($\neq, <, >$). On choisira la plus **pertinente** à la situation pratique analysée.
- 5 Dans la plupart des tests d'hypothèses, l'inégalité dans l'hypothèse H_1 dénote dans quelle **direction** est localisée la **région de rejet (critique)** de l'hypothèse H_0 .

Tests d'hypothèses paramétriques

● Tests - Hypothèses statistiques - Statistiques de test



$$\begin{cases} H_0 : m_1 \geq m_2 \Leftrightarrow m_1 - m_2 \geq 0 \\ H_1 : m_1 < m_2 \Leftrightarrow m_1 - m_2 < 0 \end{cases}$$

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \Leftrightarrow \frac{\sigma_2^2}{\sigma_1^2} = 1 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \Leftrightarrow \frac{\sigma_2^2}{\sigma_1^2} \neq 1 \end{cases}$$



$$I.C. 1 - \alpha = 95\% : \frac{\sigma_2^2}{\sigma_1^2}$$

$$\frac{\sigma_2^2}{\sigma_1^2} \frac{S_1^2}{S_2^2} \rightarrow F_{(v_1=12, v_2=14)}$$

$$\begin{cases} H_0 : p \leq p_0 = 10\% \\ H_1 : p > p_0 = 10\% \end{cases}$$

$$\frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow T_{v=22} d.d.l.$$

$$\frac{(\hat{P} - p)}{\sqrt{\frac{pq}{n}}} \hookrightarrow N(0; 1)$$

Valeur et probabilité critiques :

- La **valeur** (value en anglais) ou **valeur critique** désigne la valeur de la statistique de test **calculée sous l'hypothèse nulle H_0** .
- La **p-valeur** (p-value en anglais) ou **probabilité critique** associée à un test statistique :
 - ① Test unilatéral risque à gauche : $P(\text{Stat. Test} < \text{valeur}) = \text{p-valeur}$
 - ② Test bilatéral symétrique : $P(|\text{Stat. Test}| > \text{valeur}) = \text{p-valeur}$
 - ③ Test unilatéral risque à droite : $P(\text{Stat. Test} > \text{valeur}) = \text{p-valeur}$
- En pratique les logiciels de statistique fournissent la **valeur** et la **p-valeur** associées à un test donné.
- **Rejet** ou **Non-Rejet** de l'hypothèse nulle H_0 : il suffit de comparer la **p-valeur** fournie avec le seuil de signification ou **risque d'erreur α choisi**.
- Critère de décision : **rejet de H_0** si **p-valeur $\leq \alpha$** .
- Plus la **p-valeur** est **proche de 0**, plus forte est la contradiction entre l'hypothèse nulle H_0 et le résultat observé avec l'échantillon.

Démarche d'un test statistique

- Les principales étapes à suivre dans l'élaboration d'un test statistique :
 - 1 Hypothèses statistiques,
 - 2 Seuil de signification et conditions d'application du test,
 - 3 La statistique de test qui convient,
 - 4 Calcul de la statistique sous H_0 .
 - 5 Règle de décision et conclusion.

Risques d'un test statistique

- La règle de décision d'un test statistique comporte **2 risques ou types d'erreur possibles**. Ces deux risques varient en sens inverse : quand l'un augmente, l'autre diminue :
- ❶ **Risque de première espèce α** : c'est le seuil de signification α ; risque de rejeter à tort l'hypothèse nulle H_0 lorsque celle-ci est vraie :

$$\alpha = P(\text{Rejet } H_0 / H_0 \text{ vraie})$$

- ❷ **Risque de deuxième espèce β** : c'est le risque de ne pas rejeter l'hypothèse nulle H_0 alors que l'hypothèse H_1 vraie :

$$\beta = P(\text{Non Rejet } H_0 / H_1 \text{ vraie})$$

- Le calcul de β dépend toujours de la valeur posée pour le paramètre dans l'hypothèse H_1 , que l'on **suppose vraie**.
- Le graphique de β en fonction des diverses valeurs du paramètre posées en H_1 s'appelle la **courbe d'efficacité du test**.

Puissance d'un test statistique

- **Puissance du test** ($1 - \beta$) : c'est la probabilité complémentaire du risque de 2ème espèce β : le risque de rejeter l'hypothèse nulle H_0 alors que l'hypothèse H_1 vraie :

$$1 - \beta = P(\text{Rejet } H_0 / H_1 \text{ vraie})$$

Le graphique de $(1 - \beta)$ en fonction des diverses valeurs du paramètre posées en H_1 s'appelle la **courbe de puissance du test**.

- Exemple : En contrôle industriel, **le risque 1ère espèce α** correspond au risque pris par le **producteur** (ou le fournisseur) alors que le **risque de 2ème espèce β** correspond au risque pris par le **consommateur** (ou le client).

Risques d'un test statistique

- Les risques liés aux tests d'hypothèses peuvent se résumer comme suit :

Conclusion du test	Réalité : H_0 vraie	Réalité : H_1 vraie
Décision : Non-rejet H_0	bonne : $1 - \alpha$	mauvaise : β
Décision : Rejet de H_0	mauvaise : α	bonne : $1 - \beta$

Exemple d'application

- Calcul et représentation graphique des risques α et β , du niveau de confiance $(1 - \alpha)$ et de la puissance du test $(1 - \beta)$:
- Un procédé de remplissage est ajusté de telle sorte que les contenants pèsent en moyenne 400g. Le poids des contenants est supposé normalement distribué avec un écart-type de 8g.

On a prélevé un échantillon de 16 contenants pour vérifier si le procédé de remplissage se maintient en moyenne à 400g, on opte pour la règle de décision suivante :

$$\begin{cases} \text{si } 396,08 \text{ g} \leq \bar{X}_{16} \leq 403,92 \text{ g} & \text{le processus opère correctement} \\ \text{Sinon} & \text{arrêter le processus de remplissage} \end{cases}$$

Exemple d'application

- 1 Quelles sont les hypothèses statistiques que l'on veut tester avec cette méthode de contrôle ?
- 2 Déterminer la probabilité de commettre une erreur de première espèce.
- 3 Lors d'un récent contrôle, on a obtenu, pour un échantillon de 16 contenants, un poids moyen de 395g. Doit-on poursuivre ou arrêter la production ?
- 4 Quelle est la probabilité de commettre une erreur de deuxième espèce selon l'hypothèse alternative $H_1 : m = 394 \text{ g}$?
- 5 Avec ce plan de contrôle, quelle est la probabilité de rejeter l'hypothèse selon laquelle le procédé opère à 400g, alors qu'en réalité il opère à 394g ?
- 6 Faire de même pour les valeurs suivantes sous $H_1 : m = 395\text{g}, 396\text{g}, 397\text{g}, 398\text{g}, 399\text{g}$ et 400g. Tracer la courbe d'efficacité du test.

Exemple d'application 12

- Une entreprise fournit à un client des tiges d'acier. Le client exige que les tiges aient en moyenne, une longueur de 29 mm. La longueur des tiges est supposée normalement distribuée. On veut vérifier si le procédé de fabrication opère bien à 29 mm. Un échantillon aléatoire de 12 tiges provenant de la fabrication donne une longueur moyenne de 27,25 mm et un écart-type empirique de 2,97 mm.

Doit-on conclure, au seuil $\alpha = 5\%$, que la machine est dérégulée ?

Exemple d'application 13

- Aux dernières élections, un parti politique a obtenu 42% des suffrages. Un récent sondage a révélé que, sur 1041 personnes interrogées en âge de voter, 458 accorderaient son appui à ce parti.

Le secrétaire général du parti a déclaré que la popularité de son parti est en hausse. Que penser de cette affirmation au seuil de signification $\alpha = 5\%$, ?

Exemple d'application 14

- Le responsable de la production suggère à ses clients des tiges d'acier avec un nouvel alliage. Il semble que ceci permettrait d'obtenir une résistance à la rupture plus élevée. Les résultats d'un test de résistance à la rupture de 50 tiges avec et sans le nouvel alliage se résument comme suit.

	Sans le nouvel alliage	Avec le nouvel alliage
Nombre de tiges	50	50
Résistance moyenne	600.50	605.00
Variance empirique	148.50	137.61

Au seuil de signification $\alpha = 5\%$, est-ce que l'hypothèse selon laquelle la résistance moyenne à la rupture sans l'alliage est moins élevée que celle avec l'alliage est confirmée ?

Exemple d'application 15

- Un industriel utilise des pièces de deux constructeurs différents. Après six mois d'utilisation, il constate que sur les 80 pièces du constructeur 1, 50 ne sont jamais tombées en panne, alors que pour les 60 pièces réceptionnées du constructeur 2, la proportion est de l'ordre de 2 pièces sur 3.

Au seuil de signification $\alpha = 5\%$, peut-on considérer que les proportions de pièces qui ne sont jamais tombées en panne de ces deux constructeurs sont équivalentes ?

Exemple d'application 16

On a relevé les prix de deux titres du CAC40 du secteur des assurances des dix dernières années. Les prix de ces titres - actions exprimés en euros (€), sont supposés normalement distribués. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Titre	AGF	AXA
Taille de l'échantillon	10	10
Prix moyen observé €	29.83	26.97
Ecart-type (N) empirique	0.21	0.63

- 1 Les échantillons sont-ils indépendants ?
- 2 Peut-on affirmer avec un risque d'erreur $\alpha = 5\%$, qu'en moyenne, le titre AGF est plus cher que le titre AXA ?